DOCUMENTA MATHEMATICA

Extra Volume

PROCEEDINGS OF THE INTERNATIONAL CONGRESS OF MATHEMATICIANS

Berlin 1998, August 18 - 27

Volume I : Plenary Lectures and Ceremonies (662 pages)

Volume II : Invited Lectures (881 pages)
Invited Forty-Five Minute Lectures at the Section Meetings, Sections 1 - 9

Volume III : Invited Lectures (825 pages)
Invited Forty-Five Minute Lectures at the Section Meetings, Sections 10 - 19

ICM 1998
Contents of Volumes I, II, and III

The Work of the Fields Medalists and of the Rolf Nevanlinna Prize Winner

Invited One-Hour Plenary Lectures

LECTURES BY THE FIELDS MEDALISTS AND BY THE ROLF NEVANLINNA PRIZE WINNER

APPENDIX: INVITED FORTY-FIVE MINUTE LECTURES AT THE SECTION MEETINGS

This appendix contains three manuscripts of Invited Speakers which are not included in Volume II or III

Invited Forty-Five Minute Lectures at the Section Meetings

Contents of Volumes II and III

In case of several authors, Invited Speakers are marked with a *.
The author index is at the end of each of these two volumes.

Section 12. Probability and Statistics

Section 13. Combinatorics

Section 14. Mathematical Aspects of Computer Science

Section 1

Logic

In case of several authors, Invited Speakers are marked with a *.

Generic Large Cardinals: New Axioms for Mathematics?

Matthew Foreman

Abstract. This article discusses various attempts at strengthening the axioms for mathematics, *Zermelo-Fraenkel Set Theory with the Axiom of Choice* It focuses on a relatively recent collection of axioms, *generic large cardinals*, their success at settling well known independent problems and their relations to other strengthenings of ZFC, such as large cardinals.

1991 Mathematics Subject Classification: 3,4,5,28
Keywords and Phrases: axioms, large cardinals, ideals, generic large cardinals

Introduction. While the standard axiomatization of mathematics *Zermelo-Fraenkel Set Theory with the Axiom of Choice* (ZFC) has been extremely successful in resolving the foundational issues that arose at the turn of the century, it has some shortcomings. These shortcomings are largely due to its inability to settle various natural problems.

Most prominent among these problems are Hilbert's 1^{st} problem (the Continuum Hypothesis), and issues having to do with the use of the Axiom of Choice. The development of Forcing, in the early 1960's, led to independence results in most areas of mathematics that have a strong infinitary character, particularly including measure theory and other parts of analysis, infinite group theory, topology and combinatorics.

This paper surveys some of these independence results and the attempts at finding new axiom systems to settle these questions. It will focus on a technique that arose naturally in relating large cardinals with combinatorial and descriptive set theoretic properties of sets of size (roughly) the continuum. This technique generated plausible properties of the universe. Taken as axioms they settle most of the important independent statements of mathematics.

Without further explanation, the first few uncountable cardinals are $\aleph_1, \aleph_2, \ldots, \aleph_n, \ldots$ and the first uncountable limit cardinal and its successor are $\aleph_\omega$ and $\aleph_{\omega+1}$. The natural numbers will be denoted alternately as $\boldsymbol{N}$ or more commonly ω, the first limit ordinal. The cardinality of the real numbers will be referred to as $\boldsymbol{c}$, and the cardinality of the power set of a set X as 2^X. In particular, $2^\omega = \boldsymbol{c}$. If λ is a cardinal, $n \in \boldsymbol{N}$, then λ^{+n} will be the n^{th} cardinal past λ. Lapsing into the jargon of subfield, I will refer to the mathematical universe as V. (Due to space limitations the author has not attempted to credit appropriate authors, particularly for well-known results.)

INDEPENDENCE RESULTS. Gödel's theorems ([5]) show that any consistent axiom system $\mathcal{A}$ sufficiently strong to encompass elementary number theory and sufficiently concrete to be recognized as an axiom system (i.e. $\mathcal{A}$ is recursively enumerable) must be *incomplete*. This means that there are statements φ such that there are examples of mathematical structures satisfying the axiomatization $\mathcal{A}$ that satisfy φ and examples of structures that satisfy $\mathcal{A}$ and satisfy the negation of φ. (A simple analagous situation is that the property of being *abelian* is independent of *Group Theory* because there are examples of abelian and non-abelian groups.) Further, Gödel gave a uniform method of producing such a φ: it is a number-theoretic statement equivalent to the consistency of $\mathcal{A}$.

After the shock of this result wears off the question arises as to whether there are statements of "ordinary mathematics" that are independent of the standard axioms of set theory. On one level the answer is clearly affirmative: Matijasevič ([14]), using results of Davis, Putnam and Robinson ([2]), showed that every recursively enumerable set of natural numbers is the range of a diophantine polynomial (of several variables) applied to the natural numbers. (This gave a solution to Hilbert's 10^{th} problem. ([7])) Since the collection of inconsistencies of a recursively enumerable axiom system $\mathcal{A}$ can be coded canonically as a recursively enumerable set of natural numbers, the consistency of $\mathcal{A}$ is equivalent to the non-existence of a natural number solution to a particular diophantine equation. If we fix $\mathcal{A}$ to be our (consistent) axiom system, such as ZFC (or ZFC with large cardinals) we find that there is a diophantine equation such that the (non-)existence of an integer solution to this diophantine equation is independent of $\mathcal{A}$.

Mathematical problems that arose from motivations outside mathematical logic itself eventually were seen to be independent. The most famous of these is Hilbert's 1^{st} problem: the Continuum Hypothesis. The Continuum Hypothesis (or CH) is the statement that the real numbers have cardinality the first uncountable cardinal. Equivalently $\boldsymbol{c} = \aleph_1$. Another equivalent statement is that every infinite subset of the real numbers is either countable or has cardinality $\boldsymbol{c}$.

Gödel ([6]) discovered a canonical example of the axioms of ZFC, called the *Constructible Universe*, L. The idea behind this example is that it is built using only concrete operations, with the only non-constructive elements being the infinite ordinals in the domain of these functions. Gödel showed that if the Zermelo-Fraenkel axioms hold, then the Continuum Hypothesis held in L along with the controversial *Axiom of Choice*. Hence Gödel showed that if the Zermelo-Fraenkel axioms are consistent, then they are consistent with the Continuum Hypothesis and the Axiom of Choice.

An important breakthrough came with the advent of *Forcing* in 1963, in a paper of Cohen ([1]). In this paper, Cohen gave a general method of building new examples of ZF from old ones. (In some ways the method is analogous to adding an algebraic element to a field.) Cohen used this method to show that the Axiom of Choice and the Continuum Hypothesis are independent of ZF.

Forcing, as developed by Solovay and others, became a primary tool for showing independence results. Among the most prominent statements shown to be independent of ZFC:

- *Most statements of infinitary cardinal arithmetic such as the Generalized*

Continuum Hypothesis and the Singular Cardinals Hypothesis.

• *The existence of a* Suslin line*, a complete linear ordering with no uncountable collection of disjoint open intervals that is* NOT *isomorphic to the real line.* After this came an extensive body of work showing independence results in many parts of point-set topology.

• *The independence of the existence of a non-Lebesgue measurable set from* ZF + The Axiom of Countable Choice. This shows that the existence of a non-measurable set is inherently tied up with the use of a non-constructive uncountable set existence principle.

• *The existence of a non-free Whitehead group.* This result and related techniques led to a plethora of independence results in abelian groups and homological algebra.

• *The existence of a discontinuous homomorphism between Banach Algebras.*

• *Many infinitary combinatorial principles, particularly in infinitary Ramsey Theory.*

• *The existence of a locally finite group action on a measure space X with a unique invariant mean (positive linear functional of norm 1).*

• *The existence of a paradoxical decomposition of the sphere S^2 constructed using $\mathcal{G}_\delta$ and $\mathcal{F}_\sigma$ subsets of $\boldsymbol{R}^5$ and the operations of complement and projection.*

Is there a meaningful way of settling these questions? Is there anything more to say after they have been shown to be independent of ZFC?

A potential response is to suggest that whatever process led to the acceptance of ZFC as an axiomatization for mathematics (despite its controversial beginnings) may lead to other assumptions that settle, or partially settle most of the problems we are interested in.

The axiom V=L. Jensen ([8, 9]) realized that Gödel's Constructible Universe had a "fine structure" that made it amenable to the kind of close study that settles the types of problems mentioned above. Moreover, he discovered a technique, that when applied with suitable cleverness, appears to answer essentially any question about L. As part of this work, he discovered various combinatorial principles such as $\square_\kappa$ and $\diamond_\kappa$ that are highly applicable in domains beyond L.

While the axiom of constructibility is very effective, most people working in set theory reject it as inappropriate. This is primarily because the axiom saying "every set is constructible" is viewed as *restrictive* and thus does not account for all of the possible behavior of sets or other mathematical objects.

Further, in the constructible universe there are "pathologies" such as easily constructible paradoxical decompositions of the sphere.

Determinacy Axioms. The *Axiom of Determinacy*, proposed by Mycielski and Steinhaus ([13]) is a nonconstructive existence principle that contradicts the Axiom of Choice. It makes sense however, to assert it in limited domains such as the collection of Projective Sets or in the smallest model of ZF containing all of the real numbers. These assertions do not ostensibly contradict the Axiom of Choice for the class of all sets.

Given a set A contained in the unit interval $[0, 1]$ one can associate a game G_A where players I, II alternate playing a sequence of digits $n_0, n_1, n_2, \ldots$. (Each $n_i \in \{0, 1, \ldots, 9\}$.) The resulting play yields a number a in the unit interval whose

decimal expansion is $a = .n_0 n_1 n_2 \ldots$. We declare player II the winner if $a \in A$. The assertion that A is determined is the assertion that either player I or player II has a winning strategy in G_A. A collection $\mathbf{\Gamma}$ of subsets of $\boldsymbol{R}$ is said to be determined iff every element $A \in \mathbf{\Gamma}$ is determined.

Martin [11] showed that all Borel sets are determined. However, in L there is a subset of the real line that is the projection of a Borel set in the plane that is not determined using strategies in L. Hence one can go no further in ZFC.

Why are determinacy axioms attractive? Asserting determinacy for reasonably robust classes $\mathbf{\Gamma}$ implies that every element of $\mathbf{\Gamma}$ is nicely regular, e.g. is Lebesgue measurable, has the Property of Baire and uniformization holds in the relevant guise. So, for example, asserting determinacy for projective sets implies that there is no paradoxical decomposition using projective sets. (Projective sets are the subsets of $\boldsymbol{R}^n$ constructed from Borel sets in higher dimensions using the operations of projection and complement.)

The drawbacks of determinacy are twofold. First off, it says nothing about sets that are not in its domain. For example, while determinacy in $L(\boldsymbol{R})$ tells you that there is no Suslin line in $L(\boldsymbol{R})$, it says nothing about the *actual* existence of a Suslin line. Secondly, there appears to be no extrinsic motivating heuristic for determinacy. Its appeal and force lie in its effectiveness and the body of coherent, predictable consequences.

LARGE CARDINALS. The other main source of new axioms for the mathematical universe is a collection of ideas called *large cardinals*. These axioms were generated by intuitions about "higher infinities", sets whose relation to smaller sets were roughly similar to the relation between $\boldsymbol{N}$ and finite sets.

Another motivation for large cardinals is the idea of *reflection*: the set formation process has no natural stopping point, for at such a point we would simply take the union of all sets constructed and form a new set. Hence any property that holds in the mathematical universe should hold of many set-approximations of the mathematical universe. Moreover, since this is a property of the universe, there should be many sets that, in turn, have this property relative to smaller sets, etc. The sets that have the reflection properties relative to smaller sets are the large cardinals.

Eventually large cardinal axioms came to be stated more or less uniformly as the existence of certain kinds of symmetries. Technically these are elementary embeddings j from the universe V to transitive classes M. (An embedding is *elementary* iff for all properties ϕ and all $a_1, \ldots a_n$, if ϕ holds of $a_1, \ldots, a_n$, then ϕ holds of $j(a_1), \ldots j(a_n)$. So, e.g., if X is a manifold, $j(X)$ is a manifold.)

These axioms vary in strength according to where j sends ordinals and the closure of the class M. (We can classify M according to the least cardinality of a set X such that $X \notin M$. A theorem of Kunen proves that there always is such a set.) An important ordinal is the smallest ordinal moved by j, called the *critical point* of j, or $crit(j)$.

A well-known example of such an axiom was proposed by Ulam; the axiom of a *Measurable Cardinal*. Ulam formulated this as the statement that there is a set K and a countably additive 2-valued measure defined on all subsets of K. Using ultraproducts, this can be stated in modern language as the existence of a

non-trivial elementary embedding of V to some transitive model M with critical point κ.

The notion of supercompact and huge cardinals can also be stated as the existence of measures on sets with certain additional structure. The statement in terms of elementary embeddings is more conceptual:

Definition. A cardinal κ is λ-*supercompact* iff there is an elementary embedding $j : V \to M$, where M is a transitive class and M contains every λ sequence of ordinals. κ is supercompact iff κ is λ-supercompact for all λ.

A cardinal κ is n-*huge* iff there is an elementary embedding $j : V \to M$ with critical point κ such that M is closed under $j^n(\kappa)$-sequences.

For each elementary embedding there is an ideal object in the target model M (or system of ideal objects, in more sophisticated set ups) that determine the nature of the embedding. In particular, it determines the closure of M. Each element ι of M determines a measure and with respect to this measure every property of the ideal object holds at almost every point in the measure space determined by ι. In particular, if $S \subset \iota = crit(j)$ is stationary, then for almost every $\alpha < \iota, S \cap \alpha$ is stationary. If ι is taken to be the ideal point, then the ultraproduct of V by the measure determined by ι yields the model M.

By focusing on the ideal points one can see the reflection implied by the elementary embedding. An important example of such reflection is the statement that if κ is supercompact and $\lambda > \kappa$ is a regular cardinal then every stationary subset of λ reflects to an ordinal of cofinality less than κ. This property, while useful in its own right as a construction principle, contradicts $\square$.

Large cardinals are also significant in that many of the combinatorial properties of $\boldsymbol{N}$ hold at large cardinals. For example Rowbottom's Theorem, a direct analogue of Ramsey's theorem, states that if κ is measurable then every partition of the finite subsets of κ into less than κ colors has a homogeneous set of size κ. Baumgartner and Hajnal showed that strong partition properties hold at the cardinal successor of ω. Recent results of Hajnal and the author show that analogous partition properties hold at the successor of a measurable cardinal.

Results of Ulam (and later Tarski and Keisler) showed that large cardinals, such as measurable cardinals, must be inacessibly larger than most ordinary mathematical objects, such as the real numbers $\boldsymbol{c}$. (Recent results of Gitik and Shelah show that if $\mathcal{I}$ is a countably complete ideal on a cardinal such as $\boldsymbol{c}$ (or $P(\boldsymbol{c})$) then $P(\boldsymbol{c})/\mathcal{I}$ does not have a dense countable set; the least possible density is $\aleph_1$.)

Gödel suggested that large cardinal assumptions may eventually be a route to settling the continuum hypothesis. This hope was dashed however by a theorem of Levy and Solovay ([10]) that showed that "small forcing" does not affect large cardinals. In particular the Continuum Hypothesis is independent of of any large cardinal assumption. This theorem and the apparent remoteness of these cardinals to ordinary sets is a major drawback of large cardinal assumptions.

Large cardinals do have a coherent motivating heuristic and independent affirming intuitions. They have also proved essential for relative consistency results, such as the failure of the singular cardinals hypothesis. (e.g. Jensen's *Covering Lemma* showed that large cardinals were strictly necessary.)

GRAND UNIFICATION. In the 70's and early 80's large cardinal axioms and determinacy axioms were viewed as competing attempts at extending the axioms ZFC. Martin and Harrington had showed various connections between some of the weaker versions of the two systems of axioms, but the exact relationships weren't clear.

An important breakthrough came in 1984 ([3]), when it was realized that large cardinal axioms implied the existence of large cardinal type embeddings, where the embedding $j : V \to M$ was definable not in V, but in a forcing extension of V. These elementary embeddings have critical point $\aleph_1$, and thus the embeddings are immediately relevant to "small" sets such as the real numbers. Moreover, this discovery uncovered a new class of relatively weak large cardinals, the *Woodin cardinals*. (Named after the person who isolated the definition.)

Following on the heels of this discovery, Martin and Steel ([12]) showed that determinacy for the class of projective sets follows from the existence of sufficiently many Woodin cardinals. Woodin ([16]), using the generic large cardinal embeddings, showed that determinacy held for all sets in $L(\boldsymbol{R})$. In particular, all of the consequences of determinacy follow from large cardinals.

More recent work has exactly fixed many of the relations between large cardinal and determinacy axioms, often showing that a particular large cardinal axiom implies determinacy of a class of sets $\boldsymbol{\Gamma}$, which in turn implies the consistency of a slightly weaker large cardinal axiom.

This close relationship has become a major feature of the contemporary study of other extensions of ZFC. By and large they are all known to either follow from, or be equiconsistent with large cardinal axioms. This is viewed by many people as being suggestive that the various alternative axiom systems suggested are simply different aspects of the same phenomenon, hence confirming large cardinal axioms.

Despite this type of confirmation and large cardinals' role of calibrating the consistency strength of most independent propositions of ZFC, it remains frustrating that they cannot actually settle important problems such as the Continuum Hypothesis.

GENERIC LARGE CARDINALS. Generic large cardinals are a marriage of large cardinals and forcing. The axioms assert the existence of an elementary embedding $j : V \to M$, where M is a transitive model, where j is definable in a forcing extension of the universe $V[G]$. These embeddings can be viewed as *virtual* versions of large cardinal embeddings, whose specifics are revealed by forcing with the appropriate partial ordering. (This technique was first used by Solovay. Jech and Prikry, realizing its interest, isolated the notion of a *precipitous* ideal.)

The advantage of generic large cardinals is that the critical point of j can be a "small" cardinal such as $\aleph_1$. With some limitations this allows these cardinals to have similar reflection and resemblance properties as posited by large cardinal axioms on highly inacessible cardinals. Moreover, it allows one to state "symmetry principles" that can hold in a generic extension of the universe. By and large the motivational principles used to generate large cardinals can be restated to apply to generic large cardinal axioms, virtually verbatim.

The current study of generic large cardinal axioms now breaks into three parts: their consequences as axioms, showing their consistency relative to large cardinals

and showing that they imply the existence of inner models with large cardinals. Many research programs in the area combine one or more of these parts. In a typical example, relative consistency results of properties of $\aleph_2$ can be shown by first establishing that they follow from a generic large cardinal property and then showing that the property is consistent relative to a conventional large cardinal. They can be used in the other direction as well; an archetypical result in the area, shown by Solovay, is that the existence of a real-valued measurable cardinal implies the existence of an inner model with a 2-valued measurable cardinal. This was done by first showing that a real-valued measurable cardinal implied the existence of a generic large cardinal, which in turn implied the existence of an inner model with a measurable cardinal.

The parameters involved in determining a generic large cardinal are expanded to include the nature (in particular the density or saturation) of the partial ordering $\boldsymbol{P}$ involved in the forcing. Analogously to large cardinals, the transitive model M typically contains an ideal object, ι, whose existence implies the closure of the model M. Rather than determining a measure, this ideal object determines an ideal $\mathcal{I}$ in the ground model V on any set Z such that $\iota \in j(Z)$. Most of the relevant properties of $\boldsymbol{P}$ (particularly the stronger properties such as saturation) are inherited by the Boolean algebra $P(Z)/\mathcal{I}$, and hence we primarily discuss the saturation and density properties of $\mathcal{I}$ (or more properly $P(Z)/\mathcal{I}$.) We will refer to embeddings as generically huge, or generically λ-supercompact if the closure of M corresponds to the analogous large cardinal property. To simplify statements of theorems, we will often neglect the optimal hypothesis.

The first result is that if there is a generic huge embedding such that $j(\boldsymbol{c}) = 2^{\boldsymbol{C}}$, defined in the simplest possible forcing extension, then the continuum hypothesis holds and there is a Suslin line:

•(Foreman) *Suppose that there is a normal and fine $\aleph_1$-dense ideal on the collection of subsets of $2^{\boldsymbol{C}}$ of cardinality $\boldsymbol{c}$. Then the continuum hypothesis holds and there is a Suslin line. (Woodin has reduced the hypothesis of the first assertion to the existence of an $\aleph_1$-dense ideal on $\aleph_2$.)*

To extend this to the GCH, there are several possible axioms, one that stresses the resemblance between successor cardinals is the hypothesis of the following theorem:

•(Foreman) *Suppose that for all regular λ, $n \in \boldsymbol{N}$ there is a generic huge embedding sending $\aleph_{k+1}$ to λ^{+k} $(k \leq n)$. Then the Continuum Hypothesis implies the Generalized Continuum Hypothesis.*

Just as large cardinals imply stationary set reflection, generic large cardinals do as well. Magidor showed (in a different guise) that if for all n, $\aleph_n$ is generically supercompact by $\aleph_{n-1}$-closed forcing then every stationary subset of $\aleph_{\omega+1}$ reflects. Since Jensen's $\square$ implies the existence of non-reflecting stationary sets, generic embeddings imply the failure of $\square$. However, there are variations of $\square$, that while strictly weaker, are nearly as useful. The strongest of these is $\square_{\kappa,\omega}$. The following theorem shows that it is possible to have some of the best of $\square$ and stationary set reflection.

•(Cummings, Foreman, Magidor) *Suppose that there is an example of set theory with infinitely many supercompact cardinals. Then there is an example of*

set theory where every stationary subset of $\aleph_{\omega+1}$ *reflects and where* $\square_{\kappa,\omega}$ *holds.*

The proof of this theorem uses generic supercompactness in a subtle way. Magidor and the author showed that generic supercompactness by countably closed forcing is incompatible with Weak Square ([4]). Instead, in this proof, each $\aleph_n$ is generically supercompact by a closed forcing notion in a stationary set preserving extension of V.

As one might expect, generic large cardinals have implications for other topics in the theory of singular cardinals, such as the "PCF" theory developed by Shelah. For example, if there is a generic huge embedding, sending $\aleph_1$ to $\aleph_{\omega+1}$, then there is no "Good Scale" in the sense of the PCF theory. The flow goes the other way as well; using PCF theory one can show that there is no "generic ω-huge cardinal", an analogue to a result of Kunen for ordinary large cardinals.

Generic large cardinals have similar effects on Ramsey Theory as large cardinals:

•(Foreman, Hajnal) *Suppose that there is an* $\aleph_1$*-dense ideal on* $\aleph_2$*. Then the partition property* $\omega_2 \to (\omega_1^2 + 1, \alpha)$ *holds for all* $\alpha < \omega_2$*.*

Generic large cardinal axioms have other combinatorial consequences. For example the existence of generic huge embeddings with simple forcing notions imply that every graph on $\aleph_n$ with infinite chromatic number has subgraphs of all smaller infinite chromatic numbers (and these subgraphs have the same finite subgraphs as the original graph.)

One can postulate other properties of the forcing $\boldsymbol{P}$. Suppose that κ is a regular cardinal. Say that $\boldsymbol{P}$ is κ*-tame* if $\boldsymbol{P}$ is a regular subalgebra of the partial ordering for adding a Cohen subset of a cardinal less than κ followed by a product of κ-closed and strongly κ-c.c. partial orderings. Mitchell showed that it is consistent for $\aleph_2$ to be generically weakly compact by an $\aleph_1$-tame partial ordering. Abraham improved this to two consecutive cardinals.

•(Cummings, Foreman) *Suppose that it is consistent for there to be infinitely many supercompact cardinals. Then it is consistent that for all* $n \geq 2$*,* $\aleph_n$ *is generically weakly compact by an* $\aleph_{n-1}$*-tame* $\boldsymbol{P}$*. Moreover, this implies that for all* $n \geq 2$*, there is no Aronszajn tree on* $\aleph_n$*.*

These have applications in other parts of mathematics where infinitary combinatorics plays a role. As an example we consider the case of a vector space X over a field F, with a symmetric bilinear form ϕ. If we choose a basis $\{x_\alpha : \alpha < \kappa\}$ for X and let $X_\alpha = \text{span}\{x_\beta : \beta < \alpha\}$ we can consider $\Gamma(X, \phi) = \{\alpha : X = X_\alpha \oplus X_\alpha^\perp\}$. This set is invariant under isomorphism modulo the non-stationary ideal on κ. (This is called the Γ*-invariant.*) It makes sense to ask which sets can arise this way.

•(Foreman, Spinas) *Suppose that* $\aleph_2$ *is generically weakly compact by an* $\aleph_1$*-tame partial ordering. Then there is a subset of* $\aleph_2$ *that is not the* Γ*-invariant of any* (X, F, ϕ)*.*

In addition to the role of generic large cardinal axioms in the unification of the axiom systems of large cardinals and determinacy, Woodin has shown directly that they imply determinacy:

•(Woodin) *The axiom of determinacy in* $L(\boldsymbol{R})$ *is equiconsistent with "ZFC + there is an* $\aleph_1$*-dense ideal on* $\aleph_1$*."*

There are many open problems about which generic large cardinals can be shown to be consistent from large cardinals. However much progress has been made. A partial listing of such results includes:

•(Woodin, improving results of Kunen, Laver and Magidor) *Let $n \in \boldsymbol{N}$. Assuming the consistency of an almost-huge cardinal, it is consistent that there is an $\aleph_n$-complete, $\aleph_n$-dense ideal on $\aleph_n$.*

•(Foreman) *Assuming the consistency of a huge cardinal, it is consistent that for all regular κ, there is a κ^+-saturated ideal on κ.*

•(Foreman) *Assuming the consistency of a 2-huge cardinal, then for all n, it is consistent that there is a generic 2-huge embedding with critical point $\aleph_n$.*

•(Foreman) *Assuming the existence of a huge cardinal, it is consistent that there is a countably complete, uniform $\aleph_1$-dense ideal on $\aleph_2$.*

•(Steel-Van Wesep from determinacy assumptions, Foreman, Magidor and Shelah from large cardinal assumptions with Shelah proving the optimal theorem) *Assuming the consistency of a Woodin cardinal, it is consistent that the non-stationary ideal on $\aleph_1$ is $\aleph_2$-saturated.*

It is also possible to show that the generic large cardinal axioms form a hierarchy in consistency strength. A typical theorem includes:

•(Foreman) *Let $n > 1$. Suppose that there is a generic n-huge embedding by the partial ordering $Col(\omega, \aleph_1)$. Then it is consistent to have a generic (n-1)-huge embedding with partial ordering $Col(\omega, \aleph_1)$.*

Further it is possible, in certain cases to show from generic embeddings that large cardinals are consistent. For example:

•(Steel) *Suppose that there is a saturated ideal on $\aleph_1$ and a measurable cardinal, then there is an inner model with a Woodin cardinal.*

Using naive technology one can show that the existence of certain generic elementary embeddings imply inner models with huge cardinals. Using this fact, one can find strong Chang's conjecture principles of the $\aleph_n$'s that lie strictly between a huge cardinal and a 2-huge cardinal.

With the exception of the results mentioned in the next section, generic large cardinals give a coherent theory that settles most of the classical independent statements of mathematics. Many are known to be consistent relative to conventional large cardinals. Are all principles generated this way consistent? Are they consistent with each other? It turns out that there are non-trivial restrictions on the saturation properties of various natural ideals.

Most prominent among these are the results of Shelah, and Shelah and Gitik. Shelah's theorem states that if $\mathcal{I}$ is a saturated ideal on κ^+, then the collection of ordinals of cofinality different from the cofinality κ is an element of $\mathcal{I}$; in particular, if $\kappa > \omega$, the non-stationary ideal on κ^+ is not saturated. Shelah and Gitik showed that the non-stationary ideal on the successor of a singular cardinal κ is not saturated, even when restricted to the points having cofinality equal to the cofinality of κ. The following theorem extends work of Burke and Matsubara.

•(Foreman, Magidor) *Suppose that $\kappa < \lambda, \aleph_1 < \lambda$. Then the non-stationary ideal on $P_\kappa(\lambda)$ is not λ^+ saturated.*

Finally it is possible to show that the limitations on the closure of the target model M for a generic elementary embedding are roughly similar as they are for

conventional large cardinals.

- *There is no $\aleph_\omega$-saturated ideal on the subsets of $\aleph_\omega$ of order type $\aleph_\omega$.*

MARTIN'S MAXIMUM AND P-MAX. In [3], Magidor, Shelah and the author formulated a principle called *Martin's Maximum* and showed that it implied that the non-stationary ideal on $\aleph_1$ is $\aleph_2$-saturated, and the singular cardinals hypothesis holds.

Woodin showed (assuming a mild large cardinal hypothesis) that if the non-stationary ideal on $\aleph_1$ is $\aleph_2$-saturated then there is a fairly concrete surjection $\rho : \boldsymbol{R} \to \aleph_2$. Further, he developed a canonical theory "P-max" to describe the sets of hereditary cardinality $\aleph_1$, and showed that this theory is canonical and robust in many ways. Further it has a close connection with Martin's Maximum and its variants such as MM^+ and MM^{++}.

As of this writing, this theory appears to be particular to $\aleph_1$, as the results in the previous section (and others) show that it is inconsistent for the non-stationary ideal on $\aleph_1$ to be saturated and have an $\aleph_1$-preserving generic elementary embedding with critical point $\aleph_2$.

REFERENCES

[1] Cohen, P. *The independence of the Continuum Hypothesis.* I. PNAS 50(1963), 1143–1148.

[2] Davis, M., Putnam, H., Robinson, J. *The decision problem for exponential diophantine equations*, Annals of Mathematics 74(1961) 425–436.

[3] Foreman, M., Magidor, M., Shelah, S. *Martin's Maximum, Saturated Ideals and Non-regular Ultrafilters, Part I.* Annals of Mathematics, 127(1988), 1–47.

[4] Foreman, M., Magidor, M. *A Very Weak Square Principle.* Jour. Symb. Log., 62(1)(1997), 175–196.

[5] Gödel, K. *Über formal unentscheidbare Sätze der* Principia Mathematica *und verwandter Systeme I.* Monatshefte Für Mathematik und Physik 38 (1931) 173–198.

[6] Gödel, K. *The consistency of the Axiom of Choice and the Generalized Continuum Hypothesis.* Proc. Natl. Acad. of Sci. 24(1938) 556–557.

[7] Hilbert, D. *Mathematical Problems*, Bull. Am. Math. Soc. 8(1902) 437–479.

[8] Jensen, R. *Souslin's Hypothesis is incompatible with $V = L$*, Not. Am. Math. Soc. 15 (1968) 935.

[9] Jensen, R. *The fine structure of the Constructible hierarchy.* Ann. Math. Logic 4(1972) 229–308.

[10] Levy, A., Solovay, R. *Measurable Cardinals and the Continuum Hypothesis.* Israel Journal of Math 5 (1967), 234–248.

[11] Martin, D. *Borel determinacy.* Annals of Mathematics, 102 (1975), 363–371.

[12] Martin, D., Steel, J. *A proof of projective determinacy*, J. Am. Math. Soc. 2 (1989) 71–125.

[13] Mycielski, J., Steinhaus, H. *A mathematical axiom contradicting the axiom of choice*, Bull. Acad. Pol. Sci. 10, (1962) 1–3.

[14] Matijesevič, J. *Enumerable sets are diophantine*, Dokl. Akad. Nauk. SSSR. 191 (1970) 279–282.

[15] Solovay, R., Reinhardt W., Kanamori, A. *Strong axioms of infinity and elementary embeddings* Ann. Math. Logic 13 1978 73–116.

[16] Woodin, W. H. *Supercompact cardinals, sets of reals, and weakly homogeneous trees* Proc. Natl. Acad. of Sci. 85 (1988), 6587–6591.

Matthew Foreman
Department of Mathematics
University of California
Irvine, California 92697
USA

When is an Equivalence Relation Classifiable?

Greg Hjorth

Abstract. One finds in certain branches of analysis the idea that a classifiable equivalence relation is one for which we can assign points in a very concrete space as a complete invariant. Results by Effros, Glimm, and Mackey, and then later Harrington, Kechris, and Louveau, have given a thorough analysis of when such a classification is possible. In the last few years a similar analysis has been undertaken by descriptive set theorists regarding when an equivalence relation is classifiable by *countable structures considered up to isomorphism*. There is a kind of parallel theory of which equivalence relations can be assigned countable structures as complete invariants.

1991 Mathematics Subject Classification: 04A15
Keywords and Phrases: Equivalence relations, effective cardinality, classification, Polish group actions.

§0 One answer The question posed in the title of this talk is admittedly a vague one. Not only is the question itself vague, but moreover any answer to this question will necessarily be subjective, since a classification theorem will only be satisfactory if it is judged as such for some specific purposes.

Nevertheless, in certain branches of mathematics, especially those influenced by the works of George Mackey, one finds the idea that a *classifiable* equivalence relation is one for which points in some very concrete spaces – such as $\mathbb{R}$, $\mathbb{C}, \mathbb{T}, C([0,1])$ – can be assigned in some reasonably 'nice', preferably Borel, manner. Ultimately I will discuss some alternative notions of *classifiable* and present motivating examples for this line of research. Before continuing we should understand the following definition.

0.1 Definition Let E be an equivalence relation on a Polish space X. E is *smooth* or *tame* if there is Polish space Y and a Borel function

$$\theta : X \to Y$$

such that for all $x, y \in X$

$$xEy \Leftrightarrow \theta(x) = \theta(y).$$

Just so there are no confusions about the definitions, a *Polish space* is a separable topological space that admits a complete compatible metric – and so the class of Polish spaces includes objects like the reals, the complex numbers,

Hilbert space, and so on. A function between Polish spaces is said to *Borel* if the pullback of any open set is Borel.

It is also customary in this context to refer to a Polish space stripped down to its Borel structure as a *standard Borel space*; that is to say, $(Y, \mathcal{B})$ is a standard Borel space if there is a Polish topology τ on Y with respect to which $\mathcal{B}$ is the σ-algebra generated by the τ-open sets.

In definition 0.1 we could just as well insist that Y be $\mathbb{R}$, since any Polish space allows a Borel injection into the reals.

It may then be helpful to think of the function

$$\theta : X \to \mathbb{R}$$

from 0.1 as lifting to an injection

$$\hat{\theta} : X/E \to \mathbb{R},$$

and that in this sense the *Borel cardinality* of X/E is less than or equal to the Borel cardinality of $\mathbb{R}$. Indeed this is an important theme in this branch of descriptive set theory: Determine the *effective cardinality* of quotients of the form X/E.

An another equivalent formulation of smoothness is that the space of equivalence classes, $X/E = \{[x]_E : x \in X\}$, be a subspace of a *standard Borel space* in the quotient Borel structure – that is to say, if we let $\mathcal{B}_E$ be the collection of subsets of X/E of the form $\{[x]_E : x \in A\}$ for $A \subset X$ an E-invariant (any $x \in A$ has $[x]_E \subset A$) Borel set, then there is some standard Borel space $(Y, \mathcal{B})$ with $Y \supset X/E$ and $\mathcal{B}_E = \{A \cap X/E : A \in \mathcal{B}\}$. Finally, E is smooth if and only if there is a countable sequence $(A_n)_{n\in\mathbb{N}}$ of E-invariant such that for all $x, y \in X$

$$xEy \Leftrightarrow \forall n(x \in A_n \Leftrightarrow y \in A_n).$$

I suppose that for a mathematician approaching this from another area the restriction to the Borel category may seem rather arbitrary. It turns out that many mathematically objects can be naturally realized as either points in some Polish space or as equivalence classes in some Polish space, and in fact the context of these problems is far wider than it may initially appear. The theorems stated below in §4 for Borel functions all pass to much more general classes of *reasonably definable* functions.

Historically the notion of smoothness as classifiability is extremely important. Not only does one find the notion in papers such as [2], [3], and [5], and perhaps [15]. These papers suggest a wider project to determine which equivalence relations are smooth and which classification problems are no harder than that of the equality relation on $\mathbb{R}$.

§1 Examples: Smooth

1.1 Example: Compact Riemann surfaces A very natural classification problem is that of compact Riemann surfaces considered up conformal equivalence. In this case there exists a reduction to the equality relation on the reals. The classical theory, as at say [11], obtains points in some standard Borel space as a complete invariant.

Of course one can ask how this sits with the original definition at 0.1. Here it is routine (but see [13] for details) to obtain a standard Borel space parameterizing (separable) complex manifolds in some natural manner. In this context one has that the set of points parameterizing the *compact* complex *surfaces* is Borel and the equivalence relation of conformal equivalence restricted to this Borel set is indeed smooth in exactly the sense of 0.1.

1.2 Example: Bernoulli shifts Let $S = \{s_1, ..., s_n\}$ be a finite alphabet, $\sigma : S^{\mathbb{Z}} \to S^{\mathbb{Z}}$ be the shift map, and for $p_1, p_2, ..., p_n$ a finite sequence of positive numbers summing to 1 let μ the product measure resulting from giving s_i the weight p_i. We may choose to think of two such systems as being equivalent if there is an invertible measure preserving map that conjugates them: that is, set $(S_1, \sigma_1, \mu_1) \sim (S_2, \sigma_2, \mu_2)$ if there is a measurable preserving bijection

$$\pi : (S_1)^{\mathbb{Z}} \to (S_2)^{\mathbb{Z}}$$

such that

$$\sigma_1 = \pi^{-1} \circ \sigma_2 \circ \pi$$

$$\forall A \subset (S_2)^{\mathbb{Z}} (\mu_2(A) = \mu_1(\pi^{-1}(A))).$$

Ornstein in [16] shows a single real number, the *entropy* of the system (S, σ, μ), provides a complete invariant. Moreover in a suitable standard Borel structure, this invariant *can* be calculated in a Borel fashion. Here as a suitable Borel structure one may represent the shift by the sequence $p_1, p_2, ..., p_n \in \mathbb{R}^n$ for various n; the point is that a countable union of standard Borel spaces, such as $\bigcup_n \mathbb{R}^n$ is again standard Borel.

1.3 Example: Group representations Consider the irreducible representations of the group $\mathbb{Z}$. Given a complex Hilbert space H with associated unitary group U of all inner product respecting transformations, we can let $\mathrm{Irr}(\mathbb{Z}, H)$ be the space of homomorphisms

$$\tau : \mathbb{Z} \to U$$

where U has no non-trivial invariant subspaces under $\tau[\mathbb{Z}]$. It is natural to think of τ_1 and τ_2 as somehow presenting equivalent representations if there is some $T \in U$ with

$$\tau_1(g) = T \circ \tau_2(g) \circ T^{-1}$$

for all $g \in \mathbb{Z}$.

The space of *all* representations may be naturally identified with a closed subspace of $H^{\mathbb{Z}}$, and hence it is a Polish space. Furthermore the equivalence relation of interest here is induced by the continuous action of the group H.

Here $\mathrm{Irr}(\mathbb{Z}, H)$ is non-empty if and only if H is one dimensional. Moreover we may identify the elements of $\mathrm{Irr}(\mathbb{Z}, H)$ with characters, and thus a complete classification of these objects may be given by points in $\mathbb{T}$, and hence $\mathbb{R}$.

On the other hand if G is finite the space $\mathrm{Irr}(G, H)$ will be non-empty only when H is finite dimensional. Then the above equivalence relation will be induced by the a continuous action of the now compact group U on the Polish space

$\mathrm{Irr}(G, H)$. In general such orbit equivalence relations are always classifiable by points in $\mathbb{R}$.

§2 EXAMPLES: NON-SMOOTH

2.1 EXAMPLE: GENERAL COMPLEX DOMAINS One can view Becker, Henson, and Rubel in [1] as obtaining non-classifiability by a process tantamount to embedding the equivalence relation E_0 of eventual agreement on infinite sequences of 0's and 1's into conformal equivalence on complex domains – so that for $f, g : \mathbb{N} \to \{0, 1\}$ we have fE_0g if there is some $N \in \mathbb{N}$ such that

$$\forall n > N(f(n) = g(n)).$$

Here E_0 is an F_σ equivalence relation on $\{0, 1\}^{\mathbb{N}}$, the space of all infinite binary sequences in the product topology, and is in some ways (compare [9]) the canonical example of a non-smooth equivalence relation.

In fact if we assign $\mathcal{D}$, the space of open subsets of $\mathbb{C}$, with the *Effros standard Borel structure* – under which it does have a natural Borel structure – then their argument can be seen as showing that there is a Borel function

$$\theta : \{0, 1\}^{\mathbb{N}} \to \mathcal{D}$$

such that fE_0g if and only if $\theta(f)$ and $\theta(g)$ are biholomorphic. Since E_0 is non-smooth we obtain non-smoothness of conformal equivalence on arbitrary complex surfaces, even with respect to the Borel structure articulated in [13].

2.2 EXAMPLE: ARBITRARY MEASURE PRESERVING TRANSFORMATIONS Consider M_∞ the group of all invertible measure preserving transformations of the unit interval . In the topology it inherits from its action on $L^2([0, 1])$ this is a topological group that is Polish as a space – that is to say, it is a *Polish group*. For instance, if (U_n) enumerates the basic open subsets of $[0, 1]$ we obtain a complete metric with

$$d(\pi_1, \pi_2) = \sum_{n \in \mathbb{N}} 2^{-n} \lambda(\pi_1(U_n) \Delta \pi_2(U_n)) + \lambda(\pi_1^{-1}(U_n) \Delta \pi_2^{-1}(U_n)).$$

The obvious classification problem is for the conjugacy equivalence relation – it is natural to say that $\pi_1, \pi_2 : [0, 1] \to [0, 1]$ are *equivalent* if they are conjugate, in the sense of their being some $\sigma \in M_\infty$ such that

$$\sigma \circ \pi_1 = \pi_2 \circ \sigma \text{ a.e.}$$

This equivalence relation was observed by Feldman [5] to be non-smooth. As with 2.1 the proof rested on embedding E_0.

2.3 EXAMPLE: GROUP REPRESENTATIONS AGAIN Let G be a countable discrete group that it *not* abelian-by-finite. Let H_∞ be a separable infinite dimensional Hilbert space and U_∞ the unitary group on H_∞. Again take $\mathrm{Irr}(G, H_\infty)$ to be space of irreducible representations $\tau : G \to U_\infty$ with the equivalence relation of conjugacy –

$$\tau_1 \approx \tau_2 \Leftrightarrow \exists A \in U_\infty \forall g \in G(\tau_1(g) = A^{-1} \circ \tau_2(g) \circ A).$$

It is known from [17] and [8] that $\mathrm{Irr}(G, H_\infty)$ is non-empty and $\approx$ is not smooth: there is no Borel assignment of reals as complete invariants to $\mathrm{Irr}(G, H_\infty)/\approx$.

§3 More Examples: Puzzling cases The above were deliberately chosen with the view to supporting the intuition that *classifiable* means *smooth*. In the cases where there is a proof of smoothness, it is generally accepted as a classification theorem. In the cases where the equivalence relation does not admit points in $\mathbb{R}$ as a complete invariant, the authors seemed to take *that* as a proof of at least some manner of non-classifiability. Consequently I hope the position that takes classifiable to mean smooth will seem an initially attractive one.

This much said, let us consider some examples where there is a more generous notion of classifiability implicit; these in turn have motivated the search for new tools in the study of Borel and analytic equivalence relations.

3.1 Question: Complex surfaces Becker, Henson, and Rubel in [1] explicitly ask: is there some reasonably non-pathological way to assign to every domain $D \subset \mathbb{C}$ some countable set of complex numbers S_D such that

$$D \cong D'$$

if and only if

$$S_D = S_{D'}?$$

3.2 Example: Discrete spectrum mpt's Halmos and von Neumann in [10] showed that for *discrete spectrum* elements of M_∞, we may assign a countable collection $\{c_i(\pi) : i \in \mathbb{N}\}$ of complex numbers that completely describe the equivalence class of π. While conjugacy on discrete measure preserving transformations is not smooth, the Halmos-von Neumann theorem would seem to constitute some sort of weaker notion of classification, and it certainly appears to be accepted as such.

3.3 Example: C^*-algebras and topological dynamics (This is not quite analogous to examples 1.3 and 2.3, but derives from roughly the same area.) Giordano, Putnam, and Skau in [6] consider the problem of classifying *minimal Cantor systems up to orbit equivalence*. Two continuous

$$\varphi_1 : X_1 \to X_1,$$

$$\varphi_2 : X_2 \to X_2$$

which are *minimal* in the sense of having no non-trivial closed invariant sets and are *Cantor* in the sense of X_1, X_2 being compact, uncountable and completely disconnected metric spaces, are said to be *orbit equivalent* if there is a homeomorphism $F : X_1 \to X_2$ which respects the orbit structure set wise, in that for all x

$$\{\varphi_2^i(F(x)) : i \in \mathbb{Z}\} = F[\{\varphi_1^i(x) : i \in \mathbb{Z}\}].$$

This problem is in turn equivalent to classifying a certain class of C^*-algebras.

Here they produce countable ordered abelian groups as complete invariants. One similarly finds discussion of the classification of certain C^* by countable discrete structures considered up to isomorphism in papers such as [4].

It is important to note a link between the kind of classification one finds in 3.1-2 and 3.3: Any equivalence relation that can be classified by a countable set of reals can be classified by countable structures considered up to isomorphism. For instance, if we let (q_n) enumerate the rationals, then to a countable unordered set $A \subset \mathbb{R}$ we can associate the model $\mathcal{M}_A = \{x_a : a \in A\}$ with unary predicates (P_n) governed by the rule that

$$\mathcal{M}_A \models P_n(x_a)$$

if and only if $q_n < a$. Trivially then reduction to the equality relation on $\mathbb{R}$ implies classification by countable sets of reals, and hence classification by countable structures.

Thus we may be led to formulate a more generous notion of classifiability.

3.4 QUESTION For which E can we provide some kind of countable structure considered up to isomorphism as a complete invariant?

Letting $\mathcal{L}$ be a countable language, we form $\mathrm{Mod}(\mathcal{L})$, the space of all $\mathcal{L}$-structures on $\mathbb{N}$ with the topology generated by quantifier free formulas. This is a Polish space, and therefore there is a precise version of the question.

For which equivalence relations E on Polish X can we find a Borel $\theta : X \to \mathrm{Mod}(\mathcal{L})$ such that for all $x, y \in X$

$$xEy \Leftrightarrow \theta(x) \cong \theta(y)?$$

In very general terms these examples may illustrate the kinds of concerns driving the descriptive set theory of equivalence relations, as well as the particular problem of classification by countable structures. I should add to these general remarks that the isomorphism relation on countable structures is historically important in logic, and that for someone in my area it seems intriguing to ask which classification problems may be simply reduced to that of countable models considered up to isomorphism.

§4 SOME THEOREMS I will begin with two sufficient conditions for classifiability, the first of which is trivial.

4.1 THEOREM(folklore) Let G be a compact metrizable group acting continuously on a Polish space X with induced orbit equivalence relation E_G. Then E_G is smooth.

4.2 THEOREM (Kechris [14]) Let G be a locally compact Polish group acting continuously on a Polish space X. Then there is a countable sequence of Borel functions $(f_i)_{i\in\mathbb{N}}$ such that for all $x, y \in X$

$$xE_Gy \Leftrightarrow \{f_i(x) : i \in \mathbb{N}\} = \{f_i(y) : i \in \mathbb{N}\}.$$

In other words, we may classify by *countable unordered sets* of reals.

And two sufficient conditions for non-classifiability:

4.3 Theorem (folklore) Let G be a Polish group and X a Polish space. Suppose that

(i) some orbit is dense;

(ii) every orbit is meager (its complement includes the intersection of countably many open dense sets). Then E_G is not smooth.

4.4 Theorem (Hjorth [12]) Let G be a Polish group and X a Polish space. Suppose that

(i) some orbit is dense;

(ii) every orbit is meager (its complement includes the intersection of countably many open dense sets);

(iii) for some $x \in X$, *the local orbits of* x are all somewhere dense; that is to say, if V is an open neighborhood of 1_G, U is an open set containing x, and if $O(x, U, V)$ is the set of all $\hat{x} \in [x]_G$ such that there is a finite sequence $(x_i)_{i \leq k} \subset U$ such that $x_0 = x$, $x_k = \hat{x}$, and each $x_{i+1} \in V \cdot x_i$, then the closure of $O(x, U, V)$ contains an open set.

Then there is no Borel (or even Baire measurable) $\theta : X \to \mathrm{Mod}(\mathcal{L})$ such that for all $x, y \in X$

$$xE_G y \Leftrightarrow \theta(x) \cong \theta(y).$$

Consequently there is no sequence $(f_i)_{i \in \mathbb{N}}$ of Borel (or even *reasonably definable*) functions

$$f_i : X \to \mathbb{R}$$

such that

$$xE_G y \Leftrightarrow \{f_i(x) : i \in \mathbb{N}\} = \{f_i(y) : i \in \mathbb{N}\}.$$

A Polish group action satisfying 4.4(i)-(iii) is called *generically turbulent*.

Again I will return to the motivation and examples in the next and final section. These examples on their own may suggest that 4.4 is the *right theorem* for showing this kind of non-classifiability.

However there are also results in [12] reinforcing this view. The presence of a generically turbulent action is necessary for non-classifiability in the sense that if E_G arises from the continuous action of Polish G on Polish X then either E_G is reducible to isomorphism on countable structures (using say *universally Baire measurable* functions) or there is a generically turbulent Polish G-space Y which admits a continuous G-embedding into X. (Here a function θ is said to be universally Baire measurable if for any Borel function ρ we have that $\theta \circ \rho$ is Baire measurable – in the sense of pulling back open sets to sets with the Baire property.)

§5 Examples again

5.1 Example: Complex manifolds again By the uniformization theorem, conformal equivalence on complex surfaces may be reduced to an appropriately chosen locally compact group action.

THEOREM (Hjorth-Kechris [13]) Let $\mathcal{D}$ be the space of all complex domains. Then there is a *definable* assignment

$$M \mapsto S_M$$

of countable sets of reals to domains such that for all $M, N \in \mathcal{D}$

$$M \cong N \Leftrightarrow S_M = S_N.$$

Moreover it is Borel in the sense of there existing a countable sequence (f_n) of Borel functions from $\mathcal{D}$ to $\mathbb{R}$ such that S_M always equals the (unordered) set $\{f_n(M) : n \in \mathbb{N}\}$.

But in higher dimensions one may embed a generically turbulent orbit equivalence relation and obtain:

THEOREM (Hjorth-Kechris [13]) Let $\mathcal{M}^2$ be the space of two dimensional complex manifolds. Then there is no Borel assignment of countable structures up to isomorphism as complete invariants. Consistently with ZFC there is no *definable* assignment.

5.2 EXAMPLE: MEASURE PRESERVING TRANSFORMATIONS AGAIN

THEOREM (Hjorth) Let M_∞ be the space of invertible measure preserving transformations on the unit interval. Consider the conjugacy equivalence relation $\sim$: $\pi_1 \sim \pi_2$ if there is $\sigma \in M_\infty$ such that

$$\sigma \circ \pi_1 = \pi_2 \circ \sigma \text{ a.e.}$$

Then there is no sequence $(f_i)_{i \in \mathbb{N}}$ of Borel functions

$$f_i : M_\infty \to \mathbb{R}$$

such that

$$\pi_1 \sim \pi_2 \Leftrightarrow \{f_i(\pi_1) : i \in \mathbb{N}\} = \{f_i(\pi_2) : i \in \mathbb{N}\}.$$

In fact, $\sim$ is strictly more complicated than isomorphism on countable models: there is a Borel $\theta : \text{Mod}(\mathcal{L}) \to M_\infty$ such that for all $M, N \in \text{Mod}$

$$M \cong N \Leftrightarrow \theta(M) \sim \theta(N),$$

but (for any choice of $\mathcal{L}$) there is no Borel (or even universally Baire measurable) $\rho : M_\infty \to \text{Mod}(\mathcal{L})$ such that for all $\pi_1, \pi_2 \in M_\infty$

$$\pi_1 \sim \pi_2 \Leftrightarrow \rho(\pi_1) \cong \rho(\pi_2).$$

5.3 EXAMPLE: DISCRETE GROUP REPRESENTATIONS AGAIN

THEOREM (Hjorth) Let G be a countable group that is *not* abelian-by-finite. Let H_∞ be an separable infinite dimensional Hilbert space, let $\text{Irr}(G, H_\infty)$ be the space of irreducible representations of G in H_∞. Then there is no sequence $(f_i)_{i \in \mathbb{N}}$ of Borel functions

$$f_i : \text{Irr}(G, H_\infty) \to \mathbb{R}$$

such that

$$\tau_1 \approx \tau_2 \Leftrightarrow \{f_i(\tau_1) : i \in \mathbb{N}\} = \{f_i(\tau_2) : i \in \mathbb{N}\}.$$

In fact there is no *reasonably definable* assignment of countable models considered up to isomorphism as complete invariants.

References

[1] J. Becker, C.W. Henson, L. Rubel, *First order conformal invariants,* Annals of Mathematics, vol. 102(1980), pp. 124-178.

[2] E. Effros, *Transformation groups and C^*-algebras,* Annals of Mathematics, ser 2, vol. 81(1975), pp. 38-55.

[3] E. Effros, *Polish transformation groups and classification problems*, General topology and modern analysis (Proc. Conf., Univ. California, Riverside, Calif., 1980), pp.217-227, Academic Press, New York-London, 1981.

[4] G.A. Elliott, *On the classification of inductive limits of sequences of semisimple finite-dimensional algebras*, Journal of Algebra, vol. 38(1976), pp. 29-44.

[5] J. Feldman, *Borel structure and invariants for measurable transformations,* Proceedings of the American Mathematical Society, vol. 46(1974), pp. 383-394.

[6] T. Giordano, I.F. Putnam, C.F. Skau, *Topological orbit equivalence and C^*-crossed products*, Journal Reine und Angewandt Mathematische, vol. 469(1995), pp. 51-111.

[7] J. Glimm, *Locally compact transformations groups,* Transactions of the American Mathematical Society, vol. 101(1961), pp. 124-138.

[8] J. Glimm, *Type I C^*-algebras,* Annals of Mathematics, 1961, pp. 572-612.

[9] L.A. Harrington, A.S. Kechris, A. Louveau, *A Glimm-Effros dichotomy for Borel equivalence relations,* Journal of the American Mathematical Society, vol. 3(1990), pp. 903-928.

[10] P.R. Halmos, J. von Neumann, *Operator methods in classical mechanics, II,* Annals of Mathematics, vol. 43(1942), pp. 332-50.

[11] Y. Imayoshi, M. Taniguchi, An introduction to Teichmüller spaces, Springer-Verlag, Berlin, 1992.

[12] G. Hjorth, Classification and orbit equivalence relations, manuscript (available at www.math.ucla.edu/~greg/research.html).

[13] G. Hjorth, A.S. Kechris, The complexity of the classification of Riemann surfaces by complex manifolds, preprint, UCLA, 1998.

[14] A.S. Kechris, *Countable sections for locally compact group actions,* Journal of ergodic theory and dynamical systems, vol. 12(1992), pp. 283-295.

[15] G.W. Mackey, *Infinite-dimensional group representations,* Bulletin of the American Mathematical Society, vol. 69(1963), pp.628-686.

[16] D.S. Ornstein, *Bernoulli shifts with the same entropy are isomorphic,* ADVANCES IN MATHEMATICS, vol. 4(1970), pp. 337-52.

[17] E. Thoma, *Eine Charakterisierung diskreter Gruppen vom type I,* INVENTIONES MATHEMATIQUE, vol. 6(1968), pp. 190-196.

Greg Hjorth
Mathematics
UCLA
CA90095-1555
USA
greg@math.ucla.edu

Meager Forking and m-Independence

Ludomir Newelski

Abstract. We describe meager forking, m-independence and related notions of geometric model theory relevant for Vaught's conjecture and more generally for classifying countable models of a superstable theory.

1991 Mathematics Subject Classification: 03C45
Keywords and Phrases: Vaught's conjecture, superstable theory

0 Introduction

Throughout, $T = T^{eq}$ is a complete theory in a countable first-order language L and we work within a large saturated model $\mathfrak{C}$ of T (a monster model). Until section 5 we assume that T is stable. Often we assume that T is small, i.e. $S_n(\emptyset)$ is countable for every $n < \omega$. The general references are [Bu5, Pi].

The main motivation here is Vaught's conjecture for superstable theories. Vaught's conjecture says that if T has $< 2^{\aleph_0}$ countable models, then T has countably many of them. If T is not small, then T has $2^{\aleph_0}$ countable models. So the assumptions that T is small or even that T has $< 2^{\aleph_0}$ countable models appear naturally in many theorems in this paper. Thus far Vaught's conjecture is proved for ω-stable theories [SHM] and superstable theories of finite U-rank [Bu4]. The main tools of Shelah in [SHM] are forking of types and forking independence. These tools are combinatorial in nature. Forking is also the main tool in [Sh]. [Bu4, Ne1, Ne3] indicate that in order to approach Vaught's conjecture for superstable theories we may need some new ideas and tools, of more geometric and algebraic character.

In a series of papers I introduced meager forking, m-independence and other notions intended for a fine analysis of countable models. Meager forking relates forking to the topological structure of the space of types. It is used to show that the topological character of forking is related to the geometry of forking. An important problem arising in the context of Vaught's conjecture is to describe the ways in which a type in a superstable theory may be non-isolated, and also to describe the sets of stationarizations of such a type. Here m-independence and the calculus of traces of types are useful. Apart from their relevance to Vaught's conjecture, these notions may be important for model theory in general. Indeed, in a small stable theory m-independence is the strongest natural notion of independence (on finite tuples) refining forking independence. So there is a hope that with sharper tools we can better describe countable models. The theory of m-independence is

in many ways parallel to the theory of forking independence of Shelah [Sh]. Also, restricted to $*$-algebraic tuples, m-independence may be defined in an arbitrary (small) theory T.

Usually, $a, b, c, \ldots$ denote finite tuples and $A, B, C, \ldots$ finite sets of elements of $\mathfrak{C}$. $x, y, z, \ldots$ denote finite tuples of variables.

1 Meager forking and meager types

Assume $s(x)$ is a (possibly incomplete) type over $\mathfrak{C}$. $[s]$ denotes the class of types in variables x containing $s(x)$. $s(\mathfrak{C})$ denotes the set of tuples from $\mathfrak{C}$ realizing s. We define the trace of s over A as the set $Tr_A(s) = \{tp(a/acl(A)) : a \in s(\mathfrak{C})\}$, a closed subset of $S(acl(A))$. In particular, for $p \in S(A)$, $Tr_A(p)$ is the set of stationarizations of p over A.

Assume P is a closed subset of $S(acl(A))$. We say that forking is meager on P if for every formula $\varphi(x)$ forking over A, the set $Tr_A(\varphi) \cap P$ is nowhere dense in P (equivalently: for every finite $B \supset A$, the set of types $r \in P$ with a forking extension in $S(acl(B))$ is meager in P). For $p \in S(A)$ we say that forking is meager on p, if forking is meager on $Tr_A(p)$.

Assume r is a stationary regular type. We say that $\varphi(x) \in L(A)$ is an r-formula (over A) if

- every type in $S(acl(A)) \cap [\varphi]$ is either hereditarily orthogonal to r or regular non-orthogonal to r,
- the set $P_\varphi = \{p \in S(acl(A)) \cap [\varphi] : p \not\perp r\}$ is closed and non-empty,
- r-weight 0 is definable on φ, that is whenever $a \in \varphi(\mathfrak{C})$ and $w_r(a/Ac) = 0$, then for some formula $\psi(x, y)$ over $acl(A)$, true of (a, c), if $\psi(a', c')$ holds, then $w_r(a'/Ac') = 0$.

If $P_\varphi = \{p\}$ is a singleton, then we say that p is strongly regular. Strongly regular types were an essential ingredient in describing countable models of an ω-stable theory in [SHM].

For a stationary regular type $r \in S(B)$, forking induces a closure operator cl on $r(\mathfrak{C})$ defined by $a \in cl(X)$ iff $a \not\perp X(B)$, where $\{a\} \cup X \subseteq r(\mathfrak{C})$. cl is a (combinatorial) pregeometry on $r(\mathfrak{C})$ (this is in fact equivalent to regularity of r), which we call the forking geometry on r. We say that r is [locally] modular, if this geometry is [locally] modular. We say that r is non-trivial, if this geometry is non-trivial [Pi].

Locally modular regular types are important in geometric model theory. If r is non-trivial and locally modular, then the associated geometry is either affine or projective over some division ring [Hr1]. By [HS], in a superstable T, for any non-trivial regular type r, r-formulas exist.

Definition 1 ([Ne5]) *We say that a regular stationary type r is meager if for some (equivalently: any) r-formula φ, forking is meager on P_φ.*

For instance, every properly weakly minimal non-trivial type is meager.

Theorem 1 ([Ne5]) *Every meager type is non-trivial and locally modular.*

This theorem improves [Bu1, LP]. It shows that the topological character of forking on a regular type is relevant to its geometric properties. Hrushovski and Shelah proved in [HS] that in a superstable theory without the omitting types order property

(∗) every regular type is either locally modular or non-orthogonal to a strongly regular type.

So in this case either the forking geometry on a type is nice or the situation is similar to the ω-stable case. Hrushovski [Hr2] gave an example of a regular type in a superstable theory, for which (∗) fails.

Question 1 *Does* (∗) *hold in any superstable theory with* $< 2^{\aleph_0}$ *countable models ?*

Following [Ta] we say that a regular type $p \in S(A)$ is eventually strongly non-isolated (esn), if some non-forking extension p' of p over a finite $A' \supset A$ is strongly non-isolated, that is, for every finite $B \supset A'$, p' is almost orthogonal to any isolated type in $S(B)$. Also we say that p is almost strongly regular (asr), via $\varphi \in p$, if φ is a p-formula over A and $P_\varphi = Tr_A(p)$. Since by Theorem 1 every meager type is locally modular, the following characterization of non-trivial esn types is relevant for Question 1.

Theorem 2 ([Ne7]) *Assume T is small superstable and p is a non-trivial regular type. Then p is esn iff (1) or (2) below holds. Moreover, (1) and (2) are mutually exclusive.*
(1) p is non-orthogonal to an almost strongly regular non-isolated type.
(2) p is meager.

2 $\mathcal{M}$-rank and m-independence

In this section T is small and stable. For $p \in S(A)$, $Tr_A(p)$ is either finite or homeomorphic to the Cantor set. We measure traces of types by comparing topologically traces of their various extensions. This is done by means of $\mathcal{M}$-rank and m-independence.

Assume $q \in S(B)$ is a non-forking extension of $p \in S(A)$ ($A \subseteq B$). Then $Tr_A(q)$ is a closed subset of $Tr_A(p)$ and either is open in $Tr_A(p)$ or is nowhere dense in $Tr_A(p)$. In the former case we call q an m-free and in the latter a meager extension of p. So q is an m-free extension of p iff q is isolated in the set of non-forking extensions of p in $S(B)$.

Definition 2 ([Ne3]) *The rank function $\mathcal{M}$ is the minimal function defined on the set of all complete types over finite sets, with values in $Ord \cup \{\infty\}$, such that for every $\alpha \in Ord$ we have*
$\mathcal{M}(p) \geq \alpha + 1$ iff $\mathcal{M}(q) \geq \alpha$ for some meager non-forking extension q of p.
$\mathcal{M}(a/A)$ abbreviates $\mathcal{M}(tp(a/A))$. We say that T is m-stable if $\mathcal{M}(p) < \infty$ for every p.

DEFINITION 3 ([NE8]) *We say that a is m-independent from B over A (symbolically: $a \overset{m}{\downarrow} B(A)$) if $tp(a/A \cup B)$ is an m-free extension of $tp(a/A)$.*

m-independence has similar properties as forking independence.

PROPOSITION 1 ([NE3, NE8]) *(1)(symmetry) If $a \overset{m}{\downarrow} b(A)$, then $b \overset{m}{\downarrow} a(A)$.*
(2) (transitivity) $a \overset{m}{\downarrow} B \cup C(A)$ iff $a \overset{m}{\downarrow} B(A)$ and $a \overset{m}{\downarrow} C(A \cup B)$.
(3) $a \overset{m}{\downarrow} B(A)$ is invariant under automorphisms of $\mathfrak{C}$ and under changes of enumerations of a, A, B.
(4)(acl-triviality) If $B \subseteq acl(A)$, then $a \overset{m}{\downarrow} B(A)$.
(5) In a small theory, $\overset{m}{\downarrow}$ has an extension property, i.e. every type $p \in S(A)$ has an m-free extension over any finite $B \supset A$.

THEOREM 3 ([NE10]) *In a small stable theory m-independence is the strongest notion of independence on finite tuples and finite sets of elements of $\mathfrak{C}$, which refines forking independence and has the properties exhibited in Proposition 1.*

In a small stable theory, in the following Lascar-style inequalities

(L) $\mathcal{M}(a/Ab) + \mathcal{M}(b/A) \leq \mathcal{M}(ab/A) \leq \mathcal{M}(a/Ab) \oplus \mathcal{M}(b/A)$

the right side is always true, while the left side holds if $a \downarrow b(A)$ (that is, if a, b are forking-independent over A).

In a small superstable theory $\mathcal{M}$-rank may be used to find meager types [Ne6] (similarly as U-rank considerations lead to regular types [Ls]). To find many such types we need types of large (infinite, but $< \infty$) $\mathcal{M}$-rank, to begin with. Unfortunately, no such types are known in a small stable theory.

CONJECTURE 1 ([NE7, THE $\mathcal{M}$-GAP CONJECTURE]) *In a small stable theory there is no type p with $\omega \leq \mathcal{M}(p) < \infty$.*

This conjecture is true for superstable theories under the few models assumption.

THEOREM 4 ([NE5, NE7]) *If T is superstable with $< 2^{\aleph_0}$ countable models, then T is m-stable. Moreover, for every type p, $\mathcal{M}(p)$ is finite and $\leq U(p)$.*

The proof of this theorem relies on the construction of some meager types and the analysis of traces of some types in the associated meager groups (defined below). The special case of theorem 4, where T is weakly minimal and $U(p) = 1$, was conjectured by Saffe and proved in [Ne1]. It was decisive in the proof of Vaught's conjecture for weakly minimal theories [Bu3, Ne1].

Using the notions of $\mathcal{M}$-rank and m-independence we get the following description of traces of types.

THEOREM 5 ([NE8, THE TRACE THEOREM]) *If T is superstable with $< 2^{\aleph_0}$ countable models, then for every $p \in S(A)$ there is a formula $\varphi(x)$ (usually not in p) with $Tr_A(\varphi) = Tr_A(p)$. In particular, if p is regular and forking is meager on p, then p is isolated and meager.*

[Ne8, Theorem 2.6] contains more information on traces of meager types.

Regarding Theorem 3 we should mention that there is a notion of independence intermediate between $\overset{m}{\downarrow}$ and $\downarrow$. Namely, assume again $q \in S(B)$ is a non-forking extension of $p \in S(A)$. On $Tr_A(p)$ there is a natural probabilistic Haar measure, invariant under $Aut(\mathfrak{C}/A)$. We say that q is a μ-free extension of p if $Tr_A(q)$ has positive measure in $Tr_A(p)$. This leads to the notion of μ-independence $\overset{\mu}{\downarrow}$ (implicitly used in [LS]), having the properties from Proposition 1 [Ne8]. Also, $\overset{m}{\downarrow} \Rightarrow \overset{\mu}{\downarrow} \Rightarrow \downarrow$.

Tanovic proved that m-independence and μ-independence are equal in an m-stable theory [Ne8], and I proved there that they are equal in an m-normal theory (defined below). In particular, by Theorem 4 we could say that in a superstable theory with $< 2^{\aleph_0}$ countable models, "measure equals category". No theory is known in which these two notions of independence differ.

3 THE $\mathcal{M}$-GAP CONJECTURE AND M-NORMAL THEORIES

In this section we assume T is small stable. In an attempt to refute the $\mathcal{M}$-gap conjecture I constructed in [Ne8] small weakly minimal groups with types of various $\mathcal{M}$-ranks. However the traces of types in these groups are not complicated, they are just translates of traces of some generic subgroups. This leads to the definition of an m-normal theory.

DEFINITION 4 ([NE8]) *T is m-normal if for every finite $A \subseteq B$ and $a \in \mathfrak{C}$, for some $E \in FE(A)$, the set $Tr_A(a/B) \cap [E(x,a)]$ has finitely many conjugates over Aa.*

The idea underlying this definition is that in an m-normal theory, locally $Tr_A(a/B)$ can be almost recovered from Aa alone. This corresponds to the condition $Cb(a/A) \subseteq acl(a)$, defining 1-based theories.

There is an evident analogy between the theory of m-independence and the theory of forking independence: meager forking, $\mathcal{M}$-rank, meager types, m-stability correspond to forking, U-rank, regular types, superstability. (Unfortunately in the theory of m-independence there is no good counterpart of the notion of a stationary type.) m-normality corresponds to 1-basedness. In order to justify this we need to introduce $*$-finite tuples, which play for m-independence a role similar to imaginaries in forking.

DEFINITION 5 ([NE8]) *(1) A $*$-finite tuple is a tuple $a_I = \langle a_i, i \in I \rangle$ of elements of $\mathfrak{C}$ (with the index set I countable), such that $a_I \subseteq dcl(a)$ for some finite tuple a of elements of $\mathfrak{C}$. Moreover, we say that a_I is $*$-algebraic over A if $a_I \subseteq acl(A)$.*
(2) $S_I(A)$ denotes the space of complete types over A, in variables $x_I = \langle x_i, i \in I \rangle$. If a_I is $*$-finite [$*$-algebraic over A], then we call $tp(a_I/A)$ $*$-finite [$*$-algebraic].

EXAMPLE 1 Let $p = tp(a/A) \in S(A)$. Then $a^* = \langle a/E : E \in FE(A) \rangle$ is a $*$-finite $*$-algebraic over A tuple naming $tp(a/acl(A))$ over A.

EXAMPLE 2 Let $G \subseteq \mathfrak{C}$ be a group definable over A and let $G_n, n < \omega$, be a sequence of A-definable subgroups of finite index in G with $G^0 = \cap_n G_n$ (G^0 is the connected component of G). Then an element a/G^0 of G/G^0 may be regarded as

a $*$-finite $*$-algebraic over A tuple $\langle a/G_n, n < \omega\rangle$. So G/G^0 is a $*$-finite $*$-algebraic group.

The definitions of forking, traces of types, $\mathcal{M}$-rank and m-independence work also for $*$-finite tuples and $*$-finite types. From now on we let $a, b, c, \ldots$ denote $*$-finite tuples and $A, B, C, \ldots$ finite sets of $*$-finite tuples of elements of $\mathfrak{C}$. Finite tuples or sets of elements of $\mathfrak{C}$ will be called standard.

Most importantly, in the new set-up Proposition 1 remains valid, also (L) holds in the same way as for standard tuples. Theorem 4 is true, except that for a $*$-finite type p, $\mathcal{M}(p)$ may be larger than $U(p)$. Unfortunately Theorem 5 does not hold for $*$-finite types. The change of the set-up does not affect the value of the $\mathcal{M}$-ranks of standard types. Also, in Example 1, $\mathcal{M}(a/A) = \mathcal{M}(a^*/A)$. This is an important point, showing that $*$-algebraic tuples are the backbone of $\mathcal{M}$-rank and m-independence. If $p \in S_I(A)$ is $*$-algebraic, then there is a natural correspondence between $Tr_A(p)$ and $p(\mathfrak{C})$, inducing on $p(\mathfrak{C})$ a compact topology.

The next theorem explains the definition of an m-normal theory with the help of $*$-algebraic tuples, making it similar to the definition of a 1-based theory using imaginaries.

THEOREM 6 ([NE7, NE12]) *T is m-normal iff for every finite A and a, b $*$-algebraic over A, there is a $c \in acl_A(a) \cap acl_A(b)$ with $a \overset{m}{\downarrow} b(Ac)$.*

Here $c \in acl_A(a)$ means that c has finitely many Aa-conjugates. For an infinite set I of $*$-finite tuples, $c \in acl(I)$ means that $c \in acl(I_0)$ for some finite $I_0 \subset I$.

Buechler characterized 1-based theories among superstable theories of finite rank as those where every U-rank 1 type is locally modular [Bu2]. This explains the geometric importance of 1-basedness. In the case of m-normality we can give a similar description. Since $*$-algebraic types are the backbone of m-independence, this description refers to some geometries on $*$-algebraic types of $\mathcal{M}$-rank 1.

Assume $p \in S_I(A)$ is $*$-algebraic, of $\mathcal{M}$-rank 1. We say that $I \subseteq p(\mathfrak{C})$ is a flat Morley sequence in p if I is countably infinite, m-independent over A and dense in $p(\mathfrak{C})$ (by [Ne10], such an I is unique up to $Aut(\mathfrak{C}/A)$). Now acl_A induces a pregeometry on $p(\mathfrak{C})$ (just like acl induces the forking geometry on a U-rank 1 type). We say that p is locally modular if for some flat Morley sequence I in p, the localized acl_{AI}-geometry on $p(\mathfrak{C})$ is modular.

We define the notion of [almost] m-orthogonality analogously to the corresponding definition in the theory of forking. We say that T has weak m-coordinatization if every $*$-algebraic type of $\mathcal{M}$-rank > 0 is m-nonorthogonal to a $*$-algebraic type of $\mathcal{M}$-rank 1. We say that T has full m-coordinatization if for every A and a $*$-algebraic over A with $\mathcal{M}(a/A) > 0$, there is some $b \in acl_A(a)$ with $\mathcal{M}(b/A) = 1$.

The next three theorems justify our interest in m-normal theories.

THEOREM 7 ([NE12]) *Assume T is small, of finite $\mathcal{M}$-rank. Then the following are equivalent.*
(1) T is m-normal.
(2) T has full m-coordinatization and every $*$-algebraic $\mathcal{M}$-rank 1 type is locally modular.

(3) T has weak m-coordinatization and every $*$-algebraic $\mathcal{M}$-rank 1 type is locally modular.

THEOREM 8 ([NE8, NE12]) *In an m-normal theory there is no type p with $\omega \leq \mathcal{M}(p) < \infty$.*

So for m-normal theories the $\mathcal{M}$-gap conjecture is true. The small weakly minimal groups referred to at the beginning of this section are m-normal. I know no small theory, which is not m-normal.

THEOREM 9 ([NE11, NE12]) *If T is superstable with $< 2^{\aleph_0}$ countable models, then T is m-normal.*

Regarded as properties of m-independence, Theorems 4,7 and 9 correspond to the result from [CHL] saying that every $\aleph_0$-stable $\aleph_0$-categorical theory has finite Morley rank and is 1-based. [Ne11] contains more information on $*$-algebraic types of $\mathcal{M}$-rank 1 in superstable theories with $< 2^{\aleph_0}$ countable models.

4 MEAGER GROUPS

Meager groups are some definable groups of standard elements of $\mathfrak{C}$. First we shall define however the notion of a $*$-finite group. We say that G is a $*$-finite group if G is a type-definable group consisting of uniformly $*$-finite tuples, that is for some finite set A and a tuple $f_I = \langle f_i, i \in I \rangle$ of A-definable functions, $G = \{f_I(a) : a \in X\}$ for some set $X \subseteq \mathfrak{C}$ type-definable over A. $\mathcal{G} \subseteq S_I(acl(A))$ denotes the set of generic types of G. For $B \supseteq A$ we say that $a_I \in G$ is m-generic over B (and $tp(a_I/B)$ is m-generic) if $a_I \overset{m}{\downarrow} B(A)$, $tp(a_I/acl(A)) \in \mathcal{G}$ and $Tr_A(a_I/B)$ is open in $\mathcal{G}$. We define $\mathcal{M}(G)$ as $\mathcal{M}(p)$ for any m-generic type p of elements of G. Also there is a natural group structure on $\mathcal{G}$, given by independent multiplication of types [Ne2]. G is called $*$-algebraic if elements of G are $*$-algebraic over A (the group from Example 2 is a good example here).

Now assume $G \subseteq \mathfrak{C}$ is an A-definable regular abelian group in a stable theory. As above, $\mathcal{G} \subseteq S(acl(A))$ denotes the set of generic types of G. Let $p \in \mathcal{G}$ be the generic type of G^0, the connected component of G. Notice that G is a p-formula and $\mathcal{G} = P_G$. So p is meager iff forking is meager on $\mathcal{G}$. In this case we call G a meager group.

By [Hr1], for any locally modular regular type q there is a regular group non-orthogonal to q, so every meager type is non-orthogonal to a meager group. We will say more on such groups.

Assume G is a locally modular regular abelian group definable over A. Let $\mathcal{G}m$ denote the set of modular types in $\mathcal{G}$ (so $p \in \mathcal{G}m$ and $\mathcal{G}m$ is a subgroup of $\mathcal{G}$). Let Gm (the modular component of G) be the subgroup of G generated by the realizations of types in $\mathcal{G}m$. In a small theory, $\mathcal{G}m$ is closed in $\mathcal{G}$ and $\mathcal{G} \setminus \mathcal{G}m$ is open in $S(acl(A))$ [Ne5].

THEOREM 10 ([NE5, NE7]) *Assume T is superstable with $< 2^{\aleph_0}$ countable models and $G \subseteq \mathfrak{C}$ is a locally modular regular abelian group definable over $\emptyset$. Then:*
(1) G is meager iff $[G : Gm] = [\mathcal{G} : \mathcal{G}m]$ is infinite iff $\mathcal{G}m$ is nowhere dense in $\mathcal{G}$.

(2) If G is meager, then $\mathcal{M}(G) = \mathcal{M}(Gm) + 1$.
(3)(generalized Saffe's condition) If G is meager and $a \in G$ is generic over A, then exactly one of the following conditions holds:
(a) $Tr_\emptyset(a/A)$ is open in $\mathcal{G}$ (i.e. a is m-generic over A and $tp(a/A)$ is isolated).
(b) $Tr_\emptyset(a/A)$ is contained in finitely many cosets of $\mathcal{G}m$ (so it is nowhere dense and $tp(a/A)$ is non-isolated).

Also, with every locally modular group G we associate a division ring F_G of definable pseudo-endomorphisms of G^0, and forking dependence on G^0 is essentially the linear dependence over F_G [Hr1]. Now if G is meager, then F_G is a locally finite field and every element of F_G is definable over $acl(\emptyset)$ [Lo, Ne5].

Using the above ideas we can prove Vaught's conjecture for some superstable theories of infinite rank. For instance, we have the following theorem.

THEOREM 11 ([NE9]) *Assume $T = Th(G)$, where G is a meager group of U-rank ω and $\mathcal{M}$-rank 1, with F_G being a prime field. Then Vaught's conjecture is true for T.*

The proof of this theorem uses also ideas from [Bu3] and from [Ne3, Ne4] on describing models piece-by-piece. This leads to some "relative Vaught's conjecture" results, which consist in the following.

Suppose $\Phi(x)$ is a countable disjunction of formulas in T. Then we can consider the restricted (many-sorted) theory $T\lceil\Phi = Th(\Phi(\mathfrak{C}))$. Proving Vaught's conjecture for T *relative* to Φ means proving Vaught's conjecture for T under the assumption of Vaught's conjecture for $T\lceil\Phi$. [Ne9, Ne13] contain some results of this form. $T = Th(G)$ for some meager group G there and $\Phi(x)$ is a disjunction of formulas such that $\Phi(G) = G^- = \{a \in G : a \text{ is non-generic}\}$, or $\Phi(G) = Gm$.

5 A GENERALIZATION

As mentioned in section 3, $*$-algebraic tuples are the backbone of m-independence. Definition 3 (of m-independence) makes sense in an arbitrary theory if a is $*$-algebraic over A. m-independence restricted to $*$-algebraic tuples has all the properties from Proposition 1 (but smallness is needed to get (5)). Then (1)-(5) from Proposition 1 imply (L), which for $*$-algebraic tuples holds fully (because when a, b are $*$-algebraic over A, then $a \mathop{\smile\!\!\!\!\!\!|}_{A} b(A)$). Also, Theorems 7 and 8 hold for an arbitrary small theory (or even just for a theory, where $*$-algebraic tuples satisfy conditions (1)-(5) from Proposition 1) [Ne12]. This suggests a possibility of applying m-independence in an unstable context.

Hrushovski and Pillay prove in [HP] that every 1-based group is abelian-by-finite. In [Ne12] I develop a theory of $*$-algebraic groups in a small m-normal theory parallel in some respects to [HP].

THEOREM 12 ([NE12]) *Assume G is a $*$-algebraic group type-definable over $\emptyset$, in a small m-normal theory. Assume $a \in G$ and $p = tp(a/A)$. Then $p(G)$ is a finite union of cosets of subgroups of G definable over parameters algebraic over $\emptyset$. Also, G is abelian-by-finite.*

Any $*$-algebraic group is a topological profinite group. It would be interesting to extract the topological content of Theorem 12. Since the group G/G^0 from Example 2 is $*$-algebraic, we get the following surprising corollary.

Corollary 1 *Assume G is a (standard) group interpretable in a superstable theory with $< 2^{\aleph_0}$ countable models. Then G/G^0 is abelian-by-finite.*

Question 2 *Is any $*$-algebraic group interpretable in a small (stable) theory abelian-by-finite ?*

Regarding this question we should mention that by the results from [Ba], if G is a standard group interpretable in a superstable theory, then G/G^0 is solvable-by-finite, and if additionally $\mathcal{M}(G/G^0) = 1$, then G/G^0 is abelian-by-finite.

References

[Ba] A.Baudisch, *On superstable groups*, Journal of London Mathematical Society (2) 42(1990), 452-464.

[Bu1] S.Buechler, *The geometry of weakly minimal types*, Journal of Symbolic Logic 50(1985), 1044-1053.

[Bu2] S.Buechler, *Locally modular theories of finite rank*, Annals of Pure and Applied Logic 30(1986), 83-95.

[Bu3] S.Buechler, *Classification of small weakly minimal sets I*, in: Classification theory, Proceedings, Chicago 1985, ed. J.T.Baldwin, Springer 1987, 32-71.

[Bu4] S.Buechler, *Vaught's conjecture for superstable theories of finite rank*, Ann.Pure Appl.Logic, to appear.

[Bu5] S.Buechler, Essential Stability Theory, Springer 1996.

[CHL] G.Cherlin, L.Harrington, A.Lachlan, *$\aleph_0$-categorical $\aleph_0$-stable structures*, Ann.Pure Appl.Logic 28(1986), 103-135.

[Hr1] E.Hrushovski, *Locally modular regular types*, in: Classification theory, Proceedings, Chicago 1985, ed. J.T.Baldwin, Springer 1987, 132-164.

[Hr2] E.Hrushovski, *A new strongly minimal set*, Ann.Pure Appl.Logic 62(1993), 147-166.

[HP] E.Hrushovski, A.Pillay, *Weakly normal groups*, in: Logic Colloquium'85, ed. Paris Logic Group, North Holland 1987, 233-244.

[HS] E.Hrushovski, S.Shelah, *A dichotomy theorem for regular types*, Ann.Pure Appl.Logic 45(1989), 157-169.

[Ls] D.Lascar, *Relation entre le rang U et le poids*, Fundamenta Mathematicae 121(1984), 117-123.

[LS] M.C.Laskowski, S.Shelah, *Forcing isomorphism II*, J.Symb.Logic 61(1996), 1305-1320.

[Lo] J.Loveys, *Abelian groups with modular generic*, J.Symb.Logic 56(1991), 250-259.

[LP] L.F.Low, A.Pillay, *Superstable theories with few countable models*, Archive for Mathematical Logic 31(1992), 457-465.

[Ne1] L.Newelski, *A proof of Saffe's conjecture*, Fund.Math. 134(1990), 143-155.

[Ne2] L.Newelski, *On type-definable subgroups of a stable group*, Notre Dame Journal of Formal Logic 32(1991), 173-187.

[Ne3] L.Newelski, *A model and its subset*, J.Symb.Logic 57(1992), 644-658.

[Ne4] L.Newelski, *Scott analysis of pseudo-types*, J.Symb.Logic 58(1993), 648-663.

[Ne5] L.Newelski, *Meager forking*, Ann.Pure Appl.Logic 70(1994), 141-175.

[Ne6] L.Newelski, *$\mathcal{M}$-rank and meager types*, Fund.Math. 146(1995), 121-139.

[Ne7] L.Newelski, *$\mathcal{M}$-rank and meager groups*, Fund.Math. 150(1996), 149-171.

[Ne8] L.Newelski, *$\mathcal{M}$-gap conjecture and m-normal theories*, Israel Journal of Mathematics, to appear.

[Ne9] L.Newelski, *Vaught's conjecture for some meager groups*, Israel J.Math., to appear.

[Ne10] L.Newelski, *Flat Morley sequences*, J.Symb.Logic, to appear.

[Ne11] L.Newelski, *Geometry of $*$-finite types*, J.Symb.Logic, to appear.

[Ne12] L.Newelski, *m-normal theories*, preprint.

[Ne13] L.Newelski, *On countable meager groups*, preprint.

[Pi] A.Pillay, GEOMETRIC STABILITY THEORY, Oxford 1996.

[Sh] S.Shelah, CLASSIFICATION THEORY, 2nd ed., North Holland 1990.

[SHM] S.Shelah, L.Harrington, M.Makkai, *A proof of Vaught's conjecture for $\aleph_0$-stable theories*, Israel J.Math. 49(1984), 259-278.

[Ta] P.Tanovic, *Fundamental order and the number of countable models*, Ph.D. thesis, McGill University, December 1993.

Instytut Matematyczny
Uniwersytetu Wrocławskiego
pl.Grunwaldzki 2/4
50-384 WROCŁAW, Poland
e-mail: newelski@math.uni.wroc.pl

Basis Problems in Combinatorial Set Theory

Stevo Todorcevic

1991 Mathematics Subject Classification: 04–02

An analysis of a given class $\mathcal{S}$ of structures in this area frequently splits into two natural parts. One part consists in recognizing the critical members of $\mathcal{S}$ while the other is in showing that a given list of critical members is in some sense complete. These kinds of problems tend to be interesting even in cases when elements of $\mathcal{S}$ do not have much structure or interest in themselves as they often appear as crucial combinatorial parts of other problems abundant in structure. A typical example of such a situation is the appearance of the Hausdorff gap (a critical substructure of the reduced power $\mathbb{N}^{\mathbb{N}}/\mathrm{FIN}$; see [15]) at the crucial place in Woodin's (consistency) proof of Kaplansky's conjecture about automatic continuity in Banach algebras ([34]). The purpose of this paper is to explain some of these problems and resulting developments. Before we start describing specific Basis Problems some general remarks are in order. Critical objects are almost always some canonical members of $\mathcal{S}$ simple to describe and visualize. Sometimes, however, it may take a considerable number of years (or decades) before an old object is identified as critical, or before one finds a (simple!) definition of a new critical object. To show that a given list $\mathcal{S}_0$ of critical objects is exhaustive one needs to relate a given structure from $\mathcal{S}$ to one from the list $\mathcal{S}_0$. If the structure in question is explicitly given one usually has no problems in finding the corresponding member of $\mathcal{S}_0$ and the connecting map. However, if the given structure from $\mathcal{S}$ is "generic", while one may still be able to identify the member of $\mathcal{S}_0$ to which it is related, one can only hope for a "generic" connecting map. Whenever we use this approach to show that a given list $\mathcal{S}_0$ is in some sense complete, the corresponding Theorem or Conjecture will be marked by [PFA]. The readers interested in the metamathematical aspects of this approach will find a satisfactory explanation in the recent monograph of Woodin [35] where it is actually shown that there is a certain degree of uniqueness in this approach.

1 Distance Functions

It is not surprising that many critical objects in families of uncountable structures live on the domain ω_1 of all countable ordinals as "critical" very often means "minimal" in some sense. It is rather interesting that many such critical objects can be defined on the basis of a single transformation $\alpha \mapsto c_\alpha$ which for every countable ordinal α picks a set c_α of smaller ordinals of minimal possible order-type subject to the requirement that $\alpha = \sup(c_\alpha)$. This gives us a way to approach higher ordinals from below in various recursive definitions. For example, given two ordinals $\beta > \alpha$ one can step from β down towards α along the set c_β. More

precisely, one can define *the step* from β towards α as the minimal point ξ of c_β such that $\xi \geq \alpha$. Let $c_\beta(\alpha)$, or simply $\beta(\alpha)$, denote this ordinal. Now one can step further from $\beta(\alpha)$ towards α and get $\beta(\alpha)(\alpha)$ $(= (\beta(\alpha))(\alpha))$, and so on. This leads us to the notion of a *minimal walk* from β to α

$$\beta > \beta(\alpha) > \beta(\alpha)(\alpha) > \cdots > \beta(\alpha)(\alpha)\cdots(\alpha) = \alpha.$$

Let $\lceil \beta(\alpha) \rceil$ denote *the weight* of the step from β towards α, the cardinality of the set of all $\xi \in c_\beta$ such that that $\xi < \alpha$. This gives us a way to define various *distances* between α and β:

1. $\|\alpha\beta\| = \max\{\lceil \beta(\alpha) \rceil,\ \|\alpha\beta(\alpha)\|, \|\xi\alpha\| : \xi \in c_\beta, \xi < \alpha\}$,
2. $\|\alpha\beta\|_1 = \max\{\lceil \beta(\alpha) \rceil,\ \|\alpha\beta(\alpha)\|_1\}$,
3. $\|\alpha\beta\|_2 = \|\alpha\beta(\alpha)\| + 1$.

Thus, $\|\alpha\beta\|_2$ is the number of steps in the minimal walk from β towards α, and $\|\alpha\beta\|_1$ is the maximal weight of a single step in that walk. On the other hand, $\|\alpha\beta\|$ is a much finer distance function which has the following interesting *subadditivity properties* for every triple $\gamma > \beta > \alpha$ of countable ordinals:

4. $\|\alpha\gamma\| \leq \max\{\|\alpha\beta\|, \|\beta\gamma\|\}$,
5. $\|\alpha\beta\| \leq \max\{\|\alpha\gamma\|, \|\beta\gamma\|\}$.

Moreover, we also have the following important *coherence properties* for every pair $\beta > \alpha$ of countable ordinals and every integer n (see §4 below where this is used):

6. $\|\xi\alpha\| = \|\xi\beta\|$ and $\|\xi\alpha\|_1 = \|\xi\beta\|_1$ for all but finitely many $\xi < \alpha$.
7. $\|\xi\alpha\| > n$ and $\|\xi\alpha\|_1 > n$ for all but finitely many $\xi < \alpha$,

The minimal walk from β to α can be coded by the sequence $\rho_0(\alpha, \beta)$ of weights of the corresponding steps, or more precisely:

8. $\rho_0(\alpha, \beta) = \lceil \beta(\alpha) \rceil^\frown \rho_0(\alpha, \beta(\alpha))$.

This leads us to another distance function whose values are countable ordinals rather than non-negative integers:

9. $\Delta_0(\alpha, \beta) = \min\{\xi : \rho_0(\xi, \alpha) \neq \rho_0(\xi, \beta)\}$.

Let $\mathrm{Tr}(\alpha, \beta)$ denote the places visited during the walk from β to α i.e., the set of all $\xi \leq \beta$ for which $\rho_0(\xi, \beta)$ is an initial segment of $\rho_0(\alpha, \beta)$. This leads us now to the first basic *square-bracket operation* on ω_1:

10. $[\alpha\beta] = \min(\mathrm{Tr}(\xi, \beta) \setminus \alpha)$ where $\xi = \Delta_0(\alpha, \beta)$.

Thus $[\alpha\beta]$ is the member β_i on the path $\mathrm{Tr}(\alpha, \beta) = \{\beta = \beta_0 > \beta_1 > \ldots > \beta_n = \alpha\}$ furthest from β subject to the requirement that there exists $\alpha = \alpha_0 > \alpha_1 > \cdots > \alpha_i$ such that $\rho_0(\alpha_j, \alpha) = \rho_0(\beta_j, \beta)$ and $c_{\beta j} \cap \alpha = c_{\alpha_j} \cap \alpha_i$ for all $j < i$.

Theorem 1 . [24] *For every uncountable subset X of ω_1, the set of all ordinals of the form $[\alpha\beta]$ for some $\alpha < \beta$ in X contains a closed and unbounded subset of ω_1.*

This operation has been used in constructions of various mathematical objects of complex behavior such as groups, geometries, and Banach spaces ([20, 21], [7, 8]). The usefulness of $[\cdot]$ in these constructions is based on the fact that $[\cdot]$ reduces questions about uncountable subsets of ω_1 (the subsets one usually talks about) to questions about closed and unbounded subsets of ω_1 which are much easier to handle. Recent metamathematical results of Woodin [35] give some explanation to this phenomenon.

2 Binary relations

For a given subset A of ω_1, let R_A denote the set of all pairs (α, β) of countable ordinals such that $[\alpha\beta] \in A$. Then one can show that the family R_A $(A \subseteq \omega_1)$ of binary relations exhibits a too complex behavior if we are to choose isomorphic embeddings as connecting maps. It turns out that in this context the right choice of connecting maps is a reduction introduced long time ago by J.W. Tukey [32] for quite a different purpose. Given two binary relations R and S, we say that R is *Tukey reducible* to S, and write $R \leq_T S$, if there exist maps $f : \mathrm{dom}(R) \to \mathrm{dom}(S)$ and $g : \mathrm{ran}(S) \to \mathrm{ran}(R)$ such that for every $r \in R$ and $s \in S$,

11. $(f(r), s) \in S$ implies $(r, g(s)) \in R$.

Tukey considered this reduction only in the case of directed sets as only they are relevant to the theory of Moore–Smith convergence he was studying. The definition is, however, as meaningful in the general case (see [33] and [31] for other variations). While the square bracket operation $[\cdot]$ defined in the previous section can be used to show the extreme complexity also in this generality, the critical objects of the subclass of all transitive binary relations seem to remain critical also in this bigger class. Some examples of critical transitive relations are the usual well-ordering relation on ω_1, which we denote by ω_1, or the direct sum $\omega \cdot \omega_1$ of countably many copies of ω_1. The equality relation $=$ on ω_1 is of course the maximal binary reflexive relation on ω_1. Another critical structure is the family FIN_{ω_1} of all finite subsets of ω_1 ordered by inclusion. That these are indeed some of the critical structures for the whole class of binary relations would follow from the positive answer to the following problem.

Conjecture 1 . [PFA] *For every binary relation R on ω_1, either $R \leq_T \omega \cdot \omega_1$ or $\mathrm{FIN}_{\omega_1} \leq_T R$.*

This seems to be a rather strong conjecture but it may not be so unreasonable since we were able to prove it in the case of transitive relations ([27]). An essentially equivalent Ramsey-theoretic reformulation of this conjecture has been around since the early 1970's in various correspondences between F. Galvin, K. Kunen, R. Laver and others (see [12]): For every family G of unordered pairs of countable ordinals either there exist an uncountable subset of ω_1 which avoids G, or else there exist

uncountable subsets A and B of ω_1 such that $\{\alpha, \beta\} \in G$ whenever $\alpha \in A$, $\beta \in B$ and $\alpha < \beta$. There are a number of well-known open problems in other areas of mathematics which are awaiting the solution to this conjecture. One of them is the following duality conjecture between the closure and covering properties of subsets of an arbitrary regular topological space X (see [25] or section 5 below).

CONJECTURE 2. [PFA] *A family of open subsets of X contains a countable subfamily with the same union if and only if an arbitrary subset of X contains a countable subset with the same closure.*

3 TRANSITIVE RELATIONS

Tukey introduced his reduction in order to illuminate the theory of Moore-Smith convergence, so he was concerned only with upwards-directed partially ordered sets. He was already able to isolate the following five directed sets as pairwise inequivalent under the equivalence relation induced by his reducibility:

$$1,\ \omega,\ \omega_1,\ \omega \times \omega_1 \quad \text{and} \quad \mathrm{FIN}_{\omega_1}.$$

It turns out that this is indeed the list of *all* critical directed sets on the domain ω_1 as the following result shows.

THEOREM 2. [23][PFA] *Every directed set on ω_1 is Tukey equivalent to one of the basic five 1, ω, ω_1, $\omega \times \omega_1$, FIN_{ω_1}.*

A number of years later we were able to extend this result to arbitrary transitive relations on ω_1. To simplify the notation, let $D_0 = 1$, $D_1 = \omega$, $D_2 = \omega_1$, $D_3 = \omega \times \omega_1$ and $D_4 = \mathrm{FIN}_{\omega_1}$, and let $m \cdot D$ denote the direct sum of m copies of D.

THEOREM 3. [27][PFA] *Every transitive relation on ω_1 is Tukey equivalent to one of the following where n_i's are all non-negative integers:*

(a) $n_0 \cdot D_0 \oplus n_1 \cdot D_1 \oplus n_2 \cdot D_2 \oplus n_3 \cdot D_3 \oplus n_4 D_4$,

(b) $\omega \cdot D_0 \oplus n_2 \cdot D_2 \oplus n_3 \cdot D_3 \oplus n_4 \cdot D_4$,

(c) $\omega \cdot D_2 \oplus n_4 \cdot D_4$,

(d) $\omega \cdot D_4$,

(e) $=$.

The class of transitive relations that one can associate with the reals is considerably richer than the class of all transitive relations on the domain ω_1 and the analogue of Theorem 3 for this domain is false. For example, Isbell [16] showed that the Banach lattice ℓ^1 and the lattice $\mathbb{N}^{\mathbb{N}}$ are not equivalent to either of the five basic directed sets (and moreover, not equivalent to each other). In [10], Fremlin realized that Tukey reductions (or non-reductions) between the classical objects of Real Analysis and Measure Theory are meaningful even from the point of view of these

areas of mathematics. Mathematical structures that one finds in these areas are often associated with studies of certain notions of smallness, or more precisely, *ideals* in Boolean rings such as, for example, the power-set of the reals or the integers. Many of them can in fact be represented as *analytic P-ideals* on $\mathbb{N}$ i.e., ideals of the power-set of $\mathbb{N}$, that are σ-directed modulo FIN, the ideal of finite subsets of $\mathbb{N}$, and given in some explicit way (or more precisely representable as continuous images of the irrationals when viewed as subspaces of the Cantor set $2^{\mathbb{N}}$). Recently, a number of unexpected connections in this class of ideals have been discovered (see, for example, [22], [28]). One of them is the following result which shows that FIN, $\mathbb{N}^{\mathbb{N}}$ and ℓ^1 (all representable as analytic P-ideals on $\mathbb{N}$) are indeed critical members of this class.

Theorem 4 . [29] *If J is an analytic P-ideal on $\mathbb{N}$, then either J is generated over* FIN *by a single subset of $\mathbb{N}$ or else $\mathbb{N}^{\mathbb{N}} \leq_T J \leq_T \ell^1$.*

4 Linear orderings

A *basis* for a class $\mathcal{X}$ of linear orderings is any of its subclasses $\mathcal{Y}$ with the property that every member of $\mathcal{X}$ contains an isomorphic copy of a member of $\mathcal{Y}$. Clearly ω_1 and its converse ω_1^* will be members of any basis for uncountable linear orderings so we may restrict our attention to the class $\mathcal{R}$ of uncountable linear orderings orthogonal to both ω_1 and ω_1^*. The class $\mathcal{R}$ itself naturally splits into the subclass $\mathcal{S}$ of separable orderings and its relative orthogonal $\mathcal{A} = \mathcal{S}^{\perp} \cap \mathcal{R}$ which turns out to be nonempty. The Basis Problem for $\mathcal{S}$ was solved by Baumgartner [4] who has actually proved the following more precise result where $\mathcal{S}^d$ denotes the family of all $L \in \mathcal{S}$ with the property that every nontrivial interval of L has exactly $\aleph_1$ many elements.

Theorem 5 . [4][PFA] *Every two orderings from $\mathcal{S}^d$ are isomorphic.*

The Ramsey-theoretic analysis of Baumgartner's proof turned out to be quite rewarding. Out of a number of closely related coloring principles discovered over the years (see [1], [25]), the following asymmetric principle of open colorings turned out to be quite useful even in problems far beyond the original scope (see e.g. [25], [9]):

[OCA] For every separable metric space X and every open symmetric and irreflexive relation R on X, either X can be decomposed into countably many sets that avoid R, or else X contains an uncountable subset Y such that every two distinct members of Y are related in R.

To see the relevance of OCA to the Basis Problem of $\mathcal{S}$ consider two uncountable separable (dense) linear orderings A and B. Let $X = A \times B$ and let $R = \{((a_0, b_0), (a_1, b_1)) : a_0 \neq a_1,\ b_0 \neq b_1, (a_0 <_A a_1 \equiv b_0 <_B b_1)\}$. This is indeed an open relation with respect to the natural order topology on X. It is a general fact that the cartesian product of two orderings from $\mathcal{S}(\cup\{\omega_1, \omega_1^*\})$ cannot be decomposed into countably many chains so the first alternative of OCA fails in this situation. The second alternative of OCA gives us an embedding of

an uncountable subset of A into B. This shows that no two members of $\mathcal{S}$ are orthogonal to each other which is a half way towards the solution of the Basis Problem for separable linear orderings. The progress on the Basis Problem for the orthogonal $\mathcal{A} = (\mathcal{S} \cup \{\omega_1, \omega_1^*\})^\perp$ has been much slower. It was initiated by the following brilliant question of R.S. Countryman [6]: Is there an uncountable linear ordering whose cartesian square is the union of countably many chains? We have already remarked that the class $\mathcal{C}$ of Countryman's orderings (if nonempty) must be included in $\mathcal{A}$. Note also that every $C \in \mathcal{C}$ is orthogonal to its reverse C^* (which also belongs to $\mathcal{C}$). Thus, unlike to the case of separable orderings, if the class $\mathcal{C}$ is nonempty, we cannot hope for a single-element basis in this case. In [19], Shelah established that $\mathcal{C}$ is indeed a nonempty class of orderings and posed the following interesting conjecture.

CONJECTURE 3. [PFA] *The class $\mathcal{C}$ is a basis for $\mathcal{A}$.*

This together with Baumgartner's result about the class of separable orderings leads us to the following equivalent conjecture.

CONJECTURE 4. [PFA] *The class of all uncountable linear orderings has a 5-element basis $\omega_1, \omega_1^*, B, C, C^*$ where B is some uncountable set of reals and where C is any uncountable linear ordering whose cartesian square is the union of countably many chains.*

While Shelah's conjecture is still widely open one can still try to find the Ramsey-theoretic principle that lies behind. This search turned out to be quite simple and (unlike the case of OCA above) the resulting coloring principle turned out to be equivalent to the statement that $\mathcal{A}$ has a 2-element basis. The analysis is based on a fundamental concept introduced more than 60 years ago by Đ. Kurepa [17], a concept whose relevance in constructing critical uncountable structures has been realized only in recent times. This is the concept of a (special) *Aronszajn tree* (A-tree, in short). An A-tree is simply a transformation a which to every countable ordinal ξ associates its enumeration $a_\xi : \xi \to \omega$ (one-to-one or finite-to-one map) with the property that for a given countable ordinal α the set of restrictions $\{a_\xi \restriction \alpha : \xi < \omega_1\}$ is at most countable. The set $A_a = \{a_\xi : \xi < \omega_1\}$, ordered lexicographically, is a typical member of the class $\mathcal{A}$. Clearly we can view the transformation a also as a two-place distance function $a(\alpha, \beta) = a_\beta(\alpha)$ which makes this concept relevant in descriptions of other critical structures as well. For example, it can be seen that the distance functions $\|\cdot\|$, $\|\cdot\|_1$ and ρ_0 considered in the first section are all Aronszajn. However, our analysis from that section also suggests considering the notion of a *coherent A-tree* i.e., an A-tree $a_\xi : \xi \to \omega_1$ of finite-to-one mappings which has the following property for all $\alpha < \beta$:

12. $a_\alpha(\xi) = a_\beta(\xi)$ for all but finitely many $\xi < \alpha$.

The importance of this notion can be seen from the following

THEOREM 6. [24] *The cartesian square of any lexicographically ordered coherent A-tree is the union of countably many chains.*

In other words, a coherent A-tree immediately gives us a critical member of the class of uncountable linear orderings. It is therefore not surprising that this notion will also give us a Ramsey-theoretic reformulation of Shelah's Conjecture. Recall the notion of distance function $\Delta(\alpha,\beta) = \min\{\xi : a_\alpha(\xi) \neq a_\beta(\xi)\}$ that one associates to an A-tree $a_\xi : \xi \to \omega$ of enumerations. Thus, $\Delta(\beta,\gamma) > \Delta(\alpha,\beta)$ reads as "β is closer to γ than to α". So it is natural to call a binary relation R on ω_1 an *a-open relation* if

13. $R(\alpha,\beta)$ and $\Delta(\beta,\gamma) > \Delta(\alpha,\beta)$ imply $R(\alpha,\gamma)$,

whenever α,β and γ are pairwise distinct countable ordinals. However, this is not quite analogous to the situation in the Cantor set $2^{\mathbb{N}}$ since it easily follows that in the present case the complement of an a-open relation on ω_1 is also a-open.

Theorem 7. [2][PFA] *The class $\mathcal{A}$ of linear orderings has a 2-element basis if and only if for every a-open symmetric relation R on ω_1 there is an uncountable subset X of ω_1 such that $X^2\backslash$diagonal is included either in R or in its complement.*

It should be remarked that if in this Ramsey-theoretic principle we use another A-tree as a parameter which describes the notion of openness we get an equivalent formulation.

5 Topological spaces

While this is an area of considerable generality and wealth of examples there seem to be some patterns in descriptions of these examples. Pathological spaces almost always contain uncountable discrete subspace (a copy of $D(\omega_1)$, the discrete space on ω_1) and this is usually at the root of their complexity. On the other hand, spaces that do not contain $D(\omega_1)$ are usually obtained as mild modifications of separable metric topologies. A typical such example is the split-interval of Alexandroff and Urysohn [3] or its subspaces. It is obtained by doubling each point of the unit interval $I = [0,1]$, or more precisely the space $I \times 2$ with the lexicographic order topology. Note that the split-interval is a $2-$to-1-preimage of the unit interval so the two spaces share many properties in common. On the other hand, they are orthogonal to each other since clearly the split-interval contains no uncountable metrizable subspace.

Conjecture 5. [PFA] *The class of uncountable regular spaces has a 3-element basis consisting of $D(\omega_1), B$ and $B \times \{0\}$, where B is some uncountable subset of the unit interval and where $B\times\{0\}$ is considered as a subspace of the split-interval.*

This is a rather bold conjecture based on a question first considered by Gruenhage [14]. Note that Conjecture 2, about the equivalence of certain closure and covering properties in regular spaces considered above, is an immediate consequence of Conjecture 5. In fact, Conjecture 5 has several other weakenings which if true would still be of considerable interest. For example, if we restrict ourselves to compact spaces we get the following consequence of Conjecture 5 which is related to a problem first asked by D.H. Fremlin (see [11] or [14]).

CONJECTURE 6. [PFA] *Every compact space which does not contain $D(\omega_1)$ admits an at most 2−to−1 continuous map onto a compact metric space.*

It is easily seen that this conjecture is in fact equivalent to Conjecture 5 restricted to regular spaces that can be compactified avoiding copies of $D(\omega_1)$. Note also that this conjecture solves the Basis Problem for compact spaces: A compact space K is either metrizable, or it contains a copy of $D(\omega_1)$, or an uncountable subspace of the split interval of the form $B \times 2$. Note that from such a subspace $B \times 2$ of K one can easily build an uncountable *biorthogonal system* in the Banach space $\mathcal{C}(K)$ of continuous real-valued functions on K i.e., a system (x_b, x_b^*) $(b \in B)$ of elements of $\mathcal{C}(K) \times \mathcal{C}(K)^*$ with uniformly bounded norms such that

14. $x_b^*(x_b) = 1$ and $x_b^*(x_a) = 0$ whenever $a \neq b$.

Another interesting consequence of Conjecture 6 is the fact that if the product of two compact spaces does not contain a copy of $D(\omega_1)$ then one of the factors must be metrizable.

It is interesting that the Ramsey-theoretic principles needed to solve these two conjectures are some forms of OCA discussed above in connection with the Basis Problem for separable linear orderings. This is not surprising since a separable linear ordering shows up in Conjecture 5 as a member of a basis for uncountable regular spaces. However, to solve these two conjectures one needs a much stronger form of OCA valid for a class of spaces larger than the class of separable metric spaces occurring in the original form (see [14], [25]). Lacking the methods to attack these Ramsey-theoretic problems, it is natural to try to test these two conjectures by either proving some of the consequences or by restricting ourselves to some concrete class of spaces. We have two results of this sort that show a surprising degree of accuracy in these conjectures.

THEOREM 8. [26][PFA] *A compact space K is metrizable if and only if the Banach space $\mathcal{C}(K)$ contains no uncountable biorthogonal system.*

Pointwise compact sets of Baire class-1 functions showed up perhaps for the first time in the two selection theorems of E. Helly about families of monotonic functions on the unit intervals. In more recent years the interest was renewed after Odell and Rosenthal [18] proved that a separable Banach space E contains no copy of ℓ^1 if and only if the unit ball of E^{**} with the weak* topology is such a compactum when considered as a family of functions defined on the unit ball of E^*. A number of deep general results about this class of spaces were established soon afterwards by Bourgain, Fremlin, Talagrand [5] and Godefroy [13]. Since the split-interval can be represented as a compactum lying inside the first Baire class, it is natural to try to test the validity of Conjecture 6 on this class of compact spaces. The following result shows that for this class of compact spaces Conjecture 6 is indeed true even in some stronger form.

THEOREM 9. [30] *Every pointwise compact subset of the first Baire class which does not contain a copy of $D(\omega_1)$ admits an at most 2−to−1 continuous map onto a metric compactum, and moreover, it is either metric itself or it contains a full copy of the split-interval.*

References

[1] U. Abraham, M. Rubin, and S. Shelah, On the consistency of some partition theorems for continuous colorings, *Ann. Pure. Appl. Logic* 29 (1985), 123–206.

[2] U. Abraham and S. Shelah, Isomorphism types of Aronszajn trees, *Israel J. Math.* 50 (1985), 75–113.

[3] P. Alexandroff and P. Urysohn, Mémoire sur les espaces topologiques compacts, *Verh. Akad. Wetensch., Amsterdam* 14 (1929).

[4] J. Baumgartner, All $\aleph_1$-dense sets of reals can be isomorphic, *Fund. Math.* 79 (1973), 101–106.

[5] J. Bourgain, D.H. Fremlin and M. Talagrand, Pointwise compact sets of Baire-measurable functions, *Amer. J. Math.* 100 (1978), 845–886.

[6] R.S. Countryman, Spaces having a σ-monotone base, preprint 1970.

[7] P. Erdös, S. Jackson and R.D. Mauldin, On partitions of lines and spaces, *Fund. Math.* 145 (1994), 101–119.

[8] _____, On infinite partitions of lines and spaces, *Fund. Math.* 152 (1997), 75–95.

[9] I. Farah, Analytic quotients, submitted to Memoirs Amer. Math. Soc.

[10] D.H. Fremlin, The partially ordered sets of measure theory and Tukey's ordering, *Note di Matematica* 11 (1991), 177–219.

[11] D.H. Fremlin, Problems, March 1, 1998.

[12] F. Galvin, letter of November 12, 1980.

[13] G. Godefroy, Compacts de Rosenthal, *Pacific. J. Math.* 91 (1980), 293–306.

[14] G. Gruenhage, Perfectly normal compacta and some partition problems, in: Open Problems in Topology (J. van Mill et al ed) Elsevier Sci. Publ. 1990.

[15] F. Hausdorff, Summen von $\aleph_1$ Mengen, *Fund. Math.* 26 (1936), 241–255.

[16] J. R. Isbell, Seven cofinal types, *J. London Math. Soc.* 4 (1972), 651–654.

[17] Đ. Kurepa, Ensembles linéaires et une classe de tableaux ramifiés (tableaux ramifiés de M. Aronszajn), *Publ. Math. Univ. Belgrade* 6 (1936), 241–255.

[18] E. Odell and H.P. Rosenthal, A double-dual characterization of separable Banach spaces containing ℓ^1, *Israel J. Math.* 20 (1975), 375–384.

[19] S. Shelah, Decomposing uncountable square into countably many chains, *J. Comb. Theory (A)* 21 (1976), 110–114.

[20] S. Shelah and J. Steprans, Extraspecial p-groups, *Ann. Pure Appl. Logic,* 34 (1987), 87–97.

[21] S. Shelah and J. Steprans, A Banach space on which there are few operators, *Proc. Amer. Math. Soc.*, 104 (1988), 101–105.

[22] S. Solecki, Analytic ideals, *Bull. Symb. Logic* 2 (1996), 339–348.

[23] S. Todorcevic, Directed sets and cofinal types, *Trans. Amer. Math. Soc.* 290 (1985), 711–723.

[24] _____, Partitioning pairs of countable ordinals, *Acta Mathematica*, 159 (1987), 261–294.

[25] _____, Partition problems in topology, Amer. Math. Soc., Providence, 1989.

[26] _____, Irrendundant sets in Boolean algebras, *Trans. Amer. Math. Soc.,* 339 (1993), 35–44.

[27] _____, A classification of transitive relations on ω_1, *Proc. London Math. Soc.* 73 (1996), 501–533.

[28] _____, Analytic gaps, *Fund. Math.* 150 (1996), 55–66.

[29] _____, Definable ideals and gaps in their quotients, in: *Set Theory: Techniques and Applications* (C.A. DiPrisco et al, eds), Kluwer Acad. Press 1997, pp. 213–226.

[30] _____, Compact sets of the first Baire class, preprint 1997.

[31] S. Todorcevic and J. Zapletal, On the Alaoglu–Birkhoff equivalence of posets, preprint 1997.

[32] J.W. Tukey, Convergence and uniformity in topology, Princeton Univ. Press 1940.

[33] P. Vojtáš, Galois–Tukey connections between explicit relations on classical objects of real analysis, *Israel Math. Conf. Proc.* vol. 6 (1993), 619–643.

[34] W.H. Woodin, Discontinuous homomorphisms of $C(\Omega)$ and Set Theory, Ph.D. Thesis, University of California, Berkeley, 1984.

[35] W.H. Woodin, The axiom of determinacy, forcing axioms and the nonstationary ideal, in preparation.

C.N.R.S.,
Université Paris VII, France.

Matematicki Institut,
Beograd, Yugoslavia.

University of Toronto,
Toronto, Canada.

Geometry of Infinitesimal Group Schemes

Eric M. Friedlander

1991 Mathematics Subject Classification: 14L15, 17B50
Keywords and Phrases: group schemes, modular representation theory, Ext-groups, functor cohomology

We consider affine group schemes G over a field k of characteristic $p > 0$. Equivalently, we consider finitely generated commutative k-algebras $k[G]$ (the coordinate algebra of G) endowed with the structure of a Hopf algebra. The group scheme G is said to be *finite* if $k[G]$ is finite dimensional (over k) and a finite group scheme is said to be *infinitesimal* if the (finite dimensional) algebra $k[G]$ is local. A *rational* $\mathbf{G}$*-module* is a k-vector space endowed with the structure of a comodule for the Hopf algebra $k[G]$. The abelian category of rational G-modules has enough injectives, so that $Ext^i_G(M, N)$ is well defined for any pair of rational G-modules M, N and any non-negative integer i. Unlike the situation in characteristic 0, this category has many non-trivial extensions reflected by the cohomology groups we study.

We sketch recent results concerning the cohomology algebras $H^*(G, k)$ and the $H^*(G, k)$-modules $Ext^*_G(M, M)$ for infinitesimal group schemes G and finite dimensional rational G-modules M. These results, obtained with Andrei Suslin and others, are inspired by analogous results for finite groups. Indeed, we anticipate but have yet to realize a common generalization to the context of finite group schemes of our results and those for finite groups established by D. Quillen [Q1], J. Carlson [C], G. Avrunin and L. Scott [A-S], and others. Although there is considerable parallelism between the contexts of finite groups and infinitesimal group schemes, new techniques have been required to work with infinitesimal group schemes. Since the geometry first occuring in the context of finite groups occurs more naturally and with more structure in these recent developments, we expect these developments to offer new insights into the representation theory of finite groups.

The most natural examples of infinitesimal group schemes arise as *Frobenius kernels* of affine algebraic groups G over k (i.e., affine group schemes whose coordinate algebras are reduced). Recall that the Frobenius map

$$F : G \to G^{(1)}$$

of an affine group scheme is associated to the natural map $k[G]^{(1)} \to k[G]$ of k-algebras. (For any k-vector space V and any positive integer r, the r-th Frobenius

twist $V^{(r)}$ is the k-vector space obtained by base change by the p^r-th power map $k \to k$.) The r-th Frobenius kernel of G, denoted $G_{(r)}$, is defined to be the kernel of the r-th iterate of the Frobenius map, $ker\{F^r : G \to G^{(r)}\}$; thus,

$$k[G_{(r)}] = k[G]/(X^{p^r}; X \in \mathcal{M}_e)$$

where $\mathcal{M}_e \subset k[G]$ is the maximal ideal at the identity of G.

If M is an irreducible rational G-module for an algebraic group G, then $M^{(r)}$ is again irreducible; moreover, for $r \neq s$, $M^{(r)}$ is not isomorphic to $M^{(s)}$. It is easy to see that a rational G-module N is the r-th twist of some rational G-module M if and only if $G_{(r)}$ acts trivially on N. Thus, much of the representation theory of an algebraic group G is lost when rational G-modules are viewed by restriction as $G_{(r)}$-modules. On the other hand, in favorable cases the category of rational G-modules is equivalent to the category locally finite modules for the hyperalgebra of G, the ind-object $\{G_{(r)}, r \geq 0\}$ (see, for example, [CPS]).

The special case of the 1st infinitesimal kernel $G_{(1)}$ of an algebraic group G is a familiar object. The (k-linear) dual $k[G_{(1)}]^{\#}$ of the coordinate algebra of $G_{(1)}$ is naturally isomorphic to the *restricted enveloping algebra* of the p-restricted Lie algebra $g = Lie(G)$. Thus, the category of rational $G_{(1)}$-modules is naturally isomorphic to the category of restricted g-modules. The results we describe below are natural generalizations and refinements of results earlier obtained by the author and Brian Parshall for p-restricted Lie algebras (see, for example, [FP1], [FP2], [FP3], [FP4]).

Throughout our discussion, unless otherwise specified, k will denote an arbitrary (but fixed) field of characteristic $p > 0$ and the finite group schemes we consider will be finite over k.

§1. Finite Generation and Strict Polynomial Functors

The following theorem proved by the author and Andrei Suslin is fundamental in its own right and aspects of its proof play a key role in further developments. This result, valid for an arbitrary finite group scheme, is a common generalization of the finite generation of the cohomology of finite groups proved by L. Evens [E] and B. Venkov [V], and the finite generation of restricted Lie algebra cohomology (cf. [FP1].

THEOREM 1.1 [F-S]. *Let G be a finite group scheme over k and let M be a finite dimensional rational G-module. Then $H^*(G,k)$ is a finitely generated k-algebra and $H^*(G,M)$ is a finite $H^*(G,k)$-module.*

After base extension via some finite field extension K/k, G_K as a finte group scheme over K is a semi-direct product of a finite group by an infinitesimal group scheme. Using classical results about finite generation of cohomology of finite groups (cf. [E]) and the fact that any infinitesimal group scheme G admits an embedding in some $GL_{n(r)}$, we find that Theorem 1.1 is implied by the following more concrete assertion.

THEOREM 1.2 [F-S]. *For any $n > 1, r \geq 1$, there exist rational cohomology classes*

$$e_r \in H^{2p^{r-1}}(GL_n, gl_n^{(r)})$$

which restrict non-trivially to

$$H^{2p^{r-1}}(GL_{n(1)}, gl_n^{(r)}) = H^{2p^{r-1}}(GL_{n(1)}, k) \otimes gl_n^{(r)},$$

where gl_n denotes the adjoint representation of the algebraic general linear group GL_n. Moreover, these classes e_r induce a GL_n-equivariant map of k-algebras

$$\phi : \bigotimes_{i=1}^{r} S^*((gl_n^{(r)})^{\#}[2p^{i-1}]) \to H^*(GL_{n(r)}, k),$$

where $S^*((gl_n^{(r)})^{\#}[2p^{i-1}])$ denotes the symmetric algebra on the vector space $gl_n^{(r)\#}$ placed in degree $2p^{i-1}$, with the property that $H^*(GL_{n(r)}, k)$ is thereby a finite module over $\bigotimes_{i=1}^{r} S^*((gl_n^{(r)})^{\#}[2p^{i-1}])$.

We may interpret e_1 as the group extension associated to the general linear group over the ring $W_2(k)$ of Witt vectors of length 2 over k:

$$1 \to gl_n \to GL_{n,W_2(k)} \to GL_{n,k} \to 1,$$

where $GL_{n,k}$ denotes the algebraic general linear group GL_n over k (with k made explicit). Alternatively, we can view $e_1 \in Ext^2_{GL_n}(I_n^{(1)}, I_n^{(1)})$ as the extension of rational GL_n-modules

$$0 \to I_n^{(1)} \to S^p(I_n) \to \Gamma^p(I_n) \to I_n^{(1)} \to 0 \tag{1.2.1}$$

where I_n denotes the canonical n-dimensional representation of GL_n, $S^p(I_n)$ denotes the p-th symmetric power of I_n defined as the coinvariants under the action of the symmetric group Σ_p on the p-th tensor power $I_n^{\otimes p}$, and $\Gamma^p(I_n)$ denotes the p-th divided power of I_n defined as the invariants of Σ_p on $I_n^{\otimes p}$. It would be of considerable interest to give an explicit description for e_r for $r \geq 2$; even for e_2, this is a considerable challenge, for we require an extension of $I_n^{(2)}$ by itself of length $2p$.

The core of the proof of Theorem 2 utilizes standard complexes, the exact Koszul complex and the DeRham complex whose cohomology is known by a theorem of P. Cartier [Ca]. Our strategy is taken from V. Franjou, J. Lannes, and L. Schwartz (cf. [FLS]). It appears to be essential to first work "stably with respect to n" rather than work directly with rational GL_n-modules.

Indeed, we introduce the concept of a *strict polynomial functor* on finite dimensional k-vector spaces and our computations of Ext-groups occur in this abelian category $\mathcal{P} = \mathcal{P}_k$ (with enough projective and injective objects). There is a natural transformation

$$\mathcal{P} \to \mathcal{F}$$

from $\mathcal{P}$ to the category $\mathcal{F}$ of all functors from finite dimensional k-vector spaces to k-vector spaces. If k is a finite field, this "forgetful" natural transformation is not faithful. If $F \in \mathcal{F}$, then the difference functor $\Delta(F)$ is defined by $\Delta(F)(V) = ker\{F(V \oplus k) \to F(V)\}$. A functor $F \in \mathcal{F}$ is said to be polynomial if $\Delta^N(F) = 0$ for $N >> 0$; each strict polynomial functor when viewed in $\mathcal{F}$ is a polynomial functor. There is a well defined formulation of the degree of $P \in \mathcal{P}$ which has the property that this is greater than or equal to the degree of P when viewed as a polynomial functor in $\mathcal{F}$. One very useful property of $\mathcal{P}$ is that it splits as a direct sum of categories $\mathcal{P}_d$ of strict polynomial functors homogeneous of degree d.

The extension (1.2.1) arises from the extension of strict polynomial functors of degree p

$$0 \to I^{(1)} \to S^p \to \Gamma^p \to I^{(1)} \to 0$$

by evaluation on the vector space k^n. We prove that Ext_{GL_n}-groups can be computed as Ext-groups in the category of strict polynomial functors. Indeed, in a recent paper with V. Franjou and A. Scorichenko and A. Suslin, we prove the following theorem (a weak version of which was proved independently by N. Kuhn [K]). This theorem incorporates earlier results of the author and A. Suslin [A-S] as well as W. Dwyer's stability theorem [D] for the cohomology of the finite groups $GL_n(\mathbb{F}_q)$.

THEOREM 1.3 [FFSS]. *Set k equal to the finite field $\mathbb{F}_q$ for q a power of p. Let $\mathcal{P}_{\mathbb{F}_q}$ denote the category of strict polynomial functors on finite dimensional $\mathbb{F}_q$-vector spaces and let $\mathcal{F}_{\mathbb{F}_q}$ denote the category of polynomial functors from finite dimensional $\mathbb{F}_q$-vector spaces to $\mathbb{F}_q$-vector spaces. For $P, Q \in \mathcal{P}_{\mathbb{F}_q}$ of degree d, there is a natural commutative diagram of Ext-groups*

$$\begin{array}{ccc} Ext^i_{\mathcal{P}_{\mathbb{F}_q}}(P^{(r)}, Q^{(r)}) & \longrightarrow & Ext^i_{GL_{n,\mathbb{F}_q}}(P^{(r)}(\mathbb{F}_q^n), Q^{(r)}(\mathbb{F}_q^n)) \\ \downarrow & & \downarrow \\ Ext^i_{\mathcal{F}_{\mathbb{F}_q}}(P, Q) & \longrightarrow & Ext^i_{GL_n(\mathbb{F}_q)}(P(\mathbb{F}_q^n), Q(\mathbb{F}_q^n)) \end{array}$$

which satisfies the following:

- *(a.) The upper horizontal arrow is an isomorphism provided that $n \geq dp^r$. (This is valid with $\mathbb{F}_q$ replaced by an arbitary field k of characteristic p.)*
- *(b.) $Ext^i_{\mathcal{P}_{\mathbb{F}_q}}(P^{(r)}, Q^{(r)}) \equiv Ext^i_{\mathcal{P}_k}(P^{(r)}, Q^{(r)}) \otimes_{\mathbb{F}_q} k$ for any field extension $k/\mathbb{F}_q$.*
- *(c.) The lower horizontal arrow is an isomorphism for $n >> i, d$.*
- *(d.) The left vertical map is an isomorphism for $r \geq log_p(\frac{i+1}{2})$ provided that $q \geq d$.*

In proving part (c.) of Theorem 1.3 in [FFSS,App.1], A. Suslin verifies a conjecture of S. Betley and T. Pirashvili asserting that "stable K-theory equals topological Hochschild homology" for finite functors on $\mathbb{F}_q$-vector spaces.

We say that a sequence $A^0, A^1, \ldots, A^n, \ldots$ of functors (respectively, strict polynomial functors) is *exponential* (resp., exponential strict polynomial) if

$$A^0(k) = k, \quad A^n(V \oplus W) \equiv \bigoplus_{m=0}^{n} A^m(V) \otimes A^{n-m}(W).$$

Examples of exponential strict polynomial functors are the identity functor, the symmetric power, the divided power, the tensor power and their Frobenius twists. The following proposition, which first arose in the context of additive functors in the work of T. Pirashvili [P], appears to distill an essential feature of Ext-groups in categories of functors which does not hold for Ext-groups for rational G-modules. This property (and the injectivity of symmetric functors in the category $\mathcal{P}$) much facilitates computations.

PROPOSITION 1.4 [FFSS]. *Let A^* be an exponential strict polynomial functor and let B, C be strict polynomial functors. Then we have a natural isomorphism*

$$Ext^*_{\mathcal{P}}(A^n, B \otimes C) \simeq \bigoplus_{m=0}^{n} Ext^*_{\mathcal{P}}(A^m, B) \otimes Ext^*_{\mathcal{P}}(A^{n-m}, C).$$

Proposition 1.4 also holds if the category $\mathcal{P}$ of strict polynomial functors is replaced by the category $\mathcal{F}$. However, computations are made much easier in $\mathcal{P}$ because the splitting $\mathcal{P} = \oplus_{d \geq 0} \mathcal{P}_d$ implies that $Ext^*_{\mathcal{P}}(P, Q) = 0$ whenever P, Q are homogeneous strict polynomial functors of different degree.

For the computation of $Ext^*_{\mathcal{P}}(I^{(r)}, I^{(r)})$ needed to prove Theorem 2, we merely require the special case of Proposition 1.4 in which A is linear (i.e., $A^n = 0$ for $n > 1$): as a Hopf algebra,

$$Ext^*_{\mathcal{P}}(I^{(r)}, I^{(r)}) \simeq (k[t]/t^{p^r})^{\#},$$

generated as an algebra by the classes $e_r, e_{r-1}^{(1)}, \ldots, e_1^{(r-1)}$ each of which has p-th power equal to 0. The generality of Proposition 1.4 is employed in [FFSS] to give the complete calculation of the tri-graded Hopf algebras

$$Ext^*_{\mathcal{P}}(\Gamma^{*(j)}, S^{*(r)}), \quad Ext^*_{\mathcal{P}}(\Gamma^{*(j)}, \Lambda^{*(r)}), \quad Ext^*_{\mathcal{P}}(\Gamma^{*(j)}, \Gamma^{*(r)}),$$

$$Ext^*_{\mathcal{P}}(\Lambda^{*(j)}, S^{*(r)}), \quad Ext^*_{\mathcal{P}}(\Lambda^{*(j)}, \Lambda^{*(r)}), \quad Ext^*_{\mathcal{P}}(S^{*(j)}, S^{*(r)}),$$

as well as complete calculations of the corresponding $Ext_{\mathcal{F}_{\mathbb{F}_q}}$ tri-graded Hopf algebras.

§2. $H^*(G, k)$ AND 1-PARAMETER SUBGROUPS

For simple algebraic groups G, the author and B. Parshall [FP2], [FP4] computed $H^*(G_{(1)}, k)$ provided that the Coexter number $h(G)$ of G satisfies $h(G) < 3p - 1$ (except in type G_2, in which case the bound was $h(G) < 4p - 1$). This bound has been improved to $h(G) < p$ by H. Andersen and J. Jantzen [A-J]. The answer is intriguingly geometric: the algebra $H^*(G_{(1)}, k)$ is concentrated in even degrees and is isomorphic to the coordinate algebra of the nilpotent cone $\mathcal{N} \subset g = Lie(G)$. For "small" p, a precise determination of $H^*(G_{(1)}, k)$ appears to be quite difficult. Our model for the "computation" of $H^*(G, k)$ for an infinitesimal group scheme G is D. Quillen's identification [Q1] of the maximal ideal spectrum $SpecH^{ev}(\pi, k)$ of the cohomology of a finite group π as the colimit of the linear varieties $E \otimes_{\mathbf{F}_p} k$ indexed by the category of elementary abelian p-subgroups $E \subset \pi$.

We shall frequently use $|G|$ to denote the affine scheme associated to the commutative k-algebra $H^{ev}(G,k)$. In other words, $|G| = SpecH^{ev}(G,k)$. Following work of the author and B. Parshall, J. Jantzen [J] proved for any infinitesimal group scheme G of height 1 that the natural map $|G| \to g = Lie(G)$ has image the closed subvariety of p-nilpotent elements $X \in g$ (i.e., elements X such that $X^{[p]} = 0$).

In this section, we describe work of the author, Christopher Bendel, and Andrei Suslin which identifies the affine scheme $|G|$ up to universal finite homeomorphism for any infinitesimal group scheme G. Our identification is in terms of the scheme of "1-parameter subgroups" G. Recall that the infinitesimal group scheme G is said to have *height* r provided that r is the least integer for which G can be embedded as a closed subgroup of some $GL_{n(r)}$. By an abuse of notation, we call a homomorphism $\mathbb{G}_{a(r)} \to G$ an *infinitesimal of height* r 1*-parameter subgroup* of G. (We use the notation $\mathbb{G}_a$ to denote the additive group whose coordinate algebra is the polynomial ring in 1 variable). We verify that the functor on finite commutative k-algebras which sends the algebra A to the set of infinitesimal of height r 1-parameter subgroups of $G \otimes A$ over A is representable by an affine scheme $V_r(G)$:

$$Hom_{k-alg}(k[V_r(G)], A) = Hom_{A-group\ schemes}(\mathbb{G}_{a(r)} \otimes A, G \otimes A).$$

Here, $G \otimes A$ is the A-group scheme obtained from G over k by base change via $k \to A$. Clearly, $V_r(G) = V_r(G_{(r)})$.

For $G = GL_n$, $V_r(GL_n)$ is the closed reduced subscheme of $gl_n^{\times r}$ consisting of r-tuples of p-nilpotent, pairwise commuting matrices. A similar description applies for G equal to symplectic, orthogonal, and special linear algebraic groups and various closed subgroups of these groups [SFB1]. An explicit description of $V_r(G)$ is lacking for any arbitrary algebraic group G.

The following determination of $H^*(\mathbb{G}_{a(r)}, k)$ by E. Cline, B. Parshall, L. Scott, and W. van der Kallen is fundamental.

THEOREM 2.1 [CPSK].

1. *Assume that $p \neq 2$. Then the cohomology algebra $H^*(\mathbb{G}_a, k)$ is a tensor product of a polynomial algebra $k[x_1, x_2, ...]$ in generators x_i of degree 2 and an exterior algebra $\Lambda(\lambda_1, \lambda_2, ...)$ in generators λ_i of degree one. If $p = 2$, then $H^*(\mathbb{G}_a, k) = k[\lambda_1, \lambda_2, ...]$ is a polynomial algebra in generators λ_i of degree 1; in this case, we set $x_i = \lambda_i^2$.*
2. *Let $F : \mathbb{G}_a \to \mathbb{G}_a$ denote the Frobenius endomorphism, then $F^*(x_i) = x_{i+1}$, $F^*(\lambda_i) = \lambda_{i+1}$.*
3. *Let s be an element of k and use the same notation s for the endomorphism (multiplication by s) of $\mathbb{G}_a$. Then $s^*(x_i) = s^{p^i} x_i$, $s^*(\lambda_i) = s^{p^{i-1}} \lambda_i$.*
4. *Restriction of x_i and λ_i to $\mathbb{G}_{a(r)}$ is trivial for $i > r$. Denoting the restrictions of x_i and λ_i (for $i \leq r$) to $\mathbb{G}_{a(r)}$ by the same letter we have*

$$H^*(\mathbb{G}_{a(r)}, k) = k[x_1, ..., x_r] \otimes \Lambda(\lambda_1, ..., \lambda_r) \qquad p \neq 2$$
$$H^*(\mathbb{G}_{a(r)}, k) = k[\lambda_1, ..., \lambda_r] \qquad p = 2.$$

The class $x_r \in H^2(\mathbb{G}_{a(r)}, k)$ plays a special role for us, as can be seen both in the following proposition and in Theorem 3.1.

PROPOSITION 2.2 [SFB1]. *For any affine group scheme G, there is a natural homomorphism of graded commutative k-algebras*

$$\psi : H^{ev}(G,k) \to k[V_r(G)]$$

which multiplies degrees by $\frac{p^r}{2}$. For an element $a \in H^{2n}(G,k)$, $\psi(a)$ is the coefficient of x_r^n in the image of a under the composition

$$H^*(G,k) \to H^*(G,k) \otimes k[V_r(G)] = H^*(G \otimes k[V_r(G)], k[V_r(G)])$$

$$\xrightarrow{u^*} H^*(\mathbb{G}_{a(r)} \otimes k[V_r(G)], k[V_r(G)]) = H^*(\mathbb{G}_{a(r)}, k) \otimes k[V_r(G)],$$

where $u : \mathbb{G}_{a(r)} \otimes k[V_r(G)] \to G \otimes k[V_r(G)]$ is the universal infinitesimal of height r 1-parameter subgroup of G.

Alternatively, the map of schemes $\Psi : V_r(G) \to Spec H^{ev}(G,k)$ is obtained by sending a K-point of $V_r(G)$ coresponding to a 1-parameter subgroup $\nu : \mathbb{G}_{a(r)} \otimes K \to G \otimes K$ to the K-point of $Spec H^{ev}(G,k)$ corresponding to

$$eval_{x_1=1} \circ \epsilon_{K*} \circ \nu^* : H^{ev}(G,K) \to H^{ev}(\mathbb{G}_{a(r)}, K) \to H^{ev}(\mathbb{G}_{a(1)}, K) \to K.$$

Here, $\epsilon_* : H^*(\mathbb{G}_{a(r)}, k) \to H^*(\mathbb{G}_{a(1)}, k)$ is induced by the coalgebra map

$$\epsilon : k[\mathbb{G}_{a(r)}] = k[t]/t^{p^r} \to k[s]/s^p = k[\mathbb{G}_{a(1)}] \tag{2.2.1}$$

defined to be the k-linear map sending t^i to 0 if i is not divisible by p^{r-1} and to s^j if $i = jp^{r-1}$. (Note that ϵ_* sends $x_i \in H^2(\mathbb{G}_{a(r)}, k)$ to $x_1 \in H^2(\mathbb{G}_{a(1)}, k)$ if $i = r$ and to 0 otherwise.)

The following "geometric description" of $H^*(G,k)$ is the assertion that the homomorphism ψ of Proposition 2.2 is an isomorphism modulo nilpotents.

THEOREM 2.3 [SFB2]. *Let G be an infinitesimal group scheme of height $\leq r$. Then the kernel of the natural homomorphism*

$$\psi : H^{ev}(G,k) \to k[V_r(G)]$$

is nilpotent and its image contains all p^r-th powers of $k[V_r(G)]$.

In particular, the associated map of affine schemes

$$\psi : V_r(G) \to |G|$$

is a finite universal homeomorphism.

The proof of Theorem 2.3 splits naturally into two parts. We first prove surjectivity modulo nilpotents as stated in the following theorem.

THEOREM 2.4 [SFB1]. *The homomorphism ϕ of Theorem 1.2 factors as the composition*

$$j^* \circ \overline{\phi} : \bigotimes_{i=1}^{r} S^*(gl_n^{(r)\#}[2p^{i-1}]) \to k[V_r(GL_n)] \to H^*(GL_{n(r)}, k),$$

where j denotes the natural closed embedding $V_r(GL_n) \subset gl_n^{\times r}$. Moreover, the composition

$$\psi \circ \overline{\phi} : k[V_r(GL_n)] \to H^*(GL_{n(r)}, k) \to k[V_r(GL_n)]$$

equals F^r, the r-th iterate of Frobenius.

Theorem 2.4 provides surjectivity modulo nilpotents for any closed subgroup $G \subset GL_{n(r)}$ by the surjectivity of $k[V_r(GL_n)] \to k[V_r(G)]$ and the naturality of ψ.

The proof of Theorem 2.4 entails the study of characteristic classes

$$e_r(j)(G, V) \in Ext_G^{2j}(V^{(r)}, V^{(r)})$$

associated to a rational representation $G \to GL(V)$ and the universal class

$$e_r(j) = \frac{(e_1^{(r-1)})^{j_0}(e_2^{(r-2)})^{j_1} \ldots e_r^{j_{r-1}}}{(j_0!)(j_1!)\cdots(j_{r-1}!)} \in Ext^{2j}_{\mathcal{P}_{\mathbb{F}_p}}(I^{(r)}, I^{(r)}).$$

In particular, we determine $e_r(j)(\mathbb{G}_{a(r)} \otimes A, V_{\underline{\alpha}})$, where $V_{\underline{\alpha}}$ is the free A-module A^n made into a rational $\mathbb{G}_{a(r)} \otimes A$-module via $\underline{\alpha} : \mathbb{G}_{a(r)} \otimes A \to GL_n \otimes A$ (given by an r-tuple $\alpha_0, \ldots, \alpha_{r-1}$ of p-nilpotent, pairwise commuting matrices in $GL_n(A)$). This determination involves a careful study of coproducts to reduce the problem of identifying these characteristic classes to the special case in which $\underline{\alpha}$ consists of a single non-zero p-nilpotent matrix. These coproducts are in turn identified by investigating characteristic classes for the special case $r = 2$.

The assertion of injectivity modulo nilpotents in Theorem 2.3 is a consequence of the following detection theorem. The generality of this statement is useful when considering the algebras $Ext_G^*(M, M)$ as we do in the next section.

THEOREM 2.5 [SFB2]. *Let G be an infinitesimal group scheme of height $\leq r$ and let Λ be an associative, unital, rational G-algebra. Then $z \in H^n(G, \Lambda)$ is nilpotent if and only if for every field extension K/k and every 1-parameter subgroup $\nu : \mathbb{G}_{a(r)} \otimes K \to G \otimes K$, the class $\nu^*(z_K) \in H^n(\mathbb{G}_{a(r)} \otimes K, \Lambda_K)$ is nilpotent.*

For G unipotent, Theorem 2.5 is proved in a manner similar to that employed by D. Quillen to prove his theorem that the cohomology modulo nilpotents of a finite group is detected on elementary abelian subgroups [Q2]. Namely, analogues of the Quillen-Venkov Lemma [Q-V] and J.-P. Serre's cohomological characterization of elementary abelian p-groups [S] are proved. In contrast to the context of finite groups in which a transfer argument permits reduction to nilpotent groups (i.e., to p-Sylow subgroups), we require a new strategy to extend this detection theorem to arbitrary infinitesimal group schemes. We develop a generalization of a spectral sequence of H. Andersen and J. Jantzen [A-J] in order to relate the cohomology of G/B with coefficients in the cohomology of B to the cohomology of G, where $B \subset G$ is a Borel subgroup.

§3. GEOMETRY FOR G-MODULES

As in the case of finite groups, we investigate $Ext^*_G(M,M)$ as a $H^{ev}(G,k)$-module, where G is an infinitesimal group scheme and M a finite dimensional rational G-module. The techniques discussed in the previous section enable us to prove the following analogue of "Carlson's Conjecture", a result proved by G. Avrunin and L. Scott for $k\pi$-modules for a finite group π [A-S]. The proof of Avrunin and Scott involved the study of representations of abelian Lie algebras with trivial restriction. Their result was generalized to an arbitrary finite dimensional restricted Lie algebra by the author and B. Parshall [FP3].

THEOREM 3.1 [SFB2]. *Let G be an infinitesimal group scheme of height $\leq r$ and let M be a finite dimensional G-module. Define the closed subscheme*

$$|G|_M \ \subset \ |G| = Spec H^{ev}(G,k)$$

to be the the reduced closed subscheme defined by the radical of the annihilator ideal of $Ext^*_G(M,M)$. Define the closed subscheme

$$V_r(G)_M \ \subset \ V_r(G)$$

to be the reduced closed subscheme whose K-points for any field extension K/k are those 1-parameter subgroups $\nu : \mathbb{G}_{a(r)} \otimes K \to G \otimes K$ with the property that $\epsilon_{K*} \circ \nu^*(M)$ is a projective $\mathbb{G}_{a(1)}$-module (where ϵ is given in (2.1.1)). Then the finite universal homeomorphism

$$\Phi : V_r(G) \to |G|$$

of Theorem 2.3 satisfies

$$\Phi^{-1}(|G|_M) = V_r(G)_M.$$

Observe that $|G|_M$ is essentially cohomological in nature whereas $V_r(G)_M$ is defined without reference to cohomology. Properties of $V_r(G)_M$ (and thus of $|G|_M$) can often be verified easily.

PROPOSITION 3.2. *(cf.* [SFB2]*) Let G be an infinitesimal group scheme and let $\mathcal{G}(G)$ denote the reduced Green ring of isomorphism classes of finite dimensional rational G-modules modulo projectives. Then sending M to $V_r(G)_M$ determines a function*

$$\Theta : \mathcal{G}(G) \to \{\textit{reduced, closed, conical subschemes of } V_r(G)\}$$

such that

(a.) Θ *is surjective.*
(b.) If $\Theta(M) = pt$, *then* M *is projective.*
(c.) $\Theta(M \oplus N) = \Theta(M) \cup \Theta(N)$.
(d.) $\Theta(M \otimes N) = \Theta(M) \cap \Theta(N)$.

§4. Speculations Concerning Local Type

We briefly mention a possible formulation of the "local type" of finite dimensional rational G-modules M for infinitesimal group schemes G.

PROPOSITION 4.1. *Let G be an infinitesimal group scheme of height r and let M be a rational G-module of dimension m. Then M determines a conjugacy class of p-nilpotent matrices in $GL_m(k[V_r(G)])$ associated to the rational $G_{a(1)} \otimes k[V_r(G)]$-module $\epsilon_{k[V_r(G)]*} \circ u^*(M \otimes k[V_r(G)])$, where $u : \mathbb{G}_{a(r)} \otimes k[V_r(G)] \to G \otimes k[V_r(G)]$ is the universal infinitesimal of height r 1-parameter subgroup of G and $\epsilon : k[\mathbb{G}_{a(r)}] \to k[\mathbb{G}_{a(1)}]$ is given in (2.2.1).*

The following formulation of local type is one possible "numerical invariant" which we can derive from the matrix of Proposition 4.1.

DEFINITION 4.2. *Let G be an infinitesimal group scheme of height r. For each prime ideal x of $k[V_r(G)]$, let $\nu_x : \mathbb{G}_{a(r)} \otimes k(x) \to G \otimes k(x)$ be the 1-parameter subgroup associated to the residue homomorphism $k[V_r(G)] \to k(x)$, where $k(x) = frac\{k[V_r(G)]/x\}$. For a finite dimensional rational G-module M, we define the local type of M to be the lower semi-continuous function*

$$t_M : V_r(G) \to \mathbb{N}^p,$$

where $t_M(x) = (t_{M,p}(x), \dots, t_{M,1}(x))$ with $t_{M,i}(x)$ equal to the number of Jordan blocks of size i of $\epsilon_{k(x)*}(\nu_x^*(M \otimes k(x)))$ as a rational $\mathbb{G}_{a(1)} \otimes k(x)$-module.

Thus, the points of $V_r(G)_M$ are those points $x \in V_r(G)$ such that $t_M(x) \neq (\frac{m}{p}, 0, \dots 0)$.

We conclude with two questions, even partial answers to which would be of considerable interest.

QUESTION 4.3. *Describe in module-thoeretic terms the condition on a pair of rational G-modules M, N that implies $t_M = t_N$.*

QUESTION 4.4. *Characterize those functions $t : V_r(G) \to \mathbb{N}^{p-1}$ for which there exists some finite dimensional rational G-module M with $t = t_M$.*

References

[A-J] H Andersen and J. Jantzen, *Cohomology of induced representations for algebraic groups*, Math. Ann. **269** (1985), 487-525.

[A-S] G. Avrunin and L. Scott, *Quillen stratification for modules*, Inventiones Math. **66** (1982), 277-286.

[C] J. Carlson, *The varieties and the cohomology ring of a module*, J. Algebra **85** (1983), 104-143.

[Ca] P. Cartier, *Une nouvelle opération sur les formes différentielles*, C.-R. Acad. Sci Paris **244** (1957), 426–428.

[CPS] E. Cline, B. Parshall, and L. Scott, *Cohomology, hyperalgebras, and representations*, J. Algebra **63** (1980), 98 – 123.

[CPSK] E. Cline, B. Parshall, L. Scott, and W. van der Kallen, *Rational and generic cohomology*, Inventiones Math. **39** (1977), 143–163.

[D] W. Dwyer, *Twisted homological stability for general linear groups*, Annals of Math. **111** (1980), 239-251.

[E] L. Evens, *The cohomology ring of a finite group*, Trans. A.M.S. **101** (1961), 224–239.

[FFSS] V. Franjou, E. Friedlander, A. Scorichenko, and A. Suslin, *General linear and functor cohomology over finite fields*, Preprint.

[FLS] V. Franjou, J. Lannes, and L. Schwartz, *Autor de la cohomologie de MacLane des corps finis*, Inventiones Math. **115** (1994), 513–538.

[FP1] E. Friedlander and B. Parshall, *On the cohomology of algebraic and related finite groups*, Inventiones Math. **74** (1983), 85–117.

[FP2] E. Friedlander and B. Parshall, *Cohomology of Lie algebras and algebraic groups*, Amer. J. Math. **108** (1986), 235-253.

[FP3] E. Friedlander and B. Parshall, *Support varieties for restricted Lie algebras*, Inventiones Math. **86** (1986), 553-562.

[FP4] E. Friedlander and B. Parshall, *Geometry of p-Unipotent Lie algebras*, J. Algebra **109** (1987), 25-45.

[F-S] E. Friedlander and A. Suslin, *Cohomology of finite group schemes over a field*, Inventiones Math. **127** (1997), 209-270.

[J] J. Jantzen, *Kohomologie von p-Lie Algebren und nilpotente Elemente*, Abh. Math. Sem. Univ. Hamburg **56** (1986), 191-219.

[K] N. Kuhn, *Rational cohomology and cohomological stability in generic representation theory*, Preprint.

[P] T. Pirashvili, *Higher additivisations [Russian, English summary]*, Trudy Tbiliss. Mat. Inst. Razmodze Akad. Nauk Gruzin. SSR **91** (1988), 44-54.

[Q1] D. Quillen, *The spectrum of an equivariant cohomology ring: I, II*, Ann. Math. **94** (1971), 549-572, 573-602.

[Q2] D. Quillen, *On the cohomology and K-theory of the general linear group over a finite field*, Annals of Math. **96** (1972), 552–586.

[Q-V] D. Quillen and B. Venkov, *Cohomology of finite groups and elementary abelian subgroups*, Topology **11** (1972), 317-318.

[S] J.-P. Serre, *Sur la dimension cohomologique des groupes profinis*, Topology **3** (1965), 413-420.

[SFB1] A. Suslin, E. Friedlander, and C. Bendel, *Infinitesimal 1-parameter subgroups and cohomology*, Journal of the A.M.S. **10** (1997), 693-728.

[SFB2] A. Suslin, E. Friedlander, and C. Bendel, *Support varieties for infinitesimal group schemes*, Journal of the A.M.S. **10** (1997), 729 -759.

[V] B. Venkov, *Cohomology algebras for some classifying spaces*, Dokl. Akad. Nauk. SSSR **127** (1959), 943–944.

Eric Mark Friedlander
Department of Mathematics
Northwestern University
Evanston, IL 60208-2730, USA
eric@math.nwu.edu

On the Burnside Problem for Groups of Even Exponent

Sergei V. Ivanov[1]

Abstract. The Burnside problem about periodic groups asks whether any finitely generated group with the law $x^n \equiv 1$ is necessarily finite. This is proven only for $n \leq 4$ and $n = 6$. A negative solution to the Burnside problem for odd $n \gg 1$ was given by Novikov and Adian. The article presents a discussion of a recent solution of the Burnside problem for even exponents $n \gg 1$ and related results.

1991 Mathematics Subject Classification: Primary 20F05, 20F06, 20F10, 20F50

Recall that the notorious Burnside problem about periodic groups (posed in 1902, see [B]) asks whether any finitely generated group that satisfies the law $x^n \equiv 1$ (n is a fixed positive integer called the *exponent* of G) is necessarily finite. A positive solution to this problem is obtained only for $n \leq 4$ and $n = 6$. Note the case $n \leq 2$ is obvious, the case $n = 3$ is due to Burnside [B], $n = 4$ is due to Sanov [S], and $n = 6$ to M. Hall [Hl] (see also [MKS]). A negative solution to the Burnside problem for odd exponents was given in 1968 by Novikov and Adian [NA] (see also [Ad]) who constructed infinite m-generator groups with $m \geq 2$ of any odd exponent $n \geq 4381$ (later Adian [Ad] improved on this estimate bringing it down to odd $n \geq 665$). A simpler geometric solution to this problem for odd $n > 10^{10}$ was later given by Ol'shanskii [Ol1] (see also [Ol2]). We remark that attempts to approach the Burnside problem via finite groups gave rise to a restricted version of the Burnside problem [M] that asks whether there exists a number $f(m, n)$ so that the order of any finite m-generator group of exponent n is less than $f(m, n)$. The existence of such a bound $f(m, n)$ was proven for prime n by Kostrikin [K1] (see also [K2]) and for $n = p^\ell$ with prime p by Zelmanov [Z1]-[Z2]. By a reduction theorem due to Ph. Hall and Higman [HH] it then follows from this Zelmanov result that, modulo the classification of finite simple groups, the function $f(m, n)$ does exist for all m, n.

However, the Burnside problem for even exponents n without odd divisor ≥ 665, being especially interesting for $n = 2^\ell \gg 1$, remained open. The principal difference between odd and even exponents in the Burnside problem can be illustrated by pointing out that, on the one hand, for every odd $n \gg 1$ there are infinite

[1]Supported in part by the Alfred P. Sloan Foundation and the National Science Foundation of the United States

2-generator groups of exponent n all of whose proper subgroups are cyclic [IA] (see also [Ol2]) and, on the other hand, any 2-group the orders of whose finite (or abelian) subgroups are bounded is itself finite [Hd].

A negative solution to the Burnside problem for even exponents $n \gg 1$ is given in recent author's article [Iv] and based on the following inductive construction (which is analogous to Ol'shanskii's construction [Ol1] for odd $n > 10^{10}$).

Let F_m be a free group of rank m over an alphabet $\mathcal{A} = \{a_1^{\pm 1}, \dots, a_m^{\pm 1}\}$, $m > 1$, $n \geq 2^{48}$ and n be divisible by 2^9 provided n is even (from now on we impose these restrictions on m and n unless otherwise stated; note this estimate $n \geq 2^{48}$ has been improved on by Lysënok [L] to $n \geq 2^{13}$). By induction on i, let $B(m, n, 0) = F_m$ and, assuming that the group $B(m, n, i-1)$ with $i \geq 1$ is already constructed as a quotient group of F_m, define A_i to be a shortest element of F_m (if any) the order of whose image (under the natural epimorphism $\psi_{i-1} : F_m \to B(m, n, i-1)$) is infinite. Then $B(m, n, i)$ is constructed as a quotient group of $B(m, n, i-1)$ by the normal closure of $\psi_{i-1}(A_i^n)$. Clearly, $B(m, n, i)$ has a presentation of the form

$$B(m, n, i) = \langle\, a_1, \dots, a_m \parallel A_1^n, \dots, A_{i-1}^n, A_i^n \,\rangle, \tag{1}$$

where $A_1^n, \dots, A_{i-1}^n, A_i^n$ are the defining relators of $B(m, n, i)$.

The quotient group F_m/F_m^n, where F_m^n is the subgroup of the free group F_m generated by all nth powers, is denoted by $B(m, n)$ and called the free m-generator Burnside group of exponent n. Now we give a summary of basic results of [Iv].

THEOREM 1 ([IV]). *Let $m > 1$, $n \geq 2^{48}$, and 2^9 divide n provided n is even. Then the following hold.*

(a) *The free m-generator Burnside group $B(m, n)$ of exponent n is infinite.*
(b) *The word A_i does exist for each $i \geq 1$.*
(c) *The direct limit $B(m, n, \infty)$ of the groups $B(m, n, i)$, $i = 1, 2, \dots$, is the free m-generator Burnside group $B(m, n)$ of exponent n, that is, $B(m, n)$ has the presentation*

$$B(m, n) = \langle\, a_1, \dots, a_m \parallel A_1^n, \dots, A_i^n, A_{i+1}^n, \dots \,\rangle. \tag{2}$$

(d) *There are algorithms that solve the word and conjugacy problems for the group $B(m, n)$ given by presentation* (2).
(e) *Let $n = n_1 n_2$, where n_1 is odd and n_2 is a power of 2. If n is odd, then every finite subgroup of $B(m, n)$ is cyclic. If n is even, then every finite subgroup of $B(m, n)$ is isomorphic to a subgroup of the direct product $D(2n_1) \times D(2n_2)^\ell$ for some ℓ, where $D(2k)$ is a dihedral group of order $2k$.*
(f) *For every $i \geq 0$ the group $B(m, n, i)$ given by presentation* (1) *is hyperbolic (in the sense of Gromov [G]).*

Note that the part (a) of Theorem 1 is immediate from part (b) because if $B(m, n)$ were finite it could be given by finitely many defining relators and so A_i would fail to exist for sufficiently large i. To prove part (d) the word and conjugacy problems for the group $B(m, n)$ are effectively reduced to the word problem for some $B(m, n, i)$ and it is shown that every $B(m, n, i)$ satisfies a linear isoperimetric

inequality and so is hyperbolic. It should be pointed out that for odd exponents $n \gg 1$ all parts of Theorem 1 had been known due to Novikov and Adian (parts (a), (e), (f); see [Ad]) and Ol'shanskii (parts (b), (c), (f); see [Ol1], [Ol2]).

It is worth noting that the structure of finite subgroups of the free Burnside group $B(m,n)$ and $B(m,n,i)$ is very complex when the exponent n is even and, in fact, finite subgroups of groups $B(m,n,i)$, $B(m,n)$ turn out to be so important in proofs of [Iv] that at least a third of article [Iv] is an investigation of their various properties and another third is a preparation of necessary techniques to conduct this investigation. Part (e) of Theorem 1 may be regarded as a central result on finite subgroups of $B(m,n)$. To state more results on finite subgroups of groups $B(m,n,i)$, $B(m,n)$, denote by $\mathcal{F}(A_i)$ a maximal finite subgroup of $B(m,n,i-1)$ relative to the property that A_i (that is, the image of A_i in $B(m,n,i-1)$) normalizes $\mathcal{F}(A_i)$. A word U is called an $\mathcal{F}(A_i)$-*involution* if $U^2 \in \mathcal{F}(A_i)$, U normalizes the subgroup $\mathcal{F}(A_i)$ of $B(m,n,i-1)$, and

$$UA_iU^{-1} = A_i^{-1}F$$

in $B(m,n,i-1)$, where $F \in \mathcal{F}(A_i)$.

For example, if A_i is a letter then $\mathcal{F}(A_i) = \{1\}$ and there are no $\mathcal{F}(A_i)$-involutions. If $n \gg 1$ is odd then for every i one has $\mathcal{F}(A_i) = \{1\}$ and there are no $\mathcal{F}(A_i)$-involutions. If now $A_i = a_1^{n/2}a_2^{n/2}$ then $\mathcal{F}(A_i) = \{1\}$ and $a_1^{n/2}$, $a_2^{n/2}$ are $\mathcal{F}(A_i)$-involutions. An example when $\mathcal{F}(A_i) \neq \{1\}$ is provided by

$$A_i = (a_1^{n/2}(a_1a_2)^{n/2})^{n/2}(a_2^{n/2}(a_1a_2)^{n/2})^{n/2}$$

for $(a_1a_2)^{n/2} \in \mathcal{F}(A_i)$.

Next, define

$$\mathcal{G}(A_i) = \langle\, U_i, A_i, \mathcal{F}(A_i) \,\rangle$$

to be a subgroup in $B(m,n,i-1)$ generated by A_i, by the subgroup $\mathcal{F}(A_i)$, and by a word U_i, where U_i is an $\mathcal{F}(A_i)$-involution provided there are $\mathcal{F}(A_i)$-involutions and $U_i = 1$ otherwise. It follows from definitions that $\mathcal{G}(A_i)$ is either an extension of $\mathcal{F}(A_i)$ by an infinite dihedral group generated by elements U_i, A_i of order $2, \infty$, respectively, modulo $\mathcal{F}(A_i)$ provided there are $\mathcal{F}(A_i)$-involutions or $\mathcal{G}(A_i)$ is an extension of $\mathcal{F}(A_i)$ by an infinite cyclic group generated by A_i provided there are no $\mathcal{F}(A_i)$-involutions. Basic properties of finite subgroups of groups $B(m,n,i)$, $B(m,n)$ are collected in the following.

Theorem 2 ([Iv]). *Let $B(m,n)$ be a free m-generator Burnside group of even exponent $n \geq 2^{48}$ given by presentation* (2)*, where n is divisible by 2^9. Then the following are true.*

(a) *The subgroup $\mathcal{F}(A_i)$ is defined uniquely and is a 2-group.*

(b) *The word $A_i^{n/2}$ centralizes in $B(m,n,i-1)$ the subgroup $\mathcal{F}(A_i)$ and hence the quotient $\mathcal{G}(A_i)/\langle A_i^n\rangle$, denoted by $\mathcal{K}(A_i)$ is either an extension of $\mathcal{F}(A_i)$ by a dihedral group of order $2n$ generated by elements U_i, A_i, or $\mathcal{K}(A_i)$ is an extension of $\mathcal{F}(A_i)$ by a cyclic group of order n generated by A_i. In addition, the group $\mathcal{K}(A_i)$ naturally embeds in $B(m,n,i)$ and $B(m,n)$.*

(c) *Every word W of finite order in $B(m,n,i-1)$ is conjugate in $B(m,n,i-1)$ to a word of the form $A_j^k T$ with some integers k, $j<i$ and $T\in\mathcal{F}(A_i)$. Moreover, the conjugacy in $B(m,n,i-1)$ of nontrivial in $B(m,n,i-1)$ words $A_{j_1}^{k_1}T_1$ and $A_{j_2}^{k_2}T_2$, where $T_1\in\mathcal{F}(A_{j_1})$ and $T_2\in\mathcal{F}(A_{j_2})$, $j_1,j_2<i$, yields $j_1=j_2$ and $k_1\equiv\pm k_2 \pmod n$. (Therefore, given a nontrivial in $B(m,n,i-1)$ word W of finite order such number j is defined uniquely in $B(m,n,i-1)$ as well as in $B(m,n)$ and called the height of the word W.)*

(d) *Every finite subgroup of $B(m,n)$, consisting of words of heights $\le i$ and containing a word of height i, is conjugate to a subgroup of $\mathcal{K}(A_i)=\langle\, U_i, A_i, \mathcal{F}(A_i)\,\rangle\subseteq B(m,n)$.*

(e) *The words A_i and U_i act on the subgroup $\mathcal{F}(A_i)$ of $B(m,n,i-1)$ by conjugations in the same way as some words V_{A_i} and V_{U_i} act respectively, where V_{A_i} and V_{U_i} are such that the subgroup $\langle\, V_{A_i}, V_{U_i}, \mathcal{F}(A_i)\,\rangle$ of $B(m,n,i-1)$ is finite and the equality $U_i^2=V_{U_i}^2$ (as well as $(U_iA_i)^2=(V_{U_i}V_{A_i})^2$ provided $U_i\ne 1$) holds in $B(m,n,i-1)$.*

Let us see how the algebraic description of finite subgroups of $B(m,n)$ of Theorem 1 (e) can be derived from Theorem 2 by induction on the height $h(G)$ of a finite subgroup G of $B(m,n)$ ($h(G)$ is the maximum of heights of elements of G). Let G be a finite subgroup of $B(m,n)$ with $h(G)=i$. By Theorem 2 (d), one may assume that G is a subgroup of $\mathcal{K}(A_i)$. It follows from definitions and Theorem 2 (e) that there are homomorphisms

$$\kappa_1:\mathcal{K}(A_i)\to D(2n),\quad \kappa_2:\mathcal{K}(A_i)\to G_0$$

such that

$$\operatorname{Ker}\kappa_1=\mathcal{F}(A_i),\quad G_0=\langle\,\mathcal{F}(A_i),V_{A_i},V_{U_i}\,\rangle\subseteq B(m,n,i-1)$$

and $h(G_0)<i$. Since $\operatorname{Ker}\kappa_1\cap\operatorname{Ker}\kappa_2=\{1\}$, the group G embeds in the direct product $D(2n)\times G_0$. By the induction hypothesis, G_0 embeds in $D(2n_1)\times D(2n_2)^\ell$ for some ℓ. By Theorem 2 (a), $\mathcal{F}(A_i)$ is a 2-group, therefore the subgroup of $D(2n_1)$ of index 2 has the trivial intersection with the image of $\mathcal{F}(A_i)$ in $D(2n)\times D(2n_1)\times D(2n_2)^\ell$ and hence $D(2n_1)$ can be replaced by $D(2n_2)$. Since $D(2n)$ embeds in $D(2n_1)\times D(2n_2)$, we have that G is embeddable in $D(2n_1)\times D(2n_2)^{\ell+1}$ as required. This algebraic description of finite subgroups is very important in many parts of article [Iv], especially, in making the inductive step from the group $B(m,n,i-1)$ to $B(m,n,i)$. For example, this description helps a great deal in proving one of the hardest and absolutely crucial for the whole work technical results: If the subgroup

$$\langle\, A_i^kTA_i^{-k}\mid k=0,1,\dots,7\,\rangle$$

of $B(m,n,i-1)$ is finite then A_i normalizes this subgroup. We will make an informal remark that if a similar claim (where instead of 7 one could put a number as large as $n^{1/2}$) were false then A_i^n would not have to centralize $\mathcal{F}(A_i)$ and imposing the relation $A_i^n=1$ on $B(m,n,i-1)$ would result in extra relations of type $R=1$ where R is a nontrivial element of $\mathcal{F}(A_i)$ (R has the form $R=$

$A_i^n F A_i^{-n} F^{-1} \neq 1$ with $F \in \mathcal{F}(A_i)$). This secondary factorization would make a complete mess implying that one could be far better off trying to solve the Burnside problem for $n = 2^\ell$ in the affirmative.

The proofs in [Iv] are based on geometric techniques of van Kampen diagrams (which are labelled planar 2-complexes representing consequences of defining relators of group presentations; see [Ol2], [LS], [IO1]) and may be regarded as a further development of Ol'shanskii's method [Ol1] for solving the Burnside problem for odd exponents $n \gg 1$. The main obstacle in carrying over Ol'shanskii's proof to even exponents is in making the inductive step from $B(m, n, i-1)$ to $B(m, n, i)$. Curiously, in odd case the inductive step from $B(m, n, i-1)$ to $B(m, n, i)$ is being made in [Ol1] by boiling everything down to an elementary fact that the fundamental group of an annulus is cyclic (however, the reduction itself is highly nontrivial). The same fact is ultimately responsible for the cyclicity of finite subgroups of $B(m, n)$ with odd $n \gg 1$. This reduction naturally fails in even case due to the existence of self-compatible cells in nonsimply connected diagrams over $B(m, n, i-1)$. (A self-compatible cell is a 2-cells that surrounds hole(s) of a diagram and has two long arcs of its boundary with a narrow strip squeezed between the arcs.) Informally, turning these self-compatible cells from the main obstacle into a source of new information is what the article [Iv] is all about. However, extracting gems from this mine is quite a challenge and that partially explains an extraordinary length (over 300 pages) of the article and its complex logical structure (over 110 lemmas are proved by simultaneous induction on the parameter i with quite a few back references; note a similar simultaneous induction is carried out in [NA], [Ol1]).

It is implied by results of [Iv] that finite (as well as locally finite) subgroups of $B(m, n)$ with even n are very interesting subject for investigation on their own. In particular, one might wonder if their description given in Theorem 1 (e) is complete, that is, every group $D(2n_1) \times D(2n_2)^\ell$ embeds in $B(m, n)$. Another natural question is to ask whether every locally finite subgroup of $B(m, n)$ is an FC-group. Recall that a group G is *locally finite* if every finitely generated subgroup of G is finite. A group G is termed an *FC-group* if every conjugacy class of G is finite. Also, let D_i, $i = 1, 2, \ldots$, be groups isomorphic to $D(2n_2)$, $\mathcal{D}$ be the cartesian product of D_i, $i = 1, 2, \ldots$, C_i be the normal cyclic subgroup of D_i of order n_2, and $b_i \in D_i$ be an element of order 2 that together with C_i generate $D_i = \langle C_i, b_i \rangle$. By $\mathcal{B}$ denote the subgroup of $\mathcal{D}$ that consists of all elements whose projection on every D_i is either b_i or 1. By $\mathcal{C}$ denote the direct product of groups C_i naturally embedded in $\mathcal{D}$. At last, let $\mathcal{E} = \langle \mathcal{B}, \mathcal{C} \rangle$. Clearly, $\mathcal{E} = \mathcal{B}\mathcal{C}$ is a semidirect product of $\mathcal{B}$ and $\mathcal{C}$.

The following Theorem 3 is a summary of joint with Ol'shanskii results on (locally) finite subgroups of $B(m, n)$.

Theorem 3 ([IO2]). *Let $B(m, n)$ be a free m-generator Burnside group of even exponent n, where $m > 1$ and $n \geq 2^{48}$, $n = n_1 n_2$, n_1 is odd, n_2 is a power of 2, $n_2 \geq 2^9$. Then the following hold:*

(a) *Suppose $\mathcal{G}$ is a finite 2-subgroup of $B(m, n)$. Then the centralizer $C_{B(m,n)}(\mathcal{G})$ of $\mathcal{G}$ in $B(m, n)$ contains a subgroup $\mathcal{M}$ isomorphic to a free*

Burnside group $B(\infty, n)$ of infinite countable rank such that $\mathcal{G} \cap \mathcal{M} = \{1\}$. In particular, $\langle \mathcal{G}, \mathcal{M} \rangle = \mathcal{G} \times \mathcal{M}$.

(b) *The centralizer $C_{B(m,n)}(\mathcal{H})$ of a subgroup $\mathcal{H}$ of $B(m, n)$ is infinite if and only if $\mathcal{H}$ is a locally finite 2-subgroup. In particular, $C_{B(m,n)}(\mathcal{H})$ is finite provided $\mathcal{H}$ is not locally finite.*

(c) *An arbitrary infinite group G embeds in $B(m, n)$ as a locally finite subgroup if and only if G is isomorphic to a countable subgroup of $\mathcal{E}$.*

(d) *An arbitrary infinite group G embeds in $B(m, n)$ as a maximal locally finite subgroup if and only if G is isomorphic to a countable subgroup of $\mathcal{E}$.*

(e) *An infinite locally finite subgroup $\mathcal{L}$ of $B(m, n)$ is contained in a unique maximal locally finite subgroup. That is, the intersection of two distinct maximal locally finite subgroups of $B(m, n)$ is always finite.*

(f) *Given a finite 2-subgroup $\mathcal{G}$ of $B(m, n)$ there are continuously many pairwise nonisomorphic maximal locally finite subgroups that contain $\mathcal{G}$.*

(g) *If a finite subgroup $\mathcal{G}$ of $B(m, n)$ contains a nontrivial element of odd order, then $\mathcal{G}$ is contained in a unique maximal finite subgroup. In particular, the intersection of two distinct maximal finite subgroups of $B(m, n)$ is always a 2-group.*

Note the mutual disposition of infinite maximal locally finite subgroups stated in Theorem 3 (e), (f) is reminiscent of a known puzzle-type problem: Find, in a countably infinite set, continuously many subsets whose pairwise intersections are all finite (note this is impossible if the cardinalities of the intersections are bounded).

A couple of questions mentioned above can now be easily answered:

COROLLARY. *Let $B(m, n)$ be defined as in Theorem 2. Then the following are true.*

(a) *A finite group G embeds in $B(m, n)$ if and only if G is isomorphic to a subgroup of the direct product $D(2n_1) \times D(2n_2)^\ell$ for some $\ell > 0$, where $D(2k)$ is a dihedral group of order $2k$.*

(b) *The group $B(m, n)$ contains (maximal) locally finite subgroups that are not FC-groups.*

(c) *A subgroup $\mathcal{S}$ of $B(m, n)$ is locally finite if and only if every 2-generator subgroup of $\mathcal{S}$ is finite.*

The machinery developed in [Iv] for solving the Burnside problem for even exponents $n \gg 1$ has made it possible to prove a conjecture of Gromov on quotients of hyperbolic groups of bounded exponent.

To state the results we recall several definitions. Let G be a finitely generated group, $\mathcal{A}$ be a finite set of generators for G. By $|g| = |W|$ denote the length of a shortest word W in the alphabet $\mathcal{A}$ that represents an element $g \in G$. One of definitions of a hyperbolic group G is given by means of the Gromov product

$$(g \cdot h) = \tfrac{1}{2}(|g| + |h| - |g^{-1}h|)$$

as follows: A group G is called *hyperbolic* [G] if there exists a constant $\delta \geq 0$ such that for every triple $g, h, f \in G$

$$(g \cdot h) \geq \min((g \cdot f), (h \cdot f)) - \delta. \tag{3}$$

In turns out [G], [GH] that the property of being hyperbolic does not depend on a particular generating set $\mathcal{A}$ (but the constant δ does depend on $\mathcal{A}$). A trivial example of hyperbolic group is the free group $F = F(\mathcal{A})$ over $\mathcal{A}$ for which inequality (3) is satisfied with $\delta = 0$ because in this case $(g \cdot h)$ is the length of the maximal common beginning of reduced words g, h in $\mathcal{A}$. Perhaps, the most complicated (in terms of proving that) examples of hyperbolic groups are provided by the series of groups $B(m, n, i)$ of Theorems 1–2.

Similar to free groups, an arbitrary nonelementary hyperbolic group has many homomorphic images (recall a group G is termed in [G] *elementary* if G has a cyclic subgroup of finite index). Discussing an approach to construction of an infinite periodic quotient group $\bar{G}$ of a nonelementary hyperbolic group G, Gromov [G] (see also [GH], [Ol4]) points out that his approach does not let bound the orders of elements in $\bar{G}$ (and so $\bar{G}$ will not be of finite exponent). Nevertheless, Gromov conjectures (see 5.5E, 5.5F in [G]) that it is possible in principle to bound the orders of elements in $\bar{G}$ and obtain $\bar{G}$ of finite exponent n, that is, he conjectures that for every nonelementary hyperbolic group G there is an $n = n(G)$ such that the quotient G/G^n is infinite. Thus Gromov suggested a natural expansion of the Burnside problem to nonelementary hyperbolic groups. This Gromov conjecture was proven by Ol'shanskii [Ol3] for torsion free hyperbolic groups. However, in the general case of a hyperbolic group with torsion there are serious obstacles connected with nonelementary centralizers of elements of G and noncyclic finite subgroups in G/G^n that are essentially the same as those in solving the classical Burnside problem for even exponents n.

The solution [Iv] of the "even" Burnside problem discussed above combined with ideas of [Ol3] enabled Ol'shanskii and the author to prove the Gromov conjecture in full generality.

Theorem 4 ([IO3]). *For every nonelementary hyperbolic group G there exists a positive even integer $n = n(G)$ such that the following are true:*

(a) *The quotient group G/G^n is infinite.*
(b) *The word and conjugacy problems are solvable in G/G^n.*
(c) *Let $n = n_1 n_2$, where n_1 is odd and n_2 is a power of 2. Then every finite subgroup of G/G^n is isomorphic to an extension of a finite subgroup K of G by a subgroup of the direct product $D(2n_1) \times D(2n_2)^\ell$ for some ℓ, where $D(2k)$ is a dihedral group of order $2k$.*
(d) *The subgroup G^n is torsion free and $\bigcap_{k=1}^{\infty} G^{kn} = \{1\}$.*

When proving Theorem 4, we encounter several restrictions to be imposed on n and end up with that n must be divisible by $2^{k_0+5} n_0$ (to say nothing of that $n \gg \delta = \delta(G)$), where $\frac{n_0}{2}$ is the least common multiple of the exponents of the holomorphs $\mathrm{Hol}(K)$ over all finite subgroups K of G and k_0 is the minimal integer with $2^{k_0-3} > \max |K|$ over all finite subgroups K of G. We note that almost all lemmas of [Iv] are reproved in [IO3] with necessary modifications which

are analogous to those made by Ol'shanskii [Ol3] to adjust his solution [Ol1] of "odd" Burnside problem for proving Gromov conjecture for torsion free hyperbolic groups. We also note that, just like in [Iv], information on finite subgroups of G/G^n is very important in proofs of [IO3] and their description (Theorem 4 (c)) is given as in [Iv] (Theorem 1 (e)) modulo finite subgroups of G. Naturally, proofs in [IO3] also make use of various facts of general theory of hyperbolic groups, see [G], [GH], [CDP].

References

[Ad] S.I. Adian, *The Burnside problem and identities in groups*, Nauka, Moscow, 1975; English translation: Springer-Verlag, 1979.

[AI] V.S. Atabekian and S.V. Ivanov, Two remarks on groups of bounded exponent, # 2243-B87, VINITI, Moscow, 1987 (this is kept in the Depot of VINITI, Moscow, and is available upon request), 23 pp.

[B] W. Burnside, On unsettled question in the theory of discontinuous groups, *Quart. J. Pure and Appl. Math.* **33**(1902), 230-238.

[CDP] E. Coornaert, T. Delzant, and A. Papadopoulos (ed's), *Géométric et théorie des groupes: Les groupes, huperboliques de Gromov*, Lecture Notes in Math. **1441** (1991), Springer-Verlag.

[GH] E. Ghys and P. de la Harpe (ed's), *Sur les groupes huperboliques d'aprés Mikhael Gromov*, Birkhaüser, 1990.

[G] M. Gromov, Hyperbolic groups, in *Essays in Group Theory*, ed. S.M. Gersten, M.S.R.I. Pub. 8, Springer, 1987, 75–263.

[Hl] M. Hall, Jr., Solution of the Burnside problem for exponent 6, *Proceedings Nat. Acad. Sci. USA* **43**(1957), 751–753.

[Hd] D. Held, On abelian subgroups of an infinite 2-group, *Acta Sci. Math. (Szeged)* **27**(1966), 97–98.

[HH] Ph. Hall and G. Higman, On the p-length of p-soluble groups and reduction theorems for Burnside's problem, *Proc. London Math. Soc.* **6**(1956), 1-42.

[Iv] S.V. Ivanov, The free Burnside groups of sufficiently large exponents, *Intern. J. Algebra Comp.* **4** (1994), 1–308.

[IO1] S.V. Ivanov and A.Yu. Ol'shanskii, Some applications of graded diagrams in combinatorial group theory, *London Math. Soc. Lecture Notes Ser.*, vol.160(1991), Cambridge Univ. Press, Cambridge and New York, 1991, 258–308

[IO2] S.V. Ivanov and A.Yu. Ol'shanskii, On finite and locally finite subgroups of free Burnside groups of large even exponents, *J. Algebra* **195** (1997), 241-284.

[IO3] S.V. Ivanov and A.Yu. Ol'shanskii, Hyperbolic groups and their quotients of bounded exponents, *Trans. Amer. Math. Soc.* **348** (1996), 2091–2138.

[K1] A.I. Kostrikin, On the Burnside problem, *Math. USSR Izv.* **23** (1959), 3-34.

[K2] A.I. Kostrikin, *Around Burnside*, Nauka, Moscow, 1986.

[L] I.G. Lysënok, Infinite Burnside groups of even period, *Math. Ross. Izv.* **60** (1996), 3-224.

[LS] R.C. Lyndon and P.E. Schupp, *Combinatorial group theory*, Springer-Verlag, 1977.

[M] W. Magnus, A connection between the Baker-Hausdorff formula and a problem of Burnside, *Ann. Math.* **52** (1950), 11-26; **57** (1953), 606.

[MKS] W. Magnus, J. Karras, and D. Solitar, *Combinatorial group theory*, Interscience Pub., John Wiley and Sons, 1966.

[NA] P.S. Novikov and S.I. Adian, On infinite periodic groups, I, II, III, *Math. USSR Izv.* **32** (1968), 212–244, 251–524, 709–731.

[Ol1] A.Yu. Ol'shanskii, On the Novikov-Adian theorem, *Math. USSR Sbornik* **118** (1982), 203–235.

[Ol2] A.Yu. Ol'shanskii, *Geometry of defining relations in groups*, Nauka, Moscow, 1989; English translation in *Math. and Its Applications* (Soviet series), **70** (Kluwer Acad. Publishers, 1991).

[Ol3] A.Yu. Ol'shanskii, Periodic quotient groups of hyperbolic groups, *Math. USSR Sbornik* **72**(1992), 519–541.

[Ol4] A.Yu. Ol'shanskii, On residualizing homomorphisms and G-subgroups of hyperbolic groups, *Intern. J. Algebra Comp.* **3**(1993), 365–409.

[S] I. N. Sanov, Solution of the Burnside problem for exponent 4, *Uch. Zapiski Leningrad State Univ., Ser. Matem.* **10** (1940), 166–170.

[Zl1] E.I. Zelmanov, Solution of the restricted Burnside problem for groups of odd exponent, *Math. USSR Izv.* **36**(1991), 41-60.

[Zl2] E.I. Zelmanov, A solution of the restricted Burnside problem for 2-groups, *Math. USSR Sbornik* **72**(1992), 543-565.

Sergei Ivanov
Department of Mathematics
University of Illinois at Urbana-Champaign
1409 West Green Street,
Urbana, IL 61801, U.S.A.

Simple Groups in Computational Group Theory

William M. Kantor

Abstract. This note describes recent research using structural properties of finite groups to devise efficient algorithms for group computation.

1991 Mathematics Subject Classification: 20B40, 20D05, 68Q40
Keywords and Phrases: group theoretic algorithms, simple groups

1. Introduction

Numerous applications already have been made of the monumental classification of the finite simple groups (CFSG) within algebra, combinatorics, model theory and computer science. See [GK] for a recent survey. Here I will only consider applications to computational group theory, in the intersection of algebra and computer science, emphasizing results whose statements make it far from clear how any simple group information could be applied.

Two computer systems are widely available for computational group theory: GAP (developed in Aachen by Neubüser and Schönert [Sch], but now moved to St. Andrews); and Magma (developed in Sydney by Cannon [CP]). These systems have been used for important applications within group theory and other parts of mathematics (cf. [Ser1]), occasionally with help from CFSG. One of the most important relationships between computer computations and simple groups was the construction and study of many sporadic (and other specific) simple groups. However, this brief survey focuses on the mathematics behind the algorithms: while practicality is certainly a very important aspect, additional important ones are the discovery of new ways to view standard group-theoretic results, the need for new results about groups, and complexity questions within computer science.

Introductory example The following purely mathematical result can be explained to undergraduates:

Theorem 1 [IKS]. *If p is a prime divisor of the order of a subgroup G of S_n, then the probability that a random element of G has order divisible by p is at least $1/n$. This bound is tight if and only if n is a power of p.*

A similar result is in [Ga]. There is a very practical motivation: assuming that there is a mechanism for finding random elements of G (cf. §§2,5), the theorem states that it "only" takes $O(n)$ samples in order very likely to obtain an element of order divisible by p, and hence one of order p. The proof reduces to the case of a simple group G, in which case much more precise information is obtained about the proportion of elements of G of order divisible by any given prime.

It should be emphasized that the standard proofs of Cauchy's Theorem, or more generally of Sylow's Theorem, are not likely to be used in actual computations

to obtain elements or subgroups of complicated groups. Therefore, other ideas are needed for computations.

2. The permutation group setting; polynomial time

The basic computational situation discussed here is as follows: a group is given, specified as $G = \langle S \rangle$ in terms of some (arbitrary) generating set S of its elements[1]. The goal is then to find properties of G efficiently, such as $|G|$, the derived series, a composition series, Sylow subgroups, and so on. In this section S will be a set of permutations of an n-element set, and then the word "efficiently" might mean "in time polynomial (or nearly linear) in the input length $|S|n$ of the problem".

Many permutation group algorithms are described in detail in a new book by Seress [Ser2]. Numerous other aspects of computational group theory are surveyed in [Si3,Ser1], which also discuss different ways groups are commonly input into computers, e.g., via presentations. See [Ba1] for additional background, especially regarding complexity questions within various models of computation.

The development of efficient computer algorithms for permutation groups was begun by Sims [Si1,Si2] (who then used his ideas for existence proofs for sporadic simple groups). Of fundamental importance was his use of a *base* $B = \{\alpha_1, \dots, \alpha_b\}$ for G: any set of points whose pointwise stabilizer is 1 (possibly B consists of $n-1$ points). Let $G_{(i)}$ be the pointwise stabilizer[2] of $\alpha_1, \dots, \alpha_i$, so that $G = G_{(0)} \geq G_{(1)} \geq \cdots \geq G_{(b)} = 1$ and $|G| = \Pi_1^b |G_{(i-1)} : G_{(i)}|$; here $|G_{(i-1)} : G_{(i)}|$ is the length of the orbit $\mathcal{O}_i$ of α_i under $G_{(i-1)}$. Sims developed a data structure to find a base and (generators for) all of these subgroups $G_{(i)}$ and orbits $\mathcal{O}_i$ simultaneously and efficiently. This yielded $|G|$ using only elementary group theory: it did not involve structural properties of groups. The first version of Sims's order algorithm analyzed in polynomial time is in [FHL];[3] $O(|S|n^2 + n^5)$ versions are in [Kn,Je], and these methods cannot decrease the exponent 5 [Kn].

Once $|G|$ can be found, many other properties of G, such as the derived series, solvability and nilpotence, can be determined in polynomial time. More important from an algorithmic point of view was a MEMBERSHIP TEST: *given $h \in S_n$, decide whether or not $h \in G$; and if it is, obtain h from the generating set S.* The first of these is easy: one could test whether or not $|G| = |\langle S \cup \{h\} \rangle|$; the second depends on the data structure in the order algorithm. A random element of G is now easily obtained as $t_b t_{b-1} \cdots t_1$ where, for each i, t_i is a random element of a transversal for $G_{(i)}$ in $G_{(i-1)}$. The above ideas were implemented in GAP and MAGMA.

OBSTACLES TO POLYNOMIAL-TIME COMPUTATION In polynomial time it is not possible to list the elements of any given permutation group. There are more serious obstacles to the algorithmic study of permutation groups. Consider the following problems for a given $G = \langle S \rangle \leq S_n$.

CENTRALIZER: Given $t \in G$ of order 2, find its centralizer $C_G(t)$.

INTERSECTION: Given subgroups $H, K \leq G$, find their intersection $H \cap K$.

---

[1] It is standard to have groups specified by generating sets. A familiar example is the group of Rubik's cube.

[2] Such subgroups are typically involved in "solving" Rubik's cube.

[3] Sims has informed me that his original version also was a polynomial-time algorithm.

Here $C_G(t)$, H, K and $H \cap K$ are specified by generating sets. Luks observed that Centralizer and Intersection are polynomial-time equivalent and, what is more surprising, that the following problem reduces to these in polynomial time.

Graph Isomorphism: Given two n-vertex graphs, are they isomorphic?

There are practical algorithms for Graph Isomorphism; the main one is due to B. McKay and is contained in both GAP and Magma. However, it is a long-standing open question whether there is a polynomial-time algorithm for Graph Isomorphism. This would be settled by a polynomial-time algorithm for Centralizer, but it seems unlikely that one exists. Thus, this leaves the awkward problem that centralizers, normalizers and intersections of arbitrary subgroups cannot be used in any of the algorithms considered here. This is discussed at length in [Lu3].

A remarkable result of Luks [Lu1] combined "elementary" group theory with group-theoretic algorithms (e.g., for intersecting a *solvable* group with any subgroup of S_n) to obtain a polynomial-time Graph Isomorphism algorithm assuming that the valences of the vertices are bounded.

3. Polynomial-time computation using CFSG

Structural properties of finite groups can help lead to algorithms and then be involved in proving their validity and timing. This is where CFSG enters. The breakthrough in the complexity of permutation group algorithms was Luks's use of CFSG to determine a composition series:

Theorem 2 [Lu2,Be]. *There is a polynomial-time algorithm that determines a composition series of any given $G = \langle S \rangle \leq S_n$.*

The proof used a familiar consequence of CFSG, the validity of "Schreier's conjecture": the outer automorphism group of every finite simple group is solvable. A dozen years after it was first obtained, as part of his thesis Beals observed that a slight modification of Luks's algorithm eliminates any need for CFSG! This is the premier example of a consequence of CFSG in which CFSG was eventually removed. It seems as if there should be more instances of other consequences of CFSG (in various areas) which, once known to be true, can then be proved in better or simpler ways. An example begging for such a new proof is Theorem 1.

Standard methods for finding Sylow subgroups use exponential time in the worst case. However:

Theorem 3 [Ka1]. *There are polynomial-time algorithms for the following problems. Given $G = \langle S \rangle \leq S_n$ and a prime p dividing $|G|$,*

(i) *Find a Sylow p-subgroup of G containing any given p-subgroup of G;*
(ii) *Given Sylow p-subgroups P_1 and P_2 of G, find $g \in G$ with $P_1^g = P_2$; and*
(iii) *Find the normalizer of a Sylow p-subgroup.*

The original arguments in [Ka1] were streamlined in [Ka2,KLM]. The basic idea is to reduce finding Sylow subgroups to finding them in simple groups and to conjugating them in arbitrary groups; then to reduce conjugating them to the simple group case; and finally to solve these types of problems for simple groups on a case-by-case basis. The algorithms in the theorem are impractical, but versions will go into GAP based on [Mo].

QUOTIENT GROUPS If $N \trianglelefteq G \leq S_n$, one can ask for properties and subgroups of G/N. For some questions (a composition series, Sylow subgroups) the transitions from algorithms for G to ones for G/N are elementary. However, it is not necessarily possible for G/N to be represented as a permutation group of degree $< 2^{n/4}$, as P. Neumann has observed when G is merely a direct product of dihedral groups. Hence, there may not be a permutation representation of a quotient group G/N to which previous results could be applied, so that some algorithms have to be developed from scratch for G/N. One of the most unexpected examples of this is the center. It is not difficult to find $Z(G)$ efficiently, but no elementary method is known for $Z(G/N)$. In [KL] this was computed in polynomial time, using the preceding algorithmic version of Sylow's Theorem via the following result: *There is a polynomial-time algorithm for computing* $\mathrm{Core}_G(H)$*, given* $H \leq G \leq S_n$.

Here, the *core* $\mathrm{Core}_G(H) = \cap\{H^g \mid g \in G\}$ is the largest normal subgroup of G contained in H. As noted earlier, it is not known how to intersect two subgroups of an arbitrary group in polynomial time, and this is probably impossible, presenting an apparent obstacle to computing $C = \mathrm{Core}_G(H)$. Nevertheless, C can be found as follows: for each prime $p \big| |G|$ find a Sylow p-subgroup Q of G and let $P := Q$; while $\langle P^G \rangle \not\leq H$ find $g \in G$ with $P^g \not\leq H$ and replace P by $P \cap H^{g^{-1}}$; this only involves intersecting with p-groups (cf. [Lu1]). Then C is generated by these subgroups P, one for each p. Now $L/N = Z(G/N)$ can be computed as follows: let $\hat{G} = \{(g, g) \mid g \in G\}$, acting on the disjoint union of two copies of our n-set, and let $A = \hat{G}(1 \times G)$, $B = \hat{G}(1 \times N)$ and $C = \mathrm{Core}_A(B)$; then L is the projection of C onto the first copy of G.

The results and methods in [KL] led to the QUOTIENT GROUP THESIS: *If a problem is in polynomial time for permutation groups then it is also in polynomial time for quotients of permutation groups.* It is not clear how this thesis could ever be proved, but it holds for all "standard" questions concerning groups given as permutation groups. However, ridiculously different methods, involving simple groups, appear to be needed in the cases of permutation groups and their quotients.

FASTER ALGORITHMS As already mentioned, finding $|G|$ requires time $O(|S|n^2 + n^5)$ by the methods in [Si1,Si2,FHL,Kn,Je]. The exponent 5 can be decreased by a very different method based on a nonalgorithmic CFSG-based property of primitive permutation groups [Ca]: *If* $G \leq S_n$ *is primitive and* $|G| > n^{2 \log n}$*, then* $n = \binom{m}{l}^k$ *for some* m, l, k*, and* G *is a subgroup of* $S_m \mathrm{wr} S_k$ *with socle* $(A_m)^k$ *acting on the ordered* k*-tuples of* l*-subsets of an* m*-set.* Thus, reductions to primitive groups lead either to groups that are not too big or to ones that are easily understood. This in turn led to an $O(|S|n^3 \log^c n)$ *time algorithm for finding* $|G|$ [BLS2].

This faster method arose from the study of the theoretical feasibility of parallel computation with permutation groups, in which one allows a polynomial number of processors running in time $O(\log^c(|S|n))$ (the complexity class NC). Sims's method for order and membership testing is unavailable: it requires sequentially using a pointwise stabilizer sequence. Nevertheless, methods that later decreased the exponent 5 had already put those problems into NC [BLS1]—and are also used in new methodology described in the next section and employed in GAP. These algorithms rely heavily on CFSG, as do parallel algorithms for finding a

composition series [BLS1] and Sylow subgroups [KLM].

4. NEARLY LINEAR ALGORITHMS

It takes time at least $|S|n$ to read $|S|$ permutations. It is remarkable that quite a few algorithms run in time not too far from this linear lower bound. A *nearly linear* algorithm for a permutation group $G = \langle S \rangle \leq S_n$ is one running in time $O(|S|n \log^c |G|)$ for some constant c. When applied to groups having a base of size $O(\log^{c'} n)$ for some c', $O(|S|n \log^c |G|)$ becomes $O(|S|n \log^{c+c'} n)$, which is very close to linear in the input length. Note that the classical simple groups, in all of their permutation representations, have bases of size $O(\log^2 n)$. It appears that most practical group-theoretic computations involve either small-base groups or alternating or symmetric groups.

RANDOMIZED ALGORITHMS I will need informal definitions of Las Vegas and Monte Carlo algorithms. Both are randomized algorithms. *The output of a* Monte Carlo algorithm *may be incorrect*, but that only can happen with a small, user-prescribed probability. So there is always an output, but there is also the uncomfortable possibility of error. *The output of a* Las Vegas algorithm *is correct*, but there is a small, user-prescribed probability that nothing is output. This is more comforting.

Most known nearly linear algorithms are Monte Carlo. These are of more than theoretical interest, since a large part of the permutation group library in GAP is based on implementations of nearly linear algorithms for many of the problems discussed so far: finding $|G|$, the derived series and a composition series; and soon, finding and conjugating Sylow subgroups with some restrictions on the noncyclic composition factors [Mo]. Many of these algorithms are described in detail in [Ser2]. In GAP the possibility of erroneous output presently is avoided by applying an $O(|S|n^2 \log^c |G|)$ algorithm essentially due to Sims in order to check the correctness of point stabilizer constructions [Ser2]. Now, under mild restrictions, all of these nearly linear Monte Carlo algorithms can be upgraded to nearly linear Las Vegas ones, using algorithms Seress will program into GAP:

THEOREM 4 [KS1,KS2,KM]. *There are* nearly linear Las Vegas *algorithms which, when given $G = \langle S \rangle \leq S_n$ with no composition factor isomorphic to any $PSU(3,q)$, ${}^2B_2(q)$, ${}^2G_2(q)$ or ${}^2F_4(q)$, determine the following*: $|G|$, *membership in G, a composition series for G, and everything else previously found only by Monte Carlo algorithms.*

The proof starts with a known Monte Carlo algorithm that finds a composition series for G. For an alleged simple group that is a (composition) factor of this series, determine its order and isomorphism type and use a very fast constructive recognition test (cf. §5). If the test fails, output nothing (the probability of this is small); otherwise verify the precise composition factors of G, hence find $|G|$. Knowing $|G|$ with certainty in turn allows the outputs of all other known nearly linear algorithms to be verified with certainty.

Following [BLS1], this approach departs significantly from the standard methods based on Sims's point stabilizer ideas: by the time we have $|G|$ we have a composition series for G. Very fast simple group recognition algorithms were essential.

These have much wider applicability, as will be seen in the next section.

5. Black box groups

In §2 I mentioned a general computational setting for a group $G = \langle S \rangle$. The most-studied case of that setting is permutation groups. Another very natural setting is matrix groups: S is a set of invertible matrices over some field, which here is always a finite field. The questions remain the same: efficiently find properties of G, such as $|G|$, solvability, a composition series, etc. If $S \subset GL(d, q)$ then the input length is $|S|d^2 \log q$ (since $\log q$ bits are required to write each of the d^2 entries). These problems seems to be very hard. However, under reasonable additional conditions, and allowing probabilistic algorithms, this has become an actively studied area. Some of the most interesting results have stemmed from ignoring the representation of G on $\mathbb{F}_q^d$ implicit in the above description (hence ignoring eigenvalues, minimal polynomials and so on), and abstracting to the following notion.

A *black box group* is a group G, whose elements are *encoded by binary strings* of the same length N, such that routines ("oracles") are provided for

$$\begin{cases} \text{multiplying two elements,} \\ \text{inverting an element,} \\ \text{deciding whether an element} = 1. \end{cases}$$

Here G is specified as $G = \langle S \rangle$ for some set S of elements. Note that $|G| \leq 2^N$; not all strings correspond to group elements.

The basic examples are permutation groups and matrix groups. However, considering permutation groups as black box groups, so that group operations can be performed much faster than permutation multiplication, has recently become a crucial tool within the study of permutation groups [Ser2,KS1]. Moreover, experience has shown that allowing fewer tools sometimes forces new and better methods.

According to an amazing result of Babai [Ba2], *one can find a nearly uniformly distributed random element of a black box group $G = \langle S \rangle$ using $O(|S|N^5)$ group operations.* This tour de force involves combinatorial methods but nothing about the structure of G; note that $|G|$ is never known here. A practical heuristic algorithm in [CLMNO] for finding random group elements is adequate for Las Vegas algorithms, in which correctness of the output is ultimately verified (cf. [Ba3]).

At present the most general theorem concerning black box groups is

THEOREM 5 [BB,KS1]. *There is a Las Vegas algorithm for the following problem. Suppose that a black box group $G = \langle S \rangle$ is given, together with a list of primes that contains all prime divisors of $|G|$ as well as oracles for handling elementary abelian subgroups of G and discrete logarithms. Then $|G|$ and a composition series for G can be computed in time polynomial in both the input length and the size of the largest field involved in defining the Lie type composition factors of G.*

The additional oracle hypothesis presumes access to "discrete logarithms" (*write any given element of $\mathbb{F}_p^*$ as a power of a given generator*) and "handling elementary abelian subgroups" E (i.e., *the ability to do all standard linear algebra in E*). As originally stated in [BB] the polynomial timing in the theorem also involved a polynomial in the smallest integer $\nu(G)$ such that all nonabelian composition factors of G have faithful permutation representations of degree at most

$\nu(G)$. The slightly stronger result stated above was obtained in [KS1] using simple group recognition algorithms, discussed below. Sylow subgroups can also be found in the situation of the theorem, but the permutation group literature [Ka1] is no longer available for this as it was in [BB].

Recognizing simple groups All modern Sylow subgroup algorithms for permutation groups reduce to the case of simple groups [Ka1,Ka2,Mo,KLM,CCH]. For any given simple permutation group one first determines an explicit isomorphism with a known simple group, afterwards studying Sylow subgroups of the concrete simple groups. Deterministic algorithms producing such isomorphisms are in [Ka1,Ka2,KLM].

Outside the permutation group setting, the problem of *recognizing* simple groups began with algorithms [NeP,NiP,CLG1] for deciding whether a given subgroup $G = \langle S \rangle \leq GL(d,q)$ contains $SL(d,q)$ or a classical group as a normal subgroup. These were *nonconstructive* recognition algorithms, outputting either "G contains a normal classical group", or "G probably does not contain any classical group of $d \times d$ matrices as a normal subgroup". These rely heavily on CFSG: they search for certain matrices in G that occur with high probability in the relevant classical groups, and then use a far-reaching nonalgorithmic consequence of CFSG determining the subgroups of $GL(d,q)$ containing such elements [GPPS].

This suggested the need for *constructive* recognition algorithms: given $G = \langle S \rangle \leq GL(d,q)$ containing $SL(d,q)$, the algorithm in [CLG2] writes any given element of $SL(d,q)$ in terms of S. This has also been done for the symplectic groups [Ce]. The nonconstructive recognition algorithms are Monte Carlo (more precisely, *one-sided Monte Carlo* [Ba3], since an output such as "contains $SL(d,q)$" is guaranteed to be correct), and run in time polynomial in the input length; the constructive ones can be viewed as Las Vegas algorithms whose timing also depends on a small power of q, so that these do not run in polynomial time.

The black box version of constructive recognition is conceptually harder: there is no longer a vector space available, hence no linear algebra to rely on. The goal is an effective isomorphism[4] $\varphi: G \to H$ to a concrete version H of the simple black box group G, whereas only the name of H is output in nonconstructive recognition. It was not at all clear that one could recognize a black box simple group in this manner, but in [CFL] this was shown to be possible when $G \cong PSL(d,2)$.

More generally, *there is a Las Vegas algorithm which, when given a black box group G isomorphic to a simple group of Lie type of known characteristic, constructively recognizes G in time polynomial in the input length and field size* [KS1] (cf. [KM]). The characteristic assumption is removed in [KS3] by assuming instead that there is a method (an oracle) for finding the order of any given element of G. When the characteristic is known, the idea is to (probably) construct an element in a large conjugacy class, one of whose powers is a (long) root element; then construct larger subgroups using random conjugates of these root elements; and ultimately make a recursive call to a group of rank one less than that of G (if G does not already have rank 1). The final step of the algorithm verifies the correctness of the output isomorphism by computing a presentation for G (see below).

[4] Able to find $g\varphi$ or $h\varphi^{-1}$ for any *given* $g \in G$ or $h \in H$.

Here the isomorphism type of G is not part of the input. Note that the Lie rank l and logarithm of the field size q are polynomial in N, since $N \geq \log|G| \geq (l^2 \log q)/4$. While q appears in timing estimates of all known constructive recognition algorithms, it should not be needed, at least when l is large. Analogous results for alternating groups are in [BLNPS].

PRESENTATIONS Las Vegas algorithms recognizing simple black box groups eventually need to verify a presentation of a known simple group, which for use in Theorem 4 must be written in time polynomial only in the input length. In view of the time required to multiply out a permutation g given as a product of generators, verifying that $g = 1$ involves the lengths of presentations[5]. For all simple groups except, perhaps, $PSU(3,q)$, ${}^2B_2(q)$ and ${}^2G_2(q)$, there is a presentation of length $O(\log^c|G|)$, using $c = 2$ and in most cases even $c = 1$; the proof in [BGKLP] uses simple tricks to adapt the usual Curtis-Steinberg-Tits presentations for these groups. It is perhaps surprising that the case of the very familiar groups $PSU(3,q)$ has remained open for almost 10 years. Short presentations have the following nonalgorithmic consequence needed in the proof of Theorem 4: *Every finite group G, with no composition factor of the form $PSU(3,q)$, ${}^2B_2(q)$ or ${}^2G_2(q)$, has a presentation of total length $O(\log^3|G|)$.* The exponent 3 is best possible.

These short presentations also were used in [Ma] to prove the existence of a constant c such that there are at most $n^{cd\log n}$ d-generator groups of order n with no composition factor $PSU(3,q)$, ${}^2B_2(q)$ or ${}^2G_2(q)$, and each such group can be defined by means of $cd\log n$ relations in those d generators. As with Theorem 1, we see that algorithmic needs have led to a result about finite groups. Of course, it is not surprising that many applications of CFSG have needed new properties of simple groups (cf. [GK]).

ACKNOWLEDGMENT: I am grateful to L. Babai, E. Luks, J. Neubüser, Á. Seress and C. Sims for very helpful comments.

REFERENCES

[Ba1] L. Babai, Computational complexity in finite groups, pp. 1479-1489 in: Proc. ICM, Kyoto 1990.

[Ba2] L. Babai, Local expansion of vertex-transitive graphs and random generation in finite groups, pp. 164–174 in: Proc. ACM STOC 1991.

[Ba3] L. Babai, Randomization in group algorithms: conceptual questions, [FK] 1–17.

[BB] R. Beals and L. Babai, Las Vegas algorithms for matrix groups, pp. 427–436 in: Proc. IEEE FOCS 1993.

[BCFLS] L. Babai, G. Cooperman, L. Finkelstein, E. M. Luks and Á. Seress, Fast Monte Carlo algorithms for permutation groups, pp. 90–100 in: Proc. ACM STOC 1991.

[BCFS] L. Babai, G. Cooperman, L. Finkelstein and Á Seress, Nearly linear time algorithms for permutation groups with a small base, pp. 200–209 in: Proc. ISSAC 1991.

[Be] R. Beals, An elementary algorithm for computing the composition factors of a permutation group, pp. 127-134 in: Proc. ISSAC 1993.

[5] The *length* of a presentation $\langle X \mid R\rangle$ is $|X| + \Sigma_{r\in R}\, l_X(r)$.

[BGKLP] L. Babai, A. J. Goodman, W. M. Kantor, E. M. Luks and P. P. Pálfy, Short presentations for finite groups. J. Algebra 194 (1997) 79–112.

[BLNPS] R. Beals, C. R. Leedham-Green, A. C. Niemeyer, C. E. Praeger and Á. Seress, A mélange of black box algorithms for recognising finite symmetric and alternating groups (in preparation).

[BLS1] L. Babai, E. M. Luks and Á. Seress, Permutation groups in NC, pp. 409–420 in Proc. ACM STOC 1987.

[BLS2] L. Babai, E. M. Luks and Á. Seress, Fast management of permutation groups I. SIAM J. Comput. 26 (1997) 1310–1342; II (in preparation).

[Ca] P. Cameron, Finite permutation groups and finite simple groups. BLMS 13 (1981) 1–22.

[CCH] J. Cannon, B. Cox and D. F. Holt, Computing Sylow subgroups in permutation groups (to appear).

[Ce] F. Celler, Matrixgruppenalgorithmen in GAP. Ph. D. thesis, RWTH Aachen 1997.

[CFL] G. Cooperman, L. Finkelstein and S. Linton, Recognizing $GL_n(2)$ in non-standard representation, [FK] 85–100.

[CLG1] F. Celler and C. R. Leedham-Green, A non-constructive recognition algorithm for the special linear and other classical groups, [FK] 61–67.

[CLG2] F. Celler and C. R. Leedham-Green, A constructive recognition algorithm for the special linear group (to appear in Proc. ATLAS Conf.).

[CLMNO] F. Celler, C. R. Leedham-Green, S. H. Murray, A. C. Niemeyer and E. A. O'Brien, Generating random elements of a finite group. Comm. in Alg. 23 (1995) 4931–4948.

[CP] J. Cannon and C. Playoust, An introduction to Magma. School of Math. and Stat., Univ. Sydney, 1993. magma@maths.usyd.edu.au

[FHL] M. Furst, J. Hopcroft and E. Luks, Polynomial–time algorithms for permutation groups, pp. 36–41 in Proc. IEEE FOCS 1980.

[FK] L. Finkelstein and W. M. Kantor, Groups and Computation II, AMS 1997.

[Ga] A. Gambini, Zur Komplexität einiger gruppentheoretischer Algorithmen, Ph. D. thesis, Univ. Freiburg 1992.

[GK] R. M. Guralnick and W. M. Kantor, Some applications of the classification of finite simple groups (in preparation).

[GPPS] R. M. Guralnick, T. Penttila, C. E. Praeger and J. Saxl, Linear groups with orders having certain primitive prime divisors (submitted).

[IKS] I. M. Isaacs, W. M. Kantor and N. Spaltenstein, On the probability that a group element is p–singular. J. Algebra 176 (1995) 139–181.

[Je] M. R. Jerrum, A compact representation for permutation groups. J. Algorithms 7 (1986) 60–78.

[Ka1] W. M. Kantor, Sylow's theorem in polynomial time. J. Comp. Syst. Sci. 30 (1985) 359–394.

[Ka2] W. M. Kantor, Finding Sylow normalizers in polynomial time. J. Algor. 11 (1990) 523–563.

[KL] W. M. Kantor and E. M. Luks, Computing in quotient groups, pp. 524–534 in: Proc. ACM STOC 1990.

[KLM] W. M. Kantor, E. M. Luks and P. D. Mark, Sylow subgroups in parallel (to appear in J. Algorithms).

[KM] W. M. Kantor and K. Magaard, Black box exceptional groups of Lie type (in preparation).

[Kn] D. E. Knuth, Efficient representation of perm groups. Combinatorica 11 (1991) 33–43.

[KS1] W. M. Kantor and Á. Seress, Black box classical groups (submitted).

[KS2] W. M. Kantor and Á. Seress, Permutation group algorithms via black box recognition algorithms (to appear in Proc. Bath Conf.).

[KS3] W. M. Kantor and Á. Seress (in preparation).

[Lu1] E. M. Luks, Isomorphism of graphs of bounded valence can be tested in polynomial time. J. Comp. Syst. Sci. 25 (1982) 42–65.

[Lu2] E. M. Luks, Computing the composition factors of a permutation group in polynomial time. Combinatorica 7 (1987) 87–99.

[Lu3] E. M. Luks, Permutation groups and polynomial time computation, pp. 139–175 in: Groups and Computation (eds. L. Finkelstein, W. M. Kantor), AMS 1993.

[Ma] A. Mann, Enumerating finite groups and their defining relations (to appear in J. Group Theory).

[Mo] P. Morje, A nearly linear algorithm for Sylow subgroups of permutation groups. Ph.D. thesis, Ohio State U. 1995.

[NeP] P. M. Neumann and C. E. Praeger, A recognition algorithm for special linear groups, PLMS 65 (1992) 555–603.

[NiP] A. C. Niemeyer and C. E. Praeger, Implementing a recognition algorithm for classical groups, [FK] 273–296.

[Sch] M. Schönert et. al., GAP: Groups, Algorithms, and Programming. Lehrstuhl D für Mathematik, RWTH Aachen, 1994. gap@dcs.st-and.ac.uk

[Ser1] Á. Seress, An introduction to computational group theory. Notices AMS 44 (1997) 671–679.

[Ser2] Á. Seress, Permutation group algorithms (Cambridge University Press, to appear).

[Si1] C. C. Sims, Computational methods in the study of permutation groups, pp. 169–183 in: Computational problems in abstract algebra (ed. J. Leech), Pergamon 1970.

[Si2] C. C. Sims, Computation with permutation groups, pp. 23–28 in: Proc. Symp. Symb. Alg. Manipulation (ed. S. R. Petrick), ACM 1971.

[Si3] C. C. Sims, Group-theoretic algorithms, a survey, pp. 979–985 in: Proc. ICM, Helsinki 1978.

William M. Kantor
U. of Oregon
Eugene, OR 97403, USA

Spetses

Gunter Malle

Abstract. We report on the properties of unipotent degrees of complex reflection groups.

1991 Mathematics Subject Classification: Primary 20F; Secondary 20G.
Keywords and Phrases: complex reflection groups, spetses, unipotent degrees, fusion rules.

1. Introduction

About two hours by boat south of Athens in the Aegean sea lies the small island Spetses. On a conference there in July 1993 Michel Broué first asked whether maybe every finite complex reflection group occurs as Weyl group of some object which is an analogue of a finite group of Lie type. This question has instigated some fruitful research and led to the discovery of fascinating structures associated to finite reflection groups, for example the so-called unipotent degrees (see [14]). It therefore seems appropriate to call these yet unknown objects *spetses*.

It was first noted by Springer [20] that non-real reflection groups naturally appear inside Weyl groups as what is now called relative Weyl groups, that is, normalizers modulo centralizers of subspaces. The importance of this construction was revealed in the work [3] of Broué, Michel and the author on the ℓ-blocks of characters of finite groups of Lie type. There it was shown that the unipotent characters inside an ℓ-block are parametrized by the irreducible characters of a relative Weyl group, which in general is a non-real reflection group. A possible interpretation of this result was subsequently proposed by Broué and the author [2] in terms of the so-called cyclotomic Hecke algebra attached to a finite complex reflection group (see also [6]).

The results of Lusztig show that the set of unipotent characters of a finite group of Lie type only depends on the Weyl group of the associated algebraic group, together with the action of the Frobenius endomorphism on it. In the course of this classification Lusztig observed that similar sets can formally also be attached to those finite real reflection groups which are not Weyl groups [11,12].

It is the purpose of this article to give a brief introduction into a similar construction, for complex reflection groups, of 'unipotent characters', Fourier transform matrices, and eigenvalues of Frobenius, which satisfy combinatorial properties like unipotent characters of actual finite groups of Lie type, just as if there existed an algebraic group whose Weyl group is non-real. It is tempting to speculate about an underlying algebraic structure giving rise to the unipotent characters attached to complex reflection groups. Until now, such an object has not been found, but a lot of intriguing evidence for its existence has been collected.

We would like to conclude this introduction by stating that many of the properties of unipotent degrees were found by computer experiments, and some of the results for large exceptional spetses could only be verified using computer algebra systems. We think that this might serve as a good illustration of the power of experimental/computational algebra.

2. Reflection data

2.1. Complex reflection groups. Let V be a finite dimensional vector space over a subfield k of the field of complex numbers $\mathbb{C}$. An element $1 \neq \sigma \in \mathrm{GL}(V)$ is called a *(complex) reflection* if σ pointwise fixes some hyperplane in V. A finite subgroup $W \leq \mathrm{GL}(V)$ generated by complex reflections will be called a *complex reflection group*. Thus the finite Coxeter groups, being real reflection groups, are examples of complex reflection groups, and in particular all finite Weyl groups fall into this class.

A *parabolic subgroup* of a complex reflection group W is the centralizer (pointwise stabilizer) in W of some subspace of V. It is a remarkable result of Steinberg that all parabolic subgroups of a complex reflection group are again generated by reflections.

Let $S(V)$ denote the symmetric algebra of V. By the theorem of Shephard-Todd and Chevalley, the ring of invariants $S(V)^W$ of W in $S(V)$ is a polynomial ring. The quotient $S(V)_W$ of $S(V)$ by the ideal generated by the invariants of strictly positive degree is called the *coinvariant algebra*. Chevalley has shown that the W-module $S(V)_W$ affords a graded version of the regular representation. For an irreducible character $\chi \in \mathrm{Irr}(W)$ the *fake degree* is defined as the graded multiplicity of χ in $S(V)_W$ (the generating function for the embedding degrees),

$$R_\chi := \langle S(V)_W, \chi \rangle_W = (x-1)^{\dim(V)} P_W \frac{1}{|W|} \sum_{w \in W} \frac{\det_V(w)\chi(w)}{\det_V(x-w)} \ \in \mathbb{Z}[x]\,,$$

where P_W denotes the Poincaré polynomial of the coinvariant algebra $S(V)_W$. The fake degrees can be considered as a first elementary approximation of the unipotent degrees of W from which they differ in a subtle way (see Section 5.1).

2.2. Families of groups of Lie type. Our aim is to introduce for a complex reflection group W an object which behaves like a family of finite groups of Lie type with Weyl group W. To motivate this, let first $\mathbf{G}$ be a connected reductive algebraic group over the algebraic closure of a finite field of positive characteristic. We assume that $\mathbf{G}$ is already defined over the finite field $\mathbb{F}_q$ and let $F : \mathbf{G} \to \mathbf{G}$ be the corresponding Frobenius morphism. The group of fixed points $\mathbb{G}(q) := \mathbf{G}^F$ is then a finite group of Lie type. Let $\mathbf{T}$ be an F-stable maximal torus of $\mathbf{G}$ contained in an F-stable Borel subgroup of $\mathbf{G}$, and Y the cocharacter group of $\mathbf{T}$. Via its action on Y, W can be considered as a subgroup of the automorphism group of the real vector space $Y_\mathbb{R} := Y \otimes_\mathbb{Z} \mathbb{R}$. The Frobenius endomorphism F also acts on $Y_\mathbb{R}$ as a product $q\phi$ where ϕ is an automorphism of finite order normalizing W. Replacing the Borel subgroup by another one containing $\mathbf{T}$ changes ϕ by an element of the Weyl group W of $\mathbf{G}$ with respect to $\mathbf{T}$. Hence ϕ is determined as automorphism of $Y_\mathbb{R}$ up to elements of W.

Conversely, the real vector space $Y_{\mathbb{R}}$ together with the actions of W and ϕ determines F and $\mathbf{G}$ up to isogeny. In this way, a whole series $\{\mathbb{G}(q) \mid q \text{ a prime power}\}$ of finite groups of Lie type can be encoded by the data $(Y_{\mathbb{R}}, W\phi)$.

2.3. REFLECTION DATA. This leads us to consider complex reflection groups together with certain automorphisms. We define a *reflection datum with Weyl group* W to be a pair $\mathbb{G} = (V, W\phi)$ where W is a complex reflection group on V and $\phi \in \mathrm{GL}(V)$ normalizes W. A *Levi subdatum of* $\mathbb{G}$ is a reflection datum $\mathbb{L} = (V, W_{\mathbb{L}}w\phi)$ where $w \in W$ and $W_{\mathbb{L}}$ is a $w\phi$-stable parabolic subgroup of W. A reflection datum with trivial Weyl group $W = 1$ is called a *torus*. For a torus $\mathbb{S} = (V', (w\phi)|_{V'})$ of $\mathbb{G}$ we define its *centralizer* to be the Levi subdatum

$$C_{\mathbb{G}}(\mathbb{S}) := (V, C_W(V')w\phi)$$

(note that $C_W(V')$ is a reflection subgroup of W by Steinberg's theorem).

2.4. SYLOW THEORY. Let $\mathbb{G}$ be a reflection datum. The *(polynomial) order of* $\mathbb{G}$ is defined as

$$|\mathbb{G}| := \frac{x^N}{\frac{1}{|W|}\sum_{w\in W} \frac{\det_V(w)}{\det_V(x-w\phi)}},$$

where N is the number of reflecting hyperplanes of W in V. For example, if $\phi = 1$ then $|\mathbb{G}| = (x-1)^{\dim(V)}P_W$, and if $\mathbb{G} = (V, \phi)$ is a torus then $|\mathbb{G}| = \det_V(x - \phi)$ is the characteristic polynomial of ϕ on V. Steinberg has shown that in the case of groups of Lie type, $|\mathbb{G}|(q)$ gives the order of $\mathbb{G}(q)$. By an extension of Molien's formula for the ring of invariants $S(V)^W$ it follows that $|\mathbb{G}|$ is a product of cyclotomic polynomials over k, that is, of k-irreducible polynomials whose roots are roots of unity.

Let Φ be a cyclotomic polynomial over k. A torus $\mathbb{S}$ is called a Φ-*torus* if its order $|\mathbb{S}|$ is a power of Φ. Thus $\mathbb{S} = (V, \phi)$ is a Φ-torus if and only if all eigenvalues of ϕ on V are roots of Φ. The centralizers of generic Φ-tori of $\mathbb{G}$ are called Φ-*split* Levi subdata. Note that, in particular, $\mathbb{G}$ itself is Φ-split. The results of Springer [20, 3.4 and 6.2] on eigenspaces of elements in complex reflection groups imply that the Φ-tori of reflection data satisfy an analogue of Sylow theory:

THEOREM 2.5. *Let $\mathbb{G} = (V, W\phi)$ be a reflection datum and $\Phi \neq x$ a prime divisor of $|\mathbb{G}|$ over k.*

(a) There exist non-trivial Φ-tori of $\mathbb{G}$.

(b) For any maximal Φ-torus $\mathbb{S}$ of $\mathbb{G}$ we have $|\mathbb{S}| = \Phi^{a(\Phi)}$, where $a(\Phi)$ is the precise power of Φ dividing $|\mathbb{G}|$.

(c) Any two maximal Φ-tori of $\mathbb{G}$ are W-conjugate.

The concept of reflection data allows to capture much of the structural properties of groups of Lie type. The unipotent degrees play a similar role for the description of the irreducible representations.

2.6. CLASSIFICATION. The finite irreducible complex reflection groups have been classified by Shephard and Todd [19]. The irreducible groups fall into an infinite series of monomial groups $G(m, p, n)$ (with $n \geq 1$, $m \geq 2$, $p|m$), the symmetric groups, and 34 exceptional groups which all occur in dimension at most eight. All the results to be given in the sequel depend on this classification in the sense that their proofs are case-by-case.

3. Cyclotomic Hecke algebras

We now introduce a deformation of the group algebra of a complex reflection group which generalizes the usual Iwahori-Hecke algebra of a Coxeter group.

3.1. Cyclotomic algebras. Starting from the classification mentioned in 2.6 one can find for any complex reflection group a so-called *good presentation*

$$W = \langle s \in S \mid s^{d_s} = 1, \text{ certain homogeneous relations} \rangle$$

similar to the Coxeter presentation of real reflection groups with various good properties. For example, the elements of S map to reflections, and S has minimal size subject to this (equal to n or $n+1$ if $W \leq \mathrm{GL}_n(k)$ is irreducible). Moreover, any subset S' of S together with those relations involving only elements of S' gives a presentation of a parabolic subgroup of W. However, in contrast to the Coxeter case, no conceptual definition of good presentations is available at present.

Let $\mathbf{u} = (u_{s,j} \mid s \in S, 0 \leq j \leq d_s - 1)$ be transcendentals over $\mathbb{Z}$, such that $u_{s,j} = u_{t,j}$ whenever s and t are conjugate in W. Let $A := \mathbb{Z}[\mathbf{u}, \mathbf{u}^{-1}]$. The *generic cyclotomic Hecke algebra* $\mathcal{H}(W, \mathbf{u})$ *of* W *with parameter set* $\mathbf{u}$ is defined to be the A-algebra on generators $\{T_s \mid s \in S\}$ subject to the homogeneous relations from the good presentation of W, and the deformed order relations

$$\prod_{j=0}^{d_s-1} (T_s - u_{s,j}) = 0 \qquad (s \in S).$$

For a ring homomorphism $f : A \to R$ we write $\mathcal{H}_R = \mathcal{H} \otimes_A R$ for the corresponding specialization of $\mathcal{H} := \mathcal{H}(W, \mathbf{u})$. Clearly, such a homomorphism is uniquely determined by the images $f(u_{s,j})$. Under the specialization defined by

$$u_{s,j} \mapsto \exp(2\pi i j/d_s) \qquad \text{for } s \in S,\ 0 \leq j \leq d_s - 1, \tag{3.2}$$

$\mathcal{H}$ maps to the group algebra of the complex reflection group W. Any specialization through which (3.2) factors will be called *admissible*. One particularly important example is the *1-parameter specialization* $\mathcal{H}(W, x)$ of $\mathcal{H}(W, \mathbf{u})$ induced by the map

$$u_{s,j} \mapsto \begin{cases} x & j = 0, \\ \exp(2\pi i j/d_s) & j > 0, \end{cases} \tag{3.3}$$

where x is an indeterminate. This is the analogue of the classical 1-parameter Iwahori-Hecke algebra for real W.

3.4. Braid groups. Let us sketch a more conceptual construction of cyclotomic algebras (see [5]). For an irreducible complex reflection group $W \leq \mathrm{GL}(V)$ we let $\mathcal{A}$ be the set of the reflecting hyperplanes of W and denote by

$$M := V \setminus \bigcup_{H \in \mathcal{A}} H$$

the complement. For a fixed base point $x_0 \in M$ we define the *pure braid group of* W as the fundamental group $P(W) := \pi_1(M, x_0)$. Let $\bar{M}$ be the quotient of M

by W. By the theorem of Steinberg on parabolic subgroups M is an unramified Galois cover of $\bar{M}$, with group W. Thus we have the canonical exact sequence

$$1 \longrightarrow P(W) \longrightarrow B(W) \longrightarrow W \longrightarrow 1$$

with the *braid group* $B(W) := \pi_1(\bar{M}, \bar{x}_0)$, where $\bar{x}_0$ is the image of x_0. To each hyperplane $H \in \mathcal{A}$ is attached a class of elements in $B(W)$, the generators of monodromy around H, which in W map to reflections along H. It is shown in [5] (for all but six irreducible types) that $\mathcal{H}(W, \mathbf{u})$ is isomorphic to the quotient

$$\mathcal{H}(W, \mathbf{u}) := \mathbb{Z}[\mathbf{u}, \mathbf{u}^{-1}]B(W) \Big/ \Big(\prod_{j=0}^{d_s - 1} (\mathbf{s} - u_{s,j}) \mid \mathbf{s} \text{ generator of monodromy} \Big),$$

of the group algebra of $B(W)$ by the ideal generated by the deformed order relations. This approach gives a definition of cyclotomic algebras independent of the choice of a good presentation. It also allows to study $\mathcal{H}(W, \mathbf{u})$ via monodromy representations of the braid group, see [5,18]. For example, Opdam uses this to derive symmetry properties of the fake degrees of W.

3.5. STRUCTURE OF CYCLOTOMIC ALGEBRAS. The following important structure result for $\mathcal{H}(W, \mathbf{u})$ has been proved for all but finitely many irreducible complex reflection groups (see [1,2]); it is conjectured to hold in all cases:

THEOREM 3.6. *The cyclotomic Hecke algebra $\mathcal{H}(W, \mathbf{u})$ is free over $A = \mathbb{Z}[\mathbf{u}, \mathbf{u}^{-1}]$ of rank $|W|$.*

Assume that W is defined over the number field k. Let $\mu(k)$ denote the group of roots of unity in k and let $\mathbf{v} = (v_{s,j} \mid s \in S,\ 0 \le j \le d_s - 1)$ where $v_{s,j}^{|\mu(k)|} = \exp(-2\pi i j/d_s) u_{s,j}$. In terms of this, one has ([15, Theorem 5.2]):

THEOREM 3.7. *The field $K_W := k(\mathbf{v})$ is a splitting field for $\mathcal{H}(W, \mathbf{u})$.*

In the rational case, when W is a Weyl group and $k = \mathbb{Q}$, we have $|\mu(k)| = 2$ and thus recover the classical result of Benson/Curtis and Lusztig.

It follows from Theorem 3.6, the fact that k is a splitting field for W, and Tits' deformation theorem that $\mathcal{H}_{K_W}$ is isomorphic to the group algebra $K_W W$. Thus any extension to $\mathbb{Z}[\mathbf{v}, \mathbf{v}^{-1}]$ of the specialization (3.2) defines a bijection $\mathrm{Irr}(W) \xrightarrow{\sim} \mathrm{Irr}(\mathcal{H}_{K_W})$, $\chi \mapsto \chi_{\mathbf{v}}$, between $\mathrm{Irr}(W)$ and $\mathrm{Irr}(\mathcal{H}_{K_W})$.

Since the group algebra $K_W W$ is symmetric, the same is true for $\mathcal{H}_{K_W}$. But in fact, it is known for all but finitely many irreducible W that this statement already holds over A (see [17]):

THEOREM 3.8. *The cyclotomic algebra $\mathcal{H}(W, \mathbf{u})$ is a symmetric algebra over A.*

Let us choose a symmetric form $\langle\ ,\ \rangle : \mathcal{H} \otimes \mathcal{H} \to A$ on $\mathcal{H}$ with Gram matrix invertible over A such that the associated trace form $t_{\mathbf{u}} : \mathcal{H}(W, \mathbf{u}) \to A$ defined by $t_{\mathbf{u}}(h) := \langle 1, h \rangle$ under (3.2) specializes to the canonical trace form on the group algebra of W. Over the splitting field K_W of $\mathcal{H}(W, \mathbf{u})$, we may write $t_{\mathbf{u}}$ as a sum over the irreducible characters of $\mathcal{H}_{K_W}$ with non-vanishing coefficients, so

$$t_{\mathbf{u}} = \sum_{\chi \in \mathrm{Irr}(W)} \frac{1}{c_\chi} \chi_{\mathbf{v}}\,,$$

where c_χ is integral over A. The c_χ are called *Schur elements* of $\mathcal{H}(W, \mathbf{u})$ (with respect to $t_\mathbf{u}$).

3.9. SPETSIAL REFLECTION GROUPS. We now come to an important property of some complex reflection groups which seems to lie at the heart of the existence of unipotent degrees. Let W be a finite complex reflection group defined over k, $\mathbb{Z}_k$ the ring of integers of k, $\mathcal{H}(W, x)$ the 1-parameter cyclotomic algebra (3.3) over $\mathbb{Z}[x, x^{-1}]$, $k(y)$ with $y^{|\mu(k)|} = x$ a splitting field for $\mathcal{H}(W, x)$ (see Theorem 3.7). Let $\chi \in \mathrm{Irr}(W)$. The *generic degree* $\delta_\chi := P(W)/c_\chi$ *of* χ is a Laurent polynomial in y. We write $a(\chi)$ for the order of zero of δ_χ at $y = 0$ divided by $|\mu(k)|$, and $b(\chi)$ for the order of zero of the fake degree R_χ at $x = 0$. If $a(\chi) = b(\chi)$ then χ is called *special*. In all cases where the generic degrees are explicitly known, the following can be checked (see [16]):

PROPOSITION 3.10. *The following are equivalent:*

(i) for all $\chi \in \mathrm{Irr}(W)$ *there exists a special* $\psi \in \mathrm{Irr}(W)$ *with* $a(\chi) = a(\psi)$*;*
(ii) (rationality) $\delta_\chi \in k(x)$ *for all* $\chi \in \mathrm{Irr}(W)$*;*
(iii) (integrality) $\delta_\chi \in k[y]$ *for all* $\chi \in \mathrm{Irr}(W)$*;*
(iv) $\delta_1 = 1$*;*
(v) (representability) the k*-subspaces* $\langle \delta_\chi \mid \chi \rangle$ *and* $\langle R_\chi \mid \chi \rangle$ *of* $k(y)$ *coincide.*

Except for the obvious implications, no a priori proof is known for any of these statements.

A reflection group satisfying the above (very special) equivalent conditions will be called *spetsial*. It can be checked that all parabolic subgroups of spetsial reflection groups are again spetsial. A reflection datum with spetsial reflection group is called a *spets*. Thus, Levi subdata of spetses are again spetses.

The spetsial reflection groups include all the real ones, as well as all those irreducible n-dimensional ones which are generated by n reflections of order two. The complete list can be found in [16].

4. UNIPOTENT DEGREES

4.1. UNIPOTENT CHARACTERS. Let $\mathbb{G} = (V, W\phi)$ be a spets with rational Weyl group W and $\{\mathbb{G}(q) \mid q \text{ prime power}\}$ an associated family of finite groups of Lie type as in Section 2.3. By the results of Lusztig the unipotent characters $\mathcal{E}(\mathbb{G}(q))$ of the groups $\mathbb{G}(q)$ can be parametrized by a set $\mathcal{E}(\mathbb{G})$ independent of q, and there is a function $\deg : \mathcal{E}(\mathbb{G}) \to k[x]$, $\gamma \mapsto \deg(\gamma)$, such that for any choice of q there is a bijection $\psi_q^\mathbb{G} : \mathcal{E}(\mathbb{G}) \to \mathcal{E}(\mathbb{G}(q))$ such that $\psi_q^\mathbb{G}(\gamma)$ has degree $\psi_q^\mathbb{G}(\gamma)(1) = \deg(\gamma)(q)$ (see [11,3]). Then $\mathcal{E}(\mathbb{G})$ is called the set of *(generic) unipotent characters of* $\mathbb{G}$. This set has many interesting combinatorial properties, some of which we will now describe. We formulate these for the case that W is a Weyl group, $k = \mathbb{Q}$, but the reader should already have in mind the case of a more general reflection group.

4.2. GENERALIZED HARISH-CHANDRA THEORY. The functors of Lusztig induction and restriction give, for any Levi subspets $\mathbb{L}$ of $\mathbb{G}$ linear maps

$$R_\mathbb{L}^\mathbb{G} : \mathbb{Z}\mathcal{E}(\mathbb{L}) \to \mathbb{Z}\mathcal{E}(\mathbb{G}), \quad {}^*R_\mathbb{L}^\mathbb{G} : \mathbb{Z}\mathcal{E}(\mathbb{G}) \to \mathbb{Z}\mathcal{E}(\mathbb{L}),$$

satisfying $\psi_q^{\mathbb{G}} \circ R_{\mathbb{L}}^{\mathbb{G}} = R_{\mathbb{L}(q)}^{\mathbb{G}(q)} \circ \psi_q^{\mathbb{L}}$ for all q (when extending $\psi_q^{\mathbb{G}}$ linearly to $\mathbb{Z}\mathcal{E}(\mathbb{G})$), where $R_{\mathbb{L}(q)}^{\mathbb{G}(q)}$ denotes Lusztig induction between finite reductive groups $\mathbb{L}(q) \leq \mathbb{G}(q)$ associated to $\mathbb{L}, \mathbb{G}$.

Let Φ be a cyclotomic polynomial over k dividing $|\mathbb{G}|$. A unipotent character $\gamma \in \mathcal{E}(\mathbb{G})$ is called Φ-*cuspidal* if ${}^*R_{\mathbb{L}}^{\mathbb{G}}(\gamma) = 0$ for any Φ-split proper Levi subspets $\mathbb{L}$ of $\mathbb{G}$. It can be shown that this is equivalent to γ being of central Φ-defect, that is, to $|\mathbb{G}_{\mathrm{ss}}|/\deg(\gamma)$ not being divisible by Φ. Here, $\mathbb{G}_{\mathrm{ss}}$ is the semisimple quotient $(V/V^W, W\phi)$ of $\mathbb{G}$.

A pair $(\mathbb{L}, \lambda)$ consisting of a Φ-split Levi subspets of $\mathbb{G}$ and a unipotent character $\lambda \in \mathcal{E}(\mathbb{L})$ is called a Φ-*split pair*. It is called Φ-*cuspidal* if moreover λ is Φ-cuspidal. Let $(\mathbb{M}_1, \mu_1)$ and $(\mathbb{M}_2, \mu_2)$ be Φ-split in $\mathbb{G}$. Then we say that $(\mathbb{M}_1, \mu_1) \leq_\Phi (\mathbb{M}_2, \mu_2)$ if $\mathbb{M}_1$ is a Φ-split Levi subspets of $\mathbb{M}_2$ and μ_2 occurs in $R_{\mathbb{M}_1}^{\mathbb{M}_2}(\mu_1)$.

For a Φ-cuspidal pair $(\mathbb{L}, \lambda)$ of $\mathbb{G}$ we write

$$\mathcal{E}(\mathbb{G}, (\mathbb{L}, \lambda)) := \{\gamma \in \mathcal{E}(\mathbb{G}) \mid (\mathbb{L}, \lambda) \leq_\Phi (\mathbb{G}, \gamma)\}$$

for the set of unipotent characters of $\mathbb{G}$ lying above $(\mathbb{L}, \lambda)$. We call $\mathcal{E}(\mathbb{G}, (\mathbb{L}, \lambda))$ the Φ-*Harish-Chandra series above* $(\mathbb{L}, \lambda)$ because of the following fundamental result (see [3]), which is a complete analogue of the usual Harish-Chandra theory (the case $\Phi = x - 1$):

THEOREM 4.3. *(a) (Disjointness) The sets $\mathcal{E}(\mathbb{G}, (\mathbb{L}, \lambda))$ (where $(\mathbb{L}, \lambda)$ runs over a system of representatives of the $W_{\mathbb{G}}$-conjugacy classes of Φ-cuspidal pairs) form a partition of $\mathcal{E}(\mathbb{G})$.*

(b) (Transitivity) Let $(\mathbb{L}, \lambda)$ be Φ-cuspidal and $(\mathbb{M}, \mu)$ be Φ-split such that $(\mathbb{L}, \lambda) \leq_\Phi (\mathbb{M}, \mu)$ and $(\mathbb{M}, \mu) \leq_\Phi (\mathbb{G}, \gamma)$. Then $(\mathbb{L}, \lambda) \leq_\Phi (\mathbb{G}, \gamma)$.

4.4. PERFECT ISOMETRIES. The only known proof of Theorem 4.3 in the case $\Phi \neq x - 1$ consists in the explicit determination of the Lusztig induced of Φ-cuspidal unipotent characters. To state this result from [3] we need to introduce an important invariant of a Φ-Harish-Chandra series. Let $(\mathbb{L}, \lambda)$ be a Φ-cuspidal pair in $\mathbb{G}$. By results of Lusztig it is possible to define an action of $N_W(W_{\mathbb{L}})/W_{\mathbb{L}}$ on $\mathcal{E}(\mathbb{L})$ which is the generic version of the corresponding actions in the series of finite groups of Lie type attached to $\mathbb{G}$ (see [3]). We then call $W_{\mathbb{G}}(\mathbb{L}, \lambda) := N_W(W_{\mathbb{L}}, \lambda)/W_{\mathbb{L}}$ the *relative Weyl group of* $(\mathbb{L}, \lambda)$ *in* $\mathbb{G}$.

THEOREM 4.5. *For each Φ there exists a collection of isometries*

$$I_{(\mathbb{L},\lambda)}^{\mathbb{M}} : \mathbb{Z}\mathrm{Irr}(W_{\mathbb{M}}(\mathbb{L}, \lambda)) \to \mathbb{Z}\mathcal{E}(\mathbb{M}, (\mathbb{L}, \lambda)),$$

such that for all $\mathbb{M}$ and all $(\mathbb{L}, \lambda)$ we have

$$R_{\mathbb{M}}^{\mathbb{G}} \circ I_{(\mathbb{L},\lambda)}^{\mathbb{M}} = I_{(\mathbb{L},\lambda)}^{\mathbb{G}} \circ \mathrm{Ind}_{W_{\mathbb{M}}(\mathbb{L},\lambda)}^{W_{\mathbb{G}}(\mathbb{L},\lambda)}.$$

Here $\mathbb{M}$ runs over the Φ-split Levi subgroups of $\mathbb{G}$ and $(\mathbb{L}, \lambda)$ over the set of Φ-cuspidal pairs of $\mathbb{M}$.

For $\Phi = x - 1$ and $\lambda = 1$, this gives an embedding $\mathrm{Irr}(W^\phi) \subseteq \mathcal{E}(\mathbb{G})$; the image consists of the so-called principal series unipotent characters.

4.6. GENERALIZED HOWLETT/LEHRER-LUSZTIG THEORY. A deeper understanding of Theorem 4.5 can be gained starting from the following surprising fact (which is verified in a case-by-case analysis, see [2]; for a general argument in the case $\lambda = 1$ see [9]):

PROPOSITION 4.7. *For any Φ-cuspidal pair $(\mathbb{L}, \lambda)$ of $\mathbb{G}$ the relative Weyl group $W_{\mathbb{G}}(\mathbb{L}, \lambda)$ is a complex reflection group. Moreover, if W acts irreducibly on V then $W_{\mathbb{G}}(\mathbb{L}, \lambda)$ is also irreducible in its natural reflection representation.*

In particular there is a cyclotomic algebra $\mathcal{H}(W_{\mathbb{G}}(\mathbb{L}, \lambda))$ attached to the relative Weyl group. In view of this, Theorem 4.5 can be considered as the shadow of an even more precise result on the decomposition of $R_{\mathbb{L}}^{\mathbb{G}}$ which was verified in [2, 14] for all the cases where the Schur elements are known:

THEOREM 4.8. *For any Φ dividing $|\mathbb{G}|$ and any Φ-cuspidal pair $(\mathbb{L}, \lambda)$ of $\mathbb{G}$ there exists an admissible specialization $f_{(\mathbb{L},\lambda)} : \mathbb{Z}[\mathbf{u}, \mathbf{u}^{-1}] \to \mathbb{C}[x]$ of the cyclotomic Hecke algebra $\mathcal{H}(W_{\mathbb{G}}(\mathbb{L}, \lambda), \mathbf{u})$, such that for all $\chi \in \mathrm{Irr}(W_{\mathbb{G}}(\mathbb{L}, \lambda))$*

$$\deg(I^{\mathbb{G}}_{(\mathbb{L},\lambda)}(\chi)) = \frac{|\mathbb{G}|_{x'}}{|\mathbb{L}|_{x'}} \frac{\deg(\lambda)}{f_{(\mathbb{L},\lambda)}(c_\chi)}\,.$$

The specialization $f_{(\mathbb{L},\lambda)}$ is determined locally, that is, by situations in which $W_{\mathbb{G}}(\mathbb{L}, \lambda)$ is a 1-dimensional reflection group. Furthermore, in the case $\phi = 1$, for $\Phi = x - 1$ and $\lambda = 1$, $f_{(\mathbb{L},\lambda)}$ is the 1-parameter specialization (3.3), which shows that the degrees of the principal series unipotent characters are the generic degrees of $\mathcal{H}(W, x)$.

This is precisely the formula one would obtain if the specialization of the cyclotomic algebra $\mathcal{H}(W_{\mathbb{G}}(\mathbb{L}, \lambda), \mathbf{u})$ of the relative Weyl group were the endomorphism algebra of $R^{\mathbb{G}(q)}_{\mathbb{L}(q)}(\psi^{\mathbb{L}}_q(\lambda))$ (see [2] and also [6] for a conjectural explanation).

4.9. UNIPOTENT DEGREES. Let now $\mathbb{G} = (V, W\phi)$ be an arbitrary spets, with spetsial reflection group W over k. Assume given for any Levi subspets $\mathbb{L} = (V, W_{\mathbb{L}}w\phi)$ of $\mathbb{G}$ a set $\mathcal{E}(\mathbb{L})$ with an action of $N_W(W_{\mathbb{L}})$, a degree map $\deg : \mathcal{E}(\mathbb{L}) \to k[x]$, and for any Φ and any pair $\mathbb{L} \leq \mathbb{M}$ of Φ-split Levi subspetses two homomorphisms

$$R^{\mathbb{M}}_{\mathbb{L}} : \mathbb{Z}\mathcal{E}(\mathbb{L}) \to \mathbb{Z}\mathcal{E}(\mathbb{M}), \quad {}^*R^{\mathbb{M}}_{\mathbb{L}} : \mathbb{Z}\mathcal{E}(\mathbb{M}) \to \mathbb{Z}\mathcal{E}(\mathbb{L}),$$

adjoint to each other with respect to the scalar products for which $\mathcal{E}(\mathbb{L})$, $\mathcal{E}(\mathbb{M})$ are orthonormal. If these data satisfy the analogues of Theorems 4.5 and 4.8 (and hence also of Theorem 4.3) then we say that $\mathbb{G}$ *has unipotent degrees* attached to it. Thus, by the above, spetses for Weyl groups have unipotent degrees. Amazingly, this property is shared by all spetsial reflection groups, even the non-real ones:

THEOREM 4.10. *Spetses have unipotent degrees.*

For the explicit construction of the sets $\mathcal{E}(\mathbb{G})$ see [11] for real reflection groups, [14] for the infinite series (where also the properties in Theorem 4.5 and 4.8 are verified for the infinite series of Weyl groups) and [4] for the exceptional spetses. Note that non spetsial reflection data cannot have unipotent degrees by Proposition 3.10.

5. Fusion rules

5.1. Fourier transforms. Let $\mathbb{G} = (V, W\phi)$ be a spets and $\mathcal{E}(\mathbb{G})$ be the associated set of unipotent degrees. By Proposition 3.10 the generic degrees lie in the space spanned by the fake degrees $\{R_\chi\}$. But more is true: all unipotent degrees lie in this space. This gives rise to further fascinating properties of unipotent degrees. To explain these, let us return to the case of rational spetses $\mathbb{G}$ originating from families of groups of Lie type, as in Section 2.3. For simplicity of exposition, let us also assume that $\phi = 1$.

By the fundamental results of Lusztig, the subspace of the space of class functions spanned by the unipotent characters coincides with the space spanned by the so-called unipotent almost characters, and both sets form orthonormal bases of this subspace. Let S denote the corresponding base change matrix, the *Fourier transform matrix*. Then S is unitary and of order 2. Moreover, S transforms the vector of unipotent degrees of $\mathbb{G}$ into the vector of degrees of almost characters, that is, a vector consisting of the fake degrees of W, extended by a suitable number of zeros. Furthermore, Lusztig associates to each unipotent character a root of unity, the so-called eigenvalue of Frobenius. Let F denote the diagonal matrix formed by these Frobenius eigenvalues. Then $(FS)^3 = 1$, hence F and S give rise to a representation of the modular group $\mathrm{PSL}_2(\mathbb{Z})$.

A similar situation occurs in the case of arbitrary spetses: There exists a matrix S over k which is symmetric, unitary, and which transforms the unipotent degrees into the fake degrees of W. Moreover, to each $\gamma \in \mathcal{E}(\mathbb{G})$ can be attached an eigenvalue of Frobenius (a root of unity), such that S together with the diagonal matrix F formed by the Frobenius eigenvalues satisfy:

$$S^4 = 1, \quad (F\,S)^3 = 1, \quad [F, S^2] = 1,$$

hence S, F give rise to a representation of $\mathrm{SL}_2(\mathbb{Z})$.

5.2. Families. In the case of spetses coming from groups of Lie type, the Fourier matrix has a block diagonal shape, with blocks given by the cells, or families, of the Weyl group. In the general case, a similar statement holds. There exists a partition of $\mathcal{E}(\mathbb{G})$ into *families*, such that the a-function is constant on families, each family contains a unique special character, and S becomes block diagonal with respect to this partition. Via the embedding in Theorem 4.5 this induces a subdivision of $\mathrm{Irr}(W)$ into families, which is not yet well understood, since the concept of cells does not seem to generalize easily to complex reflection groups.

In the case of Coxeter groups W, two characters $\chi, \chi' \in \mathrm{Irr}(W)$ lie in the same family if and only if the corresponding characters of the 1-parameter Hecke algebra $\mathcal{H}(W, x)_\mathcal{O}$ lie in the same block, with $\mathcal{O} := A[(1 + x\mathbb{Z}[x])^{-1}]$ (as was pointed out by Raphaël Rouquier; see also [8]). We expect the same statement to be true in the case of arbitrary spetsial reflection groups. It seems interesting to study the projective characters of these blocks.

Let $\mathcal{F} \subseteq \mathcal{E}(\mathbb{G})$ be a family with Fourier matrix $S_\mathcal{F} = (s_{jk})_{j,k\in\mathcal{F}}$. We define structure constants

$$n_{jk}^l := \sum_{m\in\mathcal{F}} \frac{s_{jm}s_{km}\overline{s_{lm}}}{s_{j_0m}} \qquad \text{for } j, k, l \in \mathcal{F}, \tag{5.3}$$

where $j_0 \in \mathcal{F}$ is the special character (by the above, $s_{j_0 m} \neq 0$ for all $m \in \mathcal{F}$). In the case of finite Coxeter groups, all structure constants are non-negative integers (see [10, 12, 13]). In the complex case, the following slightly weaker result holds:

THEOREM 5.4. *The structure constants of families of spetses are rational integers.*

The proof is case by case and will be published elsewhere. This result shows that (apart from the missing positivity) the data $S_{\mathcal{F}}, F_{\mathcal{F}}$ of a family satisfy the axioms of a fusion rule (see e.g. [7]). In particular, for any family $\mathcal{F}$ of a spets $\mathbb{G}$ we obtain an associative, commutative ring $B(\mathcal{F})$ with 1, free and indecomposable over $\mathbb{Z}$ with basis indexed by $\mathcal{F}$ and structure constants given by (5.3).

REFERENCES

[1] S. Ariki, *Representation theory of a Hecke algebra of $G(r, p, n)$*, J. Algebra **177** (1995), 164–185.

[2] M. Broué and G. Malle, *Zyklotomische Heckealgebren*, Astérisque **212** (1993), 119–189.

[3] M. Broué, G. Malle, J. Michel, *Generic blocks of finite reductive groups*, Astérisque **212** (1993), 7–92.

[4] M. Broué, G. Malle, J. Michel, *Towards spetses*, in preparation (1998).

[5] M. Broué, G. Malle, R. Rouquier, *Complex reflection groups, braid groups, Hecke algebras*, to appear, J. reine angew. Math. (1998).

[6] M. Broué and J. Michel, *Sur certains éléments réguliers des groupes de Weyl et les variétés de Deligne-Lusztig associées*, Progress in Mathematics (M. Cabanes, ed.), vol. 141, Birkhäuser, 1997, pp. 73–140.

[7] J. Fuchs, *Fusion rules in conformal field theory*, Fortschr. Phys. **42** (1994), 1–48.

[8] A. Gyoja, *Cells and modular representations of Hecke algebras*, Osaka J. Math. **33** (1996), 307–341.

[9] G.I. Lehrer and T.A. Springer, *Intersection multiplicities and reflection subquotients of unitary reflection groups, I*, preprint (1997).

[10] G. Lusztig, *Leading coefficients of character values of Hecke algebras*, Proc. Symp. in Pure Math. **47** (1987), 235–262.

[11] G. Lusztig, *Coxeter groups and unipotent representations*, Astérisque **212** (1993), 191–203.

[12] G. Lusztig, *Exotic Fourier transform*, Duke J. Math. **73** (1994), 243–248.

[13] G. Malle, *Appendix: An exotic Fourier transform for H_4*, Duke J. Math. **73** (1994), 243–248.

[14] G. Malle, *Unipotente Grade imprimitiver komplexer Spiegelungsgruppen*, J. Algebra **177** (1995), 768–826.

[15] G. Malle, *On the rationality and fake degrees of characters of cyclotomic algebras*, submitted (1998).

[16] G. Malle, *On the generic degrees of cyclotomic algebras*, preprint (1998).

[17] G. Malle and A. Mathas, *Symmetric cyclotomic Hecke algebras*, to appear, J. Algebra (1998).

[18] E.M. Opdam, *Complex reflection groups and fake degrees*, preprint (1998).

[19] G.C. Shephard and J.A. Todd, *Finite unitary reflection groups*, Canad. J. Math. **5** (1954), 274–304.

[20] T.A. Springer, *Regular elements of finite reflection groups*, Invent. Math. **25** (1974), 159–198.

Gunter Malle
Im Neuenheimer Feld 368
Universität Heidelberg
D 69120 Heidelberg
malle@urania.iwr.uni-heidelberg.de

Birational Automorphisms of Higher-Dimensional Algebraic Varieties

Aleksandr V. Pukhlikov

Abstract. The present survey covers the known results on the groups of birational automorphisms, rationality problem and birational classification for Fano fibrations.

1991 Mathematics Subject Classification: 14E05, 14E07, 14J45
Keywords and Phrases: Fano fibration, birational automorphism, maximal singularity, untwisting, birational rigidity

0. Birational geometry starts with M.Nöther's paper [45] on Cremona transformations. The problems of birational geometry of algebraic varieties, that is, birational classification, the rationality and unirationality problems, structure of the group of birational automorphisms, formed a subject of exclusive attention for the Italian classics, including C.Segre, Castelnuovo, Enriques, Comessatti, B.Segre, Fano, Morin, Predonzan and many others, see, for instance, [59]. Italian geometers, first of all — G.Fano, laid the foundation of the modern birational geometry, outlined solutions to certain hard problems, gave surprisingly exact forecasts and suggested some crucial ideas.

The modern period of birational geometry started with Yu.I.Manin's papers on geometry of surfaces over non-closed fields [38,39]. The breakthrough into higher dimensions was made in 1970 in the papers of V.A.Iskovskikh and Yu.I.Manin [29] and H.Clemens and Ph.Griffiths [7], where the Lüroth problem got its negative solution (both the techniques and the final results of these papers were absolutely independent of each other). In [29] Iskovskikh and Manin, using certain classical ideas of Nöther and Fano, developed a new method of study of birational correspondences between algebraic varieties (which have no nontrivial differential-geometric birational invariants) — the *method of maximal singularities*. The results which were obtained by means of this method in 70s were summed up 15 years ago in [24,25]. Since that day, a considerable progress has been made in the field. It is worth noting that, although we have now new approaches and concepts [34,35], this method up to this day is the most effective tool in birational geometry. The contemporary state of the theory form the subject of the present survey.

1. The aim of birational geometry is birational classification of algebraic varieties. In the most general sense, for two given varieties V, V' we should be able to say, whether their function fields $k(V)$ and $k(V')$ are isomorphic, and if yes, how such

an isomorphism can be obtained (here k is an algebraically closed field of characteristic zero; the principal case is $k = \mathbf{C}$). We may understand the classification problem also as the problem of investigating birational geometry of the given variety V, that is, those geometric properties which are independent of the concrete model of the field $k(V)$.

Birational geometry is most interesting, rich and also hard to study for *Fano fibrations* $\pi: V \to S$, the generic fiber F_η of which is a Fano variety over the non-closed field $k(S)$, that is, the canonical class K_{F_η} is negative. In other words, the fiber F_t over a point $t \in S$ of general position is a Fano variety over the field k. For this class of objects we can specify the following particular cases in the general problem of birational classification.

(1) Describe all the structures of a Fano fibration on the given variety V (in the *birational* sense). In many cases this problem can be transformed into the following question: is a birational map $\chi: V - - \to V'$ between two given Fano fibrations $\pi: V \to S$ and $\pi': V' \to S'$ fiber-wise?

(2) Compute the group of birational automorphisms $\operatorname{Bir} V = \operatorname{Aut} k(V)$. If $V = \mathbf{P}^m$, we get the *m-dimensional Cremona group*.

(3) The *rationality problem*: whether V is birational to $\mathbf{P}^m$ (in algebraic terms: whether $k(V)$ is a purely transcendent extension $k(t_1, \dots, t_m)$ of the field of constants)?

The problem of birational classification naturally generalizes to *rational correspondences*: for two given algebraic varieties V, V' describe the set of rational (p, q)-correspondences between them, $p, q \geq 1$. For $V' = \mathbf{P}^m, p = 1$ we get the classical *unirationality problem* (whether the field $k(V)$ can be embedded in $k(t_1, \dots, t_m)$?). Unfortunately, today we have got no methods, which could make it possible to study the subject, only direct constructions of unirationality of the type of B.Segre [64], U.Morin [44] and A.Predonzan [46], see also the modern papers [9,41].

2. Fano fibrations satisfy the classical *termination condition for canonical adjunction*, the importance of which was understood by the Italian classics: for any divisor D the linear system $|D + nK_V|$ is empty for n sufficiently high. The *threshold of canonical adjunction*

$$c(V, D) = \sup\{ \frac{b}{a} \;\Big|\; a, b \in \mathbf{Z}_+ \backslash \{0\}, |aD + bK_V| \neq \emptyset \}.$$

is a quantitative characteristic of termination. To study a birational map $\chi: V - - \to V'$, we compare the corresponding thresholds on V and V': let $|D'|$ be a linear system of divisors on V', free in codimension 1, and $|D| = |D(\chi)| = (\chi^{-1})_* |D'|$ be its proper inverse image on V, then we get two numbers $c(V, D)$ and $c(V', D')$. In a certain natural sense the threshold $c(V, D)$ characterizes the "complexity" or "size" of the linear system $|D|$. Decreasing the threshold by means of an "elementary" birational map $\tau: V_1 - - \to V$, where V_1 is, generally speaking, another model of the field $k(V)$, we "simplify" the system $|D|$ and thus the map χ itself: $c(V_1, D(\chi \circ \tau)) < c(V, D(\chi))$. This is the general idea of simplification (in the traditional terminology, *untwisting*) of a birational map.

Definition 1. (i) A Fano fibration $\pi\colon V \to S$ is said to be *birationally rigid*, if for any V', D', χ' there exists a birational automorphism of the generic fiber $\chi^* \in \operatorname{Bir} F_\eta \subset \operatorname{Bir} V$ such that the composition $\chi \circ \chi^*\colon V - - \to V'$ satisfies the monotonicity condition: $c(V, D) \leq c(V', D')$.

(ii) A Fano fibration is said to be *birationally superrigid*, if the monotonicity condition is always true (i.e. we can take $\chi^* = \mathrm{id}$.)

The property of being (super)rigid characterizes birational geometry of a variety in an exhaustive way.

Proposition 1. *Assume that the Fano fibration $\pi\colon V \to S$ satisfies the following condition: for any divisor D and the induced divisor D_η on the generic fiber F_η the thresholds $c(V, D)$ and $c(F_\eta, D_\eta)$ coincide, and, moreover, that* $\operatorname{Pic} F_\eta \cong \mathbf{Z}$. *Assume the Fano fibration $\pi\colon V \to S$ to be birationally rigid. Then any birational map $\chi\colon V - - \to V'$, where $\pi'\colon V' \to S'$ is a Fano fibration of the same dimension, is fiber-wise, that is, $\pi' \circ \chi = \alpha \circ \pi$ for some (dominant rational) map of the base $\alpha\colon S - - \to S'$.*

Corollary 1. *In the assumptions of Proposition 1 the variety V is non-rational. Any birational automorphism $\chi \in \operatorname{Bir} V$ is fiber-wise.*

Corollary 2. *Birationally rigid Fano variety V with* $\operatorname{Pic} V \cong \mathbf{Z}$ *cannot be fibered (by a rational map) into rationally connected varieties over a positive-dimensional base.*

Corollary 3. *For a birationally superrigid smooth Fano variety V with* $\operatorname{Pic} V \cong \mathbf{Z}$ *the groups of birational and biregular automorphisms coincide,* $\operatorname{Bir} V = \operatorname{Aut} V$.

3. Fix a smooth (or with $\mathbf{Q}$-factorial terminal singularities) Fano fibration $\pi\colon V \to S$. Let $\chi\colon V - - \to V'$ be a birational map onto another Fano fibration. Assume that the monotonicity condition does not hold: $n = c(V, D) > c(V', D')$.

Proposition-Definition 2. *There exists a geometric (that is, realizable by a prime Weil divisor on a certain projective model of the function field) discrete valuation $\nu\colon k(V) \to \mathbf{Z}$, which satisfies the* Nöther-Fano(-Manin-Iskovskikh) inequality

$$\nu(|D|) > na(\nu, V),$$

where $a(\cdot)$ is the discrepancy. These valuations are called maximal singularities *of the map χ or the system $|D|$. If ν is of the form $\nu_B = \operatorname{mult}_B$, where $B \subset V$ is an irreducible cycle of codimension≥ 2, then B is said to be a* maximal cycle. *Otherwise, ν is said to be* infinitely near.

The general scheme of arguments which prove (super)rigidity looks as follows. It turns out that (in all the cases that can be succefully studied by this method) the maximal singularities are an exceptional phenomenon. Only very special cycles $B \subset V$ can appear as maximal (in many cases they do not occur at all), and if there is no maximal cycle, there is no infinitely near maximal singularities, either. Exclusion of the infinitely near case is based upon the following key

Proposition 3. *Let $D_{1,2} \in |D|$ be general divisors and the centre B of ν on V be of codimension≥ 3, and assume that B is not contained in the singular locus of V. Let $Z = (D_1 \bullet D_2)$ be the algebraic cycle of their scheme-theoretic*

intersection (it is an effective cycle of codimension 2). Then

$$\operatorname{mult}_B Z \geq 4\frac{\nu(|D|)^2}{a(\nu,V)^2} > 4n^2.$$

This very inequality makes the essence of the *test class* method of Iskovskikh-Manin, which was developed in [29]. Gradually [48-55] it was discovered that this fact has a very general character.

After all the potentially maximal cycles B have been detected, for each of them one constructs an "untwisting" automorphism $\tau_B \in \operatorname{Bir} V$. Taking the composition $\chi \circ \tau_B$, we simplify the map, that is, decrease the adjunction threshold $n(\chi) = c(V, D(\chi))$. After a finite number of steps the composition $\chi \circ \tau_{B_1} \circ \ldots \tau_{B_N}$ satisfies the monotonicity condition. Simultaneously we get a copresentation of the group $\operatorname{Bir} V$ (generators and relations). If maximal singularities do not occur at all, we conclude that V is birationally superrigid.

4. Here is the list of Fano varieties, birational geometry of which has been successfully studied by means of the method of maximal singularities.

1) Double spaces $V \to \mathbf{P}^M$ of index 1, branched over a hypersurface $W_{2M} \subset \mathbf{P}^M$ of degree $2M$. The hypersurface can contain a singular point $x \in W$ of general position of multiplicity $2m$, $m \leq M-2$, $M \geq 3$.

2) Double quadrics $V \to Q \subset \mathbf{P}^{M+1}$ of index 1, branched over a divisor, which is cut out on Q by a hypersurface $W_{2(M-1)}$ of degree $2(M-1)$. In dimensions ≥ 4 the branch divisor is smooth, in dimension 3 it may contain a non-degenerate double point.

3) Hypersurfaces $V = V_M \subset \mathbf{P}^M$ of degree M, $M \geq 4$. For $M = 4$ V is either smooth or is allowed to have exactly one non-degenerate double point $x \in V$, lying on exactly 24 distinct lines on V. For $M = 5$ V is arbitrary smooth, for $M \geq 6$ V is general in the following sense: for any point $x \in V$ and any system $(z_1, \ldots, z_M)$ of affine coordinates on $\mathbf{P}^M$ with the origin at x the sequence of polynomials $(q_1, \ldots, q_{M-1})$ makes a regular sequence, where $f = q_1 + \ldots + q_{M-1} + q_M$ is the equation of V with respect to z_*, q_i are homogeneous degree i.

4) Double Veronese cone $V \to W \subset \mathbf{P}^6$ of dimension three, that is, W is the cone over the Veronese surface in $\mathbf{P}^5$, and the non-singular branch divisor is cut out by a cubic, not passing through the vertex.

5) General complete intersections $V_{2\cdot 3} = Q_2 \cap Q_3 \subset \mathbf{P}^5$ (the normal bundle of any line $L \subset V$ is $\mathcal{N}_{L/V} \cong \mathcal{O}_L \oplus \mathcal{O}_L(-1)$ and there is no plane $P \subset \mathbf{P}^5$ such that $P \cap V$ consists of three lines passing through a point).

6) General (in particular, quasismooth) hypersurfaces in the weighted projective space $V_d \subset \mathbf{P}(1, a_1, a_2, a_3, a_4)$, $d = a_1 + \ldots + a_4$, which are $\mathbf{Q}$-factorial Fano threefolds with terminal singularities. There are 95 families of these varieties [16], starting from $V_4 \subset \mathbf{P}^4$ and ending by $V_{66} \subset \mathbf{P}(1, 5, 6, 22, 33)$. From this list we should exclude the quartic V_4 (which is already present in 3)) and $V_6 \subset \mathbf{P}(1, 1, 1, 1, 3)$, which is just the double space (class 1)).

THEOREM 1. *A) The following Fano varieties are birationally superrigid:*
— all the members of the class 1) ([24] for smooth 3-folds, [49] for smooth double spaces of dimension ≥ 4, [53] for the singular case)*;*

— *all the members of 2) of dimension* ≥ 4 [49]*;*
— *all the smooth members of 3)* ([29] for the quartic, [48] for the quintic, [55] for the rest of the cases) *and 4)* [24,33].
Their groups of birational and biregular automorphisms coincide. For a general member of the class 3) it is trivial, of the classes 1), 2) and 4) it is $\mathbf{Z}/2\mathbf{Z}$.

B) All the rest of Fano varieties from the list above are birationally rigid. For each of them there is the exact sequence

$$1 \to B(V) \to \operatorname{Bir} V \to \operatorname{Aut} V \to 1,$$

where $B(V)$ *is the untwisting subgroup, that is,* χ^* *from Definition 1 can be taken from this subgroup. More exactly:*
— *for three-dimensional double quadrics (class 2),* $M = 3$*)* $B(V)$ *is the free product of the one-dimensional family of involutions* τ_L*, associated with the lines* $L \subset V$ *(i.e., irreducible rational curves with* $(L \cdot K_V) = -1$*), which do not lie in the branch divisor* ([24] for the smooth and [18] for the singular case)*;*
— *for the singular quartics* $B(V)$ *is the free product of 25 involutions* $\tau_i, i = 0, \ldots, 24$*, where* τ_0 *is the reflection from the double point* x *and* τ_i *is the reflection from* x *in the fibers of the elliptic fibration, generated by the net of planes containing* $L_i \ni x$ ([50])*;*
— *for* $V = V_{2\cdot 3}$ *the subgroup* $B(V)$ *is an "almost free" product of two one-dimensional families of involutions* α_L*, for all the lines* $L \subset V$*, and* β_Y*, for all the irreducible conics* $Y \subset V$ *such that the plane* $P(Y) \supset Y$ *lies in* Q_2*. There is a finite number of relations* $(\alpha_{L_1}\alpha_{L_2}\alpha_{L_3})^2 = 1$*, where the lines* L_i *lie in the same plane* (the proof was started in [24] and completed in [51], for a complete exposition see [31])*;*
— *for the weighted hypersurfaces (class 6))* $B(V)$ *is generated by a finite number of involutions, associated with the terminal singular points* [12]*. For some of these varieties there are no birational involutions at all, so that they are actually superrigid.*

5. Here is the list of Fano fibrations over a non-trivial base, birational geometry of which has been succefully studied by the method of maximal singularities.

1) Standard conic bundles $\pi\colon V \to S$, $\dim V \geq 3$, with a big discriminant divisor $D \subset S$: $|D + 4K_S| \neq \emptyset$.

2) Smooth threefolds $\pi\colon V \to \mathbf{P}^1$, fibered into del Pezzo surfaces, $\operatorname{Pic} V = \mathbf{Z}K_V \oplus \mathbf{Z}F$, where F is the class of a fiber, F_η is a del Pezzo surface of degree $d = 1, 2, 3$ over the non-closed field $k(t)$, satisfying the K^2*-condition*: the numerical class $MK_V^2 - f$ in not effective for any $M \in \mathbf{Z}$, $f \in A^2(V)$ is the class of a line in a fiber (for $d = 3$ it is also assumed that if F_t is a singular fiber, then it has exactly one singular point lying on exactly six lines on F_t).

3) General smooth 4-folds $\pi\colon V \to \mathbf{P}^1$, fibered into quartic threefolds, satisfying the K^2-condition: $MK_V^2 - f$ is not effective, where f is the class of a hyperplane section of a fiber, $M \in \mathbf{Z}$.

4) Smooth higher-dimensional varieties $\pi\colon V \to \mathbf{P}^1$, fibered into double spaces of index 1 (class 1) from Sec. 4 above), satisfying the K^2-condition.

5) Certain varieties with a pencil of double quadrics of index 1 (class 2) from Sec. 4), satisfying the K^2-condition.

6) The general double cone $V \to Q_2 \subset \mathbf{P}^4$, where Q_2 is the non-degenerate quadric cone and the branch divisor is cut out by a quartic, which does not pass through the vertex. The variety V has two obvious pencils of del Pezzo surfaces of degree 2, induced by the pencils of planes on Q_2.

THEOREM 2. *A) Any birational map of a variety from the class 1) above onto another conic bundle is fiber-wise* [61,62].

B) Fano fibrations from the class 2) for $d = 1$ and from the classes 3)-5) are superrigid. For a general variety from the class 3) $\operatorname{Bir} V$ *is a trivial group, otherwise (for a general variety) it is isomorphic to* $\mathbf{Z}/2\mathbf{Z}$ [54,56].

C) Fano fibrations from the class 2) for $d = 2, 3$ are birationally rigid. The following exact sequence holds

$$1 \to \operatorname{Bir} F_\eta \to \operatorname{Bir} V \to G \to 1,$$

where G is a finite, generically trivial group of fiber-wise birational automorphisms. (See [54]. The group $\operatorname{Bir} F_\eta$ was described by Yu.I.Manin [38-40]. It is generated by involutions, associated to sections of π for $d = 2$, and for $d = 3$ — to sections and bisections of π.)

D) Varieties of the type 6) are birationally rigid as Fano varieties. For any pencil $|\Lambda|$ of rational surfaces on V there is a birational automorphism χ^*, which transforms $|\Lambda|$ into one of the two "default" pencils. The group $\operatorname{Bir} V$ *is generated by the subgroups* $\operatorname{Bir} F_{\eta_i}$, $i = 1, 2$ *(F_{η_i} are the generic fibers of the "default" pencils). Their intersection* $\operatorname{Bir} F_{\eta_1} \cap \operatorname{Bir} F_{\eta_2}$ *is generated by a finite number of involutions* [19].

6. Conjectures and open problems.

1) Let $V_{m_1 \cdot \ldots \cdot m_k} \subset \mathbf{P}^{m_1+\ldots+m_k}$ be a Fano complete intersection of index 1 with sufficiently mild singularities. Then V is birationally (super)rigid.

2) By analogy with the weighted 3-fold hypersurfaces, we should expect that 1) is true for the weighted case, either.

3) The rigidity facts about Fano fibrations can be looked at as a realization of the following informal principle:

> if a Fano fibration is "sufficiently twisted" over the base, then birational geometry of V reduces to birational geometry of the generic fiber F_η.

It seems that this principle holds in a much more general situation than A)-C) of Theorem 2. For instance, if $V \hookrightarrow \mathbf{P}(\mathcal{E})$, where $\mathcal{E}$ is a locally free sheaf on S of rank $m_1 + \ldots + m_k + 1$ and the generic fiber F_η is a complete intersection $V_{m_1 \cdot \ldots \cdot m_k} \subset \mathbf{P}^{m_1+\ldots+m_k}_\eta$, then "sufficient twistedness" over the base implies that the Fano fibration $\pi\colon V \to S$ is birationally (super)rigid. As in 2) above, this statement should be true for the weighted case, too.

4) Hypersurfaces $V_m \subset \mathbf{P}^M$ of index $M + 1 - m \geq 2$ obviously have a lot of structures of a Fano fibration. It seems natural to suggest that all these structures come from the "natural" ones, the fibers of which are Fano complete intersections in linear subspaces of $\mathbf{P}^M$. For instance, linear systems $|\Lambda_i|$, $i = 1, \ldots, k$, cut out on V by hypersurfaces of degrees $m_1, \ldots, m_k$, where $m + m_1 + \ldots + m_k \leq M$, determine a structure of a Fano fibration

$$\pi = (\pi_1, \ldots, \pi_k)\colon V - - \to \mathbf{P}^{n_1} \times \ldots \times \mathbf{P}^{n_k},$$

$n_i = \dim |\Lambda_i|$. Another example: for a quartic $V = V_4 \subset \mathbf{P}^M$ of dimension ≥ 4 we suggest that all the structures of a fibration into rational surfaces come from the linear projections from the planes $P \subset V$ and V can not be fibered into rational curves (by a rational map). The general cubic $V = V_3 \subset \mathbf{P}^M$, $M \geq 5$, is non-rational. The coincidence $\operatorname{Bir} V = \operatorname{Aut} V$ is very likely to be true for all the hypersurfaces of degree 4 and higher, at least for general ones (for ceratin special smooth quartics non-trivial birational automorphisms do exist, but their construction only confirms that they represent an exceptional phenomenon). Similarly for complete intersections.

5) Computation of the Cremona group $\operatorname{Bir} \mathbf{P}^n$, even for $n = 3$, and of the group $\operatorname{Bir} V_3$ for the higher-dimensional cubic still remains an open problem, seeming to be inaccessible for the modern techniques. In [25] a complete description of the group $\operatorname{Bir} V_2$ for the double space $\mathbf{P}^3$ of index 2 (branched over a quartic) was announced. Unfortunately, it also remains an open problem (although the fact itself seems to be true).

6) We have got no rationality criterion for threefolds. The crucial problem here is to prove the well-known (conjectural) Iskovskikh-Shokurov rationality criterion for conic bundles, see [28].

7) It is important to study the structure of infinitely near maximal singularities. There is a conjecture that if a linear system $|D|$ has a maximal singularity ν, it also has another maximal singularity μ (satisfying the same Nöther-Fano inequality), which can be realized as a weighted blow up. In dimension three this conjecture describes all the extremal contractions to smooth points.

8) Up to this day we are unable to prove non-unirationality otherwise but by producing differential forms. On the other hand, the general quartics in $\mathbf{P}^4$, speaking not of double spaces or general hypersurfaces and complete intersections of a small index and high dimension, seem to be non-unirational. Recently some new direct constructions of unirationality appeared [9,41].

7. The prospects of birational classification.

The well-known achievements of the minimal model program (or Mori theory) [8,32,36,47,42,43,57,65,66,70] generated some hope to convert the three-dimensional birational geometry from a collection of separate results and constructions into a regular theory. The corresponding concept of factorization of birational maps between (three-dimensional) Mori fiber spaces was developed by Sarkisov [63] and got the name Sarkisov program [58]. It was exhaustively substantiated by A.Corti [10], see also [11]. After it had been proved that any birational map between Mori fiber spaces can be factorized into a chain of elementary links, it was natural to apply this general theory to certain families of three-folds, in order to re-think on a higher level the classical results of the method of maximal singularities. As an object for this experiment the above-mentioned 95 families [16] of weighted hypersurfaces were chosen [12]. However, the results turned out to be rather unexpected: *all* the discovered elementary links were just *involutions* of the classical type, which (so far) permits no explanation from the Mori-theoretic viewpoint. On the other hand, now we have got Mori-theoretic analogs of the crucial technical means of the classical method (A.Corti's techniques of "reduction to log canonical surfaces", which in dimension three can replace the old techniques

of counting multiplicities, although the latter still seems more transparent and natural). All in all, the result of [12] turned out to be much more in the spirit of the method of maximal singularities than the modern concepts.

Sarkisov program is different from the classical approach by its essentially "dynamical" viewpoint: simplifying (untwisting) a birational map, we replace the initial model by a new one, whereas the traditional approach makes use of automorphisms (the model is always the same). For the weighted hypersurfaces the dynamical viewpoint turned out to be useless. Of course, it goes without saying that in the general case (for instance, for the projective space $\mathbf{P}^3$) it is impossible to reduce all the ampleness of birational geometry to a single model. This can be seen even in the two-dimensional case. However, in spite of all the perfection of two-dimensional birational geometry, which can be looked at as an ideal object of realization of Sarkisov program, there is still a feeling of dissatisfaction. For instance, the modern proof of the Nöther theorem on Cremona transformations *formally* makes use of *all* the minimal rational surfaces, whereas essentially only three models are of real use: the very $\mathbf{P}^2$, $\mathbf{F}_1$ and $\mathbf{P}^1 \times \mathbf{P}^1$. This example and all the higher dimensional ones suggest that the modern concept of a minimal model is *too fine* for the rough purposes of birational classification. Sometimes (and even in the "majority" of cases) the minimal model is unique (rigidity phenomenon). But then we have no need in the dynamical viewpoint! In other cases we need some new, essentially more rough approach to the problem of choice of a suitable model for a given field of rational functions.

References

1. Alekseev V.A., Rationality conditions for three-folds with a pencil of Del Pezzo surfaces of degree 4. Mat. Zametki. 41, 5, 1987, 724-730.
2. Artin M. and Mumford D., Some elementary examples of unirational varieties which are not rational. Proc. London Math. Soc. 25, 1, 1972, 75-95.
3. Bardelli F., Polarized mixed Hodge structures: On irrationality of threefolds via degeneration. Ann. Mat. Pura et Appl. 137, 1984, 287-369.
4. Batyrev V.V., The cone of effective divisors on three-folds. Cont. Math. 131, 3, 1992, 337-352.
5. Beauville A., Variétés de Prym et Jacobiennes intermédiares. Ann. scient. Éc. Norm. Sup. 10, 1977, 309-391.
6. Birational Geometry of Algebraic Varieties. Open Problems. Katata, Japan, 1988.
7. Clemens H. and Griffiths Ph.A., The intermediate Jacobian of the cubic threefold. Ann. Math. 95, 2, 1972, 281-356.
8. Clemens H., Kollár J. and Mori S. Higher dimensional complex geometry. Astérisque 166. 1988.
9. Conte A. and Murre J.P., On a theorem of Morin on the unirationality of the quartic fivefold, to appear in: Atti Acc. Sci. Torino.
10. Corti A., Factoring birational maps of threefolds after Sarkisov. J. Alg. Geom. 4, 1995, 223-254.
11. Corti A., A survey of 3-fold birational geometry, to appear in: Proc. Symp. in Alg. Geometry.
12. Corti A., Pukhlikov A.V. and Reid M., Birational rigidity of 3-fold Fano

weighted hypersurfaces, to appear in: Proc. Symp. in Alg. Geom.
13. Fano G., Sopra alcune varieta algebriche a tre dimensione aventi tutti i generi nulli. Atti Acc. Torino. 43, 1908, 973-977.
14. Fano G. Osservazioni sopra alcune varieta non razionali aventi tutti i generi nulli. Atti Acc. Torino. 50, 1915, 1067-1072.
15. Fano G. Nuove ricerche sulle varieta algebriche a tre dimensioni a curve-sezioni canoniche. Comm. Rend. Acc. Sci. 11, 1947, 635-720.
16. Fletcher A., Working with weighted complete intersections, to appear in: Proc. Symp. in Alg. Geom.
17. Gizatullin M. Kh. Defining relations for the Cremona group of the plane. Math. USSR Izv. 21, 1983, 211-268.
18. Grinenko M.M., Birational automorphisms of the double quadric with an elementary singularity. Mat. Sbornik. 189, 1, 1998, 101-118.
19. Grinenko M.M., Birational automorphisms of the double cone, to appear in Mat. Sbornik.
20. Hironaka H. Resolution of singularities of an algebraic variety over a field of characteristic zero. I, II. Ann. Math. 79, 1964, 109-326.
21. Hudson H.P. Cremona transformations in plane and space. Cambridge: Cambridge University Press, 1927. 454 p.
22. Iskovskikh V.A., Rational surfaces with a pencil of rational curves. Math. USSR Sb. 74, 4, 1967, 133-163.
23. Iskovskikh V.A., Rational surfaces with a pencil of rational curves and a positive square of canonical class. Mat.Sbornik. 83, 1, 1970, 90-119.
24. Iskovskikh V.A., Birational automorphisms of three-dimensional algebraic varieties.J. Soviet Math. 13, 1980, 815-868.
25. Iskovskikh V.A., Algebraic threefolds with a special regard to the problem of rationality. Proc. Int. Cong. Math.(Warsawa,1983), PWN, Warsawa, 1984, 733-746.
26. Iskovskikh V.A., On the rationality problem for algebraic three-folds fibered into Del Pezzo surfaces. Proc. Steklov Institute. 208, 1995, 128-138.
27. Iskovskikh V.A., Factoring birational maps of rational surfaces from the point of view of Mori theory. Russian Math. Surveys 51, 4, 1996, 3-72.
28. Iskovskikh V.A. On the rationality problem for algebraic threefolds. Proc. Steklov Inst.218, 1997, 190-232.
29. Iskovskikh V.A. and Manin Yu.I., Three-dimensional quartics and counterexamples to the Lüroth problem.Math. USSR Sb. 15, 1, 1971, 141-166.
30. Iskovskikh V.A. and Pukhlikov A.V., Birational automorphisms of Fano varieties. In: Geometry of complex projective varieties. Seminars and Conferences. Cetraro (Italy). Mediterranean Press. 9,, 1990, 191-202.
31. Iskovskikh V.A. and Pukhlikov A.V., Birational automorphisms of multidimensional algebraic varieties. J. Math. Sci. 82, 4, 1996, 3528-3613.
32. Kawamata Y., Matsuda K., Matsuki K. Introduction to the minimal model program. Adv. Stud. in Pure Math. 10, 1987, 283-360.
33. Khashin S.I. Birational automorphisms of the double cone of dimension three. Moscow Univ. Math. Bull. 1984, 1, 13-16.
34. Kollár J. Nonrational hypersurfaces. J. Alg. Geom., 1996.

35. Kollár J. Nonrational covers of $CP^n \times CP^n$, to appear in: Proc. Symp. Alg. Geom.
36. Kollár J., Miyaoka Y. and Mori S., Rationally connected varieties. J. Alg. Geom. 1, 1992, 429-448.
37. Kollár J. et al., Flips and abundance for algebraic threefolds. Astérisque 211, 1992.
38. Manin Yu.I., Rational surfaces over perfect fields. Inst. Hautes Etudes Sci. Publ. Math. 30, 1966, 56-97.
39. Manin Yu.I., Rational surfaces over perfect fields II. Mat. Sbornik. 72, 1967, 161-192.
40. Manin Yu.I., Cubic forms: Algebra, geometry, arithmetic. Amsterdam: North Holland, 1986.
41. Marchisio M.R., Some new examples of smooth unirational quartic threefolds. Quaderni del Dipartimento di Matematica dell'Universita' di Torino, 1998.
42. Mori S., Threefolds whose canonical bundles are not numerically effective. Ann. Math. 115, 1982, 133-176.
43. Mori S., Flip theorem and the existence of minimal models for 3-folds. J. Amer. Math. Soc. 1, 1988, 117-253.
44. Morin U., Sull'irrazionalita' dell'ipersuperficie algebrica di qualunque ordine e dimensione sufficientemente alta. Atti del II Congresso dell'UMI, Bologna 1940, 298-302.
45. Nöther M. Über Flächen welche Schaaren rationaler Curven besitzen. Math. Ann. 3, 1871, 161-227.
46. Predonzan A., Sull'unirazionalita' delle varieta' intersezione completa di piu' forme. Rend. Sem. Mat. Padova. 18, 1949, 161-176.
47. Prokhorov Yu.G., On extremal contractions from threefolds to surfaces: the case of one non-Gorenstein point. Cont. Math. 207, 1997, 119-141.
48. Pukhlikov A.V., Birational isomorphisms of four-dimensional quintics. Invent. Math. 87, 1987, 303-329.
49. Pukhlikov A.V., Birational automorphisms of a double space and a double quadric. Math. USSR Izv. 32, 1989, 233-243.
50. Pukhlikov A.V., Birational automorphisms of a three-dimensional quartic with an elementary singularity. Math. USSR Sb. 63, 1989, 457-482.
51. Pukhlikov A.V., Maximal singularities on a Fano variety V_6^3. Moscow Univ. Math. Bull. 44, 1989, 70-75.
52. Pukhlikov A.V., A note on the theorem of V.A.Iskovskikh and Yu.I.Manin on the three-dimensional quartic. Proc. Steklov Inst. 208, 1995, 244-254.
53. Pukhlikov A.V., Birational automorphisms of double spaces with singularities. J. Math. Sci. 85, 4, 1997, 2128-2141.
54. Pukhlikov A.V., Birational automorpisms of three-dimensional algebraic varieties with a pencil of del Pezzo surfaces. Izvestiya, 62, 1, 1998, 123-164.
55. Pukhlikov A.V., Birational automorphisms of Fano hypersurfaces, to appear in: Invent. Math.
56. Pukhlikov A.V., Certain examples of birationally rigid varieties with a pencil of double quadrics. MPI Preprint 1998-15.
57. Reid M., Young person's guide to canonical singularities. Proc. Symp. Pure

Math. 46, 1987, 345-414.
58. Reid M., Birational geometry of 3-folds according to Sarkisov. Univ. Warwick Preprint. 1991.
59. Riposte Armonie. Lettere di Federigo Enriques a Guido Castelnuovo (a cura di U. Bottazzini, A. Conte e P. Gario). Bollati Boringhieri, Torino 1997.
60. Roth L., Algebraic threefolds with special regard to problems of rationality. Berlin-Göttingen-Heidelberg: Springer-Verlag, 1955. 142 p.
61. Sarkisov V.G., Birational automorphisms of conical fibrations. Math. USSR Izv. 17, 1981, 177-202.
62. Sarkisov V.G., On the structure of conic bundles. Math. USSR Izv. 20, 2, 1982, 354-390.
63. Sarkisov V.G., Birational maps of standard **Q**-Fano fiberings. Preprint Kurchatov Inst. Atom. Energy. 1989.
64. Segre B., Variazione continus ed omotopis in geometrie algebrice. Ann. Mat. pura ed appl. Ser. IV, L, 1960, 149-186.
65. Shokurov V.V., 3-fold log flips. Math. USSR Izv. 40, 1993, 95-202, and 41, 1994
66. Shokurov V.V., Semistable 3-fold flips. Russian Izv. Math. 42, 2, 1994, 371-425.
67. Shafarevich I.R., On the Lüroth problem. Proc. Steklov Inst. 183, 1989, 199-204.
68. Tregub S.L., Birational automorphisms of a three-dimensional cubic. Russian Math. Surveys. 39, 1, 1984, 159-160.
69. Tregub S.L., Construction of a birational isomorphism of a three-dimensional cubic and a Fano variety of the first kind with $g = 8$, connected with a rational normal curve of degree 4. Moscow Univ. Math. Bull. 40, 6, 1985, 78-80.
70. Wilson P.M.H., Towards birational classification of algebraic varieties. Bull. London Math. Soc. 19, 1987, 1-48.

Institute for Systems Analysis
Prospekt 60-letya Oktyabrya, 9
Moscow 117312

Chair of Algebra
Department of Mathematics
Moscow State University
Moscow 119899
RUSSIA
dost@dost.mccme.rssi.ru

Tilting Theory and Quasitilted Algebras

Idun Reiten

Introduction

Tilting theory is a central topic in the representation theory of artin algebras, with origins in work of Bernstein–Gelfand–Ponomarev from the early seventies. There has been extensive interaction with various research directions in representation theory, as well as in other branches of algebra. In this paper we survey the development of tilting theory, and in particular we discuss quasitilted algebras, a recent outgrowth of tilting theory.

We consider for simplicity finite dimensional algebras over an algebraically closed field k, and we will often just say that Λ is an algebra. We deal with the category $\operatorname{mod}\Lambda$ of finitely generated Λ-modules. A Λ-module T of projective dimension at most one is a tilting module if $\operatorname{Ext}^1_\Lambda(T,T)=0$ and there is an exact sequence $0\to\Lambda\to T_0\to T_1\to 0$, where T_0 and T_1 are summands of finite direct sums of copies of T. In Section 1 we give some basic properties of such tilting modules. This includes associated torsion pairs together with induced equivalences of subcategories of $\operatorname{mod}\Lambda$ and $\operatorname{mod}\Gamma$ belonging to the torsion pairs, where Γ is the endomorphism algebra $\operatorname{End}_\Lambda(T)^{\mathrm{op}}$ [BB, HRi]. We also discuss predecessors of the theory [BGP, APR].

The material included in Section 1 was developed around 1980. In Section 2 we treat three main lines of further developments. The first two go via a generalization to tilting modules of finite projective dimension [M, H1]. One direction is concerned with the correspondence between tilting modules and a certain type of subcategories of $\operatorname{mod}\Lambda$ [AR]. The second line of development goes via the discovery of the connection with derived categories [H1]. In the third direction a tilting theory with respect to torsion pairs in abelian categories is developed [HRS].

When Λ is hereditary, the algebras $\Gamma=\operatorname{End}_\Lambda(T)^{\mathrm{op}}$, where T is a tilting Λ-module, are by definition the tilted algebras. This is an important class of finite dimensional algebras, since many questions about arbitrary algebras can be reduced to questions on tilted algebras. As a by-product of the general theory of tilting with respect to torsion pairs, the quasitilted algebras are introduced in [HRS], as a generalization of tilted algebras. Central properties of this class of algebras, which also contains the canonical algebras of Ringel [Rin1], are discussed in Section 4.

The quasitilted algebras are defined in terms of tilting objects in hereditary abelian $k-$categories $\mathcal{H}$ with finite dimensional homomorphism and extension spaces over the algebraically closed field k. The last two sections are devoted to

investigating such categories $\mathcal{H}$ with tilting objects, mainly motivated by wanting to obtain information on quasitilted algebras. In Section 5 we deal with the noetherian case. It is proved in [L] that the noetherian $\mathcal{H}$ are exactly $\operatorname{mod} H$ for a finite dimensional hereditary k-algebra H and the categories $\operatorname{coh} \mathcal{X}$ of coherent sheaves on weighted projective lines introduced in [GL]. We also investigate the relationship with the problem of when the Grothendieck group of $\mathcal{H}$ is free abelian of finite rank [RV2].

In Section 6 we deal with the question of what the hereditary abelian categories with tilting object $\mathcal{H}$ look like in general. It is conjectured that $\mathcal{H}$ (connected) must be derived equivalent to one of the categories $\operatorname{mod} H$ or $\operatorname{coh} \mathcal{X}$ above, and we prove this conjecture when there is at least one simple object and also when there is a directing object [HRe3, HRe2]. We end with a discussion of related problems about quasitilted algebras.

Due to limited space, several important results and developments related to tilting theory and quasitilted algebras are not included. For additional references we refer to the bibliography in the cited papers.

I would like to thank Dieter Happel and Sverre O. Smalø for helpful comments.

1 Classical tilting theory and historical predecessors

Let Λ be a finite dimensional algebra. In this section T is a tilting Λ-module of projective dimension at most one. We shall investigate various subcategories associated with T, along with induced equivalences of subcategories of $\operatorname{mod} \Lambda$ and $\operatorname{mod} \operatorname{End}_\Lambda(T)^{\mathrm{op}}$.

The subcategory $\mathcal{T} = \operatorname{Fac} T$ of $\operatorname{mod} \Lambda$ plays an important role in the theory, where the objects of $\operatorname{Fac} T$ are the factors of finite direct sums of copies of T. Under our assumptions, one can show that $\mathcal{T}$ is equal to $\{\, C; \operatorname{Ext}^1_\Lambda(T, C) = 0 \,\}$, and the category $\mathcal{T}$ is a *torsion class* in $\operatorname{mod} \Lambda$, that is, $\mathcal{T}$ is closed under factor modules and extensions. Associated with $\mathcal{T}$ is the *torsion free* class $\mathcal{F} = \{\, C; \operatorname{Hom}_\Lambda(T, C) = 0 \,\}$, and $(\mathcal{T}, \mathcal{F})$ is a *torsion pair* associated with T. Dually, when U is a cotilting module of injective dimension at most one, that is, the dual $D(U)$ of U is a tilting module of projective dimension at most one in $\operatorname{mod} \Lambda^{\mathrm{op}}$, where D denotes the duality $\operatorname{Hom}_k(\ , k)$, there is associated with U the torsion free class $\mathcal{Y} = \operatorname{Sub} U = \{\, C; \operatorname{Ext}^1_\Lambda(C, U) = 0 \,\}$. The objects of $\operatorname{Sub} U$ are submodules of finite direct sums of copies of U. Then there is an associated torsion pair $(\mathcal{X}, \mathcal{Y})$, where $\mathcal{X} = \{\, C; \operatorname{Hom}_\Lambda(C, U) = 0 \,\}$.

A basic feature of tilting theory is the interplay between $\operatorname{mod} \Lambda$ and $\operatorname{mod} \Gamma$, when $\Gamma = \operatorname{End}_\Lambda(T)^{\mathrm{op}}$. When T is a tilting Λ-module, T is also a tilting Γ^{op}-module, and hence $D(T)$ is a cotilting Γ-module. If $(\mathcal{T}, \mathcal{F})$ denotes the torsion pair in $\operatorname{mod} \Lambda$ associated with T and $(\mathcal{X}, \mathcal{Y})$ the torsion pair in $\operatorname{mod} \Gamma$ associated with $U = D(T)$, there are induced equivalences of categories $\operatorname{Hom}_\Lambda(T, \ ) \colon \mathcal{T} \to \mathcal{Y}$ and $\operatorname{Ext}^1_\Lambda(T, \ ) \colon \mathcal{F} \to \mathcal{X}$. This gives the possibility of transforming information between $\operatorname{mod} \Lambda$ and $\operatorname{mod} \Gamma$ in case one of the module categories is better known than the other one. This point of view has been particularly successful when one of the algebras, say Λ, is hereditary, so that Γ is a tilted algebra. Then the torsion theory $(\mathcal{X}, \mathcal{Y})$ *splits*, that is, each indecomposable object in $\operatorname{mod} \Gamma$ is in $\mathcal{X}$ or in $\mathcal{Y}$.

It is an important property of a tilting Λ-module T that $\mathrm{Hom}_\Lambda(T, \ )$ induces an equivalence between $\mathcal{T}$ and $\mathcal{Y}$. If conversely there is an equivalence between subcategories of $\mathrm{mod}\,\Lambda$ and $\mathrm{mod}\,\Gamma$, for two algebras Λ and Γ, one can ask if there is an associated tilting module T such that $\mathrm{Hom}_\Lambda(T, \ )$ (or $\mathrm{Ext}^1_\Lambda(T, \ )$) induces the given equivalence. Actually, the origin of tilting theory comes from [BGP], through the occurrence of an interesting equivalence of subcategories for two module categories, in the setting of representations of quivers. This equivalence was interpreted module theoretically in [APR] as $\mathrm{Hom}_\Lambda(T, \ )$ for a special type of what is now called a tilting module T, and extended to more general settings. Further generalizations were made in [BB, HRi], leading to the foundations of the classical tilting theory, with basic setup as discussed above.

2 Three lines of further developments

We discuss three not entirely independent directions of further developments. The first two go via a generalization of tilting and cotilting modules, dropping the requirements that the projective (or injective) dimension is at most one.

A module T in $\mathrm{mod}\,\Lambda$ for an algebra Λ is a tilting module if $\mathrm{pd}_\Lambda T$ (the projective dimension of T) is finite, $\mathrm{Ext}^i_\Lambda(T,T) = 0$ for $i > 0$, and there is an exact sequence of Λ-modules $0 \to \Lambda \to T_0 \to T_1 \to \cdots \to T_r \to 0$ with each T_i a summand of a finite direct sum of copies of T (see [M, H1]). A cotilting module is defined dually. Let $T = T^{(1)} \oplus \cdots \oplus T^{(m)}$ be a direct sum of indecomposable modules. We say that the tilting module T is basic if the $T^{(i)}$ are pairwise nonisomorphic, and in this case m is the rank of the Grothendieck group $K_0(\mathrm{mod}\,\Lambda)$.

The first direction we deal with is only concerned with modules over Λ. We have already seen that associated with a tilting module T of projective dimension at most one is the subcategory $\mathcal{T} = \{\, C; \mathrm{Ext}^1_\Lambda(T,C) = 0 \,\}$ of $\mathrm{mod}\,\Lambda$, and more generally we associate $\mathcal{T} = \{\, C; \mathrm{Ext}^i_\Lambda(T,C) = 0 \text{ for } i > 0 \,\}$ with an arbitrary tilting module T, and dually $\mathcal{Y} = \{\, B; \mathrm{Ext}^i_\Lambda(B,U) = 0 \text{ for } i > 0 \,\}$ with a cotilting module U. In order to formulate the crucial properties of $\mathcal{T}$ and $\mathcal{Y}$ we recall some important terminology. A full subcategory $\mathcal{C}$ of $\mathrm{mod}\,\Lambda$ is *covariantly finite* in $\mathrm{mod}\,\Lambda$ if for each X in $\mathrm{mod}\,\Lambda$ there is a map $g\colon X \to C$ with C in $\mathcal{C}$ such that for any map $h\colon X \to C'$ with C' in $\mathcal{C}$, there is a map $t\colon C \to C'$ with $tg = h$ [ASm1]. Further, $\mathcal{C}$ is *coresolving* if it is closed under extensions and cokernels of monomorphisms. The notions of *contravariantly finite* and *resolving* subcategories are defined dually. For simplicity we only give the main results for finite global dimension, in which case the notions of tilting and cotilting module coincide [AR]. The case of projective (or injective) dimension at most one was already done in [ASm2].

Theorem 1. *Let Λ be an algebra of finite global dimension, and let T be in* $\mathrm{mod}\,\Lambda$.

(a) The assignment $T \mapsto \mathcal{T} = \{\, C; \mathrm{Ext}^i_\Lambda(T,C) = 0$ for $i > 0\}$ induces a one-one correspondence between basic tilting modules and covariantly finite coresolving subcategories of $\mathrm{mod}\,\Lambda$. *The module T is reconstructed via taking* Ext*-projective objects in $\mathcal{T}$.*

(b) The assignment $U \mapsto \mathcal{Y} = \{ C; \mathrm{Ext}^i(C,U) = 0 \text{ for } i > 0 \}$ *induces a one-one correspondence between basic (co-)tilting modules and contravariantly finite resolving subcategories of* $\mathrm{mod}\,\Lambda$. *The module* T *is reconstructed via taking* Ext*-injective objects in* $\mathcal{Y}$.

The other two directions are concerned with the interplay between Λ and $\Gamma = \mathrm{End}_\Lambda(T)^{\mathrm{op}}$, where T is a tilting module, including induced equivalences between subcategories. A major breakthrough was the discovery of the connection with derived categories [H1]. We cite the following [H1, CPS].

THEOREM 2. *Let* T *be a tilting module over an algebra* Λ. *The derived functor* $R\mathrm{Hom}(T, \ )$ *induced by* $\mathrm{Hom}_\Lambda(T, \ )\colon \mathrm{mod}\,\Lambda \to \mathrm{mod}\,\Gamma$ *gives an equivalence* $D^b(\Lambda) \to D^b(\Gamma)$ *between the bounded derived categories for* Λ *and* Γ *if (and only if)* T *is a tilting module.*

Subsequently, dealing with the category of coherent sheaves on a weighted projective space, a similar result was obtained in [GL, Ba], introducing a notion of tilting sheaf analogous to the notion of tilting module. In this formulation, a previous result from [Be] on establishing a derived equivalence between the category $\mathrm{coh}\,\mathbb{P}^n$ of coherent sheaves on the n-dimensional projective space and some finite dimensional algebras, was incorporated in this setting, the crucial sheaves in [Be] being interpreted as special cases of tilting sheaves.

Through a further generalization of tilting modules to tilting complexes, a Morita theory for derived categories was developed in [Ric] in order to describe exactly when two algebras are derived equivalent.

The third direction has its starting point in the theory of tilting (or cotilting) modules of projective (or injective) dimension at most one, with a strong influence of the associated equivalence of derived categories in this setting [HRS]. The crucial basis for generalization is the torsion pair $(\mathcal{T}, \mathcal{F})$ associated with a tilting module, where it is known that $\mathcal{T}$ contains all injective modules. We consider torsion pairs $(\mathcal{T}, \mathcal{F})$ in $\mathrm{mod}\,\Lambda$ where $\mathcal{T}$ contains all injective modules (equivalently, $\mathcal{T}$ is a cogenerator), but which do not necessarily come from a tilting module. We call them *tilting torsion pairs*. Then we "tilt" with respect to the torsion pair $(\mathcal{T}, \mathcal{F})$ to obtain an abelian category, which is equivalent to $\mathrm{mod}\,\Gamma$ with $\Gamma = \mathrm{End}_\Lambda(T)^{\mathrm{op}}$ when $(\mathcal{T}, \mathcal{F})$ is induced by a tilting module T. The idea is to perform the construction inside the bounded derived category $D^b(\Lambda)$. Let more generally $\mathcal{A}$ be an abelian category with a torsion pair $(\mathcal{T}, \mathcal{F})$. There is an abelian category $\mathcal{B} \subset D^b(\mathcal{A})$ with torsion pair $(\mathcal{F}[1], \mathcal{T})$, and we have the following [HRS].

THEOREM 3. *If* $(\mathcal{T}, \mathcal{F})$ *is a tilting torsion pair (that is,* $\mathcal{T}$ *is a cogenerator for* $\mathcal{A}$*), and either* $\mathcal{A}$ *has enough injectives or* $\mathcal{B}$ *has enough projectives, there is induced a triangle equivalence between* $D^b(\mathcal{A})$ *and* $D^b(\mathcal{B})$.

In order for the new category $\mathcal{B}$ to be equivalent to $\mathrm{mod}\,\Gamma$ for some algebra Γ, we need that the torsion pair is induced by what we call a tilting object T in $\mathcal{A}$, generalizing the notion of tilting module of projective dimension at most one. Motivated by the fact that the endomorphism algebras $\mathrm{End}_\Lambda(T)^{\mathrm{op}}$ play a main role when T is a tilting module over a hereditary algebra Λ we introduce the more

general class of algebras $\mathrm{End}_{\mathcal{H}}(T)^{\mathrm{op}}$, called *quasitilted* algebras, when T is a tilting object in a hereditary abelian k-category with finite dimensional homomorphism and extension spaces [HRS]. Note that $\mathcal{A}$ is said to be hereditary if the Yoneda $\mathrm{Ext}^2(\ ,\ )$ is zero, and in this case T is a tilting object if $\mathrm{Ext}^1_{\mathcal{A}}(T,T) = 0$ and $\mathrm{Hom}_{\mathcal{A}}(T,X) = 0 = \mathrm{Ext}^1_{\mathcal{A}}(T,X)$ implies $X = 0$ [H2].

3 External connections

Tilting theory has played, and continues to play, a central role in the representation theory of algebras. Many questions about arbitrary algebras can be reduced to a problem about tilted algebras, where the theory is much more developed. For example, there is a useful criterion for finite representation type based on a class of tilted algebras. In addition, there are connections and interrelationships with most of the main topics and directions in representation theory. There are connections with relative homological algebra, as the concepts can be formulated in a relative setting [ASo], with stable equivalence [TW], with the generalized Nakayama conjecture and the finitistic dimension conjecture [BS, HU] and with Koszul algebras [GRS]. The connection with derived categories opened up new interesting directions. There are also interrelationships with other parts of algebra, which we discuss in this section.

A characteristic feature of finite dimensional algebras is the wealth of examples of various types available. For example, there are numerous nontrivial examples of derived equivalences, of interest in other areas where such equivalences occur.

The study of many classes of algebras has been motivated by which types of algebras are interesting in other fields. One such example is the *quasihereditary algebras*. Associated with a quasihereditary algebra is a canonical subcategory of modules $\mathcal{C}$ having a so-called $\triangle$-filtration (and also one with modules having ∇-filtration). As a beautiful illustration of Theorem 1 it was proved that $\mathcal{C}$ is contravariantly finite and resolving, and hence has an associated tilting module [Rin2]. This special tilting module associated with a quasihereditary algebra now plays an important role in the representation theory of algebraic groups, where by abuse of terminology, the word tilting module is used for an indecomposable summand of this particular tilting module [D].

There has also been a fruitful interplay between tilting theory and the theory of maximal Cohen–Macaulay modules over a complete local noetherian Cohen–Macaulay ring R. Here the dualizing module ω is the analogue of a cotilting module. Actually, the definition of a cotilting module for an algebra can be rephrased in such a way that ω becomes a cotilting module [AR]. Then the category $\mathcal{Y} = \{\, C : \mathrm{Ext}^i_\Lambda(C,U) = 0 \text{ for } i > 0 \,\}$ associated with a cotilting module U is the category $\mathrm{MCM}(R)$ of maximal Cohen–Macaulay modules over R. The theory of (maximal) Cohen–Macaulay approximations expresses amongst other things that the category $\mathcal{C} = \mathrm{MCM}(R)$ is contravariantly finite resolving [ABu], and the well known duality $\mathrm{Hom}_R(\ ,\omega)\colon \mathrm{MCM}(R) \to \mathrm{MCM}(R)$ corresponds to a similar one for algebras. Here there was mutual interplay between the developments within finite dimensional algebras and higher dimensional theory [ABr, ASm1, ASm2, ABu, AR]. In particular, the work on Cohen-Macaulay

approximations in [ABu] influenced the work on tilting and cotilting modules and their associated subcategories in [AR], where the point of view of the dualizing module being a cotilting module was stressed. Accordingly, the dualizing module and the special tilting (or equivalently, cotilting) module for quasihereditary algebras are special cases of the same common framework, hence also maximal Cohen–Macaulay modules and modules with $\triangle$-filtrations. Also dualizing complexes from algebraic geometry are similar to tilting complexes. Other connections with algebraic geometry via derived equivalence were discussed in Section 2.

After the description of derived equivalences via tilting complexes in [Ric], there has been a lot of activity on this topic in the representation theory of finite groups (see [Br]). The general theory of tilting with respect to torsion pairs has been applied to abstract blowing down in [V].

4 Quasitilted algebras

In this section we give some main results on quasitilted algebras. The type of questions investigated for this class of algebras illustrates the kind of information one is usually looking for about algebras in general. In particular, since quasitilted algebras generalize tilted algebras, established properties of tilted algebras serve as a guideline, as well as the properties of another important class of quasitilted algebras: the canonical algebras of Ringel [Rin1].

We start by giving some interesting and useful characterizations of quasitilted algebras [HRS].

THEOREM 4. *The following are equivalent for an algebra* Λ.

(a) Λ *is quasitilted.*

(b) $\operatorname{gl.dim}\Lambda \leq 2$ *and for each indecomposable* Λ*-module* C *we have* $\operatorname{pd}_\Lambda C \leq 1$ *or* $\operatorname{id}_\Lambda C \leq 1$*, where* $\operatorname{id}_\Lambda C$ *denotes the injective dimension of* C.

(c) If there is a sequence $X \to \cdots \to P$ *of nonzero maps between indecomposable* Λ*-modules and* P *is projective, then* $\operatorname{pd}_\Lambda X \leq 1$.

An interesting feature of the quasitilted algebras is that they contain the canonical algebras. The canonical k -algebras are special triangular matrix algebras of the form

$$H[M] = \begin{pmatrix} k & 0 \\ M & H \end{pmatrix},$$

called one-point extension of H by M, where H is hereditary and M is an H-module. The AR-quiver of an algebra is built from information given by almost split sequences, and tubes are important types of components occurring (see [ARS]). The canonical algebras provide examples of algebras with families of tubes of arbitrary type $(n_1, \ldots, n_t)$, where the n_i are greater than one. In addition, there is a curious trisection of the indecomposable modules into subcategories $\mathcal{P}$, $\mathcal{Q}$, and $\mathcal{R}$, where $\mathcal{Q}$ consists of what is called a sincere family of standard stable

tubes, $\mathrm{Hom}(\mathcal{R}, \mathcal{Q}) = 0 = \mathrm{Hom}(\mathcal{Q}, \mathcal{P}) = \mathrm{Hom}(\mathcal{R}, \mathcal{P})$, and any map $f\colon P \to R$ with P in $\mathcal{P}$ and R in $\mathcal{R}$ factors through any tube in $\mathcal{Q}$ [Rin1].

A natural related question is to investigate when a one-point extension $H[M]$ of a hereditary algebra H is quasitilted. We give the following result in this direction [HRS].

Theorem 5. *Let H be a indecomposable tame hereditary algebra, and let M be a nonzero regular module in* $\mathrm{mod}\, H$. *Then $H[M]$ is quasitilted if and only if M is quasisimple* (*that is, M is indecomposable and the middle term of the almost split sequence with M on the right is indecomposable*).

A central question for algebras is to describe the structure of the AR-quiver. For tilted algebras there is such a description, and also for canonical algebras, but there is yet no general description for quasitilted algebras. Of the information available, we cite the following (see [CH, CS, HRe1]).

Theorem 6. *(a) A quasitilted algebra has a preprojective component.*

(b) No component of the AR-quiver of a quasitilted non-tilted algebra Λ contains both a projective and an injective module.

Interesting open questions are whether the regular components for quasitilted non-tilted algebras are always tubes or of the form $\mathbb{Z}A_\infty$ (see [ARS]), and whether there is only one preprojective component.

A lot of effort in the representation theory of algebras has been given to classification of algebras of finite or tame representation type. For the quasitilted algebras the ones of finite type are already tilted [HRS], and there is a description for the tame quasitilted algebras [S].

5 Noetherian hereditary categories

Let throughout the rest of the paper $\mathcal{H}$ denote a hereditary abelian k-category with finite dimensional homomorphism and extension spaces. Since the quasitilted algebras are defined as endomorphism algebras of tilting objects in such hereditary k-categories, it is a central problem, in connection with understanding the whole class of quasitilted algebras, to classify the possible $\mathcal{H}$ which have a tilting object. In this section we discuss the noetherian case.

If $\mathcal{H}$ has a tilting object, then $\mathcal{H}$ has almost split sequences, and the Grothendieck group $K_0(\mathcal{H})$ is free abelian of finite rank [HRS]. We consider a natural class of categories $\mathcal{H}$ where $K_0(\mathcal{H})$ being free abelian of finite rank implies the existence of a tilting object [RV2].

A first example of a desired $\mathcal{H}$ with tilting object, which is not equivalent to $\mathrm{mod}\, H$ for a hereditary algebra H, is the category $\mathrm{coh}\, \mathbb{P}^1(k)$ of coherent sheaves on the projective line. More generally, there is introduced in [GL] the category $\mathrm{coh}\, \mathcal{X}$ of coherent sheaves on a weighted projective line $\mathcal{X}$. It was shown in [GL] that the canonical algebras could be realized as endomorphism algebras of particular tilting sheaves in $\mathrm{coh}\, \mathcal{X}$. This work was used to give an alternative approach to studying the module theory for canonical algebras. The following gives a complete description of the noetherian $\mathcal{H}$ with tilting object [L].

THEOREM 7. *The* coh $\mathcal{X}$ *and the* mod H *where* H *is a hereditary algebra constitute all connected noetherian hereditary* $\mathcal{H}$ *with tilting object.*

The category coh $\mathbb{P}^1(k)$ has an alternative description as the quotient category of the finitely generated $\mathbb{Z}$-graded $k[X,Y]$-modules modulo those of finite length. More generally, there is an interesting source of hereditary categories $\mathcal{H}_S$ (containing the coh $\mathcal{X}$) arising from two-dimensional $\mathbb{Z}$-graded isolated singularities S, finitely generated as a module over the center (see [RV2]). Interpreting $\mathcal{H}_S$ as the category of coherent modules over a sheaf of hereditary orders with center a nonsingular projective curve X, we have the following [RV2].

THEOREM 8. *Let* $S = k + S_1 + S_2 + \cdots$ *be a* $\mathbb{Z}$*-graded two-dimensional isolated singularity, with each* S_i *finite dimensional, and* S *finitely generated as a module over its center. Let* $\mathcal{H}_S$ *be the quotient category of finitely generated* $\mathbb{Z}$*-graded* S*-modules with degree zero maps, modulo the full subcategory of objects of finite length. Then the following are equivalent.*

(a) $K_0(\mathcal{H}_S) \simeq \mathbb{Z}^n$ *for some* n*.*

(b) The projective curve X *is a finite product of copies of* $\mathbb{P}^1(k)$*.*

(c) $\mathcal{H}_S$ *has a tilting object.*

(d) $\mathcal{H}_S$ *is equivalent to some* coh $\mathcal{X}$*.*

Possible choices for S with $K_0(\mathcal{H}_S) \simeq \mathbb{Z}^n$ for some n are two-dimensional $\mathbb{Z}$-graded Cohen–Macaulay isolated singularities of finite (graded) representation type, a complete classification of which is given in [RV1]. Other examples are $S = k[X,Y,Z]/(X^i + Y^j + Z^t)$, where i, j, t are pairwise relatively prime positive integers, and then $K_0(\mathcal{H}_S) \simeq \mathbb{Z}^{i+j+t-1}$ (see [GL]). The noetherian categories $\mathcal{H}_S$ with $K_0(\mathcal{H}_S) \simeq \mathbb{Z}^n$ form in a sense a bridge between some isolated Cohen–Macaulay two-dimensional singularities and a class of finite dimensional algebras, providing an additional connection between the areas.

6 HEREDITARY CATEGORIES WITH TILTING OBJECTS

We have seen in the previous section that the hereditary categories $\mathcal{H}$ with tilting object can be described in the noetherian case. In this section we discuss what can be said in general.

Since $\mathcal{H}$ is hereditary, the bounded derived category $D^b(\mathcal{H})$ has a simple description, as the indecomposable objects in this case are isomorphic to stalk complexes. When $\mathcal{H}$ has a tilting object, any hereditary abelian k-category $\mathcal{H}'$ derived equivalent to $\mathcal{H}$ also has a tilting object (and finite dimensional homomorphism and extension spaces) [HRe2]. Hence we obtain new hereditary categories $\mathcal{H}$ with tilting object by describing those in the same derived equivalence class as coh $\mathcal{X}$ and mod H (see [LS, H2]).

An interesting open problem is whether there are more hereditary categories $\mathcal{H}$ with tilting object than those derived equivalent to mod H or coh $\mathcal{X}$, or formulated differently, to finite dimensional hereditary or canonical algebras. We have the following information [HRe3].

Theorem 9. *Let $\mathcal{H}$ be a connected hereditary abelian k-category with tilting object. If $\mathcal{H}$ has some simple object, then $\mathcal{H}$ is derived equivalent to a hereditary or to a canonical algebra.*

Since every noetherian object has a simple quotient, it is also sufficient to require the existence of some noetherian object. If all objects are noetherian, the result follows from [L].

For hereditary algebras each indecomposable projective module is directing, that is does not lie on a nontrivial cycle of nonisomorphisms. We also have the following [HRe2].

Theorem 10. *Let $\mathcal{H}$ be a connected hereditary abelian k-category with tilting object. If $\mathcal{H}$ has some directing object, then $\mathcal{H}$ is derived equivalent to a finite dimensional hereditary k-algebra.*

The following provides further information along these lines [S].

Theorem 11. *If $\mathcal{H}$ is a connected hereditary abelian k-category with tilting object T such that $\operatorname{End}_{\mathcal{H}}(T)^{\mathrm{op}}$ is a tame algebra, then $\mathcal{H}$ is derived equivalent to a hereditary or to a canonical algebra.*

An important feature of hereditary categories $\mathcal{H}$ playing an essential role in the proof of Theorem 9, but also of more general interest, is the following result from [HRe3].

Theorem 12. *Let $\mathcal{H}$ be a connected hereditary abelian k-category with tilting object. Then for each exceptional object E in $\mathcal{H}$ (that is, $\operatorname{Ext}^1_{\mathcal{H}}(E,E) = 0$ and $\operatorname{End}_{\mathcal{H}}(E) \simeq k$) which is of infinite length and in $\operatorname{Fac} T$ for a tilting object T, the perpendicular category $E^{\perp}$ is equivalent to $\operatorname{mod} H$ for a finite dimensional hereditary k-algebra H.*

It is a consequence of Theorem 12 that any quasitilted algebra is derived equivalent to some one-point extension algebra $H[M]$ of a hereditary algebra H (see also [H2]). Hence a thorough investigation of such algebras $H[M]$ would also shed light on the problem of describing the $\mathcal{H}$ with tilting object.

We mention some open problems about quasitilted algebras, which would be answered if it is proved there are no more hereditary categories $\mathcal{H}$ with tilting object than those discussed above. A trisection for the canonical algebras was discussed in Section 4, and this trisection property characterizes a larger class of quasitilted algebras [LP]. Weakening the requirements on the middle part, other classes of quasitilted algebras can be characterized in such terms [LS, PR]. It is not known if this is the case for the quasitilted algebras. Another problem is formulated in terms of Hochschild cohomology. Is there some quasitilted algebra Λ with $H^1(\Lambda) \neq 0$ and $H^2(\Lambda) \neq 0$ [H3]? Denote by $\mathcal{D}$ the indecomposable Λ-modules C such that C and all its predecessors with respect to paths of nonzero maps have projective dimension at most one, and by $\mathcal{C}$ the indecomposable Λ-modules C such that C and all its successors with respect to paths of nonzero maps have injective dimension at most one. Is $\mathcal{C} \cap \mathcal{D}$ not empty for a quasitilted algebra Λ [HRS]?

A natural enlargement of the class of quasitilted algebras is the class of algebras derived equivalent to some hereditary category $\mathcal{H}$ with tilting object (or equivalently to a quasitilted algebra). It would be interesting to find a homological characterization of these algebras, called piecewise hereditary algebras (see [HRe3]).

REFERENCES

[ABr] M. Auslander and M. Bridger, *Stable module theory*, Memoirs of AMS 94 (1969), Amer. Math. Soc., Providence, R.I. 1969, 146pp.

[ABu] M. Auslander and R.O. Buchweitz, *Homological theory of maximal Cohen–Macaulay approximations*, Colloque en l'honneur de Pierre Samuel (Orsay 1987), Mem. Soc. Math. France (N.S.) No. 38 (1989), 5–37.

[APR] M. Auslander, M.I. Platzeck, and I. Reiten, *Coxeter functors without diagrams*, Trans. Amer. Math. Soc. 250 (1979), 1–46.

[AR] M. Auslander and I. Reiten, *Applications of contravariantly finite subcategories*, Adv. in Math. 86, No. 1 (1991), 111–152.

[ARS] M. Auslander, I. Reiten, and S.O. Smalø, *Representation theory of artin algebras*, Cambridge Univ. Press, Cambridge 1995.

[ASm1] M. Auslander and S.O. Smalø, *Preprojective modules over artin algebras*, J. Algebra 66 (1980), No. 1, 61–122.

[ASm2] M. Auslander and S.O. Smalø, *Almost split sequences in subcategories*, J. Alg. 69 (1981), 426–454, Addendum J. Alg. 71 (1981), 592–594.

[ASo] M. Auslander and Ø. Solberg, *Relative homology and representation theory II, Relative cotilting theory*, Comm. Algebra 21 (9) (1993), 3033–3079.

[Ba] D. Baer, *Tilting sheaves in representation theory of algebras*, Manuser. Math. 60 (1988), 323–347.

[Be] A.A. Beilinson, *Coherent sheaves on $\mathbb{P}^n$ and problems of linear algebra*, Func. An. and Appl. 12 (1978), 212–214.

[Br] M. Broué, *Equivalences of blocks of group algebras, Finite dimensional algebras and related topics*, Kluwer 1994, SerC-Vol. 424.

[BB] S. Brenner and M.C.R. Butler, *Generalization of Bernstein–Gelfand–Ponomarev reflection functors*, Springer LNM 832 (1980), 103–169.

[BGP] I.N. Bernstein, I.M. Gelfand, and V.A. Ponomarev, *Coxeter functors and Gabriel's theorem*, Russian Math. Surveys 28 (1973), 17–32.

[BS] A. Buan and Ø. Solberg, *Relative cotilting theory and almost complete cotilting modules*, Proc. ICRA VIII (Geiranger), CMS Conf. proc., Vol. 24, Algebras and modules II (1998), 77–93.

[CPS] E. Cline, B. Parshall, and L. Scott, *Derived categories and Morita theory*, J. Alg. 104 (1986), 397–409.

[CH] F. Coelho and D. Happel, *Quasitilted algebras admit a preprojective component*, Proc. Amer. Math. Soc., Vol. 125, No. 5 (1997), 1283–1291.

[CS] F. Coelho and A. Skowronski, *On Auslander–Reiten components for quasitilted algebras*, Fund. Math. 149 (1996), 67–82.

[D] S. Donkin, *On tilting modules for algebraic groups*, Math. Zeit. 212 (1993), 39–60.

[GL] W. Geigle and H. Lenzing, *A class of weighted projective curves arising in representation theory of finite dimensional algebras*, In: *Singularities, Representations of Algebras and Vector Bundles*, Springer Lecture Notes in Math. 1273 (1987), 265–297.

[GRS] E.L. Green, I. Reiten, and Ø. Solberg, *Koszul and Yoneda algebras*, in preparation.

[H1] D. Happel, *Triangulated Categories in the Representation Theory of Finite Dimensional Algebras*, LMS Lecture Note Series 119, Cambridge 1988.

[H2] D. Happel, *Quasitilted algebras*, Proc. ICRA VIII (Trondheim), CMS Conf. proc., Vol. 23, Algebras and modules I (1998), 55–83.

[H3] D. Happel, *Hochschild cohomology of piecewise hereditary algebras*, Colloq. Math., to appear.

[HRe1] D. Happel and I. Reiten, *An introduction to quasitilted algebras*, An. St. Univ. Ovidius Constantza Vol. 4 (1996), 137–149.

[HRe2] D. Happel and I. Reiten, *Directing objects in hereditary categories*, Proc. Seattle Conf. (1997), to appear.

[HRe3] D. Happel and I. Reiten, *Hereditary categories with tilting object*, Math. Zeit., to appear.

[HRS] D. Happel, I. Reiten, and S.O. Smalø, *Tilting in abelian categories and quasitilted algebras*, Memoirs Amer. Math. Soc. 575, AMS 1996.

[HRi] D. Happel and C.M. Ringel, *Tilted algebras*, Trans. Amer. Math. Soc. 274 (1982), 399–443.

[HU] D. Happel and L. Unger, *Complements and the generalized Nakayama conjecture*, Proc. ICRA VIII (Geiranger), CMS Conf. Proc. Vol. 24, Algebras and modules II (1998), 293–331.

[L] H. Lenzing, *Hereditary noetherian categories with a tilting complex*, Proc. Amer. Math. Soc. 125 (1997), 1893–1901.

[LP] H. Lenzing and J.A. de la Peña, *Concealed canonical algebras with separating tubular families*, preprint.

[LS] H. Lenzing and A. Skowronski, *Quasitilted algebras of canonical type*, Colloq. Math. 71 (1996), 161–181.

[M] T. Miayshita, *Tilting modules of finite projective dimension*, Math. Zeit. 193 (1986), 112–146.

[PR] J.A. de la Peña and I. Reiten, in preparation.

[Ric] J. Rickard, *Morita theory for derived categories*, J. London Math. Soc. 39 (1989), 436–456.

[Rin1] C.M. Ringel, *Tame Algebras and Integral Quadratic Forms*, Springer Lecture Notes in Math. 1099, 1994.

[Rin2] C.M. Ringel, *The category of modules with good filtrations over a quasi-hereditary algebra has almost split sequences*, Math. Zeit. 208 (1991), 209–225.

[RV1] I. Reiten and M. Van den Bergh, *Two-dimensional tame and maximal orders of finite representation type*, Memoirs AMS, Vol. 80, No. 408, 1989.

[RV2] I. Reiten and M. Van den Bergh, in preparation.

[S] A. Skowronski, *Tame quasitilted algebras*, J. Alg., 1998.

[TW] H. Tachikawa and T. Wakamatsu, *Tilting functors and stable equivalences for self-injective algebras*, J. Alg. 109 (1987), 138–165.

[V] M. Van den Bergh, *Abstract blowing down*, preprint 1998.

Idun Reiten
Department of Mathematical Sciences
Norwegian University of Science and Technology
7034 Trondheim, Norway
idunr@math.ntnu.no

The Abelian Defect Group Conjecture

Jeremy Rickard

Abstract. Let G be a finite group and k an algebraically closed field of characteristic $p > 0$. If B is a block of the group algebra kG with defect group D, the Brauer correspondent of B is a block b of $kN_G(D)$. When D is abelian, the blocks B and b, although they are rarely isomorphic or even Morita equivalent, seem to be very closely related. For example, Alperin's Weight Conjecture predicts that they should have the same number of simple modules. Broué's Abelian Defect Group Conjecture gives a more precise prediction of the relationship between B and b: their module categories should have equivalent derived categories. In this article we survey this conjecture, some of its consequences, and some of the recent progress that has been made in verifying it in special cases.

1991 Mathematics Subject Classification: 20C20, 18E30
Keywords and Phrases: modular representation theory, derived category, abelian defect group conjecture

1 Notation and terminology

Throughout this article, G will denote a finite group.

We shall be dealing with the characteristic p representation theory of G, where p is a prime. We shall use three coefficient rings. The ring $\mathcal{O}$ will be a complete discrete valuation ring with residue field k of characteristic p and field of fractions K of characteristic zero. Since we shall not be concerned with rationality questions, we shall assume that these coefficient rings are all 'large enough' in that they contain enough roots of unity.

As well as the group algebras $\mathcal{O}G$, kG and KG, we shall be concerned with various direct factors. We shall choose our notation so that if we denote an $\mathcal{O}$-algebra by $\mathcal{O}A$, we shall use the notation kA and KA for $\mathcal{O}A \otimes_{\mathcal{O}} k$ and $\mathcal{O}A \otimes_{\mathcal{O}} K$ respectively. It is well-known that the natural surjection $\mathcal{O}G \to kG$ induces a bijection between the primitive central idempotents of $\mathcal{O}G$ and kG, so that if

$$\mathcal{O}G \cong \mathcal{O}A_1 \times \cdots \times \mathcal{O}A_n,$$

where $\mathcal{O}A_1, \ldots, \mathcal{O}A_n$ are the blocks (i.e., the minimal direct factors) of $\mathcal{O}G$, then

$$kG \cong kA_1 \times \cdots \times kA_n,$$

where $kA_1, \ldots, kA_n$ are the blocks of kG.

Here we shall only be concerned with finitely-generated modules, and if R is any ring, we shall denote the category of finitely-generated right R-modules by $\mathrm{mod}(R)$. By a 'module' for a ring R we shall always mean a right module.

2 Local representation theory and Alperin's Weight Conjecture

There are two major themes running through much of the modular representation theory of general finite groups that relate the representation theory of a group G to that of smaller groups.

The first of these is Clifford theory, which relates the representation theory of a group G with a normal subgroup N to that of N and G/N. This is an area that has been studied systematically and intensively, but we shall not say much about it here.

The second theme is sometimes known as local representation theory, which describes the relationship between the representation theory of G and of *local subgroups*: normalizers of non-trivial p-subgroups in G. This relationship has been exploited in a more ad hoc way than Clifford theory; this is partly because, as we shall see, the precise relationship is unclear or at best conjectural.

One classical example of a theorem of local representation theory is Brauer's First Main Theorem. Recall that the *defect group* of a block kA of kG is a minimal subgroup D of G such that every kA-module is a direct summand of a module induced from kD. The defect group is always a p-subgroup and is determined uniquely up to conjugacy in G.

Theorem 2.1 (Brauer's First Main Theorem) *If D is a p-subgroup of G, there is a natural bijection between the blocks of kG with defect group D and the blocks of $kN_G(D)$ with defect group D.*

The block of $kN_G(D)$ corresponding to a block kA of kG with defect group D is called the *Brauer correspondent* of kA. The *principal block* of kG (i.e., the unique block that is not contained in the augmentation ideal of kG) has a Sylow p-subgroup P of G as its defect group, and its Brauer correspondent is the principal block of $kN_G(P)$.

The most famous example of a general conjecture in local representation theory is the following, due to Alperin [A], known as Alperin's Weight Conjecture, which has inspired a great deal of interest since it was formulated in the 1980s.

Conjecture 2.2 (Alperin's Weight Conjecture) *The number of isomorphism classes of simple kG-modules is equal to the number of pairs (Q, S), where Q runs over a set of representatives of conjugacy classes of p-subgroups of G and, for each Q, S runs over a set of representatives of isomorphism classes of simple projective $k[N_G(Q)/Q]$-modules.*

If we ignore the simple kG-modules that are projective, this conjecture claims that the number of non-projective simple kG-modules is equal to the number of pairs (Q, S) where Q is a non-trivial p-subgroup of G. In other words, it claims

that the number of non-projective simple kG-modules is 'locally determined' (i.e., determined by local subgroups) in a precise fashion.

There is a more precise 'blockwise' version [A] of Alperin's Weight Conjecture, dealing with the number of simple modules for a single block of kG in terms of suitable blocks of local subgroups. We shall not state the general conjecture here, but only the special case for a block with abelian defect group, which has a particularly simple form.

Conjecture 2.3 *Let kA be a block of kG with an abelian defect group D. Then kA and its Brauer correspondent have the same number of isomorphism classes of simple modules.*

For principal blocks, this has the following special case.

Conjecture 2.4 *Suppose G has an abelian Sylow p-subgroup P. Then the principal blocks of kG and $kN_G(P)$ have the same number of isomorphism classes of simple modules.*

3 Broué's Abelian Defect Group Conjecture

Since Alperin's Weight Conjecture reduces, for a block with abelian defect group, to the claim that the block and its Brauer correspondent have the same number of simple modules, it is natural to wonder whether there is some structural relationship between the two blocks that explains this. It is certainly not true in general that the blocks are isomorphic or even Morita equivalent. Broué [B] conjectured such a relationship in terms of derived categories.

There are now several accessible introductions to the theory of derived categories, such as the one contained in Weibel's book [W]. If R is a noetherian ring, we shall denote by $D^b(R)$ the bounded derived category of $\mathrm{mod}(R)$. Recall that the objects of $D^b(R)$ are the chain complexes of finitely-generated R-modules with only finitely many non-zero terms. As usual we shall think of $\mathrm{mod}(R)$ as embedded in $D^b(R)$ by identifying an R-module M with the complex whose only non-zero term is M in degree zero. The morphisms of $D^b(R)$ are obtained from the chain maps by formally adjoining inverses to all chain maps that induce isomorphisms in homology. Recall finally that $D^b(R)$ has the structure of a 'triangulated category': in particular, for each object X of $D^b(R)$ we can form an object $X[n]$ for $n \in \mathbb{Z}$ by shifting the complex X to the left by n places.

Conjecture 3.1 (Broué's Abelian Defect Group Conjecture) *Let $\mathcal{O}A$ be a block of $\mathcal{O}G$ with abelian defect group D and let $\mathcal{O}B$ be its Brauer correspondent (hence a block of $\mathcal{O}N_G(D)$). Then $D^b(\mathcal{O}A)$ and $D^b(\mathcal{O}B)$ are equivalent as triangulated categories.*

If R and S are noetherian rings such that $D^b(R)$ and $D^b(S)$ are equivalent as triangulated categories, we say that R and S are *derived equivalent*. Derived equivalence is clearly implied by Morita equivalence, but the converse is not true.

We have stated the conjecture over $\mathcal{O}$; it is not hard to prove that this implies the corresponding statement over k.

The Grothendieck group $K_0(\mathcal{T})$ of a triangulated category $\mathcal{T}$ can be defined [G] in a similar way to that of an abelian category, and if $\mathcal{T} = D^b(R)$ for some finite-dimensional k-algebra R, then $K_0(\mathcal{T})$ is a free abelian group whose rank is equal to the number of isomorphism classes of simple R-modules. Hence the Abelian Defect Group Conjecture implies the blockwise version of the Weight Conjecture for blocks with abelian defect group. An important open problem is to formulate a generalization of Broué's conjecture that would imply the Weight Conjecture for a general block.

4 Proving derived equivalence

Given two rings R and S, how does one go about proving that they are derived equivalent? I shall assume that either R and S are both finite-dimensional k-algebras or they are both $\mathcal{O}$-free $\mathcal{O}$-algebras of finite rank over $\mathcal{O}$, although everything that follows applies much more generally.

Most of the classical theory of Morita equivalence has generalizations to derived equivalence.

Recall first that R and S have equivalent module categories if and only if S is isomorphic to the endomorphism algebra of a finitely-generated projective generator for R. This has the following analogue [R1] for derived equivalence.

THEOREM 4.1 *R and S are derived equivalent if and only if S is isomorphic to the endomorphism algebra, in $D^b(R)$, of an object T such that*

(i) T is a bounded complex of finitely-generated projective R-modules,

(ii) $\mathrm{Hom}_{D^b(R)}(T, T[i]) = 0$ for $i \neq 0$, and

(iii) If X is an object of $D^b(R)$ such that $\mathrm{Hom}_{D^b(R)}(T, X[i]) = 0$ for all $i \in \mathbb{Z}$, then $X \cong 0$.

An object T satisfying conditions (i) to (iii) of the theorem is called a (one-sided) *tilting complex*. Condition (iii) has equivalent forms that are easier to check directly in practice.

Another well-known criterion for R and S to be Morita equivalent is that there should be an R-S-bimodule X and an S-R-bimodule Y (which is in fact isomorphic to $\mathrm{Hom}_R(X, R)$) such that X and Y are finitely-generated and projective as right modules and as left modules (but not usually projective as bimodules) and such that $X \otimes_S Y \cong R$ and $Y \otimes_R X \cong S$ as bimodules. Then the functor $? \otimes_R X$ is an equivalence of module categories. This also has an analogy for derived categories, first proved in [R2] but with a better subsequent proof by Keller [K].

THEOREM 4.2 *R and S are derived equivalent if and only if there is a bounded complex X of R-S-bimodules and a bounded complex $Y = \mathrm{Hom}_R(X, R)$ of S-R-bimodules such that*

(i) All the terms of X and Y are finitely-generated and projective as left modules and as right modules,

(ii) As a complex of R-bimodules, $X \otimes_S Y \cong R \oplus C$ for some acyclic complex C, and

(iii) As a complex of S-bimodules, $Y \otimes_R X \cong S \oplus C'$ for some acyclic complex C'.

A complex X that satisfies the conditions of the theorem is called a *two-sided tilting complex*. If we forget the left action of S on Y, then Y becomes a one-sided tilting complex for R.

If X is a two-sided tilting complex, then the functor

$$? \otimes_R X : D^b(R) \longrightarrow D^b(S)$$

is an equivalence of triangulated categories.

As we shall see in Section 6, a two-sided tilting complex seems to be the natural object to seek in order to prove the Abelian Defect Group Conjecture, although in small examples it is easier to do calculations with one-sided tilting complexes.

5 Character-theoretic consequences

If $\mathcal{O}A$ and $\mathcal{O}B$ are derived equivalent blocks and X is a two-sided tilting complex, then it is easy to check that $X \otimes_{\mathcal{O}} K$ is also a two-sided tilting complex for the semisimple algebras KA and KB.

The Grothendieck group of $D^b(KA)$ can be naturally identified with the group $K_0(KA)$ of generalized characters of KA, so X induces an isomorphism

$$\theta : K_0(KA) \cong K_0(KB).$$

The indecomposable objects of $D^b(R)$ for a semisimple K-algebra R are all of the form $M[i]$ for some irreducible R-module M and some integer i. It follows that θ maps each irreducible character χ of KA to $\pm\phi$ for some irreducible character ϕ of KB: in other words, θ is an isometry. Since the functors $? \otimes_R X$ and $? \otimes_S \mathrm{Hom}_R(X, R)$ take projective modules to complexes of projective modules, it follows that θ restricts to an isomorphism

$$\theta_p : K_{0,p}(KA) \longrightarrow K_{0,p}(KB)$$

between the subgroups of the groups of generalized characters generated by the characters of projective modules for $\mathcal{O}A$ and $\mathcal{O}B$. Such an isometry is called a *perfect isometry* by Broué [B] and can be characterized in terms of arithmetic properties of character values.

A consequence of the Abelian Defect Group Conjecture is therefore the following weaker character-theoretic conjecture, which is however still strong enough to imply the Weight Conjecture for blocks with abelian defect group.

Conjecture 5.1 *If $\mathcal{O}A$ is a block of $\mathcal{O}G$ with abelian defect group and with Brauer correspondent $\mathcal{O}B$, then there is a perfect isometry*

$$K_0(KA) \longrightarrow K_0(KB).$$

It is easier to perform calculations with characters than with derived categories, and so it is no surprise that this weaker conjecture has been verified in many more cases than the Abelian Defect Group Conjecture. One of the most impressive examples is the following, proved by Fong and Harris [FH] using the classification of finite simple groups. In fact, they proved an even stronger character-theoretic statement.

THEOREM 5.2 (FONG, HARRIS) *If $p = 2$ and G has an abelian Sylow p-subgroup P, there is a perfect isometry between the principal blocks of $\mathcal{O}G$ and $\mathcal{O}N_G(P)$.*

6 Splendid equivalences

Here we shall briefly summarize some of the main results of [R3], giving some extra conditions that the two-sided tilting complexes predicted by the Abelian Defect Group Conjecture are expected to satisfy. We shall restrict our attention to the case of principal blocks, although Harris [H] and Puig [P] have given generalizations to non-principal blocks.

Suppose then that G has an abelian Sylow p-subgroup and that $\mathcal{O}A$ and $\mathcal{O}B$ are the principal blocks of $\mathcal{O}G$ and $\mathcal{O}N_G(P)$ respectively. We can consider a two-sided tilting complex X for $\mathcal{O}A$ and $\mathcal{O}B$ as a complex of $\mathcal{O}[G \times N_G(P)]$-modules. We say that X is a *splendid* tilting complex if it satisfies the following conditions, where as before $Y = \mathrm{Hom}_{\mathcal{O}A}(X, \mathcal{O}A)$.

- X is a complex of $\mathcal{O}[G \times N_G(P)]$-modules whose restrictions to $\mathcal{O}[P \times P]$ are permutation modules of the form $\mathcal{O}\Omega$, where the point stabilizers of Ω are conjugate to subgroups of the diagonal embedding of P in $P \times P$.
- $X \otimes_{\mathcal{O}B} Y \cong \mathcal{O}A \oplus C$, where C is a contractible complex of $\mathcal{O}A$-bimodules.
- $Y \otimes_{\mathcal{O}A} X \cong \mathcal{O}B \oplus C'$, where C' is a contractible complex of $\mathcal{O}B$-bimodules.

A derived equivalence induced by a splendid tilting complex is called a *splendid equivalence*.

Of course, we can make a similar definition over k. The second and third conditions are of course stronger than the conditions in the definition of a two-sided tilting complex, where C and C' were only required to be acyclic. Known examples suggest that Broué's Abelian Defect Group Conjecture should still be true if we require the derived equivalences it predicts to be splendid.

The main property that motivates the introduction of the idea of splendid equivalence is given in the next theorem [R3].

THEOREM 6.1 *If G has an abelian Sylow p-subgroup P and there is a splendid equivalence between the principal blocks of $\mathcal{O}G$ and $N = \mathcal{O}N_G(P)$, then for each subgroup $Q \leq P$ there is a splendid equivalence between the principal blocks of $\mathcal{O}C_G(Q)$ and $\mathcal{O}C_N(Q)$.*

In fact, a more precise statement can be made about the relationship between the two perfect isometries induced by the splendid equivalences: a splendid

equivalence induces what Broué [B] calls an 'isotypy': a compatible family of perfect isometries between principal blocks of $\mathcal{O}C_G(Q)$ and $\mathcal{O}C_N(Q)$, one for each subgroup $Q \leq P$.

7 Recent progress in verifying the Abelian Defect Group Conjecture

A complete proof of the conjecture still seems a long way off. For several years after Broué formulated the conjecture, it could only be proved for fairly simple blocks, such as those with cyclic defect group, where a lot was known about the precise structure of the blocks. However, in the last few years there has been significant progress in developing techniques to verify it for particular groups.

The most complex infinite family of examples for which the conjecture has been verified is given by the following theorem [C].

Theorem 7.1 (Chuang) *The Abelian Defect Group Conjecture is true for all blocks of symmetric groups whose defect group has order p^2. Moreover, the derived equivalence is splendid.*

In particular, Chuang's theorem proves the conjecture for all blocks of the symmetric group S_n if $n < 3p$.

Consider, for simplicity, the principal block of kG, where G has an abelian Sylow p-subgroup P. The main obstacle to performing calculations to verify the Abelian Defect Group Conjecture in this case has been that the precise structure of the projective kG-modules is hard to calculate for all but the simplest examples, so it is hard to calculate one-sided tilting complexes. In contrast, the structure of projective $kN_G(P)$-modules is relatively easy to understand. In as yet unpublished work, Okuyama has introduced an ingenious technique, based on a theorem of Linckelmann [L], that allows him to verify the conjecture for several groups G without knowing the precise structure of the projective kG-modules. In fact, as a byproduct of his verifications, it is possible to calculate the structure of these modules.

Here are a few examples of the cases that Okuyama has settled.

Theorem 7.2 (Okuyama, 1997) *For $p = 3$, the Abelian Defect Group Conjecture is true for the principal blocks of the groups $M_{11}, M_{21}, M_{22}, M_{23}$ and HS.*

References

[A] J. L. Alperin, *Weights for finite groups.* The Arcata conference on representations of finite groups (Arcata, Calif., 1986), Proc. Sympos. Pure Math. 47, Part 1, Amer. Math, Soc., Providence, RI, 1987, 369–379.

[B] M. Broué, *Isométries parfaites, types de blocs, catégories dérivées.* Astérisque 181–182 (1990), 61–92.

[C] J. Chuang, *Broué's conjecture for symmetric group blocks of defect two.* Preprint, Chicago, 1997.

[FH] P. Fong and M. E. Harris, *On perfect isometries and isotypies in finite groups.* Invent. Math. 114 (1993), 139–191.

[G] A. Grothendieck, *Groupes des classes des catégories abeliennes et triangulées. Complexes parfait.* Lecture Notes in Mathematics 589 (Springer, Berlin, 1972), 351–371.

[H] M. E. Harris, *Splendid derived equivalences for arbitrary blocks of finite groups.* Preprint, Minnesota, 1997.

[K] B. Keller, *Deriving DG categories.* Ann. Scient. Éc. Norm. Sup. 27 (1994), 63–102.

[L] M. Linckelmann, *Stable equivalences of Morita type for self-injective algebras and p-groups.* Math. Zeit. 223 (1996), 87–100.

[P] L. Puig, *On the local structure of Morita and Rickard equivalences between Brauer blocks.* Preprint, Paris, 1996.

[R1] J. Rickard, *Morita theory for derived categories.* J. London Math. Soc. (2) 39 (1989), 436–456.

[R2] J. Rickard, *Derived equivalences as derived functors.* J. London Math. Soc. (2) 43 (1991), 37–48.

[R3] J. Rickard, *Splendid equivalences: derived categories and permutation modules.* Proc. London Math. Soc. (3) 72 (1996), 331–358.

[W] C. A. Weibel, *An introduction to homological algebra.* Cambridge Studies in Advanced Mathematics vol. 38, Cambridge University Press, 1994.

Jeremy Rickard
School of Mathematics
University of Bristol
University Walk
Bristol BS8 1TW
ENGLAND

Simple Groups, Permutation Groups, and Probability

Aner Shalev

Abstract. We survey recent progress, made using probabilistic methods, on several conjectures concerning finite groups.

1991 Mathematics Subject Classification: Primary 20D06; Secondary 20P05, 20B15.

1 Random generation

In recent years probabilistic methods have proved useful in the solution of several difficult problems concerning finite groups; these involve conjectures on finite simple groups and on finite permutation groups. In some cases the probabilistic nature of the problem is apparent from its very formulation; but in other cases the use of probability, or counting, seems surprising, and cannot be anticipated by the nature of the problem.

In some branches of mathematics it is quite common to use probabilistic methods, or related non-constructive methods, in order to prove existence theorems (Cantor's proof of the existence of transcendental numbers is a classical example). However, this is less common in group theory. Indeed, it is our hope that the probabilistic approach will have sufficiently many group-theoretic applications so as to become a standard tool in group theory.

The roots of the subject lie in a series of 7 papers by Erdős and Turán (starting with [ET1]) in which they study the properties of random permutations. For example they show that most permutations in the symmetric group S_n have order about $n^{\frac{1}{2}\log n}$, and have about $\log n$ cycles.

Dixon [D] used the Erdős-Turán theory to settle an old conjecture of Netto, proving that two randomly chosen elements of the alternating group A_n generate A_n with probability $\to 1$ as $n \to \infty$. He proposed the following generalization:

Conjecture 1 (Dixon, 1969): Two randomly chosen elements of a finite simple group G generate G with probability $\to 1$ as $|G| \to \infty$.

Here G and its Cartesian powers are regarded as probability spaces with respect to the uniform distribution.

At the time this was a rather daring conjecture, since the Classification of Finite Simple Groups was not yet available. Invoking the Classification Theorem (which we do throughout) it remained to prove the conjecture for the simple groups of Lie type.

A breakthrough was made in 1990 by Kantor and Lubotzky, who proved Dixon's conjecture for classical groups and for some small rank exceptional groups of Lie type [KL]. The remaining exceptional groups were handled by Liebeck and myself in 1995 [LiSh1], so we have:

THEOREM 1 (Dixon, Kantor, Lubotzky, Liebeck, Shalev): *Dixon's conjecture holds.*

This result has quantitative versions. Let $m(G)$ denote the minimal index of a proper subgroup of G. Then it turns out that the probability that two randomly chosen elements of a finite simple group G do not generate G is approximately $m(G)^{-1}$ (up to a multiplicative constant); see [Ba], [K], [LiSh3] for this and for more refined estimates.

We also obtain results on random generation by special pairs of elements. In their paper [KL] Kantor and Lubotzky pose the following:

CONJECTURE 2 (Kantor-Lubotzky, 1990): A randomly chosen involution and a randomly chosen additional element of a finite simple group G generate G with probability $\to 1$ as $|G| \to \infty$.

This was settled in [LiSh3], so we have:

THEOREM 2 (Liebeck-Shalev, 1996): *Kantor-Lubotzky's conjecture holds.*

We also show in [LiSh3] that a finite simple group which is not a Suzuki group is almost surely generated by a random element of order 3 and a random additional element.

A related question raised in [KL] is as follows. Let G be a finite simple group, and let $x \in G$ be a non-identity element. Let $P_x(G)$ denote the probability that x and a randomly chosen element of G generate G. What can be said about $P_x(G)$? Guralnick, Kantor and Saxl constructed examples where $P_x(G) \to 0$. It is shown in [GKS], [Sh1], [LiSh5], [GLSSh] that, unless G is alternating or a classical group over a field of bounded size (in which case $P_x(G)$ may be bounded away from 1), we have $P_x(G) \to 1$ as $|G| \to \infty$ (regardless of the choices of x). Another interesting result which was just established in [GK] is that $P_x(G)$ is always positive, namely, every non-identity element of a finite simple group sits in some generating pair.

Recently G. Robinson asked if a finite simple group is randomly generated by two conjugate elements. By [Sh2], [LiSh5], [GLSSh] we have:

THEOREM 3 (Guralnick-Liebeck-Saxl-Shalev, 1998) *Let G be a finite simple group and let $x, y \in G$ be randomly chosen elements. Then the elements x and $y^{-1}xy$ generate G with probability tending to 1 as $|G| \to \infty$.*

We conclude this section with a remark on profinite groups. A profinite group G has a canonical normalized Haar measure which turns it into a probability space. If, for some positive integer k, G is generated with positive probability by k randomly chosen elements, we say that G is positively finitely generated. The first examples of such groups occurred in the context of field arithmetic, and their research was continued in [KL], [Bh], [M], [MSh], [BPSh]. Positively finitely generated groups have been characterized as profinite groups in which the number

of index n maximal subgroups grows polynomially with n [MSh]. However, we are still unable to find a structural characterization of such groups, or even to formulate a reasonable conjecture.

2 The modular group

We now turn to some recent applications of the probabilistic approach. The first concerns the longstanding problem of finding the finite simple quotients of the modular group $\mathrm{PSL}_2(\mathbb{Z})$, namely the finite simple groups that can be generated by two elements of orders 2 and 3 respectively. Groups with this property are termed $(2,3)$-generated. The interest in this problem arose in geometric contexts, namely actions of finite groups on Riemann surfaces. Partial answers were provided throughout this century. For example, Miller showed in 1901 that the alternating groups of degree at least 9 are $(2,3)$-generated. The $(2,3)$-generation problem for $\mathrm{PSL}_2(q)$ was studied by Brahana and Sinkov in the 20s and 30s and was solved by Macbeath in the 60s. Some classical groups with large Lie rank were handled by Tamburini and others. In 1996 Di Martino and Vavilov showed that the simple groups $\mathrm{PSL}_n(q)$ are $(2,3)$-generated provided q is odd and $(n,q) \neq (2,9)$.

The proofs of these and many other results in the field are based on explicit constructions of generators of orders 2 and 3. This approach seems to fail for various families of classical groups, for example those with "intermediate" Lie rank. While some simple groups are not $(2,3)$-generated, the following conjecture was recently posed (see [W]).

Conjecture 3 (Di Martino-Vavilov, Wilson): All finite simple groups of Lie type except some of low rank in characteristic 2 or 3 are quotients of $\mathrm{PSL}_2(\mathbb{Z})$.

In [LiSh2] we address this problem for classical groups, using a probabilistic approach. Let $P_{2,3}(G)$ denote the probability that a random involution and a random element of order 3 generate G.

Theorem 4 (Liebeck-Shalev, 1996): *Let $G \neq \mathrm{PSp}_4(q)$ be a finite simple classical group. Then $P_{2,3}(G) \to 1$ as $|G| \to \infty$. If $G = \mathrm{PSp}_4(p^k)$ $(p \geq 5)$ then $P_{2,3}(G) \to 1/2$ as $|G| \to \infty$.*

This gives rise to the following.

Theorem 5 (Liebeck-Shalev, 1996): *Except for $\mathrm{PSp}_4(2^k)$, $\mathrm{PSp}_4(3^k)$ and finitely many other groups, all finite simple classical groups can be obtained as quotients of $\mathrm{PSL}_2(\mathbb{Z})$.*

The groups $\mathrm{PSp}_4(2^k)$ and $\mathrm{PSp}_4(3^k)$ turn out to be genuine exceptions.

The $(2,3)$-generation problem for exceptional groups of Lie type has just been solved by Lübeck and Malle [LM]. Using character theory (and computer calculations) they show that, except for the Suzuki groups and the group $G_2(2)'$, all simple exceptional groups of Lie type are obtained as quotients of the modular group. Combining this with Theorem 5 we see that Conjecture 3 is now confirmed up to finitely many exceptions.

3 Free groups

It is interesting that results on random generation, and Theorem 1 in particular, can be applied in the study of residual properties of free groups.

Let F_d be the free group on d generators ($d \geq 2$). It is well known that F_d is residually finite, and even residually p for any prime p. The following problem concerning residual properties of free groups was raised by Magnus, and then by Gorchakov and Levchuk.

MAGNUS PROBLEM: is F_d residually X for any infinite collection X of finite simple groups?

In other words, suppose X is an infinite collection of finite simple groups; does it follow that

$$\cap\{N \lhd F_d : F_d/N \in X\} = 1?$$

Since F_d is residually F_2, the question is reduced to the case $d = 2$. Several partial answers were given in the past three decades, and a complete positive solution to the problem was given by T. Weigel in 1993 [We1-We3].

In order to outline our approach to the problem, first note that it suffices to show that for every $1 \neq w = w(u, v) \in F_2$, almost all finite simple groups G have a generating pair x, y such that $w(x, y) \neq 1$. To prove this we establish a stronger result of a probabilistic nature [DPSSh].

THEOREM 6 (Dixon-Pyber-Seress-Shalev): *Fix $1 \neq w = w(u, v) \in F_2$. Let G be a finite simple group, and let $x, y \in G$ be randomly chosen elements. Then, as $|G| \to \infty$ we have* $\mathrm{Prob}(\langle x, y\rangle = G \ \wedge \ w(x, y) \neq 1) \to 1$.

The proof of Theorem 6 starts with the following reduction. Applying Theorem 1, we know that $\mathrm{Prob}(\langle x, y\rangle = G) \to 1$. Hence it suffices to prove that the probability that $w(x, y) \neq 1$ tends to 1 as $|G| \to \infty$. The last statement has the advantage that it no longer deals with generating pairs. We just have to show that (as $|G| \to \infty$) most pairs in G^2 do not satisfy a given relation. This can be proved using some algebraic geometry and suitable combinatorial tricks. Recently Pyber developed these ideas further in his study of dense free subgroups of profinite groups.

4 Permutation groups

Several of the recent applications of the probabilistic approach involve permutation groups. Let me start with a counting problem. How many maximal subgroups does the symmetric group S_n have up to conjugacy? In 1989 Babai showed, using the Classification Theorem, that S_n has at most $n^{(1+o(1))\log^3 n}$ conjugacy classes of maximal subgroups [Ba].

In [LiSh4] this is improved as follows.

THEOREM 7 (Liebeck-Shalev, 1996): *S_n has $n/2 + o(n)$ conjugacy classes of maximal subgroups.*

Note that the intransitive subgroups, which have the form $S_k \times S_{n-k}$, already yield $n/2$ classes of maximal subgroups. Therefore Theorem 7 asserts that, in some sense, almost all maximal subgroups of S_n are the obvious intransitive ones.

The methods of [LiSh4] are also relevant in counting all maximal subgroups of S_n. In this context let me mention the following general conjecture.

Conjecture 4 (Wall, 1961): The number of maximal subgroups of a finite group G is less than $|G|$.

This conjecture was confirmed by Wall for soluble groups [Wa]. We show in [LiSh4] that it also holds for symmetric groups of sufficiently large degree.

We now turn to other applications involving permutation groups. Recall that a base for a permutation group is a subset of the permutation domain whose pointwise stabilizer is trivial. Bases play an important role in computational group theory and in estimating orders of primitive permutation groups. The base size $b(G)$ of G is defined as the minimal size of a base for G, and is the subject of several conjectures. We start with

Conjecture 5 (Babai, 1982): There is a function f such that, if $G \leq S_n$ is a primitive group not involving A_d as a section, then $b(G) \leq f(d)$.

See Pyber's excellent survey [P1]. First positive evidence was provided in 1996 by Seress, who showed that $b(G) \leq 4$ for G soluble. Then, in the joint work [GSSh] with Gluck and Seress, we show the following.

Theorem 8 (Gluck-Seress-Shalev, 1998): *Babai's conjecture holds.*

This provides a structural explanation for the celebrated Babai-Cameron-Pálfy theorem, stating that the order of the groups above is polynomial in n [BCP]. The original proof in [GSSh] yields $f(d) = O(d^2)$. A modified proof from [LiSh5] yields $f(d) = O(d)$; this implies the best bounds in the Babai-Cameron-Pálfy theorem, recently obtained by Pyber [P2].

We also settle another base conjecture, posed by Cameron in [Ca].

Conjecture 6 (Cameron, 1990): Let G be an almost simple primitive permutation group. Then $b(G) \leq c$ with known exceptions.

Here c denotes an absolute constant (not depending on G). The exceptions are A_m, S_m acting on subsets or partitions, and subspace actions of classical groups. Conjecture 6 has just been settled in [LiSh5].

Theorem 9 (Liebeck-Shalev, 1998): *Cameron's conjecture holds. Moreover, there is an absolute constant c such that, excluding the prescribed exceptions, almost all c-tuples from the permutation domain form a base for G.*

This establishes a probabilistic version of the conjecture, posed in the paper [CK] by Cameron and Kantor, where the cases $G = A_m, S_m$ are settled.

The following challenging base conjecture is still open [P1].

Conjecture 7 (Pyber, 1993): The base size of a primitive subgroup G of S_n is at most $c \log |G| / \log n$.

5 Hints of proofs

Since a subset X of a group G generates G if and only if it is not contained in any maximal subgroup M of G, the proofs of the results on random generation are intimately related with information concerning the subgroup structure of finite simple groups. More specifically, for a real number s and a finite simple group G, define

$$\zeta_G(s) = \sum_{M \max G} |G : M|^{-s}.$$

Then it is easy to see that the probability that two randomly chosen elements x, y of G do not generate G is bounded above by $\zeta_G(2)$. Hence, to prove Theorem 1 it suffices to show that $\zeta_G(2) \to 0$ as $|G| \to \infty$, which is what we do. Aschbacher's theorem for classical groups (see [A], [KLi]), and its analogs for exceptional groups (see [LiSe1], [LST]), are the main tools in this proof.

The asymptotic behavior of $\zeta_G(s)$ for other values of s is crucial in proving additional results on random generation. For example, the proof of Theorem 4, which involves counting elements of orders 2 and 3 in classical groups and in their maximal subgroups, eventually boils down to estimating $\zeta_G(66/65)$. Once Theorem 4 is proved, it serves as an essential tool in the proofs of other results, such as Theorems 2 and 3

Our results on base size rely on information concerning fixed point ratios for permutation groups. This is a classical field of research which has been very active in the past 120 years or so, since the days of Jordan [J]. Denote the number of fixed points of a permutation x by $\mathrm{fix}(x)$. The basic question is how large $\mathrm{fix}(x)$ can be, assuming x is a non-identity element of a primitive permutation group (satisfying some mild conditions). The main tool in the proof of Cameron's conjecture is the following result from [LiSh5].

THEOREM 10 (Liebeck-Shalev, 1998): *There is a constant $\epsilon > 0$ such that if G is an almost simple classical group over a field with q elements with an n-dimensional natural module, and G acts primitively on a set Ω in a non subspace action, then*

(i) $\mathrm{fix}(x)/|\Omega| < |x^G|^{-\epsilon}$ *for all elements $x \in G$ of prime order, and*

(ii) $\mathrm{fix}(x)/|\Omega| < q^{-\epsilon n}$ *for all non-trivial elements $x \in G$.*

Here $|x^G|$ denotes the size of the conjugacy class of x in G. For large n the bound in part (ii) improves the $4/3q$ upper bound of [LS] (which holds with fewer exceptions).

To demonstrate the relevance of Theorem 10 in the context of Cameron's conjecture let G be as above, and let $B(G,k)$ denote the probability that a randomly chosen k-tuple $(\omega_1, \ldots, \omega_k)$ of elements of Ω forms a base for G. Given a permutation $x \in G$, the probability the x fixes a randomly chosen letter $\omega \in \Omega$ is $\mathrm{fix}(x)/|\Omega|$. Hence the probability that x fixes $\omega_1, \ldots, \omega_k$ is $(\mathrm{fix}(x)/|\Omega|)^k$. Now, if $(\omega_1, \ldots, \omega_k)$ is not a base for G, then some element $x \in G$ of prime order fixes $\omega_1, \ldots, \omega_k$. Letting P denote the set of elements of prime order in G, and applying part (i) of Theorem 10, we obtain

$$1 - B(G,k) \leq \sum_{x \in P} (\mathrm{fix}(x)/|\Omega|)^k < \sum_{x \in P} |x^G|^{-k\epsilon}.$$

Invoking information on conjugacy classes in classical groups, one can then show that, with a suitable choice of k, the right hand side of the above inequality tends to 0 as $|G| \to \infty$; therefore $B(G,k) \to 1$.

Theorem 10 has several other applications. For example, it is used in proving Theorem 3 for classical groups. It also reduces the genus conjecture of Thompson and Guralnick (see [GT]) to the case of subspace actions of classical groups.

The interested reader is referred to the more detailed survey [Sh3] and the references therein.

References

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

[Ba] L. Babai, The probability of generating the symmetric group, *J. Comb. Th. Ser. A* 52 (1989), 148-153.

[BCP] L. Babai, P.J. Cameron and P.P. Pálfy, On the orders of primitive groups with restricted nonabelian composition factors, *J. Algebra* 79 (1982), 161-168.

[Bh] M. Bhattacharjee, The probability of generating certain profinite groups by two elements, *Israel J. Math.* 86 (1994), 311-320.

[BPSh] A. Borovik, L. Pyber and A. Shalev, Maximal subgroups in finite and profinite groups, *Trans. Amer. Math. Soc.* 348 (1996), 3745-3761.

[Ca] P.J. Cameron, Some open problems on permutation groups, in *Groups, Combinatorics and Geometry* (eds: M.W. Liebeck and J. Saxl), London Math. Soc. Lecture Note Series 165, Cambridge University Press, Cambridge, 1992, 340-350.

[CK] P.J. Cameron and W.M. Kantor, Random permutations: some group-theoretic aspects, *Combinatorics, Probability and Computing* 2 (1993), 257-262.

[D] J.D. Dixon, The probability of generating the symmetric group, *Math. Z.* 110 (1969), 199-205.

[DPSSh] J.D. Dixon, L. Pyber, Á. Seress and A. Shalev, Residual properties of free groups: a probabilistic approach, in preparation.

[ET1] P. Erdős and P. Turán, On some problems of a statistical group theory. I, *Z. Wahrscheinlichkeitstheorie Verw. Gabiete* 4 (1965), 175-186.

[GSSh] D. Gluck, Á. Seress and A. Shalev, Bases for primitive permutation groups and a conjecture of Babai, *J. Algebra* 199 (1998), 367-378.

[GK] R.M. Guralnick and W.M. Kantor, Probabilistic generation of finite simple groups, to appear.

[GKS] R.M. Guralnick, W.M. Kantor and J. Saxl, The probability of generating a classical group, *Comm. in Alg.* 22 (1994), 1395-1402.

[GLSSh] R.M. Guralnick, M.W. Liebeck, J. Saxl and A. Shalev, Random generation of finite simple groups, Preprint, 1998.

[GT] R.M. Guralnick and J.G. Thompson, Finite groups of genus zero, *J. Algebra* 131 (1990), 303-341.

[J] C. Jordan, Théorèmes sur les groupes primitifs, *J. Math. Pures Appl.* 16 (1871), 383-408.

[K] W.M. Kantor, Some topics in asymptotic group theory, in *Groups, Combinatorics and Geometry* (eds: M.W. Liebeck and J. Saxl), London Math. Soc. Lecture Note Series 165, Cambridge University Press, Cambridge, 1992, 403-421.

[KL] W.M. Kantor and A. Lubotzky, The probability of generating a finite classical group, *Geom. Ded.* 36 (1990), 67-87.

[KLi] P.B. Kleidman and M.W. Liebeck, *The Subgroup Structure of the Finite Classical Groups*, London Math. Soc. Lecture Note Series 129, Cambridge University Press, 1990.

[LS] M.W. Liebeck and J. Saxl, Minimal degrees of primitive permutation groups, with an application to monodromy groups of covers of Riemann surfaces, *Proc. London Math. Soc.* (3) 63 (1991), 266-314.

[LST] M.W. Liebeck, J. Saxl and D. Testerman, Simple subgroups of large rank in groups of Lie type, *Proc. London Math. Soc.* 72 (1996), 425-457.

[LiSe1] M.W. Liebeck and G.M. Seitz, Maximal subgroups of exceptional groups of Lie type, finite and algebraic, *Geom. Ded.* 35 (1990), 353-387.

[LiSe2] M.W. Liebeck and G.M. Seitz, On the subgroup structure of exceptional groups of Lie type, to appear in *Trans. Amer. Math. Soc.*

[LiSh1] M.W. Liebeck and A. Shalev, The probability of generating a finite simple group, *Geom. Ded.* 56 (1995), 103-113.

[LiSh2] M.W. Liebeck and A. Shalev, Classical groups, probabilistic methods, and the (2,3)-generation problem, *Annals of Math.* 144 (1996), 77-125.

[LiSh3] M.W. Liebeck and A. Shalev, Simple groups, probabilistic methods, and a conjecture of Kantor and Lubotzky, *J. Algebra* 184 (1996), 31-57.

[LiSh4] M.W. Liebeck and A. Shalev, Maximal subgroups of symmetric groups, *J. Comb. Th. Ser. A* 75 (1996), 341-352.

[LiSh5] M.W. Liebeck and A. Shalev, Permutation groups, simple groups, and probability, submitted.

[LM] F. Lübeck and G. Malle, $(2,3)$-generation of exceptional groups, to appear in *J. London Math. Soc.*

[ŁP] T. Łuczak and L. Pyber, On random generation of the symmetric group, *Combinatorics, Probability and Computing* 2 (1993), 505-512.

[M] A. Mann, Positively finitely generated groups, *Forum Math.* 8 (1996), 429-459.

[MSh] A. Mann and A. Shalev, Simple groups, maximal subgroups, and probabilistic aspects of profinite groups, *Israel. J. Math.* 96 (1997), 449-468 (Amitsur memorial issue).

[P1] L. Pyber, Asymptotic results for permutation groups, in *Groups and Computation* (eds: L. Finkelstein and W.M. Kantor), DIMACS Series on Discrete Math. and Theor. Computer Science 11, AMS 1993, 197-219.

[P2] L. Pyber, Palfy-Wolf type theorems for completely reducible subgroups of $\mathrm{GL}(n,p^a)$, in preparation.

[Sh1] A. Shalev, A theorem on random matrices and some applications, *J. Algebra* 199 (1998), 124-141.

[Sh2] A. Shalev, Random generation of simple groups by two conjugate elements, *Bull. London Math. Soc.* 29 (1997), 571-576.

[Sh3] A. Shalev, Probabilistic group theory, to appear in *Groups '97 - Bath/St Andrews*, London Math. Soc. Lecture Note Series, Cambridge University Press.

[Wa] G.E. Wall, Some applications of the Eulerian function of a finite group, *J. Austral. Math. Soc.* 2 (1961), 35-59.

[We1] T. Weigel, Residual properties of free groups, *J. Algebra* 160 (1993), 16-41.

[We2] T. Weigel, Residual properties of free groups, II, *Comm. Alg.* 20 (1992), 1395-1425.

[We3] T. Weigel, Residual properties of free groups, III, *Israel J. Math.* 77 (1992), 65-81.

[W] J.S. Wilson, Economical generating sets for finite simple groups, in *Groups of Lie type and Their Geometries* (eds: W.M. Kantor and L. Di Martino), London Math. Soc. Lecture Note Series 207, Cambridge University Press, Cambridge, 1995, 289-302.

Aner Shalev
Institute of Mathematics
Hebrew University
Jerusalem 91904, Israel

Section 3

Number Theory and Arithmetic Algebraic Geometry

In case of several authors, Invited Speakers are marked with a *.

p-Adic Analytic Spaces

Vladimir G. Berkovich[1]

Abstract. This report is a review of results in p-adic analytic geometry based on a new notion of analytic spaces. I'll explain the definition of analytic spaces, basic ideas of étale cohomology for them, an application to a conjecture of Deligne on vanishing cycles, the homotopy description of certain analytic spaces, and a relation between the étale cohomology of an algebraic variety and the topological cohomology of the associated analytic space.

1991 Mathematics Subject Classification: 14G20, 14F20, 11G25, 32P05, 32C37
Keywords and Phrases: p-adic analytic spaces, étale cohomology, vanishing cycles

§1. Introduction. At the beginning of the 1960's, J. Tate discovered p-adic uniformization of elliptic curves with totally degenerate reduction. This led him to introduce rigid analytic spaces in the framework of which the above uniformization actually takes place. Basics of rigid analytic geometry were developed by him in the paper [Ta] (released in 1961) and completed by R. Kiehl in [Ki1]-[Ki2] and L. Gerritzen and H. Grauert in [GG]. Rigid analytic spaces over a field k complete with respect to a non-trivial non-Archimedean valuation are glued from local objects, affinoid spaces, which are the maximal spectra of affinoid algebras, the algebras of topologically finite type over k. The natural topology on these spaces is totally disconnected, and one has to work with a certain Grothendieck topology instead. The framework of rigid analytic geometry enables one to construct an analog of the complex analytic theory of coherent sheaves and their cohomology, but does not allow a direct application of the intuitive idea of continuity and, in particular, of the homotopy and singular homology notions.

At the beginning of the 1970's, M. Raynaud introduced a new point of view to rigid analytic spaces. Namely, they can be considered as the generic fibres of formal schemes locally finitely presented over the ring of integers k° of k, and the category of quasi-compact quasi-separated rigid spaces is equivalent to the localization of the category of formal schemes finitely presented over k° with respect to the family of formal blow-ups (see [Ra], [BL1]-[BL2]). This provided additional algebraic tools to rigid analytic geometry, but did not make it more geometric.

[1] Supported by US-Israel Binational Science Foundation

In 1986, I found that p-adic analytic spaces, to which the homotopy and singular homology notions can be directly applied, do exist. They are retrieved through spectra of affinoid algebras, where the spectrum is a generalization of the Gelfand spectrum of a complex commutative Banach algebra and, in general, is different from the space of maximal ideals. The new definition is simpler than that of rigid analytic spaces, does not require the use of Grothendieck topologies, works over fields with trivial valuation as well, and is in a sense a natural generalization of the definition of complex analytic spaces. The main advantage of the new analytic spaces is their nice topology which makes geometrical considerations relevant and useful over p-adic fields too.

In [Hu1]-[Hu4], R. Huber develops another approach to rigid analytic (and more general adic) spaces. It is based on a different notion of the spectrum of an affinoid algebra which coincides, a posteriori, with the space of points of the topos generated by the corresponding rigid affinoid space, and whose maximal Hausdorff quotient is the spectrum we consider. The relation between various approaches to p-adic analytic geometry is explained in simple terms at the end of §2.

§2. ANALYTIC SPACES. First of all, let $\mathcal{A}$ be a commutative Banach ring with unity. (Besides affinoid algebras we are going to consider, a good example is the ring of integers $\mathbf{Z}$ endowed with the absolute value $|\ |_\infty$.) The *spectrum* $\mathcal{M}(\mathcal{A})$ of $\mathcal{A}$ is the set of all bounded multiplicative seminorms on $\mathcal{A}$, i.e., functions $|\ |: \mathcal{A} \to \mathbf{R}_+$ with $|1| = 1$, $|f+g| \le |f|+|g|$, $|fg| = |f|\cdot|g|$ and $|f| \le ||f||$. Each point $x \in \mathcal{M}(\mathcal{A})$ gives rise to a bounded character $\chi_x : \mathcal{A} \to \mathcal{H}(x)$, where $\mathcal{H}(x)$ is the completion of the fraction field of the quotient ring of $\mathcal{A}$ by the kernel of the corresponding seminorm. The image of an element $f \in \mathcal{A}$ under χ_x is denoted by $f(x)$. The spectrum $\mathcal{M}(\mathcal{A})$ is endowed with the weakest topology with respect to which all real valued functions of the form $x \mapsto |f(x)|$ are continuous. For example, if the algebra $\mathcal{A}$ contains the field of complex numbers $\mathbf{C}$ then, by Gelfand-Mazur's theorem, all of the fields $\mathcal{H}(x)$ coincide with $\mathbf{C}$ and, therefore, the spectrum $\mathcal{M}(\mathcal{A})$ is the Gelfand space of maximal ideals. A basic fact is that $\mathcal{M}(\mathcal{A})$ is always a non-empty compact space.

Let k be a non-Archimedean field, i.e., a field complete with respect to a non-Archimedean valuation which is not assumed to be non-trivial. Given positive numbers $r_1, \dots, r_n$, one sets $k\{r_1^{-1}T_1, \dots, r_n^{-1}T_n\} = \{f = \sum_\nu a_\nu T^\nu \big| |a_\nu| r^\nu \to 0 \text{ as } |\nu| \to \infty\}$. It is a commutative Banach k-algebra with the norm $||f|| = \max |a_\nu| r^\nu$. A *$k$-affinoid algebra* is a commutative Banach k-algebra $\mathcal{A}$ for which there exists an epimorphism $k\{r_1^{-1}T_1, \dots, r_n^{-1}T_n\} \to \mathcal{A}$ which is admissible in the sense that the norm on $\mathcal{A}$ is equivalent to the quotient norm. The algebras which are affinoid in the usual sense, i.e., for which such an epimorphism can be found with $r_i = 1$, $1 \le i \le n$, are said to be *strictly k-affinoid*. One shows that k-affinoid algebras are Noetherian, and all their ideals are closed. The category of *k-affinoid spaces* is, by definition, the category anti-equivalent to that of k-affinoid algebras (with bounded homomorphisms between them). To define global objects, k-analytic spaces, one uses the classical language of charts and atlases (which, by the way, can also be used to define schemes and formal schemes).

Given a k-affinoid space $X = \mathcal{M}(\mathcal{A})$, a closed subset $V \subset X$ is an *affinoid*

domain if there exists a bounded homomorphism of k-affinoid algebras $\mathcal{A} \to \mathcal{A}_V$ such that the image of $\mathcal{M}(\mathcal{A}_V)$ in X is contained in V and any bounded homomorphism $\mathcal{A} \to \mathcal{B}$ with the same property, where $\mathcal{B}$ is a K-affinoid algebra for some bigger non-Archimedean field K, factors through a unique bounded homomorphism $\mathcal{A}_V \to \mathcal{B}$. One shows that $\mathcal{A}_V$ is flat over $\mathcal{A}$, $\mathcal{M}(\mathcal{A}_V) \stackrel{\sim}{\to} V$, $\mathcal{H}(x) \stackrel{\sim}{\to} \mathcal{H}_V(x)$ for any affinoid domain V that contains a point x, and affinoid neighborhoods of x form a fundamental system of its compact neighborhoods. Affinoid domains possess other nice properties (similar to those of open affine subschemes of affine schemes) which justify the following definitions.

A family τ of subsets of a topological space X is said to be a *quasi-net* if, for each point $x \in X$, there exist $V_1, \ldots, V_n \in \tau$ such that $x \in V_1 \cap \ldots \cap V_n$ and the set $V_1 \cup \ldots \cup V_n$ is a neighborhood of x. A quasi-net τ is said to be a *net* if, for any pair $U, V \in \tau$, the family $\tau\big|_{U \cap V}$ is a quasi-net on $U \cap V$.

Let X be a locally Hausdorff topological space, and let τ be a net of compact subsets on X. A *k-affinoid atlas $\mathcal{A}$ on X with the net τ* is a map which assigns, to each $V \in \tau$, a k-affinoid algebra $\mathcal{A}_V$ and a homeomorphism $V \stackrel{\sim}{\to} \mathcal{M}(\mathcal{A}_V)$ and, to each pair $U, V \in \tau$ with $U \subset V$, a bounded homomorphism of k-affinoid algebras $\mathcal{A}_V \to \mathcal{A}_U$ that identifies $(U, \mathcal{A}_U)$ with an affinoid domain in $(V, \mathcal{A}_U)$. A *k-analytic space* is a triple $(X, \mathcal{A}, \tau)$ of the above form. A *strong morphism* of k-analytic spaces $\varphi : (X, \mathcal{A}, \tau) \to (X', \mathcal{A}', \tau')$ is a pair which consists of a continuous map $\varphi : X \to X'$, such that for each $V \in \tau$ there exists $V' \in \tau'$ with $\varphi(V) \subset V'$, and of a system of compatible morphisms of k-affinoid spaces $\varphi_{V/V'} : (V, \mathcal{A}_V) \to (V', \mathcal{A}'_{V'})$ for all pairs $V \in \tau$ and $V' \in \tau'$ with $\varphi(V) \subset V'$. One gets a category k-$\widetilde{\mathcal{A}n}$. Furthermore, a strong morphism $\varphi : (X, \mathcal{A}, \tau) \to (X', \mathcal{A}', \tau')$ is said to be a *quasi-isomorphism* if φ induces a homeomorphism between X and X' and, for any pair $V \in \tau$ and $V' \in \tau'$ with $\varphi(V) \subset V'$, $\varphi_{V/V'}$ identifies V with an affinoid domain in V'. One shows that the family of quasi-isomorphisms admits calculus of right fractions. The *category of k-analytic spaces* k-$\mathcal{A}n$ is the category of fractions of k-$\widetilde{\mathcal{A}n}$ with respect to the system of quasi-isomorphisms. If one assumes that all of the k-affinoid spaces used in the definition of k-$\mathcal{A}n$ are strictly k-affinoid, one gets the *category of strictly k-analytic spaces*. We mention several properties of k-analytic spaces.

(1) The functor $X = \mathcal{M}(\mathcal{A}) \mapsto (X, \mathcal{A}, \{X\})$ from the category of k-affinoid spaces to k-$\mathcal{A}n$ is fully faithful.

(2) Each k-analytic space X has a maximal k-affinoid atlas whose elements are called *affinoid domains* in X.

(3) A subset Y of a k-analytic space X is said to be an *analytic domain* if, for any point $y \in Y$, there exist affinoid domains $V_1, \ldots, V_n$ that are contained in Y and such that $y \in V_1 \cap \ldots \cap V_n$ and the set $V_1 \cup \ldots \cup V_n$ is a neighborhood of y in Y. An analytic domain Y has a natural structure of a k-analytic space, and the family of analytic domains gives rise to a Grothendieck topology on X, called the G-*topology*.

(4) The category k-$\mathcal{A}n$ admits fibre products and, for each non-Archimedean field K over k, there is the *ground field extension functor* $X \mapsto X \widehat{\otimes} K$.

(5) Given a point $x \in X$, there is an associated non-Archimedean field $\mathcal{H}(x)$ over k and, for for each morphism $\varphi : Y \to X$, there is a *fibre Y_x of φ at x* which

is an $\mathcal{H}(x)$-analytic space. The field $\mathcal{H}(x)$ is the completion (with respect to a valuation) of a field of transcendence degree at most $\dim(X)$ over k.

(6) For each morphism $\varphi : Y \to X$, one can define its *interior* $\mathrm{Int}(Y/X)$ and the *boundary* $\partial(Y/X) = Y \backslash \mathrm{Int}(Y/X)$ so that, if Y is an analytic domain in X, then $\mathrm{Int}(Y/X)$ coincides with the topological interior of Y in X. φ is said to be *closed* if $\partial(Y/X) = \emptyset$. φ is said to be *proper* if it is proper in the topological sense and closed in the above sense.

(7) Each point of a k-analytic space has a fundamental system of open neighborhoods which are locally compact, countable at infinity and arc-wise connected. The topological dimension of a paracompact k-analytic space is at most its dimension and, if the space is strictly k-analytic, both numbers are equal. The projective space and all its Zariski open subsets are contractible, and Tate's elliptic curve is homotopy equivalent to a circle (see also §5).

(8) One can associate with each scheme $\mathcal{X}$ of locally finite type over k a closed k-analytic space $\mathcal{X}^{\mathrm{an}}$. The scheme $\mathcal{X}$ is separated (resp. proper, resp. connected) if and only if the underlying topological space of $\mathcal{X}^{\mathrm{an}}$ is Hausdorff (resp. compact, resp. arc-wise connected) and, if $\mathcal{X}$ is separated, its dimension is equal to the topological dimension of $\mathcal{X}^{\mathrm{an}}$.

(9) Given a formal scheme $\mathfrak{X}$ locally finitely presented over k°, i.e., $\mathfrak{X}$ is a locally finite union of formal schemes of the form $\mathrm{Spf}(k^\circ\{T_1, \ldots, T_n\}/(f_1, \ldots, f_m))$, one can associate with it the generic fibre $\mathfrak{X}_\eta$, which is a paracompact strictly k-analytic space, and construct a reduction map $\pi : \mathfrak{X}_\eta \to \mathfrak{X}_s$, where $\mathfrak{X}_s$ is the closed fibre of $\mathfrak{X}$.

(10) Assume that the valuation on k is non-trivial. For each Hausdorff strictly k-analytic space X, one can provide the subset $X_0 = \{x \in X \big| [\mathcal{H}(x) : k] < \infty\}$ with the structure of a rigid analytic space, and one can construct a morphism of topoi $X_0^{\sim} \to X^{\sim}$. The functor $X \mapsto X_0$ is fully faithful and induces an equivalence between the category of paracompact strictly k-analytic spaces and that of quasi-separated rigid analytic spaces which have an admissible affinoid covering of finite type. (Both categories contain all the spaces needed in practice.)

Remark. One can represent the relation between different approaches to p-adic analytic geometry in a metaphoric way on the model of real numbers as follows. In rigid analytic geometry, one does not know about the existence of irrational numbers, but is given functions on $\mathbf{Q}$ which are restrictions of continuous functions from $\mathbf{R}$. To work in such a situation one is led to provide $\mathbf{Q}$ with a Grothendieck topology (generated by the closed intervals with rational ends). In the approach of R. Huber (and essentially in that of M. Raynaud), one works with the space of points of the topos of sheaves in the above Grothendieck topology. In the approach described here, one works with the space of real numbers $\mathbf{R}$ itself.

§3. ÉTALE COHOMOLOGY FOR ANALYTIC SPACES. The necessity of constructing étale cohomology theory for p-adic analytic spaces arose in V. Drinfeld's work ([Dr1], [Dr2]) for needs of problems related to the local Langlands conjecture (see the end of §4), and one of the main requirements was to extend étale cohomology theory of schemes. Such a theory was developed in [Ber2]. In this section we

explain some basic ideas that use the nice topology of analytic spaces and make the whole theory easier than that for schemes.

Recall that the analog of étale cohomology for complex analytic spaces is the usual topological cohomology with coefficients in sheaves, and the reason is that an étale morphism between complex analytic spaces is a local isomorphism. The topological cohomology of k-analytic spaces cannot be a good analog of étale cohomology since, for example, the projective space is contractible, and so one may try to work with a class of étale morphisms which naturally generalizes that for complex analytic spaces and coincides with it over $\mathbf{C}$. Having in mind the nice topology of analytic spaces, one is easily led to the following definition.

A morphism of k-analytic spaces $\varphi : Y \to X$ is said to be *étale* if for each point $y \in Y$ there exist open neighborhoods $\mathcal{V}$ of y and $\mathcal{U}$ of $\varphi(y)$ such that φ induces a finite étale morphism $\mathcal{V} \to \mathcal{U}$. (The latter means that, for each affinoid domain $U = \mathcal{M}(\mathcal{A})$ in $\mathcal{U}$, the preimage $\varphi^{-1}(U) = \mathcal{M}(\mathcal{B})$ is an affinoid domain and $\mathcal{B}$ is a finite étale $\mathcal{A}$-algebra.) An important fact is that a morphism $\mathcal{Y} \to \mathcal{X}$ between schemes of locally finite type over k is étale if and only if the induced morphism $\mathcal{Y}^{\mathrm{an}} \to \mathcal{X}^{\mathrm{an}}$ is étale. Another important fact is the following. For a k-analytic space X and a point $x \in X$, let $\mathrm{Fét}(X, x)$ be the category of germs of k-analytic spaces finite étale over the germ of X at x, and let $\mathrm{Fét}(\mathcal{H}(x))$ be the category of schemes finite étale over the spectrum of $\mathcal{H}(x)$ (the latter is anti-equivalent to the category of finite separable $\mathcal{H}(x)$-algebras). The remarkable fact is that there is an equivalence of categories $\mathrm{Fét}(X, x) \xrightarrow{\sim} \mathrm{Fét}(\mathcal{H}(x))$. In other words, locally over the point x étale morphisms to X correspond to finite separable extensions of $\mathcal{H}(x)$. Notice that over $\mathbf{C}$ the latter means that an étale morphism is a local isomorphism.

The *étale topology* $X_{\text{ét}}$ on a k-analytic space X is the Grothendieck topology on the category of étale morphisms $U \to X$ generated by the pretopology for which the set of coverings of $U \to X$ is formed by the families $\{U_i \xrightarrow{f_i} U\}_{i \in I}$ with $U = \cup_{i \in I} f_i(U_i)$. This topology gives rise to the étale cohomology groups $H^q(X, F)$ with coefficients in an abelian étale sheaf F. A global section of F over X has the support which is a closed subset of X and, if X is Hausdorff, the étale cohomology groups with compact support $H^q_c(X, F)$ are defined as the right derived functors of the functor of global sections with compact support. In the same way one defines, for a Hausdorff morphism $\varphi : Y \to X$, the functors $F \mapsto R^q\varphi_!F$.

Consider the morphism of sites $\pi : X_{\text{ét}} \to |X|$, where $|X|$ is the underlying topological space of X. The equivalence of categories $\mathrm{Fét}(X, x) \xrightarrow{\sim} \mathrm{Fét}(\mathcal{H}(x))$ easily implies that, for any abelian étale sheaf F, the stalk $(R^q\pi_*F)_x$ coincides with the cohomology group $H^q(G_{\mathcal{H}(x)}, F_x)$, where $G_{\mathcal{H}(x)}$ is the Galois group of $\mathcal{H}(x)$. Assume that F is torsion. By property (5) from §2, $H^q(G_{\mathcal{H}(x)}, F_x) = 0$ for q bigger than $\dim(X)$ plus the cohomological dimension of k. On the other hand, if X is paracompact, the topological dimension of X is at most $\dim(X)$. Thus, the spectral sequence of the morphism π implies that $H^q(X, F) = 0$ for q bigger than $2 \cdot \dim(X)$ plus the cohomological dimension of k. In a similar way, using properties of cohomology of topological spaces and of profinite groups one describes, for a Hausdorff morphism $\varphi : Y \to X$, the stalks of the sheaves $R^q\varphi_!F$ in terms of the cohomology groups with compact support of the fibres of φ. The

proof of the corresponding fact for schemes is highly non-trivial.

Among results of [Ber2] (and [Ber4]) are the invariance of cohomology under algebraically closed extensions of the ground field, a Poincaré Duality theorem, a cohomological purity theorem, a base change theorem for cohomology with compact support, a smooth base change theorem, and comparison theorems. The latter state that, given a compactifiable morphism (resp. a morphism of finite type) $\varphi : \mathcal{Y} \to \mathcal{X}$ between schemes of locally finite type over k and an étale abelian sheaf $\mathcal{F}$ on $\mathcal{Y}$ which is torsion (resp. constructible with torsion orders prime to $\mathrm{char}(k)$), there are canonical isomorphisms $(R^q\varphi_!\mathcal{F})^{\mathrm{an}} \xrightarrow{\sim} R^q\varphi_!^{\mathrm{an}}\mathcal{F}^{\mathrm{an}}$ (resp. $(R^q\varphi_*\mathcal{F})^{\mathrm{an}} \xrightarrow{\sim} R^q\varphi_*^{\mathrm{an}}\mathcal{F}^{\mathrm{an}}$).

By the way, the notion of a *smooth* morphism we work with is as follows. It is a morphism $Y \to X$ which factors locally through an étale morphism $Y \to X \times \mathbf{A}^d$. In [Ber3], we also proved that if k is algebraically closed and X is a compact quasi-algebraic k-analytic space, i.e., X is a finite union of affinoid domains isomorphic to affinoid domains in the analytification of a scheme, then for any integer n prime to $\mathrm{char}(\widetilde{k})$, the characteristic of the residue field $\widetilde{k}$ of k, the cohomology groups $H^q(X, \mathbf{Z}/n\mathbf{Z})$ are finite.

In [Hu2]-[Hu4], R. Huber develops étale cohomology in the framework of his adic spaces. Besides the results mentioned above, he got finiteness results which imply, for example, that in the case, when k is of characteristic zero, the above fact is true without the assumption that X is quasi-algebraic.

§4. VANISHING CYCLES FOR FORMAL SCHEMES. In this section we describe an application of étale cohomology of analytic spaces to a conjecture of Deligne from [Del]. Let $\mathcal{X}$ be a scheme of finite type over a Henselian discrete valuation ring R, $\mathcal{Y}$ a subscheme of the closed fibre $\mathcal{X}_s$ of $\mathcal{X}$, l a prime different from the characteristic of the residue field of R. The conjecture states that (a) the restrictions of the vanishing cycles sheaves $R^q\Psi_\eta(\mathbf{Q}_l)$ of $\mathcal{X}$ to the subscheme $\mathcal{Y}$ depends only on the formal completion $\widehat{\mathcal{X}}_{/\mathcal{Y}}$ of $\mathcal{X}$ along $\mathcal{Y}$ and, in particular, the automorphism group of $\widehat{\mathcal{X}}_{/\mathcal{Y}}$ acts on them, and (b) there exists an ideal of definition of $\widehat{\mathcal{X}}_{/\mathcal{Y}}$ such that any automorphism of $\widehat{\mathcal{X}}_{/\mathcal{Y}}$ trivial modulo this ideal acts trivially on the above sheaves.

Some partial results were obtained earlier by J.-L. Brylinski in [Bry] (the case when R is of mixed characteristic, $\mathcal{X}$ is of dimension one over R and $\mathcal{Y}$ is a closed point of $\mathcal{X}_s$), G. Laumon in [La] and the author in [Ber6] (the case when R is equicharacteristic and $\mathcal{Y}$ is a closed point of $\mathcal{X}_s$), and in [Ber3] (the case when $\mathcal{Y}$ is an open subscheme of $\mathcal{X}_s$). We describe here the results from [Ber7] which give a positive answer in the general case.

Let k be a field complete with respect to a discrete valuation (which is not assumed to be non-trivial). A formal scheme over k° is said to be *special* if it is a locally finite union of affine formal schemes of the form $\mathrm{Spf}(A)$, where A is a quotient of the adic ring $k^\circ\{T_1, \ldots, T_n\}[[S_1, \ldots, S_m]]$ by an ideal. (All ideals of that ring are closed in the adic topology.) Given a special formal scheme $\mathfrak{X}$, its *closed fibre* is the scheme of locally finite type over $\widetilde{k}$, $(\mathfrak{X}, \mathcal{O}_{\mathfrak{X}}/\mathcal{J})$, where $\mathcal{J}$ is an ideal of definition of $\mathfrak{X}$ that contains the maximal ideal of k°. Due to P. Berthelot,

one can associate with $\mathfrak{X}$ its *generic fibre* $\mathfrak{X}_\eta$, which is a paracompact strictly k-analytic space, and a *reduction map* $\pi : \mathfrak{X}_\eta \to \mathfrak{X}_s$ so that, for any subscheme $\mathcal{Y} \subset \mathfrak{X}_s$, there is a canonical isomorphism $(\mathfrak{X}_{/\mathcal{Y}})_\eta \xrightarrow{\sim} \pi^{-1}(\mathcal{Y})$, where $\mathfrak{X}_{/\mathcal{Y}}$ is the formal completion of $\mathfrak{X}$ along $\mathcal{Y}$ (it is also a special formal scheme). In [Ber7] we constructed a *vanishing cycles functor* Ψ_η from the category of étale sheaves on $\mathfrak{X}_\eta$ to the category of étale sheaves on $\mathfrak{X}_{\overline{s}}$, where $\mathfrak{X}_{\overline{s}}$ is the lift of $\mathfrak{X}_s$ to the algebraic closure of $\widetilde{k}$, and proved the following results.

Theorem 1. *Given a scheme $\mathcal{X}$ of finite type over a local Henselian ring with the completion k°, a subscheme $\mathcal{Y} \subset \mathcal{X}_s$ and an étale abelian constructible sheaf $\mathcal{F}$ on $\mathcal{X}_\eta$ with torsion orders prime to* $\mathrm{char}(\widetilde{k})$*, there are canonical isomorphisms $(R^q\Psi_\eta\mathcal{F})\big|_{\overline{\mathcal{Y}}} \xrightarrow{\sim} R^q\Psi_\eta(\widehat{\mathcal{F}}_{/\mathcal{Y}})$, where $\widehat{\mathcal{F}}_{/\mathcal{Y}}$ is the pullback of $\mathcal{F}$ on $(\widehat{\mathcal{X}}_{/\mathcal{Y}})_\eta$.*

In [Hu4], a similar result is proven for any special formal scheme (instead of $\mathcal{X}$) under the assumption that the characteristic of k is zero.

Theorem 1 gives a precise meaning to the part (a) of Deligne's conjecture and implies that, given a second scheme $\mathcal{X}'$ of finite type over k°, a subscheme $\mathcal{Y}' \subset \mathcal{X}'_s$ and an integer n prime $\mathrm{char}(\widetilde{k})$, any morphism of formal schemes $\varphi : \widehat{\mathcal{X}}'_{/\mathcal{Y}'} \to \widehat{\mathcal{X}}_{/\mathcal{Y}}$ induces a homomorphism $\theta_n^q(\varphi)$ from the pullback of $(R^q\Psi_\eta(\mathbf{Z}/n\mathbf{Z})_{\mathcal{X}_\eta})\big|_{\overline{\mathcal{Y}}}$ to $(R^q\Psi_\eta(\mathbf{Z}/n\mathbf{Z})_{\mathcal{X}'_\eta})\big|_{\overline{\mathcal{Y}}'}$. In particular, given a prime l different from $\mathrm{char}(\widetilde{k})$, the automorphism group of $\widehat{\mathcal{X}}_{/\mathcal{Y}}$ acts on $(R^q\Psi_\eta(\mathbf{Q}_l)_{\mathcal{X}_\eta})\big|_{\overline{\mathcal{Y}}}$.

Theorem 2. *(i) Given $\widehat{\mathcal{X}}_{/\mathcal{Y}}$, $\widehat{\mathcal{X}}'_{/\mathcal{Y}'}$ and n as above, there exists an ideal of definition $\mathcal{J}'$ of $\widehat{\mathcal{X}}'_{/\mathcal{Y}'}$ such that for any pair of morphisms $\varphi, \psi : \widehat{\mathcal{X}}'_{/\mathcal{Y}'} \to \widehat{\mathcal{X}}_{/\mathcal{Y}}$, which coincide modulo $\mathcal{J}'$, one has $\theta_n^q(\varphi) = \theta_n^q(\psi)$.*

(ii) Given $\widehat{\mathcal{X}}_{/\mathcal{Y}}$ and l as above, there exists an ideal of definition $\mathcal{J}$ of $\widehat{\mathcal{X}}_{/\mathcal{Y}}$ such that any automorphism of $\widehat{\mathcal{X}}_{/\mathcal{Y}}$, trivial modulo $\mathcal{J}$, acts trivially on $(R^q\Psi_\eta(\mathbf{Q}_l)_{\mathcal{X}_\eta})\big|_{\overline{\mathcal{Y}}}$.

The proof of Theorem 2 uses a result from [Ber3] on the continuity of the action of a topological group on the étale cohomology groups of a k-analytic space if the original action of the group on the space is continuous.

The results from [Ber3] and [Ber7], described above, have been used by G. Faltings ([Fa]) and M. Harris ([Ha]) in their work on a conjecture of V. Drinfeld, and by M. Harris and R. Taylor ([HT]) in their work on the local Langlands conjecture over a p-adic field.

§5. The homotopy structure of analytic spaces. In this section we describe algebraic and homotopy topology results from [Ber8] obtained in an attempt to prove local contractibility of analytic spaces. To simplify the exposition, we do not formulate the results in the strongest possible form.

A morphism $\varphi : \mathfrak{Y} \to \mathfrak{X}$ between formal schemes locally finitely presented over k° is said to be *poly-stable* if locally in the étale topology it is of the form $\mathrm{Spf}(B_0\widehat{\otimes}_A \ldots \widehat{\otimes}_A B_p) \to \mathrm{Spf}(A)$, where each B_i is of the form $A\{T_0, \ldots, T_n\}/(T_0 \cdot \ldots \cdot T_n - a)$ with $a \in A$. A *poly-stable fibration of length l over k°* is a sequence of poly-stable morphisms $\underline{\mathfrak{X}} = (\mathfrak{X}_l \to \mathfrak{X}_{l-1} \to \ldots \to \mathfrak{X}_1 \to \mathfrak{X}_0 = \mathrm{Spf}(k^\circ))$. Such

objects form a category in the evident way. To take into account morphisms which are non-trivial on the ground field, we introduce a category $\mathcal{P}stf_l^{\text{ét}}$ whose objects are pairs $(k, \underline{\mathfrak{X}})$, where k is a non-Archimedean field and $\underline{\mathfrak{X}}$ is a poly-stable fibration of length l over k°, and morphisms $(K, \underline{\mathfrak{Y}}) \to (k, \underline{\mathfrak{X}})$ are pairs consisting of an isometric embedding of fields $k \hookrightarrow K$ and an étale morphism of poly-stable fibrations over K°, $\underline{\mathfrak{Y}} \to \underline{\mathfrak{X}} \widehat{\otimes}_{k^\circ} K^\circ$. (For brevity the pair $(k, \underline{\mathfrak{X}})$ is denoted by $\underline{\mathfrak{X}}$.)

Consider first the case when the valuation on k is trivial, i.e., all the formal schemes considered are in fact schemes of locally finite type over k. For such a (reduced) scheme $\mathcal{X}$, we set $\mathcal{X}^{(0)} = \mathcal{X}$ and, for $i \geq 0$, denote by $\mathcal{X}^{(i+1)}$ the non-normality locus of $\mathcal{X}^{(i)}$. The irreducible components of the locally closed subsets $\mathcal{X}^{(i)} \backslash \mathcal{X}^{(i+1)}$ are called *strata of* $\mathcal{X}$. One shows that, given a poly-stable fibration $\underline{\mathcal{X}} = (\mathcal{X}_l \to \ldots \to \mathcal{X}_1 \to \mathcal{X}_0 = \mathrm{Spec}(k))$, the closure of any stratum of the scheme $\mathcal{X}_l$ is a union of strata, and one associates with $\underline{\mathcal{X}}$ a simplicial set $C(\underline{\mathcal{X}})$ which encodes combinatorics of mutual inclusions between strata. (The construction of the latter is too involved to be given here, but in the case, when $\mathcal{X}_l$ is smooth and connected, $C(\underline{\mathcal{X}})$ is a point.) In this way one gets a functor C from $\mathcal{P}stf_l^{\text{ét}}$ to the category of simplicial sets that takes a poly-stable fibration $\underline{\mathfrak{X}}$ to the simplicial set $C(\underline{\mathfrak{X}}_s)$ associated with the closed fibre of $\underline{\mathfrak{X}}$. Its composition with the geometric realization functor gives a functor $|C|$ from $\mathcal{P}stf_l^{\text{ét}}$ to the category of locally compact spaces.

THEOREM 1. *For every poly-stable fibration* $\underline{\mathfrak{X}} = (\mathfrak{X}_l \stackrel{f_{l-1}}{\to} \ldots \stackrel{f_1}{\to} \mathfrak{X}_1)$ *of length* l*, one can construct a proper strong deformation retraction* $\Phi : \mathfrak{X}_{l,\eta} \times [0, l] \to \mathfrak{X}_{l,\eta} : (x, t) \mapsto x_t$ *of* $\mathfrak{X}_{l,\eta}$ *to a closed subset* $S(\underline{\mathfrak{X}})$*, the* skeleton of $\underline{\mathfrak{X}}$*, so that the following holds:*

(i) $(x_t)_{t'} = x_{\max(t,t')}$ *for all* $0 \leq t, t' \leq l$*;*

(ii) $f_{l-1,\eta}(x_t) = f_{l-1,\eta}(x)_{t-1}$ *for all* $1 \leq t \leq l$*;*

(iii) the homotopy Φ *induces a strong deformation retraction of each Zariski open subset* $\mathcal{U}$ *of* $\mathfrak{X}_{l,\eta}$ *to* $S(\underline{\mathfrak{X}}) \cap \mathcal{U}$*; if* $\mathfrak{X}_{l,\eta}$ *is normal and* $\mathcal{U}$ *is dense, the intersection coincides with* $S(\underline{\mathfrak{X}})$*;*

(iv) given a morphism $\varphi : \underline{\mathfrak{Y}} \to \underline{\mathfrak{X}}$ *in* $\mathcal{P}stf_l^{\text{ét}}$*, one has* $\varphi_{l,\eta}(y_t) = \varphi_{l,\eta}(y)_t$*.*

The latter property implies that the correspondence $\underline{\mathfrak{X}} \mapsto S(\underline{\mathfrak{X}})$ is a functor from $\mathcal{P}stf_l^{\text{ét}}$ to the category of locally compact spaces.

THEOREM 2. *There is a canonical isomorphism of functors* $|C| \stackrel{\sim}{\to} S$*.*

The simplest consequence of Theorems 1 and 2 tells that the analytification of any Zariski open subset of a proper scheme with good reduction is contractible. In the case of the Drinfeld upper half-plane Ω^d over a local non-Archimedean field K, which is the generic fibre of a formal scheme $\widehat{\Omega}^d$ (see [Dr2]), the space $|C(\widehat{\Omega}^d)|$ is the Bruhat-Tits building of the group $\mathrm{SL}_d(K)$. The embedding of the latter in Ω^d was used in [Ber5] in the proof of the fact that the group of analytic automorphisms of Ω^d coincides with $\mathrm{PGL}_d(K)$.

Theorems 1 and 2 and results of J. de Jong on alterations from [deJ2]-[deJ3] are used to prove the following results.

THEOREM 3. *Assume that the valuation on* k *is non-trivial. Let* X *be a* k*-analytic space* locally embeddable in a smooth space*, i.e., each point of* X *has*

an open neighborhood isomorphic to a strictly k-analytic domain in a smooth k-analytic space (for example, it is true if X is a smooth k-analytic space.) Then X is locally contractible.

Let X be a separated connected k-analytic space locally embeddable in a smooth space. Theorem 3 implies that X has a universal covering, which is a strictly k-analytic space and is a Galois covering of X with the Galois group isomorphic to the fundamental group of the underlying topological space $|X|$. Furthermore, if X is paracompact, the cohomology groups $H^q(|X|, \mathbf{Z})$ (which are the same as those of the associated rigid analytic space) coincide with the singular cohomology groups.

Theorem 4. *Let $\mathcal{X}$ be a separated scheme of finite type over a non-Archimedean field k. Then*

(i) the groups $H^i(|\mathcal{X}^{\mathrm{an}}|, \mathbf{Z})$ are finitely generated;

(ii) there exists a finite separable extension k' of k such that for any non-Archimedean field K over k one has $H^i(|(\mathcal{X} \otimes k')^{\mathrm{an}}|, \mathbf{Z}) \xrightarrow{\sim} H^i(|(\mathcal{X} \otimes K)^{\mathrm{an}}|, \mathbf{Z})$.

§6. An analytic analog of Tate's conjecture over finite and local fields. This section is a report on the work in progress [Ber9]. Assume that k is a finite or a local non-Archimedean field. (Finite fields are considered as non-Archimedean ones endowed with the trivial valuation.) For a separated scheme $\mathcal{X}$ of finite type over k, we set $\overline{\mathcal{X}} = \mathcal{X} \otimes k^a$, where k^a is an algebraic closure of k, and denote by $\overline{\mathcal{X}}^{\mathrm{an}}$ the $\widehat{k^a}$-analytic space $(\mathcal{X} \widehat{\otimes} \widehat{k^a})^{\mathrm{an}}$. Let l be a prime different from $\mathrm{char}(k)$. The representation of the Galois group G of k^a on the l-adic étale cohomology groups $H^i(\overline{\mathcal{X}}, \mathbf{Q}_l)$ is continuous and, by Theorem 4 from §5, on the groups $H^i(|\overline{\mathcal{X}}^{\mathrm{an}}|, \mathbf{Z})$ is smooth in the sense that the stabilizer of any element is open in G.

The homomorphisms $H^i(|\overline{\mathcal{X}}^{\mathrm{an}}|, \mathbf{Z}) \to H^i(|\overline{\mathcal{X}}^{\mathrm{an}}|, \mathbf{Z}/l^n\mathbf{Z}) \to H^i(\overline{\mathcal{X}}^{\mathrm{an}}, \mathbf{Z}/l^n\mathbf{Z})$ and the isomorphism of the comparison theorem $H^i(\overline{\mathcal{X}}, \mathbf{Z}/l^n\mathbf{Z}) \xrightarrow{\sim} H^i(\overline{\mathcal{X}}^{\mathrm{an}}, \mathbf{Z}/l^n\mathbf{Z})$ give rise to a homomorphism $H^i(|\overline{\mathcal{X}}^{\mathrm{an}}|, \mathbf{Z}) \to H^i(\overline{\mathcal{X}}, \mathbf{Q}_l)$. Since it is Galois equivariant, its image is contained in $H^i(\overline{\mathcal{X}}, \mathbf{Q}_l)^{\mathrm{sm}}$, where for an l-adic representation V of G we denote by V^{sm} the subspace consisting of the elements with open stabilizer in G. The above homomorphism gives rise to a homomorphism $H^i(|\mathcal{X}^{\mathrm{an}}|, \mathbf{Z}) \to H^i(\overline{\mathcal{X}}, \mathbf{Q}_l)$ whose image is contained in $H^i(\overline{\mathcal{X}}, \mathbf{Q}_l)^G$.

If k is a finite field, let F be the Frobenius automorphism of k^a. Otherwise, let F be a fixed element of G that lifts the Frobenius of the residue field of k. For an l-adic representation V of G, let V_μ denote the maximal F-invariant subspace of V, where all eigenvalues of F are roots of unity. One evidently has $V^{\mathrm{sm}} \subset V_\mu$.

Theorem. $H^i(|\overline{\mathcal{X}}^{\mathrm{an}}|, \mathbf{Z}) \otimes \mathbf{Q}_l \xrightarrow{\sim} H^i(\overline{\mathcal{X}}, \mathbf{Q}_l)_\mu$.

The first corollary justifies the title of this section.

Corollary 1. $H^i(|\mathcal{X}^{\mathrm{an}}|, \mathbf{Z}) \otimes \mathbf{Q}_l \xrightarrow{\sim} H^i(\overline{\mathcal{X}}, \mathbf{Q}_l)^G$.

Corollary 2. $H^i_c(|\overline{\mathcal{X}}^{\mathrm{an}}|, \mathbf{Z}) \otimes \mathbf{Q}_l \xrightarrow{\sim} H^i_c(\overline{\mathcal{X}}, \mathbf{Q}_l)_\mu$ *and* $H^i_c(|\mathcal{X}^{\mathrm{an}}|, \mathbf{Z}) \otimes \mathbf{Q}_l \xrightarrow{\sim} H^i_c(\overline{\mathcal{X}}, \mathbf{Q}_l)^G$.

Notice that the above results imply that $V_\mu = V^{\mathrm{sm}}$ for $V = H^i(\overline{\mathcal{X}}, \mathbf{Q}_l)$ and $H^i_c(\overline{\mathcal{X}}, \mathbf{Q}_l)$. Recall also that in the case of positive characteristic of k it is not yet known that the dimensions of the groups $H^i(\overline{\mathcal{X}}, \mathbf{Q}_l)$ and $H^i_c(\overline{\mathcal{X}}, \mathbf{Q}_l)$ do not depend on l.

References

[Ber1] Berkovich, V. G.: *Spectral theory and analytic geometry over non-Archimedean fields*, Mathematical Surveys and Monographs, vol. 33, American Mathematical Society, Providence, R.I., 1990.

[Ber2] Berkovich, V. G.: *Étale cohomology for non-Archimedean analytic spaces*, Publ. Math. IHES **78** (1993), 5-161.

[Ber3] Berkovich, V. G.: *Vanishing cycles for formal schemes*, Invent. Math. **115** (1994), 539-571.

[Ber4] Berkovich, V. G.: *On the comparison theorem for étale cohomology of non-Archimedean analytic spaces*, Israel J. Math. **92** (1995), 45-60.

[Ber5] Berkovich, V. G.: *The automorphism group of the Drinfeld half-plane*, C. R. Acad. Sci. Paris Sér. I Math. **321** (1995), 1127-1132.

[Ber6] Berkovich, V. G.: *Vanishing cycles for non-Archimedean analytic spaces*, J. Amer. Math. Soc. **9** (1996), 1187-1209.

[Ber7] Berkovich, V. G.: *Vanishing cycles for formal schemes. II*, Invent. Math. **125** (1996), 367-390.

[Ber8] Berkovich, V. G.: *Smooth p-adic analytic spaces are locally contractible*, Preprint, March 1998.

[Ber9] Berkovich, V. G.: *An analytic analog of Tate's conjecture over finite and local fields*, (in preparation).

[BGR] Bosch, S; Güntzer, U.; Remmert, R.: *Non-Archimedean analysis. A systematic approach to rigid analytic geometry*, Grundlehren der Mathematischen Wissenschaften, Bd. 261, Springer, Berlin-Heidelberg-New York, 1984.

[BL1] Bosch, S; Lütkebohmert, W.: *Formal and rigid geometry. I. Rigid spaces*, Math.Ann. **295** (1993), 291-317.

[BL2] Bosch, S; Lütkebohmert, W.: *Formal and rigid geometry. II. Flattening techniques*, Math.Ann. **296** (1993), 403-429.

[Bry] Brylinski, J.-L.: *Un lemme sur les cycles évanescents en dimension relative 1*, Ann. Scient. Éc. Norm. Sup. **19** (1986), 460-467.

[Del] Deligne, P.: *Sur les représentations l-adiques liées aux formes modulaires*, Letter to Piatetski-Shapiro, 1973.

[Dr1] Drinfeld, V.G.: *Elliptic modules*, Math. USSR Sbornik, **23** (1974), 561-592.

[Dr2] Drinfeld, V.G.: *Coverings of p-adic symmetric domains*, Funct. Anal. Appl. **10** (1976), 107-115.

[Fa] Faltings, G.: *The trace formula and Drinfeld's upper half-plane*, Duke Math. J. **76** (1994), 467-481.

[GG] Gerritzen, L.; Grauert, H.: *Die Azyklizität der affinoiden Überdeckungen*, in Global Analysis (Papers in Honor of K. Kodaira), Univ. Tokyo Press, Tokyo, 1969, 159-184.

[Ha] Harris, M.: *Supercuspidal representations in the cohomology of Drinfeld's upper half plane: elaboration of Carayol's program*, Invent. Math. **129** (1997), 75-120.

[HT] Harris, M., Taylor. R.: *On the geometry and cohomology of Kottwitz's simple Shimura varieties*, Preprint, June 1998.

[Hu1] Huber, R.: *A generalization of formal schemes and rigid analytic varieties*, Math. Z. **217** (1994), 513-551.

[Hu2] Huber, R.: *Étale Cohomology of Rigid Analytic Varieties and Adic Spaces*, Aspects of Mathematics, Vol. 30, Vieweg, 1996.

[Hu3] Huber, R.: *A finiteness result for the compactly supported cohomology of rigid analytic varieties*, J. Alg. Geom. **7** (1998), 313-357.

[Hu4] Huber, R.: *A finiteness result for direct image sheaves on the étale site of rigid analytic varieties*, J. Alg. Geom. **7** (1998), 359-403.

[deJ1] de Jong, A. J.: *Étale fundamental group of non-Archimedean analytic spaces*, Compositio Math. **97** (1995), 89-118.

[deJ2] de Jong, A. J.: *Smoothness, semi-stability and alterations*, Publ. Math. IHES **83** (1996), 51-93.

[deJ3] de Jong, A. J.: *Families of curves and alterations*, Ann. Inst. Fourier (Grenoble), **47** (1997), no. 2, 599-621.

[Ki1] Kiehl, R.: *Der Endlichkeitssatz für eigentliche Abbildungen in der nichtarchimedischen Funktionentheorie*, Invent. Math. **2** (1967), 191-214.

[Ki2] Kiehl, R.:*Theorem A und Theorem B in der nichtarchimedischen Funktionentheorie*, Invent. Math. **2** (1967), 256-273.

[La] Laumon, G.: *Charactéristique d'Euler-Poincaré et sommes exponentielles*, Thése, Université de Paris-Sud, Orsay, 1983.

[Ra] Raynaud, M. *Géométrie analytique rigide d'après Tate, Kiehl, ...*, Bull. Soc. Math. France, Mém. No. 39-40 (1974), 319-327.

[Ta] Tate, J.: *Rigid analytic spaces*, Invent. Math. **12** (1971), 257-289.

Vladimir G. Berkovich
Dept. of Theoretical Mathematics
The Weizmann Institute of Science
P.O.B. 26, 76100 Rehovot
ISRAEL
vova@wisdom.weizmann.ac.il

Représentations p-Adiques d'un Corps Local

Pierre Colmez[1]

Abstract. We discuss applications of the theory of (φ,Γ)-modules to the study of p-adic representations of the Galois group of a local field and in particular to Iwasawa theory and explicit reciprocity laws.

Notations

On fixe une clôture algébrique $\overline{\mathbf{Q}}_p$ de $\mathbf{Q}_p$ et un système compatible $\varepsilon = (1, \varepsilon^{(1)}, \dots, \varepsilon^{(n)}, \dots)$ de racines de l'unité avec $\varepsilon^{(1)} \neq 1$ et $(\varepsilon^{(n+1)})^p = \varepsilon^{(n)}$ si $n \in \mathbf{N}$ de telle sorte que $\varepsilon^{(n)}$ est une racine primitive p^n-ième de l'unité si $n \in \mathbf{N}$. Si K est une extension finie de $\mathbf{Q}_p$, on note $\mathscr{G}_K$ le groupe de Galois $\mathrm{Gal}(\overline{\mathbf{Q}}_p/K)$ et $\mathscr{H}_K \subset \mathscr{G}_K$ le noyau du caractère cyclotomique χ. On pose aussi $\Gamma_K = \mathscr{G}_K/\mathscr{H}_K$ de telle sorte que Γ_K est le groupe de Galois de l'extension cyclotomique $K_\infty = \cup_{n\in\mathbf{N}} K_n$ de K, où l'on a noté K_n le corps $K(\varepsilon^{(n)})$ si $n \in \mathbf{N}$.

Un $\mathbf{Q}_p$-espace vectoriel de dimension finie muni d'une action de $\mathscr{H}_K$ (resp. $\mathscr{G}_K$) est appelé une représentation p-adique de $\mathscr{H}_K$ (resp. $\mathscr{G}_K$). Si V est une représentation p-adique de $\mathscr{G}_K$ et $k \in \mathbf{Z}$, on note $V(k)$ la tordue de V par la puissance k-ième du caractère cyclotomique.

I Introduction

Soit G un groupe topologique (comme $\mathscr{H}_K$ ou $\mathscr{G}_K$). Pour mettre un peu d'ordre dans les représentations p-adiques de G, on dispose d'une stratégie, introduite et amplement utilisée par Fontaine, qui consiste à construire des $\mathbf{Q}_p$-algèbres topologiques munies d'une action continue de G et de structures additionnelles respectées par cette action. Chacune de ces algèbres B permet de découper dans l'ensemble des représentations p-adiques de G celles qui sont B-admissibles (i.e. qui deviennent triviales quand on étend les scalaires à B). Si V est une représentation B-admissible de $\mathscr{G}_K$, le B^G-module $(B\otimes V)^G$ est libre de rang $\dim_{\mathbf{Q}_p} V$ et est muni de toutes les structures additionnelles de B respectées par l'action de $\mathscr{G}_K$. Ceci permet d'associer aux représentations de G des invariants plus maniables (en général des objets provenant de l'algèbre linéaire) et, si l'anneau B est assez fin (i.e a suffisamment de structures respectées par G), de classifier les représentations B-admissibles en termes de ces invariants. Cette approche a l'avantage de ramener l'étude de toutes les représentations B-admissibles à celle de l'anneau B.

[1]Recherche financée par le C.N.R.S

Si on injecte dans cette stratégie l'idée, utilisée avec profit par Tate[2] et Sen[3], selon laquelle on a intérêt[4] à dévisser la situation en regardant $\mathscr{G}_K$ comme une extension de Γ_K par $\mathscr{H}_K$ et la théorie du corps des normes de Fontaine et Wintenberger[5] qui associe à l'extension K_∞/K un corps local $\mathbf{E}_K$ de caractéristique p, on aboutit à la théorie des (φ,Γ)-modules[6]. Le point crucial de cette théorie est que l'on peut reconstruire une représentation V de $\mathscr{G}_K$ à partir de son (φ,Γ_K)-module $D(V)$ qui est a priori un objet beaucoup plus maniable[7] et que l'on doit donc être capable de lire sur $D(V)$ toutes les propriétés de V. Dans ce texte, nous donnons quelques applications de ce principe et en particulier la construction d'une vaste généralisation de l'isomorphisme de Coleman et de l'exponentielle de Perrin-Riou qui devrait être utile pour l'étude des fonctions-L p-adiques des motifs.

II Les anneaux $\widetilde{\mathbf{E}}$ et $\widetilde{\mathbf{A}}^+$

Soit $\mathbf{C}_p$ le complété de $\overline{\mathbf{Q}}_p$ pour la topologie p-adique. Soit $\widetilde{\mathbf{E}}$ l'ensemble des suites $x=(x^{(0)},\dots,x^{(n)},\dots)$ d'éléments de $\mathbf{C}_p$ vérifiant $(x^{(n+1)})^p=x^{(n)}$. On munit $\widetilde{\mathbf{E}}$ des lois $+$ et $\cdot$ définies par $x+y=s$ où $s^{(n)}=\lim\limits_{m\to+\infty}(x^{(n+m)}+y^{(n+m)})^{p^m}$ et $x\cdot y=t$, avec $t^{(n)}=x^{(n)}y^{(n)}$, ce qui fait de $\widetilde{\mathbf{E}}$ un corps de caractéristique p algébriquement clos et complet pour la valuation $v_{\mathbf{E}}$ définie par $v_{\mathbf{E}}(x)=v_p(x^{(0)})$. On note $\widetilde{\mathbf{E}}^+$ l'anneau des entiers de $\widetilde{\mathbf{E}}$. Soit $\widetilde{\mathbf{A}}^+=W(\widetilde{\mathbf{E}}^+)$ l'anneau des vecteurs de Witt à coefficients dans $\widetilde{\mathbf{E}}^+$ [8]. Si $x\in\widetilde{\mathbf{E}}^+$, soit $[x]$ son représentant de Teichmüller dans $\widetilde{\mathbf{A}}^+$. Notre système ε de racines de l'unité peut être vu comme un élément de $\widetilde{\mathbf{E}}$, ce qui nous permet d'introduire les éléments $\pi=[\varepsilon]-1$ et $\omega=\frac{\pi}{\varphi^{-1}(\pi)}=1+[\varepsilon^{\frac{1}{p}}]+\cdots+[\varepsilon^{\frac{p-1}{p}}]$ de $\widetilde{\mathbf{A}}^+$. Tous les anneaux que nous aurons à considérer dans ce texte s'obtiennent à partir de l'anneau $\widetilde{\mathbf{A}}^+$ en introduisant plus ou moins de dénominateurs en p ou ω et en complétant [9].

---

[2] J. Tate, dans "*Proc. of a conf. on local fields*", Driebergen, 158-183, Springer 1967.

[3] S. Sen, Inv. Math. 62, 89-116, 1980.

[4] Si l'anneau B est assez gros, les représentations de $\mathscr{H}_K$ sont automatiquement B-admissibles et on est ramené à étudier l'anneau $B^{\mathscr{H}_K}$. C'est ce qu'a remarqué Sen dans le cas $B=\mathbf{C}_p$.

[5] J.-P. Wintenberger, Ann. Sci. E.N.S. 16, 59-89, 1983.

[6] J.-M. Fontaine, dans "*The Grothendieck Festschrift*", vol II, 249-309, Birkhaüser 1991.

[7] C'est un espace vectoriel de dimension finie sur un corps local de dimension 2 muni de deux opérateurs semi-linéaires commutant entre eux

[8] L'anneau $\widetilde{\mathbf{E}}^+$ est habituellement noté R ou $\mathscr{R}$ dans la théorie des périodes p-adiques et $\widetilde{\mathbf{A}}^+$ est souvent noté $\mathbf{A}_{\mathrm{inf}}$

[9] L'application qui à $\sum_{n=0}^{+\infty}p^n[x_n]$ associe $\sum_{n=0}^{+\infty}p^nx_n^{(0)}$ est un morphisme surjectif d'anneaux de $\widetilde{\mathbf{A}}^+$ sur $\mathscr{O}_{\mathbf{C}_p}$ dont le noyau est l'idéal engendré par ω qui est donc premier

III $\mathbf{B}_{\mathrm{dR}}$ et les représentations de de Rham

On note $\mathbf{B}_{\mathrm{dR}}^{+}$ le complété de $\widetilde{\mathbf{B}}^{+} = \widetilde{\mathbf{A}}^{+}[\frac{1}{p}]$ pour la topologie ω-adique. Cet anneau peut aussi s'obtenir en complétant $\overline{\mathbf{Q}}_p$ pour une topologie adéquate[10]. L'anneau $\mathbf{B}_{\mathrm{dR}} = \mathbf{B}_{\mathrm{dR}}^{+}[\frac{1}{\omega}]$ est le corps des fractions de $\mathbf{B}_{\mathrm{dR}}^{+}$ et est muni d'une filtration décroissante stable par l'action de Galois et définie par $\mathrm{Fil}^{i}\mathbf{B}_{\mathrm{dR}} = \omega^{i}\mathbf{B}_{\mathrm{dR}}^{+}$ si $i \in \mathbf{Z}$. La série $\log[\varepsilon] = \sum_{n=1}^{+\infty} \frac{(-1)^{n-1}}{n}\pi^{n}$ converge dans $\mathbf{B}_{\mathrm{dR}}^{+}$ vers un élément que nous noterons t sur lequel $\sigma \in \mathscr{G}_{\mathbf{Q}_p}$ agit via la formule $\sigma(t) = \chi(\sigma)t$ et qui peut être vu comme un analogue p-adique de $2i\pi$. Si $x \in K_{\infty}((t))$ et $n \in \mathbf{N}$, alors la suite $\frac{1}{p^m}\mathrm{Tr}_{K_m((t))/K_n((t))}(x)$ est stationnaire pour $m \geq n$ assez grand. On note $\mathrm{T}_{K,n}$ l'application de $K_{\infty}((t))$ dans $K_n((t))$ ainsi définie.

Si V est une représentation p-adique de $\mathscr{G}_K$, on note $\mathrm{D}_{\mathrm{dR}}(V)$ le module $(\mathbf{B}_{\mathrm{dR}} \otimes V)^{\mathscr{G}_K}$. C'est un K-espace vectoriel de dimension finie muni d'une filtration décroissante par des sous-K-espaces vectoriels. Une représentation $\mathbf{B}_{\mathrm{dR}}$-admissible de $\mathscr{G}_K$ est dite "de de Rham". Les représentations de $\mathscr{H}_K$ sont toutes $\mathbf{B}_{\mathrm{dR}}$-admissibles et les aplications $\mathrm{T}_{K,n}$ donnent une bonne idée de ce à quoi $\mathbf{B}_{\mathrm{dR}}^{\mathscr{H}_K}$ ressemble.

Proposition 1. *$K_{\infty}((t))$ est dense dans $\mathbf{B}_{\mathrm{dR}}^{\mathscr{H}_K}$ et $\mathrm{T}_{K,n}$ s'étend par continuité en une application $\mathbf{Q}_p$-linéaire de $\mathbf{B}_{\mathrm{dR}}^{\mathscr{H}_K}$ dans $K_n((t))$.*

IV $\mathbf{B}_{\mathrm{cont}}$ et les représentations cristallines

On note $\mathbf{A}_{\max}$ le complété de $\widetilde{\mathbf{A}}^{+}[\frac{\omega}{p}]$ pour la topologie p-adique et $\mathbf{B}_{\max}^{+} = \mathbf{A}_{\max}[\frac{1}{p}]$. Comme l'idéal (p,ω) de $\widetilde{\mathbf{A}}^{+}$ est stable par φ, l'action de φ s'étend par continuité à $\mathbf{A}_{\max}$ et $\mathbf{B}_{\max}^{+}$ mais n'est plus une bijection et on pose $\mathbf{B}_{\mathrm{cont}}^{+} = \cap_{n\in\mathbf{N}}\varphi^{n}(\mathbf{B}_{\max}^{+})$. D'autre part, $\mathbf{B}_{\mathrm{cont}}^{+}$ s'identifie naturellement à un sous-anneau de $\mathbf{B}_{\mathrm{dR}}^{+}$ contenant t (on a $\varphi(t) = pt$) et on pose $\mathbf{B}_{\mathrm{cont}} = \mathbf{B}_{\mathrm{cont}}^{+}[1/t]$.

Si V est une représentation p-adique de $\mathscr{G}_K$, on note $\mathrm{D}_{\mathrm{cris}}(V)$ le module $(\mathbf{B}_{\mathrm{cont}} \otimes V)^{\mathscr{G}_K}$. C'est un $K \cap \mathbf{Q}_p^{\mathrm{nr}}$-espace vectoriel de dimension finie muni d'une action de φ et $K \otimes_{K\cap\mathbf{Q}_p^{\mathrm{nr}}} \mathrm{D}_{\mathrm{cris}}(V)$ s'identifie à un sous-K-espace vectoriel de $\mathrm{D}_{\mathrm{dR}}(V)$ et donc est muni d'une filtration décroissante. Une représentation $\mathbf{B}_{\mathrm{cont}}$-admissible de $\mathscr{G}_K$ est dite "cristalline". Une représentation de $\mathscr{H}_K$ est "presque" $\mathbf{B}_{\mathrm{cont}}$-admissible et même "presque" $\mathbf{B}_{\mathrm{cont}}^{\varphi=1}$-admissible et la proposition suivante nous donne une description de $\mathbf{B}_{\mathrm{cont}}^{\mathscr{H}_K}$ dans le cas ou K est non ramifié sur $\mathbf{Q}_p$.

Proposition 2. *Si K est non ramifié sur $\mathbf{Q}_p$ et $x \in (\mathbf{B}_{\mathrm{cont}}^{+})^{\mathscr{H}_K}$, il existe une unique distribution μ sur $\mathbf{Q}_p$ telle que l'on ait $x = \int_{\mathbf{Q}_p}[\varepsilon^{x}]\mu$. On dit que x est la transformée de Fourier de μ. D'autre part, si $n \geq 1$, alors $\mathrm{T}_{K,n}(x)$ est la transformée de Fourier de la restriction de μ à $p^{-n}\mathbf{Z}_p$.*

[10] On renvoie à *Périodes p-adiques* exposés II et III, Astérisque 223, 1994 pour les détails concernant cette section et la suivante.

V L'application exponentielle de Bloch-Kato

Les anneaux $\mathbf{B}_{\mathrm{cont}}$ et $\mathbf{B}_{\mathrm{dR}}$ sont reliés par la suite exacte fondamentale

$$0 \longrightarrow \mathbf{Q}_p \longrightarrow \mathbf{B}_{\mathrm{cont}}^{\varphi=1} \longrightarrow \mathbf{B}_{\mathrm{dR}}/\mathbf{B}_{\mathrm{dR}}^{+} \longrightarrow 0.$$

Soient K une extension finie de $\mathbf{Q}_p$ et V une représentation p-adique de $\mathscr{G}_K$. Tensorisant la suite exacte fondamentale avec V et prenant la suite exacte de cohomologie associée, on en déduit une application de $\mathrm{D}_{\mathrm{dR}}(V)$ dans $H^1(K,V)$ appelée exponentielle de Bloch-Kato[11] et notée $\exp_V$. Cette application se factorise à travers $\mathrm{D}_{\mathrm{dR}}(V)/\mathrm{Fil}^0\mathrm{D}_{\mathrm{dR}}(V)$ et son image est incluse dans le noyau $H^1_e(K,V)$ de l'application naturelle de $H^1(K,V)$ dans $H^1(K,\mathbf{B}_{\mathrm{cont}}^{\varphi=1}\otimes V)$.

D'autre part, si V est de de Rham, l'image de $\exp_V$ est $H^1_e(K,V)$ tout entier. et si $k \gg 0$, alors $\exp_{V(k)}$ est un isomorphisme de $\mathrm{D}_{\mathrm{dR}}(V(k))$ sur $H^1(K,V(k))$. Par dualité, on définit[12] une application $\exp^*_V : H^1(K,V^*(1)) \to \mathrm{D}_{\mathrm{dR}}(V^*(1))$.

VI Les anneaux $\mathbf{E}$, $\mathbf{A}$ et $\mathbf{B}$

On note $\widetilde{\mathbf{A}}$ le complété de $\widetilde{\mathbf{A}}^+[\frac{1}{p}]$ pour la topologie p-adique. L'anneau $\widetilde{\mathbf{A}}$ est aussi l'anneau $W(\widetilde{\mathbf{E}})$ des vecteurs de Witt à coefficients dans $\widetilde{\mathbf{E}}$ et $\widetilde{\mathbf{B}} = \widetilde{\mathbf{A}}[\frac{1}{p}]$ en est le corps des fractions. Si K est une extension finie de $\mathbf{Q}_p$, les anneaux $\widetilde{\mathbf{E}}_K = \widetilde{\mathbf{E}}^{\mathscr{H}_K}$, $\widetilde{\mathbf{A}}_K = \widetilde{\mathbf{A}}^{\mathscr{H}_K}$ et $\widetilde{\mathbf{B}}_K = \widetilde{\mathbf{B}}^{\mathscr{H}_K}$ ont des structures un peu désagréables, ce qui a amené Fontaine à introduire des sous-anneaux $\mathbf{E}$, $\mathbf{A}$ et $\mathbf{B}$[13] de $\widetilde{\mathbf{E}}$, $\widetilde{\mathbf{A}}$ et $\widetilde{\mathbf{B}}$ respectivement qui sont stables par φ et $\mathscr{G}_{\mathbf{Q}_p}$. Si K est une extension finie de $\mathbf{Q}_p$, on pose[14] $\mathbf{E}_K = \mathbf{E}^{\mathscr{H}_K}$, $\mathbf{A}_K = \mathbf{A}^{\mathscr{H}_K}$ et $\mathbf{B}_K = \mathbf{B}^{\mathscr{H}_K}$.

PROPOSITION 3. *(i)* $\mathbf{B}$ *est un corps valué complet dont* $\mathbf{A} = \mathbf{B} \cap \widetilde{\mathbf{A}}$ *est l'anneau des entiers et* $\mathbf{E}$ *est le corps résiduel. De plus* $\mathbf{E}$ *est la clôture séparable de* $\mathbf{E}_{\mathbf{Q}_p} = \mathbf{F}_p((\varepsilon-1))$ *dans* $\widetilde{\mathbf{E}}$ *et* $\mathrm{Gal}(\mathbf{E}/\mathbf{E}_K) = \mathscr{H}_K$ *si* K *est une extension finie de* $\mathbf{Q}_p$.

(ii) $\mathbf{E}_K$ *est un corps local de caractéristique* p, $\widetilde{\mathbf{E}}_K$ *est le complété de sa clôture radicielle et* $\mathbf{B}_K$ *est un corps local de dimension* 2 *dont* $\mathbf{A}_K$ *est l'anneau des entiers et* $\mathbf{E}_K$ *le corps résiduel.*

Le lien entre $\varphi^{-n}(\mathbf{E}_K)$ et $\widetilde{\mathbf{E}}_K$ ou $\varphi^{-n}(\mathbf{B}_K)$ et $\widetilde{\mathbf{B}}_K$ est à peu près le même que celui entre $K_n((t))$ et $\mathbf{B}_{\mathrm{dR}}^{\mathscr{H}_K}$ comme le montre la proposition 7. En particulier, les applications $\mathrm{T}_{K,n} : \mathbf{B}_{\mathrm{dR}}^{\mathscr{H}_K} \to K_n((t))$ de la proposition 1 ont des analogues[15] très utiles pour démontrer le théorème 8 par exemple.

[11] S. Bloch et K. Kato, dans "*The Grothendieck Fesschrift*", vol. I, 333-400, Birkhaüser 1990.

[12] Cette définition de l'exponentielle duale est un peu détournée, mais K. Kato (Springer Lect. Notes 1553, 50-163, 1993), en a trouvé une construction directe.

[13] Il les note respectivement $E^{\mathrm{sép}}$, $\mathscr{O}_{\widehat{\mathscr{E}^{\mathrm{nr}}}}$ et $\widehat{\mathscr{E}^{\mathrm{nr}}}$.

[14] $\mathbf{E}_K$ est le corps des normes de l'extension K_∞/K et la théorie du corps des normes est l'ingrédient principal de la démonstration de la proposition 3.

[15] Du point de vue des distributions (cf. prop. 2), passer de $\widetilde{\mathbf{B}}$ à $\mathbf{B}$ revient à ne regarder que les distributions à support dans $\mathbf{Z}_p$ qui a le bon goût d'être compact.

VII Le (φ,Γ_K)-module associé à une représentation de $\mathscr{G}_K$

L'image de $H^1(\mathscr{H}_K,\mathrm{GL}_d(\mathbf{F}_p))$ dans $H^1(\mathscr{H}_K,\mathrm{GL}_d(\mathbf{E}))$ est triviale d'après le théorème de Hilbert 90. Un petit argument de dévissage permet d'en déduire que l'image de $H^1(\mathscr{H}_K,\mathrm{GL}_d(\mathbf{Z}_p))$ dans $H^1(\mathscr{H}_K,\mathrm{GL}_d(\mathbf{A}))$ est triviale puis que l'image de $H^1(\mathscr{H}_K,\mathrm{GL}_d(\mathbf{Q}_p))$ dans $H^1(\mathscr{H}_K,\mathrm{GL}_d(\mathbf{B}))$ est triviale. On obtient donc la proposition suivante.

Proposition 4. *Toute représentation p-adique de $\mathscr{H}_K$ est $\mathbf{B}$-admissible.*

Cette proposition peut être grandement précisée grâce à l'introduction des notions de φ-module et de (φ,Γ)-module.

Définition 5. Soit K une extension finie de $\mathbf{Q}_p$.

(i) On appelle φ-module sur $\mathbf{B}_K$ tout $\mathbf{B}_K$-espace vectoriel de dimension finie muni d'une action semi-linéaire de φ.

(ii) On dit qu'un φ-module est étale ou de pente 0 s'il possède une base sur $\mathbf{B}_K$ dans laquelle la matrice de φ appartient à $\mathrm{GL}_d(\mathbf{A}_K)$.

(iii) On appelle (φ,Γ_K)-module sur $\mathbf{B}_K$ tout $\mathbf{B}_K$-espace vectoriel de dimension finie muni d'actions semi-linéaires de Γ_K et φ commutant entre elles. On dit qu'un (φ,Γ_K)-module est étale ou de pente 0 s'il l'est en tant que φ-module.

Si K est une extension finie de $\mathbf{Q}_p$ et V est une représentation p-adique de $\mathscr{H}_K$, on pose $D(V)=(\mathbf{B}\otimes_{\mathbf{Z}_p}V)^{\mathscr{H}_K}$. L'action de φ sur $\mathbf{B}$ commutant à celle de $\mathscr{G}_K$, $D(V)$ est muni d'une action de φ. Si de plus V est la restriction à $\mathscr{H}_K$ d'une représentation de $\mathscr{G}_K$, le module $D(V)$ est muni d'une l'action résiduelle de $\mathscr{G}_K/\mathscr{H}_K=\Gamma_K$ qui commute à celle de φ.

Proposition 6. *L'application qui à V associe $D(V)$ est une équivalence*[16] *de catégories de la catégorie des représentations p-adiques de $\mathscr{H}_K$ (resp. $\mathscr{G}_K$) sur celle des φ-modules (resp. (φ,Γ)-modules) étales sur $\mathbf{B}_K$.*

VIII $\mathbf{B}^\dagger$ et les représentations surconvergentes

Si $n\in\mathbf{N}$, on note $\widetilde{\mathbf{A}}^{\dagger,n}$ le complété de $\widetilde{\mathbf{A}}^+[\frac{p}{\omega^{p^n}}]$ pour la topologie p-adique et $\widetilde{\mathbf{B}}^{\dagger,n}=\widetilde{\mathbf{A}}^{\dagger,n}[\frac{1}{p}]$. Ces anneaux s'identifient à des sous-anneaux de $\widetilde{\mathbf{B}}$ et $\widetilde{\mathbf{B}}^\dagger=\cup_{n\in\mathbf{N}}\widetilde{\mathbf{B}}^{\dagger,n}$ est un sous-corps de $\widetilde{\mathbf{B}}$ stable par φ. Si $(a_k)_{k\in\mathbf{N}}$ est une suite d'éléments de $\widetilde{\mathbf{A}}^+$ tendant p-adiquement vers 0, alors la série $\sum_{k=0}^{+\infty}\varphi^{-n}(a_k)\big(\frac{p}{\varphi^{-n}(\omega)^{p^n}}\big)^k$ converge dans $\mathbf{B}^+_{\mathrm{dR}}$, ce qui nous permet de définir un morphisme[17] d'anneaux φ^{-n} de $\widetilde{\mathbf{B}}^{\dagger,n}$ dans $\mathbf{B}^+_{\mathrm{dR}}$ qui est injectif et commute à l'action de Galois.

[16] Comme $\mathbf{B}^{\varphi=1}=\mathbf{Q}_p$, si V est une représentation p-adique de $\mathscr{G}_K$, alors $(\mathbf{B}\otimes_{\mathbf{B}_K}D(V))^{\varphi=1}$ est canoniquement isomorphe à V en tant que représentation de $\mathscr{G}_K$

[17] Ce morphisme permet de relier les invariants de V obtenus via la théorie des (φ,Γ)-modules à ceux obtenus via les anneaux des périodes p-adiques ; c'est ce qui justifie l'introduction de la notion de représentation surconvergente.

On définit un sous-corps $\mathbf{B}^\dagger$ de $\mathbf{B}$ stable par φ et $\mathscr{G}_{\mathbf{Q}_p}$ et, si $n \in \mathbf{N}$, un sous-anneau $\mathbf{B}^{\dagger,n}$ de $\mathbf{B}$ stable par $\mathscr{G}_K$ en posant $\mathbf{B}^\dagger = \mathbf{B} \cap \widetilde{\mathbf{B}}^\dagger$ et $\mathbf{B}^{\dagger,n} = \mathbf{B} \cap \widetilde{\mathbf{B}}^{\dagger,n}$. Finalement, si K est une extension finie de $\mathbf{Q}_p$, on pose $\mathbf{B}_K^\dagger = (\mathbf{B}^\dagger)^{\mathscr{H}_K}$ et $\mathbf{B}_K^{\dagger,n} = (\mathbf{B}^{\dagger,n})^{\mathscr{H}_K}$. Les éléments de $\mathbf{B}_K^\dagger$ peuvent se décrire en termes de séries de Laurent surconvergentes et on a le résultat suivant.

PROPOSITION 7. *Si K est une extension finie de $\mathbf{Q}_p$ et si n est assez grand, alors $\varphi^{-n}(\mathbf{B}_K^{\dagger,n}) \subset K_n((t))$.*

Si V est une représentation p-adique de $\mathscr{H}_K$, on pose $D^\dagger(V) = (\mathbf{B}^\dagger \otimes_{\mathbf{Q}_p} V)^{\mathscr{H}_K}$ et $D^{\dagger,n}(V) = (\mathbf{B}^{\dagger,n} \otimes_{\mathbf{Q}_p} V)^{\mathscr{H}_K}$ si $n \in \mathbf{N}$. Une représentation de $\mathscr{H}_K$ qui est $\mathbf{B}^\dagger$-admissible est dite "surconvergente". On peut trouver des représentations de $\mathscr{H}_K$ qui ne sont pas surconvergentes (c'est même le cas général), mais on a le théorème suivant[18] qui montre que l'on n'a pas besoin d'introduire trop de dénominateurs (en π ou ω) pour décrire les représentations $\mathscr{G}_K$.

THÉORÈME 8. *Si K est une extension finie de $\mathbf{Q}_p$, toute représentation p-adique de $\mathscr{G}_K$ est surconvergente.*

IX $\mathbf{B}^+$ ET LES REPRÉSENTATIONS DE HAUTEUR FINIE

On pose $\widetilde{\mathbf{B}}^+ = \widetilde{\mathbf{A}}^+[\frac{1}{p}]$ et $\mathbf{B}^+ = \mathbf{B} \cap \widetilde{\mathbf{B}}^+$. Si V et une représentation p-adique de $\mathscr{G}_K$, on pose $D^+(V) = (\mathbf{B}^+ \otimes V)^{\mathscr{H}_K}$ et on dit[19] que V est "de hauteur finie" si on n'a pas besoin de dénominateurs pour la décrire, c'est-à-dire si $D^+(V)$ contient une base de $D(V)$ sur $\mathbf{B}_K$.

Un telle représentation est particulièrement sympathique et, dans le cas où K est non ramifié, on a le résultat suivant[20] qui avait été conjecturé par Fontaine.

THÉORÈME 9. *Si K est non ramifié[21] sur $\mathbf{Q}_p$, toute représentation cristalline de $\mathscr{G}_K$ est de hauteur finie.*

---

[18] F. Cherbonnier et P. Colmez, Représentations p-adiques surconvergentes, Inv. Math. Le point de départ de la démonstration est le résultat de Sen (*loc. cit.*) qui permet de montrer que toute représentation de $\mathscr{H}_K$ est $\widetilde{\mathbf{B}}^\dagger$-admissible. Pour redescendre de $\widetilde{\mathbf{B}}^\dagger$ à $\mathbf{B}^\dagger$, on utilise les opérateurs $\mathrm{T}_{K,n}$ et une étude fine de l'action de Γ_K sur $\widetilde{\mathbf{B}}_K^\dagger$.

[19] N. Wach, Bull. de la S.M.F. 124, 375-400, 1996.

[20] P. Colmez, Représentations cristallines et représentations de hauteur finie, 1997. La démonstration qui se trouve dans cette prépublication est très tortueuse. Une démonstration plus directe fournissant une description de $D^+(V)$ serait la bienvenue ; cela a été fait par N. Wach (*loc. cit.*) dans le cas où la longueur de la filtration de $\mathrm{D}_{\mathrm{cris}}(V)$ est inférieure ou égale à $p-1$.

[21] Cette hypothèse peut être remplacée par K_∞ non ramifié sur $\mathbf{Q}_p(\mu_{p^\infty})$, mais ne peut être totalement supprimée : il existe des représentations cristallines qui ne sont pas de hauteur finie. D'autre part, on dispose d'un critère simple portant sur l'action de Γ_K sur $D^+(V)$ pour qu'une représentation de hauteur finie soit cristalline (Wach (*loc. cit.*)).

X Modules d'Iwasawa associés à une représentation p-adique

Si V est une représentation p-adique de $\mathscr{G}_K$, on note $H^i_{\mathrm{Iw}}(K,V)$ le groupe de cohomologie continue $H^i(\mathscr{G}_K, \mathbf{Z}_p[[\Gamma_K]] \otimes V)$. On peut aussi voir $\mathbf{Z}_p[[\Gamma_K]] \otimes V$ comme l'ensemble des mesures sur Γ_K à valeurs dans V et comme l'application $\mu \to \chi(x)^k\mu$ est un isomorphisme $\mathscr{G}_K$-équivariant de $\mathbf{Z}_p[[\Gamma_K]] \otimes V$ sur $\mathbf{Z}_p[[\Gamma_K]] \otimes V(k)$, on en déduit des isomorphismes $H^i_{\mathrm{Iw}}(K,V(k)) \cong H^i_{\mathrm{Iw}}(K,V)$ et des applications[22] $\mu \to \int_{\Gamma_{K_n}} \chi(x)^k\mu$ de $H^i_{\mathrm{Iw}}(K,V)$ dans $H^i(K_n,V(k))$ pour tout $k \in \mathbf{Z}$ et $n \in \mathbf{N}$.

Les groupes $H^i_{\mathrm{Iw}}(K,V)$ ont été étudiés en détail par Perrin-Riou[23]. On a en particulier le résultat suivant.

Proposition 10. *Soit V une représentation p-adique de $\mathscr{G}_K$.*

i) $H^i_{\mathrm{Iw}}(K,V) = 0$ si $i \neq 1,2$.

ii) $H^1_{\mathrm{Iw}}(K,V)$ est un $\mathbf{Q}_p \otimes \mathbf{Z}_p[[\Gamma_K]]$-module de type fini dont le sous-module de torsion est naturellement isomorphe à $V^{\mathscr{H}_K}$ et $H^1_{\mathrm{Iw}}(K,V)/V^{\mathscr{H}_K}$ est libre de rang $[K:\mathbf{Q}_p]\dim_{\mathbf{Q}_p} V$.

iii) $H^2_{\mathrm{Iw}}(K,V)$ est isomorphe à $V(-1)^{\mathscr{H}_K}$ en tant que $\mathbf{Q}_p \otimes \mathbf{Z}_p[[\Gamma_K]]$-module ; en particulier, il est de torsion.

XI La machine à fonctions-L p-adiques

Afin de mieux comprendre la construction par Coates et Wiles[24] de la fonction-L p-adique d'une courbe elliptique à multiplication complexe à partir des unités elliptiques, Coleman[25] a montré comment associer à tout $u \in \varprojlim \mathscr{O}^*_{K_n}$ une mesure λ_u sur $\mathbf{Z}_p^*$ dans le cas où K est non ramifié sur $\mathbf{Q}_p$. L'application qui à u associe λ_u est presque un isomorphisme de $\mathbf{Z}_p[[\Gamma_K]]$-modules et est appelé l'isomorphisme de Coleman. Si on prend pour u le système des unités cyclotomiques, la mesure λ_u que l'on obtient donne la fonction zêta de Kubota-Leopoldt. Quand on a la chance de disposer d'une telle construction pour une fonction-L p-adique, il y a toujours des retombées arithmétiques spectaculaires et il semble donc intéressant d'essayer de généraliser la construction de Coleman à d'autres représentations que $\mathbf{Q}_p(1)$[26]. Cela a été fait par Perrin-Riou[27] dans le cas d'une représentation cristalline d'une extension non ramifiée de $\mathbf{Q}_p$, ce qui lui a permis[28] de donner une définition (conjecturale) de la fonction-L p-adique d'un motif ayant bonne réduction en p. Sa construction repose sur une interpolation p-adique des exponentielles de Bloch-Kato

[22] Utilisant ces applications, on montre que $H^i_{\mathrm{Iw}}(K,V)$ est isomorphe à $\mathbf{Q}_p \otimes_{\mathbf{Z}_p} \varprojlim H^i(K_n,T)$, où T est un $\mathbf{Z}_p$-réseau de V stable par $\mathscr{G}_K$ et la limite projective est prise relativement aux applications de corestriction. On retombe donc sur la définition usuelle des modules d'Iwasawa.

[23] B. Perrin-Riou, Inv. Math. 115, 81-149, 1994

[24] J. Coates et A. Wiles, J. Australian Math. Soc., A 26, 1-25, 1978

[25] R. Coleman, Inv. Math. 53, 91-116, 1979

[26] La théorie de Kummer nous fournit une application δ de $\varprojlim \mathscr{O}^*_{K_n}$ dans $H^1_{\mathrm{Iw}}(K,\mathbf{Q}_p(1))$

[27] *loc. cit.*

[28] B. Perrin-Riou, Astérisque 229, 1995

pour les représentations $V(k)$ avec $k \in \mathbf{Z}$ et fournit une application "exponentielle" qui, dans le cas de $\mathbf{Q}_p(1)$ donne l'inverse de l'isomorphisme de Coleman. Dans la suite de ce texte, nous allons présenter deux généralisations de sa construction.

XII L'application logarithme

Soit V une représentation de de Rham telle que $(\mathbf{B}_{\rm cont}^{\varphi=1}\otimes V)^{\mathscr{G}_{K_n}}=\{0\}$ quel que soit $n\in\mathbf{N}$[29]. Notons $H^1_{{\rm Iw},e}(K,V)$ le sous-ensemble des éléments μ de $H^1_{\rm Iw}(K,V)$ tels que $\int_{\Gamma_{K_n}}\mu\in H^1_e(K_n,V)$ quel que soit $n\in\mathbf{N}$. L'ensemble $H^1_{{\rm Iw},e}(K,V)$ peut très bien être réduit à 0, mais il existe $k(V)\in\mathbf{Z}$ tel que l'on ait $H^1_{{\rm Iw},e}(K,V)=H^1_{\rm Iw}(K_n,V(k))$ si $k\geq k(V)$.

Si $\mu\in H^1_{{\rm Iw},e}(K,V)$ et $\tau\to\mu_\tau$ est un cocycle continu représentant μ, il existe, quel que soit $n\in\mathbf{N}$, un élément $c_n\in\mathbf{B}_{\rm cont}^{\varphi=1}\otimes V$ tel que l'on ait $(1-\tau)c_n=\int_{\Gamma_{K_n}}\mu_\tau$ quel que soit $\tau\in\mathscr{G}_{K_n}$. L'élément c_n est bien déterminé grâce à l'hypothèse faite sur V.

THÉORÈME 11. *Si $\mu\in H^1_{{\rm Iw},e}(K,V)$, alors la suite de terme général p^nc_n converge dans $\mathbf{B}_{\rm cont}^{\varphi=1}\otimes V$ vers un élément de $(\mathbf{B}_{\rm cont}^{\varphi=1}\otimes V)^{\mathscr{H}_K}$ qui ne dépend pas du choix du cocycle $\tau\to\mu_\tau$; il est noté* $\mathrm{Log}(\mu)$.

Cette application logarithme[30] est une généralisation de l'isomorphisme de Coleman et, dans le cas où V est cristalline, est, à normalisation près, un inverse de l'application exponentielle introduite par Perrin-Riou. Plus précisément, utilisant la transformée de Fourier des distributions et les résultats de Perrin-Riou, on démontre le résultat suivant.

THÉORÈME 12. *Soit K une extension finie non ramifiée de $\mathbf{Q}_p$.*

(i) Si V est une représentation cristalline de $\mathscr{G}_K$ telle que $\mathrm{Fil}^1\mathrm{D}_{\rm cris}(V)=\{0\}$ *et $\mu\in H^1_{{\rm Iw},e}(K,V)$, alors il existe une (unique) distribution*[31] *$\lambda_V(\mu)$ sur $\mathbf{Q}_p$ à valeurs dans* $\mathrm{D}_{\rm cris}(V)$ *dont* $\mathrm{Log}_V(\mu)$ *est la transformée de Fourier et si k est un entier suffisamment grand, alors*

$$\frac{1-p^{-1}\varphi^{-1}}{1-\varphi}\Big(\int_{\mathbf{Z}_p^*}\frac{k!}{(-tx)^k}\lambda_V(\mu)\Big)=\exp_{V(k)}^{-1}\Big(\int_{\Gamma_K}\chi(x)^k\mu\Big)$$

(ii) Si $u\in\varprojlim\mathscr{O}_{K_n}^*$, la mesure λ_u que l'on obtient via l'isomorphisme de Coleman est la restriction à $\mathbf{Z}_p^*$ de $\lambda_{\mathbf{Q}_p(1)}(\delta(u))$.

[29] Cette hypothèse n'est là que pour simplifier les énoncés qui suivent et devient automatique si on remplace V par $V(k)$ sauf pour un nombre fini de $k\in\mathbf{Z}$.

[30] P. Colmez, Théorie d'Iwasawa des représentations de de Rham d'un corps local, Ann. of Math.

[31] L'existence de cette distribution traduit une propriété de continuité p-adique de l'application $k\to\exp_{V(k)}$. L'idée qu'une telle continuité devait exister a d'ailleurs été le point de départ de Perrin-Riou.

XIII La loi de réciprocité explicite de Perrin-Riou

Revenons au cas où K est une extension finie quelconque de $\mathbf{Q}_p$ et V une représentation de de Rham de $\mathscr{G}_K$. Si $n \in \mathbf{N}$, on étend l'application $\mathrm{T}_{K,n}$ par linéarité en une application de $(\mathbf{B}_{\mathrm{dR}} \otimes V)^{\mathscr{H}_K} = \mathbf{B}_{\mathrm{dR}}^{\mathscr{H}_K} \otimes \mathrm{D}_{\mathrm{dR}}(V)$ dans $K_n((t)) \otimes \mathrm{D}_{\mathrm{dR}}(V)$. Un élément x de $K_n((t)) \otimes \mathrm{D}_{\mathrm{dR}}(V)$ s'écrit de manière unique sous la forme $\sum_{k\in\mathbf{Z}} \partial_k(x)t^k$ avec $\partial_k(x) \in K_n \otimes \mathrm{D}_{\mathrm{dR}}(V)$. Tout cela nous permet de définir pour chaque $k \in \mathbf{Z}$ et $n \in \mathbf{N}$, un morphisme $\mathrm{CW}_{k,n}$ de $H^1_{\mathrm{Iw},e}(K,V)$ dans $K_n \otimes D_{\mathrm{dR}}(V)$ en posant

$$\mathrm{CW}_{k,n}(\mu) = \partial_k(\mathrm{T}_{K,n}(\mathrm{Log}_V(\mu))).$$

Ces morphismes sont des généralisations des morphismes de Coates-Wiles et le théorème suivant montre qu'ils sont liés aux exponentielles de Bloch-Kato.

Théorème 13. *Si $\mu \in H^1_{\mathrm{Iw},e}(K,V)$, si $n \in \mathbf{N}$ et si $k \in \mathbf{Z}$, alors*

$$\mathrm{CW}_{k,n}(\mu) = -\exp^*_{V^*(1+k)}\Big(\int_{\Gamma_{K_n}} \chi(x)^{-k}\mu\Big).$$

Si on suppose K non ramifié sur $\mathbf{Q}_p$ et V cristalline, on peut retraduire ce théorème en termes de distributions et on obtient la proposition suivante qui est une des formes équivalentes de la loi de réciprocité conjecturée par Perrin-Riou[32].

Proposition 14. *Sous les hypothèses du théorème 12, si $k \gg 0$, alors*

$$\frac{1-p^{-1}\varphi^{-1}}{1-\varphi}\Big(\int_{\mathbf{Z}_p^*} \frac{(tx)^k}{(k-1)!}\lambda_V(\mu)\Big) = -\exp^*_{V^*(1+k)}\Big(\int_{\Gamma_K} \chi(x)^{-k}\mu\Big)$$

XIV (φ,Γ)-modules et cohomologie galoisienne

Le corps $\mathbf{B}$ est une extension de degré p de $\varphi(\mathbf{B})$, (totalement ramifiée car l'extension résiduelle est radicielle). Ceci nous permet de définir une application $\psi : \mathbf{B} \to \mathbf{B}$ par la formule $\psi(x) = \varphi^{-1}(\mathrm{Tr}_{\mathbf{B}/\varphi(\mathbf{B})}(x))$. Ceci fait de ψ un inverse à gauche de φ qui commute à l'action de $\mathscr{G}_K$.

Soient K une extension finie de $\mathbf{Q}_p$ et Δ_K le sous-groupe de torsion de Γ_K de telle sorte que $\Gamma'_K = \Gamma_K/\Delta_K$ est isomorphe à $\mathbf{Z}_p$. Soit γ un générateur de Γ'_K. Si V est une représentation p-adique de $\mathscr{G}_K$, soit $D'(V) = D(V)^{\Delta_K}$. Considérons le complexe

$$0 \longrightarrow D'(V) \longrightarrow D'(V) \oplus D'(V) \longrightarrow D'(V) \longrightarrow 0,$$

où les applications de $D'(V)$ dans $D'(V)\oplus D'(V)$ et de $D'(V)\oplus D'(V)$ dans $D'(V)$ sont respectivement définies par $x \to ((\psi-1)x, (\gamma-1)x)$ et $(a,b) \to (\gamma-1)a - (\psi-1)b$.

[32]Cette loi est une généralisation de celle de Bloch-Kato pour $\mathbf{Q}_p(r)$; une démonstration complètement différente a été obtenue par Kato, Kurihara et Tsuji.

On a le résultat suivant[33]

THÉORÈME 15. *Si $i \in \mathbf{N}$, le i-ème groupe de cohomologie du complexe ci-dessus s'identifie fonctoriellement au groupe de cohomologie galoisienne $H^i(K,V)$.*

XV (φ,Γ)-MODULES ET THÉORIE D'IWASAWA

Les résultats mentionnés ci-dessus mènent naturellement[34] à une description des groupes $H^i_{\mathrm{Iw}}(K,V)$ en termes de $D(V)$.

THÉORÈME 16. *Soit V une représentation p-adique de $\mathscr{G}_K$.*

i) $H^1_{\mathrm{Iw}}(K,V)$ s'identifie fonctoriellement à $D(V)^{\psi=1}$ via une application Exp^*.

ii) $H^2_{\mathrm{Iw}}(K,V)$ s'identifie fonctoriellement à $\frac{D(V)}{\psi-1}$.

Remarquons que l'on n'a fait aucune hypothèse restrictive sur V ou sur K pour définir Exp^*. Dans le cas où K est non ramifié sur $\mathbf{Q}_p$ et $V=\mathbf{Q}_p(1)$, un petit calcul montre que $\mathrm{Exp}^*(\delta(u))$ est la transformée de Fourier de la mesure $x\lambda_u$, ce qui permet de voir l'application Exp^* comme une vaste généralisation de l'isomorphisme de Coleman.

On peut utiliser le fait que toute représentation p-adique de $\mathscr{G}_K$ est surconvergente pour relier[35], dans le cas des représentations de de Rham, les applications Exp^* et Log et retrouver les homomorphismes de Coates-Wiles généralisés via la théorie des (φ,Γ)-modules. De manière précise, on a la loi de réciprocité explicite suivante que l'on pourra comparer avec le théorème 13.

THÉORÈME 17. *Si K est une extension finie de $\mathbf{Q}_p$ et V une représentation de de Rham de $\mathscr{G}_K$, il existe $n(V)\in\mathbf{N}$ tel que si $\mu\in H^1_{\mathrm{Iw}}(K,V)$, alors $\mathrm{Exp}^*(\mu)\in D^{\dagger,n(V)}(V)$ et si $n\geq n(V)$, on a l'égalité suivante dans $K_n((t))\otimes \mathrm{D}_{\mathrm{dR}}(V)$*

$$p^{-n}\varphi^{-n}\Big(\mathrm{Exp}^*(\mu)\Big)=\sum_{k\in\mathbf{Z}}\exp^*_{V^*(1+k)}\Big(\int_{\Gamma_{K_n}}\chi(x)^{-k}\mu\Big).$$

D.M.I., École Normale Supérieure, 45 rue d'Ulm, 75005 Paris, France
Institut de Mathématiques, 4 place Jussieu, 75005 Paris, France

[33] L.Herr, *Cohomologie Galoisienne des corps p-adiques*, thèse de l'université d'Orsay, 1995. Le point de départ de la démonstration est la suite exacte $0\to\mathbf{Q}_p\to\mathbf{B}\xrightarrow{1-\varphi}\mathbf{B}\to 0$. La thèse de Herr contient en outre une démonstration du théorème de dualité locale via la théorie des (φ,Γ)-modules dont l'ingrédient principal est une description de l'isomorphisme canonique $H^2(K,\mathbf{Q}_p(1))\cong\mathbf{Q}_p$ grâce à une application résidu.

[34] Il s'agit d'un résultat non publié de J.-M. Fontaine ; on en trouvera une démonstration dans F. Cherbonnier et P. Colmez, Théorie d'Iwasawa des représentations p-adiques d'un corps local, Journal de l'A.M.S.

[35] Cette comparaison est d'ailleurs le point de départ de la démonstration du théorème 9. D. Benois a entrepris le chemin inverse et obtenu (On Iwasawa theory of crystalline representations, preprint 1998) une démonstration de la loi de réciprocité explicite de Perrin-Riou via la théorie des (φ,Γ)-modules dans le cas des représentations cristallines de hauteur finie.

Bounds for Arithmetic Multiplicities

W. Duke[1]

Abstract. This paper will describe some recent applications of techniques giving non-trivial upper bounds for multiplicities in certain arithmetic instances. These applications include estimates for dimensions of spaces of cusp forms of weight one, multiplicities of number fields with a given degree and discriminant, and the number of elliptic curves over the rationals whose reductions have the same number of points for a few small primes. The techniques share a common strategy, which combines approximate orthogonality with rigidity properties of arithmetic Fourier coefficients.

1991 Mathematics Subject Classification: 11F,11N
Keywords and Phrases: modular forms of weight one, number fields, class groups, elliptic curves

1 Introduction

A set of problems in Number Theory where analytic and algebraic techniques combine fruitfully concern finding upper bounds for arithmetic multiplicities. These problems are perhaps best introduced through a series of particular examples. They involve counting automorphic forms, number fields, class groups and elliptic curves under various conditions on associated eigenvalues. In most cases natural conjectures arise which appear to be quite difficult and the analytic method introduced provides non-trivial information but certainly not the final answer. The corresponding existence questions are left untreated here but present fascinating challenges.

2 Modular forms of weight one

In a variety of situations it is of interest to bound the multiplicity of an automorphic representation in an appropriate family (see [S-X]). Conjectured bounds for the multiplicities of certain Maass eigenvalues for congruence subgroups are crucial assumptions in the works of Philips and Sarnak [P-S] and of Wolpert [Wol] on the disappearance of cusp forms under perturbations. The main analytic tool which has been applied is the trace formula, but, in cases when the eigenvalue is

[1]Research supported in part by NSF Grant DMS-9500797.

not isolated, it only yields rough information since it is not capable by itself of effectively separating neighboring spectrum.

Perhaps the most classical instance of this problem is in determining the dimension of the space of holomorphic cusp forms of weight one for congruence subgroups as a function of the level. For forms of integral weight larger than one the dimension is well understood by means of either the Riemann–Roch theorem or the Selberg trace formula but the eigenvalue for weight one, $1/4$, is an accumulation point for the discrete spectrum when the level increases and thus the above problem intervenes.

For a positive integer N and χ a Dirichlet character (mod N) let $S_1(N,\chi)$ denote the space of holomorphic cusp forms for $\Gamma_0(N)$ of weight 1 with Nebentypus χ. If the order of χ is fixed a direct application of the trace formula gives

$$\dim S_1(N,\chi) \ll N$$

while Deshouillers/Iwaniec and Sarnak observed (unpublished) that a clever choice of test function yields the improvement

$$\dim S_1(N,\chi) \ll \frac{N}{\log N}.$$

Early on Hecke pointed out that weight one cusp forms for real χ may be constructed from non-real characters of class groups of imaginary quadratic fields. More generally, let ρ be a two-dimensional irreducible odd Galois representation and $\tilde{\rho}$ be the induced projective representation into $PGL(2,\mathbf{C})$. The image of $\tilde{\rho}$ is dihedral or isomorphic to one of A_4, S_4, or A_5. Langland's program predicts the existence of a newform $f = \sum a_f(n)e(nz) \in S_1(N,\chi)$ with

$$a_f(p) = \mathrm{tr}(\rho(\mathrm{Frob}_p)) \text{ and } \chi(p) = \det(\rho(\mathrm{Frob}_p))$$

for p not dividing N. The dihedral case corresponds to Hecke's construction. Langlands and Tunnell (see [Tu]) proved the existence of such a form in all but the A_5 case. Deligne and Serre [D-S] proved that every newform arises in this way.

Suppose for simplicity that N is prime and that χ is real. It can be shown that there are $(h-1)/2$ independent forms of dihedral type, where h is the class number of $\mathbf{Q}(\sqrt{-N})$ and thus there are $\ll N^{1/2}\log N$ such forms. Serre raised the question of bounding from above the number of non-dihedral forms. The following was proved in [Du].

THEOREM 1 *For N prime*

$$\dim S_1(N,\chi) \ll_\varepsilon N^{11/12+\varepsilon}.$$

It appears reasonable to expect that in fact

$$\dim S_1(N,\chi) = \frac{1}{2}(h-1) + O(N^\varepsilon).$$

In particular, this would imply that $\dim S_1(N,\chi) \ll N^{1/2}\log N$. Since by Siegel's theorem

$$\dim S_1(N,\chi) \underset{\varepsilon}{\gg} N^{1/2-\varepsilon}$$

this would be essentially best-possible.

The proof of Theorem 1 extends to general N and χ as long as the order of χ is fixed. Recently S. Wong [Wo1] has carried this out and extended the arguments to apply to general χ as well. The idea behind the proof is to take advantage of two properties of the Fourier coefficients of non-dihedral newforms which cannot co-exist if there are too many of them. These are their approximate orthogonality, which is a consequence of their belonging to automorphic forms, and the finiteness of the number of their possible values at primes, which is a consequence of their coming from Galois representations of a known type. The technique could in principle be applied to estimate other eigenvalue multiplicities in other Galois cases. For example, it would bound nontrivially the multiplicity of the eigenvalue $1/4$ of the weight zero Laplacian for congruence subroups if it were known that they come from Galois representations.

3 Number fields and class groups

A closely related problem concerns bounding from above the multiplicity of number fields of a given degree as a function of its discriminant. This in turn is, in cases covered by class field theory, tied to estimating the ranks of class groups.

For a positive integer n and an integer D let $M_n(D)$ be the number of number fields of degree n with discriminant D. Hermite showed that $M_n(D)$ is finite and Stickelberger observed that $M_n(D) = 0$ unless $D \equiv 0, 1 (\mathrm{mod}\ 4)$. Except for the case $n = 2$ when $M_n(D) = 1$ exactly for D fundamental, little is known about the size of $M_n(D)$.

On average over $|D| \le X$ a little more is known. Let

$$S_n(X) = \sum_{|D|\le X} M_n(D) :$$

for $n = 2, 3$ we have that

$$S_n(X) \sim c_n X$$

where $c_n = 1/\zeta(n)$, the case $n = 3$ being a famous result of Davenport-Heilbronn [D-H]. The best general upper bound is due to Schmidt [Sch]

$$S_n(X) \ll X^{(n+2)/4}$$

using the geometry of numbers. Wright and Yukie have announced an asymptotic in the case of quartic fields.

Such results have motivated the conjectured bound for fixed n:

$$M_n(D) \ll |D|^{\varepsilon}$$

but, except for the case $n = 2$, this is open. A non-trivial upper bound for the multiplicity of quartic fields was obtained in [Du].

THEOREM 2 *For $-D$ prime*

$$M_4(D) \ll_\varepsilon |D|^{7/8+\varepsilon}.$$

As in the case of Theorem 1, this result extends to more general quartic fields (see [Wo1].) Before Theorem 2, the only known upper bounds followed from trivial bounds for class numbers. By means of class field theory Heilbronn [He] showed that

$$M_4(D) = \frac{4}{3} \sum_k h_2(K)\,,$$

where K runs over all cubic number fields of discriminant D. Here, for any ℓ and any number field K, $h_\ell(K)$ denotes the number of ideal classes of K of (exact) order ℓ. Furthermore, the number of cubic fields in the sum is $\frac{3}{2}\, h_3(\mathbf{Q}(\sqrt{D}))$. For the class number $h(K)$ of any number field K of degree $n > 1$ and discriminant D we have the bound

$$h(k) \ll |D|^{1/2} \log^{n-1} |D|$$

where the implied constant depends only on n. Since $h_\ell(K) \le h(K)$ we deduce the "trivial" bound

$$M_4(D) \ll |D|^{1+\varepsilon}.$$

The improvement of this given in Theorem 2 requires both the classification of quartic fields of discriminant D by odd S_4– Galois representations of conductor $|D|$ and the proof in this case of the Artin conjecture given in [Tu]. If we assume the Artin conjecture for icosahedral representations then similarly we can prove that the number of non-real quintic fields of discriminant D^2 whose normal closure has Galois group A_5 is $O(|D|^{11/12+\varepsilon})$.

This discussion motivates another problem, which is to bound $h_\ell(K)$ and again, very little seems to be known. A famous exception is for quadratic fields $n = 2$ when $\ell = 2$, where Gauss' genus theory gives the formula

$$h_2(K) = 2^{\nu(D)-1} - 1$$

where $\nu(D)$ is the number of primes dividing D. Once again, it is suspected that in general for a given n, ℓ

$$h_\ell(K) \ll |D|^{\varepsilon}.$$

One may also a formulate the more precise possible bound (see [B-S])

$$\log h_\ell(K) \ll \log |D| / \log\log |D|.$$

4 Elliptic Curves

A basic multiplicity problem for elliptic curves is to bound the number $M(N)$ of elliptic curves over $\mathbf{Q}$ with conductor N. Recently Brumer and Silverman [B-S] have shown that

$$M(N) \ll N^{1/2+\varepsilon}.$$

They use that solutions to the discriminant equation for elliptic curves correspond to S-integral points on a curve

$$y^2 = x^3 + A$$

and that the number of such points is $\ll h_3(K)|N|^{\varepsilon}$ for some quadratic K with discriminant $\ll N$. Thus any improvement on the trivial bound for $h_3(K)$ would improve the bound for $M(N)$. In this case it was observed in [D-K] that on average

$$\sum_{N \leq X} M(N) \ll X^{1+\varepsilon}$$

since the Davenport-Heilbronn Theorem is applicable. Brumer and Silverman also showed that under standard conjectures about L-functions for elliptic curves (GRH, BSD) that

$$M(N) \ll N^{\varepsilon}.$$

Wong [Wo2] observed that under these hypotheses one may deduce that

$$h_3(k) \ll |D|^{1/4+\varepsilon}.$$

We turn to an enrichment of the question of counting all elliptic curves. It is connected to the problem of determining the extent to which an elliptic curve defined over $\mathbf{Q}$ is determined by the trace of Frobenius for a few small primes. This problem is analogous to bounding the least quadratic non-residue, a venerable problem in classical analytic number theory. Assuming the Riemann-Hypothesis for Artin L-functions, Serre [Se2] showed that $O((\log N)^2)$ primes suffice, where N is the conductor of the curve. No nontrivial unconditional results are known for this problem due in part to the difficulty in breaking convexity for the associated Rankin-Selberg L-function (see [D-F-I]).

The associated multiplicity problem is to estimate the maximal number of isogeny classes of curves which have the same trace of Frobenius for a few small primes, in terms of the conductor. Recently in a joint work with E. Kowalski we obtained an estimate which shows that "most" curves are determined by very few primes. Our proof uses modularity of the curves and hence we must restrict ourselves to curves for which the theorem of Wiles [Wi] or a generalization applies.

For example, let $M(X, \alpha)$ be the maximal number of isogeny classes of semi-stable elliptic curves over $\mathbf{Q}$ with conductor less than or equal to X which for every prime $p \leq (\log X)^{\alpha}$ have a fixed number of points modulo p. The following is proved in [D-K].

Theorem 3 *We have for any* $\varepsilon > 0$

$$M(X, \alpha) \ll_{\varepsilon} X^{8/\alpha+\varepsilon}.$$

It follows from this and the lower bound [F-N-T]

$$Ell(X) \gg X^{5/6}$$

for the number of isogeny classes of semi-stable elliptic curves with conductor less than X that the probability that two such elliptic curves have the same number of points (mod p) for all primes $p \le (\log X)^\alpha$ tends to zero as X tends to infinity, if α is large enough. It may be viewed as an analogue of the classical result of Linnik bounding the number of primes with no small quadratic non-residues.

5 Approximate orthogonality

A unifying feature of the results outlined is the use of mean-value theorems which display in a quantitative form the orthogonality of the Fourier coefficients of newforms. Such theorems, already in extremely sophisticated form, were introduced and applied by Deshouillers and Iwaniec [D-I] and have been used extensively in the analytic theory of automorphic L-functions. The uses we are describing are more rudimentary in the sense that direct use is made of the coefficients.

Let $S(N) = S_k^+(N, \chi)$ denote the set of newforms of integral weight k for $\Gamma_0(N)$ with character χ. Each $f \in S$ has the Fourier expansion at ∞

$$f(z) = \sum_{n \ge 1} a_f(n)\, e(nz)\,.$$

The Hecke eigenvalues are

$$\lambda_f(n) = n^{-(k-1)/2} a_f(n).$$

The simplest mean-value result is that applied in the proof of Theorem 1 and is the following. For arbitrary $c_n \in \mathbf{C}$ with $1 \le n \le X$ we have

$$\sum_{f \in S(N)} \Big| \sum_{n \le X} c_n\, \lambda_f(n) \Big|^2 \ll (X+N) N^\varepsilon \sum_{n \le X} |c_n|^2. \tag{1}$$

This result is proved by using a form of duality and the following estimate for any cusp form f, not necessarily a newform:

$$\sum_{n \le X} |a_f(n)|^2 \ll (1 + X/N) \langle f, f \rangle$$

where $\langle f, f \rangle$ is the Petersson inner product.

For the proof of Theorem 3 we need more sophisticated mean value theorems which average also over the level and are thus reminiscent of the classical large sieve inequality for primitive Dirichlet characters:

$$\sum_{q \le Q} \sideset{}{^*}\sum_{\chi (\mathrm{mod}\ q)} \Big| \sum_{n \le X} c_n \chi(n) \Big|^2 \le (X + Q^2) \sum_{n \le X} |c_n|^2$$

which gives a kind of approximate orthogonality for the truncated sequences $(\chi(n))_{1 \le n \le X}$ considered as elements of a finite dimensional Hilbert space.

Suppose for simplicity that the Nebentypus character χ is trivial. The first inequality is

$$\sum_{N\le X}^{\flat}\sum_{f\in S(N)}\Big|\sum_{n\le X^{\beta}}c_n\lambda_f(n)\Big|^2\ll X^{\beta+\varepsilon}\sum_n|c_n|^2 \tag{2}$$

for any $\varepsilon>0$, and $\beta>4$, where $\sum^{\flat}$ indicates a sum over squarefree integers.

We also needed to use another inequality which is similar to the previous one except that it detects orthogonality along the squares:

$$\sum_{N\le X}^{\flat}\sum_{f\in S(N)}\Big|\sum_{n\le X^{\beta}}^{\flat}c_n\lambda_f(n^2)\Big|^2\ll X^{\beta+\varepsilon}\sum_n|c_n|^2 \tag{3}$$

for any $\varepsilon>0$, and this time $\beta>10$. This may be interpreted as a partial large-sieve inequality for the symmetric squares of the new-forms, which are $GL(3)$-automorphic forms defined by Gelbart and Jacquet [G-J].

These results are also proved using duality, but in a different form. The second one, which is by far the more difficult, reduces to proving a smoothed version of

$$\sum_{n\le X^{\beta}}\lambda_f^{(2)}(n)\lambda_g^{(2)}(n)\ll\begin{cases}X^{\beta-2+\varepsilon},\text{ if } f\neq g\\ X^{\beta+\varepsilon},\text{ if } f=g\end{cases}$$

where $\lambda_f^{(2)}$ denotes the coefficients of the L-function of the symmetric square $f^{(2)}$ of f. We are led to study the analytic properties of the "bilinear convolution" L-function

$$L_b(f^{(2)}\otimes g^{(2)},s)=\sum_{n\ge1}\lambda_f^{(2)}(n)\lambda_g^{(2)}(n)n^{-s}$$

which we do by relating it to the true Rankin-Selberg convolution $L(f^{(2)}\otimes g^{(2)},s)$, defined by Jacquet, Piatetskii-Shapiro and Shalika [J-P-S]. This comparison lemma gives us the analytic continuation of L_b up to the critical line, which is sufficient to get the result. Also used is the determination of the location of the poles of the Rankin-Selberg convolution, due to Moeglin and Waldspurger [M-W], and a result of Ramakrishnan according to which two newforms with squarefree levels cannot have the same symmetric square unless they are the same.

6 Rigidity of arithmetic coefficients

The essential idea behind the techniques for giving non-trivial upper bounds for arithmetic multiplicities is to show that the general approximate orthogonality of Fourier coefficients reflected in the mean value theorems of the previous section is not compatable with rigidity properties they possess by virtue of their arithmetic nature.

In the proof of Theorem 1 this rigidity comes from the finiteness of the set of possible values of $\lambda_f(p)$ when the associated Galois representation, whose existence was proved by Deligne and Serre, is not dihedral. For example, if it is of type A_5 then it is shown for $p\nmid N$ that

$$\lambda_f(p^{12})-\lambda_f(p^8)-\chi(p)\lambda_f(p^2)=1. \tag{4}$$

This relation comes from the recurrence relations satisfied by the Hecke operators and the fact that

$$\chi(p)\lambda_f(p^2) \in \left\{-1, 0, 3, \frac{1+\sqrt{5}}{2}, \frac{1-\sqrt{5}}{2}\right\}.$$

Letting $S(A_5)$ denote the set of f of type A_5 we deduce from (1) taking $X = N$ and using positivity that

$$\sum_{f\in S(A_5)} |\sum_{n\le X} c_n\,\lambda_f(n)|^2 \ll N^{1+\varepsilon} \sum_{n\le N} |c_n|^2. \tag{5}$$

Choosing $c_{p^{12}} = 1, c_{p^8} = -1, c_{p^2} = -\chi(p)$ for primes p and all other $c_n = 0$, by means of (4) the prime number theorem gives

$$\sum_{n\le N} c_n\lambda_f(n) \sim \frac{12N^{1/12}}{\log N} \qquad \text{for } f \in S(A_5),$$

while

$$\sum_{n\le N} |c_n|^2 \sim \frac{36N^{1/12}}{\log N}.$$

Hence we get from (5) the bound

$$\#S(A_5) \ll N^{11/12+\varepsilon}.$$

Similar arguments give better bounds for the other non-dihedral forms. In particular, we get that the number of S_4–forms is $\ll N^{7/8+\varepsilon}$ and then Theorem 2 follows from the classification of quartic fields by S_4–Galois representations (see [Se1]) and Tunnell's proof of the Artin conjecture for them.

The proof of Theorem 3 makes use of the simpler Hecke relation

$$\lambda_f(p)^2 - \lambda_f(p^2) = 1$$

for unramified p in combination with (2) and (3). Of crucial importance is the essential independence of the level of these relations. After using positivity to restrict to modular elliptic curves, assuming the equality of just a few traces of Frobenius is enough to produce a contradiction in the approximate orthogonality of (2) and (3) by expanding their number through multiplicativity.

REFERENCES

[B-S] A. Brumer and J. Silverman, *The number of elliptic curves over Q with conductor N*, Manuscripta Math. 91, 95–102 (1996).

[D-H] H. Davenport and H. Heilbronn, *On the density of discriminants of cubic fields. II.* Proc. Roy. Soc. London Ser. A 322 (1971), no. 1551, 405–420.

[D-S] P. Deligne and J–P. Serre, *Formes modulaires de poids 1,* Ann. Sci. Ec. Norm. Sup. 7 (1974) 507–530, in Serre's Collected Papers, III, 193–216.

[D-I] J.-M. Deshouillers and H. Iwaniec, *Kloosterman sums and Fourier coefficients of cusp forms.* Invent. Math. 70 (1982/83), no. 2, 219–288.

[Du] W. Duke, *The dimension of the space of cusp forms of weight one*, Internat. Math. Res. Notices 1995, no. 2, 99–109

[D-F-I] W. Duke, J.B. Friedlander and H. Iwaniec, *H. Bounds for automorphic L-functions. I*, Invent. Math. 112 (1993), no. 1, 1–8, II. Invent. Math. 115 (1994), no. 2, 219–239.

[D-K] W. Duke and E. Kowalski, *A problem of Linnik for elliptic curves and mean-value estimates for automorphic representations*, appendix by D. Ramakrishnan, to appear in Inventiones Math..

[F-N-T] É. Fouvry, M. Nair and G. Tenenbaum, *L'ensemble exceptionnel dans la conjecture de Szpiro*, Bull. Soc. Math. France 120 (1992) no 4, 483-506.

[G-J] Gelbart, S. and Jacquet, H.: A relation between automorphic representations of $GL(2)$ and $GL(3)$, Ann. Sci. E.N.S 4ème série 11, 471-552 (1978).

[He] H. Heilbronn, *On the* 2*–classgroup of cubic fields,* in Studies in Pure Math. (L. Mirsky, ed.), Academic Press 1971, 117–119.

[J-P-S] H. Jacquet, I.I. Piatetskii-Shapiro, and J.A. Shalika, *Rankin-Selberg convolutions*, Amer. Jour. of Math. 105, 367-464 (1983).

[M-W] C. Moeglin and J.L. Waldspurger, *Pôles des fonctions L de paires pour $GL(N)$, appendice to Le spectre résiduel de $GL(n)$*, Ann. Sci. ENS (4ème série) 22, 605-674 (1989).

[P-S] R. Phillips and P. Sarnak, *Cusp forms for character varieties*, Geom. Funct. Anal. 4 (1994), no. 1, 93–118.

[S-X] P. Sarnak, and X. Xue, *Bounds for multiplicities of automorphic representations*, Duke Math. J. 64 (1991), no. 1, 207–227.

[Sch] W. Schmidt, *Number fields of given degree and bounded discriminant*, Columbia University Number Theory Seminar (New York, 1992). Astérisque No. 228 (1995), 4, 189–195.

[Se1] J–P Serre, *Modular forms of weight one and Galois representations*, Algebraic Number Fields, ed. by A. Fröhlich, Academic Press 1977, 193–268, in Serre's Collected Papers, III, 292–367.

[Se2] J–P. Serre, *Quelques applications du théorème de densité de Chebotarev*, Pub. Math. I.H.E.S 54, 123-201 (1981).

[Tu] J. Tunnell, *Artin's conjecture for representations of octahedral type,* Bull. A.M.S. 5 (1981), 173–175.

[Wi] A. Wiles, *Modular elliptic curves and Fermat's last theorem*, Ann. of Math. (2) 141, 443-551 (1995).

[Wol] S.A. Wolpert, *Disappearance of cusp forms in special families*, Ann. of Math. (2) 139 (1994), no. 2, 239–291.

[Wo1] S. Wong, *Automorphic forms on $GL(2)$ and the rank of class groups*, Preprint.

[Wo2] S. Wong, *On the rank of ideal class groups*, Preprint.

William Duke
Department of Mathematics
Hill Center
Rutgers University
110 Frelinghuysen RD
Piscataway, NJ 08854-8019 USA
duke@math.rutgers.edu

Quelques Résultats d'Indépendance Algébrique

François Gramain

Abstract. We describe a theorem of Yu. Nesterenko [Nes1] on algebraic independence of values of Eisenstein series, together with some corollaries. The most impressive of these is the algebraic independence of π, e^π and $\Gamma(1/4)$. Finally we give some indications on new methods of proving algebraic independence.

1991 Mathematics Subject Classification: 11J85, 11J89, 11J91, 11F11
Keywords and Phrases: Algebraic Independence, Modular Functions

La théorie des nombres transcendants a donné lieu ces dernières années à de nombreux résultats importants. Cependant, nous nous limiterons ici à la description du théorème de Yu. Nesterenko sur l'indépendance algébrique de valeurs de fonctions modulaires et à quelques corollaires. Un dernier paragraphe sera consacré aux méthodes apparues récemment dans les preuves d'indépendance algébrique. Bien que beaucoup des résultats cités aient des analogues $p-$adiques, nous parlerons surtout du cas complexe.

1. Fonctions elliptiques et modulaires : notations

Les résultats classiques rappelés dans ce paragraphe sont traités en détail dans [Cha], [Ser], [Lan1] et [Lan2].

Soit $L = \mathbf{Z}\omega_1 + \mathbf{Z}\omega_2$ le réseau de $\mathbf{C}$ engendré par les nombres complexes ω_1 et ω_2 linéairement indépendants sur $\mathbf{R}$. Dans toute la suite on supposera que $\tau = \omega_2/\omega_1$ est dans le demi-plan de Poincaré $\mathcal{H} = \{\tau \in \mathbf{C}\,;\, \mathrm{Im}(\tau) > 0\}$. Le groupe quotient $\mathbf{C}/L$ est la courbe elliptique paramétrée par $(1 : \wp(z) : \wp'(z))$ dans $\mathbf{P}_2(\mathbf{C})$, où $\wp(z) = z^{-2} + \sum_{\omega\in L\setminus\{0\}} \left((z-\omega)^{-2} - \omega^{-2}\right)$ est la fonction de Weierstrass associée au réseau L. Son équation affine est

$$y^2 = 4\,x^3 - g_2(L)\,x - g_3(L),$$

où les invariants $g_2(L)$ et $g_3(L)$ sont donnés par

$$g_2(L) = 60 \sum_{\omega\in L\setminus\{0\}} \omega^{-4} \quad \text{et} \quad g_3(L) = 140 \sum_{\omega\in L\setminus\{0\}} \omega^{-6}\,.$$

L'ensemble des périodes de la fonction $\wp$ est le réseau L et, à toute période $\omega \in L$ est associée une quasi-période $\eta = \eta(\omega)$ par $\eta = \zeta(z+\omega) - \zeta(z)$, où la fonction ζ est une primitive de $-\wp$.

À chaque réseau L est associé son invariant $j(L) = 1728\,\dfrac{g_2^3(L)}{g_2^3(L) - 27\,g_3^2(L)}$,

et deux courbes elliptiques $\mathbf{C}/L$ et $\mathbf{C}/M$ sont (analytiquement) isomorphes si et seulement si $j(L) = j(M)$.

Par homogénéité, l'invariant j est en fait une fonction de $\tau \in \mathcal{H}$ qui satisfait

$$j(\tau+1) = j(\tau) \quad \text{et} \quad j(-1/\tau) = j(\tau)\,, \text{ donc}$$
$$j\left(\frac{a\tau+b}{c\tau+d}\right) = j(\tau) \text{ pour tout } \tau \in \mathcal{H} \text{ et tout } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbf{Z}).$$

Plus généralement, une fonction f holomorphe sur le demi-plan supérieur $\mathcal{H} = \{\tau \in \mathbf{C}\,;\, \mathrm{Im}(\tau) > 0\}$ est modulaire de poids $2k$ ($k \in \mathbf{N}$) si

$$f\left(\frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^{2k} f(\tau) \text{ pour tout } \tau \in \mathcal{H} \text{ et tout } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbf{Z}).$$

En particulier on a $f(\tau+1) = f(\tau)$, de sorte que f possède un développement de Fourier à l'infini $f(\tau) = F(z) = \sum_{n\in\mathbf{Z}} a_n\, z^n$, où $z = e^{2i\pi\tau}$.
Si la série F a seulement un pôle en 0, on dit que f est une fonction modulaire. Ainsi l'invariant modulaire j est une fonction modulaire de poids 0 et

$$j(\tau) = J(z) = z^{-1} + 744 + 196\,884\, z + 21\,493\,760\, z^2 + \textstyle\sum_{n\geq 3} c(n) z^n,$$

où la série J est convergente dans le disque unité pointé $\mathcal{D} = \{z \in \overline{\mathbf{C}}\,;\, 0 < |z| < 1\}$.

Une forme modulaire est une fonction modulaire qui est holomorphe sur $\mathcal{H}$ et à l'infini de sorte que la série F est entière. L'algèbre graduée des formes modulaires est engendrée par les séries d'Eisenstein Q et R (notations de Ramanujan) :
Si, pour $n \in \mathbf{N}$, on pose $\sigma_k(n) = \sum_{d|n} d^k$, alors, en notant $z = e^{2i\pi\tau}$ où $\tau \in \mathcal{H}$, $Q(z) = E_4(z) = 1 + 240 \sum_{n\geq 1} \sigma_3(n)\, z^n = 12\,(2\pi)^{-4}\, g_2(\tau)$ est une forme modulaire de poids 4 et $R(z) = E_6(z) = 1 - 504 \sum_{n\geq 1} \sigma_5(n)\, z^n = 216\,(2\pi)^{-6}\, g_3(\tau)$ est une forme modulaire de poids 6. On a noté $g_i(\tau) = g_i(\mathbf{Z} + \tau\mathbf{Z})$, de sorte que la formule donnant j en fonction de g_2 et g_3 s'écrit $J = 1728\, Q^3\,/(Q^3 - R^2)$.

L'opérateur différentiel $\Theta = z\,\dfrac{d}{dz} = \dfrac{1}{2i\pi}\,\dfrac{d}{d\tau}$ a une grande importance dans la théorie des formes modulaires bien qu'il fasse intervenir une série d'Eisenstein qui n'est pas une forme modulaire : $P(z) = E_2(z) = 1 - 24 \sum_{n\geq 1} \sigma_1(n)\, z^n$ ne définit pas une forme modulaire de poids 2, mais presque, comme le montre la formule $P(e^{-2i\pi/\tau}) = \tau^2\, P(e^{2i\pi\tau}) + 6\,\tau/(i\,\pi)$.

Les séries d'Eisenstein sont liées par le système différentiel

$$12\,\Theta P = P^2 - Q\,, \qquad 3\,\Theta Q = PQ - R\,, \qquad 2\,\Theta R = PR - Q^2.$$

Comme conséquence immédiate de ces relations on obtient

$$\frac{\Theta J}{J} = -\frac{R}{Q} \qquad\qquad 6\,\frac{\Theta^2 J}{\Theta J} = P - \frac{4R}{Q} - \frac{3Q^2}{R}$$

et les relations réciproques

$$P = 6\,\frac{\Theta^2 J}{\Theta J} - 4\,\frac{\Theta J}{J} - 3\,\frac{\Theta J}{J - 1728}\,, \quad Q = \frac{(\Theta J)^2}{J\,(J-1728)}\,, \quad R = \frac{-(\Theta J)^3}{J^2\,(J-1728)}\,,$$

de sorte que, en notant $f(z)$ la fonction $z \mapsto f(z)$, on a l'identité des corps
$\mathbf{Q}(z,\, P(z),\, Q(z),\, R(z)) = \mathbf{Q}(z,\, J(z),\, \Theta J(z),\, \Theta^2 J(z)) = \mathbf{Q}(z,\, J(z),\, J'(z),\, J''(z))$.
Quand on passe des fonctions aux valeurs qu'elles prennent en un point, les facteurs J et $J - 1728$ en dénominateur obligent à exclure quelques points : pour $\tau \not\equiv i$ et $\tau \not\equiv \rho = e^{2i\pi/3}$ mod $SL_2(\mathbf{Z})$, les corps engendrés sur $\mathbf{Q}$ par les nombres $q = e^{2i\pi\tau}$, $P(q)$, $Q(q)$ et $R(q)$ ou par $e^{2i\pi\tau}$, $j(\tau)$, $j'(\tau)/\pi$ et $j''(\tau)/\pi^2$ ont le même degré de transcendance sur $\mathbf{Q}$.

Enfin, les valeurs des fonctions P, Q et R sont liées aux périodes et quasi-périodes des fonctions de Weierstrass : Soient $\wp$ la fonction elliptique de Weierstrass associée au réseau $L = \mathbf{Z}\omega_1 + \mathbf{Z}\omega_2$, où $\tau = \omega_2/\omega_1$ est dans $\mathcal{H}$, et η_1 la quasi-période associée à ω_1, alors on a :

$$P(q) = 3\,(\omega_1/\pi)\,(\eta_1/\pi)\,,\ 4\,Q(q) = 3\,(\omega_1/\pi)^4\,g_2(L)\,,\ 8\,R(q) = 27\,(\omega_1/\pi)^6\,g_3(L)\,.$$

2. LE THÉORÈME DE YU. NESTERENKO

Le premier résultat de transcendance concernant l'invariant modulaire j (ou J) obtenu par des méthodes modulaires est le

THÉORÈME STÉPHANOIS ([BDGP] 1995). *Si $q \in \mathcal{D}$ est algébrique, alors $J(q)$ est transcendant.*

Ce résultat avait été conjecturé par K. Mahler ([Mah1] 1969) et sa preuve s'adapte sans difficulté au cas $p-$adique conjecturé en 1971 par Yu. V. Manin [Man]. Il est contenu dans le théorème de Yu. Nesterenko obtenu moins d'un an plus tard :

THÉORÈME 1 (YU. NESTERENKO 1996). *Pour $q \in \mathcal{D} = \{z \in \mathbf{C}\,;\, 0 < |z| < 1\}$ on a* $\operatorname{degtr}_{\mathbf{Q}}\,\mathbf{Q}\,(q,\,P(q),\,Q(q),\,R(q)) \geq 3$.

D'après le paragraphe précédent, cet énoncé est équivalent au suivant :

THÉORÈME 1'. *Pour tout réseau $L = \mathbf{Z}\omega_1 + \mathbf{Z}\omega_2$ on a*

$$\operatorname{degtr}_{\mathbf{Q}}\,\mathbf{Q}\left(g_2(L),\,g_3(L),\,\omega_1/\pi,\,\eta_1/\pi,\,e^{2i\pi\omega_2/\omega_1}\right) \geq 3\,.$$

Comme pour le théorème stéphanois, la preuve du Théorème 1 n'utilise que les fonctions modulaires (et pas du tout les fonctions elliptiques), alors que les autres résultats antérieurs qu'il contient ont été démontrés par des procédés elliptiques. En voici deux exemples : le résultat suivant (dû à Th. Schneider, 1937, voir [Sch]) *Si $\omega \neq 0$ est une période d'une fonction $\wp$ de Weierstrass d'invariants algébriques g_2 et g_3, alors ω/π est un nombre transcendant*
est équivalent à ce corollaire du théorème de Nesterenko :
Pour $q \in \mathcal{D}$ les nombres $Q(q)$ et $R(q)$ ne sont pas tous deux algébriques
et il implique le suivant :
Pour $q \in \mathcal{D}$ et $J(q) \notin \{0,\,1728\}$ les nombres $J(q)$ et $qJ'(q)$ ne sont pas simultanément algébriques.

D. Bertrand a été le premier à étudier systématiquement la correspondance entre énoncés elliptiques et énoncés modulaires. C'est l'équivalence précédente qui l'a inspiré en 1975 ([Ber1]) pour obtenir l'analogue $p-$adique du résultat de Th. Schneider en utilisant les fonctions elliptiques de Jacobi-Tate au lieu des fonctions de Weierstrass. Notons que K. Barré [Bar1] a obtenu en 1995 le dernier résultat cité dans les cas complexe et $p-$adique par une preuve purement modulaire.

De façon analogue, ce résultat de G. V. Chudnovsky (1977, voir [Chu])
Si η est la quasi-période associée à la période $\omega \neq 0$ de la fonction elliptique $\wp$ de Weierstrass d'invariants g_2 et g_3, alors $\operatorname{degtr}_{\mathbf{Q}}\,\mathbf{Q}\,(g_2,\,g_3,\,\omega/\pi,\,\eta/\omega) \geq 2$
est équivalent à
Pour $q \in \mathcal{D}$ on a $\operatorname{degtr}_{\mathbf{Q}}\,\mathbf{Q}\,(P(q),\,Q(q),\,R(q)) \geq 2$.
dont la version $p-$adique est due à D. Bertrand [Ber2]. Pour un exposé complet de la correspondance entre théorèmes (et conjectures) elliptiques et modulaires, on pourra consulter [Dia2], ainsi que [Dia1] pour des liens étonnants avec la fonction exponentielle.

Venons-en à quelques corollaires qui donnent des résultats nouveaux :

COROLLAIRE 1. *Si $\tau \not\equiv i,\,\rho \bmod SL_2(\mathbf{Z})$, alors le degré de transcendance sur $\mathbf{Q}$ des corps $\mathbf{Q}(q,\,J(q),\,J'(q),\,J''(q))$ et $\mathbf{Q}\left(e^{2i\pi\tau},\,j(\tau),\,j'(\tau)/\pi,\,j''(\tau)/\pi^2\right)$ est au moins égal à 3.*

En particulier

Corollaire 2. *Si $q \in \mathcal{D}$ est algébrique (dans ce cas τ est transcendant d'après le théorème de Gel'fond-Schneider), alors les trois nombres $P(q)$, $Q(q)$ et $R(q)$ sont algébriquement indépendants sur $\mathbf{Q}$, de même que $J(q)$, $J'(q)$ et $J''(q)$.*

Corollaire 3. *Si $\wp$ est une fonction elliptique d'invariants g_2 et g_3 algébriques, alors les 3 nombres $e^{2i\pi\tau}$, ω_1/π et η_1/π sont algébriquement indépendants sur $\mathbf{Q}$.*

Dans le cas de la multiplication complexe (c'est-à-dire lorsque τ est un nombre algébrique quadratique), grâce aux relations de Legendre et de D. W. Masser ([Mas], lemme 3.1), on obtient le

Corollaire 4. *Si $\wp$ est une fonction elliptique d'invariants g_2 et g_3 algébriques et à multiplication complexe, pour toute période $\omega \neq 0$ de $\wp$, chacun des deux triplets $\left\{e^{2i\pi\tau}, \omega, \eta\right\}$ et $\left\{e^{2i\pi\tau}, \omega, \pi\right\}$ est constitué de nombres algébriquement indépendants sur $\mathbf{Q}$.*

En particulier, pour $\tau = i$ (resp. $\tau = \rho$) la courbe elliptique d'équation $y^2 = 4\,x^3 - 4\,x$ (resp. $y^2 = 4\,x^3 - 4$) est associée à un réseau dont une période est $\omega = \Gamma(1/4)^2 \,/\, \sqrt{8\pi}$ (resp. $\omega = \Gamma(1/3)^3 \,/\, (2^{4/3}\pi)$), de sorte que :

Corollaire 5. *Les trois nombres π, e^{π} et $\Gamma(1/4)$ sont algébriquement indépendants sur $\mathbf{Q}$, de même que π, $e^{\pi\sqrt{3}}$ et $\Gamma(1/3)$.*

C'est par ce biais un peu surprenant que l'on obtient

l'indépendance algébrique de π et e^{π}

grâce à une démonstration qui ne fait pas intervenir la fonction exponentielle !

L'usage de la fonction thêta de Weierstrass-Jacobi (voir [Ber3] et [Ber4])

$$\theta(\tau, w) = \sum_{n\in\mathbf{Z}} e^{i\pi n^2\tau}\, e^{2i\pi n w}$$

donne d'autres corollaires. Par spécialisation de cette fonction thêta on obtient les classiques fonctions thêta de Jacobi

$$\theta_2(z) = 2\,z^{1/4} \sum_{n\geq 0} z^{n(n+1)}, \quad \theta_3(z) = \sum_{n\in\mathbf{Z}} z^{n^2} \text{ et } \quad \theta_4(z) = \theta_3(-z)$$

où, comme plus haut, $z = e^{2i\pi\tau}$ est dans le disque unité. Ces fonctions sont liées aux séries d'Eisenstein par des relations du type

$$2\,Q(z^2) = \theta_2(z)^8 + \theta_3(z)^8 + \theta_4(z)^8\,,$$
$$P(z^2) = 4\left(\frac{\Theta\theta_2}{\theta_2} + \frac{\Theta\theta_3}{\theta_3} + \frac{\Theta\theta_4}{\theta_4}\right)(z)\,,$$

qui font que tout résultat d'indépendance algébrique sur des valeurs des fonctions P, Q, R donne un résultat analogue concernant les valeurs des fonctions θ_i ou de leurs dérivées. Ainsi l'énoncé du théorème 1 est équivalent au suivant [Ber3] :

Théorème 1". *Pour i, j et $k \in \{2, 3, 4\}$ tels que $i \neq j$ et pour $q \in \mathcal{D}$ on a*

$$\operatorname{degtr}_{\mathbf{Q}}\, \mathbf{Q}\,(q, \theta_i(q), \theta_j(q), \Theta\theta_k(q)) \geq 3$$
$$\textit{resp.}\ \operatorname{degtr}_{\mathbf{Q}}\, \mathbf{Q}\left(q, \theta_k(q), \Theta\theta_k(q), \Theta^2\theta_k(q)\right) \geq 3\,.$$

dont un cas particulier est le

Corollaire 6. *Si $\alpha \in \mathcal{D}$ est algébrique, les trois nombres $\sum_{n\geq 1} \alpha^{n^2}$, $\sum_{n\geq 1} n^2\,\alpha^{n^2}$ et $\sum_{n\geq 1} n^4\,\alpha^{n^2}$ sont algébriquement indépendants sur $\mathbf{Q}$.*

L'utilisation de la fonction de 2 variables $\theta(\tau, w)$, où τ prend en compte l'aspect modulaire et w l'aspect elliptique des nombres étudiés est peut-être un moyen de ne pas dissocier ces deux aspects des problèmes d'indépendance algébrique. D. Bertrand [Ber4] propose plusieurs conjectures dans cette direction.

Enfin, grâce à un argument de spécialisation dû à A. Weil, D. Duverney,

K. et K. Nishioka et I. Shiokawa [DNNS] retrouvent la deuxième assertion du théorème 1" ; dans le même article, ils prouvent la transcendance d'un certain nombre de sommes de séries construites à partir de suites récurrentes en les liant à des valeurs de la fonction modulaire $\Delta = (Q^3 - R^2)/1728$, par exemple :

COROLLAIRE 7. *Si $\{u_n\}$ est la suite de Fibonacci ($u_0 = 0$, $u_1 = 1$, $u_{n+2} = u_{n+1} + u_n$), les nombres $\sum_{n\geq 1} u_n^{-2}$, $\sum_{n\geq 1}(-1)^n u_n^{-2}$, $\sum_{n\geq 1} u_{2n-1}^{-1}$ et $\sum_{n\geq 1} n\, u_{2n}^{-1}$ sont transcendants.*

3. La preuve de Yu. Nesterenko

Une preuve complète du théorème de Yuri Nesterenko se trouve, en dehors de l'article original ([Nes1] pour l'annonce et [Nes2] pour les démonstrations), dans les exposés de Michel Waldschmidt au Séminaire Bourbaki [Wal1] et à Carleton [Wal2]. Pour les résultats quantitatifs les plus récents, on pourra consulter [Nes3].

Pour appliquer un critère d'indépendance algébrique de Patrice Philippon [Phi1], on construit une suite de polynômes $A_N \in \mathbf{Z}[z, X_1, X_2, X_3]$ telle que $|A_N(q, P(q), Q(q), R(q))|$ soit petit, avec un contrôle des degrés et des hauteurs (*la hauteur $H(P)$ du polynôme P est le maximum des modules de ses coefficients*) des A_N :

PREMIER PAS. Une fonction auxiliaire

Pour $N \in \mathbf{N}$ suffisamment grand, le principe des tiroirs (lemme de Siegel) fournit un polynôme non nul $A \in \mathbf{Z}[z, X_1, X_2, X_3]$ tel que:
les degrés partiels de A sont $\leq N$; $\log H(A) \leq 116\, N \log N$;
et la fonction F définie par $F(z) = A(z, P(z), Q(z), R(z))$ possède en 0 un zéro d'ordre $M \geq \frac{1}{2} N^4$.

La fonction F n'est pas identiquement nulle car les fonctions z, $P(z)$, $Q(z)$ et $R(z)$ sont algébriquement indépendantes sur $\mathbf{C}$ ([Mah2]), de sorte que M est correctement défini. On utilise alors le fait que les séries d'Eisenstein ont des coefficients entiers dont la croissance est polynomiale : la borne (grossière) $\sigma_k(n) = \sum_{d|n} d^k \leq (\sum_{d|n} d)^k \leq n^{2k}$ permet de majorer le module des coefficients de Taylor de $z^{k_0} P(z)^{k_1} Q(z)^{k_2} R(z)^{k_3}$ par les coefficients de Taylor (de même indice) de $c^N (1-z)^{-22N}$, où $N \geq k_i$ et où c est une constante absolue.

Si $A = \displaystyle\sum_{0\leq k_i\leq N} a(k_0, \underline{k})\, z^{k_0} \underline{X}^{\underline{k}}$ et $F(z) = A(z, P(z), Q(z), R(z)) = \displaystyle\sum_{n\geq 0} b_n\, z^n$,

le système linéaire des $[(N+1)^4/2]$ équations $b_n = 0$ $(0 \leq n < [(N+1)^4/2])$ en les $(N+1)^4$ inconnues $a(k_0, \underline{k})$ a ses coefficients entiers et bornés par le calcul précédent, il possède donc une solution entière non nulle et assez petite.

On a ainsi construit une fonction auxiliaire $F(z) = b_M\, z^M + \sum_{n>M} b_n\, z^n$, où b_M est un entier rationnel non nul.

DEUXIÈME PAS. Majoration de $|F(z)|$

Soit $q \in \mathcal{D}$. La majoration des $|a(k_0, \underline{k})|$ permet de borner les $|b_n|$ et d'obtenir :

Pour N assez grand et $|z| \leq r = \min(\frac{1+|q|}{2}, 2|q|)$, on a $|F(z)| \leq |z|^M\, M^{187N}$.

La condition sur $|z|$ est purement technique ; le fait important est que $|z| \leq r$ avec $|q| < r < 1$ et Yu. Nesterenko choisit un tel r.

TROISIÈME PAS. Minoration d'un $|F^{(T)}(q)|$

Il existe un entier naturel $T \leq c_1(q)\, N \log M$ tel que $|F^{(T)}(q)| > (|q|/2)^{2M}$.

Il suffit d'appliquer la formule des résidus à

$$G(z) \,=\, z^{-M-1}\, F(z)\, \left(\frac{r^2 - q\bar{z}}{r\,(z-q)}\right)^T .$$

En effet, le module de F est borné par le deuxième pas et, par construction, le résidu en 0 de G est $b_M\,(-r/q)^T$, où $|b_M| \,\geq\, 1$ car c'est un nombre entier non nul.

QUATRIÈME PAS. Lemme de zéros

LEMME DE ZÉROS (NESTERENKO). *Soient L_0 et L des nombres entiers ≥ 1. Si $A \in \mathbf{C}[z,\, X_1,\, X_2,\, X_3]$ est un polynôme non nul de degrés $\leq\, L_0$ en z et $\leq\, L$ en chacun des X_i, alors* $\mathrm{ord}_0\, A(z,\, P(z),\, Q(z),\, R(z)) \,\leq\, 2.10^{45}\, L_0\, L^3$.

Le point crucial qui permet d'obtenir un tel lemme de zéros est le système différentiel satisfait par P, Q et R :

$$12\,\Theta P = \, P^2 - Q\,, \qquad 3\,\Theta Q = \, PQ - R\,, \qquad 2\,\Theta R = \, PR - Q^2,$$

mais le z dans $\Theta \,=\overset{\downarrow}{z}\, \dfrac{d}{dz}$ est cause d'une (la ?) sérieuse difficulté : il interdit l'usage des lemmes antérieurs de Yu. Nesterenko et oblige à démontrer un résultat qui est loin d'être trivial :

Tout idéal premier non nul de $\mathbf{C}[z,\, X_1,\, X_2,\, X_3]$ ayant un zéro au point $(0, 1, 1, 1)$ et stable par l'opérateur

$$D \,=\, z\frac{d}{dz} + \frac{1}{12}(X_1^2 - X_2)\frac{\partial}{\partial X_1} + \frac{1}{3}(X_1X_2 - X_3)\frac{\partial}{\partial X_2} + \frac{1}{2}(X_1X_3 - X_2^2)\frac{\partial}{\partial X_3}$$

contient le polynôme $z\,(X_2^3 - X_3^2)$.

La conclusion de ce quatrième pas est que *le paramètre M est majoré par $M \,\leq\, c\,N^4$, où c est une constante absolue.*

CINQUIÈME PAS. Qui est A_N ?

L'opérateur différentiel D ci-dessus a été étudié pour que, compte tenu du système différentiel satisfait par P, Q et R, pour tout $B \in \mathbf{C}[z,\, X_1,\, X_2,\, X_3]$, on ait $\dfrac{d}{dz}\, B(z,\, P(z),\, Q(z),\, R(z)) \,=\, \dfrac{1}{z}\,(DB)(z,\, P(z),\, Q(z),\, R(z))$.

En particulier, pour $F(z) \,=\, A(z,\, P(z),\, Q(z),\, R(z))$, par récurrence sur $t \in \mathbf{N}$, on obtient $z^t\, F^{(t)}(z) \,=\, D_tA(z,\, P(z),\, Q(z),\, R(z))$, où l'opérateur différentiel D_t est défini par $D_t \,=\, \prod_{0\leq k<t}\,(D - k)$.

De plus, il est clair que $12^t\, D_tA$ est, comme A, un polynôme à coefficients dans $\mathbf{Z}$. Ainsi, avec les notations du troisième pas, il existe $A_N \in \mathbf{Z}[z,\, X_1,\, X_2,\, X_3]$ tel que $(12\,z)^T\, F^{(T)}(z) \,=\, A_N(z,\, P(z),\, Q(z),\, R(z))$.

Enfin, la formule $A_N \,=\, 12^T\, D_TA$ permet de majorer le degré $\deg A_N$ et la hauteur $H(A_N)$ de A_N :

Pour tout entier N suffisamment grand, il existe $A_N \in \mathbf{Z}[z,\, X_1,\, X_2,\, X_3]$ tel que

$$\deg A_N \,\leq\, c_2(q)\, N\, \log N,\; \log H(A_N) \,\leq\, c_2(q)\, N\, (\log N)^2 \;\; et$$
$$\exp(-\kappa_2(q)\, N^4) \,\leq\, |A_N(q,\, P(q),\, Q(q),\, R(q))| \,\leq\, \exp(-\kappa_1(q)\, N^4),$$

où les constantes c_2 et κ_i ne dépendent que de q.

SIXIÈME PAS. Conclusion

Un cas particulier du critère d'indépendance algébrique de P. Philippon [Phi1] permet alors de conclure :

CRITÈRE. *Soit $\underline{x} \in \mathbf{C}^m$; s'il existe une suite de polynômes $A_N \in \mathbf{Z}[\underline{X}]$ telle que*

$$\deg A_N \,\leq\, \sigma(N),\; \log H(A_N) \,\leq\, \sigma(N) \;\; et$$
$$\exp(-\kappa_2\, \lambda(N)) \,\leq\, |A_N(\underline{x})| \,\leq\, \exp(-\kappa_1\, \lambda(N)),$$

où les κ_i sont des constantes positives, σ et λ sont des fonctions croissant vers l'infini et satisfaisant $\sigma(N+1)\,/\,\sigma(N) \longrightarrow 1$ et $\lambda(N)\,/\,(\sigma(N))^k \longrightarrow \infty$ quand $N \longrightarrow \infty$, alors $\operatorname{degtr}_{\mathbf{Q}} \mathbf{Q}(\underline{x}) \geq k$.

MESURES. Si l'on choisit avec plus de soin les degrés partiels du polynôme A construit au premier pas de la démonstration, les estimations du cinquième pas sont assez précises pour que l'on puisse utiliser un critère fournissant des mesures. On obtient ainsi des mesures d'indépendance algébrique et des mesures d'approximation, par exemple :

Soit $q \in \mathcal{D}$, si $\underline{\theta} = (\theta_1, \theta_2, \theta_3)$ est une base de transcendance du corps $\mathbf{Q}(q, P(q), Q(q), R(q))$, alors il existe une constante $C > 0$ telle que, pour $B \in \mathbf{Z}[X, Y, Z] \setminus \{0\}$, on ait

$$\log |B(\underline{\theta})| > -C\,(t(B) + \deg B\, \log t(B))^4\,(\log t(B))^9,$$

où $t(B) = \max(e, \deg B + \log H(B))$.

Soient $q \in \mathcal{D}$ et $\underline{x} = (q, P(q), Q(q), R(q))$, alors il existe une constante $C > 0$ telle que, pour tout point algébrique $\underline{\alpha} = (\alpha_i)_{1\leq i\leq 4}$, on ait

$$\log \sum_{1\leq i\leq 4} |x_i - \alpha_i| > -C\,(t(\underline{\alpha})\, \deg(\underline{\alpha}))^{4/3} \log(t(\underline{\alpha})\, \deg(\underline{\alpha})),$$

où $\deg(\underline{\alpha}) = [\mathbf{Q}(\underline{\alpha}) : \mathbf{Q}]$, $t(\underline{\alpha}) = h(\underline{\alpha}) + \log \deg(\underline{\alpha})$ *et $h(\underline{\alpha})$ est la hauteur logarithmique absolue de Weil de $\underline{\alpha}$:* $h(\underline{\alpha}) = \dfrac{1}{\deg(\underline{\alpha})} \displaystyle\sum_v d_v \log^+ \max_{1\leq i\leq 4} |\alpha_i|_v$.

VARIANTE. D'autre part, dans [Phi2] et [Phi3], P. Philippon propose une autre façon de conclure à partir de la construction transcendante pour obtenir un cas particulier (contenant l'indépendance algébrique de π, e^π et $\Gamma(1/4)$) du théorème 1 : le lemme de zéros et le critère sont remplacés par une mesure de transcendance, ce qui introduit d'ailleurs une composante elliptique dans la preuve ; la construction de transcendance est celle qui a été présentée ci-dessus, mais en un peu plus simple : au troisième pas, il suffit d'avoir $F^{(T)}(q) \neq 0$ et, sans utiliser le lemme de zéros, la conclusion du cinquième pas est

$$\deg A_N \leq c(q)\, N \log M,\ \log H(A_N) \leq c(q)\, N\, (\log M)^2 \text{ et}$$
$$0 < |A_N(q, P(q), Q(q), R(q))| \leq \exp(-\kappa(q)\, M).$$

La mesure d'indépendance qu'on utilise alors est une version quantitative [GPhi1] du théorème de G.V. Chudnovsky dont on a parlé plus haut :

THÉORÈME 2. *Soit $\omega \neq 0$ une période d'une fonction elliptique $\wp$ de Weierstrass d'invariants algébriques g_2 et g_3 et soit η la quasi-période associée. Pour tout $\varepsilon > 0$, il existe une constante $c(\varepsilon) > 0$ telle que, pour tout $B \in \mathbf{Z}[X, Y] \setminus \{0\}$ on ait* $|B(\frac{\pi}{\omega}, \frac{\eta}{\omega})| > \exp\left(-c(\varepsilon)\, t(B)^{3+\varepsilon}\right)$.

La mesure annoncée par G.V. Chudnovsky était un peu meilleure que cela et P. Philippon [Phi4] vient d'en donner la première démonstration. Elle prouve, en particulier, que les nombres π/ω et η/ω, et donc $\Gamma(1/4)$, ne sont pas des nombres de Liouville.

Pour utiliser le théorème 2, on suppose que g_2 et g_3 sont algébriques et que $P(q)$, $Q(q)$ et $R(q)$ sont algébriques sur $\mathbf{Q}(\pi/\omega, \eta/\omega)$; alors, on élimine $P(q)$, $Q(q)$ et $R(q)$ entre A_N et leurs polynômes minimaux sur $\mathbf{Q}(\pi/\omega, \eta/\omega)$. On obtient ainsi un $B_N(\pi/\omega, \eta/\omega)$ qui, pour N suffisamment grand, contredit la mesure de G. Philibert. Cela prouve que

Si $J(q)$ est algébrique, alors $\mathrm{degtr}_{\mathbf{Q}}\, \mathbf{Q}\,(q,\, P(q),\, Q(q),\, R(q)) \geq 3$.

Notons enfin que, grâce à un lemme de zéros [GPhi2] pour les fonctions polynomiales en z et $J(z)$, K. Barré obtient [Bar2] des mesures d'approximation simultanée $|q-\alpha| + |J(q)-\beta| > \ldots$ meilleures que celles que l'on déduit des travaux de Yu. Nesterenko, dans le sens qu'elles séparent les contributions de α et β.

4. Nouvelles approches pour l'indépendance algébrique

La plupart des résultats classiques d'indépendance algébrique se déduisent du critère d'indépendance algébrique de P. Philippon [Phi1] dont voici un cas particulier assez représentatif.

CRITÈRE. *Soit $n \geq 1$ un nombre entier. Il existe une constante $C > 0$ ayant la propriété suivante : Soient $\underline{\theta} \in \mathbf{C}^n$ et $\eta > 0$; si, pour tout N entier suffisamment grand il existe un entier $m = m(N)$ et des polynômes $Q_{N_j} \in \mathbf{Z}[X_1, \ldots, X_n]$ $(1 \leq j \leq m)$ tels que*

$$\deg Q_{N_j} \leq N,\ H(Q_{N_j}) \leq e^N \text{ et } 0 < |Q_{N_j}(\underline{\theta})| \leq \exp(-C\,N^{\eta}),$$

et que, pour chaque N, le nombre des zéros communs aux Q_{N_j} $(1 \leq j \leq m)$ dans la boule $\{z \in \mathbf{C}^n\,;\, \max_{1\leq i\leq n} |z_i - \theta_i| \leq \exp(-3\,C\,N^{\eta})\}$ soit fini, alors

$$\mathrm{degtr}_{\mathbf{Q}}\, \mathbf{Q}(\underline{\theta}) > \eta - 1.$$

Un tel résultat est obtenu par des méthodes d'élimination (algèbre commutative). Les travaux récents de M. Laurent, D. Roy et M. Waldschmidt ont introduit un autre point de vue, qui apparaît aussi dans [Phi2] sous une forme un peu différente. Il consiste à remplacer le critère par des propriétés d'approximation du type suivant :

CONJECTURE. *Soient a et b des nombres réels ≥ 1. Il existe un nombre réel c ayant la propriété suivante : soient $\underline{\theta} \in \mathbf{C}^n$ et $t = \mathrm{degtr}_{\mathbf{Q}}\, \mathbf{Q}(\underline{\theta})$; soient $(D_N)_{N\in\mathbf{N}}$ et $(h_N)_{N\in\mathbf{N}}$ des suites de nombres réels positifs telles que $D_N + h_N$ n'est pas borné et que $c \leq h_N \leq h_{N+1} \leq a\,h_N$ et $c \leq D_N \leq D_{N+1} \leq b\,D_N$, alors pour une infinité de N il existe un point $\underline{\alpha} \in \mathbf{C}^n$ à coordonnées algébriques tel que $c^{-1}\,D_N \leq \deg(\underline{\alpha}) \leq D_N$, $h(\underline{\alpha}) \leq h_N$ et*

$$\max_{1\leq i\leq n} |\theta_i - \alpha_i| \leq \exp\left(-c^{-1}\,h_N\,D_N^{1+1/t}\right).$$

En fait, un tel énoncé permet de retrouver le critère : supposons les hypothèses du critère réalisées et le degré de transcendance de $\mathbf{Q}(\underline{\theta})$ petit. Alors la conjecture fournit de bonnes approximations algébriques de $\underline{\theta}$ par des nombres algébriques $\underline{\alpha}$, de sorte que les nombres algébriques $|Q_{N_j}(\underline{\alpha})|$, qui sont proches des $|Q_{N_j}(\underline{\theta})|$, sont petits. L'inégalité de Liouville montre alors que les $Q_{N_j}(\underline{\alpha})$ sont nuls. Cela contredit l'hypothèse de la version faible du critère où l'on suppose que les Q_{N_j} n'ont pas de zéro commun dans la boule considérée. La minoration de $\deg(\underline{\alpha})$ permet de traiter le cas général d'un nombre fini de zéros communs.

Actuellement la conjecture est démontrée pour $t = 1$, ce qui donne l'indépendance algébrique de deux nombres à partir de mesures d'approximation simultanée (voir en particulier [Roy-Wal1]). L'approche de P. Philippon (approximation par des cycles au lieu de points) permet d'atteindre, pour l'instant, le degré de transcendance 3.

Ces approches ont l'avantage de permettre l'utilisation de déterminants d'interpolation au lieu de fonctions auxiliaires pour obtenir des résultats d'indépendance algébrique. Il en est de même de la généralisation du critère (obtenue par M.

Laurent et D. Roy) faisant intervenir des multiplicités, c'est-à-dire tenant compte de petites valeurs des dérivées des polynômes Q_{N_j} au point $\underline{\theta}$. Gageons que le fruit de ces travaux en cours fera l'objet d'un exposé à ICM'02.

Bibliographie

[Bar1] K. Barré. – *Propriétés de transcendance des séries d'Eisenstein* ; Séminaire de Théorie des nombres de Paris 1994-1995 (à paraître).
[Bar2] K. Barré. – *Mesure d'approximation simultanée de q et J(q)* ; J. Number Theory 66 (1997), 102–128.
[BDGP] K. Barré-Sirieix, G. Diaz, F. Gramain et G. Philibert. – *Une preuve de la conjecture de Mahler-Manin* ; Invent. Math. 124 (1996), 1–9.
[Ber1] D. Bertrand. – *Séries d'Eisenstein et transcendance* ; Bull. Soc. Math. France 104 (1976), 309–321.
[Ber2] D. Bertrand. – *Fonctions modulaires, courbes de Tate et indépendance algébrique* ; Séminaire Delange-Pisot-Poitou, Paris, 19ème année (1977/78), exposé 36, 11p.
[Ber3] D. Bertrand. – *Theta functions and transcendence* ; Madras Number Theory Symposium 1996, The Ramanujan J. Math. 1 (1997), 339–350.
[Ber4] D. Bertrand. – *$\theta(\tau, z)$ and transcendence* ; in *Introduction to algebraic independence theory*, Yu. Nesterenko et P. Philippon éd., en préparation.
[Cha] K. Chandrasekharan. – Elliptic functions ; Grund. der math. Wiss. 281, Springer-Verlag 1985.
[Chu] G.V. Chudnovsky. – Contributions to the theory of transcendental numbers ; Math. Surveys and Monographs 19, Amer. Math. Soc., 1984, 450p.
[Dia1] G. Diaz. – *La conjecture des quatre exponentielles et les conjectures de D. Bertrand sur la fonction modulaire* ; J. Théor. Nombres Bordeaux 9 (1997), 229–245.
[Dia2] G. Diaz. – *Transcendence and algebraic independence : elliptic and modular points of vue* ; en préparation.
[DNNS] D. Duverney, K. Nishioka, K. Nishioka and I. Shiokawa. – *Transcendence of Jacobi's theta series and related results* ; Proc. Conf. Number Theory Eger 1996, K. Györy, A. Pethö and V.T. Sós eds, de Gruyter, Berlin (1998).
[GPhi1] G. Philibert. – *Une mesure d'indépendance algébrique* ; Ann. Inst. Fourier (Grenoble), 38 (1988), 85–103.
[GPhi2] G. Philibert. – *Un lemme de zéros modulaire* ; J. Number Theory 66 (1997), 306–313.
[Lan1] S. Lang. – Elliptic functions ; Addison-Wesley, Reading, MA, 1973 ; seconde édition, GTM 112, Springer-Verlag 1987.
[Lan2] S. Lang. – Introduction to Modular Forms ; Grund. der math. Wiss. 222, Springer-Verlag 1976.
[Lau-Roy] M. Laurent et D. Roy. – *Sur l'approximation algébrique en degré de transcendance un* ; à paraître.
[Mah1] K. Mahler. – *Remarks on a paper by Wolfgang Schwarz* ; J. Number Theory 1 (1969), 512–521.
[Mah2] K. Mahler. – *On algebraic differential equations satisfied by automorphic functions* ; J. Austral. Math. Soc. 10 (1969), 445–450.

[Man] Yu.I. Manin. – *Cyclotomic fields and modular curves* ; Uspekhi Mat. Nauk 26 (1971), 7–71 [en russe] ; trad. angl. : Russian Math. Surveys 26 (1971), 7–78.
[Mas] D.W. Masser. – Elliptic Functions and Transcendence ; L. N. in Math. 437, Springer-Verlag 1975.
[Nes1] Yu.V. Nesterenko. – *Modular functions and transcendence problems – Un théorème de transcendance sur les fonctions modulaires* ; C. R. Acad. Sci. Paris, Série I, 322 (1996), 909–914.
[Nes2] Yu.V. Nesterenko. – *Modular functions and transcendence questions* ; Math. Sb. 187 N° 9 (1996), 65–96 [en russe] ; trad. angl. : Math. USSR Sb. 187 (1996) 1319–1348.
[Nes3] Yu.V. Nesterenko. – *On a measure of algebraic independence of values of Ramanujan's functions* ; Trudy Math. Inst. Steklov, 218, 1997, 299–334 ; trad. angl. : Proc. Steklov Inst. Math., 218, 1997, 294–331.
[Phi1] P. Philippon. – *Critères pour l'indépendance algébrique* ; Inst. Hautes Études Sci. Publ. Math. 64 (1987), 5–52.
[Phi2] P. Philippon. – *Une approche méthodique pour la transcendance et l'indépendance algébrique de valeurs de fonctions analytiques* ; J. Number Theory 64 (1997), 291–338.
[Phi3] P. Philippon. – *Indépendance algébrique et K-fonctions* ; J. reine angew. Math., 497 (1998), 1–15.
[Phi4] P. Philippon. – *Mesures d'approximation de valeurs de fonctions analytiques* ; soumis.
[Roy-Wal1] D. Roy et M. Waldschmidt. – *Approximation diophantienne et indépendance algébrique de logarithmes* ; Ann. scient. Éc. Norm. Sup. 30 (1997), n° 6, 753–796.
[Roy-Wal2] D. Roy and M. Waldschmidt. – *Simultaneous approximation and algebraic independence* ; The Ramanujan Journal 1 (1997), n° 4, 379–430.
[Sch] Th. Schneider. – Einführung in die transzendenten Zahlen ; Springer-Verlag 1957 ; trad. franç. : Introduction aux nombres transcendants, Gauthier-Villars, Paris, 1959.
[Ser] J.-P. Serre. – Cours d'arithmétique ; Presses Univ. France, Paris, 1970 ; trad. angl. : A course in arithmetic, GTM 7, Springer-Verlag 1973.
[Wal1] M. Waldschmidt. – *Sur la nature arithmétique des valeurs de fonctions modulaires* ; Sém. Bourbaki, 49ème année, 1996/97, n○ 824 ; Soc. Math. France, Astérisque 245 (1997), 105–140.
[Wal2] M. Waldschmidt. – *Transcendance et indépendance algébrique de valeurs de fonctions modulaires* ; CNTA5, Carleton 1996 ; Proceedings of the fifth Conference of the Canadian Number Theory Association, éd. : R. Gupta et K. Williams, (à paraître).

François Gramain
Faculté des Sciences
23, rue du Docteur Paul Michelon
F-42023 St Etienne CEDEX 2
Mél. : gramain@univ-st-etienne.fr

Points Rationnels et Séries de Dirichlet

Loïc Merel

1 Le problème de la torsion des courbes elliptiques

En 1908, lors du congrès international des mathématiciens, B. Levi a proposé (au vocabulaire près) la conjecture suivante : les points $\mathbf{Q}$-rationnels d'ordre fini d'une courbe elliptique forment un groupe isomorphe à l'un des groupes suivants : $\mathbf{Z}/n\mathbf{Z}$ (avec $n \in \{1,2,3,4,5,6,7,8,10\}$), $\mathbf{Z}/n\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$ (avec $n \in \{1,2,3,4\}$). La précision de cette conjecture est d'autant plus étonnante que cette formulation précède le théorème de Mordell. Plus de soixante ans plus tard, cette conjecture fut attribuée a A. Ogg jusqu'à la redécouverte récente par N. Schappacher et R. Schoof des travaux de Levi. Bien entendu l'énoncé de Levi est depuis 1977 un théorème de B. Mazur [6].

Comme c'est souvent le cas en arithmétique, un théorème démontré pour les nombres rationnels est un théorème démontré pour seulement un corps de nombres. Mazur a produit en 1978 une seconde preuve de son théorème [7]. Cette dernière preuve peut-être considérée comme le point de départ des travaux ultérieurs sur les points de torsion des courbes elliptiques sur les corps de nombres.

On dispose depuis 1994 du théorème suivant [8] :

Théorème 1 *Soit d un entier > 0. Il existe un nombre fini de groupes, à isomorphisme près, qui sont constitués par la partie de torsion du groupe des points K-rationnels d'une courbe elliptique sur K, où K parcourt les corps de nombres de degré d sur $\mathbf{Q}$.*

Indiquons brièvement comment la recherche s'est developpée à partir de 1978. Les méthodes de Mazur ont été généralisées par S. Kamienny, ce qui a permis de démontrer le théorème ci-dessus pour $d = 2$ (Kamienny [2]), puis pour $d \leq 8$ (Kamienny et Mazur [4]), puis pour $d \leq 14$ (D. Abramovich [1]). La démonstration du cas général repose sur l'approche de Kamienny et Mazur (mais est indépendante de la première preuve de Mazur). Soulignons le rôle central joué (pour $d > 14$) par les travaux de Kolyvagin, Logachev, Gross, Zagier ... en direction de la conjecture de Birch et Swinnerton-Dyer.

Un énoncé plus faible que le théorème (finitude dépendant de K et non seulement du degré d), constituait un problème ouvert au moins depuis les années 60. La dépendance en le degré a été mise en évidence par Kamienny [3]. Comme l'a démontré Abramovich, le théorème est une des multiples conséquence étonnantes de conjectures très générales de S. Lang.

On dispose maintenant de versions précises du théorème 1. Indiquons-en deux dues à J. Oesterlé et P. Parent.

THÉORÈME 2 (OESTERLÉ) *Soit d un entier > 0. Soit K un corps de nombres de degré d sur $\mathbf{Q}$. Soit E une courbe elliptique sur K munie d'un point K-rationnel d'ordre premier p. Alors on a*

$$p \leq (1 + 3^{\frac{d}{2}})^2.$$

C'est un cas particulier des résultats ultérieurement obtenus par Parent. Il implique, à la suite de remarques d'Abramovich, Frey, Kamienny et Mazur et à l'aide d'un théorème non effectif de G. Faltings le théorème 1. Le théorème suivant se démontre de façon analogue au théorème 2 apour conséquence facile le théorème 1.

THÉORÈME 3 (PARENT) *Soit d un entier > 0. Soit K un corps de nombres de degré d sur $\mathbf{Q}$. Soit E une courbe elliptique sur K munie d'un point K-rationnel d'ordre une puissance n d'un nombre premier p. Alors on a*

$$n \leq 129(5^d - 1)(3d)^6.$$

Le théorème d'Oesterlé (malheureusement non publié) donne une idée fidèle des limites des méthodes actuelles. Un examen du principe de la démonstration convainc rapidement que les inégalités numériques obtenues par Oesterlé et Parent n'ont guère de raisons d'être satisfaisantes. Une avancée importante consisterait désormais à établir des des inégalités analogues dépendant polynomialement du degré d. Pour cela on aimerait combiner les méthodes modulaires avec les méthodes issues de la théories des nombres transcendants (voir les travaux de D. Masser, G. Wüstholz, S. David, M. Hindry, F. Pellarin).

Le seul degré d où on dispose d'un analogue satisfaisant du théorème de Mazur est le degré $d = 2$, pour lequel la liste complète des sous-groupes possibles a été établie par Kamienny [2], à l'aide de travaux antérieurs de M. A. Kenku et F. Momose [5].

Nous nous proposons d'évoquer les grandes lignes de la démonstration du théorème 2.

2 ANALYSE ÉLÉMENTAIRE

Soit d un entier > 0. Soit K un corps de nombres de degré d sur $\mathbf{Q}$. Soit E une courbe elliptique sur K munie d'un point P qui est K-rationnel et d'ordre premier p.

Considérons un nombre premier auxiliaire l. Soit λ un idéal premier de l'anneau des entiers de K au dessus de l. Des arguments élémentaires (essentiellement le théorème de Hasse-Weil) montrent qu'on a $p \leq (1 + l^{d/2})^2$ ou que E a réduction multiplicative déployée en λ et que P est d'ordre p dans le groupe des composantes de la fibre en p du modèle de Néron de E (Nous dirons que cette dernière situation constitue le cas critique en λ).

La difficulté de la démonstration du théorème est donc concentrée dans le cas totalement critique en l (c'est-à-dire le cas critique en λ pour tout idéal λ au dessus de l). Tout cela a été constaté, notamment par J. Tate, il y a plus de 40 ans.

Oesterlé choisit le nombre premier $l = 3$ pour sa démonstration. La démonstration originale du théorème 1 utilisait une idée de Kamienny, et montrait l'existence d'un nombre premier l (sans donner de valeur précise pour l) borné en fonction de d seulement pour lequel le cas totalement critique est exclu.

3 Courbes modulaires

Considérons la courbe modulaire $X_0(p)$ qui classifie grossièrement les courbes elliptiques généralisées munies d'un sous-groupe cyclique d'ordre p. Le couple $(E/ < P >, E[p]/ < P >)$ définit un point K rationnel de $X_0(p)$ et donc un point $\mathbf{Q}$-rationnel Q de la puissance symétrique d-ième $X_0(p)^{(d)}$ de $X_0(p)$.

On est donc ramené à étudier les points rationnels de $X_0(p)^{(d)}$.

4 De la courbe à sa jacobienne

Après, entre autres, A. Weil, C. Chabauty, A. Parshin, Faltings, R. Coleman, il est classique d'étudier les points rationnels d'une courbe X (resp. de la puissance symétrique d-ième de X) en combinant l'étude de la géomeétrie $\phi : X \longrightarrow A$ (resp. $X^{(d)} \longrightarrow A$), où A est une variété abélienne, avec l'étude du groupe des points rationnels de A (ce qui typiquement consiste à établir la finitude du groupe des points rationnels de A).

Considérons la plus grande variété abélienne quotient J_e de $J_0(p)$ dont la fonction L ne s'annule pas en 1. On considère le morphisme (convenablement normalisé) $X_0(p)^{(d)} \longrightarrow J_e$. Par un théorème de Kolyvagin et Logachev (dont des démonstrations alternatives ont été proposées par K. Kato d'une part et par M. Bertolini et H. Darmon d'autre part) la variété abélienne J_e n'a qu'un nombre fini de points $\mathbf{Q}$-rationnels. C'est un argument de nature profondément arithmétique qui est au centre de la démonstration du théorème 1.

On peut résumer la fin de la démonstration du théorème 2 de la façon suivante. Par des arguments dûs à Kamienny il suffit pour conclure de démontrer que les d premiers opérateurs de Hecke sont linéairement indépendants en caractéristique 3 dans J_e (Le même résultat a été ultérieurement obtenu par M. Baker grâce à des méthodes reposant sur l'intégration p-adique de Coleman).

La démonstration de l'indépendance linéaire cherchée se démontre par la théorie des symboles modulaires. (Un énoncé analogue d'indépendance linéaire en caractéristique 0 a été établi par J. Van der Kam par des méthodes fondées sur la théorie analytique des fonctions L ; Cela suffit pour démontrer le théorème 1 mais pas le théorème 2.)

Références

[1] Abramovich, D. *Formal finiteness and the torsion conjecture on elliptic curves. A footnote to a paper: "Rational torsion of prime order in elliptic curves over number fields"* Columbia University Number Theory Seminar (New York, 1992). Astrisque No. 228 (1995), 3, 5–17.

[2] Kamienny, S. *Torsion points on elliptic curves and q-coefficients of modular forms.* Invent. Math. 109 (1992), no. 2, 221–229.

[3] Kamienny, S. *Torsion points on elliptic curves over fields of higher degree.* Internat. Math. Res. Notices (1992), no. 6, 129–133.

[4] Kamienny, S.; Mazur, B. *Rational torsion of prime order in elliptic curves over number fields.* Columbia University Number Theory Seminar (New York, 1992). Astrisque No. 228 (1995), 3, 81–100.

[5] Kenku, M. A.; Momose, F. *Torsion points on elliptic curves defined over quadratic fields.* Nagoya Math. J. 109 (1988), 125–149.

[6] Mazur, B. *Modular curves and the Eisenstein ideal.* Inst. Hautes Études Sci. Publ. Math. No. 47 (1977), 33–186.

[7] Mazur, B. *Rational isogenies of prime degree (with an appendix by D. Goldfeld).* Invent. Math. 44 (1978), no. 2, 129–162.

[8] Merel, L. *Bornes pour la torsion des courbes elliptiques sur les corps de nombres.* Invent. Math. 124 (1996), no. 1-3, 437–449.

Loïc Merel
UFR de Mathématiques
case 7012
Université Denis Diderot
2, place Jussieu
75251 Paris cedex 05
France
merel@math.jussieu.fr

The Intrinsic Hodge Theory of p-Adic Hyperbolic Curves

Shinichi Mochizuki

1991 Mathematics Subject Classification: 14F30
Keywords and Phrases: hyperbolic curves, p-adic, fundamental group, Hodge theory

§1. Introduction

(A.) The Fuchsian Uniformization

A *hyperbolic curve* is an algebraic curve obtained by removing r points from a smooth, proper curve of genus g, where g and r are nonnegative integers such that $2g - 2 + r > 0$. If X is a hyperbolic curve over the field of complex numbers C, then X gives rise in a natural way to a Riemann surface $\mathcal{X}$. As one knows from complex analysis, the most fundamental fact concerning such a Riemann surface (due to Köbe) is that it may be *uniformized by the upper half-plane*, i.e.,

$$\mathcal{X} \cong \mathfrak{H}/\Gamma$$

where $\mathfrak{H} \stackrel{\text{def}}{=} \{z \in C \mid \text{Im}(z) > 0\}$, and $\Gamma \cong \pi_1(\mathcal{X})$ (the topological fundamental group of $\mathcal{X}$) is a discontinuous group acting on $\mathfrak{H}$. Note that the action of Γ on $\mathfrak{H}$ defines a canonical representation

$$\rho_{\mathcal{X}} : \pi_1(\mathcal{X}) \to PSL_2(R) \stackrel{\text{def}}{=} SL_2(R)/\{\pm 1\} = \text{Aut}_{\text{Holomorphic}}(\mathfrak{H})$$

The goal of the present manuscript is to survey various work ([Mzk1-5]) devoted to generalizing Köbe's uniformization to the p-adic case.

First, we observe that it is not realistic to expect that hyperbolic curves over p-adic fields may be literally uniformized by some sort of p-adic upper half-plane in the fashion of the Köbe uniformization. Of course, one has the theory of Mumford ([Mumf]), but this theory furnishes a p-adic analogue not of Köbe's *Fuchsian uniformization* (i.e., uniformization by a Fuchsian group), but rather of what in the complex case is known as the *Schottky uniformization*. Even in the complex case, the Fuchsian and Schottky uniformizations are fundamentally different: For instance, as the moduli of the curve vary, its Schottky periods vary holomorphically, whereas its Fuchsian periods vary only *real analytically*. This fact already suggests that the Fuchsian uniformization is of a more *arithmetic nature* than the Schottky uniformization, i.e., it involves

real analytic structures $\Longleftrightarrow$ complex conjugation $\Longleftrightarrow$ Frobenius at the infinite prime

Thus, since one cannot expect a p-adic analogue in the form of a literal global uniformization of the curve, the first order of business is to reinterpret the Fuchsian uniformization in more abstract terms that generalize naturally to the p-adic setting.

(B.) THE PHYSICAL INTERPRETATION

The first and most obvious approach is to observe that the Fuchsian uniformization gives a new physical, geometric way to reconstruct the original *algebraic curve* X. Namely, one may think of the Fuchsian uniformization as defining a *canonical arithmetic structure* $\rho_{\mathcal{X}} : \pi_1(\mathcal{X}) \to PSL_2(R)$ on the purely *topological* invariant $\pi_1(\mathcal{X})$. Alternatively (and essentially equivalently), one may think of the Fuchsian uniformization as the datum of a *metric* (given by descending to $\mathcal{X} \cong \mathfrak{H}/\Gamma$ the Poincaré metric on $\mathfrak{H}$) – i.e., an *arithmetic* (in the sense of arithmetic at the infinite prime) structure – on the differential manifold underlying $\mathcal{X}$ (which is a purely *topological* invariant). Then the equivalence

$$X \Longleftrightarrow SO(2)\backslash PSL_2(R)/\Gamma$$

between the *algebraic* curve X and the physical/analytic object $SO(2)\backslash PSL_2(R)/\Gamma$ obtained from $\rho_{\mathcal{X}}$ is given by considering *modular forms* on $\mathfrak{H} = SO(2)\backslash PSL_2(R)$, which define a projective (hence, *algebraizing*) embedding of $\mathcal{X}$.

(C.) THE MODULAR INTERPRETATION

Note that $\rho_{\mathcal{X}}$ may also be regarded as a representation into $PGL_2(C) = GL_2(C)/C^{\times}$, hence as defining an action of $\pi_1(\mathcal{X})$ on P^1_C. Taking the quotient of $\mathfrak{H} \times P^1_C$ by the action of $\pi_1(\mathcal{X})$ on both factors then gives rise to a projective bundle with connection on $\mathcal{X}$. It is immediate that this projective bundle and connection may be algebraized, so we thus obtain a projective bundle and connection $(P \to X, \nabla_P)$ on X. This pair (P, ∇_P) has certain properties which make it an *indigenous bundle* (terminology due to Gunning). More generally, an indigenous bundle on $\mathcal{X}$ may be thought of as a *projective structure* on $\mathcal{X}$, i.e., a subsheaf of the sheaf of holomorphic functions on $\mathcal{X}$ such that locally any two sections of this subsheaf are related by a linear fractional transformation. Thus, the Fuchsian uniformization defines a special *canonical indigenous bundle* on X.

In fact, the notion of an indigenous bundle is entirely *algebraic*. Thus, one has a natural moduli stack $\mathcal{S}_{g,r} \to \mathcal{M}_{g,r}$ of indigenous bundles, which forms a torsor (under the affine group given by the sheaf of differentials on $\mathcal{M}_{g,r}$) – called the *Schwarz torsor* – over the moduli stack $\mathcal{M}_{g,r}$ of hyperbolic curves of type (g, r). Moreover, $\mathcal{S}_{g,r}$ is not only algebraic, it is defined over $Z[\frac{1}{2}]$. Thus, the canonical indigenous bundle defines a *canonical real analytic section*

$$s : \mathcal{M}_{g,r}(C) \to \mathcal{S}_{g,r}(C)$$

of the Schwarz torsor at the infinite prime. Moreover, not only does s "contain" all the information that one needs to define the Fuchsian uniformization of an individual hyperbolic curve (indeed, this much is obvious from the definition of s!), it also essentially "is" (interpreted properly) the Bers uniformization of the universal covering space (i.e., "Teichmüller space") of $\mathcal{M}_{g,r}(C)$ (cf. the discussions in the Introductions of [Mzk1,4]). That is to say, from this point of view, one may regard the uniformization theory of hyperbolic curves and their moduli as the study of the canonical section s. Alternatively, from the point of view of Teichmüller theory, one may regard the uniformization theory of hyperbolic curves and their moduli as the theory of (so-called) *quasi-fuchsian deformations of the representation* $\rho_{\mathcal{X}}$.

(D.) The Notion of "Intrinsic Hodge Theory"

Note that both the physical and modular approaches to the Fuchsian uniformization assert that there is a certain equivalence

$$\text{algebraic geometry} \Longleftrightarrow \text{topology endowed with an arithmetic structure}$$

That is, on the algebraic geometry side, we have the scheme (respectively, stack) given by the curve X itself in the physical approach (respectively, its moduli $\mathcal{M}_{g,r}$ in the modular approach), whereas on the "topology plus arithmetic structure" side, we have the theory of the canonical representation $\rho_{\mathcal{X}}$ of $\pi_1(\mathcal{X})$ (i.e., $SO(2)\backslash PSL_2(R)/\Gamma$ in the physical approach; quasi-fuchsian deformations of $\rho_{\mathcal{X}}$ in the modular approach). This sort of equivalence is reminiscent of that given by classical or p-adic *Hodge theory* between the de Rham or Hodge cohomology of an algebraic variety (on the algebraic geometry side), and the singular or étale cohomology (equipped with Galois action) on the topology plus arithmetic side. In our case, however, instead of dealing with the cohomology of the curve, we are dealing with "the curve itself" and its moduli. It is for this reason that we refer to this sort of theory as the *intrinsic Hodge theory* of the curve X.

Finally, we note that this formal analogy with classical/p-adic Hodge theory is by no means merely philosophical. Indeed, even in the classical theory reviewed in (B.) and (C.) above, the methods of classical Hodge theory play an important technical role in the proofs of the main theorems. Similarly, in the theory of [Mzk1-5] – which constitute our main examples of *intrinsic Hodge theory for hyperbolic curves* – the more recently developed techniques of p-adic Hodge theory play a crucial technical role in the proofs of the main results.

§2. The Physical Approach in the p-adic Case

(A.) The Arithmetic Fundamental Group

Let K be a field of characteristic zero. Let us denote by $\overline{K}$ an algebraic closure of K. Let $\Gamma_K \stackrel{\text{def}}{=} \text{Gal}(\overline{K}/K)$. Let X_K be a *hyperbolic curve over* K; write $X_{\overline{K}} \stackrel{\text{def}}{=} X \times_K \overline{K}$. Then one has an exact sequence

$$1 \to \pi_1(X_{\overline{K}}) \to \pi_1(X_K) \to \Gamma_K \to 1$$

of *algebraic fundamental groups*. (Here, we omit the base-points from the notation for the various fundamental groups.)

We shall refer to $\pi_1(X_{\overline{K}})$ as the *geometric fundamental group* of X_K. Note that the structure of $\pi_1(X_{\overline{K}})$ is determined entirely by (g, r) (i.e., the "type" of the hyperbolic curve X_K). In particular, $\pi_1(X_{\overline{K}})$ *does not depend on the moduli of* X_K. Of course, this results from the fact that K is of *characteristic zero*; in positive characteristic, on the other hand, preliminary evidence ([Tama2]) suggests that the fundamental group of a hyperbolic curve over an algebraically closed field (far from being independent of the moduli of the curve!) may in fact *completely determine* the moduli of the curve.

On the other hand, we shall refer to $\pi_1(X_K)$ (equipped with its augmentation to Γ_K) as the *arithmetic fundamental group* of X_K. Although it is made up of two "parts" – i.e., $\pi_1(X_{\overline{K}})$ and Γ_K – which do not depend on the moduli of X_K, it is not unreasonable to expect that the extension class defined by the above exact sequence, i.e., the structure of $\pi_1(X_K)$ as a group equipped with augmentation to Γ_K, may in fact depend quite strongly on the moduli of X_K. Indeed, according to the *anabelian philosophy* of Grothendieck (cf. [LS]), for "sufficiently arithmetic" K, one expects that *the structure of the arithmetic fundamental group* $\pi_1(X_K)$ *should be enough to determine the moduli of* X_K. Although many important versions of Grothendieck's anabelian conjectures remain unsolved (most notably the so-called *Section Conjecture* (cf., e.g., [LS], p. 289, 2)), in the remainder of this §, we shall discuss various versions that have been resolved in the affirmative. Finally, we note that this anabelian philosophy is a special case of the notion of "intrinsic Hodge theory" discussed above: indeed, on the algebraic geometry side, one has "the curve itself," whereas on the topology plus arithmetic side, one has the arithmetic fundamental group, i.e., the purely (étale) topological $\pi_1(X_{\overline{K}})$, equipped with the structure of extension given by the above exact sequence.

(B.) The Main Theorem

Building on earlier work of H. Nakamura and A. Tamagawa (see, especially, [Tama1]), the author applied the p-adic Hodge theory of [Falt2] and [BK] to prove the following result (cf. Theorem A of [Mzk5]):

THEOREM 1. *Let p be a prime number. Let K be a subfield of a finitely generated field extension of Q_p. Let X_K be a hyperbolic curve over K. Then for any smooth variety S_K over K, the natural map*

$$X_K(S_K)^{\mathrm{dom}} \to \mathrm{Hom}^{\mathrm{open}}_{\Gamma_K}(\pi_1(S_K), \pi_1(X_K))$$

*is bijective. Here, the superscripted "*dom*" denotes dominant ($\Longleftrightarrow$ nonconstant) K-morphisms, while* $\mathrm{Hom}^{\mathrm{open}}_{\Gamma_K}$ *denotes open, continuous homomorphisms compatible with the augmentations to Γ_K, and considered up to composition with an inner automorphism arising from $\pi_1(X_{\overline{K}})$.*

Note that this result constitutes an analogue of the "physical aspect" of the Fuchsian uniformization, i.e., it exhibits the *scheme* X_K (in the sense of the functor

defined by considering (nonconstant) K-morphisms from arbitrary smooth S_K to X_K) as equivalent to the "physical/analytic object"

$$\mathrm{Hom}^{\mathrm{open}}_{\Gamma_K}(-, \pi_1(X_K))$$

defined by the topological $\pi_1(X_{\overline{K}})$ together with some additional canonical arithmetic structure (i.e., $\pi_1(X_K)$).

In fact, the proof of Theorem 1 was also motivated by this point of view: That is to say, just as one may regard the algebraic structure of a hyperbolic curve over C as being defined by certain (a priori) *analytic* modular forms on $\mathfrak{H}$, the proof of Theorem 1 proceeds by considering certain p-adic analytic representations of differential forms on X_K. In the p-adic case, however, the domain of definition of these analytic forms (i.e., the analogue to the upper half-plane) is the spectrum of the p-adic completion of the maximal tame extension of the function field of X_K along various irreducible components of the special fiber of a stable model $\mathcal{X} \to \mathrm{Spec}(\mathcal{O}_K)$ of X_K (where $\mathcal{O}_K$ is the ring of integers of a finite extension K of Q_p). It turns out that this object is, just like the upper half-plane, independent of the moduli of X_K.

In fact, various slightly stronger versions of Theorem 1 hold. For instance, instead of the whole geometric fundamental group $\pi_1(X_{\overline{K}})$, it suffices to consider its maximal pro-p quotient $\pi_1(X_{\overline{K}})^{(p)}$. Another strengthening allows one to prove the following result (cf. Theorem B of [Mzk5]), which generalizes a result of Pop ([Pop]):

Corollary 2. *Let p be a prime number. Let K be a subfield of a finitely generated field extension of Q_p. Let L and M be function fields of arbitrary dimension over K. Then the natural map*

$$\mathrm{Hom}_K(\mathrm{Spec}(L), \mathrm{Spec}(M)) \to \mathrm{Hom}^{\mathrm{open}}_{\Gamma_K}(\Gamma_L, \Gamma_M)$$

is bijective. Here, $\mathrm{Hom}^{\mathrm{open}}_{\Gamma_K}(\Gamma_L, \Gamma_M)$ is the set of open, continuous group homomorphisms $\Gamma_L \to \Gamma_M$ over Γ_K, considered up to composition with an inner homomorphism arising from $\mathrm{Ker}(\Gamma_M \to \Gamma_K)$.

(C.) Comparison with the Case of Abelian Varieties

Note that there is an obvious formal analogy between Theorem 1 above and Tate's conjecture on homomorphisms between abelian varieties (cf., e.g., [Falt1]). Indeed, in discussions of Grothendieck's anabelian philosophy, it was common to refer to statements such as that of Theorem 1 as the "anabelian Tate conjecture," or the "Tate conjecture for hyperbolic curves." In fact, however, there is an important difference between Theorem 1 and the "Tate conjecture" of, say, [Falt1]: Namely, *the Tate conjecture for abelian varieties is false over local fields (i.e., finite extensions of Q_p).* Moreover, until the proof of Theorem 1, it was generally thought that, just like its abelian cousin, the "anabelian Tate conjecture" was essentially global in nature. That is to say, it appears that the point of view of the author, i.e., that Theorem 1 should be regarded as a p-adic version of the "physical aspect" of

the Fuchsian uniformization of a hyperbolic curve, does not exist in the literature (prior to the work of the author).

§3. The Modular Approach in the p-adic Case

(A.) The Example of Shimura Curves

As discussed in §1, (C.), classical complex Teichmüller theory may be formulated as the study of the canonical real analytic section s of the Schwarz torsor $\mathcal{S}_{g,r} \to \mathcal{M}_{g,r}$. Thus, it is natural suppose that the p-adic analogue of classical Teichmüller theory should revolve around some sort of *canonical p-adic section* of the Schwarz torsor. Then the question arises:

How does one define a canonical p-adic section of the Schwarz torsor?

Put another way, for each (or at least most) p-adic hyperbolic curves, we would like to associate a (or at least a finite, bounded number of) canonical indigenous bundles. Thus, we would like to know what sort of properties such a "canonical indigenous bundle" should have.

The model that provides the answer to this question is the theory of *Shimura curves*. In fact, the theory of canonical Schwarz structures, canonical differentials, and canonical coordinates on Shimura curves localized at finite primes has been extensively studied by Y. Ihara (see, e.g., [Ihara]). In some sense, *Ihara's theory provides the prototype for the "p-adic Teichmüller theory" of arbitrary hyperbolic curves* ([Mzk1-4]) to be discussed in (B.) and (C.) below. The easiest example of a Shimura curve is $\mathcal{M}_{1,0}$, the moduli stack of elliptic curves. In this case, the projectivization of the rank two bundle on $\mathcal{M}_{1,0}$ defined by the first de Rham cohomology module of the universal elliptic curve on $\mathcal{M}_{1,0}$ gives rise (when equipped with the Gauss-Manin connection) to the canonical indigenous bundle on $\mathcal{M}_{1,0}$. Moreover, it is well-known that the *p-curvature* (a canonical invariant of bundles with connection in positive characteristic which measures the extent to which the connection is compatible with Frobenius) of this bundle has the following property:

The p-curvature of the canonical indigenous bundle on $\mathcal{M}_{1,0}$ (reduced mod p) is square nilpotent.

It was this observation that was the key to the development of the theory of [Mzk1-4].

(B.) The Stack of Nilcurves

Let p be an *odd* prime. Let $\mathcal{N}_{g,r} \subseteq (\mathcal{S}_{g,r})_{F_p}$ denote the closed algebraic substack of indigenous bundles with square nilpotent p-curvature. Then one has the following key result ([Mzk1], Chapter II, Theorem 2.3):

THEOREM 3. *The natural map $\mathcal{N}_{g,r} \to (\mathcal{M}_{g,r})_{F_p}$ is a finite, flat, local complete intersection morphism of degree p^{3g-3+r}. Thus, up to "isogeny" (i.e., up to the fact that this degree is not equal to one), $\mathcal{N}_{g,r}$ defines a canonical section of the Schwarz torsor $(\mathcal{S}_{g,r})_{F_p} \to (\mathcal{M}_{g,r})_{F_p}$ in characteristic p.*

It is this stack $\mathcal{N}_{g,r}$ of *nilcurves* – i.e., hyperbolic curves in characteristic p equipped with an indigenous bundle with square nilpotent p-curvature – which is the central object of study in the theory of [Mzk1-4].

Once one has the above Theorem, next it is natural to ask if one can say more about the fine structure of $\mathcal{N}_{g,r}$. Although many interesting and natural questions concerning the structure of $\mathcal{N}_{g,r}$ remain unsolved at the time of writing, a certain amount can be understood by analyzing certain *substacks, or strata,* of $\mathcal{N}_{g,r}$ defined by considering the loci of nilcurves whose p-curvature vanishes to a certain degree. For instance, nilcurves whose p-curvature vanishes identically are called *dormant*. The locus of dormant nilcurves is denoted $\mathcal{N}_{g,r}[\infty] \subseteq \mathcal{N}_{g,r}$. If a nilcurve is not dormant, then its p-curvature vanishes on some divisor in the curve. We denote by $\mathcal{N}_{g,r}[d] \subseteq \mathcal{N}_{g,r}$ the locus of nilcurves for which this divisor is of degree d. The zeroes of the p-curvature are referred to as *spikes*. Now we have the following result (cf. Theorems 1.2, 1.6 of the Introduction of [Mzk4]):

Theorem 4. *The $\mathcal{N}_{g,r}[d]$ are all smooth over F_p and either empty or of dimension $3g-3+r$. Moreover, $\mathcal{N}_{g,r}[0]$ is* affine.

It turns out that *this affineness of $\mathcal{N}_{g,r}[0]$, interpreted properly, gives a new proof of the connectedness of $(\mathcal{M}_{g,r})_{F_p}$* (for p large relative to g). This fact is interesting (relative to the claim that this theory is a p-adic version of Teichmüller theory) in that one of the first applications of classical complex Teichmüller theory is to prove the connectedness of $\mathcal{M}_{g,r}$. Also, it is interesting to note that F. Oort has succeeded in giving a proof of the connnectedness of the moduli stack of principally polarized abelian varieties by using affineness properties of certain natural substacks of this moduli stack in characteristic p.

Despite the fact that the $\mathcal{N}_{g,r}[d]$ are smooth and of the same dimension as $\mathcal{N}_{g,r}$, we remark that in most cases *$\mathcal{N}_{g,r}$ is not reduced at $\mathcal{N}_{g,r}[d]$*. In fact, roughly speaking, the larger d is, the less reduced $\mathcal{N}_{g,r}$ is at $\mathcal{N}_{g,r}[d]$. In order to give sharp quantitative answers to such questions as:

> *How reduced is $\mathcal{N}_{g,r}$ at the generic point of $\mathcal{N}_{g,r}[d]$? Or, what is the generic degree of $\mathcal{N}_{g,r}[d]$ over $(\mathcal{M}_{g,r})_{F_p}$?*

it is necessary to study what happens to a nilcurve as the underlying curve degenerates to a *totally degenerate stable curve* (i.e., a stable curve each of whose irreducible components is P^1, with a total of precisely three marked points/nodes). To do this, one must formulate the theory (using "log structures") in such a way that it applies to stable curves, as well.

Once one formulates the theory for stable curves, one sees that the answers to the questions just posed will follow as soon as one:

(i.) Classifies all *molecules* – i.e., nilcurves whose underlying curve is a totally degenerate stable curve.

(ii.) Understands how molecules *deform*.

The answer to (i.) and (ii.) depends on an extensive analysis of molecules (cf. [Mzk2-4]), and, although combinatorially quite complicated, is, in some sense, complete. Although we do not have enough space here to discuss this answer

in detail, we pause to remark the following: It turns out that the answer to (i.) consists of regarding molecules as concatenations of *atoms* – i.e., *toral nilcurves* (a slight generalization of nilcurves) whose underlying curve is P^1 with three marked points – and then classifying atoms. The difference between a toral nilcurve and a (nontoral) nilcurve is that unlike the nontoral case, where the "radii" at the three marked points are assumed to be zero, in the toral case, one allows these radii to be arbitrary elements of $F_p/\{\pm 1\}$ (i.e., the quotient of the set F_p obtained by identifying λ and $-\lambda$ for all $\lambda \in F_p$). Then it turns out that considering the three radii of an atom defines a natural bijection between the isomorphism classes of atoms and the set of (ordered) triples of elements of $F_p/\{\pm 1\}$.

The reason that we digressed to discuss the theory of atoms is that it is interesting (relative to the analogy with classical complex Teichmüller theory) in that it is reminiscent of the fact that a Riemann surface may be analyzed by decomposing it into *pants* (i.e., Riemann surfaces which are topologically isomorphic to $P^1 - \{0, 1, \infty\}$). Moreover, the isomorphism class of a "pants" is completely determined by the radii of its three holes.

(C.) Canonical Liftings

So far, we have been discussing the characteristic p theory. Ultimately, however, we would like to know if the various characteristic p objects discussed in (B.) lift canonically to objects which are flat over Z_p. Unfortunately, it seems that it is unlikely that $\mathcal{N}_{g,r}$ itself lifts canonically to some sort of natural Z_p-flat object. If, however, we consider the open substack – called the *ordinary locus* – $(\mathcal{N}^{\mathrm{ord}}_{g,r})_{F_p} \subseteq \mathcal{N}_{g,r}$ which is the étale locus of the morphism $\mathcal{N}_{g,r} \to (\mathcal{M}_{g,r})_{F_p}$, then (since the étale site is invariant under nilpotent thickenings) we get a canonical lifting, i.e., an étale morphism

$$\mathcal{N}^{\mathrm{ord}}_{g,r} \to (\mathcal{M}_{g,r})_{Z_p}$$

of p-adic formal stacks. Over $\mathcal{N}^{\mathrm{ord}}_{g,r}$, one has the sought-after canonical p-adic splitting of the Schwarz torsor (cf. Theorem 0.1 of the Introduction of [Mzk1]):

THEOREM 5. *There is a canonical section $\mathcal{N}^{\mathrm{ord}}_{g,r} \to \mathcal{S}_{g,r}$ of the Schwarz torsor over $\mathcal{N}^{\mathrm{ord}}_{g,r}$ which is the unique section having the following property: There exists a lifting of Frobenius $\Phi_{\mathcal{N}} : \mathcal{N}^{\mathrm{ord}}_{g,r} \to \mathcal{N}^{\mathrm{ord}}_{g,r}$ such that the indigenous bundle on the tautological hyperbolic curve over $\mathcal{N}^{\mathrm{ord}}_{g,r}$ defined by the section $\mathcal{N}^{\mathrm{ord}}_{g,r} \to \mathcal{S}_{g,r}$ is* invariant *with respect to the Frobenius action defined by $\Phi_{\mathcal{N}}$.*

Moreover, it turns out that the *Frobenius lifting* $\Phi_{\mathcal{N}} : \mathcal{N}^{\mathrm{ord}}_{g,r} \to \mathcal{N}^{\mathrm{ord}}_{g,r}$ (i.e., morphism whose reduction modulo p is the Frobenius morphism) has the special property that $\frac{1}{p} \cdot \mathrm{d}\Phi_{\mathcal{N}}$ induces an isomorphism $\Phi^*_{\mathcal{N}} \Omega_{\mathcal{N}^{\mathrm{ord}}_{g,r}} \cong \Omega_{\mathcal{N}^{\mathrm{ord}}_{g,r}}$. Such a Frobenius lifting is called *ordinary*. It turns out that any ordinary Frobenius lifting (i.e., not just $\Phi_{\mathcal{N}}$) defines a set of *canonical multiplicative coordinates* in a formal neighborhood of any point α valued in an algebraically closed field k of characteristic p, as well as a *canonical lifting* of α to a point valued in $W(k)$ (Witt vectors with coefficients in k). Moreover, there is a certain analogy between this general theory

of ordinary Frobenius liftings and the theory of *real analytic Kähler metrics* (which also define canonical coordinates). Relative to this analogy, the canonical Frobenius lifting $\Phi_{\mathcal{N}}$ on $\mathcal{N}_{g,r}^{\mathrm{ord}}$ may be regarded as corresponding to the *Weil-Petersson metric* on complex Teichmüller space (a metric whose canonical coordinates are the coordinates arising from the Bers uniformization of Teichmüller space). Thus, $\Phi_{\mathcal{N}}$ is, in a very real sense, a p-adic analogue of the Bers uniformization in the complex case. Moreover, there is, in fact, a canonical ordinary Frobenius lifting on the "ordinary locus" of the tautological curve over $\mathcal{N}_{g,r}^{\mathrm{ord}}$ whose relative canonical coordinate is analogous to the canonical coordinate arising from the Köbe uniformization of a hyperbolic curve.

Next, we observe that Serre-Tate theory for ordinary (principally polarized) abelian varieties may also be formulated as arising from a certain canonical ordinary Frobenius lifting. Thus, the Serre-Tate parameters (respectively, Serre-Tate canonical lifting) may be identified with the canonical multiplicative parameters (respectively, canonical lifting to the Witt vectors) of this Frobenius lifting. That is to say, in a very concrete and rigorous sense, Theorem 5 may be regarded as the analogue of Serre-Tate theory for hyperbolic curves. Nevertheless, we remark that it is *not* the case that the condition that a nilcurve be ordinary (i.e., defines a point of $(\mathcal{N}_{g,r}^{\mathrm{ord}})_{F_p} \subseteq \mathcal{N}_{g,r}$) either implies or is implied by the condition that its Jacobian be ordinary. Although this fact may disappoint some readers, it is in fact very natural when viewed relative to the general analogy between ordinary Frobenius liftings and real analytic Kähler metrics discussed above. Indeed, relative to this analogy, we see that it corresponds to the fact that, when one equips $\mathcal{M}_g$ with the Weil-Petersson metric and $\mathcal{A}_g$ (the moduli stack of principally polarized abelian varieties) with its natural metric arising from the Siegel upper half-plane uniformization, *the Torelli map $\mathcal{M}_g \to \mathcal{A}_g$ is not isometric.*

Next, we remark that $(\mathcal{N}_{g,r}^{\mathrm{ord}})_{F_p} \subseteq \mathcal{N}_{g,r}[0]$. Thus, the other $\mathcal{N}_{g,r}[d]$'s are left out of the theory of canonical liftings arising from Theorem 5. Nevertheless, in [Mzk2,4], a more general theory of canonical liftings is developed that includes arbitrary $\mathcal{N}_{g,r}[d]$. In this more general theory, instead of getting local uniformizations by multiplicative canonical parameters, i.e., uniformizations by $\widehat{G}_{\mathrm{m}}$, we get uniformizations by more general types of *Lubin-Tate groups*, or twisted products of such groups. Roughly speaking, the more "spikes" in the nilcurves involved – i.e., the larger the d of $\mathcal{N}_{g,r}[d]$ – the more Lubin-Tate the uniformization becomes.

Finally, we remark that once one develops these theories of canonical liftings, one also gets accompanying canonical (crystalline) Galois representations of the arithmetic fundamental group of the tautological curve over $\mathcal{N}_{g,r}^{\mathrm{ord}}$ (and its Lubin-Tate generalizations) into PGL_2 of various complicated rings with Galois action. It turns out that *these Galois representations are the analogues of the canonical representation $\rho_{\mathcal{X}}$* (of §1, (A.)) – which was the starting point of our entire discussion.

BIBLIOGRAPHY

[BK] Bloch, S. and Kato, K., L-Functions and Tamagawa Numbers of Motives in *The Grothendieck Festschrift*, Volume I, Birkhäuser (1990), pp. 333-400.

[Falt1] Faltings, G., Endlichkeitssätze für Abelschen Varietäten über Zahlkörpern, *Inv. Math.* 73 (1983), pp. 349-366.

[Falt2] Faltings, G., p-adic Hodge Theory, *Journal of the Amer. Math. Soc.* 1, No. 1 (1988), pp. 255-299.

[Ihara] Ihara, Y., On the Differentials Associated to Congruence Relations and the Schwarzian Equations Defining Uniformizations, *Jour. Fac. Sci. Univ. Tokyo*, Sect. IA Math. 21 (1974), pp. 309-332.

[LS] Lochak, P. and Schneps, L., Geometric Galois Actions: 1. Around Grothendieck's Esquisse d'un Programme, London Math. Soc. Lect. Note Ser. 242, Cambridge Univ. Press, 1997.

[Mumf] Mumford, D., An Analytic Construction of Degenerating Curves over Complete Local Rings, *Comp. Math.* 24 (1972), pp. 129-174.

[Mzk1] Mochizuki, S., A Theory of Ordinary p-adic Curves, RIMS Preprint 1033 (September 1995); *Publ. of RIMS* 32, No. 6 (1996), pp. 957-1151.

[Mzk2] Mochizuki, S., The Generalized Ordinary Moduli of p-adic Hyperbolic Curves, RIMS Preprint 1051 (December 1995); 281 pp.

[Mzk3] Mochizuki, S., Combinatorialization of p-adic Teichmüller Theory, RIMS Preprint 1076 (April 1996); 32 pp.

[Mzk4] Mochizuki, S., Foundations of p-adic Teichmüller Theory, *in preparation.*

[Mzk5] Mochizuki, S., The Local Pro-p Anabelian Geometry of Curves, RIMS Preprint 1097 (August 1996); 84 pp.

[Pop] Pop, F., On Grothendieck's conjecture of birational anabelian geometry II, Preprint (1995).

[Tama1] Tamagawa, A., The Grothendieck conjecture for affine curves, *Compositio Math.* 109, No. 2 (1997), pp. 135-194.

[Tama2] Tamagawa, A., On the fundamental groups of curves over algebraically closed fields of characteristic > 0, RIMS Preprint 1182 (January 1998).

Shinichi Mochizuki
Research Institute
for Mathematical Sciences
Kyoto University
Kyoto 606-01, Japan
Motizuki@kurims.kyoto-u.ac.jp

The Subspace Theorem and Applications

Hans Peter Schlickewei

Abstract. We discuss recent results on simultaneous approximation of algebraic numbers by rationals and applications to diophantine equations.

1991 Mathematics Subject Classification: 11J68,11D61
Keywords and Phrases: Subspace Theorem, Unit Equations, Recurrence Sequences

In 1955 K. F. Roth [15] proved: *Suppose α is an algebraic number and suppose $\varepsilon > 0$. Then the inequality*

$$\left|\alpha - \frac{x}{y}\right| < y^{-2-\varepsilon} \tag{0.1}$$

has only finitely many rational solutions $\frac{x}{y}$. This result is best possible since by Dirichlet's classical theorem any real irrational number α has infinitely many rational approximations satisfying

$$\left|\alpha - \frac{x}{y}\right| < y^{-2}.$$

In 1972 W. M. Schmidt [22] generalized Roth's Theorem to n dimensions. He proved the following:

Subspace Theorem. *Let $L_i = \alpha_{i1}X_1 + \cdots + \alpha_{in}X_n$ $(i = 1, \ldots, n)$ be linearly independent linear forms with algebraic coefficients. Suppose $\varepsilon > 0$. Consider the inequality*

$$|L_1(\mathbf{x}) \cdots L_n(\mathbf{x})| < \| \mathbf{x} \|^{-\varepsilon}, \quad \mathbf{x} \in \mathbb{Z}^n, \tag{0.2}$$

where $\| \mathbf{x} \| = (x_1^2 + \cdots + x_n^2)^{\frac{1}{2}}$.
Then there exist proper linear subspaces $T_1, \ldots, T_t$ of $\mathbb{Q}^n$ such that the set of solutions of (0.2) is contained in the union

$$T_1 \cup \cdots \cup T_t. \tag{0.3}$$

Recently an alternative proof of the Subspace Theorem has been given by Faltings and Wüstholz [12].

It is an easy consequence of a theorem of Minkowski that there exist forms $L_1, \ldots, L_n$ as above with the following property: For any finite collection of proper linear subspaces $S_1, \ldots, S_t$ of $\mathbb{Q}^n$ the set of solutions of

$$|L_1(\mathbf{x}) \cdots L_n(\mathbf{x})| < 1, \quad \mathbf{x} \in \mathbb{Z}^n$$

is not contained in $S_1 \cup \cdots \cup S_t$. So the Subspace Theorem, just as Roth's Theorem is best possible.

W. M. SCHMIDT in 1975 has extended his theorem to the case when the variables $\mathbf{x} = (x_1, \cdots, x_n)$ are integers of a fixed number field K. Moreover the theorem has been generalized by DUBOIS and RHIN [4] and independently by SCHLICKEWEI [17] to include p-adic valuations.

The results mentioned so far are all qualitative, and we may ask the following two questions:

i) Given $\varepsilon > 0$ and linear forms $L_1, \ldots, L_n$ as in the Subspace Theorem, is it possible to determine the subspaces $T_1, \ldots, T_t$ in (0.3) effectively, i.e., to give an algorithm to compute $T_1, \ldots, T_t$?

ii) What can be said about the number t of subspaces $T_1, \ldots, T_t$ needed in (0.3) to cover the set of solutions of (0.2)?

Question (i) is one of the most famous open problems in Diophantine Approximations. Indeed the method of proof for the Subspace Theorem, the so called Thue-Siegel-Roth-Schmidt method, is highly ineffective.

As for question (ii), in the last 15 years quite some progress was made. So in the remainder of the talk we will discuss results on question (ii).

1 The quantitative Subspace Theorem

The Thue-Siegel-Roth-Schmidt method does not provide an algorithm to determine the set of solutions of (0.1) or the subspaces occurring in (0.2), (0.3). However it does give *upper bounds* for the *number* of solutions of (0.1) or of subspaces in (0.2), (0.3). One of the main tools in giving such upper bounds are "gap principles". We illustrate the easiest case:
Let us consider the inequality

$$\left|\alpha - \frac{x}{y}\right| < y^{-2-\varepsilon} \tag{1.1}$$

in rational numbers $\frac{x}{y}$ with $y > 0$. For any two different solutions $\frac{x_1}{y_1}, \frac{x_2}{y_2}$ of (1.1) with $y_1 < y_2$ we get

$$\frac{1}{y_1 y_2} \leq \left|\frac{x_1}{y_1} - \frac{x_2}{y_2}\right| \leq \left|\frac{x_1}{y_1} - \alpha\right| + \left|\alpha - \frac{x_2}{y_2}\right| \leq 2y_1^{-2-\varepsilon}.$$

So if $\frac{x_1}{y_1}, \ldots, \frac{x_k}{y_k}$ are different solutions of (1.1) with $y_1 < y_2 < \ldots < y_k$ and with the y_i-s in an interval of the type $(Q, Q^E]$ with $Q^{\frac{\varepsilon}{2}} > 2$ and $E > 1$, then

$$k \leq 1 + \frac{\log E}{\log(1 + \frac{\varepsilon}{2})}.$$

In the proof of Roth's Theorem we have the following situation: There exists a certain value Q_0, depending upon α and ε, such that for solutions $\frac{x}{y}$ of (0.1) with $y > Q_0$ we can find m disjoint intervals $(Q_1, Q_1^E], \ldots, (Q_m, Q_m^E]$ having

$$y \in (Q_1, Q_1^E] \cup \cdots \cup (Q_m, Q_m^E].$$

So the above gap principle shows that we cannot have more than

$$m\left(1+\frac{\log E}{\log(1+\frac{\varepsilon}{2})}\right)$$

large solutions. Now Roth's method gives

$$m \le c_1(\varepsilon, d) \quad \text{and} \quad E \le c_2(\varepsilon, d),$$

where d is the degree of α. Thus the number of solutions $\frac{x}{y}$ of (0.1) with y "large" does not exceed a certain function $c(d,\varepsilon)$ depending only upon the degree d of α and upon the parameter ε.

To derive a similar statement in the situation of the Subspace Theorem, we first need a generalization of the gap principle to higher dimensions. Apart from this, there are a number of rather delicate problems in the geometry of numbers to be dealt with for dimension > 2. The pioneering work in this context is due to W. M. Schmidt [23] (1989). He proved:

Let $0 < \varepsilon < 1$. Suppose that the forms $L_1, \dots, L_n$ have $\det(L_1, \dots, L_n) = 1$ and that the coefficients of the forms are contained in a number field K with $[K : \mathbb{Q}] = d$. Then there are proper linear subspaces $T_1, \dots, T_t$ of $\mathbb{Q}^n$ where

$$t \le (2d)^{2^{26n}\varepsilon^{-2}} \tag{1.2}$$

such that the set of solutions $\mathbf{x}$ of (0.2) is contained in the union of $T_1, \dots, T_t$ and the ball

$$\| \mathbf{x} \| \le \max\{(n!)^{8/\varepsilon}, H(L_1), \dots, H(L_n)\}, \tag{1.3}$$

where $H(L_i)$ is the height of the coefficient vector of the form L_i $(1 \le i \le n)$.

Vojta [27] has shown that there exist finitely many subspaces $T_1, \dots, T_l$ which are effectively computable and which do not depend upon ε, such that all but finitely many solutions $\mathbf{x}$ of (0.2) are contained in the union $T_1 \cup \dots \cup T_l$. It seems to be very difficult to give an upper bound for the number of exceptional solutions (which in fact will depend upon ε).

Neither one of Schmidt's and Vojta's results implies the other one.

Schmidt's result (1.2), (1.3) has been extended by Schlickewei [18] to the case when the variables $(x_1, \dots, x_n)$ are taken from the field K instead of $\mathbb{Q}$ and also to a finite set S of absolute values on K. To cover the "large" solutions we do not need more than

$$c(n, \varepsilon, d, s) \tag{1.4}$$

proper linear subspaces of K^n. Here s is the cardinality of the set S of absolute values under consideration.

2 Improvements on the quantitative Subspace Theorem.

i) The bound for the number t of subspaces given in (1.2) is doubly exponential in n and exponential in ε^{-1}. Its origin essentially may be found in Roth's Lemma:

This is a criterion to guarantee that a polynomial $P(X_1, \dots, X_m)$ with integer coefficients does not vanish with too high order at a rational point $(\frac{x_1}{y_1}, \dots, \frac{x_m}{y_m})$. A much more powerful multiplicity estimate has been provided by FALTINGS [11] with his product theorem. Explicit versions of FALTINGS' result were derived independently by EVERTSE [6] and by FERRETTI [13]. As a consequence EVERTSE [7] obtained a substantial improvement of the bounds (1.2) and (1.4) in terms of the dependence upon n and ε^{-1}.

ii) A completely different problem is the question which parameters in the bounds (1.2) and (1.4) are really necessary. As will be seen, to minimize the number of parameters showing up in the bounds is a quite relevant task for applications to diophantine equations. In dealing with inequality (0.2), it turns out that it suffices to study a problem on simultaneous inequalities. It is clear that for any solution $\mathbf{x}$ of (0.2) there exist real numbers $c_1, \dots, c_n$ with

$$c_1 + \dots + c_n \leq -\varepsilon \tag{2.1}$$

and

$$|L_1(\mathbf{x})| \leq \| \mathbf{x} \|^{c_1}, \dots, |L_n(\mathbf{x})| \leq \| \mathbf{x} \|^{c_n} . \tag{2.2}$$

Indeed it suffices to study (2.1), (2.2) for a fixed tuple $c_1, \dots, c_n$. Such inequalities have been investigated by SCHLICKEWEI [19]. In the current situation he was able to replace the bound $c(n, \varepsilon, d, s)$ from (1.4) for the number of subspaces by a bound

$$c(n, \varepsilon). \tag{2.3}$$

So here in comparison with (1.4) the dependence on d and s is avoided. However (2.3) still is only valid for the "large" solutions $\mathbf{x}$, and the definition of "large" is in terms of a function

$$c(n, \varepsilon, d, L_i). \tag{2.4}$$

Let us briefly discuss why in (2.4) the parameter d shows up. In the proof of the Subspace Theorem an important ingredient is the theorem of Minkowski on the successive minima $\lambda_1, \dots, \lambda_n$ of convex bodies. By Minkowski we have

$$\frac{2^n}{n!} \leq \lambda_1 \dots \lambda_n V \leq 2^n, \tag{2.5}$$

where V is the volume of the convex body. If we deal with the Subspace Theorem for a number field K, we use the generalization of Minkowski's estimate (2.5) to number fields given by MCFEAT [14] and by BOMBIERI and VAALER [2]. However the analogue of (2.5) involves the discriminant of the field as a factor in the upper bound.

In a recent paper, ROY and THUNDER [16] have proved a version of Minkowski's theorem where they do not restrict the variables $\mathbf{x}$ anymore to a number field. They allow arbitrary elements $\mathbf{x} \in \overline{\mathbb{Q}}^n$, where $\overline{\mathbb{Q}}$ is the algebraic closure of $\mathbb{Q}$. They derive an inequality which essentially is of the same shape as

(2.5). In particular, with their approach they get rid of the discriminant factor in the upper bound.

This new result of Roy and Thunder turns out to be extremely useful in our context. It has been applied in a joint paper by Evertse and Schlickewei [9].

Roughly speaking, the main consequences derived in [9] are as follows.

a) We can now consider inequalities such as (0.2) (or the p-adic generalization) allowing arbitrary solutions $\mathbf{x} \in \overline{\mathbb{Q}}^n$ (instead of only solutions in K^n). The assertion of the theorem then is that the set of all large solutions is contained in the union of finitely many proper subspaces of $\overline{\mathbb{Q}}^n$ (*Absolute Subspace Theorem*).

b) If we consider simultaneous inequalities such as (2.1), (2.2) with solutions $\mathbf{x} \in \overline{\mathbb{Q}}^n$, then again we have to distinguish small and large solutions. The set of large solutions may be covered by

$$c(n, \varepsilon) \tag{2.6}$$

proper subspaces of $\overline{\mathbb{Q}}^n$ (similar bound as in (2.3)). However, in contrast with (2.4) the large solutions now are defined in terms of a function

$$c(n, \varepsilon, L_i) \tag{2.7}$$

only. So in (2.7), in comparison with (2.4), the dependence on the parameter d is avoided.

3 Applications to norm form equations.

Let $L(X_1, \ldots, X_n) = \alpha_1 X_1 + \ldots + \alpha_n X_n$ be a linear form with coefficients in a number field K of degree d. Denote the embeddings of K into $\overline{\mathbb{Q}}$ by $\alpha \to \alpha^{(i)}$ and write $L^{(i)}(\mathbf{X}) = \alpha_1 X_1 + \ldots + \alpha_n X_n$. Put

$$N(L(\mathbf{X})) = \prod_{i=1}^{d} L^{(i)}(\mathbf{X}).$$

By a norm form equation we mean an equation of the type

$$N(L(\mathbf{x})) = m \quad \text{in} \quad \mathbf{x} \in \mathbb{Z}^n. \tag{3.1}$$

Here m is a fixed nonzero rational number. Note that the left hand side of (3.1) is a homogeneous polynomial of degree d in the variables x_i with rational coefficients.

Under suitable and rather natural hypotheses about the linear form L, which we summarize briefly by saying that $N(L(\mathbf{X}))$ is a "nondegenerate" norm form, Schmidt [22] has shown as a consequence of the Subspace Theorem that equation (3.1) has only finitely many solutions $\mathbf{x}$. Using his quantitative result (1.2), in [24] he derived an explicit uniform upper bound for the number of solutions of (3.1) of the shape $c(n, d, m)$. Here the significant feature is that the bound does not depend

upon the coefficients of the form L. This proves the n-dimensional analogue of a conjecture made by SIEGEL in 1929. The corresponding result for $n = 2$ had been proved by EVERTSE in 1984 already. EVERTSE [5], applying his version of the quantitative Subspace Theorem, obtained a considerable improvement on the bound given by SCHMIDT. Further EVERTSE and GYÖRY [8] have studied equations with more general forms, the so called decomposable form equations.

4 UNIT EQUATIONS

Let K be a field of characteristic zero. Let $a_1, \ldots, a_n$ be fixed nonzero elements in K. Consider the equation

$$a_1x_1 + \ldots + a_nx_n = 1. \tag{4.1}$$

We call a solution $(x_1, \ldots, x_n)$ of (4.1) nondegenerate if no nonempty subsum on the left hand side of (4.1) vanishes. Applying the absolute quantitative Subspace Theorem by EVERTSE and SCHLICKEWEI, discussed in section 2, in a recent paper EVERTSE, SCHLICKEWEI and W. M. SCHMIDT [10] proved the following:

Let G be a finitely generated subgroup of the multiplicative group K^* of nonzero elements of K. Suppose G has rank r. Then the number of nondegenerate solutions $(x_1, \ldots, x_n) \in G^n$ of equation (4.1) does not exceed

$$\exp\left(n^{cn}(r+1)\right). \tag{4.2}$$

Here c is an absolute constant.

To prove such a result, we first observe that using a specialization argument, it suffices to deal with the case when K is a number field. Once we have reached this situation, after the transformation $Y_i = a_iX_i$, we may apply the Subspace Theorem to the linear forms in $Y_1, \ldots, Y_n$ given by $L_1(Y_1, \ldots, Y_n) = Y_1, \ldots, L_n(Y_1, \ldots, Y_n) = Y_n, L_{n+1}(Y_1, \ldots, Y_n) = Y_1 + \ldots + Y_n$. Actually we need the p-adic version of the Subspace Theorem, where S, the set of absolute values, consists of all archimedean absolute values of K together with those finite absolute values corresponding to the prime ideals dividing the coefficients a_i and the generators of the group G. In this application ε turns out to be a function of n only.

The results given in section 0 simply imply that we get only finitely many solutions. The results of section 1, in view of (1.4) give a bound depending upon the degree d of the number field K and upon the cardinality s of the set S. In particular, if at the beginning K is not a number field, our result will depend upon the specialization. Moreover in general, the parameter s will be much larger than the rank r of the group G. In [19] SCHLICKEWEI introduced a method which in conjunction with (2.3) allows it to derive a bound for the number of large solutions of (4.1) which in fact does not involve the cardinality s of S but only the rank r of the original group G. So (2.3) already would give a bound of type (4.2) for the number of large solutions.

There remain the small solutions. Before we had the bound (2.7) from the Absolute Subspace Theorem, the definition of the small solutions always depended

on the degree of the number field K. Clearly then the specialization argument has a deadly impact, as then the degree of the number field, we end up with after the specialization, appears in the final result.

It is at this point where the Absolute Subspace Theorem comes in. Here the small solutions are defined in terms of the forms L_i, of n and of ε only. In view of the particular shape of our forms L_i and as ε is a function of n only, the small solutions by (2.7) now are defined in terms of n only. In particular the definition of the size of the small solutions is completely independent of the number field obtained with the specialization argument.

To exploit successfully this bound for the small solutions, we needed a new gap principle. Here the results of Zhang [29], [30] on lower bounds for the heights of points on varieties are crucial. Using the elementary method introduced in this context by Zagier [28], in dimension 2 such a new gap principle was first given in a paper by Schlickewei and Wirsing [21]. For general n, Bombieri and Zannier [3] gave an elementary proof of Zhang's result and obtained a gap principle which is suitable for our purposes. This has been improved substantially by W. M. Schmidt [25].

Results on equation (4.1) apply in particular to linear recurrence sequences, i.e., to sequences $\{u_n\}_{n\in\mathbb{Z}}$ satisfying a relation

$$u_{n+k} = a_{k-1}u_{n+k-1} + \ldots + a_0u_n. \tag{4.3}$$

Here we assume that $a_0 \neq 0$ and that we have initial values $(u_0, \ldots, u_{k-1}) \neq (0, \ldots, 0)$. Writing

$$G(z) = z^k - a_{k-1}z^{k-1} - \ldots - a_0 = \prod_{i=1}^{r}(z-\alpha_i)^{\rho_i} \tag{4.4}$$

with distinct roots α_i of multiplicity ρ_i, it is well known that we have

$$u_n = \sum_{i=1}^{r} f_i(n)\alpha_i^n, \tag{4.5}$$

where the f_i are polynomials of respective degrees $\leq \rho_i - 1$. An old conjecture says that for a nondegenerate sequence u_n of order k the equation

$$u_n = 0 \quad (n \in \mathbb{Z})$$

does not have more than $c(k)$ solutions, where $c(k)$ is a function depending on k only.

For $k = 3$, this conjecture has been proved by Schlickewei [20]. Later Beukers and Schlickewei [1] derived the estimate $c(3) \leq 61$. For general k and for sequences u_n such that the companion polynomial $G(z)$ given in (4.4) has only simple zeros, in view of (4.5), the conjecture is an easy consequence of the theorem of Evertse, Schlickewei and W. M. Schmidt [10] on equation (4.1). In fact, if the zeros α_i in (4.4) are simple the polynomials f_i in (4.5) reduce to constants. The general case of the conjecture with arbitrary polynomial coefficients has been settled recently by W. M. Schmidt [26].

References

[1] Beukers, F., Schlickewei, H.P.: *The equation $x + y = 1$ in finitely generated groups*, Acta Arith. 78 (1996), 189–199.

[2] Bombieri, E., Vaaler, J.: *On Siegel's Lemma*, Invent. Math. 73 (1983), 11–32.

[3] Bombieri, E., Zannier, U.: *Algebraic Points on Subvarieties of $\mathbb{G}_m^n$*, Intern. Math. Research Notices 7 (1995), 333–347.

[4] Dubois, E., Rhin, G.: *Approximations rationnelles simultanées de nombres algébriques réels et de nombres algébriques p-adiques*, Journées Arithmétiques de Bordeaux, Astérisque 24–25 (1975), 211–227.

[5] Evertse, J.-H.: *The number of solutions of decomposable form equations*, Invent. Math. 122 (1995), 559–601.

[6] Evertse, J.-H.: *An explicit version of Faltings' Product Theorem and an improvement of Roth's Lemma*, Acta Arith. 73 (1995), 215–248.

[7] Evertse, J.-H.: *An improvement of the quantitative Subspace theorem*, Comp. Math. 101 (1996), 225–311.

[8] Evertse, J.-H., Győry: *The number of families of solutions of decomposable form equations*, Acta Arith. 80 (1997), 367–394.

[9] Evertse, J.-H., Schlickewei, H.P.: *A quantitative version of the absolute Subspace Theorem*, to appear.

[10] Evertse, J.-H., Schlickewei, H. P., Schmidt, W. M.: *Linear equations with variables which lie in a multiplicative group*, to appear.

[11] Faltings, G.: *Diophantine approximation on abelian varieties*, Ann. of Math. 133 (1991), 549–576.

[12] Faltings, G., Wüstholz, G.: *Diophantine approximations on projective spaces*, Inv. math. 116 (1994), 109–138.

[13] Ferretti, R.: *An effective version of Faltings' Product Theorem*, Forum Math. 8 (1996), 401–427

[14] McFeat, R.B.: *Geometry of numbers in adèle spaces*, Dissertationes Mathematicae 88, PWN Polish Scientific Publishers, Warsaw 1971

[15] Roth, K. F.: *Rational approximations to algebraic numbers*, Mathematika 2 (1955), 1–20.

[16] Roy, D., Thunder, J.L.: *An absolute Siegel's Lemma*, J. reine angew. Math. 476 (1996), 1–26

[17] Schlickewei, H. P.: *The $\wp$-adic Thue-Siegel-Roth-Schmidt theorem*, Arch. Math. 29 (1977), 267–270.

[18] Schlickewei, H. P.: *The quantitative Subspace Theorem for number fields*, Compos. Math. 82 (1992), 245–274.

[19] Schlickewei, H. P.: *Multiplicities of recurrence sequences*, Acta Math. 176 (1996), 171–243.

[20] Schlickewei, H. P.: *The multiplicity of binary recurrences*, Invent. Math. 129 (1997), 11–36.

[21] Schlickewei, H.P., Wirsing, E.: *Lower bounds for the heights of solutions of linear equations*, Invent. Math. 129 (1997), 1–10.

[22] Schmidt, W. M.: *Norm form equations*, Ann. of Math. 96 (1972), 526–551.

[23] Schmidt, W. M.: *The Subspace Theorem in diophantine approximation*, Compos. Math 69 (1989), 121–173.

[24] Schmidt, W. M.: *The number of solutions of norm form equations*, Transactions A.M.S. 317 (1990), 197–227.

[25] Schmidt, W. M.: *Heights of points on subvarieties of* $\mathbb{G}_m^n$, Séminaire de théorie des nombres de Paris, 1993-1994 (ed. by S. David), 157–187, Cambridge Un. Press, 1996.

[26] Schmidt, W. M.: *The zero multiplicity of linear recurrence sequences*, to appear.

[27] Vojta, P.: *A refinement of Schmidt's subspace theorem*, Amer. J. Math. 111 (1989), 489–518.

[28] Zagier, D.: *Algebraic numbers close to both* 0 *and* 1, Math. Computation 61 (1993), 485–491.

[29] Zhang, S.: *Positive line bundles on arithmetic surfaces*, Annals of Math. 136 (1992), 569–587.

[30] Zhang, S.: *Positive line bundles on arithmetic varieties*, Journal A.M.S. 8 (1995), 187–221.

Hans Peter Schlickewei
Fachbereich Mathematik
Philipps-Universität Marburg
Lahnberge
D-35032 Marburg
Germany

p-Adic Hodge Theory in the Semi-Stable Reduction Case

Takeshi Tsuji

Abstract. We survey the statement and the proof (by K. Kato and the author) of the semi-stable conjecture of Fontaine-Jannsen on p-adic étale cohomology and crystalline cohomology, generalizing it to truncated simplicial schemes. Thanks to the alteration of de Jong, this generalization especially implies that the p-adic étale cohomology of any proper variety (which may have singularity) is potentially semi-stable.

1991 Mathematics Subject Classification: 14F30, 14F20
Keywords and Phrases: p-adic étale cohomology, crystalline cohomology, semi-stable reduction, simplicial scheme

§1. Introduction.

The p-adic Hodge theory is an analogue of the Hodge theory for a variety X over a p-adic field K (a complete discrete valuation field of mixed characteristic $(0, p)$ with perfect residue field) and it compares p-adic étale cohomology with the action of the Galois group and de Rham cohomology with some additional structures (depending on how good the reduction of the variety is). In the semi-stable reduction case, it was formulated by J.-M. Fontaine and U. Jannsen [Fo3], [Fo4] as a conjecture (called the semi-stable conjecture or C_{st} for short), and it asserts that p-adic étale cohomology with the action of the Galois group and de Rham cohomology with the Hodge filtration and certain additional structures coming from log crystalline cohomology of the special fiber can be constructed from each other. (See §2 for more details).

The conjecture was studied by many people such as J.-M. Fontaine, W. Messing, S. Bloch, K. Kato, G. Faltings and O. Hyodo (cf. [Fo-M], [Bl-K], [Fa1], [Fa2], [H], [H-K], [K]), and finally it was completely solved by the author [T1]. It was also proved by G. Faltings by a different method [Fa4] afterwards together with its generalization to relative cohomology, non-constant coefficients and an open variety. A new proof was also given by W. Niziol using K-theory at least in the good reduction case [Ni]. Also a theory for p-torsion cohomology in the semi-stable reduction case was established by G. Faltings and C. Breuil ([Fa3], [Br2], [Br3]) by generalizing the theory of Fontaine-Laffaille in the good reduction case [Fo-L].

In these notes, I give a survey of the statement and the proof [K], [T1] of C_{st}, generalizing it to truncated simplicial schemes (an analogue of [D2]). Thanks

to the alteration of de Jong [dJ], this generalization implies that the p-adic étale cohomology of any proper variety (which may have singularity) is potentially semi-stable. The details of the proof for simplicial schemes will be given elsewhere. Unlike [K], [T1], here we use the log version of the syntomic and the syntomic-étale sites [Fo-M] to define log syntomic cohomology [Br3] and to construct the map from log syntomic cohomology to p-adic étale cohomology since it is easily generalized to simplicial log schemes.

NOTATION: Let K be a complete discrete valuation field of mixed characteristic $(0, p)$ whose residue field k is perfect. Let W be the ring of Witt-vectors with coefficients in k and let K_0 denote the field of fractions of W. Let $\overline{K}$ be an algebraic closure of K and set $G_K := \mathrm{Gal}(\overline{K}/K)$. We choose and fix a uniformizer π of K. For a ring, a scheme or a log scheme over W, we denote its reduction mod p^n by the subscript n.

§2. THE SEMI-STABLE CONJECTURE.

(2.1) We first recall the theory of semi-stable representations by Fontaine ([Fo1], [Fo2], [Fo3], [Fo4]) briefly. We need the rings $B_{\mathrm{st}} \subset B_{\mathrm{dR}}$ associated to K, which have the following structures and properties. The ring B_{dR} is a complete discrete valuation field with residue field $\hat{\overline{K}}$ endowed with an action of G_K. B_{dR} is filtered by the discrete valuation and it contains $\mathbb{Q}_p(r)$ $(r \in \mathbb{Z})$ and $\overline{K}$. We have $B_{\mathrm{dR}}^{G_K} = K$ and the image of a non-zero element of $\mathbb{Q}_p(1)$ is a uniformizer of B_{dR}. B_{st} is a G_K-stable subring of B_{dR} containing $\mathbb{Q}_p(r)$ $(r \in \mathbb{Z})$ and K_0 and endowed, additionally, with the Frobenius φ and the monodromy operator N satisfying $N\varphi = p\varphi N$. The natural homomorphism $B_{\mathrm{st}} \otimes_{K_0} K \to B_{\mathrm{dR}}$ is injective, $B_{\mathrm{st}}^{G_K} = K_0$ and $Fil^r B_{\mathrm{dR}} \cap B_{\mathrm{st}}^{\varphi=p^r,\, N=0} = \mathbb{Q}_p(r)$ $(r \in \mathbb{Z})$. The ring B_{st} with the actions of G_K, φ, N is independent of π, but the embedding $B_{\mathrm{st}} \hookrightarrow B_{\mathrm{dR}}$ depends on it.

By a p-adic representation of G_K, we mean a $\mathbb{Q}_p$-vector space of finite dimension endowed with a continuous linear action of G_K, and, by a filtered (φ, N)-module, a K_0-vector space of finite dimension D endowed with a semi-linear automorphism φ, a linear endomorphism N and an exhaustive and separated filtration Fil^i $(i \in \mathbb{Z})$ on $D_K := D \otimes_{K_0} K$ such that $N\varphi = p\varphi N$. Then, to a p-adic representation V, one can associate a filtered (φ, N)-module $D_{\mathrm{st}}(V) := (B_{\mathrm{st}} \otimes_{\mathbb{Q}_p} V)^{G_K}$ functorially. We have $\dim_{K_0} D_{\mathrm{st}}(V) \leq \dim_{\mathbb{Q}_p} V$ and we say V is *semi-stable* if the equality holds. The restriction of D_{st} to the category of semi-stable representations is fully faithful and exact; its quasi-inverse is given by $V_{\mathrm{st}}(D) := Fil^0(B_{\mathrm{dR}} \otimes_K D_K) \cap (B_{\mathrm{st}} \otimes_{K_0} D)^{\varphi=1,\, N=0}$.

(2.2) Let X be a scheme over O_K isomorphic to the finite base change of a proper semi-stable scheme. Then, using log crystalline cohomology, one can give a canonical (φ, N)-module structure D^q on $H^q_{\mathrm{dR}}(X_K/K)$ (depending on π) (see §3). We have the following theorem conjectured by J.-M. Fontaine and U. Jannsen ([T1], [Fa4], see also [Fo-M], [Fa2], [K-M], [K]) .

THEOREM 2.2.1. (C_{st}). *With the notation and the assumption as above, the p-adic representation $H^q_{\text{ét}}(X_{\overline{K}}, \mathbb{Q}_p)$ is semi-stable, and there exists a canonical isomorphism of filtered (φ, N)-modules $D_{\mathrm{st}}(H^q_{\text{ét}}(X_{\overline{K}}, \mathbb{Q}_p)) \cong D^q$ functorial on X and*

compatible with the product structures, Chern classes of vector bundles on X_K and cycle classes of cycles on X_K.

We need more arguments than [T1] for the compatibility with Chern classes and cycle classes, which will be given elsewhere.

Our generalization to truncated simplicial schemes is the following:

Theorem 2.2.2. (=Theorem 7.1.1). *Let m be a non-negative integer and let $X^{\cdot}$ be an m-truncated simplicial scheme such that each X^i $(0 \leq i \leq m)$ satisfies the assumption on X above. Then $\tilde{H}^q_{\text{ét}}(X^{\cdot}_{\overline{K}}, \mathbb{Q}_p)$ is semi-stable and there exists a canonical isomorphism of filtered (φ, N)-modules $D_{\text{st}}(\tilde{H}^q_{\text{ét}}(X^{\cdot}_{\overline{K}}, \mathbb{Q}_p)) \cong \tilde{H}^q_{\text{log-crys}}(Y^{\cdot}/W)$. (See §6 for the definition of $\tilde{H}^q_{\text{ét}}$ and $\tilde{H}^q_{\text{log-crys}}$.)*

Thanks to the compactification theorem of Nagata [Na], the alteration of de Jong [dJ] and cohomological descent [SD] (cf. [D2]), we obtain the following corollary (cf. (6.4.1)).

Corollary 2.2.3. *For any proper scheme X_K over K, $H^q_{\text{ét}}(X_{\overline{K}}, \mathbb{Q}_p)$ is potentially semi-stable.*

§3. Log crystalline cohomology and de Rham cohomology.

We will survey the log crystalline cohomology defined and studied in [H-K] briefly (see also [T1] §4).

(3.1) Let S denote Spec (O_K) endowed with the log structure defined by the closed point. Let $i_{n,\pi}\colon S_n \to E_n$ be the PD-envelope of the exact closed immersion of S_n into an affine line Spec $(W_n[T])$ with the log structure on the origin defined by sending T to the chosen π. We have $\Gamma(E_n, \mathcal{O}_{E_n}) \cong W[T, T^{me}/m!\ (m \geq 1)] \otimes W_n$, which we denote by R_{E_n}, where $e := [K : K_0]$. Put $R_E := \mathbb{Q} \otimes \varprojlim_n R_{E_n}$. Let $i_{n,0}\colon \underline{W}_n \hookrightarrow E_n$ be the exact closed immersion defined by the ideal generated by T, $T^{me}/m!$ $(m \geq 1)$. $\underline{W}_n$ and E_n have canonical liftings of Frobenius compatible with $i_{n,0}$ defined by $T \mapsto T^p$.

(3.2) Let X be a smooth fine saturated log scheme over S whose underlying scheme is proper over O_K. Let Y denote the special fiber of X (as a log scheme) and assume its underlying scheme is reduced. We consider three kinds of crystalline cohomology $H^*_{\text{crys}}(Y/\underline{W}_n)$, $H^*_{\text{crys}}(X_n/E_n)$, $H^*_{\text{crys}}(X_n/S_n) \cong H^*_{\text{dR}}(X_n/S_n)$, which we write without crys in the following. The first two are endowed with φ, N satisfying $N\varphi = p\varphi N$ compatible with the pull-back by $\{i_{n,0}\}$. We denote the $\varprojlim_n$ of these cohomology groups by the symbols without the subscript n.

Theorem 3.2.1. (Hyodo-Kato [H-K] §5, cf. [T1] §4.4). *There exists a unique K_0-linear section $s\colon H^q(Y/\underline{W})_{\mathbb{Q}} \to H^q(X/E)_{\mathbb{Q}}$ compatible with φ of the pull-back by $\{i_{0,n}\}$. It is also compatible with N and induces an isomorphism*

$$R_E \otimes_{K_0} H^q(Y/\underline{W})_{\mathbb{Q}} \xrightarrow{\sim} H^q(X/E)_{\mathbb{Q}}. \tag{3.2.2}$$

Furthermore the composite with the pull-back by $\{i_{\pi,n}\}$ induces an isomorphism

$$\rho_\pi\colon K \otimes_{K_0} H^q(Y/\underline{W})_{\mathbb{Q}} \xrightarrow{\sim} H^q(X/S)_{\mathbb{Q}} \cong H^q_{\text{dR}}(X_K/K). \tag{3.2.3}$$

Thus $H^*_{\mathrm{dR}}(X_K/K)$ is endowed naturally with a (φ, N)-module structure.

§4. Log syntomic cohomology.

We will survey the log version [Br1], [Br3] of the theory of syntomic cohomology [Fo-M]. See [Br1], [Br3] for details except the proof of Proposition 4.4.1 in the case $r \geq p$.

(4.1) To make the theory compatible with the theory of the log syntomic-étale site in §5, we change the topology slightly; we define the big and the small syntomic site X_{SYN}, X_{syn} of a fine log scheme X using syntomic morphisms $f\colon Y \to Z$ of fine log schemes in the sense of Kato [K] (2.5) such that the underlying morphisms of schemes of f are locally quasi-finite and that the cokernels of $(f^*M_Z)^{\mathrm{gp}} \to M_Y^{\mathrm{gp}}$ are torsion, which we will call *strictly syntomic* morphisms in these notes. Every proof in [Br1], [Br3] still works for this modified syntomic site. The big syntomic topos is functorial, but the small syntomic topos is not. However the small syntomic site is functorial as a topology in the sense of Artin [A], and it is sufficient in our application. For an exact nilimmersion, the direct image functors of the big and the small syntomic topos are exact.

(4.2) For a fine log scheme with a quasi-coherent PD-ideal (T, I, γ) such that $n\mathcal{O}_T = 0$ for some positive integer n and a fine log scheme X over T, we have the big crystalline site with syntomic topology $(X/T, I, \gamma)_{\mathrm{CRYS,\,SYN}}$ (or $(X/T)_{\mathrm{CRYS,\,SYN}}$ for short), and we have a commutative diagram of topos:

$$\begin{array}{ccccc} (X/T)^{\sim}_{\mathrm{CRYS,\,SYN}} & \xrightarrow{\alpha} & (X/T)^{\sim}_{\mathrm{CRYS}} & \xrightarrow{\beta} & (X/T)^{\sim}_{\mathrm{crys}} \\ \downarrow{\scriptstyle U_{X/T,\,\mathrm{SYN}}} & & \downarrow{\scriptstyle U_{X/T}} & & \downarrow{\scriptstyle u_{X/T}} \\ X^{\sim}_{\mathrm{SYN}} & \longrightarrow & X^{\sim}_{\mathrm{\acute{E}T}} & \longrightarrow & X^{\sim}_{\mathrm{\acute{e}t}}. \end{array} \tag{4.2.1}$$

It is easy to see that β_* is exact.

PROPOSITION 4.2.2. ([Br1] §3). *We have* $RU_{X/T,\,\mathrm{SYN}*}J^{[r]}_{X/T} = U_{X/T,\,\mathrm{SYN}*}J^{[r]}_{X/T}$ *and* $R\alpha_*J^{[r]}_{X/T} = J^{[r]}_{X/T}$.

(4.3) Let us return to the situation in (3.2). We denote by $\mathcal{O}^{\mathrm{crys}}_n$ and $J^{[r]}_n$ the restriction of $U_{X_n/W_n,\,\mathrm{SYN}*}\mathcal{O}_{X_n/W_n}$ and $U_{X_n/W_n,\,\mathrm{SYN}*}J^{[r]}_{X_n/W_n}$ to $(X_n)_{\mathrm{syn}}$ and also their direct images in $(X_m)_{\mathrm{syn}}$ $(m \geq n)$. Here W_n is endowed with the trivial log structure. By Proposition 4.2.2, the cohomology of these sheaves give us $H^*_{\mathrm{crys}}(X_n/W_n, \mathcal{O}_{X_n/W_n})$ and $H^*_{\mathrm{crys}}(X_n/W_n, J^{[r]}_{X_n/W_n})$, and we have $\Gamma(Y, \mathcal{O}^{\mathrm{crys}}_n) = \Gamma_{\mathrm{crys}}(Y/W_n, \mathcal{O}_{Y/W_n}) = \Gamma_{\mathrm{crys}}(Y_1/W_n, \mathcal{O}_{Y_1/W_n})$. By the last equality, $\mathcal{O}^{\mathrm{crys}}_n$ is naturally endowed with the Frobenius endomorphism φ. $\mathcal{O}^{\mathrm{crys}}_n$ and $J^{[r]}_n$ are flat over $\mathbb{Z}/p^n\mathbb{Z}$ and $\mathcal{O}^{\mathrm{crys}}_{n+1} \otimes \mathbb{Z}/p^n\mathbb{Z} \cong \mathcal{O}^{\mathrm{crys}}_n$, $J^{[r]}_{n+1} \otimes \mathbb{Z}/p^n\mathbb{Z} \cong J^{[r]}_n$ ([Br3] 3.1.4).

(4.4) We see easily $\varphi(J^{[r]}_n) \subset p^r\mathcal{O}^{\mathrm{crys}}_n$ if $r \leq p-1$. However, this is false in general and we use the following modification of $J^{[r]}_n$: $J^{<r>}_n := \{x \in J^{[r]}_{n+s} | \varphi(x) \in p^r\mathcal{O}^{\mathrm{crys}}_{n+s}\}/p^n$ $(s \geq r)$. The right hand side is independent of s, $J^{<r>}_n$ is flat over

$\mathbb{Z}/p^n\mathbb{Z}$ and $J_{n+1}^{<r>} \otimes \mathbb{Z}/p^n\mathbb{Z} \cong J_n^{<r>}$. Define $\varphi_r \colon J_n^{<r>} \to \mathcal{O}_n^{\mathrm{crys}}$ by setting $\varphi_r(x \bmod p^n) = y \bmod p^n$ for $x \in J_{n+r}^{[r]}$, $y \in \mathcal{O}_{n+r}^{\mathrm{crys}}$ such that $\varphi(x) = p^r y$. Set $S_n^r := \mathrm{Ker}(1 - \varphi_r \colon J_n^{<r>} \to \mathcal{O}_n^{\mathrm{crys}})$.

Proposition 4.4.1. (cf. [Fo-M] III 1.1, [Br3] 3.1.4). *The following sequence is exact for $r \geq 0$:* $0 \longrightarrow S_n^r \longrightarrow J_n^{<r>} \overset{1-\varphi_r}{\longrightarrow} \mathcal{O}_n^{\mathrm{crys}} \longrightarrow 0$.

We have a natural product structure $S_n^r \otimes S_n^{r'} \to S_n^{r+r'}$. The presheaf $Y \mapsto \Gamma(Y, M_Y^{\mathrm{gp}})$ on $(X_n)_{\mathrm{syn}}$ is a sheaf, and we denote it by M_n^{gp}. We have a symbol map $M_{n+1}^{\mathrm{gp}} \to S_n^1[1]$ (in the derived category) (cf. [Fo-M] III 6.3).

(4.5) We define $H^q(X, S_n^r)$ to be $H^q((X_{n+s})_{\mathrm{syn}}, S_n^r)$ $(s \geq r)$ and $H^q(\overline{X}, S_n^r)$ to be the inductive limit of $H^q(X', S_n^r)$, where $X' = X \times_S S'$ with S' the log scheme associated to a finite extension K' of K contained in $\overline{K}$. We write the $\mathbb{Q} \otimes \varprojlim_n$ of these cohomology groups by the same symbols with S_n^r replaced by $S_{\mathbb{Q}_p}^r$ (cf. [Fo-M] III 1.2).

§5. The log syntomic-étale site.

We will give a log version of the theory of the syntomic-étale site in [Fo-M]. In this section, by a formal scheme, we mean a locally noetherian formal scheme locally of finite type over Spf (W).

(5.1) The notion of log structure is easily extended to formal schemes. We say that a morphism $f \colon \mathfrak{X} \to \mathfrak{Y}$ of fine log formal schemes is étale, smooth, syntomic, strictly syntomic and an exact closed immersion if, for every integer $n \geq 1$, its reduction mod p^n is étale, smooth, ... respectively. We say f is *étale on the generic fibers*, if étale locally on the underlying formal scheme of $\mathfrak{X}$, f has a factorization $\mathfrak{X} \overset{i}{\to} \mathfrak{Z} \overset{g}{\to} \mathfrak{Y}$ with $\mathfrak{Z}$ affine, i an exact closed immersion and g smooth such that $K_0 \otimes_W \Gamma(\mathfrak{X}, \mathcal{I}/\mathcal{I}^2) \to K_0 \otimes_W \Gamma(\mathfrak{X}, i^*\Omega^1_{\mathfrak{Z}/\mathfrak{Y}})$ is an isomorphism, where $\mathcal{I}$ is the ideal of $\mathcal{O}_{\mathfrak{Z}}$ defining $\mathfrak{X}$. We say f is *syntomic-étale* if it is strictly syntomic and étale on the generic fibers. For a fine log scheme $\mathfrak{X}$, we define the small syntomic-étale site $\mathfrak{X}_{\mathrm{sé}}$ using syntomic-étale morphisms. We define $X_{\mathrm{sé}}$ similarly for a fine log formal scheme X over W. These sites are functorial only as topologies in the sense of Artin [A].

(5.2) Let us return to the situation in (3.2). Let $\hat{X}$ denote the p-adic completion of X. Then we have the following commutative diagram of topos, where the subscript ét denotes the étale site of the underlying scheme or formal scheme.

$$\begin{array}{ccccc}
\hat{X}^{\sim}_{\mathrm{sé}} & \xrightarrow{i_{\mathrm{sé}}} & X^{\sim}_{\mathrm{sé}} & \xleftarrow{j_{\mathrm{sé}}} & (X_K)^{\sim}_{\mathrm{sé}} \\
\hat{\varepsilon}\downarrow & & \varepsilon\downarrow & & \varepsilon_K\downarrow \\
Y^{\sim}_{\mathrm{ét}} = \hat{X}^{\sim}_{\mathrm{ét}} & \xrightarrow{i_{\mathrm{ét}}} & X^{\sim}_{\mathrm{ét}} & \xleftarrow{j_{\mathrm{ét}}} & (X_K)^{\sim}_{\mathrm{ét}}
\end{array}$$

Lemma 5.2.1. (cf. [Fo-M] III 4.1). *The direct image functor $i_{n*} \colon (X_n)^{\sim}_{\mathrm{syn}} \to \hat{X}^{\sim}_{\mathrm{sé}}$ is exact for any integer $n \geq 1$.*

Here we need the additional condition on log structures in the definition of strictly syntomic morphisms. We also denote by the same letter the direct image of S_n^r on $\hat{X}_{\text{sé}}$, whose cohomology coincides with $H^*(X, S_n^r)$.

PROPOSITION 5.2.2. (cf. [Fo-M] III 4.4). *The functor* $\mathcal{F} \mapsto (i_{\text{sé}}^*\mathcal{F}, j_{\text{sé}}^*\mathcal{F}, i_{\text{sé}}^*\mathcal{F} \to i_{\text{sé}}^* j_{\text{sé}*} j_{\text{sé}}^*\mathcal{F})$ *from the category of sheaves on* $X_{\text{sé}}$ *to the category of triples* $(\mathcal{G}, \mathcal{H}, \mathcal{G} \to i_{\text{sé}}^* j_{\text{sé}*}\mathcal{H})$ *where* $\mathcal{G}$ *(resp.* $\mathcal{H}$*) are sheaves on* $\hat{X}_{\text{sé}}$ *(resp.* $(X_K)_{\text{sé}}$*) is an equivalence of categories.*

Using A_{crys} of a sufficiently small $Y \in \text{Ob}(X_{\text{sé}})$ and the exact sequence [T1] A3.26 for this A_{crys}, we can construct a natural homomorphism $S_n^r \to i_{\text{sé}}^* j_{\text{sé}*} j'_* \mathbb{Z}/p^n\mathbb{Z}(r)'$ compatible with the product structures and with the symbol maps $M_{n+1}^{\text{gp}} \to S_n^1[1]$, $\mathcal{O}_{X_{\text{triv}}}^* \to \mathbb{Z}/p^n\mathbb{Z}(1)[1]$. Here $\mathbb{Z}/p^n\mathbb{Z}(r)' = (p^a a!)^{-1}\mathbb{Z}_p(r)/p^n$ $(r = (p-1)a + b,\ a, b \in \mathbb{Z},\ 0 \le b < p-1)$ (cf. [Fo-M] III §5), X_{triv} is the locus on X_K where the log structure is trivial, and j' denotes $(X_{\text{triv}})_{\text{ét}}^{\sim} \to (X_K)_{\text{sé}}^{\sim}$. By Proposition 5.2.2, we can glue S_n^r and $j'_*\mathbb{Z}/p^n\mathbb{Z}(r)'$ by this homomorphism and obtain a sheaf $\mathcal{S}_n^r$ on $X_{\text{sé}}$.

PROPOSITION 5.2.3. *The base change morphism* $i_{\text{ét}}^* R\varepsilon_* \to R\hat{\varepsilon}_* i_{\text{sé}}^*$ *is an isomorphism.*

This is stated in [K-M] in the case of schemes, but its proof is not sufficient. We need to prove that, for an injective sheaf on $X_{\text{sé}}$, $i_{\text{sé}}^*\mathcal{F}$ is $R\hat{\varepsilon}_*$-acyclic and the étale sheafification of the pre-sheaf pull-back of $\mathcal{F}$ on $\hat{X}_{\text{sé}}$ is a syntomic-étale sheaf. From this proposition and the proper base change theorem for étale cohomology, we obtain a homomorphism $H^q(X, S_n^r) \to H_{\text{ét}}^q(X_{\text{triv}}, \mathbb{Z}/p^n\mathbb{Z}(r)')$. Taking the inductive limit with respect to the finite base changes of X, we obtain a homomorphism

$$H^q(\overline{X}, S_n^r) \longrightarrow H_{\text{ét}}^q((X_{\text{triv}})_{\overline{K}}, \mathbb{Z}/p^n\mathbb{Z}(r)') \tag{5.2.4}$$

compatible with the action of G_K. We can prove the following theorem by modifying the argument in [T1] slightly and using Proposition 4.4.1.

THEOREM 5.2.5. *If* X *is isomorphic to the finite base change of a proper semistable scheme endowed with the log structure defined by the special fiber. Then, for* $q \le r$*, there exists* $N \ge 0$ *such that the kernel and the cokernel of* (5.2.4) *are killed by* p^N *for every* $n \ge 1$.

(5.2.4) was proven to be an isomorphism if $r \le p-2$ by M. Kurihara and K. Kato before [T1]. In fact, we can prove the theorem without the assumption on X.

§6. Cohomology of truncated simplicial topos.

(6.1) We can relax the condition (b) in the definition of D-topos in [SD] (1.2.1) as follows. For a bifibered category E over a $\mathcal{U}$-small category D whose fibers are $\mathcal{U}$-topos, $\underline{\Gamma}(E)$ is a $\mathcal{U}$-topos (cf. [SD] (1.2.12)), and, for a functor $f\colon D' \to D$ with D' $\mathcal{U}$-small, $f^*\colon \underline{\Gamma}(E) \to \underline{\Gamma}(D' \times_D E)$ has a left and a right adjoints (cf. [SD] (1.2.9)).

For a D-functor $\varphi_*\colon E \to E'$ induced by a cartesian D^0-functor $\psi\colon T' \to T$ of fibered categories over D° whose fibers are topologies in the sense of Artin [A] such that ψ_i $(i \in \mathrm{Ob}(D))$ and $m^*\colon T_i \to T_j$, $m^* : T'_i \to T'_j$ $(m\colon i \to j \in \mathrm{Mor}(D))$ are morphisms of topologies, one can calculate $R^+\underline{\Gamma}(\varphi_*)$ "fiber by fiber" (cf. [SD] (1.3.12)). These facts are necessary to apply [SD] to the syntomic and the syntomic-étale sites.

(6.2) Let Δ (resp. $\Delta[m]$) denotes the category of the ordered sets $[n] = \{0, \ldots, n\}$ (resp. such that $n \le m$) and the increasing maps. For a ringed topos $(S, \mathcal{O}_S)$, we can define the triangulated functor $R^+\varepsilon_{\cdot *}\colon D^+(\underline{\Gamma}(S \times \Delta[m]), \mathcal{O}_S) \to D^+(S, \mathcal{O}_S)$ by associating to $\Delta[m] \to C^+(S, \mathcal{O}_S); [n] \mapsto K^{\cdot n}$ the simple complex associated to $K^{\cdot 0} \to K^{\cdot 1} \to \ldots \to K^{\cdot m} \to 0 \to \ldots$ with $d^n = \sum_{0 \le i \le n+1} (-1)^i \partial^i$ $(n \le m)$ (cf. [SD] (2.3.9)). By the filtration bête with respect to the second index, the above functor factors through $D^+F(S, \mathcal{O}_S)$ the derived category of filtered complexes (cf. [SD] (2.5.3)), and we have a spectral sequence

$$(6.2.1)\quad E_1^{a,b} = H^b(K^{\cdot a}) \quad (\text{if } 0 \le a \le m),\ 0 \quad (\text{otherwise}) \Longrightarrow H^{a+b}(R^+\varepsilon_{\cdot *}K^{\cdot\cdot})$$

For $K^\cdot \in D^+(\underline{\Gamma}(S \times \Delta), \mathcal{O}_S)$, we have a natural morphism $R^+\varepsilon_*(K^\cdot) \to R^+\varepsilon_{\cdot *}(L^+i_m^* K^\cdot)$, which is a quasi-isomorphism in degree $\le m-1$ if $H^q(K^\cdot) = 0$ $(q < 0)$. Here ε_* is as in [SD] (2.1.1) and i_m^* denotes the functor $\underline{\Gamma}(S \times \Delta) \to \underline{\Gamma}(S \times \Delta[m])$: the composite with $\Delta[m] \to \Delta$.

(6.3) Now we will discuss the simplicial version of (3.2). Let $X^\cdot \to S$ be an m-truncated simplicial fine log scheme whose components satisfy the assumption on X in (3.2). We denote by $R\Gamma^\cdot(Y^\cdot/\underline{W}_n)$ the derived direct image of the structure sheaf under the morphism of topos $(Y^\cdot/\underline{W}_n)^\sim_{\mathrm{crys}} \to \underline{\Gamma}((\mathrm{Sets}) \times \Delta[m])$ defined by taking global sections on each component. We define $R\Gamma^\cdot(X_n^\cdot/S_n)$ and $R\Gamma^\cdot(X_n^\cdot/E_n)$ similarly. Then, by generalizing the argument in [H-K], we can define φ and N satisfying $N\varphi = p\varphi N$ on the first and the last complexes and show, with the notation of [H-K](4.11), (4.12), that φ on $\mathbb{Q} \otimes \{R\Gamma^\cdot(Y^\cdot/\underline{W}_n)\}_n$ is an automorphism and that there exists a unique section compatible with φ of the pull-back by $\{i_{n,0}\}$ (3.1) :

$$(6.3.1)\qquad \mathbb{Q} \otimes \{R_{E_n} \otimes_{W_n} \{R\Gamma^\cdot(Y^\cdot/\underline{W}_n)\}_n \xrightarrow{\sim} \mathbb{Q} \otimes \{R\Gamma^\cdot(X_n^\cdot/E_n)\}_n$$

(cf. Theorem 3.2.1). (We can avoid to use an embedding system in the proof of [H-K] (2.24) by generalizing [Br3] (2.2.1.1), (2.2.1.2) and using the argument of [T2] 3.8.)

Define $\tilde{H}^q(Y^\cdot/\underline{W}_n)$ to be $H^q(R^+\varepsilon_{\cdot *}R\Gamma^\cdot(Y^\cdot/\underline{W}_n))$. Then, from (6.2.1), we obtain a spectral sequence converging to this cohomology such that $E_1^{a,b} = H^b(Y^a/\underline{W}_n)$ if $0 \le a \le m$ and 0 otherwise, which we denote by $\tilde{E}(Y^\cdot/\underline{W}_n)$. We define the cohomology $\tilde{H}^q$ and the spectral sequence $\tilde{E}$ in the same way for the other two complexes. Then φ and N and the isomorphism (6.3.1) for the complexes on $\underline{\Gamma}((\mathrm{Sets}) \times \Delta[m])$ induce those for the corresponding spectral sequences.

We define $\tilde{H}^q(Y^\cdot/\underline{W})_{\mathbb{Q}}$, $\tilde{E}(Y^\cdot/\underline{W})_{\mathbb{Q}}$ by taking $\mathbb{Q} \otimes \varprojlim_n$. Note that all terms of the spectral sequences $\tilde{E}(Y^\cdot/\underline{W}_n)$ are finitely generated over W_n. Then, every

term of $\tilde{E}(Y^{\cdot}/\underline{W})_{\mathbb{Q}}$ is a (φ, N)-module of finite dimension. We define $\tilde{H}^q(X^{\cdot}/S)_{\mathbb{Q}}$ and $\tilde{E}(X^{\cdot}/S)_{\mathbb{Q}}$ in the same way. Then these are isomorphic to $\tilde{H}^q(X^{\cdot}_K, \Omega^{\cdot}_{X^{\cdot}_K})$ and $\tilde{E}(X^{\cdot}_K, \Omega^{\cdot}_{X^{\cdot}_K})$ (defined similarly as above) and hence $E_1^{a,b}$ and E_∞^c of the spectral sequence have the Hodge filtrations induced by $\sigma_{\geq i}\Omega^{\cdot}_{X^{\cdot}_K}$. We endow $E_r^{a,b}$ $(r \geq 2)$ with the filtration induced by that on $E_1^{a,b}$ as a sub-quotient. As in Theorem 3.2.1, (6.3.1) and the pull-back by $\{i_{n,\pi}\}$ (3.1) induce an isomorphism

$$\rho_\pi \colon K \otimes_{K_0} \tilde{E}(Y^{\cdot}/\underline{W})_{\mathbb{Q}} \xrightarrow{\sim} \tilde{E}(X^{\cdot}_K, \Omega^{\cdot}_{X^{\cdot}_K}). \tag{6.3.2}$$

Especially, each term of $\tilde{E}(Y^{\cdot}/\underline{W})_{\mathbb{Q}}$ has a natural filtered (φ, N)-module structure.

(6.4) For an m-truncated simplicial K-scheme $X^{\cdot}$, we define $R\Gamma_{\text{ét}}(X^{\cdot}, \mathbb{Z}/p^n\mathbb{Z}(r))$, $\tilde{H}^q_{\text{ét}}(X^{\cdot}, \mathbb{Z}/p^n\mathbb{Z}(r))$ and $\tilde{E}_{\text{ét}}(X^{\cdot}, \mathbb{Z}/p^n\mathbb{Z}(r))$ in the same way as (6.3). If $X^{\cdot} \to X$ is an m-truncated proper hypercovering, by cohomological descent ([SD] (3.3.3), (4.3.2)) and the remark in the end of (6.2), we have an isomorphism for $q \leq m-1$:

$$H^q_{\text{ét}}(X, \mathbb{Z}/p^n\mathbb{Z}(r)) \cong H^q_{\text{ét}}(\mathrm{cosk}_m(X^{\cdot}), \mathbb{Z}/p^n\mathbb{Z}(r)) \cong \tilde{H}^q_{\text{ét}}(X^{\cdot}, \mathbb{Z}/p^n\mathbb{Z}(r)). \tag{6.4.1}$$

§7. p-adic Hodge theory for truncated simplicial proper semi-stable schemes.

(7.1) Let $X^{\cdot} \to S$ be an m-truncated simplicial fine log formal scheme over S such that X^i $(i \leq m)$ is isomorphic to the finite base change of a proper semi-stable scheme endowed with the log structure defined by the special fiber. The result of (6.3) is applicable to $X^{\cdot}$. Set $\tilde{H}^q_{\text{ét}}(X^{\cdot}_{\overline{K}}, \mathbb{Q}_p) := \mathbb{Q}_p \otimes_{\mathbb{Z}_p} \varprojlim_n \tilde{H}^q_{\text{ét}}(X^{\cdot}_{\overline{K}}, \mathbb{Z}/p^n\mathbb{Z})$. In this section, we give an outline of the proof of the following theorem.

THEOREM 7.1.1. *The étale cohomology $\tilde{H}^q_{\text{ét}}(X^{\cdot}_{\overline{K}}, \mathbb{Q}_p)$ with the action of G_K is semi-stable and there exists a canonical isomorphism of filtered (φ, N)-modules: $D_{\text{st}}(\tilde{H}^q(X^{\cdot}_{\overline{K}}, \mathbb{Q}_p)) \cong \tilde{H}^q(Y^{\cdot}/\underline{W})_{\mathbb{Q}}$.*

(7.2) We can define the simplicial analogue $\tilde{H}^q(\overline{X}^{\cdot}, S_n^r)$, $\tilde{H}^q(\overline{X}^{\cdot}, S^r_{\mathbb{Q}_p})$ of $H^q(\overline{X}, S_n^r)$ and $H^q(\overline{X}, S^r_{\mathbb{Q}_p})$, and a spectral sequence $\tilde{E}(\overline{X}^{\cdot}, S_n^r)$ converging to $\tilde{H}^q(\overline{X}^{\cdot}, S_n^r)$ in the same way as (6.3). Here we use (6.1). Let d be the maximum of $\dim(X^i_K)$ $(0 \leq i \leq m)$. Then, for $r \geq 2d$, by §5, we have a canonical morphism $\tilde{E}(\overline{X}^{\cdot}, S_n^r) \to \tilde{E}_{\text{ét}}(X^{\cdot}_{\overline{K}}, \mathbb{Z}/p^n\mathbb{Z}(r)')$ whose kernel and cokernel are killed by a p^N independent of $n \geq 1$. Hence $\tilde{E}(\overline{X}^{\cdot}, S^r_{\mathbb{Q}_p}) := \mathbb{Q} \otimes \varprojlim_n \tilde{E}(\overline{X}^{\cdot}, S_n^r)$ is well-defined and $\tilde{E}(\overline{X}^{\cdot}, S^r_{\mathbb{Q}_p}) \cong \tilde{E}_{\text{ét}}(X^{\cdot}_{\overline{K}}, \mathbb{Q}_p(r))$.

(7.3) On the other hand, similarly as [K], [T1] §4, we have natural morphisms of spectral sequences

$$\tilde{E}(\overline{X}^{\cdot}, S_n^r) \longrightarrow \tilde{E}(\overline{X}^{\cdot}_n/W_n) \longrightarrow \tilde{E}(\overline{X}^{\cdot}_n/E_n) \cong \Gamma(\overline{S}_n/E_n) \otimes_{R_{E_n}} \tilde{E}(X^{\cdot}_n/E_n)$$

and, using (6.3.1) and [K] (3.7), we obtain a morphism $\tilde{E}(\overline{X}^{\cdot}, S^r_{\mathbb{Q}_p}) \to B^+_{\text{st}} \otimes_{K_0} \tilde{E}(Y^{\cdot}/\underline{W})_{\mathbb{Q}}$ whose image is contained in the part where $N = 0$ and $\varphi = p^r$.

(7.4) Using $\mathbb{Q}_p(-r) \hookrightarrow B_{\mathrm{st}}$, we obtain from (7.2) and (7.3) a morphism of spectral sequences

$$B_{\mathrm{st}} \otimes_{\mathbb{Q}_p} \tilde{E}_{\text{ét}}(X_{\overline{K}}, \mathbb{Q}_p) \longrightarrow B_{\mathrm{st}} \otimes_{K_0} \tilde{E}(Y^{\cdot}/\underline{W})_{\mathbb{Q}} \tag{7.4.1}$$

compatible with the action of G_K, φ and N. The E_1-term of (7.4.1) is nothing but the comparison map of C_{st} obtained by using (5.2.4) ([K] §6, [T1] §4). Hence (7.4.1) is an isomorphism. Now it remains to prove that this induces a filtered isomorphism after $B_{\mathrm{dR}} \otimes_{B_{\mathrm{st}}}$.

(7.5) A morphism between admissible filtered (φ, N)-modules is strictly compatible with the filtrations and the kernel and the cokernel are again admissible. Hence, by using [D1] (1.3.13), (1.3.16) and the fact that the $E_1^{a,b}$-term of $\tilde{E}(Y^{\cdot}/\underline{W})_{\mathbb{Q}}$ is admissible, we see, by induction on r, that $E_r^{a,b}$ are admissible and $d_r^{a,b}$ are strictly compatible with the filtrations. By [D1] (1.3.17), the filtration on $E_\infty^{a,b}$ coincides with the filtration induced by that on $\tilde{H}^{a+b}(Y^{\cdot}/\underline{W})_{\mathbb{Q}}$, and, by [D2] (7.2.8), the Hodge spectral sequence for $\tilde{H}^*(X_K^{\cdot}, \Omega_{X_K^{\cdot}}^{\cdot})$ degenerates. From the last fact, we obtain an isomorphism

$$\tilde{H}^q(\overline{X}^{\cdot}/S, J^{[r]})_{\mathbb{Q}} \cong Fil^r(B_{\mathrm{dR}}^+ \otimes_K \tilde{H}^q(X_K^{\cdot}, \Omega_{X_K^{\cdot}}^{\cdot}))) \tag{7.5.1}$$

in the same way as [T1] 4.7. This implies that the E_∞^c-term of (7.4.1) is compatible with the filtrations (cf. [T1] 4.8.5) after $B_{\mathrm{dR}} \otimes_{B_{\mathrm{st}}}$. Now the claim follows from the fact that (7.4.1) is a filtered isomorphism in the E_1-term after $B_{\mathrm{dR}} \otimes_{B_{\mathrm{st}}}$.

Acknowledgment: I would like to thank Professor A. J. de Jong who asked me whether one can prove the potential semi-stability of the p-adic étale cohomology of a singular variety by using his alteration. That is the starting point of the study in the simplicial case. I also would like to thank Professor B. Edixhoven who told me the paper [Na].

References

[A] Artin, M., *Grothendieck Topologies*, (Notes on a Seminar by M. Artin 1962), Harvard University.

[Bl-K] Bloch, S. and Kato, K., *p-adic etale cohomology*, Publ. Math. IHES **63** (1986), 107–152.

[Br1] Breuil, C., *Topologie log-syntomique, cohomologie log-cristalline et cohomologie de Čech*, Bull. Soc. math. France **124** (1996), 587–647.

[Br2] Breuil, C., *Construction de représentations p-adiques semi-stables*, Ann. Scient. E. N. S. **31** (1998), 281–327.

[Br3] Breuil, C., *Cohomologie étale de p-torsion et cohomologie cristalline en réduction semi-stable*, to appear in Duke Math. J.

[dJ] de Jong, A. J., *Smoothness, semi-stability and alterations*, Publ. Math. IHES **83** (1996), 51–93.

[D1] Deligne, P., *Théorie de Hodge, II*, Publ. Math. IHES **40** (1971), 5–58.

[D2] Deligne, P., *Théorie de Hodge, III*, Publ. Math. IHES **44** (1975), 5–77.

[Fa1] Faltings, G., *p-adic Hodge theory*, Journal of the AMS **1** (1988), 255–299.

[Fa2] Faltings, G., *Crystalline cohomology and p-adic Galois representations*, Algebraic Analysis, Geometry and Number Theory, Johns Hopkins Univ. Press, Baltimore, 1989, pp. 25–80.

[Fa3] Faltings, G., *Crystalline cohomology of semi-stable curves, and p-adic Galois representations*, Journal of Algebraic Geometry **1** (1992), 61–82.

[Fa4] Faltings, G., *Almost étale extensions*, preprint, MPI Bonn 1998.

[Fo1] Fontaine, J.-M., *Sur certains types de représentations p-adiques du groupe de Galois d'un corps local; construction d'un anneau de Barsotti-Tate*, Ann. of Math. **115** (1982), 529–577.

[Fo2] Fontaine, J.-M., *Cohomologie de de Rham, cohomologie cristalline et représentations p-adiques*, Algebraic Geometry, Lecture Notes in Math. 1016, Springer, 1983, pp. 86–108.

[Fo3] Fontaine, J.-M., *Le corps des périodes p-adiques,* Périodes p-adiques, Séminaire de Bures 1988, Astérisque **223** (1994), 59–111.

[Fo4] Fontaine, J.-M., *Représentations p-adiques semi-stables,* Périodes p-adiques, Séminaire de Bures 1988, Astérisque **223** (1994), 113–183.

[Fo-L] Fontaine, J.-M. and Laffaille, G., *Construction de représentations p-adiques*, Ann. Scient. E. N. S. **15** (1982), 547–608.

[Fo-M] Fontaine, J.-M. and Messing, W., *p-adic periods and p-adic etale cohomology*, Contemporary Math. **67** (1987), 179–207.

[H] Hyodo, O., *A note on p-adic etale cohomology in the semi-stable reduction case*, Inv. Math. **91** (1988), 543–557.

[H-K] Hyodo, O. and Kato, K., *Semi-stable reduction and crystalline cohomology with logarithmic poles,* Périodes p-adiques, Séminaire de Bures 1988, Astérisque **223** (1994), 221–268.

[K] Kato, K., *Semi-stable reduction and p-adic etale cohomology,* Périodes p-adiques, Séminaire de Bures 1988, Astérisque **223** (1994), 269–293.

[K-M] Kato, K. and Messing, W., *Syntomic cohomology and p-adic étale cohomology*, Tôhoku Math. J. **44** (1992), 1–9.

[Na] Nagata, M., *A generalization of the imbedding problem*, J. Math. Kyoto **3** (1963), 89–102.

[Ni] Niziol, W., *Crystalline conjecture via K-theory*, preprint.

[SD] Saint-Donat, B., *Théorie des topos et cohomologie étale des schémas* (*SGA4*) *Exposé* V^{bis}, Lecture Notes in Math. 270, Springer, 1972.

[T1] Tsuji, T., *p-adic étale cohomology and crystalline cohomology in the semi-stable reduction case*, to appear in Inv. Math.

[T2] Tsuji, T., *Frobenius, the Hodge filtration and the syntomic site*, preprint 1997.

Research Institute
for Mathematical Sciences
Kyoto University
Kyoto 606-8502
Japan
e-mail: tsuji@kurims.kyoto-u.ac.jp

Small Points and Arakelov Theory

Shou-Wu Zhang

Abstract. In this talk, I will explain the recent applications of Arakelov theory to the Bogomolov cobjecture.

1991 Mathematics Subject Classification: 11, 14
Keywords and Phrases: Néron Tate heights, Bogomolov conjecture, Arakelov theory

Height of a solution (resp. point) of a diophantine system (resp. variety) will measure the complexity of the solution (resp. point). For an abelian variety, one can define heights for its algebraic points to respect its group structure. For example, the torsion points will be only those who have zero heights. Such a normalization is called Néron-Tate height. The solutions with big heights, zero heights, or near zero heights are all interesting and important in Diophantine geometry. The Arakelov theory, an intersection theory on arithmetic varieties, has played an important role in the study of the Néron-Tate heights in recent years. In this talk, I will explain the recent applications of Arakelov theory to the Bogomolov conjecture on small points. For more details about the proof of this conjecture using Arakelov theory, one should see [Ab, U, Zh5]. (For other recent developments in Arakelov theory, see [So].)

1. Neron-Tate heights and Bogomolov conjecture

Let $\bar{\mathbb{Q}}$ denote the algebraic closure of $\mathbb{Q}$ in $\mathbb{C}$. For each place $p = \infty, 2, 3, 5, \cdots$, let $|\ |_p$ denote a p-adic norm over $\bar{\mathbb{Q}}$ normalized by $|p|_p = 1/p$ if $p \neq \infty$ and $|\ |_\infty$ is the usual absolute value on $\mathbb{C}$. For a point $x = (x_0, \cdots, x_n) \in \mathbb{P}^n(\bar{\mathbb{Q}})$, the naive height $h_{naive}(x)$ of x is defined by

$$h_{naive}(x) = \frac{1}{[K:\mathbb{Q}]} \sum_{\sigma: K \to \bar{\mathbb{Q}}} \sum_p \log \max(|\sigma(x_0)|_p, \cdots, |\sigma(x_n)|_p)$$

where K is a number field in $\mathbb{C}$ containing x_i, and σ are embeddings from K into $\bar{\mathbb{Q}}$, and p are places of $\mathbb{Q}$. If x is a rational point represented by an $(n+1)$-tuple of integers $(x_0, \cdots, x_n)$ with no common divisor, then $h_{naive}(x)$ is $\log \max(|x_0|, \cdots, |x_n|)$. If we define the complexity $c(x)$ of x as the maximum of numbers of digits of x_i which measures the time spent to write a number down, then $h_{naive}(x) - c(x)/\log 10$ is bounded on the set of rational points of $\mathbb{P}^n$. A basic property of h_{naive} is the following Northcott Theorem: for any given number D and H, the set of points in A with height $\leq H$ and degree $\leq D$ is finite.

Let $A \to \mathbb{P}^n$ be an abelian variety embedded in $\mathbb{P}^n$ defined over $\bar{\mathbb{Q}}$. Assume that the embedding is symmetric; this means that there is an automorphism ϕ of $\mathbb{P}^n$ such that $\phi(A) = A$ and $\phi|_A = [-1]_A$. The Neron-Tate height $\hat{h}(x)$ of a point x in $A(\bar{\mathbb{Q}})$ is defined by the formula:

$$\hat{h}(x) = \lim_{m\to\infty} \frac{h_{naive}(mx)}{m^2}.$$

There are two properties of $\hat{h}$ besides Northcott's Theorem:
1. $\hat{h}(x) \geq 0$, and $\hat{h}(x) = 0$ if and only if x is a torsion point;
2. the induced function $\hat{h}$ on the $\mathbb{Q}$-vector space $A(\bar{\mathbb{Q}})/A_{tor}$ is quadratic.

THEOREM A (BOGOMOLOV CONJECTURE). *Let X be an irreducible, closed subvariety of an abelian variety A defined over $\bar{\mathbb{Q}}$. Let $h : A(\bar{\mathbb{Q}}) \to \mathbb{R}$ be a Néron-Tate height function (with respect to a symmetric projective embedding of A). Assume that X is not a translation of an abelian subvariety by a torsion point. Then there is a positive number ϵ such that the set*

$$\{x \in X(\bar{\mathbb{Q}}) : \quad h(x) < \epsilon\}$$

is not Zariski dense in X.

REMARKS:
1. As the above set contains all torsion points in X, the above theorem implies a theorem of Raynaud [Ra] on Lang's conjecture.
2. The original Bogomolov conjecture stated in [Bo] p. 70 has the following form: Assume A is the Jacobian of a curve C of genus ≥ 2. For each $x \in C$ define an embedding $C \to A$ by sending p to the class of $x - p$. Let $r(x)$ denote the maximal number r such that there are only finitely many $p \in C$ such that $x - p$ has height less than r. Define $R(C)$ as the infimum of $r(x)$. Then he conjectured that $R(C) > 0$ and that $R(C)$ should be a certain height function on the moduli space of curves C. Theorem A implies that $r(x) > 0$ in general and that $R(C) > 0$ if the subvariety

$$X = \{x - y \in A : \quad x, y \in C\}$$

does not contain any translations of elliptic curves by torsion points. In the next section, we will explain a variant of our thorem which will implies that $R(C) > 0$ and that $R(C)$ is certainly a height function on the moduli space of C's.

2. ARITHMETIC AMPLENESS AND THE THEOREM OF SUCCESSIVE MINIMA

Using Arakelov theory [A, F1, GS1, Zh4], we can express the Neron-Tate heights as the degrees of hermitian lines on arithmetic curves. We illustrate the idea in the case that A is an abelian variety defined over a number field F and can be extended to an abelian scheme $\pi : \mathcal{A} \to B = \mathrm{Spec}\mathcal{O}_F$. Let $\mathcal{L}$ be a line bundle on $\mathcal{A}$ which extends the restriction of $\mathcal{O}(1)$ on A. Replacing $\mathcal{L}$ by $\mathcal{L} \otimes [-1]^*\mathcal{L}$, we may

assume that $\mathcal{L}$ is symmetric: $[-1]^*\mathcal{L} = \mathcal{L}$. Also replacing $\mathcal{L}$ by $\mathcal{L} \otimes \pi^* e^*(\mathcal{L}^{-1})$ we may fix a rigidification $r : e^*\mathcal{L} \simeq \mathcal{O}_e$ where e denote the unit section of $\mathcal{A}$. On the complex bundle $L(\mathbb{C})$ over $A(\mathbb{C})$ we choose a hermitian metric $\| \ \|$ such that

1. The curvature of $\| \ \|$ is an invariant form on $A(\mathbb{C})$;
2. The map $r_\sigma : e_\sigma^* L(\mathbb{C}) \to \mathbb{C}$ is isometric for each archimedean place σ.

Let $\bar{\mathcal{L}}$ denote the pair $(\mathcal{L}, \| \ \|)$. Let x be a point in $A(\bar{\mathbb{Q}})$. Then the Zariski closure $\bar{x}$ of x has a normalizaton $f : \mathrm{Spec}\mathcal{O}_K \to \mathcal{A}$ where $\mathcal{O}_K$ is the ring of integers of some number field K. The $\mathcal{O}_K$ invertible module $\mathcal{N} := f^*\mathcal{L}$ is equipped with hermitian metric on $\mathcal{N} \otimes_\sigma \mathbb{C}$ for each embedding $\sigma : K \to \mathbb{C}$. Then we define the degree of $\bar{L}$ on $\bar{x}$ by

$$\deg_{\bar{\mathcal{L}}} \bar{x} = \log \frac{\#\mathcal{N}/n\mathcal{O}_K}{\prod_{\sigma K \to \mathbb{C}} \|n \otimes_\sigma 1\|}$$

where n is any nonzero element in $\mathcal{N}$. One can show that

$$\hat{h}(x) = \frac{1}{[K : \mathbb{Q}]} \deg_{\bar{\mathcal{L}}}(\bar{x}).$$

One immediate advantage of Arakelov theory is to extend the definition of the degree linearly to arbitrary cycles Z of A by dimension induction:

1. if Z is a closed point of $\mathcal{A}$ then $\deg_{\bar{\mathcal{L}}}(Z) = \log \#\kappa(Z)$ where $\kappa(Z)$ is the residue field of Z;
2. if Z is a closed subvariety of $\mathcal{A}$, then

$$\deg_{\bar{\mathcal{L}}}(Z) = \deg_{\bar{\mathcal{L}}}(\mathrm{div}\ell|_Z) - \int_{Z(\mathbb{C})} \log \|\ell\| c_1(\bar{\mathcal{L}}_{\mathbb{C}})^{\dim Z}$$

where ℓ is a section of $\mathcal{L}$ which is nonzero on Z and $c_1(\bar{\mathcal{L}}_{\mathbb{C}})$ is the curvature form of $\bar{\mathcal{L}}$ which at any point where $\ell \neq 0$ can be given by

$$c_1(\bar{\mathcal{L}}_{\mathbb{C}}) = \frac{\partial\bar{\partial}}{\pi i} \log \|\ell\|.$$

If X is a closed subvariety of A, then the (Néron-Tate) height $\hat{h}(X)$ is defined by the formula

$$\hat{h}(X) = \frac{\deg_{\bar{\mathcal{L}}}(\mathcal{X})}{(\dim X + 1)\deg_{\mathcal{L}}(X)}$$

where $\mathcal{X}$ is the Zariski closure of X in $\mathcal{A}$. As for the Neron-Tate heights for points, $\hat{h}(X)$ in general will be nonnegative.

As an example, let us consider the case that X is a curve over a number field F of genus $g \geq 2$ with a smooth model $\mathcal{X}$ over $B = \mathrm{Spec}\mathcal{O}_F$ and that A is the Jacobian of X, and that the embedding is $\phi_D : x \to \mathrm{class}(x - D)$, where D is a divisor of degree 1. Then one can show that

$$\hat{h}(X) = \frac{1}{8(g-1)[F : \mathbb{Q}]} c_1(\Omega^1_{\mathcal{X}/B})^2 + \left(1 - \frac{1}{g}\right) \hat{h}(x - D)$$

where $\Omega^1_{\mathcal{X}/B}$ is equipped with the Arakelov metric. In this case, $\hat{h}(X) \geq 0$ as $c_1(\Omega^1_{\mathcal{X}/B})^2 \geq 0$ is proved by Faltings.

Now assume that $(2g-2)D - c_1(\Omega^1_X)$ is a torsion point in A. The first breakthrough step which brings the Bogomolov conjecture into the context of Arakelov theory is the following observation of L. Szpiro [Sz 1-3]: if $c_1(\Omega^1_{\mathcal{X}/B})^2 > 0$, then one can deduce from Faltings' Riemann-Roch theorem [F1] that some positive power of $\Omega^1_{\mathcal{X}/B}$ will have a section ℓ with norm $\|\ell\|_{\sup} < 1$. Using this section to compute the height then one obtains that the points of $X(\bar{\mathbb{Q}})\backslash|\mathrm{div}\ell|$ will have height bigger than $-\log\|\ell\|_{\sup}$. Szpiro also noticed that the truth of the Bogomolov conjecture would imply the positivity $c_1(\Omega^1_{\mathcal{X}/B})^2 > 0$, if one had a Nakai-Moishezon type criterion for ampleness of $\Omega^1_{\mathcal{X}/B}$. In [K1-2], M. Kim proved a Nakai-Moishezon type result and deduced the equivalence of Bogomolov conjecture and $c_1(\Omega^1_{\mathcal{X}/B})^2 > 0$ in this case.

In [Z1, 3], using the arithmetic Hilbert-Samuel formula of Gillet and Soulé [GS 2-3, AB], a general Nakai-Moishezon type theorem for arithmetic variety has been proved. One immediate consequence is the following relation between $\hat{h}(X)$ and the heights of points:

THEOREM OF SUCCESSIVE MINIMA.

$$\frac{1}{\dim X+1}\sum_{i=1}^{\dim X+1} e_i(X) \leq \hat{h}(X) \leq e_1(X)$$

where

$$e_i = \sup_{\substack{Y\subset X\\ \mathrm{codim}Y=i}} \inf_{x\in X\backslash Y} \hat{h}(x).$$

It follows that $\hat{h}(X) = 0$ if and only if $e_1(X) = 0$, or, the set of small points is dense. In particular, if X is the translate $T + x$ of an abelian subvariety T by a torsion point x then $\hat{h}(X) = 0$. Assuming that X is not the translate of an abelian subvariety by a torsion point, then following three are equivalent:

1. the Bogomolov conjecture for X;
2. $e_1(X) > 0$;
3. $\hat{h}(X) > 0$.

One immediate consequence is the Bogomolov conjecture for the embedding $\phi_D : X \to A$ defined by a divisor D such that the class $\Omega^1_X - (2g-2)D$ is not torsion. Going back to Bogomolov's original conjecture, we have

$$\kappa_1 c_1(\Omega^1_{\mathcal{X}/B})^2 \leq R(X) \leq \kappa_2 c_1(\Omega^1_{\mathcal{X}/B})^2$$

where κ_1, κ_2 are two positive constants. All these results can be generalized to the case where X and A may have bad reduction by introducing adelic metrics, and admissible relative sheaf ω_a. See [Zh 2, 4] for more details.

3. Equidistribution theorems

The question of wether $c_1(\Omega_{\mathcal{X}/B})^1 > 0$ was very challenging in Arakelov theory because in the geometric case it is proved by deformation theory and it measures how far it differs from the constant fibration. The first example with $c_1(\Omega_{\mathcal{X}/B})^2 > 0$ when X has good reduction is given by J.-F. Burnol [Bu]. He proved this positivity for curves whose Jacobians has complex multiplication by a CM-field of of degree $2g(X)$. He uses two properties of the Weierstrass divisors $\mathcal{W}_d$ ($d \in \mathbb{N}$) for powers of $\Omega^1_{\mathcal{X}/B}$:

1. The set $\mathcal{W}_d(\mathbb{C})$ of Weierstrass points of fixed degree d in $X(\mathbb{C})$ has the uniform probability measure converges to the Arakelov measure on $X(\mathbb{C})$ as d goes to infinity.
2. The Weierstrass divisor $\mathcal{W}_d$ will contain a vertical component whe $d >> 0$, if the Jacobian of X has complex multiplication.

The results of Burnol are generalized in [Zh4] to arbitrary subvariety X of A such that A is generated by $\{x-y : x, y \in X\}$ and that the morphism $\mathrm{NS}(A) \to \mathrm{NS}(X)$ is not injective. In this case, we can prove the Bogomolov conjecture by applying the Faltings' Hodge index theorem. In curves case, this will imply the positivity $c_1(\Omega^1_{\mathcal{X}/B})^2$ when $\mathrm{End}(\mathrm{Jac}(X)) \otimes \mathbb{R}$ is not a division algebra. For example all modular curves of genus ≥ 2 will satisfy this condition.

If $\mathrm{Jac}(X)$ does not have complex multiplication, Burnol's proof implies the following important fact: if $c_1(\Omega^1_{\mathcal{X}/B})^2 = 0$, then Weierstrass points will produce small points whose probability measure converges to the Arakelov measure. This turns to be a general property of small points [SUZ, Zh5] and can be easily deduced from the Theorem of Successive Minima:

Equidistribution Theorem. *Let X be a subvariety of A defined over a number field K and let x_n be a sequence of points X wich converges to the generic point of X and such that $\hat{h}(x_n) \to 0$. Then the uniform probability measure of the Galois orbit $O(x_n)$ tends to the measure $dx := c_1(\bar{\mathcal{L}})^{\dim X}|_X/deg(X)$ in the following sense: for any continuous function f on $X(\mathbb{C})$*

$$\lim_{n\to\infty} \frac{1}{\#O(x_n)} \sum_{y\in O(x_n)} f(y) = \int_{X(\mathbb{C})} f(x)dx.$$

To prove this, one just applies the right side of the inequality of successive minima

$$h_\lambda(X) \leq e_{1,\lambda}(X)$$

where $h_\lambda(X)$ and $e_{1,\lambda}(X)$ are defined in the same way as $\hat{h}(X)$ and $e_1(X)$ but with metric $\| \ \|$ replaced by

$$\| \ \|_\lambda = \| \ \| \exp(\lambda f).$$

The final step for the proof of $c_1(\Omega^1_{\mathcal{X}/B})^2 > 0$ for general curve X is due to E. Ullmo [U]. His marvelous idea is to use the equidistribution theorem twice which will produce two different metrics and therefore produce a contradiction.

His construction is as follows: for the canonical embedding $X \to A = \mathrm{Jac}(X)$, consider the induced map $\phi : X^g \to A$. If $c_1(\Omega^1_{\mathcal{X}/B})^2 = 0$ then X has a sequence $(x_n, n \in \mathbb{N})$ of distinct points such that $\hat{h}(x_n) \to 0$, and the Galois orbits of these points will have probability measures converging to the Arakelov measure. Then one can produce a sequence $(y_n, n \in \mathbb{N})$ of X^g such that

1. y_n converges to the generic point of X^g;
2. $\hat{h}(y_n)$ converges to 0;
3. The set $\{y_n : n \in \mathbb{N}\}$ is invariant under permutation action on X^g.

Then again the Galois orbits of y_n will have probability measures converge to the product of the Arakelov measure on X^g. However, the sequence $\phi(y_n)$ in A satisfies the condition of the Equidistribution Theorem, therefore the corresponding probability measure converges to the Haar measure on A. It follows that the product measure of Arakelov measure on X^g is the pullback of the Haar measure on A. This is impossible: as the map ϕ is non-smooth, the pullback of the Haar measure as a differential form will vanishes along the singular locus of ϕ!

Ullmo's idea is generalized to prove the general Bogomolov conjecture in [Zh5] by a modified Faltings' construction in [F2]: first it is easy to reduce the Bogomolov conjecture to the case that X has the trivial Ueno fiberation:

$$\{x \in A : \quad x + X = X\}$$

is finite. Then for any positive integer m we consider the map:

$$\alpha_m : X^m \to A^{m-1}$$

$$\alpha_m(x_1, \cdots, x_m) = (x_1 - x_2, \cdots, x_{m-1} - x_m).$$

Then one can show that for m large, ϕ_m will induce a birational but not smooth map $X^m \to \alpha_m(X^m)$. Now we apply the equidistribution theorem to maps: $X^m \to A^m$ and $\alpha_m(X^m) \to A^{m-1}$ for sequences of small points $(y_n, n \in \mathbb{N})$ and $(\alpha_m(y_n), n \in \mathbb{N})$. Then we obtain the equality of two forms α and β on X^m induced respectively from the map $X^m \to A^m$ and $X^m \to A^{m-1}$. But this is impossible as β vanishes along the singular locus of α_m.

Combining the Bogomolov conjecture and the equidistribution theorem, one obtains the following stronger statement about small points in probability measure rather than Zariski topology:

THEOREM B. *Let $(x_n, n \in \mathbb{N})$ be a sequence of points in $A(\bar{\mathbb{Q}})$ such that the following conditions are verified:*

1. *There is no subsequence of $(x_n, n \in \mathbb{N})$ contained in a translation of proper abelian subvariety by a torsion point;*
2. *$\lim_{n\to\infty} h(x_n) = 0$.*

Then the probability measures of the Galois orbits of x_n converge to the Haar measure of $A(\mathbb{C})$.

One can use this to show that the set of torsion points on A over the maximal totally real fields is finite; this has been previously proved by Zarhin [Za] by using Faltings' theorem on Tate's conjecture.

REMARKS:

1. There are different approaches to the Bogomolov conjecture other than Arakelov theory. Notably the diophantine approximation method used in David and Philippon [DP] and Bombieri and Zannier [BZ] produce the lower bound for $\hat{h}(X)$ effectively in terms of the degree of X.
2. The Bogomolov conjecture also has analogues for multiplicative groups $\mathbb{G}_m^n$ (or even a dynamical system [Zh4]). For $\mathbb{G}_m^n$, the Bogomolov conjecture is proved in [Zh3] and the equidistribution theorem is proved by Bilu [Bi]. His approach is very original.

4. A CONJECTURE

Let $A \to C$ be a family of abelian varieties over a curve (may be open) over $\bar{\mathbb{Q}}$. Let Λ be a finitely generated torsion free subgroup of $A(C)$. Let $\mathcal{L}$ be a relative ample symmetric line bundle on $\mathcal{A}$ which induces Neron-Tate height pairings $< >$ on $A_x(\bar{\mathbb{Q}})$ for each point $x \in C(\bar{\mathbb{Q}})$. We define the number $h_\Lambda(x)$ for each $x \in C(\bar{\mathbb{Q}})$ by the formula:

$$h_\Lambda(x) = \det(< t_i(x), t_j(x) >)$$

where $\{t_1, t_2, \cdots\}$ is a basis of Λ and $t_i(x)$ is the specialization of t_i in A_x.

CONJECTURE. *Assume that the generic fiber A_η of A over C is geometrically simple and has dimension ≥ 2. Then there is a positive number ϵ such that the set*

$$\{s \in C(\bar{\mathbb{Q}}) : h_\Lambda(s) < \epsilon\}$$

is finite.

REMARKS:

1. If $A = A_0 \times C$ is constant family with fiber A_0 and Λ is generated by the graphs of one embedding $C \to A_0$ then the above conjecture is the Bogomolov conjecture for the embedding $C \to A_0$.
2. If $A = A_0 \times C$ is a constant family, e_1 is the graph of an embedding, but $e_i (i > 1)$ are graphs of constant maps $C \to x_i \in A_0$ whose images a_i generate $A(K)$ modulo torsion, where K is a number field over which $C \to A$ is defined, then the above conjecture implies the Mordell-Lang Conjecture for $e_1(C) \subset A$, as
$$h_\Lambda(x)^{1/2} = \text{distance}(x, \Gamma \otimes_{\mathbb{Z}} \mathbb{R})/\text{volume}(\Gamma)$$
where the distance is taken in $A(\bar{\mathbb{Q}}) \otimes \mathbb{R}$ with respect to the Néron-Tate height, and Γ is the lattice generated by a_i. A related conjecture has been formulated independently by Poonen [P] and proved by him in some special cases.
3. The dimension assumption in the Conjecture is necessary as Poonen showed to me the following argument: if $A \to C$ has relative dimension one and Λ is generated by one section s then $s \cap A[N]$ will have a lot of intersection as $N \to \infty$, unless either $s(\eta)$ is a torsion point at the generic fiber, or $A \to C$ is a constant family and s is a constant section. Of course, the simpleness condition in the conjecture could be removed if we require that for any geometrically

simple component B of A, either $\dim B_\eta \geq 2$ or B is a constant family and $B \cap \Lambda$ consists of constant sections.

4. Besides the case considered as above, the next two special cases are also interesting:
   a . $\phi : A \to C$ is not a constant family and Λ is generated by one section; The conjecture implies that if $\phi(\eta)$ is not torsion, then there are only finitely many points $x \in C$ such that $\phi(x)$ is torsion.
   b. $A = A_0 \times C$ but Λ is generated by the graphs of two embeddings $\phi_i : C \to A_0$. The conjecture implies that if $\phi_1(\eta)$ and $\phi_2(\eta)$ are linearly independent, then there are at most finitely many $x \in C$ such that $\phi_1(x)$ and $\phi_2(x)$ are linearly dependent. A different formulation is as follows: the wedge product $\phi_1 \wedge \phi_2$ defines a map
   $$C(\bar{\mathbb{Q}}) \to \wedge^2 A(\bar{\mathbb{Q}})$$
   $$x \to \phi_1(x) \wedge \phi_2(x).$$
   Then the height h_Λ is induced by the norm on $\wedge^2 A(\bar{\mathbb{Q}})$. So we have a Bogomolov type conjecture for small points in $\wedge^2 A(\bar{\mathbb{Q}})$!
5. The height h_Λ is unlikely the Weil height for some positive line bundle on C, but the Northcott type theorem follows from some works of Silverman [Si] on the specializations of heights in the case that A over C has no fixed part.

REFERENCES

[Ab] A. Abbes, *Hauteurs et Discrétude*, Séminaire Bourbaki, 49éme, 1996-97, No. 825.

[AB] A. Abbes and T. Bouche, *Théorèm de Hilbert-Samuel " arithmétique"*, Ann. Inst. Fourier (Grenoble) **45** (1995), no 2, 375-401.

[Ar] S. Ju. Arakelov, *An intersection theory for divisors on an arithmetic surface,*, Math. USSR Izverstija, Vol. **8**(1974) No 6, 1167-1180.

[Bi] Y. Bilu, *Limit distribution of small points on algebraic tori*, Duke Math. J. **89** (1997), no 3, 465-476.

[Bo] F. A. Bogomolov, *Points of finite order on an abelian variety,*, Math. USSR Izv., **17** , (1981).

[BZ] E. Bombieri and U. Zannier,, *Heights of algebraic points on subvarieties of abelian varieties*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **24** (1997), no., 2, 205–225.

[Bu] J. -F. Burnol,, *Weierstrass points on arithmetic surfaces,*, Invent. Math., **107** (1992).

[DP] S. David and P. Philippon, *Minorations des hauteurs normalisées des sous-variétés abéliennes*, Number theory (Tiruchirapalli, 1996), 333-364, Contemp. Math. 210 Amer. Math. Soc., Providence, RI, 1998.

[F1] G. Faltings, *Calculus on arithmetic surfaces*, Ann. of Math. (2) **119** (1984), no 2 387-347.

[F2] ———, *Diophantine approximation on abelian varieties*, Ann. of Math. (2) **133** (1991), no. 3, 549-576.

[GS1] H. Gillet and C. Soulé, *Arithmetic intersection theory*, Inst. Hautes Études Sci. Publ. Math. No. 72 (1990), 93-174 (1991).

[GS2] ———, *Amplitude arithmétique*, C. R. Acad Sci. Paris Sér. I Math **307** (1988), no 17, 887-890.

[GS3] ———, *An arithmetic Riemann-Roch theorem*, Invent. Math. **110** (1992), no. 3, 473-543.

[K1] M. Kim, *Numerically positive line bundles on an Arakelov variety*, Duke Math. J. **61** (1990), no 3, 823-833.

[K2] ———, *Small points on constant arithmetic surfaces*, Duke Math. J. **61** (1990), no 3, 823-833.

[Po] B. Poonen, *Mordell-Lang plus Bogomolov*, Manuscript 1998.

[Ra] M. Raynaud, *Sous-variétés d'une variété abélienne et points de torsion*, Arithmetic and Geometry 1 (Ed: J. Coates and S. Helgason), Birkauser (1983).

[Si] J. Silverman, *Heights and the specialization map for families of abelian varieties*, J. Reine Angew. math. **342** (1983), 197-211.

[So] C. Soulé, *Hermitian vector bundles on arithmetic varieties*, Algebraic geometry—Santa Cruz 1995, 383-419, Proc. Sympos. Pure Math., 62, Part 1, Amer. Math. Soc., Providence, RI, 1997.

[Sz1] L. Szpiro, *Small points and torsion points*, The Lefschetz centennial conference, Part I (Mexico City, 1984), 251-260, Contemp. Math., 58, Amer. Math. Soc., Providence, R. I., 1986.

[Sz2] ———, *Présentation de la théorie d'Arakelov*, Current trends in arithmetical algebraic geometry (Arcata, Calif., 1985), 279-293, Contemp. Math., 67, Amer. Math. Soc., rovidence, RI, 1987.

[Sz3] ———, *Sur les propriétés numériques du dualisant relatif d'une surface arithmétique*, The Grothendieck Festschrift, Vol. III, 229-246, Progr. Math., 88, Birkhäuser Boston, Boston, MA, 1990.

[SUZ] L. Szpiro, E. Ullmo, and S. Zhang,, *Équirépartition des petits points*, Invent. Math. **127** (1997), no. 2, 337-347.

[U] E. Ullmo,, *Positivité et discrétion des points algébriques des courbes*, Ann. of Math. (2) **147** (1998), no. 1, 167-179.

[Za] Y. I. Zarhin,, *Endomorphisms and torsion of abelian varieties, Duke Math. J.,* **54** *(1987).*

[Zh1] S. Zhang, *Positive line bundles on arithmetic surfaces Ann. of Math. (2)* **136** *(1992), no. 3, 569-587.*

[Zh2] ———, *Admissible pairing on a curve*, Invent. Math., **112**, (1993).

[Zh3] ———, *Positive line bundles on arithmetic varieties*, J. Amer. Math. Soc. **8** (1995), no. 1, 187–221.

[Zh4] ———, *Small points and adelic metrics J. Alg. Geom.* **4** *(1995).*

[Zh5] ———, *Equidistribution of small points on abelian varieties*, Ann. of Math. (2) **147** (1998), no. 1, 159-165.

Shou-Wu Zhang
Department of Mathematics
Columbia University
New York City, New York 10027
U. S. A.

String Theory and Duality

Paul S. Aspinwall

Abstract. String Duality is the statement that one kind of string theory compactified on one space is equivalent in some sense to another string theory compactified on a second space. This draws a connection between two quite different spaces. Mirror symmetry is an example of this. Here we discuss mirror symmetry and another "heterotic/type II" duality which relates vector bundles on a K3 surface to a Calabi–Yau threefold.

1991 Mathematics Subject Classification: 81T30, 14J32, 14J28, 14J60

1 Introduction

Superstring theory does not currently have a complete definition. What we have instead are a set of incomplete definitions each of which fill in some of the unknown aspects of the other partial definitions. Naturally two questions immediately arise given this state of affairs:

1. Is each partial definition consistent with the others?
2. How completely do the partial definitions combine to define string theory?

Neither of these questions has yet to be answered and indeed both questions appear to be quite deep. The first of these concerns the subject of "string duality". Let us list the set of known manifestations of string theory each of which leads to a partial definition:

1. Type I open superstring
2. Type IIA superstring
3. Type IIB superstring
4. $E_8 \times E_8$ heterotic string
5. $\mathrm{Spin}(32)/\mathbb{Z}_2$ heterotic string
6. Eleven-dimensional supergravity (or "M-theory")

The first five of these "theories" describe a string, which is closed in all cases except the first, propagating in flat ten-dimensional Minkowski space $\mathbb{R}^{9,1}$. The last theory is more like that of a membrane propagating in eleven-dimensional Minkowski space $\mathbb{R}^{10,1}$. (Note that many people like to think of string theory as a manifestation of M-theory rather than the other way around.)

Instead of using a completely flat Minkowski space, one may try to "compactify" these string theories by replacing the Minkowski space by $X \times M$, where X is some compact $(10-d)$-dimensional manifold (or $(11-d)$-dimensional in the case of M-theory) and $M \cong \mathbb{R}^{d-1,1}$. So long as all length scales of X are large with respect to any natural length scale intrinsic to the string theory, we can see that $X \times M$ may approximate the original flat Minkowski space. This is called the "large radius limit" of X. One of the most fascinating aspects of string theory is that frequently we may also make sense of compactifications when X is small, or contains a small subspace in some sense. An extreme case of this is when X is singular. In particular, X need not be a manifold in general.

The key ingredient to be able to analyze string theories on general spaces, X, is *supersymmetry*. For our purposes we may simply regard a supersymmetry as a spinor representation of the orthogonal group of the Minkowski space in which the string theory lives. In general a theory may have more than one supersymmetry in which case the letter "N" is commonly used to denote this number. In the above theories the type I and heterotic strings together with M-theory each have $N = 1$ while the type II strings correspond to $N = 2$.

Upon compactification, the value of N will change depending upon the global holonomy of the Levi-Cevita connection of the tangent bundle of X. The new supersymmetries of M are constructed from the components of the old spinor representations of the original Minkowski space which are invariant under this holonomy. We will give some examples of this process shortly.

The general rule is that the more supersymmetry one has, the more tightly constrained the string theory is and the easier it is to analyze away from the large radius limit. Note that this rule really depends upon the total number of components of all the supersymmetries and so a large d has the same effect as a large N (since M has d dimensions and so its spinor representation would have a large dimension).

As well as constraining the string theory so that it may be more easily analyzed, supersymmetry can be regarded as a coarse classification of compactifications. A knowledge of d and N provides a great deal of information about the resulting system. Almost every possibility for d and N is worth at least one long lecture in itself. We will deal with the case $d = 4$ and $N = 2$ about which probably the most is known at this present time.

The principle of duality can now be stated as follows. Given a specific string and its compactification on X can one find another string theory compactified on another space, Y, such that the "physics" in the uncompactified space, M, is isomorphic between the two compactifications? This is important if our first question of this introduction is to be answered. In particular it should always be true for any pair of string theories in our list unless there is a good reason for a "failure" of one of the strings in some sense. We will see an example of this below.

A mathematical analysis of duality requires a precise definition of the physics of a compactification. This is not yet known in generality. What we do know is a set of objects which are determined by the physics, such as moduli spaces, partition functions, correlation functions, BPS soliton spectra etc., which may be compared to find necessary conditions for duality.

The most basic object one may study to identify the physics of two dual theories is their moduli spaces. Roughly speaking this should correspond to the moduli spaces of X and Y although one always requires "extra data" beyond this. It is the extra data which leads to the mathematical richness of the subject. Clearly if two theories are to be identified, one must be able to identify their moduli spaces point by point. This will be the focus of this talk.

It is a pleasure to thank my collaborators R. Donagi and D. Morrison for many useful discussions which were key to the results of section 4.

2 String Data

In order to be able to describe the moduli space of each string theory we are required to give the necessary data which goes into constructing each one. Unfortunately, we do not have anywhere near enough space to describe the origin of what follows. We refer to [1, 12, 19] for more details.

The theories which yield $d = 4$ and $N = 2$ in which we will interested are specified by the following

- The type IIA string is compactified on a Calabi–Yau threefold X (which has SU(3) holonomy). The following data specifies the theory.
  1. A Ricci-flat metric on X.
  2. A B-field $\in H^2(X, \mathbb{R}/\mathbb{Z})$.
  3. A Ramond-Ramond (RR) field $\in H^{\text{odd}}(X, \mathbb{R}/\mathbb{Z})$.
  4. A dilaton+axion, $\Phi \in \mathbb{C}$.
- The type IIB string is compactified on a Calabi–Yau threefold Y (which also has SU(3) holonomy). The following data specifies the theory.
  1. A Ricci-flat metric on Y.
  2. A B-field $\in H^2(Y, \mathbb{R}/\mathbb{Z})$.
  3. A Ramond-Ramond (RR) field $\in H^{\text{even}}(Y, \mathbb{R}/\mathbb{Z})$.
  4. A dilaton+axion, $\Phi \in \mathbb{C}$.
- The $E_8 \times E_8$ heterotic string is compactified on a product of a K3 surface, Z, and an elliptic curve, E_H. This product has SU(2) holonomy. The following data specifies the theory.
  1. A Ricci-flat metric on $Z \times E_H$.
  2. A B-field $\in H^2(Z \times E_H, \mathbb{R}/\mathbb{Z})$.

3. A vector bundle $V \to (Z \times E_H)$ with a connection satisfying the Yang–Mills equations and whose structure group $\subseteq E_8 \times E_8$. The respective characteristic classes in H^4 for V and the tangent bundle of $Z \times E_H$ are fixed to be equal.
4. A dilaton+axion, $\Phi \in \mathbb{C}$.

In each case we can only expect the data to provide a faithful coordinate system in some limit. This is a consequence of the the fact that we only really have a partial definition of each string theory. A sufficient condition for faithfulness is that the target space is large — i.e., all minimal cycles have a large volume, *and* $|\Phi| \gg 1$. Beyond this we may expect "quantum corrections". In general the global structure of the moduli space can be quite incompatible with this parameterization — it is only reliable near some boundary.

On general holonomy arguments (see, for example, [2,10]) one can argue that the moduli space factorizes locally

$$\mathscr{M} \cong \mathscr{M}_H \times \mathscr{M}_V, \tag{1}$$

where (at least at smooth points) $\mathscr{M}_H$ is a quaternionic Kähler manifold and $\mathscr{M}_V$ is a special Kähler manifold. We refer the reader to [16] for the definition of a special Kähler manifold. These restricted holonomy types are expected to remain *exact* after quantum corrections have been taken into account.

We may now organize the above parameters into how they span $\mathscr{M}_H$ and $\mathscr{M}_V$. First we note that Yau's theorem [28] tells us that the Ricci-flat metric on a Calabi–Yau manifold is uniquely determined by a choice of complex structure and by fixing the cohomology class of the Kähler form, $J \in H^2(\bullet, \mathbb{R})$. We may combine J and B to form the "complexified Kähler form" $B + iJ \in H^2(\bullet, \mathbb{C}/\mathbb{Z})$. We then organize as follows

- The Type IIA string: $\mathscr{M}_V$ is parametrized by the complexified Kähler form of X. $H^{\text{odd}}(X, \mathbb{R}/\mathbb{Z}) \cong H^3(X, \mathbb{R}/\mathbb{Z})$ is the *intermediate Jacobian* of X and is thus an abelian variety. We then expect a factorization $\mathscr{M}_H \cong \mathbb{C} \times \mathscr{M}'_H$, where Φ is the coordinate along the $\mathbb{C}$ factor. Finally we have a fibration $\mathscr{M}'_H \to \mathscr{M}_{\text{cx}}(X)$ with fibre given by the intermediate Jacobian, and $\mathscr{M}_{\text{cx}}(X)$ is the moduli space of complex structures on X.

- The Type IIB string: $\mathscr{M}_V$ is now parametrized by the complex structure of Y. $H^{\text{even}}(Y, \mathbb{R}/\mathbb{Z}) \cong H^0(Y, \mathbb{R}/\mathbb{Z}) \oplus H^2(Y, \mathbb{R}/\mathbb{Z}) \oplus H^4(Y, \mathbb{R}/\mathbb{Z}) \oplus H^6(Y, \mathbb{R}/\mathbb{Z})$ may be viewed as an abelian variety. We again expect a factorization $\mathscr{M}_H \cong \mathbb{C} \times \mathscr{M}'_H$, where Φ is the coordinate along the $\mathbb{C}$ factor. Finally we have a fibration $\mathscr{M}'_H \to \mathscr{M}_{\text{Kf}}(Y)$ with fibre given by the RR fields, and $\mathscr{M}_{\text{Kf}}(Y)$ is the moduli space of the complexified Kähler form of Y.

- The $E_8 \times E_8$ heterotic string: Let us first assume that the bundle $V \to (Z \times E_H)$ factorizes as $(V_Z \to Z) \times (V_E \to E_H)$. Thus the structure group of V_Z times the structure group of V_E is a subgroup of $E_8 \times E_8$. We now expect $\mathscr{M}_V$ to factorize as $\mathbb{C} \times \mathscr{M}'_V$, where Φ is the coordinate along the $\mathbb{C}$ factor (see [15] for a more precise statement). $\mathscr{M}'_V$ is then the total moduli

space of $V_E \to E_H$ including deformations of the complex structure and the complexified Kähler form of E_H. $\mathscr{M}_H$ is the total moduli space of the fibration $V_Z \to Z$ including deformations of the Ricci-flat metric of Z.

Again we emphasize that the above statements are approximate and only valid when the target space is large and $|\Phi| \gg 1$. They should be exact only at the boundary of the moduli space corresponding to these limits. It is important to see that factorization of the moduli space will restrict the way that the quantum corrections may act. For example, in the type IIA string the dilaton, Φ, lives in $\mathscr{M}_H$. This means that $\mathscr{M}_V$ cannot be subject to corrections related to having a finite $|\Phi|$. Equally, the Kähler form parameter governs the size of X and so $\mathscr{M}_H$ will not be subject to corrections due to finite size.

It is this property that some parts of the moduli space can be free from quantum corrections and that the interpretation of this part can vary from string theory to string theory which lies at the heart of the power of string duality. If two theories are simultaneously exact at some point in the moduli space then we may address the first question in our introduction. If at every point in the moduli space some theory (perhaps as yet unknown) is in some sense exact then we may address the second question.

3 Mirror Duality

Mirror symmetry as first suggested in [9, 20] was a duality between "conformal field theories". We may make a different version of mirror symmetry, a little more in the spirit of "full" string theories, by proposing the following [4]:

Definition 1 *The pair (X, Y) of Calabi–Yau threefolds is said to be a* mirror *pair if and only if the type IIA string compactified on X is physically equivalent to the type IIB string compactified on Y.*

Of course, this definition is mathematically somewhat unsatisfying as it depends on physics. However, it encompasses previous definitions of mirror symmetry. We also assume the following

Proposition 1 *If (X, Y) is a mirror pair then so is (Y, X).*

While this proposal is obvious from the old definitions it is not completely clear that we may establish it rigorously using the above definition.

Applying this to the moduli space description in the previous section we immediately see that, ignoring quantum corrections, $\mathscr{M}_{\mathrm{Kf}}(X)$ should be identified with $\mathscr{M}_{\mathrm{cx}}(Y)$ and equally $\mathscr{M}_{\mathrm{Kf}}(Y)$ should be identified with $\mathscr{M}_{\mathrm{cx}}(X)$. We know that $\mathscr{M}_V$ is unaffected by Φ corrections and we expect $\mathscr{M}_{\mathrm{cx}}(Y)$ to be *exact* since it is also unaffected by size corrections.

We expect that $\mathscr{M}_{\mathrm{Kf}}(X)$ be affected by size corrections. Similarly, given proposition 1, $\mathscr{M}_{\mathrm{cx}}(X)$ is exact and $\mathscr{M}_{\mathrm{Kf}}(Y)$ will suffer from size corrections. We will use the notation $\mathscr{Q}$ to refer to a fully corrected moduli space. Thus $\mathscr{Q}_{\mathrm{Kf}}(X) \cong \mathscr{Q}_{\mathrm{cx}}(Y) \cong \mathscr{M}_{\mathrm{cx}}(Y)$ but $\mathscr{Q}_{\mathrm{Kf}}(X) \ncong \mathscr{M}_{\mathrm{Kf}}(X)$.

The corrections to $\mathscr{M}_{\mathrm{Kf}}(X)$ take the form of "world-sheet" instantons and were studied in detail in celebrated work of Candelas et al [8]. In particular, the assertion that $\mathscr{Q}_{\mathrm{Kf}}(X) \cong \mathscr{Q}_{\mathrm{cx}}(Y)$ allows one to count the numbers of rational curves on X. Subsequently a great deal of work (see for example [18, 23, 24, 26]) has been done which has made this curve counting much more rigorous.

As well as $\mathscr{M}_{\mathrm{Kf}}$ and $\mathscr{M}_{\mathrm{cx}}$, it is instructive to look at the abelian fibres of $\mathscr{M}_H$ in the context of mirror symmetry. The effect of equating $\mathscr{M}_{\mathrm{cx}}(X)$ with $\mathscr{M}_{\mathrm{Kf}}(Y)$ is to equate

$$H^3(X,\mathbb{Z}) \sim H^0(Y,\mathbb{Z}) \oplus H^2(Y,\mathbb{Z}) \oplus H^4(Y,\mathbb{Z}) \oplus H^6(Y,\mathbb{Z}), \tag{2}$$

but that we expect this correspondence to make sense only if Y is very large. Note that by going around closed loops in $\mathscr{M}_{\mathrm{cx}}(X)$ we expect to have an action on $H^3(X,\mathbb{Z})$ induced by monodromy. If we were to take (2) to be literally true then we have to say the same thing about the action of closed loops in $\mathscr{M}_{\mathrm{Kf}}(X)$ acting on the even integral cycles in Y. That is to say, we would be claiming that if one begins with, say, a point representing an element of $H^0(Y,\mathbb{Z})$ we could smoothly shrink Y down to some small size and then smoothly let it reëxpand in some inequivalent way such that our point had magically transformed itself into, say, a 2-cycle! Clearly this does not happen in classical geometry.

The suggestion therefore [3, 7] is that quantum corrections should be applied to the notion of integral cycles so that, in the context of stringy geometry, 0-cycles *can* turn into 2-cycles when Y is small. Thus the notion of dimensionality must be uncertain for small cycles.

Of central importance to the study of mirror pairs is being in a region of moduli space where the quantum corrections are small. That is we require Y to be large. This amounts to a specification of the Kähler form on Y and must therefore specify some condition on the complex structure of X. This was analyzed by Morrison:

PROPOSITION 2 *If Y is at its large radius limit then X is at a degeneration of complex structure corresponding to maximal unipotent monodromy.*

We refer the reader to [25] for an exact statement of this. The idea is that X degenerates such that a variation of mixed Hodge Structures around this point leads to monodromy compatible with (2).

The point we wish to emphasize here is that when X is very large then the complex structure of Y is restricted to be very near a particular point in $\mathscr{M}_{\mathrm{cx}}(X)$. We only really expect mirror symmetry to be "classically" true at this degeneration. Close to this degeneration we may measure quantum perturbations leading to such effects as counting rational curves. A long way from this degeneration mirror symmetry is much more obscure from the point of view of classical geometry.

It is possible to have a Calabi–Yau threefold, X, whose moduli space $\mathscr{M}_{\mathrm{cx}}(X)$ contains no points of maximal unipotency. In this case, its mirror, Y, can have no large radius limit. Since clearly any classical Calabi–Yau threefold may be taken to be any size, Y cannot have an interpretation as a Calabi–Yau threefold. This is the sense in which duality can sometimes break down.

4 Heterotic/Type IIA Duality

Having discussed mirror duality between the type IIA and the type IIB string we will now try to repeat the above analysis for the duality between the type IIA and the $E_8 \times E_8$ heterotic string. This duality was first suggested in [14, 22] following the key work of [21, 27].

In this case $\mathscr{M}_V$ is currently fairly well-understood (see, for example [2] and references therein). Here we will discuss $\mathscr{M}_H$ which provides a much richer structure.

First let us discuss the quantum corrections. On the heterotic side, $\mathscr{M}_H$ contains the deformations of Z as well as the vector bundle over it. Note that in the the case of K3 surfaces we may not factorize the moduli space of Ricci-flat metrics into a moduli space of complex structures and the Kähler cone. This follows from the fact that given a fixed Ricci-flat metric, we have an S^2 of complex structures. The size of the K3 surface is a parameter of $\mathscr{M}_H$ and so we expect $\mathscr{M}_H$ to suffer from quantum corrections due to size effects for the heterotic string.

We also know that on the type IIA side, the dilaton is contained in $\mathscr{M}_H$. Thus we expect $\mathscr{M}_H$ to suffer from corrections due to Φ for the type IIA string. We managed to evade worrying about such effects in our discussion of mirror symmetry but here we are not so lucky.

Let us now attempt to find the place in the moduli space where we may ignore the quantum effects both due to Φ and due to size. To do this we require the following:

Proposition 3 *If a type IIA string compactified on a Calabi–Yau threefold X is dual to a heterotic string compactified on a factorized bundle over a product of a K3 surface, Z, and an elliptic curve E_H, then X must be in the form of an elliptic fibration $\pi_F : X \to \Sigma$ with a section and a K3 fibration $\pi_A : X \to B$. Here Σ is a birationally ruled surface and $B \cong \mathbb{P}^1$.*

Note that these fibrations may contain degenerate fibres. We refer to [2] for details.

Let us now assume that Z is in the form of an elliptic fibration over B with a section. Given this, we claim the following:

Proposition 4 *The limit of large Z automatically ensures that $\Phi \to \infty$ for the type IIA string. In this limit, X also undergoes a degeneration to $X_1 \cup_{Z_*} X_2$, where X_1 and X_2 are each elliptic fibrations over a birationally ruled surface and are each fibrations over $B \cong \mathbb{P}^1$ with generic fibre given by a rational elliptic surface (RES). $Z_* = X_1 \cap X_2$ is isomorphic to Z as a complex variety.*

We refer to [6, 17] for a proof.

Recall that a RES is a complex surface given by $\mathbb{P}^2$ blown up at nine points given by the intersection of two cubic curves. In a sense, for elliptic fibrations a RES is "half of a K3 surface". This degeneration is viewed as each K3 fibre of the fibration $\pi_A : X \to B$ breaking up into two RES's.

This degeneration therefore provides the analogue of the "maximally unipotent" degeneration in the case of mirror symmetry. There are important differences

however. Note that while the maximally unipotent degeneration of mirror symmetry essentially corresponds to a point in the moduli space of complex structures, the degeneration given by proposition 4 is not rigid — it corresponds a family of dual theories. In the case of mirror symmetry, by taking Y to be large we needed to fix a point in $\mathscr{M}_{\mathrm{Kf}}(Y)$ and thus $\mathscr{M}_{\mathrm{cx}}(X)$. Here we need to take the K3 surface Z to its large radius limit but this does *not* fix a point in $\mathscr{M}_H$. We may still vary the complex structure of Z (subject only to the constraint that it be an elliptic fibration with a section) and we may still vary the bundle V_Z.

We should therefore be able to see the moduli space of complex structures on the elliptic K3 surface, Z, and the moduli space of the vector bundle V_Z *exactly* from this degeneration of X. The correspondence $Z \cong Z_* = X_1 \cap X_2$ tells us how the moduli space of Z can be seen from the moduli space of the degenerated X. The moduli space of the vector bundle is a little more interesting.

V_Z may be split into a sum of two bundles $V_{Z,1}$ and $V_{Z,2}$ each of which has a structure group $\subseteq E_8$. We will identify $V_{Z,1}$ from a curve $C_1 \subset Z_*$ and $V_{Z,2}$ from $C_2 \subset Z_*$. C_1 and C_2 will form the *spectral curves* of their respective bundles in the sense of [13].

Let us consider a single RES fibre Q_b of the fibration $X_1 \to B$, where $b \in B$. Q_b is itself an elliptic fibration $\pi_Q : Q_b \to \mathbb{P}^1$. The section of the elliptic fibration $\pi_F : X \to \Sigma$ determines a distinguished section $\sigma_0 \subset Q_b$. Blowing this down gives a Del Pezzo surface with 240 lines $\sigma_1, \ldots, \sigma_{240}$.

We then have

PROPOSITION 5 *The fibre of the branched cover $C_1 \to B$ is given by the set of points $\{\sigma_i \cap Z_*; i = 1, \ldots, 240\}$,*

with an analogous construction for C_2. We refer to [5, 11] for details.

We also have the data from the abelian fibre of $\mathscr{M}_H$ corresponding to the RR fields. In the case of heterotic/type IIA duality we have [5]

PROPOSITION 6

$$\Lambda_0 \cong H^1(C_1, \mathbb{Z}) \oplus H^1(C_2, \mathbb{Z}) \oplus H^2_T(Z, \mathbb{Z}),$$

where Λ_0 is the sublattice of $H^3(X, \mathbb{Z})$ invariant under monodromy around the degeneration of proposition 4 and $H^2_T(Z, \mathbb{Z})$ is the lattice of transcendental 2-cocycles in Z.

Thus the RR-fields of the type IIA string map to the Jacobians of C_1 and C_2, required to specify the bundle data, and to the B-field on Z.

Proposition 6 should embody much of the spirit of the duality between the type IIA string and the $E_8 \times E_8$ heterotic string in a similar way that equation (2) embodies mirror symmetry. In particular Λ_0 is not invariant under monodromy around any loop in the moduli space and so the notion of what constitutes the E_8-bundles and what constitutes the K3 surface Z should be blurred in general — just as the notion of 0-cycles and 2-cycles is blurred in mirror symmetry.

The analysis of the moduli space $\mathscr{M}_H$ is very much in its infancy. In this talk we have not even mentioned how to compute quantum corrections — the

above discussion was purely for the exact classical limit. There appear to be many adventures yet to be encountered in bringing the understanding of heterotic/type IIA duality to the same level as that of mirror symmetry.

References

[1] P. S. Aspinwall, *The Moduli Space of $N = 2$ Superconformal Field Theories*, in E. Gava et al., editors, "1994 Summer School in High Energy Physics and Cosmology", pages 352–401, World Scientific, 1995, hep-th/9412115.

[2] P. S. Aspinwall, *K3 Surfaces and String Duality*, in C. Esthimiou and B. Greene, editors, "Fields, Strings and Duality, TASI 1996", pages 421–540, World Scientific, 1997, hep-th/9611137.

[3] P. S. Aspinwall and C. A. Lütken, *Quantum Algebraic Geometry of Superstring Compactifications*, Nucl. Phys. B355 (1991) 482–510.

[4] P. S. Aspinwall and D. R. Morrison, *U-Duality and Integral Structures*, Phys. Lett. 355B (1995) 141–149, hep-th/9505025.

[5] P. S. Aspinwall, *Aspects of the Hypermultiplet Moduli Space in String Duality*, J. High Energy Phys. 04 (1998) 019, hep-th/9802194.

[6] P. S. Aspinwall and D. R. Morrison, *Point-like Instantons on K3 Orbifolds*, Nucl. Phys. B503 (1997) 533–564, hep-th/9705104.

[7] P. Candelas and X. C. de la Ossa, *Moduli Space of Calabi–Yau Manifolds*, Nucl. Phys. B355 (1991) 455–481.

[8] P. Candelas, X. C. de la Ossa, P. S. Green, and L. Parkes, *A Pair of Calabi–Yau Manifolds as an Exactly Soluble Superconformal Theory*, Nucl. Phys. B359 (1991) 21–74.

[9] P. Candelas, M. Lynker, and R. Schimmrigk, *Calabi–Yau Manifolds in Weighted $\mathbb{P}_4$*, Nucl. Phys. B341 (1990) 383–402.

[10] S. Cecotti, S. Ferrara, and L. Girardello, *Geometry of Type II Superstrings and the Moduli of Superconformal Field Theories*, Int. J. Mod. Phys. A4 (1989) 2475–2529.

[11] G. Curio and R. Y. Donagi, *Moduli in $N = 1$ Heterotic/F-Theory Duality*, hep-th/9801057.

[12] E. D'Hoker, *String Theory*, in P. Deligne et al., editors, "Quantum Fields and Strings: A Course for Mathematicians", AMS, 1999, to appear, these notes are available from `http://www.cgtp.duke.edu/QFT/spring`.

[13] R. Y. Donagi, *Spectral Covers*, in "Current Topics in Complex Algebraic Geometry", Math. Sci. Res. Inst. Publ. 28, pages 65–86, Berkeley, 1992, alg-geom/9505009.

[14] S. Ferrara, J. Harvey, A. Strominger, and C. Vafa, *Second Quantized Mirror Symmetry*, Phys. Lett. 361B (1995) 59–65, hep-th/9505162.

[15] S. Ferrara and A. Van Proeyen, *A Theorem on N=2 Special Kähler Product Manifolds*, Class. Quant. Grav. 6 (1989) L243–L247.

[16] D. S. Freed, *Special Kähler Manifolds*, hep-th/9712042.

[17] R. Friedman, J. Morgan, and E. Witten, *Vector Bundles and F Theory*, Commun. Math. Phys. 187 (1997) 679–743, hep-th/9701162.

[18] A. B. Givental, *Equivariant Gromov–Witten Invariants*, Internat. Math. Res. Notices 1996 613–663, alg-geom/9603021.

[19] M. Green, J. Schwarz, and E. Witten, *Superstring Theory*, Cambridge University Press, 1987, 2 volumes.

[20] B. R. Greene and M. R. Plesser, *Duality in Calabi–Yau Moduli Space*, Nucl. Phys. B338 (1990) 15–37.

[21] C. Hull and P. Townsend, *Unity of Superstring Dualities*, Nucl. Phys. B438 (1995) 109–137, hep-th/9410167.

[22] S. Kachru and C. Vafa, *Exact Results For N=2 Compactifications of Heterotic Strings*, Nucl. Phys. B450 (1995) 69–89, hep-th/9505105.

[23] M. Kontsevich, *Homological Algebra of Mirror Symmetry*, in "Proceedings of the International Congress of Mathematicians", pages 120–139, Birkhäuser, 1995, alg-geom/9411018.

[24] B. Lian, K. Lu, and S.-T. Yau, *Mirror Principle I*, alg-geom/9712011.

[25] D. R. Morrison, *Mirror Symmetry and Rational Curves on Quintic Threefolds: A Guide For Mathematicians*, J. Amer. Math. Soc. 6 (1993) 223–247, alg-geom/9202004.

[26] Y. Ruan and G. Tian, *Higher Genus Symplectic Invariants and Sigma Model Coupled with Gravity*, Invent. Math. 130 (1997) 455–516, alg-geom/9601005.

[27] E. Witten, *String Theory Dynamics in Various Dimensions*, Nucl. Phys. B443 (1995) 85–126, hep-th/9503124.

[28] S.-T. Yau, *Calabi's Conjecture and Some New Results in Algebraic Geometry*, Proc. Natl. Acad. Sci. 74 (1977) 1798–1799.

Paul S. Aspinwall
Center for Geometry
and Theoretical Physics
Box 90318
Duke University
Durham, NC 27708-0318
USA

Mirror Symmetry and Toric Geometry

Victor V. Batyrev

Abstract. A brief survey of some recent progress towards a mathematical understanding of Mirror Symmetry is given. Using toric geometry, we can express Mirror Symmetry via an elementary duality of special polyhedra.

1991 Mathematics Subject Classification: Primary 14J32, 52B20, 81T60; Secondary 14J45, 33C50, 81T30.
Keywords and Phrases: Toric Varieties, Calabi-Yau Manifolds, Generalized Hypergeometric Functions, Rational Curves.

Introduction

Mirror Symmetry is a remarkable discovery by physicists who suggested that the partition functions of two physical theories obtained from two *different* Calabi-Yau manifolds V and V^* can be *identified* [59]. So far mathematicians couldn't find any appropriate language for a rigorous formulation of this identification (we refer the reader to Kontsevich's talk [50] for the most general conceptual framework that could help to find such a languange). Without knowing a mathematical reason for Mirror Symmetry it simply remains for one to believe in its existence. This belief is supported by many computational experiments followed by attempts to find rigorous mathematical explanations of their results.

In this talk we shall give a brief survey of some recent progress, based on toric geometry, towards a mathematical understanding of Mirror Symmetry. Loosely speaking, toric geometry provides some kind of "Platonic" approach to Mirror Symmetry, because it replaces the highly nontrivial duality between some mathematical objects, which we still don't completely know, by an elementary polar duality of special convex polyhedra. Of course, such a simplification can't reflect the whole nature of Mirror Symmetry, but it helps to form our intuition and find reasonable mathematical tests for this duality.

1 Polar Duality of Reflexive Polyhedra

Let M be a free abelian group of rank d, $N = Hom(M, \mathbf{Z})$ the dual group, and $\langle \cdot, \cdot \rangle : M \times N \to \mathbf{Z}$ the natural nondegenerate pairing. We denote by $M_{\mathbf{R}}$ (resp. by $N_{\mathbf{R}}$) the scalar extension $M \otimes_{\mathbf{Z}} \mathbf{R}$ (resp. $M \otimes_{\mathbf{Z}} \mathbf{R}$).

DEFINITION 1.1 [6] A convex d-dimensional polyhedron $\Delta \subset M_{\mathbf{R}}$ is called *reflexive* if the following conditions are satisfied:

(i) all vertices of Δ belong to the lattice $M \subset M_{\mathbf{R}}$;

(ii) the zero vector $0 \in M$ belongs to the interior of Δ;

(iii) all vertices of the polar polyhedron

$$\Delta^* := \{b \in N_{\mathbf{R}} \ : \ \langle a, b\rangle \geq -1 \ \ \forall a \in \Delta\}$$

belong to the dual lattice $N \subset N_{\mathbf{R}}$.

If $\Delta \subset M_{\mathbf{R}}$ is a reflexive polyhedron, then $\Delta^* \subset N_{\mathbf{R}}$ is again a reflexive polyhedron and $(\Delta^*)^* = \Delta$. So we obtain a natural involution $\Delta \leftrightarrow \Delta^*$ on the set of all d-dimensional reflexive polyhedra. This involution plays a crucial role in our approach to Mirror Symmetry. In the case $d = 3$, the involution $\Delta \leftrightarrow \Delta^*$ provides an interpretation of Arnold's Strange Duality [1, 30, 31, 32, 48].

Toric geometry, or theory of toric varieties, establishes remarkable relations between mathematical objects in convex geometry, e.g. convex cones and polyhedra, and algebraic varieties (see [24, 25, 34, 35, 62]). Toric varieties $\mathbf{P}_\Delta$ associated with reflexive polyhedra Δ are Fano varieties with at worst Gorenstein canonical singularities. Let $T_M = Spec\, \mathbf{C}[M]$ be the algebraic torus with lattice of characters M. Denote by $Z_f \subset T_M$ the affine hypersurface in T_M defined by the equation

$$f(x_1, \ldots, x_d) = \sum_{m \in \Delta \cap M} a_m x^m = 0,$$

where the set $\{a_m\}_{m \in \Delta \cap M}$ consists of generically choosen complex numbers. Then the projective closure of Z_f in $\mathbf{P}_\Delta$ is a normal irreducible variety $\overline{Z}_f$ having trivial canonical class. If we repeat the same procedure with the polar reflexive polyhedron Δ^*, then in the dual torus $T_N := Spec\, \mathbf{C}[N]$ we obtain another affine hypersurface $Z_g \subset T_N$ defined by an equation

$$g(y_1, \ldots, y_d) = \sum_{n \in \Delta^* \cap N} b_n y^n = 0$$

We denote by $\overline{Z}_g \subset \mathbf{P}_{\Delta^*}$ the projective compactification of Z_g in $\mathbf{P}_{\Delta^*}$. The pair $(\overline{Z}_f, \overline{Z}_g)$ is conjectured to be *mirror symmetric* [6]. If $d = 4$, then $\overline{Z}_f$ (reps. $\overline{Z}_g$) is birational to a smooth Calabi-Yau 3-fold $\widehat{Z_f}$ (resp. $\widehat{Z_g}$) and one has the equations

$$h^{1,1}(\widehat{Z_f}) = h^{2,1}(\widehat{Z_g}), \ \ h^{1,1}(\widehat{Z_g}) = h^{2,1}(\widehat{Z_f}),$$

which admit an interpretation by means of a Monomial-Divisor Mirror Map [2]. It is known that the volume of a reflexive polyhedron can be estimated by a constant depending only on d [5]. Consequently there exist only finitely many d-dimensional reflexive polyhedra Δ up to $GL(M)$-isomorphism. Some results towards a classification of reflexive polyhedra of dimension $d \leq 4$ were obtained by Kreuzer and Skarke [68, 53, 54]. It turned out that all examples of Calabi-Yau 3-folds constructed by physicists from hypersurfaces in 7555 different weighted projective spaces can be obtained from 4-dimensional reflexive polyhedra [23].

Moreover, all moduli spaces of Calabi-Yau hypersurfaces in 4-dimensional toric varieties can be connected into a web using a series of simple transformations [3, 4]. The latter confirms a conjecture of M. Reid on the connectedness of the moduli space of Calabi-Yau 3-folds [64].

The polar duality for reflexive polyhedra can be extended to a more general duality for reflexive Gorenstein cones in [11]. This generalization allowed us to express in the same way not only the construction for mirrors of Calabi-Yau complete intersections in Gorenstein toric Fano varieties [20, 56], but also the construction for mirrors of rigid Calabi-Yau 3-folds [22].

2 Topological Mirror Symmetry Test and Stringy Hodge Numbers

If two smooth Calabi-Yau $(d-1)$-folds (V, V^*) form a mirror pair, then the Hodge numbers of V and V^* are related by the equalities

$$h^{p,q}(V) = h^{d-1-p,q}(V^*), \quad 0 \le p, q \le d-1,$$

which are known as a simplest *topological Mirror Symmetry test*. A formulation of this test for projective Calabi-Yau hypersurfaces $\overline{Z}_f \subset \mathbf{P}_\Delta$ and $\overline{Z}_g \subset \mathbf{P}_{\Delta^*}$ turns out to be rather nontrivial, because these hypersurfaces are usually singular. Moreover, we can't expect that a projective *smooth* birational model $\widehat{Z_f}$ of $\overline{Z}_f$ having trivial canonical class always exists if $d \ge 5$. On the other hand, it was observed in [12] that Betti and Hodge numbers of such birational models are uniquely determined. This observation supported the idea of *stringy Hodge numbers* for singular Calabi-Yau varieties which we proposed in [9]. Denote by $E(W; u, v)$ the E-polynomial of a complex quasi-projective variety W. It is defined by the formula

$$E(W; u, v) := \sum_{p,q} e^{p,q}(W) u^p v^q,$$

where $e^{p,q}(W) := \sum_{k \ge 0} (-1)^k h^{p,q}(H^k_c(W, \mathbf{C}))$ is the Hodge-Deligne number of W (see [26]).

Definition 2.1 [13] Let X be a normal quasi-projective variety over $\mathbf{C}$ with at worst Gorenstein canonical singularities, $\rho : Y \to X$ a resolution of singularities whose exceptional locus $D \subset Y$ is a normal crossing divisor with components $D_1, \dots, D_r$, and $K_Y = \rho^* K_X + \sum_{i=1}^r a_i D_i$. We set $I = \{1, \dots, r\}$ and define the *stringy E-function* of X by the formula

$$E_{\mathrm{st}}(X; u, v) := \sum_{J \subset I} E(D_J^\circ; u, v) \prod_{j \in J} \frac{uv - 1}{(uv)^{a_j + 1} - 1},$$

where

$$D_J^\circ := \{x \in X \ : \ x \in D_j \Leftrightarrow j \in J\}.$$

If X is projective and $E_{\mathrm{st}}(X; u, v)$ is a polynomial, then we define *stringy Hodge numbers* $h^{p,q}_{\mathrm{st}}(X)$ by the formula

$$E_{\mathrm{st}}(X; u, v) := \sum_{p,q} (-1)^{p+q} h^{p,q}_{\mathrm{st}}(X) u^p v^q.$$

It is important to remark that the above definition doesn't depend on the choice of a resolution ρ [13]. A proof of this independence uses a variant of a non-archimedian integration proposed by Kontsevich [52] and developed by Denef and Loeser [28]. Using some ideas from [27], we can prove the following:

THEOREM 2.2 [10] *Let Δ and Δ^* be two dual to each other reflexive polyhedra of arbitrary dimension d. Then the stringy E-functions of the corresponding projective Calabi-Yau hypersurfaces $\overline{Z}_f \subset \mathbf{P}_\Delta$ and $\overline{Z}_g \subset \mathbf{P}_{\Delta^*}$ satisfy the duality*

$$E_{\mathrm{st}}(\overline{Z}_f; u, v) = (-u)^{d-1} E_{\mathrm{st}}(\overline{Z}_g; u^{-1}, v),$$

i.e., stringy Hodge numbers of $\overline{Z}_f$ and $\overline{Z}_g$ satisfy the topological Mirror Symmetry test:

$$h^{p,q}_{\mathrm{st}}(\overline{Z}_f) = h^{d-1-p,q}_{\mathrm{st}}(\overline{Z}_g) \quad (0 \leq p, q \leq d-1).$$

We remark that the last result holds true for all Calabi-Yau complete intersections in Gorenstein toric Fano varieties and agrees with the duality for reflexive Gorenstein cones.

Let $X := V/G$ be a quotient of a smooth Calabi-Yau manifold V modulo a regular action of a finite group G. It was shown in [17] that the *stringy Euler number*

$$e_{\mathrm{st}}(X) := \lim_{u,v \to 1} E_{\mathrm{st}}(X; u, v) = \sum_{J \subset I} e(D_J^\circ) \prod_{j \in J} \frac{1}{a_j + 1}$$

coincides with the orbifold physicists' Euler number $e(V, G)$ defined by Dixon-Harvey-Vafa-Witten formula [29]:

$$e(V, G) := \frac{1}{|G|} \sum_{gh=hg} e(V^g \cap V^h),$$

where

$$V^g \cap V^h := \{x \in V \ : \ gx = x \,\&\, hx = x\}.$$

This formula is closely related to the so-called McKay correspondence [65].

3 Counting Rational Curves and GKZ-hypergeometric Functions

Let $\partial\Delta$ be the boundary of a reflexive polyhedron $\Delta \subset M_{\mathbf{R}}$, $\{m_1, \ldots, m_r\} := \partial\Delta \cap M$, and $\{a_{m_1}, \ldots, a_{m_r}\}$ the set of coefficients in equations $f(x) = 1 - \sum_{i=1}^r a_{m_i} x^{m_i} = 0$ defining affine Calabi-Yau hypersurfaces $Z_f \subset T_M$. In [7] it was shown that the power series

$$\Phi(a_{m_1}, \ldots, a_{m_r}) = \sum \frac{(k_1 + \ldots + k_r)!}{k_1! \cdots k_r!} a_{m_1}^{k_1} \cdots a_{m_r}^{k_r},$$

where $(k_1, \ldots, k_r) \in \mathbf{Z}^r_{\geq 0}$ runs over all nonnegative integral solutions to the equation $k_1 m_1 + \cdots + k_r m_r = 0$, admits an interpretation as a period of a regular differential $(d-1)$-form $\omega \in H^0(\overline{Z}_f, \Omega^{d-1}_{\overline{Z}_f})$ and satisfies the holonomic differential

system introduced by Gelfand, Kapranov and Zelevinsky [36], i.e. Φ is a generalized GKZ-hypergeometric function. If $\Delta \subset \mathbf{R}^4$ is the convex hull of vectors $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$, $(0,0,0,1)$, $(-1,-1,-1,-1)$, then the corresponding series Φ has the form

$$\Phi_0(z) = \sum_{k\geq 0} \frac{(5k)!}{(k!)^5} z^k,$$

where $z = a_1a_2a_3a_4a_5$. The function $\Phi_0(z)$ satisfies the differential equation $L\Phi(z) = 0$, where

$$L := \left(z\frac{d}{dz}\right)^4 - 5z\left(5z\frac{d}{dz}+1\right)\left(5z\frac{d}{dz}+2\right)\left(5z\frac{d}{dz}+3\right)\left(5z\frac{d}{dz}+4\right),$$

and can completed to a natural basis $\{\Phi_0, \Phi_1, \Phi_2, \Phi_3\}$ of its solutions. Using Mirror Symmetry, Candelas, de la Ossa, Green and Parkes, in the famous paper [21], predicted that the formal power series expansion of the function

$$\mathcal{F}(q) = \frac{5}{2}\left(\frac{\Phi_1}{\Phi_0}\frac{\Phi_2}{\Phi_0} - \frac{\Phi_3}{\Phi_0}\right)$$

with respect to the variable $q = q(z) := \exp(\Phi_1/\Phi_0)$ coincides with the power series

$$F(q) := \frac{5}{2}(\log q)^3 + \sum_{j>0} K_j q^j,$$

where

$$K_j = \sum_{k|j} n_{j/k} k^{-3}$$

and n_i is the "number of rational curves" of degree i on a generic Calabi-Yau quintic 3-fold in $\mathbf{P}^4$. A mathematical verification of this exciting prediction of Mirror Symmetry demanded a lot effort by many mathematicians. As a first step one needed a rigorous mathematical definition for the "number of rational curves". Such a definition has been obtained in terms of Gromov-Witten classes introduced and investigated by Kontsevich-Manin [49], Ruan-Tian [66], and Li-Tian [57]. The second step was the idea of Kontsevich concerning an equivariant Bott's localization formula with respect to torus action on the moduli spaces of stable maps of $\mathbf{P}_1$ to $\mathbf{P}_4$ [51]. The crucial remarkable progress was obtained by Givental who succeeded in identifying solutions of quantum differential equations obtained from equivariant Gromov-Witten classes with the GKZ-hypergeometric periods of mirrors [38, 39, 40, 41]. Detailed expositions of Givental's ideas are contained in [19, 63]. Another complete mathematical proof of this famous prediction of Mirror Symmetry was obtained in 1997 by Lian, Liu and Yau in [58] using so-called *linear gauge σ-models* associated with toric varieties (see Morrison-Plesser [60]).

It was observed in [8] that GKZ-hypergeometric functions allow to make analogous predictions for the "number of rational curves" in arbitrary Calabi-Yau complete intersections in toric varieties. Many of such predictions related to GKZ-hypergeometric functions were investigated by Hosono, Klemm, Lian, Theisen and

Yau in [43, 44, 45, 46]. The most general framework for the study of resonant GKZ-hypergeometric systems associated with reflexive Gorenstein cones was developed by Stienstra [68].

4 Further developements

It is interesting to analyse possibilities for extending toric methods beyond the class of Calabi-Yau complete intersections in Gorenstein toric Fano varieties. A natural class for testing Mirror Symmetry consists of Calabi-Yau complete intersections in homogeneous manifolds, e.g. in Grassmanians, in partial flag manifolds etc. A general construction of mirrors for this class of Calabi-Yau manifolds has been proposed in [14, 15, 16]. An interesting generalization of Givental's technique for complete intersections in homogeneous spaces was obtained by Kim [47].

Another interesting direction is related to the celebrated Strominger-Yau-Zaslow interpretation of Mirror Symmetry as a T-duality using special Lagrangian torus fibrations [69] (see also [61, 42]). Recently some topological torus fibrations on Calabi-Yau hypersurfaces in toric varieties were constructed by Zharkov [70] using methods from [37]. These fibrations agree with some predictions of Leung and Vafa [55].

References

[1] V.I. Arnold, *Critical points of smooth functions*, in Proceedings of ICM-74, Vol I, Vancouver, (1974), 19-40.

[2] P. S. Aspinwall, B. R. Greene, D. R. Morrison *The Monomial-Divisor Mirror Map*, Internat. Math. Res. Notices (1993), 319-337.

[3] A.C. Avram, P. Candelas, D. Jancic, M. Mandelberg, *On the Connectedness of the Moduli Space of Calabi–Yau Manifolds*, Nucl.Phys. B465 (1996), 458-472.

[4] A.C. Avram, M. Kreuzer, M. Mandelberg, H. Skarke, *The web of Calabi-Yau hypersurfaces in toric varieties*, Nucl.Phys. B505 (1997), 625-640.

[5] V.V. Batyrev, *Boundedness of the degree of multidimensional toric Fano varieties*, Mosc. Univ. Math. Bull. 37, No.1 (1982), 28-33.

[6] V.V. Batyrev, *Dual Polyhedra and Mirror Symmetry for Calabi-Yau Hypersurfaces in Toric Varieties*, J. Alg. Geom. 3 (1994), 493-535.

[7] V.V. Batyrev, *Variations of the mixed Hodge structure of affine hypersurfaces in algebraic tori*, Duke Math. J. 69 (1993), 349-409.

[8] V.V. Batyrev, D. van Straten, *Generalized Hypergeometric Functions and Rational Curves on Calabi-Yau Complete Intersections in Toric Varieties*, Commun. Math. Phys. 168 (1995), 493-533. (alg-geom/9307010)

[9] V.V. Batyrev, D.I. Dais, *Strong McKay Correspondence, String-Theoretic Hodge Numbers and Mirror Symmetry*, Topology, 35 (1996), 901-929.

[10] V.V. Batyrev, L. A. Borisov, *Mirror duality and string-theoretic Hodge numbers*, Invent. Math. 126 (1996), 183-203.

[11] V.V. Batyrev, L. A. Borisov, *Dual Cones and Mirror Symmetry for Generalized Calabi-Yau Manifolds*, Mirror Symmetry II, AMS/IP Stud. Adv. Math 1, Amer. Math. Soc. Providence, RI (1997), 71-86.

[12] V.V. Batyrev, *Birational Calabi-Yau n-folds have equal Betti numbers*, to appear in Proc. European Algebraic Geometry Conference (Warwick, 1996), alg-geom/9710020.

[13] V.V. Batyrev, *Stringy Hodge numbers of varieties with Gorenstein canonical singularities*, to appear in Proc. Taniguchi Symposium 1997, "Integrable Systems and Algebraic Geometry, Kobe/Kyoto", alg-geom/9711008.

[14] V. V. Batyrev, *Toric Degenerations of Fano Varieties and Constructing Mirror Manifolds*, alg-geom/9712034.

[15] V. V. Batyrev, I. Ciocan-Fontanine, B. Kim, D. van Straten, *Conifold Transitions and Mirror Symmetry for Calabi-Yau Complete Intersections in Grassmannians*, Nucl. Phys. B514 (1998), 640-666.

[16] V. V. Batyrev, I. Ciocan-Fontanine, B. Kim, D. van Straten, *Mirror Symmetry and Toric Degenerations of Partial Flag Manifolds*, math.AG/9803108.

[17] V. V. Batyrev, *Non-Archimedian integrals and stringy Euler numbers of log-terminal pairs*, to appear in JEMS, math.AG/9803071.

[18] P. Berglund, S. Katz, A. Klemm, *Mirror Symmetry and the Moduli Space for Generic Hypersurfaces in Toric Varieties*, Nucl. Phys. B456 (1995) 153.

[19] G. Bini, C. De Concini, M. Polito, C. Procesi, *On the work of Givental relative to mirror symmetry*, math.AG/9805097.

[20] L. A. Borisov, *Towards the Mirror Symmetry for Calabi-Yau Complete intersections in Gorenstein Toric Fano Varieties*, alg-geom/9310001.

[21] P. Candelas, X. de la Ossa, P. Green and L. Parkes, *A pair of Calabi-Yau manifolds as an exactly soluble superconformal field theory*, Nucl. Phys. B359 (1991), 21-74.

[22] P. Candelas, E. Derrick, L. Parkes, *Generalized Calabi-Yau Manifolds and the Mirror of a Rigid Manifold*, Nucl.Phys. B407 (1993), 115-154.

[23] P. Candelas, X. de la Ossa, S. Katz, *Mirror Symmetry for Calabi-Yau Hypersurfaces in Weighted $\mathbf{P}_4$ and Extensions of Landau Ginzburg Theory*, Nucl. Phys. B450 (1995), 267-292.

[24] D. Cox, *Recent Developement in Toric Geometry*, in *Algebraic geometry*, Proceedings of the Summer Research Institute, Santa Cruz, CA, USA, July 9–29, 1995. Kollar, Janos (ed.) et al., Providence, RI: AMS, Proc. Symp. Pure Math. 62 (1997), 389-346.

[25] V.I. Danilov, *Geometry of toric varieties*, Russ. Math. Surv. 33, No.2 (1978), 97-154.

[26] V.I. Danilov, A.G. Khovansky, *Newton polyhedra and an algorithm for computing Hodge-Deligne numbers*, Math. USSR, Izv. 29 (1987), 279-298.

[27] J. Denef and F. Loeser, *Weights of exponential sums, intersection cohomology, and Newton polyhedra*, Invent. Math. 106 (1991), 275-294.

[28] J. Denef and F. Loeser, *Germs of arcs on singular algebraic varieties and motivic integration*, to appear in Inventiones Mathematicae, math.AG/980303.

[29] L. Dixon, J. Harvey, C. Vafa and E. Witten, *Strings on Orbifolds* I,II, Nucl. Phys B261 (1985), 678-686; B274 (1986), 285-314.

[30] I. Dolgachev, *Mirror symmetry for lattice polarized K3 surfaces*, J. Math. Sci., New York 81, No.3 (1996), 2599-2630.

[31] W. Ebeling, *Strange duality, mirror symmetry, and the Leech lattice*, alg-geom/9612010.

[32] W. Ebeling, *Strange duality and polar duality*, math.AG/9803066.

[33] G. Ellingsrud, S. A. Strmme, *Bott's formula and enumerative geometry*, J. Amer. Math. Soc. 9 (1996), 175-193.

[34] G. Ewald, *Combinatorial convexity and algebraic geometry*, Graduate Texts in Mathematics 168, New York, NY: Springer, 1996, 372 pp.

[35] W. Fulton, *Introduction to toric varieties*, Annals of Mathematics Studies. 131. Princeton, NJ: Princeton University Press (1993), 157pp.

[36] I.M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky, *Hypergeometric functions and toric varieties*, Funct. Anal. Appl. 23 (1989), 94-106.

[37] I.M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky, *Discriminants, resultants, and multidimensional determinants*, Boston, MA: Birkhäuser (1994), 523pp.

[38] A. Givental, *Homological geometry and mirror symmetry*, Proc. of ICM-94, August 3-11, 1994, Zürich, Chatterji, S. D. (ed.), Vol. I. Basel: Birkhäuser (1995), 472-480.

[39] A. Givental, *Homological geometry. I: Projective hypersurfaces*, Sel. Math., New Ser. 1, No.2 (1995), 325-345.

[40] A. Givental, *Equivariant Gromov - Witten Invariants*, Int. Math. Res. Not. 1996, No.1 (1996), 613-663.

[41] A. Givental, *A mirror theorem for toric complete intersections*, alg-geom/9701016.

[42] M. Gross, *Special Lagrangian Fibrations I: Topology*, alg-geom/9710006.

[43] S. Hosono, A. Klemm, S. Theisen, Yau, S.-T., *Mirror Symmetry, Mirror Map and Applications to Calabi-Yau Hypersurfaces*, Commun.Math.Phys. 167 (1995), 301-350.

[44] S.Hosono, B.H.Lian, S.-T.Yau, *GKZ-Generalized Hypergeometric Systems in Mirror Symmetry of Calabi-Yau Hypersurfaces*, Commun.Math.Phys. 182 (1996), 535-578.

[45] S.Hosono, B.H.Lian, S.-T.Yau, *Maximal Degeneracy Points of GKZ Systems*, J. Amer. Math. Soc. 10 (1997), 427-443.

[46] S. Hosono, *GKZ Systems, Gröbner Fans and Moduli Spaces of Calabi-Yau Hypersurfaces*, alg-geom/9707003.

[47] B. Kim, *Quantum Hyperplane Section Theorem For Homogeneous Spaces*, alg-geom/9712008.

[48] M. Kobayashi, *Duality of Weights, Mirror Symmetry and Arnold's Strange Duality*, alg-geom/9502004.

[49] M. Kontsevich, Yu. I. Manin, *Gromov-Witten classes, quantum cohomology, and enumerative geometry*, Commun. Math. Phys. 164 (1994), 525-562.

[50] M. Kontsevich, *Homological Algebra of Mirror Symmetry*, Proc. of ICM-94, August 3-11, 1994, Zürich, Chatterji, S. D. (ed.), Vol. I. Basel: Birkhäuser, (1995), 120-139.

[51] M. Kontsevich, *Enumeration of rational curves via torus actions*, in *The moduli space of curves*, Dijkgraaf, R. H. (ed.) et al., Prog. Math. 129 (1995), 335-368.

[52] M. Kontsevich, Lecture at Orsay (December 7, 1995).

[53] M. Kreuzer, H. Skarke, *On the Classification of Reflexive Polyhedra*, Commun.Math.Phys. 185 (1997), 495-508.

[54] M. Kreuzer, H. Skarke, *Classification of Reflexive Polyhedra in Three Dimensions*, hep-th/9805190.

[55] N. Leung, C. Vafa, *Branes and Toric Geometry*, hep-th/9711013.

[56] A. Libgober, J. Teitelbaum, *Lines on Calabi-Yau complete intersections, mirror symmetry, and Picard-Fuchs equations*, Int. Math. Res. Not. 1993, No.1 (1993), 29-39.

[57] J. Li, G. Tian, *Virtual moduli cycles and Gromov-Witten invariants of algebraic varieties*, J. Amer. Math. Soc. 11, No.1 (1998), 119-174.

[58] B. Lian, Liu, and S.-T. Yau, *Mirror principle I*, Asian J. Math. Vol. 1, no. 4 (1997), 729-763.

[59] D. R. Morrison, *Mirror symmetry and moduli spaces of superconformal field theories*, Proc. of ICM-94, August 3-11, 1994, Zürich, Chatterji, S. D. (ed.), Vol. II. Basel: Birkhäuser, (1995), 1304-1314.

[60] D. R. Morrison, M. R. Plesser, *Summing the Instantons: Quantum Cohomology and Mirror Symmetry in Toric Varieties*, Nucl. Phys. B440 (1995), 279-354

[61] D. R. Morrison, *The Geometry Underlying Mirror Symmetry*, to appear in Proc. European Algebraic Geometry Conference (Warwick, 1996), alg-geom/9608006.

[62] T. Oda, *Convex bodies and algebraic geometry. An introduction to the theory of toric varieties*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge, Bd. 15, Springer-Verlag. VIII, (1988), 212 pp.

[63] R. Pandharipande, *Rational curves on hypersurfaces (after A. Givental)*, math.AG/9806133.

[64] M. Reid, *The moduli space of 3-folds with $K = 0$ may nevertheless be irreducible*, Math. Ann. 278 (1987), 329-334.

[65] M. Reid, *The McKay correspondence and the physicists' Euler number*, Lect. Notes given at Univ. of Utah (1992) and MSRI (1992).

[66] Y. Ruan, G. Tian, *A mathematical theory of quantum cohomology*, J. Diff. Geom. 42 (1995), 259-367.

[67] H. Skarke, *Weight systems for toric Calabi-Yau varieties and reflexivity of Newton polyhedra*, Mod. Phys. Lett. A11 (1996), 1637-1652.

[68] J. Stienstra, *Resonant Hypergeometric Systems and Mirror Symmetry*, alg-geom/9711002.

[69] A. Strominger, S. T. Yau, E. Zaslov, *Mirror Symmetry is T-duality*, Nucl. Phys. B479 (1996), 243-259.

[70] I. Zharkov, *Torus Fibrations of Calabi-Yau Hypersurfaces in Toric Varieties and Mirror Symmetry*, math.AG/9806091.

Victor Batyrev
Mathematisches Institut
Universität Tübingen
72076 Tübingen, Germany
batyrev@bastau.mathematik.
uni-tuebingen.de

Cohomology of Moduli Spaces of Stable Curves

Maurizio Cornalba

Abstract. We report on recent progress towards the determination of the rational cohomology of the moduli spaces of stable curves.

1991 Mathematics Subject Classification: 14Hxx
Keywords and Phrases: Moduli, Algebraic Curves

The moduli space of smooth n-pointed genus g curves, denoted $\mathcal{M}_{g,n}$, parametrizes isomorphism classes of objects of the form $(C; p_1, \ldots, p_n)$, where C is a smooth genus g curve and $p_1, \ldots, p_n$ are distinct points of C, provided that $2g-2+n > 0$. It has been known for a long time that $\mathcal{M}_{g,n}$ is a quasi-projective variety (cf. [24] for $n = 0$); it is also known, since the work of Deligne, Mumford and Knudsen [6][22][26] that $\mathcal{M}_{g,n}$ is connected and that, although in general non complete, it admits a projective compactification $\overline{\mathcal{M}}_{g,n}$. We wish to describe some recent advances towards the determination of the rational cohomology of $\overline{\mathcal{M}}_{g,n}$, especially in low degree or in low genus. Everything will take place over the complex numbers.

1. Natural Classes

The points of $\overline{\mathcal{M}}_{g,n}$ correspond to isomorphism classes of *stable* n-pointed genus g curves; we recall what these are. Let C be a connected complete curve whose singularities are, at worst, nodes, and let $p_1, \ldots, p_n$ be *smooth* points of C. The *graph* Γ associated to these data consists, first of all, of a set $V = V(\Gamma)$ of *vertices* and a set $L = L(\Gamma)$ of *half-edges*. The set V is just the set of components of the normalization N of C, while L is the set of all points of N mapping to a node or to one of the p_i. The elements of L mapping to nodes come in pairs, the *edges* of the graph, while the remaining ones are called *legs*. For any $v \in V$, we let g_v be the genus of the corresponding component of N, L_v the set of half-edges incident to v, and l_v its cardinality. In addition, the numbering of the p_i yields a numbering of the legs.

The (arithmetic) genus of C can be read off from its graph, and is nothing but the sum of the g_v plus the number of edges minus the number of vertices plus one. The graph will be said to be *stable* if $2g_v - 2 + l_v > 0$ for any vertex v. One says that $(C; p_1, \ldots, p_n)$ is a stable n-pointed genus g curve if its graph is stable; it is easy to see that this is the same as saying that $(C; p_1, \ldots, p_n)$ has a finite automorphism group. Occasionally, it will be useful to consider stable curves whose marked points are indexed by an arbitrary finite set I, rather than by a set of the form $\{1, \ldots, n\}$; we will refer to these as I-pointed curves and shall denote the corresponding moduli space by $\overline{\mathcal{M}}_{g,I}$.

Although in general not smooth, $\mathcal{M}_{g,n}$ and $\overline{\mathcal{M}}_{g,n}$ are orbifolds; in particular, their rational cohomology satisfies Poincaré duality, and the Hodge structure on $H^k(\overline{\mathcal{M}}_{g,n},\mathbb{Q})$ is pure of weight k, for any k.

We will consider two basic types of morphisms between moduli spaces of curves. The first,

$$\pi : \overline{\mathcal{M}}_{g,I\cup\{j\}} \to \overline{\mathcal{M}}_{g,I}\,,$$

simply consists in forgetting about the point labelled by j and passing to the stable model, i.e., roughly speaking, contracting to points all the components that fail to pass the test $2g_v - 2 + l_v > 0$. This morphism has canonical sections σ_i, $i \in I$; the section σ_i associates to any I-pointed curve a new $I\cup\{j\}$-pointed curve obtained by attaching a smooth rational "tail" at the point labelled by i and labelling i and j two distinct points of the tail.

These sections enter, in two different ways, in the construction of cohomology classes on $\overline{\mathcal{M}}_{g,I}$. First of all, we may pull back via σ_i the Chern class of the relative dualizing sheaf of π; the resulting class is usually denoted ψ_i. Next denote by D_i the divisor on $\overline{\mathcal{M}}_{g,I\cup\{j\}}$ traced out by σ_i; then, following [27] and [1], we set

$$\kappa_a = \pi_*(c_1(\omega_\pi(\textstyle\sum D_i))^{a+1})$$

for any non-negative integer a. While the ψ_i are degree two classes, κ_a has degree $2a$.

Further classes can be constructed via the second basic type of map between moduli spaces. For any genus g, I-pointed graph Γ with vertex set V the morphism

$$\xi_\Gamma : X_\Gamma = \prod_{v\in V} \overline{\mathcal{M}}_{g_v,L_v} \to \overline{\mathcal{M}}_{g,I}$$

is obtained by identifying pairs of points corresponding to edges. Observe that stability implies that, for any one of the factors of the left-hand side, either $g_v < g$ or $g_v = g$ and $|L_v| < |I|$. This makes it possible to recursively define *natural*, or *tautological*, classes on $\overline{\mathcal{M}}_{g,I}$. Such a class is simply one that belongs to the subring of $H^*(\overline{\mathcal{M}}_{g,I},\mathbb{Q})$ generated by the κ_a, the ψ_i, and the pushforwards of natural classes via all the morphisms ξ_Γ (or, equivalently, via all the morphisms ξ_Γ where Γ is a graph with only one edge). Notice that all natural classes are algebraic. The image of the morphism ξ_Γ is the closure of the locus of curves whose graph is Γ; its codimension equals the number of edges of Γ. The orbifold fundamental class of this locus, defined as the pushforward via ξ_Γ of the orbifold fundamental class of $\prod_{v\in V}\overline{\mathcal{M}}_{g_v,L_v}$, divided by the order of the automorphism group of Γ, and denoted δ_Γ, is obviously a natural class. The graphs Γ with one edge, which correspond to classes δ_Γ of degree two, come in two kinds. There is the graph with one edge and one vertex of genus $g-1$ (provided g is positive), which we denote by Γ_{irr}, and there are the graphs with one edge and two vertices of genera a and $b = g - a$; if A is the subset of I indexing the legs attached to the genus a vertex we denote such a grap by $\Gamma_{a,A}$. Notice that $\Gamma_{a,A} = \Gamma_{g-a,I\setminus A}$ and that $|A| \geq 2$ if $a = 0$. For brevity, we set $\delta_{irr} = \delta_{\Gamma_{irr}}$, $\delta_{a,A} = \delta_{\Gamma_{a,A}}$.

Two questions now arise. The first is, how far is the cohomology ring of $\overline{\mathcal{M}}_{g,n}$ from the subring of natural classes. The second is, what is the structure of the

latter. It is certainly not the case that the natural classes exhaust the cohomology of $\overline{\mathcal{M}}_{g,n}$, except in special cases. In fact, it is known that $H^{11}(\overline{\mathcal{M}}_{1,11}, \mathbb{Q})$ is not zero, and Pikaart [29] has shown that this can be used to construct nonzero odd-degree cohomology classes in higher genus as well. On the other hand, to my knowledge, nobody has yet produced an even-dimensional cohomology class on some moduli space $\overline{\mathcal{M}}_{g,n}$ which is not natural. For what we know, then, although the evidence in favour of this is very weak, it might still be possible that the even-dimensional cohomology of $\overline{\mathcal{M}}_{g,n}$ is entirely made up of natural classes or, more modestly, that this is true for $H^k(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$ provided k is even and small enough relative to g.

2. Low Degree

Much of what we know about the cohomology of $\mathcal{M}_{g,n}$ for general g and n is due to Harer. In a series of papers [13][16][17] he essentially answered for $H^k(\mathcal{M}_{g,n}, \mathbb{Q})$, $k = 2, 3, 4$, the analogues of the questions we asked in section 1, the easier case of H^1 having been settled before [25]. In this context a natural class is simply a polynomial in the κ_a and the ψ_i. What turns out to be the case is that $H^k(\mathcal{M}_{g,n}, \mathbb{Q})$ vanishes for $g \geq 1$ when $k = 1$ and for $g \geq 9$ ($g \geq 6$ for $n = 0$) when $k = 3$, while $H^2(\mathcal{M}_{g,n}, \mathbb{Q})$ is freely generated by κ_1 and the ψ_i for $g \geq 3$. As for $H^4(\mathcal{M}_{g,n}, \mathbb{Q})$, what Harer shows is that it is freely generated by κ_2 and κ_1^2 for $g \geq 10$ and $n = 0$. In proving these results, Harer uses geometric topology and Teichmüller theory. He uses the same ingredients in another paper [15] to give a bound on (in effect, to compute) the cohomological dimension for constructible sheaves of $\mathcal{M}_{g,n}$, for any g and n. The bound is a direct consequence of the construction of a cellular decomposition of $\mathcal{M}_{g,n}$ by means of Strebel differentials. It would be very interesting to give a proof of this result via algebraic geometry or, alternatively, by producing an exhaustion function on $\mathcal{M}_{g,n}$ with appropriate convexity properties.

A direct calculation of the first, second, third, and fifth rational cohomology groups of $\overline{\mathcal{M}}_{g,n}$ has been carried out in [2]. The results are the following. First of all, $H^k(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$ vanishes for $k = 1, 3, 5$ and for all g and n. Secondly, $H^2(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$ is generated by κ_1, the ψ_i and the fundamental classes of the components of the boundary $\partial\mathcal{M}_{g,n} = \overline{\mathcal{M}}_{g,n} \setminus \mathcal{M}_{g,n}$, freely for $g \geq 3$, and modulo explicit relations otherwise; in particular, $H^2(\overline{\mathcal{M}}_{2,n}, \mathbb{Q})$ is freely generated by the ψ_i and the fundamental classes of the components of the boundary, while $H^2(\overline{\mathcal{M}}_{1,n}, \mathbb{Q})$ is freely generated by the fundamental classes of the components of the boundary. It should be observed that it is known that $\overline{\mathcal{M}}_{g,n}$ is always simply connected (cf. for instance [5]).

The method of proof is entirely algebro-geometric, except for the fact that Harer's bound on the cohomological dimension of $\overline{\mathcal{M}}_{g,n}$ is used; this is one of the reasons why it would be important to give an algebro-geometric proof of Harer's result. We now outline the argument. Harer's bound on the cohomological dimension of $\mathcal{M}_{g,n}$ is that this does not exceed $n - 3$ for $g = 0$, $4g - 5$ for $n = 0$, and $4g - 4 + n$ otherwise. Poincaré duality and the exact sequence of compactly supported cohomology for the inclusion of $\partial\mathcal{M}_{g,n}$ in $\overline{\mathcal{M}}_{g,n}$ immediately show that

$$H^k(\overline{\mathcal{M}}_{g,n}, \mathbb{Q}) \to H^k(\partial\mathcal{M}_{g,n}, \mathbb{Q}) \quad \text{is injective for } k \leq d(g,n),$$

where

$$d(g,n) = \begin{cases} n-4 & \text{if } g = 0, \\ 2g-2 & \text{if } n = 0, \\ 2g-3+n & \text{if } g > 0, n > 0. \end{cases}$$

The idea is to use this Lefschetz-type remark to compute $H^k(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$ inductively on g and n. Before we can do this, however, we need a further remark. The components of $\partial\mathcal{M}_{g,n}$ are precisely the images of the morphisms ξ_Γ, where Γ runs through all graphs with only one edge. We denote by X the disjoint union of the spaces X_Γ such that Γ has one edge, and by ξ the obvious map from X to $\overline{\mathcal{M}}_{g,n}$. Since X is an orbifold, $H^k(X, \mathbb{Q})$ has a Hodge structure of weight k, and the kernel of $\xi^* : H^k(\partial\mathcal{M}_{g,n}, \mathbb{Q}) \to H^k(X, \mathbb{Q})$ is $W_{k-1}H^k(\partial\mathcal{M}_{g,n}, \mathbb{Q})$. As morphisms of mixed Hodge structures are strictly compatible with the filtrations, a class in $H^k(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$ maps to zero in $H^k(X, \mathbb{Q})$ only if it maps to a class in $H^k(\partial\mathcal{M}_{g,n}, \mathbb{Q})$ which comes from $W_{k-1}H^k(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$, that is, since the Hodge structure on $H^k(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$ is pure of weight k, only if it maps to zero. The conclusion is that

$$\xi^* : H^k(\overline{\mathcal{M}}_{g,n}, \mathbb{Q}) \to \bigoplus_{\Gamma \text{ has one edge}} H^k(X_\Gamma, \mathbb{Q}) \quad \text{is injective for } k \leq d(g,n).$$

It is now straightforward to prove the vanishing of $H^k(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$ for $k = 1, 3, 5$, by induction on g and n. In fact, using Künneth (and, for $k \geq 3$, the vanishing of H^h for h less than k and odd) we see that the right-hand side of the above inclusion is a direct sum of H^k of moduli spaces $\overline{\mathcal{M}}_{g',n'}$ such that either $g' < g$ or $g' = g$ and $n' < n$. This reduces us to checking directly a finite number of cases in low genus. For instance, when $k = 1$, the moduli spaces to be examined are just $\overline{\mathcal{M}}_{0,3}$, $\overline{\mathcal{M}}_{0,4}$ and $\overline{\mathcal{M}}_{1,1}$; since the first is a point and the remaining two are isomorphic to the projective line, we are done in this case. We'll return to the initial cases of the induction for $k = 3, 5$ in the next section.

It should be remarked that the argument outlined above works just as well in higher odd degree. For instance if, as I suspect, $H^k(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$ vanishes for all g and n and for $k = 7$, or for $k = 7, 9$, then to prove this it would suffice to do "by hand" the finite number of cases when $k > d(g,n)$.

The induction step is a little more involved for $H^2(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$. Suppose $d(g,n) \geq 2$; we wish to show that $\alpha \in H^2(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$ is a natural class. For any graph Γ with one edge let α_Γ be the pullback of α to X_Γ. By induction hypothesis we know that each α_Γ is a natural class; on the other hand, for any two graphs Γ and Γ', the classes α_Γ and $\alpha_{\Gamma'}$ pull back to the same class on the fiber product of X_Γ and $X_{\Gamma'}$. The idea, roughly speaking, is to try and show that these compatibility conditions force $\xi^*(\alpha)$ to lie in the image under ξ^* of the natural classes on $H^2(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$; once this is done, we conclude by the injectivity of ξ^*. One key ingredient in making all this work is that we have a very good control on how the natural classes pull back under the maps ξ_Γ, or on how they intersect [1][7].

An important step in the proof, which is also interesting per se, is the following. Let $\varphi : \overline{\mathcal{M}}_{g-1,n+2} \to \overline{\mathcal{M}}_{g,n}$ be the morphism that one obtains by identifying the

points labelled by $n+1$ and $n+2$ (this is nothing but $\xi_{\Gamma_{irr}}$). Then

$$\varphi^* : H^k(\overline{\mathcal{M}}_{g,n}, \mathbb{Q}) \to H^k(\overline{\mathcal{M}}_{g-1,n+2}, \mathbb{Q}) \quad \text{is injective for } k \leq \min(2g-2, g+5).$$

As a toy example, we prove this for $k=2$, $g=3$, $n=0$. We need to show that, if $x \in H^2(\overline{\mathcal{M}}_3, \mathbb{Q})$ pulls back to zero under φ, then it pulls back to zero via all the morphisms ξ_Γ such that Γ is a graph with one edge. In the case at hand there is only one such graph beyond Γ_{irr}, namely the graph Γ with a vertex of genus 2 and one of genus 1. Look at the commutative diagram

$$\begin{array}{ccc} \overline{\mathcal{M}}_{1,\{i,j,h\}} \times \overline{\mathcal{M}}_{1,\{l\}} & \xrightarrow{\eta} & \overline{\mathcal{M}}_{2,2} \\ \downarrow{\scriptstyle \varphi' \times \mathbf{1}} & & \downarrow{\scriptstyle \varphi} \\ \overline{\mathcal{M}}_{2,\{h\}} \times \overline{\mathcal{M}}_{1,\{l\}} & \xrightarrow{\xi_\Gamma} & \overline{\mathcal{M}}_3 \end{array}$$

where φ' is the analogue of φ and η consists in identifying the points labelled by h and l. Then, by the vanishing of H^1, the second cohomology group of the lower left corner is just $H^2(\overline{\mathcal{M}}_{2,\{h\}}, \mathbb{Q}) \oplus H^2(\overline{\mathcal{M}}_{1,\{l\}}, \mathbb{Q})$, so we can write $\xi_\Gamma^*(x) = (y, z)$. Since $\varphi(x) = 0$, $(\varphi'(y), z)$ vanishes, showing in particular that $z = 0$. A similar argument shows that y vanishes as well, finishing the proof.

In a certain sense, the injectivity of φ^* in high enough genus can be viewed as a partial analogue, in our context, of the stability results of Harer and Ivanov [14][20][18] for the cohomology of $\mathcal{M}_{g,n}$.

In principle, one could try to treat higher even degree cohomology groups of $\overline{\mathcal{M}}_{g,n}$ along the same lines as those followed for H^2. Let us look at H^4, for instance. The initial cases of the induction are no problem at all. In performing induction, however, aside from the greater complication of the linear algebra involved, a further problem arises. The method of calculating $H^k(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$ we have outlined requires, at each stage of the induction, that we have complete control on all the relations satisfied by the natural classes in $H^k(\overline{\mathcal{M}}_{g',n'}, \mathbb{Q})$ for $g' < g$ or for $g' = g$, $n' < n$. When $k = 2$ it is a relatively easy matter to find them. Already for $k = 4$, however, it is not at all clear what precisely these relations are; new and unexpected ones in low genus have recently been discovered [10][28][3], but there may well be more. It is a very interesting problem to find all relations satisfied by the natural classes in $H^4(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$ and, in perspective, in higher even degree as well.

It would also be of considerable interest to give a proof of the induction step in even degree which is not as computational, but based on a more conceptual understanding of why natural classes on the components of the boundary which match on "intersections" of these patch together to yield a natural class on the whole space.

3. Low Genus

Keel [21] has determined the cohomology ring of $\overline{\mathcal{M}}_{0,n}$, for any $n \geq 3$, in terms of generators and relations. The ring in question is generated by the classes δ_Γ,

where Γ runs through all stable graphs with one edge, which in this case are all of the form $\Gamma_{0,A}$, where A is a subset of $\{1,\dots,n\}$ such that $2 \le |A| \le n-2$. The relations are generated by a set of linear ones and a set of quadratic ones. The linear relations are that, for any set $\{i,j,h,k\}$ of distinct indices,

$$\sum_{\substack{A \ni i,j \\ A \not\ni h,k}} \delta_{0,A} = \sum_{\substack{A \ni i,h \\ A \not\ni j,k}} \delta_{0,A} = \sum_{\substack{A \ni i,k \\ A \not\ni j,h}} \delta_{0,A}\,.$$

The quadratic relations say that $\delta_{0,A} \cdot \delta_{0,B} = 0$ unless $A \cap B = \emptyset$, $A \cap B^c = \emptyset$, $A^c \cap B = \emptyset$, or $A^c \cap B^c = \emptyset$.

In higher genus our knowledge is far less complete. Getzler [9][11] has found a generating function for the Serre characteristics (and also for their $\mathbb{S}_n$-equivariant versions) of the moduli spaces $\overline{\mathcal{M}}_{1,n}$ and the Serre characteristic of $\overline{\mathcal{M}}_{2,n}$ for $n \le 3$ (again, $\mathbb{S}_n$-equivariant or not). The Serre characteristic of a quasi-projective variety is defined as the Euler characteristic of its compactly supported cohomology in the Grothendieck group of mixed Hodge structures; it is important to notice that, since $\overline{\mathcal{M}}_{g,n}$ is an orbifold, and hence the Hodge structure on its k-th cohomology group is pure of weight k for any k, the Serre characteristic of $\overline{\mathcal{M}}_{g,n}$ determines the Hodge numbers. As an example of the results one obtains, the non-zero Hodge numbers of $\overline{\mathcal{M}}_{2,2}$ turn out to be $h^{0,0} = h^{5,5} = 1$, $h^{1,1} = h^{4,4} = 6$, and $h^{2,2} = h^{3,3} = 14$. Bini, Gaiffi and Polito [4] have found a generating function for the Euler characteristics of the spaces $\overline{\mathcal{M}}_{2,n}$, and Harer has found a generating function for the Euler characteristics $\chi(\overline{\mathcal{M}}_{g,n})$; the formula given in [4] is a closed expression in a single, recursively computable, power series (the series $A(t)$ below).

The arguments used to obtain all these results have a common basis. Let I be a finite set and let Γ be a stable genus g, I-pointed graph. We denote by $\mathcal{M}(\Gamma)$ the moduli space of those stable, genus g, I-pointed curves whose graph is Γ. This is a locally closed subspace of $\overline{\mathcal{M}}_{g,I}$ which is nothing but the image of $\prod_{v \in V(\Gamma)} \mathcal{M}_{g_v,L_v}$ under the map ξ_Γ. In fact, $\mathcal{M}(\Gamma)$ is the quotient of $\prod_{v \in V(\Gamma)} \mathcal{M}_{g_v,L_v}$ modulo the automorphism group of Γ; as such, it is an orbifold. The $\mathcal{M}(\Gamma)$ give a stratification of $\overline{\mathcal{M}}_{g,I}$, the *topological stratification*. Now suppose, for instance, that we want to calculate the Euler characteristic of $\overline{\mathcal{M}}_{g,n}$. This is just the sum of the characteristics of the open strata in the topological stratification, since these satisfy Poincaré duality. Thus it suffices to know the Euler characteristics of the open moduli spaces $\mathcal{M}_{g',n'}$ for $g' < g$ or for $g' = g$, $n' \le n$, and of certain quotients of their products. The proper setup for systematcally exploiting this phenomenon is the one of modular operads [12]. Here, however, we content ourselves with sketching the argument of [4] for the spaces $\overline{\mathcal{M}}_{1,n}$.

The top stratum of $\overline{\mathcal{M}}_{g,n}$ is $\mathcal{M}_{g,n}$. Its Euler characteristic could be calculated using the methods of [19], but for $g \le 2$ it can be computed in an elementary way. Look for instance at $\pi : \mathcal{M}_{1,n+1} \to \mathcal{M}_{1,n}$. Since any automorphism of a smooth genus 1 curve fixing five or more points is the identity, for $n \ge 5$ the fiber is a smooth genus 1 curve minus n points; by the multiplicativity of Euler characteriscs in fibrations $\chi(\mathcal{M}_{1,n+1}) = -n\chi(\mathcal{M}_{1,n})$. When $n \le 4$ this has to be modified a bit to take into account the fact that π is no longer a fibration, but some of the fibers are quotients of a smooth genus 1 curve minus n points modulo a finite group. At

any rate, it is straightforward to compute $\chi(\mathcal{M}_{1,n})$ inductively on n starting from $\chi(\mathcal{M}_{1,1}) = 1$; it turns out to be 1 for $n = 2$, 0 for $n = 3, 4$, and $(-1)^n(n-1)!/12$ for $n \geq 5$.

The goal is to compute the generating function $K_g(t) = \sum_n \chi(\overline{\mathcal{M}}_{g,n})\frac{t^n}{n!}$ (for $g = 1$). For any fixed g, the genus g stable graphs fall into a finite number of different patterns, the contribution of each of which to the generating function is handled separately. In genus 1 there are just two patterns: some graphs contain a genus 1 vertex, to which a finite number of trees are attached, while the remaining ones contain a "necklace" of edges, again with trees attached.

Let us look at a graph Γ of the first kind. It has no automorphisms. If we sever the edges stemming from the genus 1 vertex, we are left with a graph consisting of a genus 1 vertex with m legs, and stable graphs of genus zero $\Gamma_1, \ldots, \Gamma_h$, $h \leq m$, where Γ_i is (k_i+1)-pointed and $n = m - h + \sum k_i$. Thus $\chi(\mathcal{M}(\Gamma)) = \chi(\mathcal{M}_{1,m}) \prod_i \chi(\mathcal{M}(\Gamma_i))$. Now set

$$A(t) = t + \sum_{n\geq 2} \sum_G \chi(\mathcal{M}(G)) \frac{t^n}{n!},$$

where the inner sum runs through all genus zero stable $(n+1)$-pointed graphs. The contribution of the graphs we are considering to the generating function K_1 is then $\sum_n \chi(\mathcal{M}_{1,n})\frac{A^n}{n!}$. The same considerations show that $A = t + \sum_{n\geq 2} \chi(\mathcal{M}_{0,n+1})\frac{A^n}{n!}$; since $\chi(\mathcal{M}_{0,n+1})$ can be easily calculated (it equals $(-1)^n(n-2)!$) this relation makes it possible to recursively compute the coefficients of A. Since the characteristics $\chi(\mathcal{M}_{1,n})$ are known, the contribution of the graphs of the first kind to K_1 can be calculated to any given order. The contribution of the graphs of the second kind can be evaluated by similar means; the only new fact is that, when the necklace consists of a single edge, or of two edges, there is an order two automorphisms (reversing orientation of the edge, or interchanging the two edges). For instance, the contribution coming from a graph falling in the first of these two subcases is of the form $\chi(\mathcal{M}_{0,m}/\mathbb{S}_2) \prod_i \chi(\mathcal{M}(\Gamma_i))$, where $\Gamma_1, \ldots, \Gamma_h$ are stable genus zero graphs and $h \leq m-2$. If $m \geq 5$, $\mathcal{M}_{0,m} \to \mathcal{M}_{0,m}/\mathbb{S}_2$ is unramified, so $\chi(\mathcal{M}_{0,m}/\mathbb{S}_2) = \frac{1}{2}\chi(\mathcal{M}_{0,m})$, while $\chi(\mathcal{M}_{0,m}/\mathbb{S}_2) = \chi(\mathcal{M}_{0,m})$ for $m = 3, 4$. Putting everything together gives the final result

$$\sum_n \chi(\overline{\mathcal{M}}_{1,n})\frac{t^n}{n!} = \frac{19}{12}A + \frac{23}{24}A^2 + \frac{5}{18}A^3 + \frac{1}{24}A^4 - \frac{1}{12}\log(1+A) - \frac{1}{2}\log(1-\log(1+A)).$$

For Serre characteristics, the strategy is similar. That the characteristic of $\overline{\mathcal{M}}_{g,n}$ can be expressed in terms of those of the strata in the topological stratification follows from the fact that there is a spectal sequence abutting to $H^\bullet(\overline{\mathcal{M}}_{g,n}, \mathbb{Q})$ whose E_2 term is

$$E_2^{p,q} = \bigoplus_{\Gamma \text{ has } -q \text{ edges}} H_c^{p+q}(\mathcal{M}(\Gamma), \mathbb{Q}).$$

What seems really hard, in this approach, is computing Serre characteristics of open strata, in particular those of the open moduli spaces $\mathcal{M}_{g,n}$. The naive method

we used for Euler characteristics cannot be employed since Serre characteristics do not behave multiplicatively in fibrations. To treat the cases $g = 1$ [8] and $g = 2$, $n \leq 3$ [11], Getzler uses a subtle argument whose strategy is to reduce, via the Leray spectral sequence for $\mathcal{M}_{1,n} \to \mathcal{M}_{1,1}$ (resp., for $\mathcal{M}_{2,n} \to \mathcal{M}_2$) and other technology, to calculating the cohomology of certain mixed Hodge modules on $\mathcal{M}_{1,1}$ (resp., on $\mathcal{M}_2$), which is then handled via Eichler-Shimura theory (resp., via Faltings' Eichler spectral sequence). It is not clear how much farther these methods can be pushed.

Virtually all the initial cases of the induction described in section 2 and leading to the determination of the low-dimensional cohomology of $\overline{\mathcal{M}}_{g,n}$ are covered by the results of [21], [9] and [11], but for most of them simple direct proofs are also available. The only cases that escape are those of $\overline{\mathcal{M}}_3$ and $\overline{\mathcal{M}}_{3,1}$, which are needed to trigger the induction for H^5. These can be deduced, via a variant of the arguments of section 2, from the results of [23], where the Poincaré polynomials of $\mathcal{M}_3$ and $\mathcal{M}_{3,1}$ are determined.

It remains to observe that the results of [10], [11] and [3] completely describe not only the additive structure of the cohomology of $\overline{\mathcal{M}}_{1,n}$ and $\overline{\mathcal{M}}_{2,m}$ for $n \leq 4$ and $m \leq 3$, but the multiplicative structure as well.

REFERENCES

1. E. Arbarello, M. Cornalba, *Combinatorial and algebro-geometric cohomology classes on the moduli spaces of curves*, J. Alg. Geom. **5** (1996), 705–749.
2. E. Arbarello, M. Cornalba, *Calculating cohomology groups of moduli spaces of curves via algebraic geometry*, `math.AG/9803001`.
3. P. Belorousski, R. Pandharipande, *A descendent relation in genus 2*, `math.AG/9803072`.
4. G. Bini, G. Gaiffi, M. Polito, *A formula for the Euler characteristic of* $\overline{\mathcal{M}}_{2,n}$, `math.AG/9806048`.
5. M. Boggi, M. Pikaart, *Galois covers of moduli of curves*, preprint 1997.
6. P. Deligne, D. Mumford, *The irreducibility of the space of curves of given genus*, I.H.E.S. Publ. Math. **36** (1969), 75–109.
7. C. Faber, *Algorithms for computing intersection numbers on moduli spaces of curves, with an application to the class of the locus of Jacobians*, `alg-geom/9706006`.
8. E. Getzler, *Resolving mixed Hodge modules on configuration spaces*, `alg-geom/9611003`.
9. E. Getzler, *The semi-classical approximation for modular operads*, to appear in Commun. Math. Phys., `alg-geom/9612005`.
10. E. Getzler, *Intersection theory on* $\overline{\mathcal{M}}_{1,4}$ *and elliptic Gromov-Witten invariants*, J. Amer. Math. Soc. **10** (1997), 973–998.
11. E. Getzler, *Topological recursion relations in genus 2*, `math.AG/9801003`.
12. E. Getzler, M.M. Kapranov, *Modular Operads*, Compositio Math. **110** (1998), 65–126.
13. J. Harer, *The second homology group of the mapping class group of an orientable surface*, Invent. Math. **72** (1982), 221–239.

14. J. Harer, *Stability of the homology of the mapping class groups of orientable surfaces*, Ann. Math. **121** (1985), 215–249.
15. J. Harer, *The virtual cohomological dimension of the mapping class group of an orientable surface*, Inv. Math. **84** (1986), 157–176.
16. J. Harer, *The third homology group of the moduli space of curves*, Duke Math. J. **65** (1991), 25–55.
17. J. Harer, *The fourth homology group of the moduli space of curves*, to appear.
18. J. Harer, *Improved stability for the homology of the mapping class groups of surfaces*, to appear.
19. J. Harer, D. Zagier, *The Euler characteristic of the moduli space of curves*, Inv. Math. **85** (1986), 457–485.
20. N.V. Ivanov, *On the homology stability for Teichmüller modular groups: closed surfaces and twisted coefficients*, in "Mapping class groups and moduli spaces of Riemann surfaces" (C.-F. Bödigheimer and R. M. Hain, eds.), Contemp. Math. 150, Amer. Math. Soc., Providence, RI, 1993, pp. 149–194.
21. S. Keel, *Intersection theory of moduli space of stable N-pointed curves of genus zero*, Trans. AMS **330** (1992), 545–574.
22. F.F. Knudsen, *The projectivity of the moduli space of stable curves*; *I* (with D. Mumford), Math. Scand. **39** (1976), 19–55; *II, III*, Math. Scand. **52** (1983), 161–199, 200–212.
23. E. Looijenga, *Cohomology of $\mathcal{M}_3$ and $\mathcal{M}_3^1$*, in "Mapping class groups and moduli spaces of Riemann surfaces" (C.-F. Bödigheimer and R. M. Hain, eds.), Contemp. Math. 150, Amer. Math. Soc., Providence, RI, 1993, pp. 205–228.
24. D. Mumford, *Geometric Invariant Theory*, Springer-Verlag, Berlin-Heidelberg, 1965.
25. David Mumford, *Abelian quotients of the Teichmüller modular group*, J. d'Anal. Math. **18** (1967), 227–244.
26. D. Mumford, *Stability of projective varieties*, L'Ens. Math. **23** (1977), 39–110.
27. D. Mumford, *Towards an enumerative geometry of the moduli space of curves*, in "Arithmetic and Geometry" (M. Artin, J. Tate, eds.), vol. 2, Birkhäuser, Boston, 1983, pp. 271–328.
28. R. Pandharipande, *A geometric construction of Getzler's relation in $H^*(\overline{\mathcal{M}}_{1,4}, \mathbb{Q})$*, `math.AG/9705016`.
29. M. Pikaart, *An orbifold partition of $\overline{M}_g^n$*, in "The Moduli Space of Curves" (R. Dijkgraaf, C. Faber, G. van der Geer, eds.), Progress in Mathematics 129, Birkhäuser, Boston, 1995, pp. 467–482.

Maurizio Cornalba
Dipartimento di Matematica
"Felice Casorati"
Università di Pavia
via Ferrata 1
27100 Pavia, Italia
cornalba@unipv.it

Barsotti-Tate Groups and Crystals

A. J. de Jong

At the international congress of 1970 in Nice, A. Grothendieck gave a lecture with the title "Groupes de Barsotti-Tate et cristaux". Grothendieck's lecture describes the crystalline Dieudonné module functor for Barsotti-Tate (or p-divisible) groups over schemes of characteristic p. In this lecture we will see that crystalline Dieudonné module theory has progressed quite a bit since then. We have results on faithfulness, equivalence and faithfulness up to isogeny, there are applications to questions concerning abelian schemes, and we can use the theory to predict results on higher crystalline cohomology groups.

Another purpose of these lecture notes is to explain results on extensions of homomorphisms of p-divisible groups and applications that were obtained by the author.

1. Generalities

We fix a prime number p. Let S denote a scheme of characteristic p. For the definition of a p-divisible group over S we refer to Grothedieck's exposé [1]. One way to obtain a p-divisible group over S is to consider the p-divisible group associated to an abelian scheme A over S. This is basically the system $G = \{G(n)\}$ of finite locally free group schemes $G(n) := A[p^n]$ over S. It is usually denoted $A[p^\infty]$.

The crystalline Dieudonné module functor is a functor

$$\mathbb{D} : BT_S{}^\circ \longrightarrow DC_S.$$

Here BT_S stands for the category of p-divisible groups over S and DC_S stands for the category of Dieudonné crystals over S.

We would like to indicate the meaning of the term "Dieudonné crystal over S". Suppose that $S = \mathrm{Spec}(A)$ is affine. Let $J \to \mathbb{Z}_p[\{x_\alpha\}] \to A$ be a surjection of a polynomial ring over $\mathbb{Z}_p$ onto A with kernel J. Note that $p \in J$. We would like to have "divided powers" on J. This we achieve by formally adding $f^n/n!$ for every $f \in J$; we obtain a new ring D. Up to torsion one can think of this as a subring of $\mathbb{Z}_p[\{x_\alpha\}] \otimes \mathbb{Q}$. Let $\hat{D}$ denote the p-adic completion of D. There is a module of continuous differentials $\hat{\Omega}^1_{\hat{D}}$ and a differential

$$d : \hat{D} \longrightarrow \hat{\Omega}^1_{\hat{D}}.$$

Furthermore, there is still a surjection $\hat{D} \to A$ and the Frobenius endomorphism of A lifts to an endomorphism $\sigma : \hat{D} \to \hat{D}$ (for example by mapping x_α to x_α^p).

In this situation, a Dieudonné crystal over $\operatorname{Spec} A$ is given by a *crystalline Dieudonné module* over $(\hat{D}, \mathrm{d}, \sigma)$. Such a Dieudonné module is a quadruple (M, ∇, F, V) where

a) M is a finite locally free $\hat{D}$ module,
b) $\nabla : M \to M \otimes \hat{\Omega}^1_{\hat{D}}$ is a p-nilpotent connection over the differential d, and
c) $F : M \otimes_\sigma \hat{D} \to M$ and $V : M \to M \otimes_\sigma \hat{D}$ are linear maps, horizontal (for ∇) and satisfy $FV = p$ and $VF = p$.

It turns out that this notion is independent of our choices in the construction of $\hat{D}$ and functorial in A. Thus we obtain a category DC_S for every scheme of characteristic p. We obtain the category of *nondegenerate F-crystals* (see [10]) if we consider *(∇, F)-modules over $\hat{D}$*: triples (M, ∇, F), with M, ∇ and F as above such that the kernel and cokernel of F are annihilated by a power of p. If we write FC_S for the category of nondegenerate F-crystals then there is a forgetful functor $DC_S \to FC_S$. This functor is fully faithful in almost all situations and certainly fully faithful up to isogeny.

For a construction of the functor $\mathbb{D}$ we refer to [11], [12] and [13]. The functor $\mathbb{D}$ turns $\mathbb{G}_m[p^\infty]$ into the Dieudonné module $(\hat{D}, \mathrm{d}, p, 1)$ (i.e., $\nabla = \mathrm{d}, F = p, V = 1$) and it turns $\mathbb{Q}_p/\mathbb{Z}_p$ into the module $(\hat{D}, \mathrm{d}, 1, p)$.

It is clear that the definition of Dieudonné crystals (and F-crystals) given above is rather hard to work with; in fact a lot of work has been done to describe the category (for special $S = \operatorname{Spec}(A)$) in terms of more suitable rings $\hat{D}$. We will see an example of this below.

2. Properties of $\mathbb{D}$

Quite a lot is known due to work of Berthelot, Bloch, Kato, Messing and the author. Here we just list the strongest results that are known to the author. At the moment of writing these notes, the results of (vi)–(ix) have not yet been published.

(i) $\mathbb{D}$ is an equivalence over a perfect field; this follows from the classical Dieudonné theory, as was mentioned in Grothendieck's lecture.
(ii) $\mathbb{D}$ is provably faithful whenever S is reasonable; for example if S is reduced, or if S is Noetherian. The author does not know of a single counter example.
(iii) $\mathbb{D}$ is fully faithful on schemes having locally a p-basis, see [2].
(iv) $\mathbb{D}$ is an equivalence on regular schemes of finite type over a field with a finite p-basis, see [3].
(v) $\mathbb{D}$ is fully faithful up to isogeny over schemes of finite type over a field with a finite p-basis, see [3].
(vi) The finite p-basis hypothesis may be removed from the two last statements, see [4].
(vii) $\mathbb{D}$ is fully faithful in certain cases where S is a local complete intersection. For example if S is of finite type over a field and a l.c.i., and more generally if S is excellent and all of its complete local rings are complete intersections, see [4] (compare also [2]).
(viii) $\mathbb{D}$ is fully faithful up to isogeny over an excellent local ring, see [4].

(ix) $\mathbb{D}$ is essentially surjective up to isogeny over a surface, see [5].

Perhaps the functor $\mathbb{D}$ is an equivalence up to isogeny over schemes of finite type over a field? There are also some negative results:

(x) $\mathbb{D}$ is not fully faithfull in general. This fails even over the ring $\mathbb{F}_p[x,y]/(x^2,xy,y^2)$. See [2, 4.4.1].

(xi) For the experts we mention that the crystalline Dieudonné module functor on the category of *truncated* Barsotti-Tate groups is not fully faithful over $\mathbb{F}_p[t]/(t^p)$. This answers a quetion of [2, 4.4.3] in the negative.

3. Extending homomorphisms, an application of Dieudonné modules

Let G and H be p-divisible groups over a discrete valuation ring R with field of fractions K. Consider the map

$$\mathrm{Hom}(G,H) \longrightarrow \mathrm{Hom}(G_K,H_K). \tag{1}$$

In [6] Tate proved that (1) is a bijection when the characteristic of K is zero: any homomorphism between the generic fibres extends to a homomorphism over R.

In the introduction of exposé IX in SGA 7 (by A. Grothendieck, M. Raynaud and D. Rim) it was mentioned as a problem whether the same holds when the characteristic of K is p. In this case, set $S = \mathrm{Spec}\, R$ and $\eta = \mathrm{Spec}\, K$. By the results mentioned in the previous section, we can translate this, using $\mathbb{D}$, into a question on Dieudonné crystals: Is the natural restriction functor $DC_S \to DC_\eta$ fully faithful? This follows from the following stronger theorem.

Theorem 1. [7, Theorem 1.1] Assume R has a p-basis. The restriction functor on nondegenerate F-crystals $FC_S \to FC_\eta$ is fully faithful.

In the following section we will try to explain what kind of mathematics goes into the proof of this theorem. In the rest of this section we indicate a few corollaries of the result.

Theorem 2. Let R be an integrally closed, Noetherian, integral domain, with field of fractions K. Let G and H be p-divisible groups over R. A homomorphism $f : G \otimes_R K \to H \otimes_R K$ extends uniquely to a homomorphism $G \to H$.

This occurs in Tate's paper [6], with the additional assumption that K has characteristic 0. The reduction to the case where R is a complete discrete valuation ring is in [6, page 181]. The theorem then follows from Theorem 1, see [7, Introduction].

Theorem 3. [7, Theorem 2.5] Let A be an abelian variety over the discretely valued field K with valuation ring R. Let $G = A[p^\infty]$ be the associated p-divisible group. Then A has good reduction over R if and only if G has good reduction over R. Similarly for semi-stable reduction.

Of course one has to define carefully the significance of the terms "good reduction" and "semi-stable reduction" for p-divisible groups. For this see [7], compare with SGA 7 exposé IX.

Theorem 4. [7, Theorem 2.6] Let F be a field finitely generated over $\mathbb{F}_p$. Let A and B be abelian varieties over F. The natural map

$$\mathrm{Hom}(A,B) \otimes_{\mathbb{Z}} \mathbb{Z}_p \longrightarrow \mathrm{Hom}(A[p^\infty], B[p^\infty])$$

is bijective.

The case of a finite field was done by Tate [9]. The corresponding result where one replaces the p-divisible group by the ℓ-adic Tate module ($\ell \neq p$) has been known for some time now, see [14], [15] and references therein.

4. Power series and F-crystals

In this section we will try to explain what kind of algebra is used in [7] to prove Theorem 1.

Consider the ring $\Omega = \mathbb{Z}_p[[t]]$ together with the derivation $\frac{\mathrm{d}}{\mathrm{d}t}$ and "Frobenius" map $\sigma : t \mapsto t^p$. This is an example of a ring simpler than the ring $\hat{D}$ of Section 2 which can still be used to describe F-crystals over $\operatorname{Spec} \mathbb{F}_p[[t]]$. Recall that a (θ, F)-module over Ω is a triple (M, θ, F), see Section 1. Thus $\theta : M \to M$ is additive and satisfies $\theta(fm) = f\theta(m) + \frac{\mathrm{d}f}{\mathrm{d}t}m$ and F can be seen as a σ-linear map $M \to M$. The horizontality of F means that $\theta(F(m)) = pt^{p-1}F(\theta(m))$.

Our goal is to study systems of equations of the type (with $s \in \mathbb{N}$)

$$\left\{ \begin{array}{lcl} \theta(m) & = & 0 \\ F(m) & = & p^s m \end{array} \right. \tag{2}$$

Here m will be an element of $M \otimes_\Omega \Gamma$, where $\Omega \subset \Gamma$ is an extension of rings such that the derivation $\frac{\mathrm{d}}{\mathrm{d}t}$ and the "Frobenius" map σ extend to Γ. The extensions of σ and $\frac{\mathrm{d}}{\mathrm{d}t}$ will be denoted by the same symbols, and they will induce extensions $F = F \otimes \sigma$ and $\theta = \theta \otimes 1 + 1 \otimes \frac{\mathrm{d}}{\mathrm{d}t}$ on $M \otimes \Gamma$.

The question that has to be answered is of the form: Is any solution m to (2) of the form $m_0 \otimes 1$, where $m_0 \in M$. Of course this is going to depend on the ring Γ.

The specific ring in question is the following: Γ is the ring of formal Laurent series

$$f = \sum_{n \in \mathbb{Z}} a_n t^n,$$

such that $a_n \in \mathbb{Z}_p$ and such that $a_n \to 0$ as $n \to -\infty$. It is obvious how to extend σ and $\frac{\mathrm{d}}{\mathrm{d}t}$. Another description of Γ is that it is the p-adic completion of the localization of Ω at the prime ideal (p).

Thus we have to prove that any solution m to (2) does not have terms with negative exponents in its expansion with repect to some basis of M. The idea is to proceed in two steps: (a) one proves that any solution m is at least (rigid) analytic in some annulus $\eta < |t| < 1$, and (b) using horizontality prove that m extends to an analytic section over the whole disc $|t| < 1$.

More precisely, one defines a subring $\Gamma_c \subset \Gamma$ of elements

$$f = \sum_{n \in \mathbb{Z}} a_n t^n,$$

such that $\exists \eta > 1 : |a_n|\eta^{-n} \to 0$ for $n \to -\infty$. Each element of Γ_c can be thought of as a rigid analytic function on some small annulus as above. The idea to use

the ring Γ_c was first introduced by U. Zannier, who solved the case $\operatorname{rk} M = 2$. However, rings like it had already occurred in the context of Monsky-Washnitzer cohomology and overconvergent F-crystals, see next section.

The first step (a) is the harder of the two. Here we use the ring $\Gamma_{1,c}$ consisting of expressions

$$f = \sum_{\alpha \in \mathbb{Z}[1/p]} a_\alpha t^\alpha$$

with $a_\alpha \to 0$ for $\alpha \to -\infty$ and with a certain convergence condition as in the definition of Γ_c above. Note that σ does extend to $\Gamma_{1,c}$, whereas θ does not. In some sense the main new phenomenon observed in [7] is that there is always a filtration

$$M \otimes \Gamma_{1,c} = M_a \supset \ldots \supset M_1 \supset 0,$$

such that the submodules M_i are F-stable and such that on each quotient M_i/M_{i-1} the map F has pure slope s_i with $s_1 > s_2 > \ldots > s_a$. Roughly speaking F has pure slope $s \in \mathbb{Q}$ if for any element $m \in M$ the sequence of elements $F^n(m)$ become divisible by p^{ns-C} for some constant C, and $\det(F) = p^{\dim(M)s}(unit)$. This in some sense means that all "eigenvalues" of F have p-adic valuation s. For the experts we remark here that this filtration is opposite to the "usual" slope filtration on the module $M \otimes_\Omega \Gamma$.

Having proved this one can deduce step (a): any solution m of (2) lies in $M \otimes \Gamma_c$. To finish, i.e., to do step (b), one applies Dwork's trick which says that the connection on M is isomorphic to the trivial connection over the rigid analytic disc.

5. Overconvergent F-crystals

Overconvergent F-crystals are supposed to be the p-adic analogue of lisse ℓ-adic sheaves. They have been introduced by Berthelot, see [8] for example. Presently, there are more questions then answers concerning these crystals. In this section we recall the semi-stable reduction conjecture for these objects; such a conjecture occurs in work of R. Crew, N. Tsuzuki and others. It is the p-adic analog of the phenomenon of quasi-unipotent monodromy and the nilpotent orbit theorem for variations of Hodge structure.

Suppose that $\mathcal{E}$ is a nondegenerate F-crystal over $\mathbb{P}^1_{\mathbb{F}_p} \setminus \{0\}$. Then $\mathcal{E}$ will give rise to a (θ, F)-module $M(\mathcal{E}) = (M, \theta, F)$ over the ring Γ described in the previous section. Let us say that $\mathcal{E}$ is an overconvergent F-crystal on $\mathbb{P}^1 \setminus \{0\}$ if $M \cong N \otimes_{\Gamma_c} \Gamma$ for some (θ, F)-module (N, θ, F) over Γ_c. This definition is not the same as the correct definition (see [8]), but undoubtedly it is equivalent.

Of course this definition is too specialized. We leave it to the reader to formulate the meaning of overconvergence when $\mathcal{E}$ is a nondegenerate F-crystal over a smooth affine curve X over a field k of characteristic p. (Of course there will be an "overconvergence" condition at each "missing point" of X.)

Next, let us try to explain what it means for $\mathcal{E}$ over $\mathbb{P}^1_{\mathbb{F}_p} \setminus \{0\}$ to have semi-stable reduction at $t = 0$. This means that there should exist a finite free $\mathbb{Z}_p[[t]]$-module N, a connection $\theta : N \to (1/t)N$ with at worst a logarithmic pole and a horizontal

σ-linear map $F : N \to N$, all of this such that $N \otimes \Gamma \cong M(\mathcal{E})$. In more technical terms: $\mathcal{E}$ should extend to a (nondegenerate) log-F-crystal over $\mathbb{P}^1$.

Again this is too special. Let $\mathcal{E}$ be an F-crystal over a smooth curve X over a field of characteristic p. Then there is a natural notion of semi-stable reduction of $\mathcal{E}$ at each point x of a projective completion $\overline{X}$ of X. The conjecture can now be formulated as follows:

Conjecture. For any overconvergent F-crystal $\mathcal{E}$ over the curve X there exists a finite morphism of curves $\pi : Y \to X$ such that $\pi^*\mathcal{E}$ has semi-stable reduction at every point of a projective completion $\overline{Y}$ of Y.

The evidence for this conjecture is slender; it has been proved for unit root crystals by N. Tsuzuki. Assuming the conjecture one can prove finiteness for the rigid cohomology of $\mathcal{E}$ over X.

There are several natural generalizations of these notions to the case of varieties of higher dimension. For example one could define a nondegenerate F-crystal over a variety X to be overconvergent if its pullback to every curve mapping to X is overconvergent. (This is not the current definition, see [8].) Then one can ask whether every such overconvergent F-crystal pulls back to a log-F-crystal on $\overline{Y}$, where Y is an alteration of X and $Y \subset \overline{Y}$ is a nice smooth compactification of Y (as in [16]).

For all of these questions and much more on p-adic cohomology we refer the reader to work of P. Berthelot, N. Tsuzuki, R. Crew, G. Christol, Z. Mebkhout, J. Etesse, B. Le Stum, B. Chiarellotto and others.

References

[1] A. Grothendieck, *Groupes de Barsotti-Tate et cristaux*, Actes, Congrès intern. math., 1970. Tome 1, pp. 431-436.

[2] P. Berthelot and W. Messing, *Théorie de Dieudonné cristalline III,* in The Grothendieck Festschrift I, Progress in mathematics 86, Birkhäuser (1990), pp. 171-247.

[3] A.J. de Jong, *Crystalline Dieudonné module theory via formal and rigid geometry*, Publications Mathematiques 82 (1995), pp. 5-96.

[4] A.J. de Jong and W. Messing, *work in progress.*

[5] A.J. de Jong, *An application of alterations to Dieudonné modules,* to appear in: Resolutions of Singularities. A volume to be published in connection with the working week on resolution of singularities, Tirol 1997.

[6] J. Tate, *p-divisible groups,* Proceedings of a conference on local fields, Driebergen (1966), Springer-Verlag.

[7] A.J. de Jong, *Homomorphisms of Barsotti-Tate groups and crystals in positive characteristic*, to appear in Inventiones Mathematicae.

[9] J. Tate, *Endomorphisms of abelian varieties over finite fields,* Inventiones Mathematicae 2 (1966), pp. 134-144.

[8] P. Berthelot, *Cohomologie rigide et cohomologie rigide à supports propres,* Prépublications IRMAR 96-03, Université de Rennes (1996).

[10] N. Saavedra Rivano, *Catégories Tannakiennes,* Lecture notes in Mathematics 265, Springer-Verlag (1972).

[11] W. Messing, *The crystals associated to Barsotti-Tate groups: with applications to abelian schemes,* Lecture Notes in Mathematics 264, Springer-Verlag (1972).

[12] B. Mazur and W. Messing, *Universal extensions and one-dimensional crystalline cohomology,* Lecture Notes in Mathematics 370, Springer-Verlag (1974).

[13] P. Berthelot, L. Breen, W. Messing, *Théorie de Dieudonné crystalline II,* Lecture Notes in Mathematics 930, Springer-Verlag (1982).

[14] J. Zarhin, *Endomorphisms of abelian varieties over field of finite characteristics,* Math. USSR Izvestia 9 (1975), nr 2, pp. 255-260.

[15] S. Mori, *On Tate's conjecture concerning endomorphisms of abelian varieties,* Intl. Symp. of Algebraic Geometry, Kyoto 1977, pp. 219-230.

[16] A.J. de Jong, *Smoothness, semistability and alterations,* Publications Mathematiques 83 (1996), pp. 51-93.

A. J. de Jong
Massachusetts Institute
of Technology
Department of Mathematics
Building 2, room 270
Massachusetts Avenue 77
Cambridge MA 02139-4307, USA
dejong@math.mit.edu

Higher Abel-Jacobi Maps

Mark L. Green*

1991 Mathematics Subject Classification: 14C25, 14C30, 14D07, 32G20, 32J25
Keywords and Phrases: Chow group, algebraic cycle, Abel-Jacobi map, Gauss-Manin connection, Hodge structure

For a smooth projective variety X, the structure of the Chow group $CH^p(X)$ representing codimension p algebraic cycles modulo rational equivalence, is still basically a mystery when $p > 1$, even for 0-cycles on a surface. For any p, one has the (rational) *cycle class map*

$$\psi_0 : CH^p(X) \otimes \mathbf{Q} \to \mathrm{Hdg}^p(X) \otimes \mathbf{Q} \subseteq H^{2p}(X, \mathbf{Q}),$$

conjecturally surjective by the *Hodge conjecture*. By the work of Griffiths, we have the (rational) *Abel-Jacobi map*

$$\psi_1 = \mathrm{AJ}^p_X : \ker(\psi_0) \to J^p(X) \otimes \mathbf{Q}.$$

A number of beautiful results have been proved using this invariant (e.g. [Gri]), but through the work of Mumford-Roitman ([Mu],[Ro]) it was realized that the kernel of ψ_1 can be infinite-dimensional (for 0-cycles on a surface with $H^{2,0}(X) \neq 0$), while through the work of Griffiths and Clemens the image of ψ_1 may fail to be surjective [Gri] or even finitely generated [Cl] (for 1-cycles on a general quintic 3-fold) or yet for not dissimilar geometric situations, by work of Voisin [Vo1] and myself [Gre1], the image of ψ_1 may be 0 (for 1-cycles on a general 3-fold of degree ≥ 6). At present, there is no explicit description, even conjecturally, for what $\ker(\psi_1)$ and $\mathrm{im}(\psi_1)$ look like. Eventually it came to be understood through the work of Beilinson, Bloch, Deligne, and Murre, among others (see [Ja] for a discussion) that there ought to be a filtration

$$CH^p(X) \otimes \mathbf{Q} = F^0 CH^p(X) \otimes \mathbf{Q} \supseteq F^1 CH^p(X) \otimes \mathbf{Q} \supseteq \cdots \supseteq F^{p+1} CH^p(X) \otimes \mathbf{Q} = 0$$

with

$$F^1 CH^p(X) \otimes \mathbf{Q} = \ker(\psi_0)$$

and

$$F^2 CH^p(X) \otimes \mathbf{Q} = \ker(\psi_1).$$

* Research partially supported by the National Science Foundation.

This filtration has been constructed in some cases and various geometric candidates for it have been put forward, for example, by S. Saito [Sa] and Jannsen. One suggestion as to what the graded pieces of this filtration should look like is given by *Beilinson's conjectural formula* (see [Ja])

$$Gr^m CH^p(X) \otimes \mathbf{Q} \cong \mathrm{Ext}^m_{\mathcal{MM}}(1, h^{2p-m}(X)(p)),$$

where $\mathcal{MM}$ is the conjectural category of mixed motives.

One case that stands out as being well-understood is the case of the relative Chow group $CH^2(\mathbf{P}^2, T)$, where $T \subset \mathbf{P}^2$ is the triangle $z_0 z_1 z_2 = 0$, roughly described as 0-cycles on $\mathbf{P}^2 - T = \mathbf{C}^* \times \mathbf{C}^*$, modulo divisors of meromorphic functions f on curves $C \subset \mathbf{P}^2$ such that $f = 1$ on $C \cap T$. There is a series of maps

$$\begin{aligned}
&\psi_0 \colon CH^2(\mathbf{P}^2, T) \to \mathbf{Z};\\
&\psi_1 \colon \ker(\psi_0) \to \mathbf{C}^* \oplus \mathbf{C}^*;\\
&\psi_2 \colon \ker(\psi_1) \to K_2(\mathbf{C}).
\end{aligned}$$

Recall

$$K_2(\mathbf{C}) = \frac{\mathbf{C}^* \otimes_{\mathbf{Z}} \mathbf{C}^*}{\{\text{Steinberg relations}\}},$$

where the Steinberg relations are generated by $\{a \otimes (1-a) \mid a \in \mathbf{C} - \{0,1\}\}$. It is known (Bloch [Bl], Suslin [Su]) that these are all surjective and ψ_2 is an isomorphism. These all have simple algebraic descriptions—ψ_0 is degree; $\psi_1(a,b) = a \oplus b$; $\psi_2(a,b) = \{a,b\}$. The essential tool in proving this is the Suslin reciprocity theorem ([Su], see also [To]).

Another illustrative example (see [Gre2]) is $CH^2(\mathbf{P}^2, E)$, where E is a smooth plane cubic. Here we have a series of maps

$$\begin{aligned}
&\psi_0 \colon CH^2(\mathbf{P}^2, E) \to \mathbf{Z};\\
&\psi_1 \colon \ker(\psi_0) \to 0;\\
&\psi_2 \colon \ker(\psi_1) \to \frac{\mathbf{C}^* \otimes_{\mathbf{Z}} E}{\tilde{\theta}(J^4)}.
\end{aligned}$$

If $a, b \in \mathbf{P}^2 - E$, and L is the line through a and b, which meets E in $\{p_1, p_2, p_3\}$, then

$$\psi_2((a) - (b)) = \sum_{i=1}^{3} \left(\frac{a - p_i}{b - p_i} \otimes p_i\right) \in \mathbf{C}^* \otimes_{\mathbf{Z}} E.$$

Using $p_1 + p_2 + p_3 = 0$ on E (having taken 0 to be an inflection point), this has an alternative expression

$$\psi_2((a) - (b)) = \frac{(a-p_2)(b-p_1)}{(a-p_1)(b-p_2)} \otimes p_2 + \frac{(a-p_3)(b-p_1)}{(a-p_1)(b-p_3)} \otimes p_3,$$

which involves cross-ratios on L and is more clearly coordinate-independent. Once again, all three maps are surjective, and ψ_2 is an isomorphism. $\psi_0 = \deg$, $\psi_1 = 0$,

inserted to preserve the pattern. To explain the notation in ψ_2, for $a \in E - \mathrm{div}(\theta)$, let

$$\tilde{\theta}(a) = \theta(a) \otimes a \in \mathbf{C}^* \otimes_{\mathbf{Z}} E.$$

Extend the definition of $\tilde{\theta}$ to $\mathbf{Z}_E$, the group ring of E, by linearity. In $\mathbf{Z}_E$, let J be the augmentation ideal $\{\sum_i n_i(a_i) \mid n_i \in \mathbf{Z}, a_i \in E, \sum_i n_i = 0\}$. Although $\tilde{\theta}(a)$ depends on the lifting of a to $\mathbf{C}$, on J^4 it does not depend on the choice of lifting of the elements. Thus $\tilde{\theta}(J^4)$ is well-defined and constitutes a generalization of the Steinberg relations; this group has been given a motivic intepretation by Goncharov and Levin [GL].

These examples provide a model for the general case—one should think of $(\mathbf{P}^2, T)$ and $(\mathbf{P}^2, E)$ as analogous to a complete surface with $h^{2,0} = 1$. For a general X, one has

$$\psi_0 \colon CH^2(X) \to \mathrm{Hdg}^2(X);$$
$$\psi_1 \colon \ker(\psi_1) \to J^2(X);$$

where ψ_0 is the cycle class map to the Hodge classes on X, and ψ_1 is the Abel-Jacobi map. We have constructed part of the missing map ψ_2 in the case of 0-cycles on a surface, using a construction that has the potential to work more generally.

The regulator map for a curve

$$X - D \xrightarrow{(f,g)} \mathbf{C}^* \times \mathbf{C}^*$$

is a homomorphism $r \colon \pi_1(X - D) \to \mathbf{C}/(2\pi i)^2\mathbf{Z} = \mathbf{C}/\mathbf{Z}(2)$ given by

$$r(\gamma) = \int_\gamma \log(f)\frac{dg}{g} - \log(g(p)) \int_\gamma \frac{df}{f},$$

where p is a base-point on γ; the answer does not depend on p. If $\gamma = \partial U$ for U a disc in $\mathbf{C}^* \times \mathbf{C}^*$, then

$$r(\gamma) = \int_U \frac{df}{f} \wedge \frac{dg}{g}.$$

This formula generalizes to a definition in the more general situation of a non-singular curve C and a map $f \colon C \to X$ to a smooth projective surface X. If

$$\mu \in \ker(H^2(X, \mathbf{Z}) \xrightarrow{f^*} H^2(C, \mathbf{Z})),$$

then $f^*\mu = dd^c g$ for $g \in A^0(C)$, unique up to adding a constant. If $\gamma \in \ker(H_1(C, \mathbf{Z}) \to H_1(X, \mathbf{Z}))$, so that $\gamma = \partial\Gamma$ in X, then we define

$$e_{X,C}(\mu, \gamma) = \int_\Gamma \mu - \int_\gamma f^*(d^c g) \in \mathbf{C}/\mathbf{Z},$$

which does not depend on any of the choices. These quantities are known as *membrane integrals*. More intrinsically, $e_{X,C}$ is the extension class of the extension of mixed Hodge structures (see [Ca])

$$0 \to \mathrm{coker}(H^1(X) \to H^1(C)) \to H^2(X, C) \to \ker(H^2(X) \to H^2(C)) \to 0.$$

Denote the term on the left $H^1(C)_{\text{new}}$ and the term on the right $H^2(X)_C$; now

$$e_{X,C} \in \frac{\text{Hom}_{\mathbf{C}}(H^2(X)_C, H^1(C)_{\text{new}})}{\text{Hom}_{\mathbf{Z}}(H^2(X)_C, H^1(C)_{\text{new}}) + F^0\text{Hom}_{\mathbf{C}}(H^2(X)_C, H^1(C)_{\text{new}})}.$$

The class $e_{X,C}$ may also be obtained from the image under $AJ_{X\times C}$ of the graph of f minus some terms to make it homologous to 0 on $X \times C$.

We may write

$$H^2(X) = \ker(NS(X) \to H^2(C)) \oplus H^2(X)_{\text{tr}},$$

which decomposes

$$e_{X,C} = (e_{X,C})_{\text{alg}} \oplus (e_{X,C})_{\text{tr}}.$$

The class $(e_{X,C})_{\text{alg}}$ contains the same information as the map

$$\ker(NS(X) \to H^2(C)) \to \frac{J^1(C)}{\text{Alb}(X)}$$

given by

$$L \mapsto f^*L.$$

If $Z \in Z^2(X)$ and $\psi_0(Z) = 0$, $\psi_1(Z) = 0$, then if we lift Z to $\tilde{Z} \in Z^1(C)$ such that $f_*\tilde{Z} = Z$ and $\deg(Z) = 0$ on each component of C, then $AJ_C(\tilde{Z})$ is represented by the extension class $e_{C,\tilde{Z}}$ of the extension of mixed Hodge structures

$$0 \to \text{coker}(H^0(C) \to H^0(|\tilde{Z}|)) \to H^1(C, |\tilde{Z}|) \to H^1(C)_{\text{new}} \to 0;$$

and the divisor $\tilde{Z}$ gives a map $\text{coker}(H^0(C) \to H^0(|\tilde{Z}|)) \to 1$ and then

$$e_{C,\tilde{Z}} \in \frac{\text{Hom}_{\mathbf{C}}(H^1(C)_{\text{new}}, 1)}{\text{Hom}_{\mathbf{Z}}(H^1(C)_{\text{new}}, 1) + F^0\text{Hom}_{\mathbf{C}}(H^1(C)_{\text{new}}, 1)}.$$

The two extensions of MHS fit together to give a 2-step extension of MHS of $H^2(X)_{\text{tr}}$ by 1, which unfortunately cannot be used directly. By standard identifications, we may think of

$$(e_{X,C})_{\text{tr}} \in (\mathbf{R}/\mathbf{Z}) \otimes_{\mathbf{Z}} \text{Hom}_{\mathbf{Z}}(H^2(X)_{\text{tr}}, H^1(C)_{\text{new}})$$

and

$$e_{C,\tilde{Z}} \in (\mathbf{R}/\mathbf{Z}) \otimes_{\mathbf{Z}} \text{Hom}_{\mathbf{Z}}(H^1(C)_{\text{new}}, 1).$$

The tensor product followed by contraction gives an element

$$e_{X,C,\tilde{Z}} \in (\mathbf{R}/\mathbf{Z}) \otimes_{\mathbf{Z}} (\mathbf{R}/\mathbf{Z}) \otimes_{\mathbf{Z}} \text{Hom}_{\mathbf{Z}}(H^2(X)_{\text{tr}}, 1).$$

If we let $U_2^2(X) = \{e_{X,C,\tilde{Z}} \mid f_*\tilde{Z} = 0 \text{ as a } 0\text{−cycle on } X\}$ and

$$J_2^2(X) = \frac{(\mathbf{R}/\mathbf{Z}) \otimes_{\mathbf{Z}} (\mathbf{R}/\mathbf{Z}) \otimes_{\mathbf{Z}} \text{Hom}_{\mathbf{Z}}(H^2(X)_{\text{tr}}, 1)}{U_2^2(X)},$$

then $Z \mapsto [e_{X,C,\tilde{Z}}]$ gives a well-defined invariant

$$\psi_2^2 \colon \ker(\psi_1) \to J_2^2(X)$$

that is independent of the choices of C and $\tilde{Z}$, and which depends only on the rational equivalence class of Z on X. It is necessary to allow reducible curves C. Claire Voisin [Vo2] has shown that for surfaces with $h^{2,0} \neq 0$, the map ψ_2 has infinite-dimensional image, and also that it need not be injective, so that our ψ_2^2 is only part of the story.

An explanation of the role played by the extension class $(e_{X,C})_{\mathrm{alg}}$ comes from *Beilinson's conjectural formula*:

$$Gr^m CH^p(X) \otimes \mathbf{Q} \cong \mathrm{Ext}^m_{\mathcal{MM}}(1, h^{2p-m}(X)(p)),$$

where $\mathcal{MM}$ is the conjectural category of mixed motives. The map

$$f_* \colon Gr^1 CH^1(C) \to Gr^1 CH^1(X)$$

is followed by a map

$$f_*^{+1} \colon \ker(f_*) \to Gr^2 CH^2(X).$$

In terms of Beilinson's formula, this is a map

$$\mathrm{Ext}^1_{\mathcal{MM}}(1, \ker(H^1(C) \to H^3(X))) \to \mathrm{Ext}^2_{\mathcal{MM}}(1, \mathrm{coker}(H^0(C) \to H^2(X)))$$

which (see [Ja]) factors through a map

$$f_*^{+1} \colon \mathrm{Ext}^1_{\mathcal{MM}}(1, \ker(H^1(C) \to H^3(X))) \to \mathrm{Ext}^2_{\mathcal{MM}}(1, H^2(X)_{\mathrm{tr}})).$$

It is reasonable to expect that it is given by Yoneda product with an element

$$e \in \mathrm{Ext}^1_{\mathcal{MM}}(\ker(H^1(C) \to H^3(X)), H^2(X)_{\mathrm{tr}}).$$

The philosophical point here is that e should come from

$$(e_{X,C})_{\mathrm{tr}} \in \mathrm{Ext}^1_{MHS}(\ker(H^1(C) \to H^3(X)), H^2(X)_{\mathrm{tr}}).$$

In fact, one would conjecture that the map

$$\ker(J^1(C) \to J^2(X)) \to CH^2(X)$$

is zero if and only if $(e_{X,C})_{\mathrm{tr}}$ is torsion—the only if direction has been shown [Gre2].

The question then becomes how to use $(e_{X,C})_{\mathrm{tr}}$. One answer is given by ψ_2^2 above. Another piece of the puzzle is to apply the arithmetic Gauss-Manin connection ∇ to $(e_{X,C})_{\mathrm{tr}}$.

An invariant which complements the one above was obtained in joint work with Phillip Griffiths [GG]. By work of Katz [Ka2] and Grothendieck [Gro], there

is for any smooth projective variety X defined over $\mathbf{C}$ the arithmetic Gauss-Manin connection

$$\nabla_{X/\mathbf{Q}}\colon H^k(X,\mathbf{C}) \to \Omega^1_{\mathbf{C}/\mathbf{Q}} \otimes_{\mathbf{C}} H^k(X,\mathbf{C}).$$

To capture this abstractly, we define an *arithmetic Hodge structure (AHS)* to be a complex vector space V with a finite descending filtration $F^\bullet V$ and a $\mathbf{Q}$-linear connection $\nabla\colon V \to \Omega^1_{\mathbf{C}/\mathbf{Q}} \otimes_{\mathbf{C}} V$ satisfying $\nabla^2 = 0$ (flatness) and $\nabla F^p V \subseteq \Omega^1_{\mathbf{C}/\mathbf{Q}} \otimes_{\mathbf{C}} F^{p-1}V$ (Griffiths transversality) for all p.

A short exact sequence $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ of AHS (exact on each F^p) has extension class

$$e \in \mathrm{Ext}^1_{\mathrm{AHS}}(C,A) = H^1(\Omega^\bullet_{\mathbf{C}/\mathbf{Q}} \otimes_{\mathbf{C}} F^{-\bullet}\mathrm{Hom}_{\mathbf{C}}(C,A), \nabla_{\mathrm{Hom}_{\mathbf{C}}(C,A)}).$$

To obtain this, let $\phi \in F^0\mathrm{Hom}_{\mathbf{C}}(C,B)$ be a lifting of g. Now

$$g \circ \nabla_{\mathrm{Hom}_{\mathbf{C}}(C,B)}\phi = 0,$$

so

$$\nabla_{\mathrm{Hom}_{\mathbf{C}}(C,B)}\phi = f \circ e$$

for a unique $e \in F^{-1}\mathrm{Hom}_{\mathbf{C}}(C,A)$. The class of e in

$$H^1(\Omega^\bullet_{\mathbf{C}/\mathbf{Q}} \otimes_{\mathbf{C}} F^{-\bullet}\mathrm{Hom}_{\mathbf{C}}(C,A), \nabla_{\mathrm{Hom}_{\mathbf{C}}(C,A)})$$

is independent of the choice of ϕ.

A 2-step exact sequence $0 \to A \to B \to C \to D \to 0$ of AHS has a well-defined injective map from the Yoneda Ext

$$\mathrm{Ext}^2_{\mathrm{AHS}}(D,A) \to H^2(\Omega^\bullet_{\mathbf{C}/\mathbf{Q}} \otimes_{\mathbf{C}} F^{-\bullet}\mathrm{Hom}_{\mathbf{C}}(D,A), \nabla_{\mathrm{Hom}_{\mathbf{C}}(D,A)}).$$

This is obtained by composing the extension class of the two 1-step extensions $0 \to A \to B \to E \to 0$, $0 \to E \to C \to D \to 0$ it breaks into, and then using the natural map

$$\begin{aligned} &H^1(\Omega^\bullet_{\mathbf{C}/\mathbf{Q}} \otimes_{\mathbf{C}} F^{-\bullet}\mathrm{Hom}_{\mathbf{C}}(E,A), \nabla_{\mathrm{Hom}_{\mathbf{C}}(E,A)}) \otimes_{\mathbf{C}} \\ &H^1(\Omega^\bullet_{\mathbf{C}/\mathbf{Q}} \otimes_{\mathbf{C}} F^{-\bullet}\mathrm{Hom}_{\mathbf{C}}(D,E), \nabla_{\mathrm{Hom}_{\mathbf{C}}(D,E)}) \to \\ &H^2(\Omega^\bullet_{\mathbf{C}/\mathbf{Q}} \otimes_{\mathbf{C}} F^{-\bullet}\mathrm{Hom}_{\mathbf{C}}(D,A), \nabla_{\mathrm{Hom}_{\mathbf{C}}(D,A)}). \end{aligned}$$

This is exactly the obstruction to finding an AHS V with an additional increasing filtration $W_\bullet V$ by sub-AHS with $W_0 = 0$, $W_1 \cong A$, $W_2 \cong B$, $W_3 = V$, $V/W_1 \cong C$, and $V/W_2 \cong D$ and realizing the given 1-step extensions. The extension class is injective on $\mathrm{Ext}^2_{\mathrm{AHS}}(D,A)$, but it is not clear that all extension classes can occur.

This approach has much in common with the work of Carlson and Hain [CH].

This extension class theory fits in well with the pre-existing *arithmetic cycle class map* (see [EP] and work of Srinivas [Sr])

$$\eta\colon CH^p(X) \otimes_{\mathbf{Z}} \mathbf{Q} \to \mathbf{H}^{2p}(\Omega^{\geq p}_{X/\mathbf{Q}})$$

whose graded pieces are

$$\eta_m : \ker(\eta_{m-1}) \to H^m(\Omega^\bullet_{\mathbf{C}/\mathbf{Q}} \otimes_{\mathbf{C}} F^{p-\bullet} H^{2p-m}(X, \mathbf{C}), \nabla_{X/\mathbf{Q}});$$

one expects these to be consistent with the conjectural Bloch-Beilinson-Deligne-Murre filtration on $CH^p(X) \otimes_{\mathbf{Z}} \mathbf{Q}$. For 0-cycles on a surface, we are able to show that η_2 is the element of $\mathrm{Ext}^2_{\mathrm{AHS}}(H^2(X), 1)$ coming from the 2-step extension of AHS using $Z \subset C \subset X$ analogous to the construction above in the mixed Hodge structure case. A parallel construction was found independently by Asakura and Saito [AS], who have used it for some interesting geometric applications.

I would like to close by listing a few open problems that particularly appeal to me and which seem especially relevant to the next phase of the study of algebraic cycles.

(i) HODGE-THEORETIC FORMULA FOR $\nabla_{X/\mathbf{Q}}$

In cases of smooth projective varieties over $\mathbf{C}$ where Torelli's theorem holds (i.e. X is determined by Hodge-theoretic data on $H^*(X)$, at least theoretically $\nabla_{X/\mathbf{Q}}$ is determined on $H^i(X)$ by the Hodge structure of $H^i(X)$. It would be helpful to have a formula for this. Such a formula, involving Eisenstein series, was found by Katz [Ka1] for elliptic curves. For abelian varieties and K-3 surfaces, it would be very revealing to have a formula for $\nabla_{X/\mathbf{Q}}$. It would also be interesting to have an example where Torelli's theorem fails and $\nabla_{X/\mathbf{Q}}$ is different for two X's with the same Hodge structure; the alternative to this is the very attractive prospect that there is a general Hodge-theoretic formula for $\nabla_{X/\mathbf{Q}}$.

One facet of this question is the conjecture of Deligne (see [DMOS]), subordinate to the Hodge conjecture, that for X defined over $\mathbf{C}$, a Hodge class ξ necessarily satisfies

$$\nabla_{X/\mathbf{Q}} \xi = 0.$$

One possible "explanation" why this might be true is that a formula as alluded to above exists—such a formula would be expected to have the property that if $H^i(X) = H_1 \oplus H_2$ as Hodge structures, then H_1, H_2 would be $\nabla_{X/\mathbf{Q}}$-stable.

A related question is to ask whether $Gr^m CH^p(X) \otimes \mathbf{Q}$ is determined by the Hodge structure of $H^{2p-m}(X)$, or whether one definitely needs further information contained in the motive $h^{2p-m}(X)$, e.g $\nabla_{X/\mathbf{Q}}$.

(ii) $Gr^2 CH^2(A)$ FOR AN ABELIAN SURFACE A

If $\mathbf{Z}_A$ is the group ring of A with augmentation ideal J, then

$$S^2_{\mathbf{Z}} A \cong \frac{J^2}{J^3}$$

maps surjectively to $Gr^2 CH^2(A)$ by

$$a \otimes b \mapsto ((a) - (0)) * ((b) - (0)).$$

Thus

$$Gr^2 CH^2(A) = \frac{S^2_{\mathbf{Z}} A}{U}$$

for some subgroup $U \subset S^2_{\mathbf{Z}}A$. Describe U in terms of the Hodge structure of A. For $A = E_1 \times E_2$ a product of elliptic curves,

$$Gr^2CH^2(E_1 \times E_2) = \frac{E_1 \otimes_{\mathbf{Z}} E_2}{U'}$$

for some subgroup $U' \subseteq E_1 \otimes E_2$. Somekawa has given a description of U', but not in explicit Hodge-theoretic terms. The subgroups U, U' may be thought of as generalized Steinberg relations, as in the example given earlier of $Gr^2CH^2(\mathbf{P}^2, E)$. This is an excellent test case.

(iii) HIGHER REGULATORS FOR K-GROUPS

The *Borel regulator map*

$$r\colon K_3(\mathbf{C})^{\mathrm{ind}} = Gr_2K_3(\mathbf{C}) \to \mathbf{C}^*$$

is, by the work of Goncharov [Go] the Abel-Jacobi map

$$CH^2(\mathbf{P}^3, T_2)_{\mathrm{hom}} \to J^2(\mathbf{P}^3, T_2),$$

where T_2 is the tetrahedron $\{z_0z_1z_2z_3 = 0\}$. One should think of $(\mathbf{P}^3, T_2)$ as the analogue of a 3-fold with trivial canonical bundle. Conjecturally, r is injective when tensored with $\mathbf{Q}$. Its image has the same qualitative properties that Clemens showed the image of AJ^2_X possesses for the general quintic 3-fold—zero-dimensional, but not finitely generated even over $\mathbf{Q}$. One should think of r as a "toy model" model for the Abel-Jacobi map for codimension 2 cycles, in much the same was as $K_2(\mathbf{C})$ is the toy model for $Gr^2CH^2(X)$ for a surface with $H^{2,0} \neq 0$. The toy model for the higher Abel-Jacobi maps

$$Gr^mCH^p(X) \otimes \mathbf{Q} \to J^p_m(X)$$

should be maps, injective when tensored with $\mathbf{Q}$,

$$r^p_m\colon Gr_pK_{2p-m}(\mathbf{C}) \to \frac{\otimes^m_{\mathbf{Z}}\mathbf{C}^*}{U^p_m}$$

for some subgroup $U^p_m \subset \otimes^m_{\mathbf{Z}}\mathbf{C}^*$. For $m = 1$ these are the Borel regulators, while for $m = p$ they are the isomorphisms to Milnor K-theory. Can these maps, or something like them, be constructed?

(iv) EXPLICIT SUSLIN RECIPROCITY THEOREM

The Suslin Reciprocity Theorem [Su], used for example to compute $CH^2(\mathbf{P}^2, T)$ above, gives the vanishing of certain elements of $\wedge^m_{\mathbf{Z}}\mathbf{C}^*$ in $K_m(\mathbf{C})$, but does not explicitly produce the elements of the Steinberg ideal that makes them vanish. On some level, these are produced by the proof, but the *transfer map* or *norm map* N is not geometrically explicit. For example, given a rational curve $Y \subset \mathbf{P}^2$ and $f \in \mathbf{C}(Y)$ such that $f|_{Y\cap T} = 1$, we know not only that $\mathrm{div}(f) \in Z_0(\mathbf{C}^* \times \mathbf{C}^*)$ maps to 1 in $K_2(\mathbf{C})$, but in fact if we map it to $\wedge^2_{\mathbf{Z}}\mathbf{C}^*$, if $\mathrm{div}(f) = \sum_i n_ip_i$, it maps

to the product of the Steinberg symbols $(CR(p_i) \wedge (1 - CR(p_i)))^{n_i}$, where $CR(p_i)$ is the cross-ratio of p_i and one point each from the intersections of Y with each of the lines in T, with the product taken over all choices and all i (Goncharov,[Gre2]). For Y of higher genus, there is no comparably satisfying formula. In general, it would be nice to have as simple a version of N as possible for this type of geometric situation.

(v) DEFINITION OF $F^2CH^p(X) \otimes \mathbf{Q}$

Nori showed [No] that, for $p \geq 3$, $Z \equiv_{\mathrm{AJ}} 0$ does not imply $NZ \equiv_{\mathrm{alg}} 0$ for some $N > 0$. However, one might hope that for any p, $Z \equiv_{\mathrm{AJ}} 0$ implies that there exists a codimension 1 subvariety $Y \subset X$ such that $Z \subset Y$ and $NZ \equiv_{\mathrm{hom}} 0$ on Y for some $N > 0$. In many ways, a natural definition for $F^2CH^p(X) \otimes \mathbf{Q}$ is those Z such that both $NZ \equiv_{\mathrm{AJ}} 0$ and there exists a codimension 1 subvariety $Y \subset X$ such that $Z \subset Y$ and $NZ \equiv_{\mathrm{hom}} 0$ on Y for some $N > 0$. This conjecture would make the two plausible definitions of F^2 the same. It also fits in with what is needed to make the construction of ψ_2^2 go through for codimension 2 cycles in general. In particular, it would imply that for $p = 2$, $Z \equiv_{\mathrm{AJ}} 0$ implies $NZ \equiv_{\mathrm{alg}} 0$ for some $N > 0$.

The general conjecture would be that if $Z \in F^mCH^p(X) \otimes \mathbf{Q}$, then there exists a codimension 1 subvariety $Y \subset X$ such that $Z \subset Y$ and $Z \in F^{m-1}CH^{p-1}(Y) \otimes \mathbf{Q}$. This is what is needed to carry out the construction of higher Abel-Jacobi maps for arbitrary codimension.

REFERENCES

[AS] Asakura, M. and Saito, S., "Filtration on Chow groups and higher Abel-Jacobi maps," preprint.

[Bl] Bloch, Spencer, "An elementary presentation for K-groups and motivic cohomology," Motives (Seattle, WA, 1991), 239–244, Proc. Sympos. Pure Math., 55, Part 1, Amer. Math. Soc., Providence, RI, 1994.

[Ca] Carlson, James A., "Extensions of mixed Hodge structures," Journées de Géometrie Algébrique d'Angers, Juillet 1979/Algebraic Geometry, Angers, 1979, pp. 107–127, Sijthoff & Noordhoff, Alphen aan den Rijn—Germantown, Md., 1980.

[CH] Carlson, James A.; Hain, Richard M., "Extensions of variations of mixed Hodge structure," Actes du Colloque de Théorie de Hodge (Luminy 1987), Astérisque (1989), no. 179-180, 9, 39-65.

[Cl] Clemens, H., "Homological equivalence, modulo algebraic equivalence, is not finitely generated," Inst. Hautes Études Sci. Publ. Math. No. 58 (1983), 19–38 (1984).

[DMOS] Deligne, Pierre; Milne, James S.; Ogus, Arthur; Shih, Kuang-yen Hodge cycles, motives, and Shimura varieties. Lecture Notes in Mathematics, 900. Philosophical Studies Series in Philosophy, 20. Springer-Verlag, Berlin-New York, 1982.

[EP] Esnault, Hélène; Paranjape, Kapil H., "Remarks on absolute de Rham and absolute Hodge cycles," C. R. Acad. Sci. Paris Sér. I Math. 319 (1994), no. 1, 67–72.

[Go] Goncharov, A. B., "Chow polylogarithms and regulators," Math. Res. Lett. 2 (1995), no. 1, 95–112.
[GL] Goncharov, A. and Levin, A., "Zagier's conjecture on $L(E, 2)$," preprint.
[Gre1] Green, Mark L., "Griffiths' infinitesimal invariant and the Abel-Jacobi map," J. Differential Geom. 29 (1989), no. 3, 545–555.
[Gre2] Green, Mark L., "'What comes after the Abel-Jacobi map?," preprint.
[GG] Green, Mark and Griffiths, Phillip, in preparation.
[Gri] Griffiths, Phillip, "On the periods of certain rational integrals I, II," Ann. of Math. (2) 90 (1969), 460-495; ibid. (2) 90 (1969), 496–541.
[Gro] Grothendieck, A., "On the de Rham cohomology of algebraic varieties," Inst. Hautes Études Sci. Publ. Math. No. 29 (1966), 95–103.
[Ja] Jannsen, Uwe, "Motivic sheaves and filtrations on Chow groups," Motives (Seattle, WA, 1991), 245–302, Proc. Sympos. Pure Math., 55, Part 1, Amer. Math. Soc., Providence, RI, 1994.
[Ka1] Katz, Nicholas M., "p-adic interpolation of real analytic Eisenstein series," Ann. of Math. (2) 104 (1976), no. 3, 459–571.
[Ka2] Katz, Nicholas M., "Nilpotent connections and the monodromy theorem: Applications of a result of Turrittin," Inst. Hautes Études Sci. Publ. Math. No. 39 (1970), 175–232.
[Mu] Mumford, D., "Rational equivalence of 0-cycles on surfaces," J. Math. Kyoto Univ. 9 (1968), 195–204.
[No] Nori, Madhav V., "Algebraic cycles and Hodge-theoretic connectivity," Invent. Math. 111 (1993), no. 2, 349–373.
[Ro] Roĭtman, A. A., "Rational equivalence of zero-dimensional cycles," (Russian) Mat. Zametki 28 (1980), no. 1, 85–90, 169.
[Sa] Saito, Shuji, "Motives and filtrations on Chow groups," Invent. Math. 125 (1996), no. 1, 149–196.
[Sr] Srinivas, V., "Gysin maps and cycle classes for Hodge cohomology," Proc. Indian Acad. Sci. Math. Sci. 103 (1993), no. 3, 209-247.
[Su] Suslin, A. A., "Reciprocity laws and the stable rank of rings of polynomials," (Russian) Izv. Akad. Nauk SSSR Ser. Mat. 43 (1979), no. 6, 1394–1429.
[To] Totaro, Burt, "Milnor K-theory is the simplest part of algebraic K-theory," K-Theory 6 (1992), no. 2, 177–189.
[Vo1] Voisin, Claire, "Une remarque sur l'invariant infinitésimal des fonctions normales," C. R. Acad. Sci. Paris Sér. I Math. 307 (1988), no. 4, 157–160.
[Vo2] Voisin, Claire, "Some results on Green's higher Abel-Jacobi map," preprint.

Mark L. Green
Department of Mathematics
U.C.L.A.
Los Angeles, CA. 90095
USA
mlg@math.ucla.edu

Operads and Algebraic Geometry

M. Kapranov

Abstract. The study (motivated by mathematical physics) of algebraic varieties related to the moduli spaces of curves, helped to uncover important connections with the abstract algebraic theory of operads. This interaction led to new developments in both theories, and the purpose of the talk is to discuss some of them.

1991 Mathematics Subject Classification: 14H10, 18C15, 08C9
Keywords and Phrases: operads, superpositions, moduli spaces.

1. Operads.

The modern concept of an operad originated in topology [25] but the underlying ideas can be traced back at least to Hilbert's 13th problem which (in the way it came to be understood later) can be stated as follows.

(1.1) Let P be some class of functions considered in analysis (e.g., continuous, smooth, algebraic etc.). Is it possible to express any function $f : \mathbf{R}^n \to \mathbf{R}$ of class P, depending of $n \geq 3$ variables, as a superposition of functions of class P, depending on 1 or 2 variables only?

This question involves the *superposition operation* on functions of several variables: we can substitute some n functions $g_1(x_{1,1}, ..., x_{1,a_1})$, ..., $g_n(x_{n,1}, ..., x_{n,a_n})$ for arguments of another function $f(x_1, ..., x_n)$, thereby getting a new function $f(g_1, ..., g_n)$ depending on $a_1 + ... + a_n$ variables. (Here the x_{ij}, as well as the values of the functions, belong to some fixed set X.) This generalizes the observation that functions of one variable can be composed: $(f(x), g(x)) \mapsto f(g(x))$.

For a set X, maps $X^n \to X$ are also called n-ary operations on X. A collection $\mathcal{P}$ of such maps (with possibly varying n) is called an *operad*, if it is closed under arbitrary superpositions as well as under permutations of variables. Thus $\mathcal{P}(n)$, the n-ary part of $\mathcal{P}$, is acted upon by the symmetric group S_n. One can view $\mathcal{P}$ as a multivariable analog of a (semi)group of transformations of X.

As with groups, the actual working definition [25] splits the above naive one into two. First, one defines an abstract operad $\mathcal{P}$ as a collection of sets $\mathcal{P}(n), n \geq 0$ with S_n acting on $\mathcal{P}(n)$, equipped with an element $1 \in \mathcal{P}(1)$ (the unit) and maps

$$\mathcal{P}(n) \times \mathcal{P}(a_1) \times ... \times \mathcal{P}(a_1) \to \mathcal{P}(a_1, ..., a_n) \tag{1.2}$$

satisfying the natural associativity and equivariance conditions, as well as the conditions for the unit, see [25]. Then, one defines a $\mathcal{P}$-algebra as a set X together

with S_n-equivariant maps $\mathcal{P}(n) \to \mathrm{Hom}(X^n, X)$ which take (1.2) into the actual superpositions of operations.

It was soon noticed that one can replace sets in the above definitions with objects of any symmetric monoidal category $(\mathcal{C}, \otimes)$, by using $\otimes$ instead of products of sets. The categories $\mathcal{C}$ used in practice include two groups of examples:

(1.3) $\mathcal{T}op$ (topological spaces), $\mathcal{V}ar$ (algebraic varieties, say over $\mathbf{C}$), $\mathcal{S}t$ (algebraic stacks), with $\otimes$ being the Cartesian product.

(1.4) $\mathcal{V}ect_k$ (vector spaces over $\mathbf{C}$), $dg\mathcal{V}ect_k$ (dg-vector spaces, i.e., bounded cochain complexes of finite-dimensional spaces), $g\mathcal{V}ect_k$ (graded spaces, i.e., complexes with zero diferential), with $\otimes$ being the tensor product. Operads in these categories are called, respectively, linear, dg- or graded operads.

If $\mathcal{P}$ is an operad in $\mathcal{T}op$ or $\mathcal{V}ar$, then the topological homology spaces $H_\bullet(\mathcal{P}(n), \mathbf{C})$ form a graded operad denoted $H_\bullet(\mathcal{P})$. Similarly, the chain complexes of the $\mathcal{P}(n)$ form a dg-operad.

2. Role of operads in algebraic geometry.

The source for the recent interest in operads in algebraic geometry is the synthesis of several different approaches, which we recall.

A. Abstract-algebraic approach. Many familiar algebraic structures can be described as algebras over appropriate operads. This use of operads is the traditional approach of "universal algebra".

Example 2.1. Let $\mathcal{G}r(n)$ be the Weyl group of the root system B_n, i.e., the semidirect product of S_n and $\{\pm 1\}^n$; for $n = 0$ set $\mathcal{G}r(0) := \{\mathrm{pt}\}$. The collection of the $\mathcal{G}r(n)$ can be made into an operad $\mathcal{G}r$ (in the category of sets) so that any group G is naturally a $\mathcal{G}r$-algebra. The maps $G^n \to G$ corresponding to elements of $\mathcal{G}r(n)$, are of the form $(x_1, ..., x_n) \mapsto x_{\sigma(1)}^{\epsilon_1} ... x_{\sigma(n)}^{\epsilon_n}$, $\sigma \in S_n$, $\epsilon_i \in \{\pm 1\}$. More generally, an arbitrary $\mathcal{G}r$-algebra is the same as a semigroup with involution $*$ satisfying $(ab)^* = b^* a^*$.

Example 2.2. We have linear operads $\mathcal{A}s, \mathcal{C}om, \mathcal{L}ie$ whose algebras (in the category $\mathcal{V}ect$) are respectively, associative, commutative or Lie algebras. Explicitly, $\mathcal{L}ie(n)$ is the subspace in the free Lie algebra on $x_1, ..., x_n$ formed by elements multihomogeneous of degrees (1, ..., 1), and similarly for the other classes of algebras, see [11]. For example, each $\mathcal{C}om(n) \simeq \mathbf{C}$ while $\dim(\mathcal{A}s(n)) = n!$.

B. Algebro-geometric examples. These examples are elaborations on the idea of gluing Riemann surfaces, present in the string theory for a long time [30]. But the operadic approach to this idea is surprisingly useful.

Example 2.3. Let $\overline{M}_{g,n}$ be the moduli stack of stable n-pointed curves of genus g, see [2]. Set $\overline{\mathcal{M}}_0(n) = \overline{M}_{0,n+1}$ and $\overline{\mathcal{M}}(n) = \coprod_g \overline{M}_{g,n+1}$. The $\overline{\mathcal{M}}_0(n)$, $n \geq 2$ are in fact algebraic varieties. The $n+1$ marked points on $C \in \overline{\mathcal{M}}(n)$ are denoted by $(x_0, ..., x_n)$ and the group S_n acts by permutations of $x_1, ..., x_n$. The collection of $\overline{\mathcal{M}}(n), n \geq 2$, is naturally made into an operad $\overline{\mathcal{M}}$ in $\mathcal{S}t$ while $\overline{\mathcal{M}}_0 = \{\overline{\mathcal{M}}_0(n)\}$ forms an operad in $\mathcal{V}ar$. The maps (1.2) take $(C, D_1, ..., D_n)$ into the reducible curve obtained by identifying the 0th point of D_i with the ith point of C.

EXAMPLE 2.4. Denote by $\widetilde{M}_{g,n}$ be the set of isomorphism classes of Riemann surfaces of genus g with boundary consisting of n circles, together with a smooth identification of each boundary component with S^1. It is naturally an infinite-dimensional topological space. As before we set $\widetilde{\mathcal{M}}_0(n) = \widetilde{M}_{0,n+1}$, $\widetilde{\mathcal{M}}(n) = \coprod_g \widetilde{M}_{g,n+1}$. Gluing Riemann surfaces together along the boundary makes $\widetilde{M}$ and $\widetilde{M}_0$ into operads in the category of topological spaces.

Note that the above two examples are in some sense, dual to each other, as $\widetilde{M}_{g,n+1}$ projects naturally to the open stratum (the locus of smooth curves) $M_{g,n+1} \subset \overline{M}_{g,n+1}$. When $g = 0$, the complement to $M_{0,n+1}$ is precisely formed by the images of the maps (1.2).

The above does not of course exhaust all types of operads of algebro-geometric nature. A quite different class of examples was developed in [20].

C. RELATIONS OF A AND B. By taking homology of operads from B, we get graded operads which, remarkably, are related to the operads from A.

Each $\overline{\mathcal{M}}_0(n)$ is a smooth irreducible projective variety of dimension $n-2$. Let $q_n \in H_{2(n-2)}(\overline{\mathcal{M}}_0(n), \mathbf{C})$ be the fundamental class. It is clear that q_2 generates the suboperad $H_0(\overline{\mathcal{M}}_0) \subset H_\bullet(\overline{\mathcal{M}}_0)$ isomorphic to $\mathcal{C}om$. Thus an $H_\bullet(\overline{\mathcal{M}}_0)$-algebra is a commutative algebra with extra structure, and this extra structure was described explicitly by M. Kontsevich and Y.I. Manin [19].

THEOREM 2.5. *An $H_\bullet(\overline{\mathcal{M}}_0)$-algebra is the same as a graded vector space A with multilinear totally symmetric (in the graded sense) operations $(x_1, ..., x_n)$, $n \geq 2$ of degree $2(n-2)$ satisfying the generalized associativity conditions:*

$$\sum_{[n]=S_1 \coprod S_2} \pm((a, b, x_{S_1}), c, x_{S_2}) \quad = \quad \sum_{[n]=S_1 \coprod S_2} \pm(a, (b, c, x_{S_1}), x_{S_2}), \tag{2.6}$$

where $x_S, S \subset [n]$ means the unordered set of $x_i, i \in S$ and $\pm$ is given by the Koszul sign rules. In particular, (x_1, x_2) is a commutative associative multiplication.

The condition (2.6) is the well known WDVV relation. The theory of Gromov-Witten invariants such as it was developed in [1,2, 19, 24] gives the following fact.

THEOREM 2.7. *For any smooth projective variety V the homology space $H_\bullet(V, \mathbf{C})$ has a natural structure of an algebra over the operad $H_\bullet(\overline{\mathcal{M}})$, in particular, it has the structure specified in Theorem 2.5.*

By contrast, the homology of the open moduli spaces is related to the operad $\mathcal{L}ie$ and its generalizations. First, an old result of F. Cohen [3] implies that $H_{\max} M_{0,n+1} \simeq \mathcal{L}ie(n) \otimes \mathrm{sgn}_n$, as an S_n-module. More generally, E. Getzler [8] defined an operad structure on the collection of $\mathcal{G}(n) = H_\bullet(M_{0,n+1}, \mathbf{C})[2-n] \otimes \mathrm{sgn}_n$ by using the Poincaré residue maps and proved the following.

THOREM 2.8. *A $\mathcal{G}$-algebra is the same as a graded vector space A together with totally antisymmetric products $[x_1, ..., x_n]$ of degree $2-n$, $n \geq 2$, satisfying the generalized Jacobi identities:*

$$\sum_{i \leq i < j \leq k} \pm\big[[a_i, a_j], a_1, ..., \widehat{a}_i, ..., \widehat{a}_j, ..., a_k, b_1, ..., b_l\big]$$

$$= \begin{cases} \big[[a_1, ..., a_k], b_1, ...b_l\big], & \text{if } l > 0; \\ 0, & l = 0. \end{cases}$$

In particular, $(A, [x_1, x_2])$ *is a graded Lie algebra.*

D. OPERADS AND TREES. Maps (1.2) represent a single instance of superposition of elements of an operad (e.g., functions in several variables). By applying them several times, we get *iterated superpositions*, which are best described by their "flow charts", similar to those used by computer programmers.

More precisely, call an n-tree a tree T with $n+1$ external edges which are divided into n "inputs" and one output and such that the inputs are numbered by $1, ..., n$. Such T has orientation according to the flow from the inputs to the output; in particular, the edges adjacent to any vertex v, are separated into two subsets $\mathrm{In}(v)$ and $\mathrm{Out}(v)$, the latter consisting of one element. Given an operad $\mathcal{P}$, the S_m-action on $\mathcal{P}(m)$ allows us to speak about sets $\mathcal{P}(I)$, where I is any m-element set (any identification $I \to [m]$ identifies $\mathcal{P}(I)$ with $\mathcal{P}(m)$). Now, a flow chart with n inputs for $\mathcal{P}$ is an n-tree T together with assignment, for any vertex v of T, of an element $q_v \in \mathcal{P}(\mathrm{In}(v))$. This data produces an iterated superposition of the q_v, belonging to $\mathcal{P}(n)$.

So the combinatorics of trees is closely connected to all questions related to operads and superpositions. It is worth pointing out that the first paper of Kolmogoroff [16] on the Hilbert superposition problem for continuous functions used trees in an essential way.

On the other hand, J. Harer and R. Penner [12, 26] constructed a cell decomposition of the moduli space of curves in which cells are parametrized by graphs (with some extra structure). This led M. Kontsevich [17] to introduce certain purely combinatorial chain complexes formed of summands labelled by graphs, which turned out to be very important in such diverse questions as quasiclassical approximation to the Chern-Simons invariant of 3-manifolds and cohomology of infinite-dimensional Lie algebras.

An observation of V. Ginzburg and the author [11] was that the tree parts of Kontsevich's graph complexes can be very easily interpreted and generalized in the language of operads.

More precisely, if $\mathcal{A} = \mathcal{A}(n), n \geq 2$ is any collection of dg-vector spaces with S_n-actions, the *free operad* $F_{\mathcal{A}}$ generated by $\mathcal{A}$ consists of all possible formal iterated superpositions of elements of $\mathcal{A}$, i.e.,

$$F_{\mathcal{A}}(n) = \bigoplus_{n-\text{trees } T} \bigotimes_{v \in \mathrm{Vert}(T)} \mathcal{A}(\mathrm{In}(v)).$$

(A similar definition can also be given for operads in (1.3) if we replace $\bigoplus$ with $\coprod$ and $\bigotimes$ with the Cartesian product.)

For a collection $\mathcal{A}$ as above its suspension $\Sigma\mathcal{A}$ constists of shifted complexes with twisted S_n-action: $(\Sigma\mathcal{A})(n) = \mathcal{A}(n)[n-1] \otimes \mathrm{sgn}_n$. (Meaning: if $\mathcal{A}$ is in fact an operad and A is an $\mathcal{A}$-algebra, then the $A[1]$ is a $\Sigma\mathcal{A}$-algebra.)

THEOREM 2.9. *(a) Let* $\mathcal{P}$ *be a dg-operad with* $\mathcal{P}(0) = 0$, $\mathcal{P}(1) = \mathbf{C}$ *and* $\mathcal{P}^*$ *be the collection of the dual dg-spaces* $\mathcal{P}(n)^*, n \geq 2$. *Then the components of the free*

operad $F_{\Sigma\mathcal{P}^*}$ admit natural differentials with respect to which they form a new dg-operad $\mathbf{D}(\mathcal{P})$ called the cobar-dual to $\mathcal{P}$. We have a canonical quasiisomorphism $\mathbf{D}(\mathbf{D}(\mathcal{P})) \to \mathcal{P}$.
(b) We have quasi-isomorphisms $\mathbf{D}(\mathcal{A}s) \simeq \mathcal{A}s$, $\mathbf{D}(\mathcal{C}om) \simeq \mathcal{L}ie$, $\mathbf{D}(\mathcal{L}ie) \simeq \mathcal{C}om$.

3. Koszul duality

The functor $\mathbf{D}$ from Theorem 2.9 can be viewed as a kind of cohomology theory on the category of operads. In particular, when $\mathcal{P}$ is just a linear operad (has trivial dg-structure), $H^\bullet(\mathbf{D}(\mathcal{P}(n))$ provides information about generators, relations and higher syzygies of $\mathcal{P}$. There is a class of operads for which $\mathbf{D}(\mathcal{P})$ is especially simple.

In [11], a linear operad $\mathcal{P}$ with $\mathcal{P}(0) = 0, \mathcal{P}(1) = \mathbf{C}$, was called *quadratic*, if the following conditions hold:

(3.1) $\mathcal{P}$ is generated by the binary part, i.e., every element of every $\mathcal{P}(n)$ is a sum of iterated superpositions of elements of $\mathcal{P}(2)$, so that the morphism $\phi_{\mathcal{P}} : F_{\mathcal{P}(2)} \to \mathcal{P}$ is surjective.

(3.2) All the relations among the binary generators follow from those holding in the ternary part, i.e., the "ideal" $\mathrm{Ker}(\phi_{\mathcal{P}})$ is generated by $\mathrm{Ker}(\phi_{\mathcal{P}}(3) : F_{\mathcal{P}(2)}(3) \to \mathcal{P}(3))$.

Thus a quadratic operad $\mathcal{P}$ can be described by giving a vector space $V = \mathcal{P}(2)$ of generators, equipped with S_2-action and an S_3-invariant subspace of relations $R \subset \mathcal{F}_V(3)$. We will write $\mathcal{P} = Q(V, R)$. The *Koszul dual* operad $\mathcal{P}^!$ is defined as $\mathcal{P}^! = Q(V^* \otimes \mathrm{sgn}, R^\perp)$. This is a natural analog of Koszul duality for algebras as defined by Priddy [27].

Theorem 3.3. *The operads $\mathcal{A}s$, $\mathcal{C}om$, $\mathcal{L}ie$ are quadratic, and their Koszul duals are: $\mathcal{A}s^! = \mathcal{A}s$, $\mathcal{C}om^! = \mathcal{L}ie$, $\mathcal{L}ie^! = \mathcal{C}om$.*

The duality between commutative and Lie algebras, as a meta-mathematical principle, goes back at least to the work of D. Quillen on rational homotopy theory. Later, V. Drinfeld suggested to look for some tangible reasons behind this principle. The explanation provided by Theorem 3.3 is so far the most elementary: it exhibits the sought-for "reason" as the fact that certain given subspaces of given dual vector spaces are orthogonal complements of each other.

For any quadratic operad $\mathcal{P}$ there is a natural morphism of dg-operads $\mathbf{D}(\mathcal{P}) \to \mathcal{P}^!$, and $\mathcal{P}$ is called Koszul if this is a quasiisomorphism. Thus, Theorem 2.9(b) implies that the operads $\mathcal{A}s$, $\mathcal{C}om$ and $\mathcal{L}ie$ are Koszul. Similarly to Koszul quadratic algebras of Priddy [27], Koszul operads possess many nice properties allowing one to calculate the homological invariants in an elementary way.

A generalization of the theory of quadratic and Koszul operads to the case when $\mathcal{P}$ is not necessarily generated by $\mathcal{P}(2)$, was developed by E. Getzler [8]. In this case, the meaning of "quadratic" is that all the relations follow from those involving only the simplest instances of superposition of the generators. It was proved in [8] that in this more general sense, the operads $H_\bullet(\overline{\mathcal{M}}_0)$ and $\mathcal{G}$ from Theorems 2.5 and 2.8 are quadratic, Koszul and dual to each other.

4. Cyclic and modular operads.

The algebro-geometric examples of operads from §2B in fact possess more than just an operad structure. First, the division of the $n+1$ marked points (or boundary components) inth n inputs and one output is artificial procedure. So even though, say, $\overline{\mathcal{M}}_0(n)$ is the n-ary part of an operad, we have an action of $S_{n+1} = \mathrm{Aut}\{0,...,n\}$ on it. Motivated by this, E. Getzler and the author [9] called a cyclic operad an operad $\mathcal{P}$ together with S_{n+1}-action on each $\mathcal{P}_n$ which is compatible with compositions in the following sense. Denote by τ_n the cycle $(0,1,...,n) \in S_{n+1}$. Then it is required that $\tau_1(1) = 1$, and that

$$\tau_{m+n-1}(p(1,...,1,q)) = (\tau_n q)(\tau_m(p),1,...,1), \quad p \in \mathcal{P}(m), q \in \mathcal{P}(n).$$

So all the examples from §2*B* are cyclic operads in this sense.

It is not obvious that this concept should have any meaning from the point of view of §2A, but it does. Let $\mathcal{P}$ be a linear (or dg-) operad. It was shown in [9] (generalizing some observations of M. Kontsevich), that a cyclic structure on $\mathcal{P}$ is precisely the data necessary to meaningfully speak about *invariant scalar products* on $\mathcal{P}$-algebras. For example, a scalar product B on a Lie algebra is called invariant if $B([x,y],z) = B(x,[y,z])$, and similarly for the other types of algebras from Example 2.2. This indicates (and this is indeed the case) that these operads are cyclic. The operads from Theorems 2.5 and 2.8 are cyclic too. For a cyclic operad $\mathcal{P}$ it is notationally convenient to denote the S_n-module $\mathcal{P}(n-1)$ by $\mathcal{P}(\!(n)\!)$, thereby emphasizing the symmetry between the inputs and the output. We will also call a cyclic $\mathcal{P}$-algebra a pair (A,g) consisting of an algebra and an invariant scalar product. Theorems 2.5 and 2.7 have an even nicer formulation in terms of cyclic operads (see [24] for the detailed proof of (a)).

THEOREM 4.1. *(a) A finite dimensional (graded) cyclic algebra over $H_\bullet(\overline{\mathcal{M}}_0)$ is the same as a formal germ of a potential Frobenius (super-)manifold in the sense of [5, 24].*
(b) For any smooth projective variety V the intersection pairing g on $H_\bullet(V,\mathbf{C})$ makes it into a cyclic $H_\bullet(\overline{\mathcal{M}})$-algebra.

If, in §2B, we consider moduli spaces of curves of arbitrary genus, then there is still another structure present: two marked points (or boundary components) *of the same curve* can be glued together, producing a curve of genus higher by 1 and number of marked points (or boundary components) less by 2. This structure was axiomatized in [10] under the name "modular operad".

Explicitly, a modular operad $\mathcal{P}$ (in a monoidal category $\mathcal{C}$) is a collection of objects $\mathcal{P}(\!(g,n)\!)$ given for $n, g \geq 0$ such that $2g-2+n > 0$ (the number g is called genus), with S_n acting on $\mathcal{P}(\!(g,n)\!)$ and the following data:

(4.2) A structure of a cyclic operad on the collection of $\mathcal{P}(\!(n)\!) = \coprod_g \mathcal{P}(\!(g,n)\!)$ so that the genus of any superposition is equal to the sum of the genera of the elements involved. (For categories from (1.4) we should understand $\coprod$ as the direct sum). We can speak of $\mathcal{P}(\!(g,I)\!)$ for $|I| = n$, via the S_n-action.

(4.3) The contraction maps $\xi_{i,j} : \mathcal{P}(\!(g, I)\!) \to \mathcal{P}(\!(g+1, I-\{i,j\})\!)$ given for any finite set I, $i \neq j \in I$, satisfying natural equivariance and coherence conditions ([10], §3).

A "flow chart" for a modular operad $\mathcal{P}$ is given by a *stable n-graph*, that it, a connected graph G with some number n of external legs (edges not terminating in a vertex) which are numbered by $1, ..., n$, plus an assignment, to any vertex v, of a number $g(v) \geq 0$ so that $2g(v) - 2 + n(v) > 0$, where $n(v)$ is the valence of v. The (total) genus of a stable n-graph G is defined as $\sum_v g(v)$ plus the first Betti number of G. Let $\Gamma(\!(g, n)\!)$ be the set of isomorphism classes of stable n-graphs of genus g. For $G \in \Gamma(\!(g, n)\!)$ set $\mathcal{P}(\!(G)\!) = \bigotimes_{v \in \mathrm{Vert}(G)} \mathcal{P}(g(v), \mathrm{Ed}(v))$, where $\mathrm{Ed}(v)$ is the set of (half-)edges issuing from v. Then a modular operad structure on $\mathcal{P}$ gives the superposition map $\mathcal{P}(\!(G)\!) \to \mathcal{P}(\!(g, n)\!)$.

In the dg-framework there is a related concept of a *twisted modular operad* where we require superposition maps of the form $\mathcal{P}(\!(G)\!) \otimes \mathrm{Det}(\mathbf{C}^{\mathrm{Ed}(G)}) \to \mathcal{P}(\!(g, n)\!)$, where $\mathrm{Ed}(G)$ is the set of all edges of G and $\mathrm{Det}(V) = \Lambda^{\dim(V)}(V)[\dim(V)]$. As was shown in [10], the following generalization of the cobar-duality to modular operads encompasses Kontsevich's graph complexes in full generality.

Theorem 4.4. *For a modular dg-operad $\mathcal{P}$ the collection of*

$$F(\mathcal{P})(\!(g, n)\!) = \bigoplus_{G \in \Gamma(\!(g,n)\!)} \left(\mathrm{Det}(\mathbf{C}^{\mathrm{Ed}(G)}) \otimes \bigotimes_{v \in \mathrm{Vert}(G)} \mathcal{P}(\!(g, v)\!)^* \right)_{\mathrm{Aut}(G)}$$

has natural differentials and composition maps which make it into a twisted modular dg-operad $F(\mathcal{P})$ called the Feynman transform of $\mathcal{P}$.
(b) The functor $\mathcal{P}$ takes quasiisomorphisms to quasiisomorphisms and gives an equivalence between the derived categories of modular dg-operads and twisted modular dg-operads.

The inverse to F is constructed similarly to F but with a different determinantal twist.

Example 4.5. Let $\mathcal{P} = \mathcal{A}s$ is the associative operad considered as a modular operad, i.e., $\mathcal{A}s(\!(g, n)\!) = 0, g > 0$, $\mathcal{A}s(\!(0, n)\!) = \mathcal{A}s(n-1)$. Then

$$(F\mathcal{A}s)(\chi, 0) = \bigoplus_{2g-2+n=\chi} C_\bullet(|M_{g,n}|/S_n, \mathbf{C}),$$

where $|M_{g,n}|$ is the coarse moduli space of smooth curves of genus g with n punctures, and $C_\bullet$ is the chain complex with respect to Penner's cell decomposition labelled by "fat graphs", i.e., graphs with a cyclic order on each $\mathrm{Ed}(v)$. The reason is that $\mathcal{A}s(\!(0, n)\!) = \mathcal{A}s(n-1)$, as an S_n-module, can be identified with the vector space spanned by all cyclic orders on $\{1, ..., n\}$.

One of the main results of [10] is the determination of the Euler characteristics of the $F(\mathcal{P})(\!(g, n)\!)$ (as elements of the representation ring of S_n) in terms of those of $\mathcal{P}(\!(g, n)\!)$. The set of $\chi(\mathcal{P}(\!(g, n)\!))$ is encoded into a formal power series $C_{\mathcal{P}}(h, p_1, p_2, ...)$ of infinitely many variables, and $C_{F(\mathcal{P})}$ is identified with a certain formal Fourier transform of $C_{\mathcal{P}}$ with respect to a Gaussian measure on $\mathbf{R}^\infty$. In

the case when $\mathcal{P} = \mathcal{A}s$, the infinite-dimensional integral in the Fourier transform can be calculated explicitly by separating the variables, and we have the following theorem.

THEOREM 4.4. *The series*

$$\Psi(h) = \sum_{\chi=1}^{\infty} h^{\chi} \sum_{2g-2+n=\chi} e(|M_{g,n}|/S_n),$$

where e is the topological Euler characteristic, is calculated as follows:

$$\Psi(h) = \sum_{n,l=1}^{\infty} \frac{\mu(l)}{l} \Psi_n(h^l), \quad \text{where}$$

$$\Psi_n(h) = \sum_{k=1}^{\infty} \frac{\zeta(-k)}{-k} \alpha_n^{-k} + (\alpha_n + \frac{1}{2}) \ln(nh^n\alpha_n) - \alpha_n + \frac{1}{nh^n} - c(n)/2n,$$

$$\alpha_n = \alpha_n(h) = \frac{1}{n} \sum_{d|n} \frac{\phi(d)}{h^{n/d}}, \quad c_n = \frac{1}{2}(1 + (-1)^n)$$

and ϕ, μ, ζ are respectively, the Euler, Möbius and Riemann zeta functions.

Recall [13] that the *orbifold* Euler characteristic of $M_{g,1}$ is equal to the rational number $\zeta(1-2g)$.

5. Operads and curvature invariants.

The operadic point of view turned out to be useful even in such "classical" parts of geometry as the theory of characteristic classes. Let M be a complex manifold and $T = TM$ its tangent bundle. Then the $\mathcal{E}_M(n) = \{\mathrm{Hom}(T^{\otimes n}, T)\}$ form an operad in the category of holomorphic vector bundles on M and hence $\mathbf{E}_M = \{\mathbf{E}_M(n) = H^{\bullet}(M, \mathcal{E}_M(n))\}$ (holomorphic cohomology) is a graded operad. On the other hand, the curvature of any Hermitian metric h on M defines a Dolbeault cohomology class $\alpha_M \in H^1(M, \mathrm{Hom}(T^{\otimes 2}, T)) = \mathbf{E}_M(2)^1$ (the Atiyah class). This class is symmetric with respect to the S_2-action on $\mathbf{E}_M(2)$ (when h is Kähler, even the curvature form is symmetric). Consider the desuspension $\Sigma^{-1}(\alpha_M)$ (see §2) which is an element of $\Sigma^{-1}(\mathbf{E}_M)(2)^0$ anti-symmetric with respect to S_2. The following fact, inspired by [29, 18], was proved in [14].

THEOREM 5.1. *The element $\Sigma^{-1}(\alpha_M)$ satisfies the Jacobi identity in the operad $\Sigma^{-1}(\mathbf{E}_M)$, i.e., it defines a morphism of operads $\mathcal{L}ie \to \Sigma^{-1}(\mathbf{E}_M)$.*

One can say that algebraic geometry is based on the operad $\mathcal{C}om$, governing commutative associative algebras. It is a natural idea to develop some generalized geometries based on more general linear operads $\mathcal{P}$. For example, for $\mathcal{P} = \mathcal{A}s$, we get "noncommutative geometry" based on associative but not necessarily commutative algebras, and several important approaches to such geometry have been developed [4, 7, 23, 28]. The case of a general $\mathcal{P}$ presents of course, even more difficulties, but Theorem 5.1 suggests the following heuristic principle which is confirmed whenever $\mathcal{P}$-geometry can be given sense:

Pre-theorem 5.2. *Let $\mathcal{P}$ be a Koszul operad. Then, curvature invariants of a "space" in $\mathcal{P}$-geometry satisfy the constraints of the Koszul dual operad $\mathcal{P}^!$.*

Example 5.3: Formal $\mathcal{P}$-geometry. We can always speak about $D^n_{\mathcal{P}}$, the *formal n-disk* in $\mathcal{P}$-geometry, i.e., the object corresponding to the completion of the free $\mathcal{P}$-algebra on n generators, cf. [17]. Being infinitesimal, it does not by itself possess global curvature invariants. However, the question becomes interesting for group structures on $D^n_{\mathcal{P}}$, i.e., n-dimensional formal groups over $\mathcal{P}$. Such groups were studied by M. Lazard [21] who, in a work pre-dating the modern concept of an operad, developed an analog of Lie theory for them. A modern interpretation of his theory [6, 11] revealed that the analog of a Lie algebra for a formal group over $\mathcal{P}$ is in fact a $\mathcal{P}^!$-algebra.

Example 5.4: Semiformal $\mathcal{A}s$-geometry. In [15], the author developed a formalism of "noncommutative formal neighborhoods" (called NC-thickenings) of a smooth algebraic variety M. They are ringed spaces $X = (M, \mathcal{O}_X)$ where $\mathcal{O}_X$ is a sheaf of noncommutative rings with $\mathcal{O}_X/[\mathcal{O}_X, \mathcal{O}_X] = \mathcal{O}_M$ and such that its completion at any point is isomorphic to $\mathbf{C}\langle\langle x_1, ..., x_n\rangle\rangle$, the algebra of noncommutative formal power series.

Theorem 5.5. *Let M be a smooth algebraic variety. Then any NC-thickening X of M has a characteristic class $\alpha_X^- \in H^1(M, \Omega^2_M \otimes T)$. The sum $\alpha_X = \alpha_M + \alpha_X^- \in H^1(M, \mathrm{Hom}(T^{\otimes 2}, T))$ is such that $\Sigma^{-1}\alpha_X \in \Sigma^{-1}(\mathbf{E}_M)(2)^0$ is an associative element, i.e., gives rise to a morphism of operads $\mathcal{A}s \to \Sigma^{-1}\mathbf{E}_M$. Moreover, α_X is an A_∞-element in the sense of Stasheff [31].*

References

[1] K. Behrend, Gromov-Witten invariants in algebraic geometry, Invent. Math. 127 (1997), 607-617.

[2] K. Behrend, Y.I. Manin, Stacks of stable maps and Gromov-Witten invariants, Duke Math. J. 85 (1996), 1-60.

[3] F.R. Cohen, Cohomology of C_{n+1}-spaces, Lecture Notes in Math. 533, p. 207-351, Springer, 1977.

[4] A. Connes, Noncommutative Geometry, Academic Press, 1996.

[5] B. Dubrovin, Geometry of 2D topological field theories, Lecture Notes in Math. 1620, p. 120-348, Springer, 1996.

[6] B. Fresse, Lie theory of formal groups over an operad, preprint IRMA, Strasbourg 1996.

[7] I.M. Gelfand, V.S. Retakh, Quasideterminants I, Selecta Math. 3 (1997), 517-546.

[8] E. Getzler, Operads and moduli spaces of Riemann surfaces, in: "The moduli space of curves" (R. Dijkgraaf et al. Eds.), 199-230, Birkhäuser, Boston, 1995.

[9] E. Getzler, M. Kapranov, Cyclic operads and cyclic homology, in "Geometry, Topology and Physics for R. Bott" (S.T. Yau, Ed.) p.167-201, Intern. Press, Cambridge MA 1995.

[10] E. Getzler, M. Kapranov, Modular operads, Compositio Math. 110 (1998), 65-126.

[11] V. Ginzburg, M. Kapranov, Koszul duality for operads, Duke Math. J. 76 (1994), 203-272.
[12] J. Harer, The cohomology of the moduli space of curves, Lect. Notes in Math. 1337, p. 138-221, Springer, 1988.
[13] J. Harer, D. Zagier, The Euler characteristic of the moduli space of curves, Invent. Math. 85 (1986), 457-485.
[14] M. Kapranov, Rozansky-Witten invariants via Atiyah classes, preprint alg-geom/9704009, Compositio Math., to appear.
[15] M. Kapranov, Noncommutative geometry based on commutator expansions, preprint math.AG/9802041, Crelle's J., to appear.
[16] A.N. Kolmogoroff, On the representation of continuous functions of several variables as superpositions of functions of smaller number of variables, Soviet. Math. Dokl. 108 (1956), 179-182.
[17] M. Kontsevich, Formal noncommutative symplectic geometry, in "The Gelfand Mathematical Seminars" (L. Corwin et al. Eds.) p. 173-187, Birkhäuser, Boston, 1993.
[18] M. Kontsevich, Rozansky-Witten invariants via formal geometry, preprint dg-ga/9704009, Compositio Math. to appear.
[19] M. Kontsevich, Y.I. Manin, Gromov-Witten classes, quantum cohomology and enumerative geometry, Comm. Math. Phys. 164 (1994), 525-562.
[20] I. Kriz, J. P. May, Operads, Algebras, Modules and Motives, Asterisque, 233, 1985.
[21] M. Lazard, Lois de groupe et analyseurs, Ann. ENS, 72 (1955), 299-400.
[22] J.-L. Loday, La renaissance des opérades, Sém. Bourbaki, Exp. 792, 1994-5.
[23] Y.I. Manin, Topics in Noncommutative Geometry, Princeton Univ. Press, 1992.
[24] Y.I. Manin, Frobenius manifolds, quantum cohomology and moduli spaces, preprint math.QA/9801006.
[25] J.P. May, Geometry of Iterated Moduli Spaces, Lecture Notes in Math. 271, Springer, 1972.
[26] R. Penner, Perturbative series and the moduli space of punctured surfaces, J. Diff. Geom. 27 (1988), 35-53.
[27] S. Priddy, Koszul resolutions, Trans. AMS, 152 (1970), 39-60.
[28] A. Rosenberg, Noncommutative Algebraic Geometry and Representations of Quantized Algebras, Kluwer Publ. 1995.
[29] L. Rozansky, E. Witten, Hyperkähler geometry and invariants of 3-manifolds, Selecta Math. 3 (1997), 401-458.
[30] G. Segal, Geometric aspects of quantum field theory, Proc. ICM-90, p. 1387-1396, Springer, 1991.
[31] J.D. Stasheff, Homotopy associativity of H-spaces, Trans. AMS, 108 (1963), 275-312.

M. Kapranov
Department of Mathematics
Northwestern University
Evanston IL 60208 USA

Section 5

Differential Geometry and Global Analysis

In case of several authors, Invited Speakers are marked with a *.

Hard Balls Gas and Alexandrov Spaces of Curvature Bounded Above

Dmitri Burago[1]

Abstract. This lecture is an attempt to give a very elementary account of a circle of results (joint with S. Ferleger and A. Kononenko) in the theory of semi-dispersing billiard systems. These results heavily rely on the methods and ideology of the geometry of non-positively curved length spaces.

The purpose of this lecture is to give a very informal and elementary account of one geometric approach in the theory of billiard systems. Precise formulations of the results (joint with S. Ferleger and A. Kononenko), their proofs and a more detailed exposition can be found in the survey [B-F-K-4] and the papers [B-F-K-1],[B-F-K-2] and [B-F-K-3].

The approach is based on representing billiard trajectories as geodesics in a certain length space. This representation is similar to turning billiard trajectories in a square billiard table into straight lines in a plane tiled by copies of the square. It is important to understand that this construction by itself does not provide new information regarding the billiard system in question; it only converts a dynamical problem into a geometric one. Nevertheless, while a problem may seem rather difficult in its billiard clothing, its geometric counterpart may turn out to be relatively easy by the standards of the modern metric geometry. For the geometry of non-positively curved length spaces we refer to [Ba], [Gr] and [Re].

Apparently, one of the motivations to study semi-dispersing billiard systems comes from gas models in statistical physics. For instance, the hard ball model is a system of round balls moving freely and colliding elastically in a box or in empty space. Physical considerations naturally lead to several mathematical problems regarding the dynamics of such systems. The problem that served as the starting point for the research discussed in this lecture asks whether the number of collisions in time one can be estimated from above. Another well-known and still unsolved problem asks whether such dynamical systems are ergodic. A "physical" version of both problems goes back to Boltzman, while their first mathematical formulation is probably due to Ya. Sinai.

[1]The author greatfully acknowledges the support of the NSF and the Sloan Foundation in the form of a Sloan Research Fellowship and NSF grants DMS-95-05175 and DMS-98-03129. The author thanks MFI in Oberwolfach, ESI in Vienna, ETH in Zurich, IRMA in Strasbourg and Universities Paris-11&12 for his fruitful and unforgettable visits to these institutions.

Making a short digression here, I would mention that, in my opinion, the adequacy of these model problems for physical reality is quite questionable. In particular, these problems are extremely sensitive to slight changes of their formulations. Introducing particles that are arbitrarily close in shape to the round balls and that are allowed to rotate, one can produce unbounded number of collisions in unit time [Va]. It is plausible that introducing even a symmetrical and arbitrarily steep potential of interaction between particles instead of discontinuous collision "potential", one can destroy the ergodicity ([Do]). The result of Simanyi and Szasz ([Si-Sz])(seemingly, the best one can prove in support of the ergodicity of the hard balls model in the present state of the art) asserts that the ergodicity does take place ... for almost all combinations of radii and masses of the balls. Such a result should be less than satisfactory for a physicist, since a statement that is valid only for "balls of irrational radii" does not make any physical sense at all. Perhaps, one would rather hope that the existence of an ergodic component whose complement is negligibly small (at least for a system of very many balls) is a more stable property. On the other hand, hard balls gas (of even very many small balls) in a spherical or a cylindrical vessel is obviously very non-ergodic since it possesses a first integral coming from rotational symmetries of the system. This happens regardless of a good deal of hyperbolicity produced by the dynamics of colliding balls, and it is not at all clear what happens if the symmetrical shape of the vessel is slightly perturbed.

Regardless of this minor criticism of the physical meaning of mathematical problems involving gas models, the author believes that these problems are quite interesting on their own, and from now on we stick to their mathematical set-up. It is well known that, by passing to the configuration space, the dynamics of a N balls can be substituted by the dynamics of one (zero-size) particle moving in the complement of several cylinders in $\mathbf{R}^{3N}$ and experiencing elastic collisions with the cylinders. These cylinders correspond to the prohibited configurations where two of the balls intersect. Another gas model, the Lorentz gas, just begins with a dynamical system of one particle moving in the complement of a regular lattice of round scatterers; its dynamics can be studied on the quotient space, which is a torus with a scatterer in it. All these example fit in the following general scheme.

Let M be a complete Riemannian manifold M together with a (finite or at least locally-finite) collection of smooth convex subsets B_i. These convex sets B_i are bounded by (smooth, convex) hypersurfaces W_i, which (together with B_i's) will be referred to as *walls*. In most physical models, M is just a flat torus or Euclidean space (whose Euclidean structure given by the kinetic energy of the system). Throughout this lecture we assume that M has non-positive curvature and positive injectivity radius; however, local uniform bounds on the number of collisions remain valid without these restrictions. The dynamics takes place in the (semi-dispersing) billiard table, which is the complement of $\bigcup B_i$ in M. More precisely, the phase space is (a subset of) the unit tangent bundle to this complement. A point moves along a geodesic until it reaches one of the walls W_i, and then it gets reflected so that both the magnitude and the projection of its velocity on the plane tangent to the wall are conserved. For simplicity, we exclude the trajectories that ever experience a collision with two walls simultaneously.

Systematic mathematical study of such systems, called semi-dispersing billiards, was initiated by Ya. Sinai and continued by many other mathematicians and physicists.

Our discussion will be concentrated around the idea of gluing several copies of M together and then developing billiard trajectories into this new space. This idea is very old and its simplest versions arise even in elementary high-school mathematical puzzles. For instance, if the billiard table is a square, one can consider a tiling of Euclidean plane by such squares, and billiard trajectories turn into straight lines. Although this idea is rather naive, it already provides valuable information. For instance, if one wonders how close a non-periodic trajectory comes to vertices of the square, the answer is given in terms of rational approximations to the slope of the corresponding line. In this instance, a dynamical problem is transformed into a question in the arithmetic of real numbers. We plan to do an analogous reformulation with geometry of length spaces on the other side.

We are concerned with semi-dispersing billiard systems. In the early sixties V. Arnold "speculated" that "such systems can be considered as the limit case of geodesic flows on negatively curved manifolds (the curvature being concentrated on the collisions hypersurface)" [Ar]. Indeed, this is nowadays well known (due to the works of Sinai, Bunimovich, Chernov, Katok, Strelcyn, Szasz, Simanyi and many others) that a large portion of the results in the smooth theory of (semi-)hyperbolic systems can be generalized (with appropriate modifications) to (semi-)dispersing billiards. In spite of this, the construction suggested by Arnold has never been used. It also caused several serious objections; in particular, A. Katok pointed out that such approximations by geodesic flows on manifolds necessarily produce geodesics that bend around collision hypersurfaces and therefore have no analogs in the billiard system.

To illustrate both Arnold's suggestion and the difficulty noticed by Katok, let us consider a simple example of the billiard in the complement of a disc in a two-torus (or Euclidean plane). Taking two copies of the torus with (open) discs removed and gluing them along the boundary circles of the discs, one obtains a Riemannian manifold (a surface of genus 2) with a metric singularity along the gluing circle. This manifold is flat everywhere except at this circle. One can think of this circle as carrying singular negative curvature. Smoothing this metric by changing it in an (arbitrarily small) collar around the circle of gluing, one can obtain a non-positively curved metric, which is flat everywhere except in this collar. To every segment of a billiard trajectory, one can (canonically) assign a geodesic in this metric. Collisions with the disc would correspond to intersections with the circle of gluing, where the geodesic leaves one copy of the torus and goes to the other one.

Unfortunately, many geodesics do not correspond to billiard trajectories. They can be described as coming from "fake" trajectories hitting the disc at zero angle, following an arc of its boundary circle (possibly even making several rounds around it) and then leaving it along a tangent line. Dynamically, such geodesics carry "the main portion of entropy" and they cannot be disregarded. On the other hand, it is difficult to tell actual trajectories from the fake ones when analyzing the geodesic flow on this surface.

There is another difficulty arising in higher dimension. If one tries to repeat the same construction for a three-torus with a ball removed, then after gluing two copies of this torus the gluing locus defines a totally geodesic subspace. It carries positive curvature, and this positive curvature persists under smoothing of the metric in a small collar of the sphere. Thus, in this case we do not get a negatively curved manifold at all.

We will (partially) avoid these difficulties by substituting a non-positively curved manifold by a length space of non-positive curvature in the sense of A.D. Alexandrov. Unfortunately, a construction that would allow us to represent all billiard trajectories as geodesics in one compact space is unknown in dimensions higher than three. Attempts to do this lead to a striking open question: Is it possible to glue finitely many copies of a regular 4-simplex to obtain a (boundaryless) non-positive pseudo-manifold (cf. [B-F-Kl-K])?

We introduce a construction that represents trajectories from a certain combinatorial class, where by a combinatorial class of (a segment of) a billiard trajectory we mean a sequence of walls that it hits.

Fix such a sequence of walls $K = \{W_{n_i}, i = 1, 2, \dots N\}$. Consider a sequence $\{M_i, i = 0, 1, \dots N\}$ of isometric copies of M. For each i, glue M_i and M_{i+1} along B_{n_i}. Since each B_{n_i} is a convex set, the resulting space M_K has the same upper curvature bound as M due to Reshetnyak's theorem ([Re]).

There is an obvious projection $M_K \to M$, and M can be isometrically embedded into M_K by identifying it with one of M_i's (regarded as subsets of M_K). Thus every curve in M can be lifted to M_K in many ways. A billiard trajectory whose combinatorial class is K admits a canonical lifting to M_K: we lift its segment till the first collision to $M_0 \subset M_K$, the next segment between collisions to $M_1 \subset M_K$ and so on. Such lifting will be called developing of the trajectory. It is easy to see that a development of a trajectory is a geodesic in M_K.

Note that, in addition to several copies of the billiard table, M_K contains other redundant parts formed by identified copies of B_i's. For example, if we study a billiard in a curved triangle with concave walls, B_i's are not the boundary curves. Instead, we choose as B_i's some convex ovals bounded by extensions of these walls. (One may think of a billiard in a compact component of the complement to three discs.) In this case, these additional parts look like "fins" attached to our space (the term "fin" has been used by S. Alexander and R. Bishop in an analogous situation). In case of the billiard in the complement of a disc in a two-torus (see discussion above), the difference is that we do *not* remove the disc when we glue together two copies of the torus. Now a geodesic cannot follow an arc of the disc boundary, as the latter can be shorten by pushing inside the disc. Still, there are "fake" geodesics, which go through the disc. However, there are fewer of them than before and it is easier to separate them.

It might seem more natural to glue along the boundaries of W_{n_i} rather than along the whole B_{n_i}. For instance, one would do so thinking of this gluing as "reflecting in a mirror" or by analogy with the usual development of a polygonal billiard. However, gluing along the boundaries will not give us a non-positively curved space in any dimension higher than 2.

One may wonder how the interiors of B_i's may play any role here, as they are

"behind the walls" and billiard trajectories never get there. For instance, instead of convex walls in a manifold without boundary, one could begin with a manifold with several boundary components, each with a non-negative definite second fundamental form (w.r.t. the inner normal). Even for one boundary component, this new set-up cannot be reduced to the initial formulation by "filling in" the boundary by a non-positively curved manifold. Such an example was pointed out to me by J. Hass ([Ha]), and our main dynamical result does fail for this example. Thus, it is indeed important that the walls are not only locally convex surfaces, and we essentially use the fact that they are filled by convex bodies.

Let us demonstrate how the construction of M_K can be used by first reproving (and slightly generalizing) a known result. L. Stoyanow has shown that each combinatorial class of trajectories in a strictly dispersing billiard (in Euclidean space or a flat torus) contains no more than one periodic trajectory. By a strictly dispersing property we mean that all walls have positive definite fundamental forms. For a semi-dispersing billiard, L. Stoyanov proved that all periodic trajectories in the same combinatorial class form a family of parallel trajectories of the same length. Together with local bounds on the number of collisions (which were known in dimension 2, and the general case is discussed below), these results imply exponential upper bound on the growth of the number of (parallel classes of) periodic trajectories. These estimates are analogous to the estimates on the number of periodic geodesic in non-positively curved manifolds.

Assume that we have two periodic trajectories in the same combinatorial class K. Choose a point on each trajectory and connect the points by a geodesic segment $[xy]$. Let us develop one period of each trajectory into M_K, obtaining two geodesics $[x'x'']$ and $[y'y'']$ connected by two lifts $[x'y']$ and $[x''y'']$ of the segment $[xy]$. Thus, M_K contains a geodesic quadrangle with the sum of angles equal to 2π. It is well known that, in a non-positively curved space, such a quadrangle bounds a flat totally-geodesic surface; in our case it has to be a parallelogram since it has equal opposite angles. Thus, $|x'x''| = |y'y''|$ and the family of lines parallel to $[x'x'']$ and connecting the sides $[x'y']$ and $[x''y'']$ projects to a family of periodic trajectories. Moreover, this parallelogram has to intersect the walls in segments, and thus it is degenerate if the fundamental forms of the walls are positive definite. This just means that the two periodic trajectories coincide. The same is true if the sectional curvature of M is strictly negative, as it is equal to zero for any plane tangent to the parallelogram.

This argument is ideologically very close to the proof of the following result: the topological entropy of the time-one map T of the billiard flow for a compact semi-dispersing billiard table is finite. Note that the differential of the time-one map T is unbounded, and therefore the finiteness of the topological entropy is not obvious. Moreover, it is quite plausible that the following problem has an affirmative solution: if one drops the curvature restriction for M, can the topological entropy of the time-one map be infinite? Is the topological entropy of the billiard in a smooth convex curve in Euclidean plane always finite?

To estimate the topological entropy by h, it is enough to show that, given a positive ϵ, there is a constant $C(\epsilon)$ with the following property: for each N, the space of trajectories $T^i(v), i = 0, 1, \ldots, N$ can be partitioned into no more than

$C(\epsilon) \cdot \exp(hN)$ classes in such a way that every two trajectories from the same class stay ϵ-close to each other.

At first glance, such a partition seems rather evident in our situation. Indeed, first let us subdivide M into several regions of diameter less than ϵ (the number of these regions is independent of N). If M is simply connected, we can just say that two trajectories belong to the same class if they have the same combinatorial class and both trajectories start from the same region and land in the same region of the subdivision of M. If M is not simply-connected, one also requires that the trajectories have the same homotopy type (formally, lifting two corresponding segments of the flow trajectories of the same combinatorial class K to M_K and connecting their endpoints by two shortest path, one gets a rectangle; this rectangle should be contractible). Since both the number of combinatorial classes and the fundamental group of M grow at most exponentially, is rather easy to give an exponential (in N) upper bound on the number of such classes (using again the local uniform estimates on the number of collisions, see below). On the other hand, for two trajectories from the same class, their developments into the appropriate M_K have ϵ-close endpoints and the quadrangles formed by the geodesics and the shortest paths connecting their endpoints is contractible. For a non-positively curved space, this implies that these geodesics are ϵ-close everywhere between their endpoints.

There is, however, a little hidden difficulty, which the reader should be aware of. The previous argument proves the closeness between the projections of two trajectories onto M, while we need to establish this closeness in the phase space. Thus, some extra work has to be done to show that if two geodesics in M_K stay sufficiently close, then so do the directions of their tangent vectors (in some natural sense). This is a compactness-type argument, which we will not dwell upon here.

Let us come back to the example used above to illustrate Arnold's suggestion. This is 2-dimensional Lorentz gas, that is the billiard in the complement of a disc in a flat two-torus. To count the number of classes in the above sketch of the argument, one can pass to an Abelian cover of M_K (since this billiard table has just one wall, there is no ambiguity in choosing K). The latter is two copies of Euclidean plane glued together along a lattice of discs centered at integer points. A (class of) billiard trajectories naturally determines a broken line with integer vertices. While not every broken line with integer vertices arises from a billiard trajectory, the portion of such lines coming from "fake" trajectories approaches zero for small radii of scatterers. Counting such broken lines is a purely combinatorial problem, and one sees that the topological entropy of Lorentz gas converges to a number between 1 and 2 as the radius of the repeller approaches zero. This result is stable: the "limit entropy" is the same for a convex repeller of any shape. The author has no idea whether this number has any physical meaning.

Now we pass to the main problem of estimating the number of collisions. For the hard ball system, one asks whether the number of collisions that may occur in this system can be estimated from above by a bound depending only on the number of balls and their masses. If we consider the balls moving in unbounded Euclidean space, we count the total number of collisions in infinite time. For a system of balls in a box, we mean the number of collisions in unit time (for a fixed

value of kinetic energy). As far as I know, these problems have been resolved only for systems of three balls ([Th-Sa], [Mu-Co]).

It is relatively easy to establish such upper bounds on the number of "essential" collisions, opposed to collisions when two balls barely touch each other. While such "non-essential" collisions indeed do not lead to a significant exchange by energy or momentum, they nevertheless cannot be disregarded from a "physical viewpoint". Indeed, they may serve as the main cause of instability in the system: the norm of differential of the flow does not admit an upper bound just at such trajectories. In a general semi-dispersing billiard it is also easier to estimate the number of collisions that occur at an angle separated from zero. Such arguments are based on introducing a bounded function on the phase space so that the function does not decrease along each trajectory and increases by an amount separated from zero after each "essential" collision. For some cases, such as 2-dimensional and polyhedral billiard tables, one can estimate the fraction of "essential collisions" among all collisions and thus get uniform bounds on the total number of collisions (see [Va], [Ga-1], [Ga-2], [Si-1]). The simplest case that is unclear how to treat by such methods is a particle shot almost along the intersection line of two convex surfaces in 3-dimensional Euclidean spaces and hitting the surfaces at very small angles.

Contrary to dynamical arguments indicated above, we use a geometric approach based on some length comparisons. Let us first prepare the necessary notation and formulations. When one wants to obtain uniform bounds on the number of collisions for a general semi-dispersing billiard table, it is clear that an additional assumption is needed. Indeed, already for a two-dimensional billiard table bounded by several concave walls, a trajectory may experience an arbitrarily large number of collisions (in time one) in a neighborhood of a vertex if two boundary curves are tangent to each other. Thus, a non-degeneracy condition is needed. For simplisity, let us introduce the following *non-degeneracy assumption* (it can be essentially weakened for non-compact billiard tables): there exists a number C such that, if a point is ϵ-close to all sets from some sub-collection of the B_i's, then it is $C\epsilon$-close to the intersection of B_i's from this sub-collection. This assumption rules out various degenerations of the arrangements of hyperplanes tangent to walls. It is not difficult to verify that the hard ball gas model does satisfy the non-degeneracy assumption.

The main local result reads as follows: if a semi-dispersing billiard table satisfies the non-degeneracy assumption, then there exists a finite number P such that every point p in the billiard table possesses a neighborhood $U(p)$ such that every trajectory segment *contained in* $U(p)$ experiences no more than P collisions.

Passing to estimating the global number of collisions (for infinite time) we want to stay away from situations such as a particle infinitely bouncing between two disjoint walls. The result for this case reads as follows: if a semi-dispersing billiard table satisfies the non-degeneracy assumption, M is simply-connected and the intersection $\bigcap B_i$ of B_i's is non-empty, then there exists a finite number P such that every trajectory experiences no more than P collisions.

Outlining the proofs of these results, we restrict ourselves to the case of two walls W_1 and W_2 bounding two convex sets B_1 and B_2. Thus we avoid inessential

combinatorial complications and cumbersome indices.

We begin by discussing the local bound. Let us assume that M is simply-connected; otherwise, one can pass to its universal cover. Consider a billiard trajectory T connecting two points x and y and pick any point $z \in B_1 \bigcap B_2$. Denote by $K = \{W_1, W_2, W_1, W_2 \ldots\}$ the combinatorial class of T, and consider the development T' of T in M_K. This is a geodesic between two points x' and y'. By Alexandrov's theorem, every geodesic in a simply-connected non-positively curved space is the shortest path between its endpoints. Note that z canonically lifts to M_K since all copies of z in different copies of M got identified. Denoting this lift by z', we see that $|zx| = |z'x'|$ and $|zy| = |z'y'|$. Thus we conclude that the lengths of T between x and y is less that $|xz| + |zy|$ for all $z \in B_1 \bigcap B_2$. In other words, any path in M connecting x and y and visiting the intersection $B_1 \bigcap B_2$ is longer than the segment of T between x and y.

The following argument is the core of the proof. It shows that if a trajectory made too many collisions then it can be modified into a shorter curve with the same endpoints and passing through the intersection $B_1 \bigcap B_2$. This contradicts the previous assertion and thus gives a bound on the number of collisions.

Assume that T is contained in a neighborhood $U(p)$ and it collided with W_1 at points $a_1, a_2, \ldots a_N$ alternating with collisions with W_2 at $b_1, b_2, \ldots b_N$. Let z_i be the point in $B_1 \bigcap B_2$ closest to b_i and h_i be the distance from b_i to the shortest geodesic $[a_i a_{i+1}]$. By the non-degeneracy assumption, $|z_i b_i| \le C \cdot dist(b_i, B_1) \le h_i$. Thus the distance H_i from z_i to the shortest geodesic $a_i a_{i+t}$ is at most $(C+1)h_i$.

Plugging this inequality between the heights of the triangles $a_i b_i a_{i+1}$ and $a_i z_i a_{i+1}$ into a routine argument which develops these triangles on both Euclidean plane and k-plane, one concludes that $d_i \le C_1 \cdot D_i$, where $d_i = |a_i b_i| + |b_i a_{i+1}| - |a_i a_{i+1}|$, $D_i = |a_i z_i| + |z_i a_{i+1}| - |a_i a_{i+1}|$. Here k is the infinum of the sectional curvature in $U(p)$, and a constant C_1 can be chosen depending on C alone provided that $U(p)$ is sufficiently small.

Let d_j be the smallest of d_i's. Let us modify the trajectory T into a curve with the same endpoints: substitute its pieces $a_i b_i a_{i+1}$ by the shortest segments $a_i a_{i+1}$ for all i's excluding $i = j$. This new curve is shorter than T by at least $(N-1)d_p$. Let us make a final modification by replacing the piece $a_j b_j a_{j+1}$ by $a_j z_j a_{j+1}$. It makes the path longer by D_j, which is at most $C_1 d_j$. Hence, $N \le C_1 + 1$ because otherwise we would have a curve with the same endpoints as T, passing through $z_j \in B_1 \bigcap B_2$ and shorter than T. This proves the local bound on the number of collisions.

Now we are ready to estimate the global number of collisions, and here geometry works in its full power. Consider a trajectory T making N collisions with the walls $K = \{1, 2, 1, \ldots, 2, 1\}$. Reasoning by contradiction, assume that $N > 3P+1$, where P is the local bound on the number of collisions. Consider the space M_K and "close it up" by gluing $M_0 \in M_K$ and $M_N \in M_K$ along the copies of B_1. Denote the resulting space by $\tilde{M}$. We cannot use Reshetnyak's theorem to conclude that $\tilde{M}$ is a non-positively curved space any more, since we identify points in the same space and we do not glue two spaces along a *convex* set.

We recall that a space has non-positive curvature iff every point possesses a neighborhood such that, for every triangle contained in the neighborhood, its

angles are no bigger than the corresponding angles of the comparison triangle in Euclidean plane. However, using the correspondence between geodesics and billiard trajectories, one can conclude (reasoning exactly as in the proof of the *local* estimates on the number of collisions), that each side of a small triangle cannot intersect interiors of more than P copies of the billiard table. Since $N > 3P+1$, for every small triangle for which we want to verify the angle comparison property, we can undo one of the gluings without tearing the sides of the triangle. This ungluing may only increase triangle's angles, but now we find ourselves in a non-positively curved space (which is actually just M_K), and thus we get the desired comparison for the angles of the triangle.

To conclude the proof, it remains to notice that the development of T in $\tilde{M}$ is a geodesic connecting two points in the same copy of B_1. This is a contradiction since every geodesic in a simply-connected non-positively curved space is the only shortest path between its endpoints; on the other hand, there is a shortest path between the same points going inside this copy of B_1.

Let us finish with the following remark. It would be desirable if one could begin with finitely many copies of M and glue them together along walls B_i to obtain a non-positively curved space $\hat{M}$ so that each wall participates in at least one gluing. In particular, such a construction would immediately provide an alternative proof for both local and global estimates on the number of collisions. For instance, for global estimates it is enough to notice that every billiard trajectory lifts to a shortest path and hence it cannot intersect a copy of one wall in $\hat{M}$ more than once. Hence the number of collisions is bounded by the total number of copies of walls in $\hat{M}$. As it is mentioned above, it is however unclear whether such gluing exists even for a regular 4-simplex.

Acknowledgements. The author would like to use this opportunity to thank M. Brin, M. Gromov, A. Katok and B. Kleiner for interesting discussions and valuable comments.

References

[Ar] V.Arnold. Lecture given at the meeting in the Fields institute dedicated to his 60th birthday.

[Ba] W.Ballmann. Lectures on spaces of nonpositive curvature. With an appendix by Misha Brin. DMV Seminar, 25. Birkhauser Verlag, Basel, 1995.

[B-F-K-1] D.Burago, S.Ferleger, A.Kononenko. Uniform estimates on the number of collisions in semi-dispersing billiards. *Annals of Mathematics,* to appear.

[B-F-K-2] D.Burago, S.Ferleger, A.Kononenko. Topological entropy of semi-dispersing billiards. *Ergodic Theory and Dynamical Systems,* to appear.

[B-F-K-3] D.Burago, S.Ferleger, A.Kononenko. Unfoldings and global bounds on the number of collisions for generalized semi-dispersing billiards. *Asian J. of Math,* to appear.

[B-F-K-4] D.Burago, S.Ferleger, A.Kononenko. A geometric approach to semi-dispersing billiards. *Ergodic Theory and Dynamical Systems, Reviews* to appear.

[B-F-Kl-K] D.Burago, S.Ferleger, B.Kleiner and A.Kononenko. Gluing copies of a 3-dimensional polyhedron to obtain a closed nonpositively curved (pseudo)manifold. *preprint*.

[Do] V.Donnay. Elliptic islands in generalized Sinai billiards. *Ergodic Theory Dynam. Systems*, 16, no. 5, 975–1010, 1996.

[Ga-1] G.A.Gal'perin. Systems with locally interacting and repelling particles moving in space. (Russian) *Tr. MMO*, 43, 142–196, 1981.

[Ga-2] G.A.Gal'perin. Elastic collisions of particles on a line. (Russian) *Uspehi Mat. Nauk*, 33 no. 1(199), 211–212, 1978.

[Gr] M.Gromov. Structures metriques pour les varietes riemanniennes. Edited by J. Lafontaine and P. Pansu. Textes Mathematiques, 1. CEDIC, Paris, 1981.

[Ha] J.Hass, P.Scott. Bounded 3-manifold admits negatively curved metric with concave boundary. *J. Diff. Geom.* 40, no. 3, 449-459, 1994.

[Mu-Co] T.J.Murphy, E.G.D.Cohen. Maximum Number of Collisions among Identical Hard Spheres, *J. Stat. Phys.*, 71, 1063–1080, 1993.

[Re] Yu.G.Reshetnyak (ed.). Geometry 4, non-regular Riemannian geometry. *Encyclopedia of Mathematical Sciences*, Vol.70, 1993.

[Si-1] Ya.G.Sinai. Billiard trajectories in polyhedral angles. *Uspehi Mat. Nauk*, 33, No.1, 229–230, 1978.

[Si-2] Ya.G.Sinai (ed.). Dynamical Systems 2. *Encyclopedia of Mathematical Sciences*, Vol.2, 1989.

[Si-3] Ya.G.Sinai. On the foundations of the ergodic hypothesis for a dynamical system of statistical mechanics. *Soviet Math. Dokl.*, 4, 1818–1822, 1963.

[Si-4] Ya.G.Sinai. Hyperbolic billiards. Proceedings of the International Congress of Mathematicians, Vol. I, II (Kyoto, 1990), 249–260, Math. Soc. Japan, Tokyo, 1991.

[Si-Sz] N.Simanyi, D.Szasz. Lecture given at Penn State University, 1997.

[St] L.Stojanov. An estimate from above of the number of periodic orbits for semi-dispersed billiards. *Comm. Math. Phys.*, 124, No.2, 217-227, 1989.

[Th-Sa] W.Thurston, G.Sandri. Classical hard sphere 3-body problem. *Bull. Amer. Phys. Soc.*, 9, 386, 1964.

[Va] L.N.Vaserstein. On systems of particles with finite range and/or repulsive interactions. *Comm. Math. Phys.*, 69, 31–56, 1979.

Dmitri Burago
Pennsylvania State University
Deptartment of Mathematics
University Park PA-16802 USA
burago@math.psu.edu

Spaces with Ricci Curvature Bounds

Tobias H. Colding

0. Introduction

One of the early (and very important) formulas involving Ricci curvature is the so called Bochner formula from the forties. This formula asserts that if M^n is a Riemannian manifold and $u \in C^3(M)$, then

$$\frac{1}{2}\Delta|\nabla u|^2 = |\mathrm{Hess}_u|^2 + \langle \nabla \Delta u, \nabla u\rangle + \mathrm{Ric}\,(\nabla u, \nabla u)\,. \tag{0.1}$$

Bochner used this formula to conclude that a closed n-dimensional manifold with nonnegative Ricci curvature has first Betti number, b_1, at most equal to the dimension with equality if and only if the manifold is a flat torus.

From the Bochner formula one can also obtain the Ricatti equation which in turn can be seen to yield the Laplacian comparison. This important comparison principle allows one to construct cut off functions with bounds on the Laplacian on manifolds with a lower Ricci curvature bound. It also allows one to use maximum principle methods to give a priori estimates on these manifolds. The Laplacian comparison theorem says that if $\mathrm{Ric}_{M^n} \geq (n-1)\,\Lambda$, $x \in M$ is fixed, r denotes the distance function to x, and $f : \mathbf{R} \to \mathbf{R}$ is a C^2 function then the following hold. If $f' \geq 0$, then

$$\Delta f(r) \leq \Delta_\Lambda f(r)\,, \tag{0.2}$$

and if $f' \leq 0$ then

$$\Delta_\Lambda f(r) \leq \Delta f(r)\,. \tag{0.3}$$

Here $\Delta_\Lambda f(r)$ denote the corresponding quantities on the simply connected n-dimensional space form of constant sectional curvature Λ.

It is important to note (as Calabi originally emphasized in the fifties) that the Laplacian comparison holds in a useful generalized sense, even at points where the distance function fails to be smooth i.e. on the cut locus. This point was also illustrated in the gradient estimate mentioned below and the Cheeger-Gromoll splitting theorem from the early seventies.

From the Laplacian comparison together with an integration by parts argument one can get the so called volume comparison theorem (see below for an important application of this). This comparison theorem assert that if $\mathrm{Ric}_{M^n} \geq (n-1)\,\Lambda$ then for all $x \in M$ and all $0 < s \leq t$

$$\frac{\mathrm{Vol}(B_s(x))}{\mathrm{Vol}(B_t(x))} \geq \frac{\mathrm{V}^n_\Lambda(s)}{\mathrm{V}^n_\Lambda(t)}\,. \tag{0.4}$$

Here $\mathrm{V}^n_\Lambda(s)$ is the volume of a ball of radius s in the simply connected space form with constant sectional curvature Λ. For the present survey one of the most important consequences of the volume comparison theorem is that it can be thought of as claiming the monotonicity of a density type quantity. For instance if the Ricci curvature is nonnegative then the volume comparison can clearly be restated as

$$r_0^{-n}\,\mathrm{Vol}(B_{r_0}(x))\downarrow, \tag{0.5}$$

for all fixed x. A particular consequence of this volume comparison theorem is the upper bound on the volume of balls due to Bishop in the early sixties. That is if $\mathrm{Ric}_{M^n} \geq (n-1)\,\Lambda$ then for all $x \in M$ and all $0 < r_0$

$$\mathrm{V}^n_\Lambda(r_0) \geq \mathrm{Vol}(B_{r_0}(x))\,. \tag{0.6}$$

Another very crucial estimate on manifolds with a lower Ricci curvature bound is the gradient estimate of Cheng and Yau from the early seventies.

For instance if $\mathrm{Ric}_{M^n} \geq (n-1)\,\Lambda$, $x \in M$ and u is a harmonic function on $B_{2r_0}(x)$ then

$$\sup_{B_{r_0}(x)} |\nabla u| \leq C(n,\Lambda,r_0) \sup_{B_{2r_0}(x)} |u|\,. \tag{0.7}$$

The final inequality that we will recall is the Abresch-Gromoll inequality from the late eighties. This inequality was inspired by the Cheeger-Gromoll splitting theorem and it gives an important estimate for thin triangles. A specific case of this inequality is the following. If $\mathrm{Ric}_{M^n} \geq 0$, $n \geq 3$, x, y, $z \in M$ and $e_{x,y}(z) = \overline{x,z} + \overline{z,y} - \overline{x,y}$ then assuming $2\,\overline{x,z} \leq \overline{x,y}$,

$$\min\{\overline{x,z}^{\frac{n}{n-1}}, \overline{x,z}^{\frac{1}{n-1}}\, e_{x,y}(z)\} \leq C\,h^{\frac{n}{n-1}}\,. \tag{0.8}$$

Here h is the height of the triangle, that is $h = \inf_{m\in\{\gamma_{x,y}\}} \overline{m,z}$ where $\{\gamma_{x,y}\}$ is the set of minimal geodesics connecting x, y and $C = C(n)$.

There is a natural generalization of the classical Hausdorff distance between subsets of Euclidean space to a distance function on all metric spaces. This is the Gromov-Hausdorff distance and it gives a good tool for studying metric spaces.

Suppose that (X, d_X) and (Y, d_Y) are two compact metric spaces. We say that the Gromov-Hausdorff distance between them is at most $\epsilon > 0$ if there exist maps $f : X \to Y$ and $g : Y \to X$ such that

$$\forall x_1,\, x_2 \in X:\ |d_X(x_1,x_2) - d_Y(f(x_1),f(x_2))| < \epsilon\,, \tag{0.9}$$

$$\forall x \in X:\ d_X(x, g\circ f(x)) < \epsilon\,, \tag{0.10}$$

and the two symmetric properties in Y hold. The Gromov-Hausdorff distance between X and Y, denoted by $d_{GH}(X,Y)$, is then the infimum of all such ϵ.

Using this topology we can then say that a sequence of metric space converges to another metric space. For noncompact metric spaces there is a more useful notion of convergence which is essentially convergence on compact subsets. Namely, if (X_i, x_i, d_{X_i}) is a pointed sequence of metric spaces then we say that (X_i, x_i, d_{X_i}) converges to some pointed metric space (X, x, d_X) in the pointed Gromov-Hausdorff topology if for all $r_0 > 0$ fixed the compact metric spaces $B_{r_0}(x_i)$ converges to $B_{r_0}(x)$.

Gromov's compactness theorem is the statement that any pointed sequence, (M_i^n, m_i), of n-dimensional manifolds, with

$$\mathrm{Ric}_{M_i^n} \geq (n-1)\,\Lambda\,, \tag{0.11}$$

has a subsequence, (M_j^n, m_j), which converges in the pointed Gromov-Hausdorff topology to some length space (M_∞, m_∞).

The proof of this compactness theorem relies only on the volume comparison theorem. In fact it only uses a volume doubling which is implied by the volume comparison.

Finally we refer to [C5] and [Ga] for surveys related to this article and more detailed references.

1. Geometry and topology of smooth manifolds

Let us first recall some results from [C3]; see also [C1], [C2].

Theorem 1.1. [C3]. Given $\epsilon > 0$ and $n \geq 2$, there exist $\delta = \delta(\epsilon, n) > 0$ and $\rho = \rho(\epsilon, n) > 0$, such that if M^n has $\mathrm{Ric}_{M^n} \geq -(n-1)$ and $0 < r_0 \leq \rho$ with

$$\mathrm{Vol}(B_{r_0}(x)) \geq (1-\delta)\,\mathrm{V}_0^n(r_0)\,, \tag{1.2}$$

then

$$d_{GH}(B_{r_0}(x), B_{r_0}(0)) < \epsilon\, r_0\,, \tag{1.3}$$

where $B_{r_0}(0) \subset \mathbf{R}^n$.

We also have the following converse.

Theorem 1.4. [C3]. Given $\epsilon > 0$ and $n \geq 2$, there exist $\delta = \delta(\epsilon, n) > 0$ and $\rho = \rho(\epsilon, n) > 0$, such that if M^n has $\mathrm{Ric}_{M^n} \geq -(n-1)$ and $0 < r_0 \leq \rho$ with

$$d_{GH}(B_{r_0}(x), B_{r_0}(0)) < \delta\, r_0\,, \tag{1.5}$$

then

$$\mathrm{Vol}(B_{r_0}(x)) \geq (1-\epsilon)\,\mathrm{V}_0^n(r_0)\,, \tag{1.6}$$

where $B_{r_0}(0) \subset \mathbf{R}^n$.

As mentioned in the introduction Bochner used the formula (0.1) to give a bound for b_1 on closed manifolds with nonnegative Ricci curvature. In the late seventies Gromov (see also Gallot) showed that there exists an $\epsilon = \epsilon(n) > 0$ such that any closed n-manifold with $\mathrm{Ric}_M \mathrm{diam}_M^2 > -\epsilon$ has $b_1 \leq n$. Gromov also conjectured the following.

Theorem 1.7. [C3]. There exists an $\epsilon = \epsilon(n) > 0$ such that if M^n is a closed n-dimensional manifold with $\mathrm{Ric}_M\, \mathrm{diam}_M^2 > -\epsilon$ and $b_1(M) = n$, then M is homeomorphic to a torus.

For further discussion of other work related to this see [C3].

We will think of Theorems 1.1 and 1.4 as regularity results which parallel Allard's classical regularity theorem for minimal submanifolds. To explain this point further we make the following definition inspired in part by the classical work of Reifenberg.

Let (Z, d_Z) be a complete metric space. We will say that Z satisfies the (ϵ, ρ, n)-$\mathcal{G}_{\mathcal{R}}$ condition at $z \in Z$ if for all $0 < \sigma < \rho$ and all $y \in B_{\rho-\sigma}(z)$,

$$d_{GH}(B_\sigma(y), B_\sigma(0)) < \epsilon\, \sigma\,, \tag{1.8}$$

where $B_{r_0}(0) \subset \mathbf{R}^n$.

The above two theorems has the following corollary.

COROLLARY 1.9. Given $\epsilon > 0$ and $n \geq 2$, there exist $\delta = \delta(\epsilon, n) > 0$ and $\rho = \rho(\epsilon, n) > 0$, such that if M^n has $\mathrm{Ric}_{M^n} \geq -(n-1)$ and $0 < r_0 \leq \rho$ with

$$d_{GH}(B_{2r_0}(x), B_{2r_0}(0)) < \delta\, r_0\,, \tag{1.10}$$

or

$$\mathrm{Vol}(B_{2r_0}(x)) \geq (1-\delta)\, \mathrm{V}_0^n(2r_0)\,, \tag{1.11}$$

then M^n satisfies the (ϵ, r_0, n)-$\mathcal{G}_{\mathcal{R}}$ condition at x.

THEOREM 1.12. [ChC2]. Given $\epsilon > 0$ and $n \geq 2$, there exists $\delta > 0$ such that if (Z, d_Z) is a complete metric space and Z satisfies the (δ, r_0, n)-$\mathcal{G}_{\mathcal{R}}$ condition at $z \in Z$ then there exists a bi-Hölder homeomorphism $\Phi : B_{\frac{r_0}{2}}(z) \to B_{\frac{r_0}{2}}(0)$ such that for all z_1, $z_2 \in Z$,

$$r_0^{-\epsilon}\, |\Phi(z_1) - \Phi(z_2)|^{1+\epsilon} \leq d_Z(z_1, z_2) \leq r_0^{\epsilon}\, |\Phi(z_1) - \Phi(z_2)|^{1-\epsilon}\,. \tag{1.13}$$

Using the above results one can show the following stability theorem.

THEOREM 1.14. [C3], [ChC2]. If M^n is a closed n-manifold, then there exists an $\epsilon = \epsilon(M) > 0$ such that if N^n is an n-manifold with $\mathrm{Ric}_N \geq -(n-1)$ and $d_{GH}(M, N) < \epsilon$ then M and N are diffeomorphic.

A very useful result about the structure on a small but definite scale of smooth manifolds with a lower Ricci curvature bound is the following theorem whose implications will be explained in more detail in the next section.

THEOREM 1.15. [ChC1]. Given $\epsilon > 0$ and $n \geq 2$, there exist $\rho = \rho(\epsilon, n) > 0$, $\delta = \delta(\epsilon, n) > 0$ such that if M^n has $\mathrm{Ric}_{M^n} \geq -(n-1)$, $0 < r_0 \leq \rho$, and

$$(2\, r_0)^{-n}\, \mathrm{Vol}(B_{2r_0}(p)) \geq (1-\delta)\, r_0^{-n}\, \mathrm{Vol}(B_{r_0}(p))\,, \tag{1.16}$$

then

$$d_{GH}(B_{2r_0}(p) \setminus B_{r_0}(p), B_{2r_0}(v) \setminus B_{r_0}(v)) < \epsilon\, r_0\,. \tag{1.17}$$

Here v is the vertex of some metric cone $(0, \infty) \times_r X$, for some length space X.

As a particular consequence of this theorem together with the monotonicity, (0.5), we get that if M^n has nonnegative Ricci curvature and Euclidean volume growth, i.e. $V_M = \lim_{r_0 \to \infty} r_0^{-n}\, \mathrm{Vol}(B_{r_0}(x)) > 0$, then every tangent cone at infinity of M is a metric cone. Here a tangent cone at infinity is a rescaled limit of $(M, x, r_i^{-2} g)$ where $r_i \to \infty$.

Another useful result is the following almost splitting theorem.

Theorem 1.18. [ChC1]. Given $\epsilon > 0$ and $n \geq 2$, there exist $\rho = \rho(\epsilon, n) > 0$, $\delta = \delta(\epsilon, n) > 0$ such that if M^n has $\mathrm{Ric}_{M^n} \geq -(n-1)$, $0 < r_0 \leq \rho$, x, $y \in \partial B_{r_0}(z)$, and

$$e_{x,y}(z) \leq \delta\, r_0\,, \tag{1.19}$$

then there exists some metric space X, $\overline{z} \in X \times \mathbf{R}$ such that

$$d_{GH}(B_{\delta\, r_0}(\overline{z}), B_{\delta\, r_0}(z)) < \epsilon\, r_0\,. \tag{1.20}$$

2. Singular spaces

We will now describe some results concerning spaces, (M^n_∞, m_∞) which are pointed Gromov-Hausdorff limits of sequences of manifolds satisfying (0.11). If in addition

$$\mathrm{Vol}(B_1(m_i)) \geq v > 0, \tag{2.1}$$

then we say that the sequence is noncollapsing.

To begin with, recall that M^n_∞ is a *length space* of Hausdorff dimension at most n and for all $r > 0, y_i \in M^n_i$, with $y_i \to y_\infty$, we have by [C3] (see also [ChC2], Theorem 2.13) the following which was conjectured by Anderson-Cheeger

$$\lim_{i\to\infty} \mathcal{H}^n(B_r(y_i)) = \mathcal{H}^n(B_r(y_\infty)), \tag{2.2}$$

where $\mathcal{H}^\alpha$ denotes α-dimensional Hausdorff measure. (Note that for smooth n-dimensional manifolds $\mathcal{H}^n$ is just the Riemmanian volume). As a consequence of this we have that the volume comparison also holds for noncollapsed limit spaces. That is for all $x \in M_\infty$ and all $0 < s \leq t$

$$\frac{\mathcal{H}^n(B_s(x))}{\mathcal{H}^n(B_t(x))} \geq \frac{\mathrm{V}^n_\Lambda(s)}{\mathrm{V}^n_\Lambda(t)}\,. \tag{2.3}$$

By definition, a *tangent cone*, $(T_x M^n_\infty, x_\infty)$, at $x \in M^n_\infty$, is any rescaled pointed Gromov-Hausdorff limit, $(T_x M^n, x_\infty, d_\infty)$, of some sequence, $(M^n_\infty, x, r_i^{-1} d_\infty)$, where $r_i \to 0$. Here, d_i, d_∞ denotes the metric (i.e. distance function) on M^n_i, M^n_∞. It follows from Gromov's compactness theorem that tangent cones exist at all points, $x \in M^n_\infty$, but according to [ChC2], Example 5.37, they need not be unique even when (2.1) hold. However, when (2.1) holds, every tangent cone is a metric cone, $C(X)$, on some length space, X, with $\mathrm{diam}(X) \leq \pi$; see [ChC2], Theorem 2.2.

The set of points, $\mathcal{R}$, for which the tangent cone is unique and isometric to $\mathbf{R}^k$ for some k, is called the regular set. The complement of the regular set is called singular set and denoted by $\mathcal{S}$. For $k \leq n-1$, let $\mathcal{S}_k$ denote the subset of singular points of M^n_∞ consisting of points for which no tangent cone splits off a factor, $\mathbf{R}^{k+1}$, isometrically. Then one can show using [C3] that $\bigcup_k \mathcal{S}_k = \mathcal{S}$. By Theorem 1.13 of [ChC2], have $\dim \mathcal{S}_k \leq k$, where *dim* denotes Hausdorff dimension. Moreover, when (2.1) holds, we have $\mathcal{S} \subset \mathcal{S}_{n-2}$; see [ChC2], Theorem 4.1.

The ϵ-*regular* set, $\mathcal{R}_\epsilon$, is by definition, the set of points for which some tangent cone satisfies, $d_{GH}(B_1(x_\infty), B_1(0)) < \epsilon$, where x_∞ is the vertex of $T_x M^n_\infty$ and $0 \in \mathbf{R}^n$. There exists $\epsilon(n) > 0$, such that for $\epsilon < \epsilon(n)$ the subset, $\mathcal{R}$, is homeomorphic to a smooth n-dimensional manifold and the restriction of the metric

to $\mathcal{R}_\epsilon$ is bi-Hölder equivalent to a smooth Riemannian metric with exponent $\alpha(\epsilon)$ satisfying $\alpha(\epsilon) \to 1$ as $\epsilon \to 0$. Note that for all $\epsilon > 0$, we have $M^n_\infty = \mathcal{S}_{n-2} \bigcup \mathcal{R}_\epsilon$, although $\mathcal{S}_{n-2} \bigcap \mathcal{R}_\epsilon$ need not be empty.

If (2.1) hold and some tangent cone at a point is isometric to $\mathbf{R}^n$, then every tangent cone at the point has this property. Clearly, $\mathcal{R} = \cap_{\epsilon>0}\mathcal{R}_\epsilon$, but $\mathcal{R}$ need not be open. For the results of this paragraph and the preceding one, see Section 2 of [ChC2].

In case (0.11) is strengthened to

$$|\mathrm{Ric}_{M^n_i}| \leq n - 1, \tag{2.4}$$

we have, using in part a theorem of Anderson, $\mathcal{R}_{\epsilon(n)} = \mathcal{R}$, for some $\epsilon(n) > 0$. In particular, $\mathcal{R}$ is open and $\mathcal{S} \subset \mathcal{S}_{n-2}$ is closed in this case. Moreover, $\mathcal{R}$ is a $C^{1,\alpha}$ Riemannian manifold — a C^∞ Riemannian manifold if in addition to (2.1) and (2.4), we assume that M^n_i is Einstein for all i. For above mentioned results, see Section 3 of [ChC2].

If the manifolds, M^n_i, are Kähler, and (0.11), (2.1) holds, then the tangent cones, $T_x M^n_\infty$, have natural complex structures on their smooth parts. If in addition to (2.1), (2.4) M_i is a convergent sequence of Kähler Einstein metrics on a fixed complex manifold then $\mathcal{H}^{n-4}(\mathcal{S}) < \infty$. Moreover, there exists $A \subset \mathcal{S}_{n-4} \setminus \mathcal{S}_{n-5}$, such that $\mathcal{H}^{n-4}(\mathcal{S}_{n-4} \setminus A) = 0$ and at points of A, the tangent cone is unique and isometric to $\mathbf{R}^{n-4} \times C(S^4/\Gamma_x)$. Where $S^3 \setminus \Gamma_x$ is a 3-dimensional space form and the order of Γ_x is bounded by a constant, $c(n, \mathcal{H}(B_1(x))$. Moreover, the isometry, between $T_x M^n_\infty$ and $\mathbf{R}^{n-4} \times C(\mathbf{S}^3/\Gamma_x) = \mathbf{C}^{k-2} \times \mathbf{C}/\Gamma$ is complex. See [ChCT1], [ChCT2] for these results.

When (2.1) does not hold it is useful to construct renormalized limit measures, ν, on limit spaces. These measures were first constructed by Fukaya. In the noncollapsed case, the limit measure exists without the necessity of passing to a subsequence, or of renormalizing the measure. The unique limit measure is just Hausdorff measure, $\mathcal{H}^n$; see Theorem 5.9 of [ChC2]. (If, for the sake of consistency, one does renormalize the measure, then one obtains a multiple of $\mathcal{H}^n$, where as usual, the normalization factor depends on the choice of base point.) However, in the collapsed case, the renormalized limit measure on the limit space can depend on the particular choice of subsequence; see Example 1.24 of [ChC2]. The renormalized limit measures play an important role in [ChC3], [ChC4], for instance in connection with the theory of the Laplace operator on limit spaces.

Let M^n satisfy $\mathrm{Ric}_{M^n} \geq (n-1)\,\Lambda$, and define the renormalized volume function, by

$$\underline{V}(x,r) := \underline{\mathrm{Vol}}(B_r(x)) := \frac{1}{\mathrm{Vol}(B_1(p))}\mathrm{Vol}(B_r(x))\,. \tag{2.5}$$

By combining the proof of Gromov's compactness theorem with the proof of the theorem of Arzela-Ascoli, we obtain:

Given any sequence of pointed manifolds, (M^n_i, m_i), for which (0.11) holds, there is a subsequence,(M^n_j, m_j), convergent to some (M_∞, m_∞) in the pointed Gromov-Hausdorff sense, and a continuous function $\underline{V}_\infty : M_\infty \times \mathbf{R}_+ \to \mathbf{R}_+$, such

that if $y_j \in M_j^n, z \in M_\infty$ and $y_j \to z$, then for all $R > 0$,

$$\underline{V}_j(y_j, R) \to \underline{V}_\infty(z, R) \qquad \text{(uniformly on } B_{R_1}(p) \times [0, R_2]) \,. \tag{2.6}$$

One can then show that there is a unique Radon measure, ν, on M_∞ such that for all, z, R,

$$\nu(B_R(z)) = \underline{V}_\infty(z, R) \,. \tag{2.7}$$

In particular ν satisfies the inequality (0.4) with ν in place of Vol.

By a result in Section 2 of [ChC2] $\nu(\mathcal{S}) = 0$ also in the collapsed case. Even though the measure ν depend on the particular subsequence it is shown in [ChC3] that the collection, $\{\nu\}$, of all renormalized limit measures determines a well defined measure class i.e. ν_a is absolutely continuous with respect to ν_b, for all ν_a, ν_b.

We will conclude this section with some well known important open problems.

Conjecture 2.8. The interior of $M_\infty^n \setminus \mathcal{S}_{n-4}$ is a topological manifold.

Conjecture 2.9. Suppose that (M_i^n, x_i) are Einstein, (2.1) and (2.4) hold then $\mathcal{S} = \mathcal{S}_{n-4}$ and there exist some $C < \infty$ such that for all $x \in M_\infty$

$$\mathcal{H}^{n-4}(B_1(x) \cap \mathcal{S}) \le C \,. \tag{2.10}$$

Moreover outside a subset $A \subset \mathcal{S}$ with $\mathcal{H}^{n-4}(A) = 0$ the tangent cone is unique and isometric to some orbifold.

3. Analysis on manifolds

Analysis on manifolds with nonnegative (Ricci) curvature is generally believed to resemble that on Euclidean space. For instance, in the 1970's S.T. Yau showed that the classical Liouville theorem generalized to this setting; cf. the gradient estimate of Cheng and Yau. Yau conjectured that in fact the spaces of harmonic functions of polynomial growth were finite dimensional on these spaces. It is a classical fact that $\mathcal{H}_d(\mathbf{R}^n)$ consists of harmonic polynomials of degree at most d. That is they are spanned by the spherical harmonics where the eigenvalue on $\mathbf{S}^{n-1}$ is given in terms of d.

Recall that for an open manifold M, $d > 0$, and $x \in M$ a function u is in $\mathcal{H}_d(M)$, the space of harmonic functions of polynomial growth of degree at most d, if $\Delta u = 0$ and there exists $C < \infty$ such that

$$|u| \le C\,(1 + r^d) \,. \tag{3.1}$$

We note that there have been numerous interesting results in this area over the years, including work of Li and Tam, Donnelly and Fefferman, and others (see [CM1], [CM4] for detailed references).

This conjecture of Yau was settled affirmatively in [CM5]. We also obtained sharper (polynomial) estimates for the dimension and finite dimensionality in more general settings, including for certain harmonic sections of bundles and stationary varifolds.

In the case of Euclidean volume growth we obtained in [CM2] an asymptotic description of harmonic functions with polynomial growth. Namely, recall that asymptotically $u \in \mathcal{H}_d(\mathbf{R}^n)$ behaves like its highest order homogeneous part; i.e.

u is asymptotically a homogeneous separation of variables solution. We showed in [CM2] that this asymptotic description remain true in the Euclidean volume growth case. In [CM7] we showed that the Euclidean volume growth case studied in [CM2] was indeed the worst case.

In [ChCM], we showed that if a manifold with nonnegative Ricci curvature has $\dim \mathcal{H}_1(M) \geq k+1$ then any tangent cone at infinity, M_∞, splits isometrically as $\bar{M}_\infty \times \mathbf{R}^k$, where $\mathbf{R}^k$ has the standard flat metric. Combining this with the volume convergence of [C3] yields a rigidity theorem. Namely, if M^n is complete, $\mathrm{Ric}_M \geq 0$, and $\dim \mathcal{H}_1(M) = n+1$, then M^n is isometric to $\mathbf{R}^n$.

In [CM7], we gave polynomial bounds for the dimension of the space of polynomial growth L-harmonic functions, where L is a second order divergence form uniformly elliptic on a manifold with the doubling property and a scale invariant lower bound on some Neumann eigenvalue (not necessarily the first).

In the case of nonnegative Ricci curvature, we obtained the following "Weyl type" bounds.

THEOREM 3.2. [CM7]. Let $\mathrm{Ric}_M \geq 0$, and $d \geq 1$. There exist a constant C and $o(d^{n-1})$ depending on n such that

$$\dim \mathcal{H}_d(M^n) \leq C\, V_M\, d^{n-1} + o(d^{n-1})\,, \tag{3.3}$$

where $\lim_{d\to\infty} d^{1-n}\, o(d^{n-1}) = 0$.

This theorem yields immediately a Siegel type theorem for the field of rational functions on a Kähler manifold with nonnegative Ricci curvature. By looking at a cone over a compact manifold, Theorem 3.2 gives "Weyl type" eigenvalue estimates for compact manifolds; see [CM7].

In [CM6], we showed that the finite dimensionality continues to hold in other related settings; e.g., given a mean value inequality and the doubling property. As applications, in [CM6] we got finite dimensionality results for spaces of harmonic sections of bundles with nonnegative curvature and on stationary varifolds with Euclidean volume growth; see also [CM4]. The methods are flexible and give many related, but slightly different, theorems; the two following results are representative.

THEOREM 3.4. [CM6]. Let $\mathrm{Ric}_M \geq 0$, and E^k a rank k Hermitian vector bundle over M with nonnegative curvature. For all $d \geq 1$,

$$\dim \mathcal{H}_d(M^n, E) \leq C\, k\, d^{n-1}\,, \tag{3.5}$$

where $C = C(n) < \infty$.

The final result that we will mention is a weak Bernstein type theorem for stationary varifolds in Euclidean space. This is obtained from a bound on the space of harmonic functions of a given degree of growth since the restrictions of the coordinate functions are harmonic. Recall that stationary varifolds satisfies a mean value inequality.

THEOREM 3.6. [CM6]. Let $\Sigma^n \subset \mathbf{R}^N$ be a stationary n-rectifiable varifold with density at least one almost everywhere on its support and bounded from above by V_Σ, then Σ must be contained in some affine subspace of dimension at most $C = C(n)$.

References

[ChC1] J. Cheeger and T.H. Colding, Lower bounds on the Ricci curvature and the almost rigidity of warped products, Annals of Math. vol. 144, N. 1 (1996) 189-237.

[ChC2] J. Cheeger and T.H. Colding, On the structure of spaces with Ricci curvature bounded below; I, Jour. of Diff. Geometry, 46 (1997) 406-480.

[ChC3] J. Cheeger and T.H. Colding, On the structure of spaces with Ricci curvature bounded below; II (to appear).

[ChC4] J. Cheeger and T.H. Colding, On the structure of spaces with Ricci curvature bounded below; III (to appear).

[ChC5] J. Cheeger and T.H. Colding, Almost Rigidity of Warped Products and the Structure of Spaces with Ricci Curvature Bounded Below, C.R. Acad. Sci. Paris t. 320, Serie 1 (1995) 353-357.

[ChCM] J. Cheeger, T.H. Colding, and W.P. Minicozzi II, Linear Growth Harmonic Functions on Complete Manifolds with Nonnegative Ricci Curvature, GAFA v. 5, n. 6 (1995) 948-954.

[ChCT1] J. Cheeger, T.H. Colding, and G. Tian, Constraints on singularities under Ricci curvature bounds, C. R. Acad. Sci. Paris, t. 324, Serie 1 (1997) 645-649.

[ChCT2] J. Cheeger, T.H. Colding, and G. Tian, On the singularities of spaces with bounded Ricci curvature, to appear.

[C1] T.H. Colding, Shape of Manifolds with Positive Ricci Curvature, Invent. Math. 124 Fasc. 1-3 (1996) 175-191.

[C2] T.H. Colding, Large Manifolds with Positive Ricci Curvature, Invent. Math. 124 Fasc. 1-3 (1996) 193-214.

[C3] T.H. Colding, Ricci Curvature and Volume Convergence, Annals of Math. vol. 145, N. 3 (1997) 477-501.

[C4] T.H. Colding, Stability and Ricci Curvature, C.R. Acad. Sci. Paris t. 320, Serie 1 (1995) 1343-1347.

[C5] T.H. Colding, Aspects of Ricci Curvature, Comparison Geometry, MSRI Publ., vol. 30, Cambridge University Press (1997) 83-98.

[CM1] T.H. Colding and W.P. Minicozzi II, On function theory on spaces with a lower Ricci curvature bound, Math. Research Letters 3 (1996) 241-246.

[CM2] T.H. Colding and W.P. Minicozzi II, Harmonic functions with polynomial growth, J. Diff. Geom. v. 46, no. 1 (1997) 1-77.

[CM3] T.H. Colding and W.P. Minicozzi II, Large scale behavior of the kernel of Schrödinger operators, Amer. J. Math. 119 (1997) 1355-1398.

[CM4] T.H. Colding and W.P. Minicozzi II, Generalized Liouville properties of manifolds, Math. Res. Lett. 3 (1996) 723-729.

[CM5] T.H. Colding and W.P. Minicozzi II, Harmonic functions on manifolds, Annals of Math. vol. 146, no. 3 (1997) 725-747.

[CM6] T.H. Colding and W.P. Minicozzi II, Liouville theorems for harmonic sections and applications, Comm. Pure Appl. Math., vol. 51, no. 2 (1998) 113-138.

[CM7] T.H. Colding and W.P. Minicozzi II, Weyl type formulas for harmonic functions, Inventiones Math., 131, 2. (1998) 257 – 298.

[Ga] S. Gallot, Volumes, courbure de Ricci et convergence des variétes [d'après T.H. Colding et Cheeger-Colding] Séminarie Nicolas Bourbaki, 50ème année 1997-98 exposé 835, p. 1-33.

Tobias H. Colding
Courant Institute
of Mathematical Sciences
251 Mercer Street
New York, NY 10012
USA

Lefschetz Fibrations in Symplectic Geometry

S. K. Donaldson

Keywords and Phrases: Lefschetz pencil, symplectic manifolds

Section 1. Linear systems.

One of the cornerstones of complex geometry is the link between *positivity* of curvature and *ampleness*. Let X be a compact complex manifold and $L \to X$ be a holomorphic line bundle over X. Suppose that L has a unitary connection whose curvature form is $-2\pi i\omega$ where ω is a positive $(1,1)$-form on X. Then for large k the line bundle L^k has many holomorphic sections. More precisely, the holomorphic sections define a projective embedding of X (Kodaira). This provides a passage from the discussion of abstract complex manifolds to concrete algebro-geometric models. If one chooses some linear subspace of the holomorphic sections of L^k one gets birational maps into smaller projective spaces: for example, if X has complex dimension 2 then it may be immersed in $\mathbf{CP}^4$ with a finite number of double points and can be mapped to a hypersurface in $\mathbf{CP}^3$ with "ordinary singularities". Perhaps the simplest case of all is that covered by Bertini's Theorem: if the intersection of the zero sets of all the holomorphic sections is empty (i.e if the linear system has no base points), then the zero-set of a generic section is a smooth hypersurface in X.

A familiar instance of these ideas occurs when X is a compact Riemann surface and we consider two sections of L^k. The ratio of these sections is a meromorphic function on X, i.e. a branched covering map $X \to \mathbf{CP}^1$. If the sections are sufficiently general then this map has a very simple local structure. There are a finite number of critical values $b_\alpha \in S^2 \cong \mathbf{CP}^1$; for each α there is a corresponding critical point $x_\alpha \in X$; the restriction of the map to $X \setminus \{x_\alpha\}$ is a covering map and around each point x_α the map is modelled, in suitable local co-ordinates, on the standard example $z \mapsto z^2$. The Riemann surface X, and the branched covering map, can be recovered from the data consisting of the configuration of points b_α in the Riemann sphere and the *monodromy*, a homomorphism from the fundamental group of the punctured sphere $S^2 \setminus \{b_\alpha\}$ to the permutation group of d objects, the sheets of the covering.

More generally one has the notion of a "Lefschetz pencil" on a higher dimensional complex variety. The ratio of two, sufficiently general, sections of our line bundle is a meromorphic function, which defines a holomorphic map from the complement of a codimension-2 submanifold $A \subset X$. Alternatively, we get a map from the blow-up $\tilde{X}$ of X along A to $\mathbf{CP}^1$. Again there are a finite number of critical points, around which the map is modelled on the quadratic function $(z_1, \ldots, z_n) \mapsto z_1^2 + \cdots + z_n^2$. The monodromy in this situtation is more complicated:

parallel transport around loops in the punctured sphere defines a homomorphism into the mapping class group of the fibre, that is, the group of diffeomorphisms of the fibre modulo isotopy.

The main purpose of this contribution is to report on extensions of these classical ideas in complex geometry to the more general setting of *symplectic* manifolds. In the next section we will describe some of the main results, and the ideas of the proofs, and in the final section we will make some more general comments.

Section 2.1 The symplectic case: techniques.

Now let (V, ω) be a compact symplectic manifold, of dimension $2n$. We suppose that the de Rham cohomology class $[\omega]$ is an integral class, so we can choose a C^∞ line bundle $L \to V$ with $c_1(L) = [\omega]$. To mimic the classical case we can begin by choosing an almost-complex structure on V, algebraically compatible with the symplectic form. There is also a unitary connection on L with curvature $-i\omega$. This gives a notion of a "holomorphic" section of L: we can define a $\overline{\partial}$-operator on L, using the connection and the almost-complex structure and a (local) section s is holomorphic if $\overline{\partial} s = 0$. The problem is that, for $n > 1$ and for generic almost-complex structures, one expects this definition to be vacuous in that there will be no non-trivial holomorphic sections. This is because the generalised Cauchy-Riemann equation $\overline{\partial} s = 0$ is over-determined and the compatibility condition which is needed to have local solutions is precisely the integrability of the almost-complex structure. This contrasts with the much-studied theory of holomorphic maps from a Riemann surface into an almost-complex manifold, where the integrability of the almost-complex structure does not make a great difference to the local theory of solutions. The way around this problem is to study certain approximately holomorphic sections of the line bundle, or more precisely of the tensor power L^k, for large k. The integer k is the crucial parameter throughout the discussion, and it is convenient to work with the family of Riemannian metrics g_k on V, where g_k is associated to the symplectic form $k\omega$ in the usual fashion. Thus the diameter of (V, g_k) is $O(\sqrt{k})$ but on a ball of g_k-radius 1 the almost-complex structure is close to the standard flat model, as $k \to \infty$. For any $C > 0$ we set

$$H_{k,C} = \{s \in \Gamma(L^k) : \|\overline{\partial} s\|, \|\nabla \overline{\partial} s\|, \|\nabla^2 \overline{\partial} s\| \leq C\sqrt{k}^{-1} \|s\|\},$$

where all norms are L^∞, computed using the metric g_k. Elements of $H_{k,C}$ are a substitute for the holomorphic sections in the classical case. One shows that, for a suitable C depending on the geometry of X and for $k >> 0$, there is a large supply of sections in $H_{k,C}$. This is quite elementary: the sections can be constructed as linear combinations of sections concentrated in balls, of a fixed g_k-radius, in X. The fundamental model, which serves as a prototype for the influence of curvature on holomorphic geometry, is the case of $\mathbf{C}^n$, with the standard flat metric. Then there is a holomorphic section σ of the corresponding Hermitian line bundle over $\mathbf{C}^n$ which decays rapidly at infinity:

$$|\sigma(z)| = e^{-|z|^2}.$$

(In the several complex variables literature this phenomena is often described in the equivalent language of weighted L^2 norms.)

The classical theory sketched in Section 1 involves, beyond the existence of a plentiful supply of holomorphic sections, holomorphic versions of various familiar transversality statements. For example Bertini's theorem is a holomorphic version of Sard's Theorem, and the proof of the existence of Lefschetz pencils is a variant of the proof of the existence of Morse functions. The price that must be paid for the freedom to work with only approximately holomorphic sections is that one needs refinements of such transversality theorems, involving explicit estimates. These have interest in their own right. Results in this direction were obtained by Yomdin [5], although the precise statements needed are somewhat different. For simplicity consider the case of a holomorphic function f on the unit ball B^{2n} in $\mathbf{C}^n$. The familiar Sard theorem asserts that the regular values of f are dense in $\mathbf{C}$. For $\epsilon > 0$ we say that a point $w \in \mathbf{C}$ is an ϵ-regular value of f over a subset $K \subset B^{2n}$ if there are no points $z \in K$ with both $|f(z) - w| < \epsilon$ and $|\partial f(z)| < \epsilon$. The question we wish to answer is this: given any $w' \in \mathbf{C}$, how close is w' to an ϵ-regular value of f? An answer is provided by the following statement:

PROPOSITION. *There is a constant p such that for all holomorphic functions f on B^{2n} with $\|f\|_{L^\infty} \leq 1$, any $w' \in \mathbf{C}$ and $\epsilon \in (0, 1/2)$ there is an ϵ-regular value w of f over the interior ball $\frac{1}{2}B^{2n}$ with*

$$|w - w'| \leq (\log(\epsilon^{-1}))^p \epsilon.$$

One way of thinking of this result is that one would really like to have the stronger and simpler statement

$$|w - w'| \leq C.\epsilon,$$

but the factor $\log(\epsilon)^{-1}$ grows slowly as $\epsilon \to 0$, so the result stated in the Proposition serves almost as well. (The writer does not know whether the stronger statement is true or not.) The point to make is that the standard proofs of Sard's Theorem are not well-adapted to proving quantitative refinements of this kind, and the proof goes, following the idea of Yomdin, by approximating the function by polynomials and using facts about the complexity of real-algebraic sets.

SECTION 2.2: THE SYMPLECTIC CASE: MAIN RESULTS.

The first result proved using these ideas [2] is, roughly speaking, a symplectic version of Bertini's Theorem. For large k it is shown that one can choose an approximately holomorphic section $s \in H_{k,C}$ such that $|\partial s| > \delta \|s\|$ on the zero-set Z_s, for a fixed $\delta > 0$, independent of k. It follows that Z_s is a symplectic submanifold of V, i.e. the restriction of the form ω is nondegenerate on Z_s. Thus we have

THEOREM. *If (V, ω) is a compact symplectic manifold and $[\omega]$ is an integral class then for large k the Poincaré dual of $k[\omega]$ is represented by a symplectic codimension-2 submanifold.*

(This result can be compared with a much sharper but more specialised theorem of Taubes [4], proved shortly afterward using the Seiberg-Witten equations,

which asserts that if V is a symplectic 4-manifold with $b_2^+(V) > 1$ then $-c_1(V)$ is represented by a symplectic surface in V.)

This result was extended, in a number of directions, by D. Aroux [1]. One striking extension was a result about the asymptotic uniqueness of the symplectic submanifold which is constructed. Let us return to the classical case of complex geometry. Then it is clear that the discriminant set $\Delta \subset H^0(L^k)$, consisting of sections whose zero-set is not transverse, is a complex analytic variety. In particular the complement of Δ is *connected*. Thus if s_0, s_1 are two sections whose zero-sets Z_0, Z_1 are transverse, there is a isotopy of the ambient manifold taking Z_0 to Z_1. This is an important principle in complex geometry. It means, for exammple, that at the level of differential topology one can unambigously talk about "a smooth hypersurface of degree d in $\mathbf{CP}^n$", without specifying precisely which polynomial is used in the definition. Of course this contrasts with the case of real algebraic geometry, where the topological type does vary with the polynomial. Aroux's extension of this principle to the symplectic case made use of the notion of an asymptotic sequence (s_k), $s_k \in \Gamma(L^k)$, of sections of the kind whose existence is established in the result above. He proves that

THEOREM.
If J, J' are two almost-complex structures on V compatible with ω and $(s_k), (s'_k)$ are two asymptotic sequences of approximately holomorphic sections, with respect to J, J', then for large k there is a symplectic isotopy of V mapping the zero set of s_k to that of s'_k.

We now go on to consider the symplectic analogue of the classical theory of "pencils", generated by a pair of sections.

DEFINITION. *A topological Lefschetz pencil on a symplectic manifold (V, ω) is given by the following data.*

(i) a codimension-2 symplectic submanifold $A \subset V$, (ii) a finite set of points $x_\alpha \in V \setminus A$, (iii) a differentiable map $f : V \setminus A \to S^2$ such that f is a submersion on $V \setminus A \setminus \{x_\alpha\}$.

The map f is required to conform to the following standard models. At a point $a \in A$ we can choose local complex co-ordinates z_i such that A is locally defined by the equations $z_1 = z_2 = 0$ and f is given locally by the map $(z_1, \ldots, z_n) \mapsto z_1/z_2 \in \mathbf{CP}^1 \cong S^2$. At a point x_α we can choose local complex co-ordinates on V, and a complex co-ordinate centred on $f(x_\alpha) \in S^2$ such that the map is given locally by $(z_1, \ldots, z_n) \mapsto z_1^2 + \cdots + z_n^2$.

Then we have

THEOREM. *If (V, ω) is a symplectic manifold with $[\omega]$ integral then for large k V admits a topological Lefschetz pencil, in which the fibres are symplectic subvarieties, representing the Poincaré dual of $k[\omega]$.*

There is also an asymptotic uniqueness statement, in the same vein as Aroux's result.

Let us spell out more explicitly what this theorem says, concentrating on the case of a 4-dimensional symplectic manifold V. In this case A is just a finite set

of points. If $\tilde{V}$ is the blow-up of V at these points then f defines a smooth map from $\tilde{V}$ to S^2 whose generic fibre is a compact Riemann surface, of genus g say. There are a finite number of singular fibres, passing through the critical points x_α. Writing $b_\alpha = f(x_\alpha)$, we have a differentiable monodromy

$$\rho : \pi_1(S^2 \setminus \{b_\alpha\}) \to \Gamma_{g,h}$$

where h is the number of points in A and $\Gamma_{g,h}$ is the mapping class group of a surface of genus g with h marked points. The standard theory from complex geometry adapts with little change to give a detailed "local" picture of this monodromy. Fix a base point P in $S^2 \setminus \{b_\alpha\}$ and let γ_α be a standard generator of the fundamental group of the punctured sphere, winding once around b_α. Then the monodromy $\rho(\gamma_\alpha)$ is the *Dehn twist* $T[\delta_\alpha]$ of the fibre $\Sigma = f^{-1}(P)$ in a *vanishing cycle* $\delta_\alpha \subset \Sigma$, an embedded loop in Σ. The conclusion is, in brief, that any smooth 4-manifold which admits a symplectic structure with integral periods may be constructed from combinatorial data consisting of a marked Riemann surface Σ and a suitable set of loops δ_α in Σ, the essential requirement being that the product

$$T[\delta_1] \circ T[\delta_2] \circ \cdots \circ T[\delta_N]$$

be the identity in the mapping class group.

Section 3: Discussion.

There are a number of directions in which one could hope to extend and fill-out and these results. First one could look at linear systems of other dimensions, and hope to prove analogues of the classical theorems in complex geometry for these cases. There is recent work of Aroux in this direction. Second, we should mention work of Gompf which provides a converse to the discussion above of symplectic Lefschetz pencils on 4-manifolds. Gompf shows that the total space of a 4-dimensional topological Lefschetz fibration, satisfying some mild numerical conditions, admits a symplectic structure. Putting everything together, one might expect to get a completely combinatorial-topological description of symplectic 4-manifolds. One natural set of questions involves the dependence on the parameter k. For example if $Z_k \subset V$ is a symplectic hypersurface representing $k[\omega]$ one can hope to describe a hypersurface $Z_{2k} \subset V$ representing $2k[\omega]$ by considering deformations of a singular space $Z_k \cup Z'_k$, where Z'_k is obtained by applying a generic small perturbation to Z_k. There is a similar discussion for Lefschetz pencils. If this project was carried through one could refine the asymptotic uniqueness statement into an explicit "stabilisation" mechanism.

The implications of this line of work for symplectic topology are unclear at present. Although it seems quite practical to "reduce" many fundamental questions about symplectic manifolds to combinatorial-topological problems, the latter seem very difficult to attack directly. Many of the difficulties have to do with the complexity of the braid groups which act as automorphisms of the fundamental groups of punctured Riemann spheres. In the case of 4-manifolds we may consider the set $\mathcal{R}$ of the representations of $\pi_1 = \pi_1(S^2 \setminus \{b_\alpha\})$ into the mapping class group $\Gamma_{g,h}$ which correspond to topological Lefschetz fibrations (i.e. which map

each standard generator to a Dehn twist). Then the (spherical) braid group B_N acts on $\mathcal{R}$ and the natural invariants of the fibrations are the orbits under this action. Thus one would like are computable invariants which detect these orbits. One can view this problem as a higher dimensional analogue of the classical theory for branched covers of the Riemann sphere. In that case the issue is to classify transitive representations of π_1 into the permutation group on d-elements which map each generator to a transposition, modulo the action of the braid group, and a theorem of Hurwicz states that there is just one orbit.

On the positive side, it is worth pointing out that there are many similarities between the ideas that occur in this theory and those developed in the past few years by P. Seidel [3], in the framework of symplectic Floer theory. In both cases the Dehn twists, and their higher-dmensional generalisations, play a prominent role. These generalised Dehn twists are defined as follows. If L is an embedded Lagrangian m-sphere in a symplectic manifold W^{2m} a neighbourhood of L in W can be identified with a neighbourhood of the zero section in TS^m. The one can define a compactly-supported diffeomorphism of TS^m, using the geodesic flow composed with the antipodal map. This can then be transported to a symplectomorphism $\tau_L : W \to W$. Seidel shows that, when $m = 2$ in many cases the squares τ_L^2 are not symplectically isotopic to the identity, although they are so differentiably, thus revealing some of the rich structure of symplectic mapping class groups. On the other hand, these same symplectomorphisms occur as the monodromy of Lefschetz pencils of a symplectic manifold of dimension $2(m + 1)$. They may also be analysed from the point of view of the braid group action on a Lefschetz pencil for W: the diffeomorphism arises from the action of a standard braid on a pair of identical monodromies. For these, and other, reasons, it seems possible that there may be some fruitful interaction between the symplectic Floer theory and the general Lefschetz pencil description of symplectic manifolds.

References

1. D. Aroux, *Asymptotically holomorphic families of symplectic submanifolds*, Geometric and Functional Analysis (to appear).
2. S. K. Donaldson, *Symplectic submaifolds and almost-complex geometry*, Jour. Differential Geometry **44** (1996), 666-705.
3. P. Seidel, *The symplectic isotopy problem*, Oxford D.Phil. Thesis (1997).
4. C. H. Taubes, *More constraints on symplectic manifolds from the Seiberg-Witten equations*, Math. Research Letters **2** (1995), 9-14.
5. B. Yomdin, *The geometry of critical and near-critical values of differentiable mappings*, Math. Annalen **104** (1983), 495-515.

S. K. Donaldson
Department of Mathematics,
Stanford University,
CA 94305, USA

Geometry and Analytic Theory of Frobenius Manifolds

Boris Dubrovin

Abstract. Main mathematical applications of Frobenius manifolds are in the theory of Gromov - Witten invariants, in singularity theory, in differential geometry of the orbit spaces of reflection groups and of their extensions, in the hamiltonian theory of integrable hierarchies. The theory of Frobenius manifolds establishes remarkable relationships between these, sometimes rather distant, mathematical theories.

1991 Mathematics Subject Classification: 32G34, 35Q15, 35Q53, 20F55, 53B50

WDVV equations of associativity is the problem of finding of a quasihomogeneous, up to at most quadratic polynomial, function $F(t)$ of the variables $t = (t^1, \dots, t^n)$ and of a constant nondegenerate symmetric matrix $\left(\eta^{\alpha\beta}\right)$ such that the following combinations of the third derivatives $c^\gamma_{\alpha\beta}(t) := \eta^{\gamma\epsilon}\partial_\epsilon\partial_\alpha\partial_\beta F(t)$ for any t are structure constants of an asociative algebra $A_t = \text{span}\,(e_1, \dots, e_n)$, $e_\alpha \cdot e_\beta = c^\gamma_{\alpha\beta}(t)e_\gamma$, $\alpha, \beta = 1, \dots, n$ with the unity $e = e_1$ (summation w.r.t. repeated indices will be assumed). These equations were discovered by physicists E.Witten, R.Dijkgraaf, E.Verlinde and H.Verlinde in the beginning of '90s. I invented Frobenius manifolds as the coordinate-free form of WDVV.

1. Definition of Frobenius manifold (FM).

1.1. Frobenius algebra (over a field k; we mainly consider the case $k = \mathbf{C}$) is a pair $(A, <\ ,\ >)$, where A is a commutative associative k-algebra with a unity e, $<\ ,\ >$ is a symmetric nondegenerate *invariant* bilinear form $A \times A \to k$, i.e. $< a \cdot b, c >=< a, b \cdot c >$ for any $a, b \in A$. A *gradation of the charge* d on A is a k-derivation $Q : A \to A$ such that $< Q(a), b > + < a, Q(b) >= d < a, b >$, $d \in k$. More generally, graded of the charge $d \in k$ Frobenius algebra $(A, <\ ,\ >)$ over a graded commutative associative k-algebra R by definition is endowed with two k-derivations $Q_R : R \to R$ and $Q_A : A \to A$ satisfying the properties $Q_A(\alpha a) = Q_R(\alpha)a + \alpha Q_A(a)$, $\alpha \in R$, $a \in A < Q_A(a), b > + < a, Q_A(b) > -Q_R < a, b >= d < a, b >$, $a, b \in A$.

1.2. Frobenius structure of the charge d on the manifold M is a structure of a Frobenius algebra on the tangent spaces $T_tM = (A_t, <\ ,\ >_t)$ depending (smoothly, analytically etc.) on the point $t \in M$. It must satisfy the following axioms.

FM1. The metric $<\ ,\ >_t$ on M is flat (but not necessarily positive definite). Denote ∇ the Levi-Civita connection for the metric. The unity vector field e must be covariantly constant, $\nabla e = 0$.

FM2. Let c be the 3-tensor $c(u,v,w) :=< u\cdot v, w >$, u, v, $w \in T_tM$. The 4-tensor $(\nabla_z c)(u,v,w)$ must be symmetric in u, v, w, $z \in T_tM$.

FM3. A linear vector field $E \in Vect(M)$ must be fixed on M, i.e. $\nabla\nabla E = 0$, such that the derivations $Q_{Func(M)} := E$, $Q_{Vect(M)} := \mathrm{id} + \mathrm{ad}_E$ introduce in $Vect(M)$ the structure of graded Frobenius algebra of the given charge d over the graded ring $Func(M)$ of (smooth, analytic etc.) functions on M. We call E *Euler vector field*.

Locally, in the flat coordinates $t^1,\dots,t^n$ for the metric $<\ ,\ >_t$, a FM with diagonalizable (1,1)-tensor ∇E is described by a solution $F(t)$ of WDVV associativity equations, where $\partial_\alpha\partial_\beta\partial_\gamma F(t) =< \partial_\alpha\cdot\partial_\beta, \partial_\gamma >$, and vice versa. We will call $F(t)$ *the potential* of the FM (physicists call it *primary free energy*; in the setting of quantum cohomology it is called Gromov - Witten potential [KM]).

1.3. DEFORMED FLAT CONNECTION $\tilde\nabla$ on M is defined by the formula $\tilde\nabla_u v := \nabla_u v + z\,u\cdot v$. Here u, v are two vector fields on M, z is the parameter of the deformation. (In [Gi1] another normalization is used $\tilde\nabla \mapsto \hbar\tilde\nabla$, $\hbar = z^{-1}$.) We extend this to a meromorphic connection on the direct product $M\times\mathbf{C}$, $z\in\mathbf{C}$, by the formula $\tilde\nabla_{d/dz} v = \partial_z v + E\cdot v - z^{-1}\mu\, v$ with $\mu := 1/2(2-d)\cdot\mathbf{1} - \nabla E$, other covariant derivatives are trivial. Here u, v are tangent vector fields on $M\times\mathbf{C}$ having zero components along $\mathbf{C}\ni z$. The curvature of $\tilde\nabla$ is equal to zero. This can be used as a definition of FM [Du3]. So, there locally exist n independent functions $\tilde t_1(t;z),\dots,\tilde t_n(t;z)$, $z\neq 0$, such that $\tilde\nabla\, d\tilde t_\alpha(t;z) = 0$, $\alpha = 1,\dots,n$. We call these functions *deformed flat coordinates*.

2. EXAMPLES OF FMS appeared first in 2D topological field theories [W1, W2, DVV].

2.0. TRIVIAL FM: $M = A_0$ for a graded Frobenius algebra A_0. The potential is a cubic, $F_0(t) = \frac{1}{6} < 1,(t)^3 >$, $t\in A_0$. Nontrivial examples of FM are

2.1. FM WITH GOOD ANALYTIC PROPERTIES. They are analytic perturbations of the cubic. That means that, in an appropriate system of flat coordinates $t = (t', t'')$, where all the components of t' have $Lie_E t' = \mathrm{const}$, all the components of t'' have $Lie_E t'' \neq \mathrm{const}$, we have $F(t) = F_0(t) + \sum_{k,\,l\geq 0} A_{k,l}(t'')^l e^{k\,t'}$ and the series converges in some neiborghood of $t''=0$, $t' = -\infty$.

2.2. K.SAITO THEORY OF PRIMITIVE FORMS AND FROBENIUS STRUCTURES ON UNIVERSAL UNFOLDINGS OF QUASIHOMOGENEOUS SINGULARITIES. Let $f_s(x)$, $s = (s_1,\dots,s_n)$ be the universal unfolding of a quasihomogeneous isolated singularity $f(x)$, $x\in\mathbf{C}^N$, $f(0) = f'(0) = 0$. Here n is the Milnor number of the singularity. The Frobenius structure on the base $M\ni s$ of the universal unfolding can be easily constructed [BV] using the theory of primitive forms [Sai2]. For the example [DVV] of the A_n singularity $f(x) = x^{n+1}$ the universal unfolding reads $f_s(x) = x^{n+1} + s_1 x^{n-1} + \dots + s_n$, $M = \mathbf{C}^n \ni (s_1,\dots,s_n)$. On the FM $e = \partial/\partial s_n$, $E = \sum (k+1)s_k\partial/\partial s_k$, the metric has the form

$$< \partial_{s_i}, \partial_{s_j} >= -(n+1)\operatorname*{res}_{x=\infty} \frac{\partial f_s(x)/\partial s_i\, \partial f_s(x)/\partial s_j}{f_s'(x)} \tag{2.1}$$

the multiplication is defined by

$$< \partial_{s_i}\cdot\partial_{s_j}, \partial_{s_k} >= -(n+1)\operatorname*{res}_{x=\infty} \frac{\partial f_s(x)/\partial s_i\, \partial f_s(x)/\partial s_j\, \partial f_s(x)/\partial s_k}{f_s'(x)}. \tag{2.2}$$

This is a polynomial FM. The deformed flat coordinates are given by oscillatory integrals

$$\tilde{t}_c = \frac{1}{\sqrt{z}} \int_c e^{z\, f_s(x)} dx \tag{2.3}$$

Here c is any 1-cycle in $\mathbf{C}$ that goes to infinity along the direction $\operatorname{Re} z\, f_s(x) \to -\infty$.

2.3. QUANTUM COHOMOLOGY of a $2d$-dimensional smooth projective variety X is a Frobenius structure of the charge d on a domain $M \subset H^*(X, \mathbf{C})/2\pi i H^2(X, \mathbf{Z})$ (we assume that $H^{odd}(X) = 0$ to avoid working with supermanifolds, see [KM]). It is an analytic perturbation in the sense of n.2.1 of the cubic for $A_0 = H^*(X)$ defined by a generating function of the genus zero Gromov - Witten (GW) invariants of X [W1, W2, MS, RT, KM, Beh]. They are defined as intersection numbers of certain cycles on the moduli spaces of stable maps [KM]

$$X_{[\beta],l} := \left\{\beta : (S^2, p_1, \dots, p_l) \to X, \quad \text{given homotopy class } [\beta] \in H_2(X; \mathbf{Z})\right\}.$$

The holomorphic maps β of the Riemann sphere S^2 with $l \geq 1$ distinct marked points are considered up to a holomorphic change of parameter. The markings define evaluation maps $p_i : X_{[\beta],l} \to X, \quad (\beta, p_1, \dots, p_l) \mapsto \beta(p_i)$.

$$F(t) = F_0(t) + \sum_{[\beta]\neq 0} \sum_l \left\langle e^{t''} \right\rangle_{[\beta],l} \exp \int_{S^2} \beta^*(t')$$
$$\left\langle e^t \right\rangle_{[\beta],l} := \frac{1}{l!} \int_{X_{[\beta],l}} p_1^*(t) \wedge \ldots \wedge p_l^*(t) \tag{2.4}$$

for $t = (t', t'') \in H^*(X)$, $t' \in H^2(X)/2\pi i H^2(X, \mathbf{Z})$, $t'' \in H^{*\neq 2}(X)$. This potential together with the Poincaré pairing on $TM = H^*(X)$, the unity vector field $e = 1 \in H^0(X)$, the Euler vector field $E(t) = \sum (1 - q_\alpha) t^\alpha e_\alpha + c_1(X), \quad t = t^\alpha e_\alpha, \ e_\alpha \in H^{2q_\alpha}(X)$ gives the needed Frobenius structure. The deformed flat coordinates are generating functions of certain "gravitational descendents" [Du5], see also [DW, Ho, Gi1] $\tilde{t}_\alpha(t; z) = \sum_{p=0}^\infty \sum_{[\beta],l} \left\langle z^{\mu+p} z^{c_1(X)} \tau_p(e_\alpha) \otimes 1 \otimes e^{t''} \right\rangle_{[\beta],l} e^{\int_{S^2} \beta^*(t')}$, $\alpha = 1, \dots, n = \dim H^*(X)$, $\mu(e_\alpha) = (q_\alpha - d/2) e_\alpha$, The definition of the descendents $\left\langle \tau_{p_1}(a_1) \otimes \tau_{p_2}(a_2) \otimes \ldots \otimes \tau_{p_l}(a_l) \right\rangle_{[\beta],l}$ see in [W2], [KM]. The definition of GW invariants can be extended on a certain class of compact symplectic varieties X using Gromov's theory [Gr] of pseudoholomorphic curves, see [W2, MS, RT].

3. CLASSIFICATION OF SEMISIMPLE FMS.

3.1. DEFINITION. A point $t \in M$ is called *semisimple* if the algebra on $T_t M$ is semisimple. A connected FM M is called semisimple if it has at least one semisimple point. Classification of semisimple FMs can be reduced, by a nonlinear change of coordinates, to a system of ordinary differential equations. First we will describe these new coordinates.

3.2. CANONICAL COORDINATES on a semisimple FM. Denote $u_1(t), \ \dots, \ u_n(t)$ the roots of the characteristic polynomial of the operator of multiplication by the Euler vector field $E(t)$ ($n = \dim M$). Denote $M^0 \subset M$ the open subset where

all the roots are pairwise distinct. It turns out [Du2] that the functions $u_1(t)$, $\dots$, $u_n(t)$ are independent local coordinates on $M^0 \neq \emptyset$. In these coordinates $\partial_i \cdot \partial_j = \delta_{ij}\partial_i$, where $\partial_i := \partial/\partial u_i$, and $E = \sum_i u_i \partial_i$. The local coordinates u_1, $\dots$, u_n on M^0 are called *canonical*.

3.3. DEFORMED FLAT CONNECTION IN THE CANONICAL COORDINATES AND ISOMONODROMY DEFORMATIONS. Staying in a small ball on M^0, let us order the canonical coordinates and choose the signs of the square roots $\psi_{i1} := \sqrt{< \partial_i, \partial_i >}$, $i = 1, \dots, n$. The orthonormal frame of the normalized idempotents ∂_i establishes a local trivialization of the tangent bundle TM^0. The deformed flat connection $\tilde{\nabla}$ in TM^0 is recasted into the following flat connection in the trivial bundle $M^0 \times \mathbf{C} \times \mathbf{C}^n$

$$\tilde{\nabla}_i = \partial_i - z\, E_i - V_i, \qquad \tilde{\nabla}_{d/dz} = \partial_z - U - z^{-1}V, \tag{3.1}$$

other components are obvious. Here the $n \times n$ matrices E_i, U, $V = (V_{ij})$ read $(E_i)_{kl} = \delta_{ik}\delta_{il}$, $U = \operatorname{diag}(u_1, \dots, u_n)$, $V = \Psi\, \mu\, \Psi^{-1} = -V^T$ where the matrix $\Psi = (\psi_{i\alpha})$ satisfying $\Psi^T \Psi = \eta$ is defined by $\psi_{i\alpha} := \psi_{i1}^{-1} \partial t_\alpha / \partial u_i$, $i, \alpha = 1, \dots, n$. The skew-symmetric matrices V_i are determined by the equations $[U, V_i] = [E_i, V]$.

Flatness of the connection (3.1) reads as the system of commuting time-dependent Hamiltonian flows on the Lie algebra $so(n) \ni V$ equipped with the standard linear Poisson bracket

$$\partial_i V = \{V, H_i(V; u)\}, \quad i = 1, \dots, n \tag{3.2}$$

with the quadratic Hamiltonians $H_i(V; u) = \frac{1}{2} \sum_{j \neq i} \frac{V_{ij}^2}{u_i - u_j}$, $i = 1, \dots, n$. For the first nontrivial case $n = 3$ (3.2) can be reduced to a particular case of the classical Painlevé-VI equation. The monodromy of the operator $\tilde{\nabla}_{d/dz}$ (i.e., the monodromy at the origin, the Stokes matrix, and the central connection matrix, see definitions in [Du3, Du5]) does not change with small variations of a point $u = (u_1, \dots, u_n) \in M$.

3.4. PARAMETRIZATION OF SEMISIMPLE FMS BY MONODROMY DATA OF THE DEFORMED FLAT CONNECTION. We now reduce the above system of nonlinear differential equations to a linear boundary value problem of the theory of analytic functions. First we will describe the set of parameters of the boundary value problem.

3.4.1. MONODROMY AT THE ORIGIN (defined also for nonsemisimple FMs) consists of:

- a linear n-dimensional space $\mathcal{V}$ with a symmetric nondegenerate bilinear form $< , >$, a skew-symmetric linear operator $\mu : \mathcal{V} \to \mathcal{V}$, $< \mu(a), b > + < a, \mu(b) >= 0$, and a marked eigenvector e_1 of μ, $\mu(e_1) = -d/2\, e_1$. In main examples the operator μ will be diagonalizable.

- A linear operator $R : \mathcal{V} \to \mathcal{V}$ satisfying the following properties: (1) $R = R_1 + R_2 + \dots$ where $R_k(\mathcal{V}_\lambda) \subset \mathcal{V}_{\lambda+k}$ for the root decomposition of $\mathcal{V} = \oplus_\lambda \mathcal{V}_\lambda$, $\mu(v_\lambda) = \lambda v_\lambda$ for $v_\lambda \in \mathcal{V}_\lambda$. (2) $\{Rx, y\} + \{x, Ry\} = 0$ for any x, $y \in \mathcal{V}$ where $\{x, y\} := \left\langle e^{\pi i \mu} x, y \right\rangle$.

3.4.2. STOKES MATRIX is an arbitrary $n \times n$ upper triangular matrix $S = (s_{ij})$ with $s_{ii} = 1$, $i = 1, \dots, n$. We treat it as a bilinear form $< a, b >_S := a^T S b$, $a, b \in \mathbf{C}^n$.
3.4.3. CENTRAL CONNECTION MATRIX is an isomorphism $C : \mathbf{C}^n \to \mathcal{V}$ satisfying $< a, b >_S = < Ca, e^{\pi i \mu} e^{\pi i R} Cb >$ for any $a, b \in \mathbf{C}^n$. The matrices S and C are defined up to a transformation $S \mapsto DSD$, $C \mapsto CD$, $D = \operatorname{diag}(\pm 1, \dots, \pm 1)$.
3.4.4. RIEMANN - HILBERT BOUNDARY VALUE PROBLEM (RH b.v.p.). Let us fix a radius $R > 0$ and an argument $0 \le \varphi < 2\pi$. Denote $\ell = \ell_+ \cup \ell_-$ the oriented line $\ell_+ = \{z \,|\, \arg z = \varphi\}$, $\ell_- = \{z \,|\, \arg z = \varphi + \pi\}$. It divides the complex z-plane into two halfplanes Π_{right} and Π_{left}. For a given $u = (u_1, \dots, u_n)$ with $u_i \neq u_j$ for $i \neq j$ and for given monodromy data we are looking for: (1) $n \times n$ matrix-valued functions $\Phi_{\text{right}}(z)$, $\Phi_{\text{left}}(z)$ analytic for $|z| > R$ and $z \in \Pi_{\text{right}}$ and $z \in \Pi_{\text{left}}$ resp., continuous up to the boundaries $|z| = R$ or $z \in \ell$ and satisfying $\Phi_{\text{right/left}}(z) = 1 + O(1/z)$ for $|z| \to \infty$ within the correspondent half-plane $\Pi_{\text{right/left}}$; (2) $n \times n$ matrix-valued function $\Phi_0(z)$ (with values in $Hom(\mathcal{V}, \mathbf{C}^n)$) analytic for $|z| < R$ and continuous up to the boundary $|z| = R$, such that $\det \Phi_0(0) \neq 0$. The boundary values of the functions must satisfy

$$\Phi_{\text{right}}(z) e^{zU} = \Phi_{\text{left}}(z) e^{zU} S \quad \text{for } z \in \ell_+,\ |z| > R;$$

$$\Phi_{\text{right}}(z) e^{zU} = \Phi_{\text{left}}(z) e^{zU} S^T \quad \text{for } z \in \ell_-,\ |z| > R;$$

$$\Phi_{\text{right}}(z) e^{zU} = \Phi_0(z) z^\mu z^R C \quad \text{for } |z| = R,\ z \in \Pi_{\text{right}};$$

$$\Phi_{\text{left}}(z) e^{zU} S = \Phi_0(z) z^\mu z^R C \quad \text{for } |z| = R,\ z \in \Pi_{\text{left}}.$$

The branchcut in the definition of the multivalued functions z^μ and z^R is chosen along ℓ_-. For solvability of the above RH b.v.p. we have also to require the complex numbers $u_1, \dots, u_n$ to be ordered in such a way, depending on φ, that

$$\mathcal{R}_{jk} := \{z = -ir(\bar{u}_j - \bar{u}_k) | r \ge 0\} \subset \Pi_{\text{left}} \quad \text{for } j < k. \tag{3.3}$$

Denote $\mathcal{U}(\varphi) \subset \mathbf{C}^n$ the set of all points $u = (u_1, \dots, u_n)$ with $u_i \neq u_j$ for $i \neq j$ satisfying (3.3). Let $\mathcal{U}_0(\varphi)$ be the subset of points $u \in \mathcal{U}(\varphi)$ such that: (1) the RH b.v.p. is solvable and (2) all the coordinates of the vector $\Phi_0(0) e_1$ are distinct from zero. It can be shown (cf. [Mi], [Mal]) that the solution $\Phi_{\text{right/left}} = \Phi_{\text{right/left}}(z; u)$, $\Phi_0 = \Phi_0(z; u)$ of the RH b.v.p depends analytically on $u \in \mathcal{U}_0(\varphi)$. Let $\Phi_0(z; u) = \sum_{p=0}^{\infty} \phi_p(u) z^p$. Denote (only here) $(\ ,\ )$ the standard sum of squares quadratic form on $\mathbf{C}^n$. Choose a basis $e_1, e_2, \dots, e_n$ of eigenvectors of μ, $\mu(e_\alpha) = \mu_\alpha e_\alpha$, $\mu_1 = -d/2$, and put $\eta_{\alpha\beta} := < e_\alpha, e_\beta >$, $(\eta^{\alpha\beta}) := (\eta_{\alpha\beta})^{-1}$.

THEOREM 1 [Du2, Du3, Du5]. *The formulae*

$$t_\alpha(u) = (\phi_0(u) e_\alpha, \phi_1(u) e_1), \quad t^\alpha = \eta^{\alpha\beta} t_\beta,\ \alpha = 1, \dots, n,$$

$$F = 1/2\,[(\phi_0 t, \phi_1 t) - 2(\phi_0 t, \phi_1 e_1) + (\phi_1 e_1, \phi_2 e_1) - (\phi_3 e_1, \phi_0 e_1)]$$

$$E(t) = \sum_{\alpha=1}^{n} (1 + \mu_1 - \mu_\alpha) t^\alpha \partial_\alpha + \sum_\alpha (R_1)^\alpha_1 \partial_\alpha$$

define on $\mathcal{U}_0(\varphi)$ a structure of a semisimple FM $Fr(\mathcal{V}, <\ ,\ >, \mu, e_1, R, S, C)$. Any semisimple FM locally has such a form.

3.5. REMARK. The columns of the matrices $\Phi_0(z;u)z^\mu z^R$ and $\Phi_{\text{right}}(z;u)e^{zU}$ correspond to two different bases in the space of deformed flat coordinates. The first basis is a deformation, $z \to 0$, of the original flat coordinates, $\tilde{t} = [t + O(z)]z^\mu z^R$. The second one, defined only in the semisimple case, corresponds to a system of deformed flat coordinates given by oscillatory integrals (see (2.3) and Section 6 below).

3.6. GLOBAL STRUCTURE OF SEMISIMPLE FMS AND ACTION OF THE BRAID GROUP ON THE MONODROMY DATA. Let B_n be the group of braids with n strands. We will glue globally the FM from the charts described in n.3.4 with different S and C. So, for brevity, we redenote here the charts $Fr(\mathcal{V}, <\ ,\ >, \mu, e_1, R, S, C) =: Fr(S, C)$. The charts will be labelled by braids $\sigma \in B_n$. By definition in the chart $Fr(S^\sigma, C^\sigma)$ the functions $t^\alpha(u)$, $F(u)$ are obtained as the result of analytic continuation from $Fr(S, C)$ along the braid σ. The action $S \mapsto S^\sigma$, $C \mapsto C^\sigma$ of the standard generators $\sigma_1, \dots, \sigma_{n-1}$ of B_n is given by $S^{\sigma_i} = KSK$, $C^{\sigma_i} = CK$ where the only nonzero entries of the matrix $K = K^{(i)}(S)$ are $K_{kk} = 1$, $k = 1, \dots, n$, $k \neq i, i+1$, $K_{i,i+1} = K_{i+1,i} = 1$, $K_{i,i} = -s_{i,i+1}$. Let $B_n(S, C) \subset B_n$ be the subgroup of all braids σ such that $S^\sigma = DSD$, $C^\sigma = CD$, $D = \text{diag}(\pm 1, \dots, \pm 1)$.

THEOREM 2 [Du3, Du5]. *Any semisimple FM has the form*

$$M = \cup_{\sigma \in B_n / B_n(S,C)} Fr(\mathcal{V}, <\ ,\ >, \mu, e_1, R, S^\sigma, C^\sigma)$$

where the glueing of the charts is given by the above action of B_n.

3.7. TAU-FUNCTION OF THE ISOMONODROMY DEFORMATION AND ELLIPTIC GW INVARIANTS. Like in n.2.2, the genus g GW invariants can be defined in terms of the intersection theory on the moduli space $X_{[\beta],l}(g)$ of stable maps $\beta : C_g \to X$ of curves of genus g with markings [KM, Beh]. It turns out that, assuming semisimplicity of quantum cohomology of X, the elliptic (i.e., of $g = 1$) GW invariants can still be expressed via isomonodromy deformations. To this end we define, following [JM], the τ-function $\tau(u_1, \dots, u_n)$ of a solution $V(u)$ of the system (3.2) by the quadrature of a closed 1-form $d \log \tau = \sum_{i=1}^n H_i(V(u); u) du_i$. We define G-function of the FM by $G = \log(\tau / J^{1/24})$ where $J = \det(\partial t^\alpha / \partial u_i) = \pm \prod_{i=1}^n \psi_{i1}(u)$.

THEOREM 3 [DZ2]. *For an arbitrary semisimple FM the G-function is the unique, up to an additive constant, solution to the system of [Ge] for the generating function of elliptic GW invariants satisfying $Lie_e G = 0$, $Lie_E G = \text{const}$.*

3.8. PROBLEM OF SELECTION OF SEMISIMPLE FMS WITH GOOD ANALYTIC PROPERTIES of n.2.1 is still open. Experiments for small n [Du3] show that such solutions are rare exceptions among all semisimple FMs. Analyticity of the G-function near the point $t' = -\infty$, $t'' = 0$ imposes further restrictions on M [DZ2]. To solve the problem one is to study the behaviour of solutions of the RH b.v.p. in the limits when two or more among the canonical coordinates merge. At the point $t' = -\infty$, $t'' = 0$ all $u_1 = \dots = u_n = 0$.

4. EXAMPLES OF MONODROMY DATA.

4.1. UNIVERSAL UNFOLDINGS OF ISOLATED SINGULARITIES. The subspace $M_0 \subset M$ consists of the parameters s for which the versal deformation $f_s(x)$ has $n = \dim M$ distinct critical values $u_1(s)$, ..., $u_n(s)$. These will be our canonical

coordinates. The monodromy at the origin is the classical monodromy operator [AGV] of the singularity, the Stokes matrix coincides with the matrix of the variation operator in the Gabrielov's distinguished basis of vanishing cycles (see [AGV]; we may assume that $\dim x \equiv 1 \,(\mathrm{mod} 4)$).

4.2. QUANTUM COHOMOLOGY OF FANO VARIETIES. The following two questions are to be answered in order to apply the above technique to the quantum cohomology of a variety X.

PROBLEM 1. When does the generating series (2.4) converge?

PROBLEM 2. For which X the quantum cohomology of X is semisimple?

Hopefully, in the semisimple case the convergence can be proved on the basis of the differential equations of n.3. To our opinion the problem 2 is more deep. A necessary condition to have a semisimple quantum cohomology is that X must be a Fano variety. It was conjectured to be also a sufficient condition [TX], [Man1]. We analyze below one example and suggest some more modest conjecture describing also a part of the monodromy data.

4.2.1. QUANTUM COHOMOLOGY OF PROJECTIVE SPACES. For $X = \mathbf{P}^d$: (1) the monodromy at the origin is given by the bilinear form $< e_\alpha, e_\beta >= \delta_{\alpha+\beta,d+2}$ in $H^*(X) = \mathcal{V} = \mathrm{span}\,(e_1, \dots, e_{d+1})$, the matrix $\mu = 1/2\mathrm{diag}\ (-d, 1-d, \dots, d-1, d)$ and R is the matrix of multiplication by the first Chern class $R = R_1 = c_1(X)$, $R\,e_\alpha = (d+1)e_{\alpha+1}$ for $\alpha \le d$, $R\,e_{d+1} = 0$. With obvious modifications these formulae work also for any variety X with $H^{\mathrm{odd}}(X) = 0$ (see [Du3]). (2) The Stokes matrix $S = (s_{ij})$ has the form

$$s_{ij} = \begin{pmatrix} d+1 \\ j-i \end{pmatrix} \text{ for } i \le j, \quad s_{ij} = 0 \text{ for } i > j. \tag{4.1}$$

This form of Stokes matrix was conjectured in [CV], [Zas] but, to our knowledge, it was proved only in [Du5] for $d = 2$ and in [Guz] for any d. (3) The central connection matrix C has the form $C = C'C''$, $C' = \left(C'^{\alpha}_{\beta}\right)$, $C'' = \left(C''^{\beta}_{j}\right)$ where $C''^{\beta}_{j} = [2\pi i(j-1)]^{\beta-1}/(\alpha-1)!$, $j, \beta = 1, \dots, d+1$, $C'^{\alpha}_{\beta} = \frac{(-1)^{d+1}}{(2\pi)^{\frac{d+1}{2}} i^{\bar d}} \begin{cases} A_{\alpha-\beta}(d), & \alpha \ge \beta \\ 0, & \alpha < \beta \end{cases}$ with $\bar d = 1$ for d =even and $\bar d = 0$ for d =odd where the numbers $A_0(d) = 1$, $A_1(d)$, $\dots$, $A_d(d)$ are defined from the Laurent expansion for $x \to 0$: $1/x^{d+1} + A_1(d)/x^d + \dots + A_d(d)/x + O(1) = (-1)^{d+1}\Gamma^{d+1}(-x)e^{-\pi i \bar d x}$. Observe that (4.1) is the Gram matrix of the bilinear form $\chi(E,F) := \sum_k (-1)^k \dim Ext^k(E,F)$ in the basis given by a particular full system $E_j = \mathcal{O}(j-1)$, $j = 1, \dots, d+1$ of exceptional objects in the derived category $Der^b(Coh(\mathbf{P}^d))$ of coherent sheaves on $\mathbf{P}^d$ [Rud]. The columns of the matrix C'' are the components of the Chern character $\mathrm{ch}\,(E_j) = e^{2\pi i c_1(E_j)}$, $j = 1, \dots, d+1$. The geometrical meaning of the matrix C' remains unclear. In other charts of the FM S^σ and $C^\sigma = C'C''^\sigma$, $\sigma \in B_n$, have the same structure for another full system $E_1^\sigma, \dots, E_{d+1}^\sigma \in Der^b(Coh(\mathbf{P}^d))$ of exceptional objects, where the action of the braid group $(E_1, \dots, E_{d+1}) \mapsto (E_1^\sigma, \dots, E_{d+1}^\sigma)$ is described in [Rud]. Warning: the points of the FM corresponding to the restricted quantum cohomology [MM], where $t \in H^2(\mathbf{P}^d)$, do not belong to the chart $Fr(S,C)$ with the matrices S and C as above!

4.2.2. CONJECTURE. We say that a Fano variety X is *good* if $Der^b(Coh(X))$ admits, in the sense of [BP], a full system of exceptional objects $E_1, \ldots, E_n$, $n = \dim H^*(X)$. Our conjecture is that (1) the quantum cohomology of X is semisimple *iff* X is a good Fano variety; (2) the Stokes matrix $S = (s_{ij})$ is equal to $s_{ij} = \chi(E_i, E_j)$, $i, j = 1, \ldots, n$; (3) the central connection matrix has the form $C = C'C''$ when the columns of C'' are the components of $\mathrm{ch}\,(E_j) \in H^*(X)$ and $C' : H^*(X) \to H^*(X)$ is some operator satisfying $C'(c_1(X)a) = c_1(X)C'(a)$ for any $a \in H^*(X)$.

For $X = \mathbf{P}^d$ the validity of the conjecture follows from n.4.2.1 above. The conjecture probably can be derived from more general conjecture [Kon] about equivalence of $Der^b(Coh(X))$ to the Fukaya category of the mirror pair X^* of X. According to it (see also [EHX, Gi1, Gi2]) the basis of horizontal sections of $\tilde{\nabla}$ corresponding to the columns of $\Phi_{\mathrm{right}}(z; u)e^{zU}$ coincides with the oscillatory integrals of the Fukaya category of X^*. However, we do not know who is the first factor C' of the connection matrix in this general setting.

5. INTERSECTION FORM of a FM is a bilinear symmetric pairing on T^*M defined by $(\omega_1, \omega_2)|_t := i_{E(t)}(\omega_1 \cdot \omega_2)$, $\omega_1, \omega_2 \in T_t^*M$. *Discriminant* is the locus $\Sigma = \{t \in M \mid \det(\ ,\ )_t = 0\}$. On $M \setminus \Sigma$ the inverse to $(\ ,\ )$ determines a flat metric and, thus, a local isometry $\pi : \left(M \setminus \Sigma, (\ ,\ )^{-1}\right) \to \mathbf{C}^n$ where $\mathbf{C}^n$ is equipped with a constant complex Euclidean metric $(\ ,\ )_0$. This local isometry is called *period mapping* (our terminology copies that of the singularity theory where the geometrical structures with the same names live on the bases of universal unfoldings, see [AGV]). The image $\pi(\Sigma)$ is a collection of nonisotropic hyperplanes in $\mathbf{C}^n$. Multivaluedness of π is described by the *monodromy representation* $\pi_1(M \setminus \Sigma) \to Iso\,(\mathbf{C}^n, (\ ,\ )_0)$ (for $d \neq 1$ to the orthogonal group $O\,(\mathbf{C}^n, (\ ,\ )_0)$). The image $W(M)$ of the representation is called *monodromy group* of the FM M. In the semisimple case it is always an extension of a reflection group (see details in [Du5]). Our hope is that, for a semisimple FM M with good analytic properties, the monodromy group acts discretely in some domain $\Omega \subset \mathbf{C}^n$, and M is identified with a branched covering of the quotient $\Omega/W(M)$.

5.1. EXAMPLES of a FM with $W(M)$= finite irreducible Coxeter group W acting in $\mathbf{R}^n$ [Du3]. These are polynomial FMs, $M = \mathbf{C}^n/W$, constructed in terms of the theory of invariant polynomials of W. Conjecturally, all polynomial semisimple FMs are equivalent to the above and to their direct sums.

This construction was generalized in [DZ1] to certain extensions of affine Weyl groups and in [Ber] to Jacobi groups of the types A_n, B_n, G_2. For the quantum cohomology of $\mathbf{P}^2$ the monodromy group is isomorphic to $PSL_2(\mathbf{Z}) \times \{\pm 1\}$ [Du5].

6. MIRROR CONSTRUCTION represents certain system of deformed flat coordinates on a semisimple FM by oscillatory integrals $I_j(u; z) = \frac{1}{\sqrt{z}} \int_{Z_j} e^{z\lambda(p;u)} dp$, $\tilde{\nabla} I_j(u; z) = 0$, $j = 1, \ldots, n$ having the phase function $\lambda(p; u)$ depending on the parameters $u = (u_1, \ldots, u_n)$ defined on a certain family of open Riemann surfaces $\mathcal{R}_u \ni p$ realized as a finite-sheeted branched covering $\lambda : \mathcal{R}_u \to D \subset \mathbf{C}$ over a domain in the complex plain. The ramification points of $\mathcal{R}_u$, i.e., the critical values of the phase function, are $u_1, \ldots, u_n$. The 1-cycles $Z_1, \ldots, Z_n$ on $\mathcal{R}_u$ go to infinity in a way that guarantees the

convergence of the integrals. The function $\lambda(p;u)$ satisfies an important property: for any two critical points $p_i^{1,2} \in \mathcal{R}_u$ with the same critical value u_i the equality $d^2\lambda(p_i^1;u)/dp^2 = d^2\lambda(p_i^2;u)/dp^2$ must hold true. The metric $<\ ,\ >$ and the trilinear form $c(a_1,a_2,a_3) :=< a_1 \cdot a_2, a_3 >$ are given by the residue formulae similar to (2.1), (2.2). The solutions $p = p(u;\lambda)$ of the equation $\lambda(p;u) = \lambda$ are the flat coordinates of the *flat pencil of the metrics* $(\ ,\ ) - \lambda <\ ,\ >$ on T^*M [Du3-Du5].

For the case when generically there is a unique critical point p_i over u_i for each i and $\mathcal{R}_u$ can be compactfied to a Riemann surface of a finite genus g, we arrive at the Hurwitz spaces of branched coverings [Du1, Du3].

The construction of the Riemann surfaces $\mathcal{R}_u$, of the phase function $\lambda(p;u)$ and of the cycles $Z_1, \ldots, Z_n$ is given in [Du5] by universal formulae assuming $\det(S+S^T) \neq 0$. In the quantum cohomology of a d-fold X the last condition is valid for $d =$ even. For $d =$ odd one has $\det(S-S^T) \neq 0$. In this case one can represent the deformed flat coordinates by oscillatory integrals with the phase function $\lambda(p,q;u) = \nu(p;u) + q^2$ depending on two variables p, q. The details will be published elsewhere.

7. GRAVITATIONAL DESCENDENTS is a physical name for intersection numbers $< \tau_{m_1}(a_1) \otimes \ldots \otimes \tau_{m_l}(a_l) >$ of the pull-back cocycles $p_1^*(a_1), \ldots, p_l^*(a_l)$ with the Mumford - Morita - Miller cocycles $\psi_1^{m_1}, \ldots, \psi_l^{m_l} \in H^*(X_{[\beta],l})$ [W2], [DW], [KM]. We will describe first their genus $g = 0$ generating function $\mathcal{F}_0(T) = \sum_{[\beta]} \sum_l \left\langle e^{\sum_{\alpha=1}^n \sum_{p=0}^\infty T^{\alpha,p}\tau_p(e_\alpha)} \right\rangle_{[\beta],l,g=0}$. Here $T = (T^{\alpha,p})$ are indeterminates (the coordinates on the "big phase space", according to the physical terminology). This function has the form $\mathcal{F}_0(T) = 1/2 \sum \Omega_{\alpha,p;\,\beta,q}\,(t(T))\,\tilde{T}^{\alpha,p}\tilde{T}^{\beta,q}$ where $\tilde{T}^{\alpha,p} = T^{\alpha,p}$ for $(\alpha,p) \neq (1,1)$, $\tilde{T}^{1,1} = T^{1,1} - 1$, the functions $\Omega_{\alpha,p;\,\beta,q}(t)$ on M are the coefficients of the expansion of the matrix valued function $\Omega_{\alpha\beta}(z,w;t) := (z+w)^{-1}\left[\left(\Phi_0^T(w;t)\Phi_0(z;t)\right)_{\alpha\beta} - \eta_{\alpha\beta}\right] = \sum_{p,\,q\geq 0} \Omega_{\alpha,p;\,\beta,q}(t) z^p w^q$, the vector function $t(T) = \left(t^1(T), \ldots, t^n(T)\right)$

$$t^\alpha(T) = T^{\alpha,0} + \sum_{q>0} T^{\beta,q}\nabla^\alpha \Omega_{\beta,q;\,1,0}(t)|_{t^\alpha = T^{\alpha,0}} + \ldots \tag{7.1}$$

is defined as the unique solution of the following fixed point equation $t = \nabla \sum_{\alpha,p} T^{\alpha,p}\Omega_{\alpha,p;\,1,0}(t)$.

The generating function $\mathcal{F}_1(T)$ of the genus $g = 1$ descendents has the form [DZ2], [DW], [Ge] $\mathcal{F}_1(T) = \left[G(t) + \frac{1}{24}\log\det M_{\alpha\beta}(t,\dot{t})\right]_{t=t(T),\,\dot{t}=\partial_{T^{1,0}}t(T)}$ where $G(t)$ is the G-function of the FM, the matrix $M_{\alpha\beta}(t,\dot{t})$ has the form $M_{\alpha\beta}(t,\dot{t}) = \partial_\alpha\partial_\beta\partial_\gamma F(t)\dot{t}^\gamma$, the vector function $t(T)$ is the same as above. The structure of the genus $g = 2$ corrections is still unclear, although there are some interesting conjectures [EX] related, in the case of quantum cohomology, to the Virasoro constraints for the full partition function

$$Z(T;\varepsilon) = \exp \sum_{g=0}^\infty \varepsilon^{2g-2}\mathcal{F}_g(T) \tag{7.2}$$

ε is a formal small parameter called *string coupling constant*.

8. INTEGRABLE HIERARCHIES of PDEs of the KdV type and FMs. The idea that FMs may serve as moduli of integrable hierarchies of evolutionary equations (see [W2], [Du2], [Du3]) is based on

(1) the theorem of Kontsevich - Witten identifying the partition function (7.2) in the case $X =$ point as the tau-function of a particular solution of the KdV hierarchy.

(2) The construction [Du2, Du3] of bihamiltonian integrable hierarchy of the Whitham type $\partial_{T^{\alpha,p}} t = \{t(X), H_{\alpha,p}\}_1 = K^{(0)}_{\alpha,p}(t, t_X)$ (the vector function in the r.h.s. depends linearly on the derivatives t_X) such that the full genus zero partition function is the tau-function of a particular solution (7.1) to the hierarchy. The solution is specified by the symmetry constraint $t_X - \sum T^{\alpha,p} \partial_{T^{\alpha,p-1}} t = 1$. The phase space of the hierarchy is the loopspace $\mathcal{L}(M) = \big\{\big(t^1(X), \dots, t^n(X)\big) \mid X \in S^1\big\}$, the first Hamiltonian structure is $\big\{t^\alpha(X), t^\beta(Y)\big\}_1 = \eta^{\alpha\beta}\delta'(X-Y)$, the second one $\{\ ,\ \}_2$ is determined [Du3] by the flat metric $(\ ,\ )$ according to the general scheme of [DN]. The Hamiltonians are $H_{\alpha,p} = \int \Omega_{\alpha,p;\,1,0}(t)\, dX$. Actually, any linear combination $\{\ ,\ \}_2 - \lambda\{\ ,\ \}_1$ with an arbitrary λ is again a Poisson bracket on the loop space since $(\ ,\ )$ and $<\ ,\ >$ form *a flat pencil of metrics* on T^*M [Du3, Du4] (this *bihamiltonian property* is a manifestation of integrability of the hierarchy, see [Mag], [Du4]).

What we want to construct is a deformation of the hierarchy of the form $\partial_{T^{\alpha,p}} t = K^{(0)}_{\alpha,p}(t, t_X) + \sum_{g\geq 1} \varepsilon^{2g} K^{(g)}_{\alpha,p}(t, t_X, \dots, t^{(2g+1)})$ where $K^{(g)}_{\alpha,p}$ are some vector valued polynomials in $t_X, \dots, t^{(2g+1)}$ with the coefficients depending on $t \in M$. All the equations of the hierachy must commute pairwise. The full partition function must be the tau-function of a particular solution to the hierachy. The first $g = 1$ correction for an arbitrary semisimple FM was constructed in [DZ]. Its bihamiltonian structure is described, for $d \neq 1$, by a nonlinear deformation of the Virasoro algebra with the central charge $c = 6\varepsilon^2(1-d)^{-2}[n - 4\mathrm{tr}\,\mu^2]$. For the FMs corresponding to the ADE Coxeter groups this formula gives the known result [FL] for the central charge of the classical W-algebra of the ADE-type $c = 12\varepsilon^2\rho^2$, where ρ is half of the sum of positive roots of the corresponding root system.

More recently it has been proved [DZ3] for a semisimple FM that the partition function (7.2) is annihilated, within the genus one approximation, by half of a Virasoro algebra described in terms of the monodromy data of the FM.

ACKNOWLEDGMENTS. I am grateful to D.Orlov for helpful discussion of derived categories, and to S.Barannikov and M.Kontsevich for fruitful conversations.

REFERENCES

[AGV] Arnol'd, V.I., Gusein-Zade, S.M. and Varchenko, A.N.: Singularities of Differentiable Maps, volumes I, II, Birkhäuser, Boston-Basel-Berlin, 1988.

[Beh] Behrend, K.: Gromov - Witten invariants in algebraic geometry, *Inv. Math.* **124** (1997) 601 - 627.

[Ber] Bertola, M.: Jacobi groups, Hurwitz spaces, and Frobenius manifolds, Preprint SISSA 69/98/FM.

[BV] Blok, B. and Varchenko, A.: Topological conformal field theories and the flat coordinates, *Int. J. Mod. Phys.* **A7** (1992) 1467.

[BP] Bondal, A.I. and Polishchuk, A.E.: Homological properties of associative algebras: the method of helices, *Russ. Acad. Sci. Izv.* **42** (1994) 219 - 260.

[CV] Cecotti, S. and Vafa, C.: On classification of $N = 2$ supersymmetric theories, *Comm. Math. Phys.* **158** (1993), 569-644.

[DVV] Dijkgraaf, R., Verlinde, E. and Verlinde, H.: Topological strings in $d < 1$, *Nucl. Phys.* **B 352** (1991) 59.

[DW] Dijkgraaf, R., and Witten, E.: Mean field theory, topological field theory, and multimatrix models, *Nucl. Phys.* **B 342** (1990) 486-522.

[Du1] Dubrovin, B.: Hamiltonian formalism of Whitham-type hierarchies and topological Landau - Ginsburg models, *Comm. Math. Phys.* **145** (1992) 195 - 207.

[Du2] —: Integrable systems in topological field theory, *Nucl. Phys.* **B 379** (1992) 627 - 689.

[Du3] —: Geometry of 2D topological field theories, In: "Integrable Systems and Quantum Groups", Eds. M.Francaviglia, S.Greco, Springer Lecture Notes in Math. **1620** (1996) 120 - 348.

[Du4] —: Flat pencils of metrics and Frobenius manifolds, math.DG/9803106, to appear in Proceedings of 1997 Taniguchi Symposium "Integrable Systems and Algebraic Geometry".

[Du5] —: Painlevé transcendents in two-dimensional topological field theory, math.AG/9803107.

[DN] —, Novikov, S.P.: The Hamiltonian formalism of one-dimensional systems of hydrodynamic type and the Bogoliubov - Whitham averaging method, *Sov. Math. Dokl.* **27** (1983) 665 - 669.

[DZ1] —, Zhang, Y.: Extended affine Weyl groups and Frobenius manifolds, *Compositio Math.* **111** (1998) 167-219.

[DZ2] —, —: Bihamiltonian hierarchies in 2D topological field theory at one-loop approximation, Preprint SISSA 152/97/FM, hep-th/9712232, to appear in *Comm. Math. Phys.*.

[DZ3] —, —: Frobenius manifolds and Virasoro constraints, to appear

[EHX] Eguchi, T., Hori, K., Xiong, C.-S.: Gravitational quantum cohomology, *Int.J.Mod.Phys.* **A12** (1997) 1743-1782.

[EX] Eguchi, T., Xiong, C.-S.: Quantum Cohomology at Higher Genus: Topological Recursion Relations and Virasoro Conditions, hep-th/9801010.

[FL] Fateev, V., Lukyanov, S.: Additional symmetries and exactly solvable models in two-dimensional conformal field theories, Parts I, II and III, *Sov. Sci. Rev.* **A15** (1990) 1.

[Ge] Getzler, E.: Intersection theory on $\bar{M}_{1,4}$ and elliptic Gromov-Witten invariants, alg-geom/9612004.

[Gi1] Givental, A.B.: Stationary phase integrals, quantum Toda lattice, flag manifolds, and the mirror conjecture, alg-geom/9612001.

[Gi2] —: Elliptic Gromov-Witten invariants and generalized mirror conjecture, math.AG/9803053.

[Gr] Gromov, M.: Pseudo-holomorphic curves in symplectic manifolds, *Invent. Math.* **82** (1985), 307.

[Guz] Guzzetti, D.: Stokes matrices and monodromy groups of the quantum cohomology of projective space, to appear.

[Ho] Hori, K.: Constraints for topological strings in $D \geq 1$, *Nucl. Phys.* **B 439** (1995) 395 - 420.

[JM] Jimbo, M. and Miwa, T.: Monodromy preserving deformations of linear ordinary differential equations with rational coefficients. II. *Physica* **2D** (1981) 407 - 448.

[Kon] Kontsevich. M.: Talk at Scuola Normale Superiore, Pisa, April '98.

[KM] —, Manin, Yu.I.: Gromov - Witten classes, quantum cohomology and enumerative geometry, *Comm.Math. Phys.* **164** (1994) 525 - 562.

[MS]. McDuff, D. and Salamon, D.: J-holomorphic curves and quantum cohomology, Providence, RI., American Mathematical Society, 1994.

[Mag] Magri, F.: A simple model of the integrable Hamiltonian systems, *J. Math. Phys.* **19** (1978) 1156 - 1162.

[Mal] Malgrange, B.: Équations Différentielles à Coefficients Polynomiaux, Birkhäuser, 1991.

[Man1] Manin, Yu.I.: Frobenius manifolds, quantum cohomology, and moduli spaces, Preprint MPI 96-113.

[Man2] —: Three constructions of Frobenius manifolds: a comparative study, math.AG/9801006.

[MM] —, Merkulov, S.A.: Semisimple Frobenius (super)manifolds and quantum cohomology of P^r, alg-geom/9702014.

[Mi] Miwa, T.: Painlevé property of monodromy preserving equations and the analyticity of τ-functions, *Publ. RIMS* **17** (1981), 703-721.

[Rud] Rudakov, A.: Integer valued bilinear forms and vector bundles, *Math. USSR Sbornik* **66** (1989), 187 - 194.

[RT] Ruan, Y. and Tian, G.: A mathematical theory of quantum cohomology, *Math. Res. Lett.* **1** (1994), 269-278.

[Sai1] Saito, K.: On a linear structure of a quotient variety by a finite reflection group, Preprint RIMS-288 (1979), *Publ. RIMS, Kyoto Univ.*, **29** (1993) 535–579.

[Sai2] —: Period mapping associated to a primitive form, *Publ. RIMS* **19** (1983) 1231 - 1264.

[SYS] —, Yano, T. and Sekeguchi, J.: On a certain generator system of the ring of invariants of a finite reflection group, *Comm. in Algebra* **8(4)** (1980) 373 - 408.

[TX] Tian, G. and Xu, G.: On the semisimplicity of the quantum cohomology algebra of complete intersections, alg-geom/9611035.

[W1] Witten, E.: On the structure of the topological phase of two-dimensional gravity, *Nucl. Phys.* **B 340** (1990) 281-332.

[W2] —: Two-dimensional gravity and intersection theory on moduli space, *Surv. Diff. Geom.* **1** (1991) 243-210.

[Zas] Zaslow, E.: Solitons and Helices: The Search for a Math-Physics Bridge, *Comm. Math. Phys.* **175** (1996) 337-376.

B. Dubrovin, SISSA, Via Beirut, 2-4, I-34013 Trieste, Italy

Invariants in Contact Topology

Yakov Eliashberg[1]

Abstract. Contact topology studies contact manifolds and their Legendrian submanifolds up to contact diffeomorphisms. It was born, together with its sister Symplectic topology, less than 20 years ago, essentially in seminal works of D. Bennequin and M. Gromov (see [2, 18]). However, despite several remarkable successes the development of Contact topology is still significantly behind its symplectic counterpart. In this talk we will discuss the state of the art and some recent breakthroughs in this area.

1991 Mathematics Subject Classification: 53C, 55N, 57R, 58G, 81T
Keywords and Phrases: Contact manifolds, Legendrian submanifolds, holomorphic curves, contact homology algebra

1 Contact preliminaries

A 1-form α on a $(2n-1)$-dimensional manifold V is called *contact* if the restriction of $d\alpha$ to the $(2n-2)$-dimensional tangent distribution $\xi = \{\alpha = 0\}$ is non-degenerate (and hence symplectic). A codimension 1 tangent distribution ξ on V is called a *contact structure* if it can be locally (and in the co-orientable case globally) defined by the Pfaffian equation $\alpha = 0$ for some choice of a contact form α. The pair (V, ξ) is called a *contact manifold*. According to Frobenius' theorem the contact condition is a condition of maximal non-integrability of the tangent hyperplane field ξ. In particular, all integral submanifolds of ξ have dimension $\leq n-1$. On the other hand, $(n-1)$-dimensional integral submanifolds, called *Legendrian*, always exist in abundance. Any non-coorientable contact structure can be canonically double-covered by a coorientable one. If a contact form α is fixed then one can associate with it the *Reeb vector field* R_α, which is transversal to the contact structure $\xi = \{\alpha = 0\}$. The field R_α is uniquely determined by the equations $R_\alpha \lrcorner d\alpha = 0$; $\alpha(R_\alpha) = 1$.

The $2n$-dimensional manifold $M = (T(V)/\xi)^* \setminus V$, called the *symplectization* of (V, ξ), carries the natural symplectic structure ω induced by the embedding $M \to T^*(V)$ which assigns to each linear form $T(V)/\xi \to \mathbb{R}$ the corresponding form $T(V) \to T(V)/\xi \to \mathbb{R}$. A choice of a contact form α (if ξ is co-orientable) defines a splitting $M = V \times (\mathbb{R} \setminus 0)$. We will usually pick the positive half $V \times \mathbb{R}_+$ of M, and call it symplectization as well. The symplectic structure ω can be written in terms of this splitting as $d(\tau\alpha), \tau > 0$. It will be more convenient for us,

[1]Supported by the National Science Foundation

however, to use additive notations and write ω as $d(e^t\alpha)$, $t \in \mathbb{R}$, on $M = V \times \mathbb{R}$. Notice that the vector field $T = \frac{\partial}{\partial t}$ is conformally symplectic: we have $\mathcal{L}_T\omega = \omega$, as well as $\mathcal{L}_T(e^t\alpha) = e^t\alpha$. All the notions of contact geometry can be formulated as the corresponding symplectic notions, invariant or equivariant with respect to this conformal action. For instance, any contact diffeomorphism of V lifts to an equivariant symplectomorphism of M; contact vector fields on V (i.e. vector fields preserving the contact structure) are projections of $\mathbb{R}$-invariant contact symplectic (and automatically Hamiltonian) vector fields on M; Legendrian submanifolds in M correspond to cylindrical (i.e. invariant with respect to the $\mathbb{R}$-action) Lagrangian submanifolds of M.

The symplectization of a contact manifold is an example of a symplectic manifold with *cylindrical* (or rather conical) ends, which is a possibly non-compact symplectic manifold (W, ω) with ends of the form $E_+ = V_+ \times [0, \infty)$ and $E_- = V_- \times (-\infty, 0]$, such that $V_\pm$ are compact manifolds, and $\omega|_{V_\pm} = d(e^t\alpha_\pm)$, where $\alpha_\pm$ are a contact forms on $V_\pm$. In other words, the ends $E_\pm$ of (W, ω) are symplectomorphic, respectively, to the positive, or negative halves of the symplectizations of the contact manifolds $(V_\pm, \xi_\pm = \{\alpha_\pm = 0\})$. We will consider the splitting of the ends and the the contact forms $\alpha_\pm$ to be parts of the structure of a symplectic manifold with cylindrical ends. We will also call (W, ω) a *directed symplectic cobordism* between the contact manifolds (V_+, ξ_+) and (V_-, ξ_-), and denote it, sometimes, by $\overrightarrow{V_+V_-}$. Let us point out that this is not an equivalence relation, but rather a partial order. Existence of a directed symplectic cobordism $\overrightarrow{M_+M_-}$ does not imply the existence of a directed symplectic cobordism $\overrightarrow{M_-M_+}$, but directed symplectic cobordisms $\overrightarrow{M_0M_1}$ and $\overrightarrow{M_1M_2}$ can be glued, in an obvious way, into a directed symplectic cobordism $\overrightarrow{M_0M_2}$. Suppose now that the symplectic form ω is exact and equal $d\beta$, where $\beta_{|E_\pm} = e^t\alpha_\pm$, and that there exists a Morse function $\varphi : W \to \mathbb{R}$ which coincides with the function t at infinity and such that for any $c \in \mathbb{R}$ the restriction $\beta|_{\{\varphi=c\}}$ is a contact form away from the critical points of the function φ. In this case we say that (W, ω) is a *directed Stein cobordism* between the contact manifolds (V_+, ξ_+) and (V_-, ξ_-). Notice that indices of critical points of the function φ are bounded in this case by $n = \frac{1}{2}\dim W$. If there exists a directed symplectic (resp. Stein) cobordism between a contact manifold (V_+, ξ_+) and $V_- = \emptyset$, then (V_+, ξ_+) is called *symplectically* (resp. *Stein*) *fillable*.[2] The Stein filling W is called *subcritical* if the function φ can be chosen without critical points of the maximal index n.

Contact structures have no local invariants. Moreover, any contact form is locally isomorphic to the form $\alpha_0 = dz - \sum_1^{n-1} y_i dx_i$ (Darboux' normal form). The contact structure ξ_0 on $\mathbb{R}^{2n-1}$ given by the form α_0 is called *standard*. Standard contact structure on S^{2n-1} is formed by complex tangent hyperplanes to the unit sphere in $\mathbb{C}^n$. The standard contact structure on S^{2n-1} is isomorphic in the complement of a point to the standard contact structure on $\mathbb{R}^{2n-1}$. According to

[2]The Stein fillabillity of (V, ξ) is equivalent to the existence of a compact complex manifold with a strictly pseudoconvex boundary V and a Stein interior, such that ξ is the field of complex tangencies to the boundary V. See [9] for the discussion of different notions of symplectic fillabillity.

a theorem of J. Gray (see [17]) contact structures on closed manifolds have the following stability property: *Given a family ξ_t, $t \in [0,1]$, of contact structures on a closed manifold M, there exists an isotopy $f_t : M \to M$, such that $df_t(\xi_0) = \xi_t; t \in [0,1]$.* Notice that for contact *forms* the analogous statement is wrong. For instance, the topology of the 1-dimensional foliation determined by the Reeb vector field R_α is very sensitive to deformations of the contact form α.

The conformal class of the symplectic form $d\alpha|_\xi$ depends only on the cooriented contact structure ξ and not a choice of the contact form α. In particular, one can associate with ξ an almost complex structure $J : \xi \to \xi$, compatible with $d\alpha$ which means that $d\alpha(X, JY); X, Y \in \xi$, is an Hermitian metric on ξ. The space of almost complex structures J with this property is contractible, and hence the choice of J is homotopically canonical. Thus a cooriented contact structure ξ defines on M a *stable almost complex structure* $\widetilde{J} = \widetilde{J}_\xi$, i.e. a splitting of the tangent bundle $T(V)$ into the Whitney sum of a complex bundle of (complex) dimension $(n-1)$ and a trivial 1-dimensional real bundle. The existence of a stable almost complex structure is necessary for the existence of a contact structure on V. If V is open (see [19]) or $\dim V = 3$ (see [25, 24]) this property is also sufficient for the existence of a contact structure in the prescribed homotopy class. It is still unknown whether this condition is sufficient for the existence of a contact structure on a closed manifold of dimension > 3. However, the positive answer to this question is extremely unlikely. Similarly, the homotopy class of $\widetilde{J}_\xi$, which we denote by $[\xi]$ and call the *formal* homotopy class of ξ, serves as an invariant of ξ. For an open V it is a complete invariant (see [19]) up to a deformation of contact structures, but not up to a contact diffeomorphism. For closed manifolds this is known to be false in all dimensions, see the discussion below. The main goal of this talk is the construction of invariants which would allow to distinguish contact manifolds in the same formal homotopy class.

2 Invariants of open manifolds

We concentrate in this section on 3-dimensional contact manifolds, although some part of the discussion can be generalized to higher dimensions. First of all 3-dimensional contact manifolds are orientable, and any contact structure determines an orientation of M. If M is a priori oriented then contact structures can be divided into positive and negative. We will consider here only positive contact structures.

It is proven to be useful to divide all 3-dimensional contact manifolds into two complementary classes: tight and overtwisted. A contact 3-manifold (M, ξ) is called *overtwisted* if there exists an embedded disc $D^2 \subset M$ such that its boundary ∂D^2 is tangent to ξ (i.e. ∂D^2 is a Legendrian curve), while the disc itself is transverse to ξ along its boundary. A non-overtwisted contact structures are called *tight*. D. Bennequin (see [2]) was the first who discovered the phenomenon of overtwisting. He proved that the standard contact structure ξ_0 on S^3 is tight and constructed an overtwisted contact structure ξ_1 in the same formal homotopy class.

As it turned out, overtwisted contact structures on all closed 3-manifolds are classified up to isotopy by their formal homotopy classes (see [7]). On open

manifolds one should subdivide furthermore overtwisted contact structures into *overtwisted at infinity* and *tight at infinity*. A contact structure, which is overtwisted at infinity, is determined up to isotopy by its formal homotopy class (see [6]). Overtwisted, but tight at infinity contact structure on $M \setminus F$, where M is a closed 3-manifold and F is its finite subset, can always be canonically extended to M see [6]), and thus an isotopical classification of such structures coincides with the formal homotopical classification on M (and not on $M \setminus F$!).

Let us now restrict ourself to the class of tight contact structures. It is not so easy to provide non-trivial invariants of tight contact structures on open manifolds. The problem is that all standard symplectic invariants (the Gromov width, capacities, etc.) take infinite values for symplectizations of contact manifolds. One knows, for instance, that on $\mathbb{R}^3$ any tight contact structure is isotopic to the standard one (see [6]). However, on the closed half-space $\mathbb{R}^3_+ = \{y \geq 0\} \subset \mathbb{R}^3$ there are non-isomorphic contact structures, which we are describing below. Let us denote by Ξ_0 the space of tight contact structures on $\mathbb{R}^3_+$ which coincide with the standard contact structure ξ_0 near the plane $\Pi = \{y = 0\} = \partial\mathbb{R}^3_+$. We are interested in invariants of contact half-spaces $(\mathbb{R}^3_+, \xi)$, where $\xi \in \Xi_0$, up to diffeomorphisms fixed near Π.

Given a contact structure $\xi \in \Xi_0$ let us consider an embedded plane $\widetilde{\Pi} \subset \mathbb{R}^3_+$ which coincides with Π at infinity, and which is transversal to ξ_0. The 1-dimensional line field $\xi \cap T(\widetilde{\Pi})$ integrates into a 1-dimensional *characteristic foliation* $\widetilde{\Pi}_\xi$ on $\widetilde{\Pi}$. The foliation $\widetilde{\Pi}_\xi$ coincides with the foliation by lines $\{z = \text{const}\}$ at infinity, and thus the holonomy along its leaves defines a compactly supported diffeomorphism $h_{\widetilde{\Pi}} : \mathbb{R} \to \mathbb{R}$, where we identify the source $\mathbb{R}$ with the line $\{x = -N\} \subset \Pi$, and the target $\mathbb{R}$ with the line $\{x = N\} \subset \Pi$ for a sufficiently large $N > 0$. Let us define $c_\xi(z) = \sup_{\widetilde{\Pi}} \left(h_{\widetilde{\Pi}}(z) - z\right)$, and call the function $c_\xi(z)$ the *contact shape* of $(\mathbb{R}^3, \xi)$. Of course, sometimes we have $c(\xi) \equiv +\infty$. For instance this is the case for the standard contact structure $\xi = \xi_0$. On the other hand, the following construction (see [12]) shows that any positive continuous [3] function $f : \mathbb{R} \to \mathbb{R}$, such that $f(z) + z$ is a monotone function, can be realized as the invariant c_ξ for some contact structure $\xi \in \Xi_0$.

For a positive Lipschitz function φ on $\mathbb{R}^2$ we denote by S_φ its graph $\{y = \varphi(x, z)\} \subset \mathbb{R}^3$. If the function φ decays sufficiently fast when $|x| \to \infty$ (say, $\varphi(x, z) < \frac{C}{x^2}$), then the holonomy diffeomorphism $h_\varphi : \mathbb{R} \to \mathbb{R}$ along the leaves of the characteristic foliation of the graph $\Pi_\varphi = \{y = \varphi(x, z)\}$ is well defined. It is easy to find a Lipshitz function φ with the prescribed continuous holonomy $h_\varphi(z) = z + f(z)$. Consider the domain $\Omega_\varphi = \{0 \leq y < \varphi(x, z)\}$. Clearly, the contact manifold (Ω_φ, ξ_0) belongs to the class Ξ_0. We have (see [12])

PROPOSITION 2.1

$$c_{(\Omega_\varphi, \xi_0)}(z) = h_\varphi(z) - z = f(z)\,.$$

A similar invariant can be defined for open manifolds (without boundary) when $H_1(M) \neq 0$. (see [5])

[3] In fact, it need not to be even continuous.

3 Invariants of closed manifolds

Until very recently there was known only one result allowing to distinguish contact structures on manifolds of dimension > 3 within a given formal homotopy class. Namely, we have

Theorem 3.1 *For any $n > 2$ the sphere S^{2n-1} has a contact structure ξ_1 in the standard formal homotopy class, which is not isomorphic to the standard contact structure ξ_0.*

For odd values of n this was shown by the author in [4], and later extended to even values of n by H. Geiges (see [15]). [4]

We will discuss in this section some new powerful algebraic invariants of closed contact manifolds, related to Gromov-Witten invariants of symplectic manifolds, which were recently developed jointly by H. Hofer, A. Givental and the author. See also Hofer's talk at the current proceedings for the discussion of other related aspects of this theory.

Contact homology algebra. To define the invariants of a contact manifold (V, ξ) let us fix a contact form α and an almost complex structure $J : \xi \to \xi$ compatible with the symplectic form $d\alpha$. The symplectization M of (V, ξ) can be identified, as was explained in Section 1, with $(V \times \mathbb{R}, d(e^t\alpha))$. The complex structure J extends from ξ to $T(M)$ by setting $J\frac{\partial}{\partial t} = R_\alpha$, where R_α is the Reeb vector field of the contact form α. For a generic choice of α there are only countably many periodic trajectories (including multiple ones) of the vector field R_α. Moreover, these trajectories can be assumed *non-degenerate* which means that the linearized Poincaré return map along any of these trajectories has no eigenvalues equal to 1.

Let $\mathcal{P} = \mathcal{P}_\alpha$ be the set of all periodic trajectories of R_α. We do not fix initial points on periodic trajectories, and include all multiples as separate points of $\mathcal{P}$. Let us first assume that $H_1(V) = 0$. For each $\gamma \in \mathcal{P}$ let us choose and fix a surface F_γ spanning the trajectory γ in V. This enable us to define the *Conley-Zehnder index* $\mu(g)$ of γ as follows. Choose a homotopically unique trivialization of the symplectic vector bundle $(\xi, d\alpha)$ over each trajectory $\gamma \in \mathcal{P}$ which extends to $\xi|_{F_\gamma}$. The linearized flow of R_α along γ defines then a path in the group $Sp(2n-2, \mathbb{R})$ of symplectic matrices, which begins at the unit matrix and ends at a matrix with all eingenvalues different from 1. The Maslov index of this path (see [1, 26]) is, by the definition, the Conley-Zehnder index $\mu(\gamma)$ of the trajectory γ. For our purposes it will be convenient to use the *reduced* Conley-Zehnder index $\overline{\gamma} = \mu(\gamma) + n - 3$, also called the degree of γ. Notice that by changing the spanning surfaces for the trajectories from $\mathcal{P}$ one can change Conley-Zehnder indices by the values of the cohomology class $2c_1(\xi)$, where $c_1(\xi)$ is the first Chern class of the contact bundle

[4] Although this result sounds similar to Bennequin's theorem asserting that the standard contact structure on S^3 is not overtwisted, non-standard contact structures on high-dimensional spheres provided by Theorem 3.1 are quite different: they are *symplectically*, and even *Stein fillable*, while an overtwisted contact structure on S^3 is not.

ξ. In particular, mod 2 indices can be defined independently of any spanning surfaces, and even in the case when $H_1(V) \neq 0$.

Next, we consider certain moduli spaces of holomorphic curves in the manifold $M = V \times \mathbb{R}$, which are essential for all our algebraic constructions. Let us observe that for each periodic orbit $\gamma \in \mathcal{P}$ the cylinder $\gamma \times \mathbb{R}$ is a J-holomorphic curve. Let D_r be the disc of radius r in $\mathbb{C}$ centered at the origin. Given any orbit $\gamma \in \mathcal{P}$ we say that a J-holomorphic map $f : D_r \setminus 0 \to M = V \times \mathbb{R}$ converges near 0 to the periodic trajectory γ at $\pm\infty$ if $f(z) = (g(z), h(z))$, $h(z) \underset{|z|\to 0}{\to} = \pm\infty$, and there exits the limit $\bar{g}(\varphi) = \lim_{\rho\to 0} g(\rho e^{i\varphi})$ which parametrizes the periodic trajectory γ. Notice that the orientation which is defined this way on γ coincides with the orientation given by the Reeb vector field $\mathbb{R}_\alpha$ at $+\infty$, and opposite to this orientation at $-\infty$.

Let us denote by S_{sr}, $s, r = 0, 1, \dots$, the 2-sphere S^2 with $s+r$ fixed punctures $y_1, \dots, y_s, x_1, \dots, x_r$. Given $s + r$ periodic orbits $\gamma_1, \dots, \gamma_s, \delta_1, \dots, \delta_r \in \mathcal{P}$ we consider the space $\widetilde{\mathcal{M}}^A(\gamma_1, \dots, \gamma_s; \delta_1, \dots, \delta_r)$, which consists of pairs (f, j), where j is a conformal structure on S_{sr}, and $f : S_{sr} \to M$ is a (j, J)-holomorphic curve, such that near each puncture y_k, $k = 1, \dots, s$, the map f converges to γ_k at $+\infty$, and near each puncture x_l, $l = 1, \dots, r$, it converges to δ_l at $-\infty$. As usual we pass to the corresponding moduli space $\mathcal{M}(\gamma_1, \dots, \gamma_s; \delta_1, \dots, \delta_r)$ by identifying pairs (f, j) and $(\tilde{f}, \tilde{j})$ which differ by a diffeomorphism of the sphere S^2 which fixes the punctures $y_1, \dots, y_s, x_1, \dots, x_r$. The space $\mathcal{M}$ can be written as a disjoint union $\mathcal{M} = \bigcup_{A\in H_2(V)} \mathcal{M}^A$, where $\mathcal{M}^A$ consists of holomorphic curves which together with the surfaces spanning in V the trajectories $\gamma_1, \dots, \gamma_s, \delta_1, \dots, \delta_r \in \mathcal{P}$ represent the homology class $A \in H_2(M) = H_2(V)$. Then we have

PROPOSITION 3.2 *For a generic choice of J, and any periodic orbits $\gamma_1, \dots, \gamma_s, \delta_1, \dots \delta_r \in \mathcal{P}$ the moduli space $\mathcal{M}^A(\gamma_1, \dots, \gamma_s; \delta_1, \dots \delta_r)$ is an orbifold of dimension* [5]

$$\sum_{k=1}^{s} \overline{\gamma_k} - \sum_{l=1}^{r} \overline{\delta}_l + (2 - 2s)(n - 3) + 2c_1(\xi)[A].$$

REMARK 3.3 The additive group $\mathbb{R}$ acts on $M = V \times \mathbb{R}$ by J-biholomorphic translations $(x, t) \mapsto (x, t+c)$. The moduli spaces $\mathcal{M}(\gamma_1, \dots, \gamma_s; \delta_1, \dots, \delta_r)$ are invariant under this action, and hence, with the exception of trivial spaces $\mathcal{M}(\gamma; \gamma)$ (which consist of cylinders $\gamma \times \mathbb{R}$), a non-empty moduli space $\mathcal{M}(\gamma_1, \dots, \gamma_s; \delta_1, \dots, \delta_r)$ always has a positive dimension.

Let us consider now a free (super-)commutative graded algebra $\Theta = \Theta_\alpha$ over $\mathbb{C}$ with the unit element generated by elements of $\mathcal{P}_\alpha$. In other words, Θ is a polynomial algebra with complex coefficients of generators of even degree and an exterior algebra of odd degree generators. Let us recall that we count all

[5] It is a standard difficulty in the Floer homology theory and the theory of holomorphic curve invariants in general, that in the presence of multiply-covered curve it is, sometimes, impossible to achieve transversality needed for this dimension formula just by perturbing the almost complex structure J. The appropriate virtual cycles technique which works in this case and involves multivalued perturbations was recently developed by several authors, see [14], [23] et al.

multiples of a given trajectories as independent generators of Θ. Each monomial element $\theta \in \Theta$ is graded by its total degree $\bar{\theta}$. Let Θ^{H_2} be the group algebra of $H_2(V) = H_2(M)$ with coefficients in Θ. Thus elements of Θ^{H_2} can be written as polynomials $\sum\limits_{A \in H_2(V)} \theta_A t^A$, $\theta_A \in \Theta$.

We define now a sequence of operations $\underbrace{\Theta^{H^2} \otimes \dots \Theta^{H_2}}_{s} \to \Theta^{H_2}$, which make Θ into a L_∞-algebra (see, for instance, [22]), or rather a P_∞-algebra, where P stands for the Poisson structure. This is done by an appropriate counting of components of the moduli spaces $\mathcal{M}^A(\gamma_1, \dots, \gamma_s; \delta_1, \dots \delta_r)$.

Take any $s \geq 1$ periodic orbits $\gamma_1, \dots, \gamma_s \in \mathcal{P}$ and set

$$[\gamma_1, \dots, \gamma_s]_s = \sum_{A \in H_2(V)} \sum_{\Delta} a_\Delta^A t^A \Delta,$$

where $a_\Delta^A \in \mathbb{C}$, and the second sum is taken over all monomials $\Delta = \delta_1^{j_1} \cdots \delta_r^{j_r}$ of (distinct) generators $\delta_1, \dots, \delta_r, \dots \in \Theta$ (notice that we allow the case $r = 0$). We set $a_\Delta^A = 0$ if the dimension $\sum\limits_{k=1}^{s} \overline{\gamma_k} - \sum\limits_{l=1}^{r} j_l \overline{\delta_l} + (2-2s)(n-3) + 2c_1(\xi)[A]$ of the moduli space $\mathcal{M}^A = \mathcal{M}^A(\gamma_1, \dots, \gamma_s; \underbrace{\delta_1 \dots, d_1}_{j_1}, \dots, \underbrace{d_r, \dots, \delta_r}_{j_r})$ is different from 1. Otherwise, we define the coefficient a_Δ^A as the sum $\sum\limits_{C} w(C)$ of weights $w(C)$ assigned to 1-dimensional components of the moduli space $\mathcal{M}^A$ Given a component C of $\mathcal{M}$ we set

$$w(C) = \pm \frac{1}{r! d} m(\delta_1)^{j_1} \cdots m(\delta_r)^{j_r},$$

where $m(\delta_l)$ is the multiplicity of the periodic orbit δ_l, $l = 1, \dots, r$; $d = 1$ if the curves from C are not multiply-covered and d is the order of the group of deck transformations of the corresponding branched covering in the multiply-covered case. Finally the sign $\pm$ is determined by an algorithm, similar to the one used in the traditional Floer theory (see [13]). This algorithm shows, in particular, that the operations $[\cdot, \dots, \cdot]_s$ are skew-symmetric: a transposition of any two elements γ_1, γ_2' in the bracket changes the sign by $(-1)^{\overline{\gamma_1}\overline{\gamma_2}}$. It is important to point out that compactness theorems for holomorphic curves (see [18, 20, 11]) garantee that the operations $[\,\cdot, \dots, \cdot\,]_s$ take values in *polynomial* functions (and not in formal power series).

The operation $[\,\cdot, \dots, \cdot\,]_s$ which was just defined on the generators of Θ admits a unique extension to a skewsymmetric multilinear operation $\underbrace{\Theta^{H_2} \otimes \dots, \otimes \Theta^{H_2}}_{s} \to \Theta^{H_2}$ which satisfies the Leibnitz rule:

$$(-1)^t [\theta_1, \dots, \theta_l \theta_l', \dots, \theta_s]_s = \theta_l [\theta_1, \dots, \theta_l', \dots, \theta_s]_s + (-1)^{\overline{\theta_l}} \theta_l' [\theta_1, \dots, \theta_l, \dots, \theta_s]_s,$$

where $t = \sum\limits_{1}^{l-1} \overline{\theta_i}$. Let us now take a closer look to the operation $[\theta]_1$ which will also be denoted by $d\theta$, and called the *differential* of θ. Notice that it decreases the grading by 1, i.e. $\overline{d\theta} = \overline{\theta} - 1$ for any monomial $\theta \in \Theta$.

THEOREM 3.4 *1.* $d^2 = 0$.

2. Any homotopy $\alpha_t, t \in [0,1]$, *of contact forms, together with a compatible homotopy* J_t *of almost complex structures, induces a quasi-isomorphism* $\Phi_{\{a_t,J_t\}}$ *of the corresponding algebras. In particular, the graded contact homology algebra* $H\Theta^{H_2} = \operatorname{Ker} d/\operatorname{Coker} d$ *is an invariant of the contact manifold* (M,ξ).

Sometimes it is more convenient to consider the *reduced* contact homology algebra $\widetilde{H\Theta}^{H_2}(M,\xi)$ of a closed contact manifold (M,ξ), which is defined similarly to $H\Theta^{H_2}$, except that instead of contact forms for ξ on the whole M, we use contact forms on the punctured manifold $M \setminus x$, $x \in M$, which are isomorphic at infinity to the standard contact form $dz - \sum_{i=1}^{n-1} y_i dx_i$ on $\mathbb{R}^{2n-1}$.

The contact homology algebras $H\Theta^{H_2}$ and/or $\widetilde{H\Theta}^{H_2}$ can be explicitely computed in several interesting examples. Let us formulate here some of these results.

THEOREM 3.5 *1.* $\widetilde{H\Theta}(S^{2n-1},\xi_0) = \mathbb{C}$; $H\Theta(S^{2n-1},\xi_0)$ *is a graded polynomial algebra of generators* $\gamma_1,\gamma_2,\dots$, *of degrees* $\overline{\gamma_i} = 2(n+i-1)$, $i = 0,\dots$.

2. If ξ *is an overtwisted contact structure on a* 3*-manifold* V *then* $H\Theta(V,\xi) = 0$.

3. For any Stein fillable contact manifold (V,ξ) *we have* $H\Theta(V,\xi) \neq 0$.

4. Suppose that a contact manifold (V,ξ) *of dimension* $2n-1$ *with* $H_1(V) = 0$ *has a subcritical Stein filling* W. *Then* $\widetilde{H\Theta}^{H_2}(V,\xi)$ *is a group algebra of* $H^2(V)$ *over a free graded commutative algebra with generators* $\gamma_{ikl}, k = 1,\dots,n-1$, $l = 1,\dots,\dim H_k(W;\mathbb{R}), i = 0,\dots$, *of degree* $\overline{\gamma_{ikl}} = 2(n+i-1)-k$.

5. Let ξ_1 *be a non-standard contact structure on* S^{2n-1}, $n > 2$, *which is provided by Theorem 3.1 above, and the contact manifold* (S^{2n-1},ξ_k) *be the connected sum of* k *copies of* (S^{2n-1},ξ_1). *Then the contact homology algebras* $H\Theta(S^{2n-1},\xi_k)$ *are pairwise non-isomorphic for all* k, *and in particular* S^{2n-1} *has infinitely many distinct contact structures in the standard homotopy class.*

We thank Yu. Chekanov who pointed out to us the property 3.5.3. The computations in 3.5.4 were done by M.-L. Yau, and the result in 3.5.5 is due to I. Ustilovsky.

THE CASE $H_1(V) \neq 0$. For a general contact manifold V with $H_1(V) \neq 0$ one may first construct a similar contact homology algebra $H\Theta^{H_2}_{\text{contr}}$ generated by the subset $\mathcal{P}_a^{\text{contr}} \subset \mathcal{P}_a$ of *contractible* periodic orbits, and then for each free loop homotopy class Γ consider a module Θ_Γ over the algebra $\Theta^{H_2}_{\text{contr}}$, generated by elements of $\mathcal{P}$ from the homotopy class Γ. The differential $d : \Theta_\Gamma \to \Theta_\Gamma$ on this module is defined, as above, by counting components of 1-dimensional moduli spaces $\mathcal{M}(\gamma;\delta_1,\dots,\delta_r)$ with an extra condition that γ and δ_1 belong to the class Γ, while all the other trajectories $\delta_2,\dots,\delta_s$ are from $\mathcal{P}^{\text{contr}}$. Then we also have $d^2 = 0$,

and thus the homology $H\Theta_\Gamma^{H_2}$, which is a module over the contact homology algebra $H\Theta_{\text{contr}}^{H_2}$ is another invariant of the contact manifold V. For a generic α $H\Theta_\Gamma^{H_2}$ is a finite dimensional module over $H\Theta_{\text{contr}}^{H_2}$, and thus can be effectively computed, especially when the contact homology algebra $H\Theta_{\text{contr}}^{H_2}$ is isomorphic to $\mathbb{C}$ (or to the group algebra of $\mathbb{C}[H_2(V)]$). For instance, let ξ_1 be the standard contact structure on the 3-torus $V = T^3 = T^2 \times S^1$ viewed as the unit cotangent bundle of T^2. For $k = 2, \dots,$ we denote by ξ_k the pull-back of ξ_1 under the k-sheeted covering $T^3 \to T^3$ which unwinds the fiber S^1. Then

Theorem 3.6 *$H\Theta_{\text{contr}}^{H_2}(\xi_k) = \mathbb{C}[H_2(T^3)]$; $\dim_{\mathbb{C}[H_2(T^3)]}(H\Theta_\Gamma^{H_2}(\xi_k)) = 2k$ for any horizontal 1-dimensional homology class, i.e. a class from $H_1(T^2 \times \text{point}) \subset T^3$, and $H\Theta_\Gamma^{H_2}(\xi_k)) = 0$ for all other classes $\Gamma \in H_1(T^3)$.*

As a corollory of 3.6 we get a theorem of E. Giroux and Y. Kanda (see [16, 21]) which states that the contact structures ξ_k, $k = 1, \dots,$ are pairwise non-isomorphic. It seems likely that the algebra $H\Theta_{\text{contr}}^{H_2}(V, \xi)$ is trivial (i.e. isomorphic to the group algebra of $H_2(V)$ over $\mathbb{C}$) for any *strongly tight* contact manifold (V, ξ), i.e. a contact 3-manifold which is covered by $\mathbb{R}^3$ with a tight, and hence standard contact structure.

Hamiltonian formalism. It turns out that the the operations $[\cdot, \dots, \cdot]_s$ for $s > 1$ can be viewed as certain cohomological operations on the contact homology algebra. For instance, we have

Theorem 3.7 *The operation $[\cdot, \cdot]_2$ is a Poisson bracket on $\Pi(H\Theta)$, where Π is the operator of changing the parity. The quasi-isomorphism $\Phi_{\{a_t, J_t\}}$ from 3.4.2, induced by a deformation of contact forms and almost complex structures, preserves the Poisson bracket.*

Other operations $[\cdot, \dots, \cdot]_s$, $s > 2$, define secondary cohomological operations on the contact homology algebra $H\Theta^{H_2}$, which all fit into a structure of a L_∞, or rather a P_∞-algebra on Θ^{H_2}. However, the following Hamiltonian formalism provides a better algebraic framework for all these operations.

Let us associate with each periodic trajectory $\gamma \in \mathcal{P} = \mathcal{P}_\alpha$ two variables, p_γ and q_γ of the same degree $\overline{p_\gamma} = \overline{q_\gamma}$. Let $T\Theta$ be the free graded (super-)commutative algebra over $\mathbb{C}$ with the unit generated by variables p_γ, q_γ associated to each periodic orbit $\gamma \in \mathcal{P}$, and completed with respect to variables p_γ, $\gamma \in \mathcal{P}$. This means that the elements of $T\Theta$ are formal power series in p-variables with coefficients which are polynomial of q-variables. We will also consider the group algebra $T\Theta^{H_2}$ of the group $H_2(V)$ with coefficients in $T\Theta$. Informally, if one thinks about the algebra Θ as the algebra of polynomial functions on an infinite-dimensional (super-)space L with coordinates q_γ, $\gamma \in \mathcal{P}$, then $T\Theta$ is the algebra of functions on the cotangent bundle T^*L of L. This infinite-dimensional cotangent bundle is endowed with an even symplectic form $\sum_{\gamma \in \mathcal{P}} dp_\gamma \wedge dq_\gamma$, which defines, in its turn, Poisson brackets on algebras $T\Theta$ and $T\Theta^{H_2}$.

Next we construct a Hamiltonian function $H \in T\Theta^{H_2}$ which will encode all the information about the brackets $[\,\cdot, \cdots, \cdot\,]_k$ introduced above. We set

$$H(p,q) = \sum [\gamma_1, \dots, \gamma_s]_s p_{\gamma_1} \cdots p_{\gamma_s},$$

where the sum is taken over the set of all monomials in variables p_γ, $\gamma \in \mathcal{P}$, and where we assume that the brackets $[\gamma_1, \dots, \gamma_s]_s \in \Theta^{H_2}$ are expressed in terms of the variables q_γ, $\gamma \in \mathcal{P}$. Theorems 3.8 and 3.9 below generalize Theorem 3.4 and formalize properties of all considered above operations.

THEOREM 3.8 $\{H, H\} = 0$.[6]

Consider a Poisson subalgebra $Z\Theta^{H_2} = \{f \in T\Theta^{H_2};\ \{f, H\} = 0\} \subset T\Theta^{H_2}$ and its ideal $B\Theta^{H_2}$, generated by functions of the form $\{g, H\}$, $g \in T\Theta^{H_2}$. Then $P\Theta^{H_2} = Z\Theta^{H_2}/B\Theta^{H_2}$ also carries a Poisson structure.

THEOREM 3.9 *Any homotopy α_t, $t \in [0,1]$, of contact forms, together with a compatible homotopy J_t of almost complex structures, induces an isomorphism $\Psi_{\{\alpha_t, J_t\}} : P\Theta^{H_2}(V, \alpha_0, J_0) \to P\Theta^{H_2}(V, \alpha_1, J_1)$ of Poisson algebras.*

These results are only the first steps of a bigger story. For instance, a directed symplectic cobordism between two contact manifolds generate a Lagrangian correspondence between the corresponding Poisson algebras, and the composition of directed cobordisms generates the composition of Lagrangian correspondences. We hope that this would provide tools for effective computations of rational Gromov-Witten invariants of symplectic manifolds by splitting them into compositions of elementary directed symplectic cobordisms. The larger picture also incorporates moduli spaces of holomorphic curves of higher genus, as well as higher-dimensional spaces of holomorphic curves.

INVARIANTS OF LEGENDRIAN SUBMANIFOLDS. Let us briefly mention here a relative analog of the contact homology theory, which provides invariants of pairs (V, L) where $V = (V, \xi)$ is a contact manifold, and L its Legendrian submanifold. For the case when (V, ξ) is the standard contact $(\mathbb{R}^{2n-1}, \xi_0)$ this theory produces invariants of immersed Lagrangian submanifolds in $\mathbb{R}^{2n-2}$ up to contact isotopy (see [10]), i.e up to regular Lagrangian homotopy in $\mathbb{R}^{2n-2}$, which lifts to a Legendrian isotopy in $\mathbb{R}^{2n-1}$. When $n = 2$ all the involved holomorphic curves can be explicitely seen from the combinatorics of the corresponding (Lagrangian) immersion of the curve L into $\mathbb{R}^2$, and thus the theory may be developed via pure combinatorial means. The first part of this combinatorial theory, parallel to the theory of the differential d in the absolute case, was independently done by Yu. Chekanov (see [3]). However, even for $n = 2$ a (non-commutative) analog of the described above Hamiltonian formalism allows us to define many other invariants of Legendrian curves, which can also be computed and studied by pure combinatorial means.

[6] One should remember that in the super-commutative setting the bracket of a function with itself does not vanish automatically.

References

[1] V. I. Arnold, On a characteristic class entering in quantization conditions, *Funct. Anal. and Applic.*, 1(1967), 1–14.

[2] D. Bennequin, Entrelacements et équations de Pfaff, *Astérisque*, 106–107(1983).

[3] Yu. Chekanov, Differential algebra of a Legendrian link. Preprint 1997.

[4] Y. Eliashberg, On symplectic manifolds which some contact properties, *J. of Diff. Geom.*, 33(1991), 233–238.

[5] Y. Eliashberg, New invariants of symplectic and contact manifolds, *J. of AMS*, 4(1991), 513–520.

[6] Y. Eliashberg, Classification of contact structures on $\mathbb{R}^3$, *Int. Math. Res. Notices*, 3(1993), 87–91.

[7] Y. Eliashberg, Classification of overtwisted contact structures, *Invent. Math.*, 98(1989), 623–637.

[8] Y. Eliashberg, Contact 3-manifolds twenty years after J. Martinet's work, *Annales de l'Inst. Fourier*, 42(1992), 165–192.

[9] Y. Eliashberg, Unique holomorphically fillable contact structure on the 3-torus*Internat. Math. Res. Notices* 1996, no. 2, 77–82.

[10] Y. Eliashberg and M. Gromov, Lagrangian intersections theory. Finite-dimensional approach, Arnold's volume, AMS, Providence 1998. 1996.

[11] Y. Eliashberg, H. Hofer and S. Salamon, Lagrangian intersections in contact geometry, *Geom. and Funct. Anal.*, 5(1995), 244–269.

[12] Y. M. Eliashberg and W. P. Thurston, Confoliations, University Lecture series, 13(1998), AMS, Providence.

[13] A. Floer, H. Hofer, Coherent orientations for periodic orbit problems in symplectic geometry, *Math. Z.*, 212(1993), 13–38.

[14] K. Fukaya and K. Ono, Arnold conjecture and Gromov-Witten invariants, preprint 1996.

[15] H. Geiges, Applications of contact surgery, *Topology*, 36(1997), 1193–1220.

[16] E. Giroux, Une structure de contact, même tendue est plus ou moins tordue, *Ann. Scient. Ec. Norm. Sup.*, 27(1994), 697–705.

[17] J.W. Gray, Some global properties of contact structures, *Annals of Math.*, 69(1959), 421–450.

[18] M. Gromov, Pseudo-holomorphic curves in symplectic manifolds, *Invent. Math.*, 82(1985), 307–347.

[19] M. Gromov, Partial Differential Relations, Springer-Verlag, 1986.

[20] H. Hofer, Pseudo-holomorphic curves and Weinstein conjecture in dimension three, *Invent. Math.*, 114(1993), 515–563.

[21] Y. Kanda, The classification of tight contact structures on the 3-torus, preprint 1995.

[22] M. Kontsevich, Deformation quantization of Poisson manifolds, I, preprint 1997.

[23] G. Liu and G. Tian, Floer homology and Arnold conjecture, preprint 1997.

[24] R. Lutz, Structures de contact sur les fibrés principaux en cercles de dimension 3 *Ann. Inst. Fourier*, XXVII, 3(1977), 1–15.

[25] J. Martinet, Formes de contact sur les variétés de dimension 3, *Lecture Notes in Math.,* 209(1971), 142–163.

[26] J. Robbin, D. Salamon,The Maslov index for paths, *Topology*, 32(1993), 827–844.

Yakov Eliashberg
Department of Mathematics
Stanford University
Stanford, CA 94305-2125 USA
eliash@math.stanford.edu

Curvature-Decreasing Maps are Volume-Decreasing

(on joint work with G. Besson and G. Courtois)

S. Gallot

Abstract. Giving a lower bound of the minimal volume of a manifold in terms of the simplicial volume, M. Gromov obtained a generalization of the Gauss-Bonnet-Chern-Weil formulas and conjectured that the minimal volume of a hyperbolic manifold is achieved by the hyperbolic metric. We proved this conjecture via an analogue of the Schwarz's lemma in the non complex case: if the curvature of X is negative and not greater than the one of Y, then any homotopy class of maps from Y to X contains a map which contracts volumes. We give a construction of this map which, under the assumptions of Mostow's rigidity theorems, is an isometry, providing a unified proof of these theorems. It moreover proves that the moduli space of Einstein metrics, on any compact 4-dimensional hyperbolic manifold reduces to a single point.
Assuming that X is a compact negatively curved locally symmetric manifold, and without any curvature assumption on Y, another version of the real Schwarz's lemma provides a sharp inequality between the entropies of Y and X . This answers conjectures of A. Katok and M. Gromov. It implies that Y and X have the same dynamics iff they are isometric.
This also ends the proof of the Lichnerowicz's conjecture : any negatively curved compact locally harmonic manifold is a quotient of a (noncompact) rank-one-symmetric space.

1. A real Schwarz lemma :

As was remarked by Pick, the classical Schwarz lemma may be rewritten in the language of the hyperbolic geometry (i. e. on the disk B^2 endowed with the hyperbolic metric $g_o = \frac{4}{(1-\|x\|^2)^2}\ \left((dx_1)^2 + (dx_2)^2\right)$ as follows :

1.1. Schwarz lemma.

- *Any holomorphic map $f : B^2 \to B^2$, is a contracting map from (B^2, g_o) to (B^2, g_o).*

Considering now holomorphic maps between compact Kählerian manifolds of higher dimension, there have been many generalizations of this Schwarz lemma (due in particular to L. Ahlfors, S. T. Yau, N. Mok, ...). For example, the following one, which may be found in [Mok] :

1.2. PROPOSITION.- *Let X, Y be compact Kählerian manifolds of the same dimension whose Kählerian metrics are denoted by g_X and g_Y. If $Ricci_{g_Y} \geq -C^2 \geq Ricci_{g_X}$, then any holomorphic map $F : Y \to X$ satisfies $|JacF| \leq 1$. Moreover, if $|JacF| = 1$ at some point y, then d_yF is isometric.*

Let us recall that *$Ricci_g$ is the Ricci curvature tensor of the metric* g, and that *the assumption $Ricci_g \geq -C^2$ means that $Ricci_g(u,u) \geq -C^2.g(u,u)$ at any point and for any tangent vector u at this point.*
In its homotopy class, when the target-space has negative sectional curvature, a holomorphic map is unique ([Ha]) and is a good candidate for contracting the measure. Holomorphic maps are a particular case of harmonic maps between Riemannian manifolds. As, by the negativity of the curvature, each C^0 homotopy class of maps contains exactly one harmonic map (J. Eells and J. H. Sampson, [E-S]), one may ask whether it contracts volumes. Though unsuccessful, this idea underlies the attempts for a unified proof of the Mostow's rigidity theorem, where the method of harmonic maps fits very well to the hermitian cases and moreover improves Mostow's theorem (works of Y. T. Siu, K. Corlette, J. Jost and S. T. Yau, M. Gromov and R. Schoen ..., see for instance [Mok] and [Jo]), but still gives nothing in the real hyperbolic case. Substituting another canonical map to the harmonic one, we prove that the contracting property is not particular to complex manifolds and holomorphic maps :

1.3. THEOREM ([B-C-G 3]).- *Let (Y^n, g_Y), (X^m, g_X) be complete riemannian manifolds satisfying $3 \leq dim(Y) \leq dim(X)$, let us assume that $Ricci_{g_Y} \geq -(n-1)\ C^2$ and that the sectional curvature of X satisfy $K_{g_X} \leq -C^2$ for some constant $C \neq 0$. Then any continuous map $f : Y \to X$ may be deformed to a family of C^1 canonical maps F_ϵ $(\epsilon \to 0_+)$ such that $Vol[F_\epsilon(A), g_X] \leq (1+\epsilon)\ Vol(A, g_Y)$ for any measurable set A in Y. Moreover*
(i) if Y, X are compact of the same dimension and if $Vol(Y) = |degf|\ Vol(X)$, then Y, X have constant sectional curvature and the $F'_\epsilon s$ converge, when $\epsilon \to 0$, to a riemannian covering F (an isometry when $|degf| = 1$).
(ii) If Y, X are compact, homotopically equivalent, of the same dimension, and if $K_{g_Y} < 0$, then any homotopy equivalence f may be deformed to a smooth (canonically constructed) map F such that $|JacF| \leq 1$ at every point y of Y. Moreover, if $|JacF| = 1$ at some point y, then d_yF is isometric.

1.4 REMARKS : (1) Contrary to the above result of J. Eells and J. H. Sampson on harmonic maps, the theorem 1.3 is not only an existence theorem, but moreover a direct construction of the maps F_ϵ and F.
(2) The property (ii) remains valid when $dim(Y) < dim(X)$ and when X is noncompact (however, we must assume that $\pi_1(X)$ acts on the universal covering $\tilde{X}$ in a "convex cocompact" way, i. e. that X retracts to a compact submanifold with convex boundary). *In this case, any homotopy equivalence $Y \to X$ is homotopic to some (canonical) map F such that $|JacF| \leq 1$; moreover $|JacF| \equiv 1$ iff F is an isometric and totally geodesic embedding* (cf [B-C-G 3]).

2. Applications to minimal (and maximal) volume :
Let M be a compact connected manifold ; its *minimal volume* (denoted by $MinVol(M)$) is defined by M. Gromov ([Gr 1]) as the infimum of the volumes of all the metrics g on M whose sectional curvature K_g satisfies $-1 \leq K_g \leq 1$. Similarly, when the manifold admits some metric with strictly negative sectional curvature, one may define the *maximal volume* of M as the supremum of $Vol(g)$, for all the metrics g which satisfy $K_g \leq -1$.

In dimension 2, the Gauss-Bonnet formula gives $\int_M K_g dv_g = 2\pi\chi(M)$, where $\chi(M)$ is the Euler characteristic of M. When $\chi(M) < 0$, this immediately implies that $MinVol(M) = 2\pi|\chi(M)| = MaxVol(M)$ and that the minimal and the maximal volumes are achieved for (and only for) metrics with constant sectional curvature -1.
In the higher even dimensional case, the Allendœrfer-Chern-Weil formulas also provide a lower bound of the minimal volume in terms of the Euler characteristic, however this bound is not sharp.
The simplicial volume (denoted by $SimplVol$), is defined as the infimum of $\|c\|_1 = \sum |\lambda_i|$ for all the linear real combinations of simplices $c = \sum \lambda_i \sigma_i$ which are closed chains c representing the fundamental n-class. Substituting this notion to the Euler characteristic, M. Gromov obtained the:

2.1. Theorem (M. Gromov, [Gr1]).- *For any compact manifold M, one has $MinVol(M) \geq C_n\ SimplVol(M)$, where C_n is a universal constant.*

For any compact manifold which admits a *hyperbolic metric* (i. e. a metric, denoted by g_o, whose sectional curvature is constant and equal to -1), an exact computation of the simplicial volume has been given by M. Gromov and W. Thurston ([Gr1]). By the theorem 2.1, it implies that $MinVol(X) \geq C'_n\ Vol(X, g_o)$. However, this estimate was also not sharp and justifies the

2.2. Theorem ([B-C-G 1,3]).- *Let X be a compact manifold with dimension $n \geq 3$. If X admits a hyperbolic metric g_o, then*
(i) $MinVol(X) = Vol(g_o) = MaxVol(X)$.
(ii) A metric g on X (such that $|K_g| \leq 1$) realizes the minimal volume iff it is isometric to g_o.
(iii) For any other riemannian manifold (Y^n, g) satisfying $Ricci_g \geq -(n-1).g$ and any map $f : Y^n \to X^n$, one has $Vol(Y, g) \geq |deg(f)|Vol(X, g_o)$

This theorem answers a conjecture of M. Gromov and provides the first exact computations of (non trivial) minimal volumes in dimension $n \geq 3$.

Proof : We first apply the theorem 1.3 to the map $id_X : (X, g) \to (X, g_o)$. It implies the existence of homotopic maps F_ϵ, of degree 1 (and thus surjective), such that $(1+\epsilon)Vol(g) \geq Vol\ (F_\epsilon(X), g_o) = Vol(g_o)$. Making $\epsilon \to 0$, we deduce the first equality of (i).

If $Vol(g) = Vol(g_o)$, the equality case in the theorem 1.3 (i) proves that the $F'_\epsilon s$ converge to an isometry. This proves (ii).
The same proof also gives (iii) if one notices that the integral on Y of the Jacobian of the $F'_\epsilon s$ provides an upper bound for the degree of f.
On the other hand, if $K_g \leq -1$, the second equality of (i) is proved by applying the theorem 1.3 (ii) to the map $id_X : (X, g_o) \to (X, g)$. $\diamond$

3. APPLICATIONS TO EINSTEIN MANIFOLDS :
An Einstein manifold is a Riemannian manifold whose Ricci curvature tensor is proportional to the metric. As the moduli space of Einstein metrics on a given compact manifold Y may also be characterized as the set of critical metrics for the functional $g \to$ *total scalar curvature of* g, the main problem is thus to describe this moduli space. In dimensions 2 and 3, it reduces to metrics of constant sectional curvature, so this problem is relevant only when the dimension is at least 4. However, in the non Kählerian case, very little is known. Even the simplest questions :

3.1. - *Does every n-manifold admit at least one Einstein metric?*

3.2. - *If a n-manifold X admits a negatively curved locally symmetric metric, is it the only Einstein metric on X (modulo homotheties)?*

are still conjectures in dimension $n \geq 5$. In dimension 4, there were some answers to the problem 3.1, involving the Euler characteristic $\chi(Y)$, the signature $\tau(Y)$ and the simplicial volume :

3.3. - *In the 3 following cases, a 4-dimensional compact manifold Y does not admit any Einstein metric :*
(i) If $\chi(Y) < 0$ (M. Berger, [Bes2]),
(ii) *If* $\chi(Y) - \frac{3}{2}|\tau(Y)| < 0$ (J. Thorpe, [Bes2] p 210),
(iii) *If* $\chi(Y) < \frac{1}{2592\pi^2}.SimplVol(Y)$ (M. Gromov, see [Bes2] theorem 6.47).

In dimension 4, nothing was known about the problem 3.2.
If true, the conjecture 3.2 would give a strong version of the Mostow's rigidity theorem. In fact, when the sectional curvature is a negative constant, the possible local models are all homothetic. On the contrary, for negative Einstein manifolds, the possible local models are not homothetic (see [Bes 2]). Thus, one must previously find the topological (or global) reason which excludes all the possible local models except one.

Let us thus assume that (Y, g) is a Einstein 4-dimensional manifold with $Ricci_g = (n-1)k.g$. The Allendœrfer-Chern-Weil formulas for the Euler characteristic and the signature give $\frac{4\pi^2}{3}\left(\chi(Y) \pm \frac{3}{2}\tau(Y)\right) = \int_Y P_\pm(R_g)dv_g$, where $P_\pm$ is a quadratic form in R_g, which satisfies $P_\pm(R_g) \geq k^2$ when g is Einstein, the equality being achieved when g has constant sectional curvature k (see for instance [Bes 2] or [Bes 3]). From this comes :

(3.4) $\frac{4\pi^2}{3}\left(\chi(Y) - \frac{3}{2}|\tau(Y)|\right) \geq k^2\ Vol(Y,g)$,

the equality being achieved when g has constant sectional curvature k. This is the classical proof of the theorems 3.3 (i) and (ii).
Let us now assume that there exists some map f of nonzero degree from Y to some hyperbolic 4-dimensional manifold X. The corollary 2.2 (iii) and the equality-case of (3.4) imply

(3.5) $Max(0,-k)^2\ Vol(Y,g) \geq |deg f|\ Vol(X,g_o) = \frac{4\pi^2}{3}|deg f|\left(\chi(X) - \frac{3}{2}|\tau(X)|\right)$

This implies that $k < 0$. *If* $\chi(Y) - \frac{3}{2}|\tau(Y)| = \chi(X) - \frac{3}{2}|\tau(X)|$ (for example if Y is homotopically equivalent to X), the inequalities (3.4) and (3.5) are equalities, thus $|deg f| = 1$ and $Vol(Y,g) = Vol(X,g_o)$. We thus are in the equality case of the theorem 1.3 (i) and (Y,g) is isometric to (X,g_o). This applies in particular to the case where $Y = X$ and $f = id_X$ and proves the

3.6. THEOREM ([B-C-G 1]).- *Let X be a compact 4-dimensional manifold which admits a real hyperbolic metric, then this is (modulo homotheties) the only Einstein metric on X.*

If $\chi(Y) - \frac{3}{2}|\tau(Y)| < |deg\ f|\ (\chi(X) - \frac{3}{2}|\tau(X)|)$, inequalities (3.4) and (3.5) are contradicted and Y *does not admit any Einstein metric* (A. Sambusetti, [Sam]), providing new answers to the conjecture 3.1 : in fact, from theorem 3.3 (ii), one might conjecture that any manifold Y which satisfies $\chi(Y) - \frac{3}{2}|\tau(Y)| > 0$ (or some other relation between χ and τ) admits an Einstein metric. M. Gromov's theorem 3.3 (iii) provided some counter-examples ([Bes 2] example 6.48); a complete answer is the :

3.7. PROPOSITION (A. Sambusetti, [Sam]).- *To every possible values k and t of the Euler characteristic and of the signature corresponds an infinity of (non homeomorphic) 4-dimensional manifolds Y_i which satisfy $\chi(Y_i) = k$ and $\tau(Y_i) = t$ and which admit no Einstein metric.*

The Y_i's are obtained by gluing, to any compact hyperbolic manifold X (such that $\chi(X) > k$), copies of $\pm CP^2$, $S^2 \times S^2$ or $S^2 \times \mathbf{T}^2$, in order to obtain the prescribed signature and Euler characteristic. One then apply the above Sambusetti's obstruction to the map of degree one : $Y_i \to X$.
These results may be compared to those obtained simultaneously by C. LeBrun ([LeB 1,2]), using Seiberg-Witten invariants, in particular the :

3.8. THEOREM (C. LeBrun, [LeB 1]).- *Let X be a compact 4-dimensional manifold which admits a complex hyperbolic metric, then this is (modulo homotheties) the only Einstein metric on X.*

4. SKETCH OF THE PROOF OF THE REAL SCHWARZ LEMMA (see [B-C-G 1,2,3] for a complete proof) :
Rescaling the metrics g_Y and g_X of the theorem 1.3, we may assume that $Ricci_{g_Y} \geq -(n-1)\ g_Y$ and $K_{g_X} \leq -1$.
Let us consider the riemannian universal coverings $(\tilde{Y}, \tilde{g}_Y)$ and $(\tilde{X}, \tilde{g}_X)$ of the compact riemannian manifolds (Y, g_Y) and (X, g_X), whose riemannian distance and riemannian volume-measure are denoted by $\rho_{\tilde{Y}}$, $\rho_{\tilde{X}}$ and $dv_{\tilde{g}_Y}$, $dv_{\tilde{g}_X}$. Let μ_y^c be the measure on $\tilde{Y}$ defined by $\mu_y^c = e^{-c\rho_{\tilde{Y}}(y,\bullet)} dv_{\tilde{g}_Y}$.
The infimum h_Y of the values c such that this measure is finite is called the *entropy* of (Y, g_Y). Another definition is $h_Y = \lim_{R\to+\infty} \left(\frac{1}{R} Log(Vol\ \tilde{B}(y,R)) \right)$, where $\tilde{B}(y,R)$ is the ball of $(\tilde{Y}, \tilde{g}_Y)$ centered at y and of radius R.

Let us consider positive measures μ on $\tilde{X}$ which are absolutely continuous w. r. t. the riemannian measure and such that the function $D_\mu(x) = \int_{\tilde{X}} \rho_{\tilde{X}}(x,z)\ d\mu(z)$ is finite. Following an idea of H. Furstenberg ([Fu], see also [D-E]), the *barycentre* $bar(\mu)$ is defined as the unique point where the function D_μ achieves its minimum (the existence comes from the triangle inequality and the uniqueness from the convexity of $\rho_{\tilde{X}}$). The barycentre is thus given by the implicit equation $(dD_\mu)_{|_{bar(\mu)}} = 0$.

Let $\tilde{f} : \tilde{Y} \to \tilde{X}$ be the lift of f, *we define* $\tilde{F}_c$ *by* $\tilde{F}_c(y) = bar(\tilde{f}_* \mu_y^c)$, where $\tilde{f}_* \mu_y^c$ is the push-forward by $\tilde{f}$ of the measure μ_y^c. If $\rho = [f]$ is the induced representation $\pi_1(Y) \to \pi_1(X)$, $\tilde{f}$ (and thus $\tilde{f}_*$ also) satisfies the equivariance property $\tilde{f} \circ \gamma = \rho(\gamma) \circ \tilde{f}$ for any deck-transformation $\gamma \in \pi_1(Y)$. The invariance of the distance and of the riemannian measure by deck-transformations implies that $bar\,(\rho(\gamma)_* \mu) = \rho(\gamma)\,(bar(\mu))$ and $\mu_{\gamma.y}^c = \gamma_* \mu_y^c$. Thus $\tilde{F}_c$ is equivariant w. r. t. the same representation $\rho = [f]$, and goes down to a map $F_c : Y \to X$ which is homotopic to f.
Let $c = (1+\epsilon)\ h_Y$, we want to prove that, when $\epsilon \to 0_+$, F_c answers theorem 1.3.
Let us define $\Delta : \tilde{X} \times \tilde{Y} \to \mathbf{R}$ by $\Delta(x,y) = D_{\tilde{f}_* \mu_y^c}(x)$ and let ∂^1 (resp. ∂^2) be the derivatives w. r. t. the first (resp. the second) parameter.
By the definition of $\tilde{F}_c$ and by the variational characterization of the barycentre, $\tilde{F}_c$ is defined by the implicit equation : $\partial^1 \Delta_{|_{(\tilde{F}_c(y),y)}} = 0$.
By derivation, we get $\partial^1\partial^1\Delta_{|_{(\tilde{F}_c(y),y)}}(d\tilde{F}_c(u), v) = -\ \partial^2\partial^1\Delta_{|_{(\tilde{F}_c(y),y)}}(u,v)$ for any $u \in T_y\tilde{Y}$ and $v \in T_{\tilde{F}_c(y)}\tilde{X}$. This writes

$$(4.1) \qquad \int_{\tilde{Y}} Dd\rho_{\tilde{X}_{|_{(\tilde{F}_c(y),\tilde{f}(z))}}}(d\tilde{F}_c(u), v)\ d\mu_y^c(z)$$
$$= c \int_{\tilde{Y}} d\rho_{\tilde{X}_{|_{(\tilde{F}_c(y),\tilde{f}(z))}}}(v)\ d\rho_{\tilde{Y}_{|_{(y,z)}}}(u)\ d\mu_y^c(z) \leq c\ \tilde{g}_X(H_y(v), v)^{1/2}\ \tilde{g}_Y(K_y(u), u)^{1/2},$$

where the tensor $Dd\rho_{\tilde{X}}$ is computed by derivation w. r. t. the first parameter and where H_y (resp.K_y) is the symmetric endomorphism of $T_{\tilde{F}_c(y)}\tilde{X}$ (resp. of $T_y\tilde{Y}$) associated to the quadratic form $v \to \int_{\tilde{Y}} (d\rho_{\tilde{X}_{|_{(\tilde{F}_c(y),\tilde{f}(z))}}}(v))^2 d\mu_y^c(z)$ (resp.

to the quadratic form $u \to \int_{\tilde{Y}} (d\rho_{\tilde{Y}|_{(y,z)}}(u))^2 d\mu_y^c(z)$).

As the gradient of $\rho_{\tilde{X}}(\bullet, \tilde{f}(z))$ is a unit vector normal to the geodesic spheres centered at $\tilde{f}(z)$, the second fundamental form of these spheres is equal to $Dd\rho_{\tilde{X}|_{(\bullet,\tilde{f}(z))}}$. As $K_{g_X} \leq -1$, the Rauch's comparison theorem provides the lower bound $coth\ \rho_{\tilde{X}}(\bullet, \tilde{f}(z))$ for the principal curvatures of these spheres, and thus implies that $\tilde{g}_X - d\rho_{\tilde{X}} \otimes d\rho_{\tilde{X}}$ is a lower bound for $Dd\rho_{\tilde{X}}$.

First plugging this in (4.1), replacing c by its value and then writing the induced inequality for determinants, we obtain :

$$(4.2)\quad \tilde{g}_X \left((Id - H_y) \circ d_y \tilde{F}_c(u), v \right) \leq (1+\epsilon)\ h_Y\ \tilde{g}_X\ (H_y(v), v)^{1/2}\ \tilde{g}_y\ (K_y(u), u)^{1/2},$$

$$(4.3)\qquad (1+\epsilon)^{-n} h_Y^{-n}\ \frac{det(Id - H_y)}{(detH_y)^{1/2}}\ |det(d_y \tilde{F}_c)| \leq (det\ K_y)^{1/2} \leq \left(\tfrac{1}{n} Trace\ K_y\right)^{n/2}$$

As $\|d\rho_{\tilde{Y}}\| = 1 = \|d\rho_{\tilde{X}}\|$, we have Trace $K_y = 1 = TraceH_y$. On the other hand, *the function* $\delta : A \to \frac{det(I-A)}{(detA)^{1/2}}$ (defined on the set of symmetric positive definite $n \times n$ matrices ($n \geq 3$) whose trace is equal to 1) achieves its minimum at the unique point $A_o = \frac{1}{n} I$.

Plugging this in (4.3) gives : $|det(d_y \tilde{F}_c)| \leq (1+\epsilon)^n \left(\frac{h_Y}{n-1}\right)^n$. We end the proof of the general inequality of the theorem 1.3 by applying the comparison theorem of R. L. Bishop : i. e. the assumption $Ricci_{g_Y} \geq -(n-1)$ implies that $h_Y \leq n-1$. $\diamond$

When $K_{g_Y} < 0$, one may identify $\tilde{Y}$ with a ball and compactify it by addition of the sphere, called the *ideal boundary* and denoted $\partial\tilde{Y}$. One may then extend continuously $\tilde{f}$ to a map $\bar{f} : \partial\tilde{Y} \to \partial\tilde{X}$.

Let us fix an origin y_o in $\tilde{Y}$. A sequence of measures $(\mu_{y_o}^{c_n}(\tilde{Y}))^{-1} \mu_y^{c_n}$ converges, on the compact set $\tilde{Y} \cup \partial\tilde{Y}$ (when $c_n \to h_Y$), to a measure μ_y, with support in $\partial\tilde{Y}$, which is known as *the Patterson-Sullivan measure* and satisfies $\mu_y = e^{-h_Y B_{\tilde{Y}}(y,\bullet)} \mu_{y_o}$, where $B_{\tilde{Y}}(y,\theta) = \lim_{t \to +\infty} [\rho_{\tilde{Y}}(c_\theta(t), y) - t]$ and where c_θ is the normal geodesic-ray from y_o to θ.

Mimicking the previous proof (just replacing $\rho_{\tilde{Y}}$ and $\rho_{\tilde{X}}$ by $B_{\tilde{Y}}$ and $B_{\tilde{X}}$), *we define* $\tilde{F}$ *by* $\tilde{F}(y) = bar(\bar{f}_* \mu_y)$ and prove the inequality of the theorem 1.3 (ii) : $|det(d_y \tilde{F})| \leq \left(\frac{h_Y}{n-1}\right)^n \leq 1$.

When $|det(d_y \tilde{F})| \geq \left(\frac{h_Y}{n-1}\right)^n$, and a fortiori when $|Jac\tilde{F}| = 1$, the analogues of the inequalities (4.3) are equalities which imply that $K_y = \frac{1}{n} I$ and that δ achieves its minimum at the point H_y, which is thus equal to $A_o = \frac{1}{n} I$. Plugging this in (4.2) and replacing v by $d_y \tilde{F}(u)$, we deduce that $d_y \tilde{F}$ is a contracting map whose determinant is equal to 1, thus it is isometric (see [B-C-G 2,3] for more explanations).

On the contrary, when K_{g_Y} may take both signs, we have to prove that the F_c's admit a limit when $c \to h_Y$, that this limit is a contracting map and that the property of preserving global volumes implies that it is isometric (see [B-C-G 1] sections 7 and 8). $\diamond$

5. ANOTHER VERSION OF THE REAL SCHWARZ LEMMA :
The present version of the real Schwarz lemma is adapted to the case where the target-space is a compact quotient of a hyperbolic space modelled on the real or complex or quaternionic or Cayley field (the canonical basis of the field beeing denoted by $\{1, J_1, ..., J_d\}$).

5.1. THEOREM ([B-C-G 1,2,3]).- *Let (X, g_X) be a compact locally symmetric manifold with negative curvature and (Y, g_Y) be any compact riemannian manifold such that dim X = dim $Y \geq 3$, then any continuous map $f : Y \to X$ may be deformed to a family of C^1 maps $F_\epsilon(\epsilon \to 0_+)$ such that $|Jac\ F_\epsilon| \leq \left(\frac{h_Y+\epsilon}{h_X}\right)^n$. In particular, one has $(h_Y)^n Vol(Y) \geq |deg f|(h_X)^n Vol(X)$. Moreover, if $(h_Y)^n Vol(Y) = |deg f|(h_X)^n Vol(X)$, then Y is also locally symmetric and f is homotopic to a riemannian covering F (an isometry when $|deg f| = 1$).*

5.2. REMARKS.- (1) This theorem proves conjectures of A. Katok and M. Gromov about the minimal entropy.
(2) When (Y, g_Y) has negative curvature and f is a homotopy equivalence, the following proof provides a direct construction of $F : Y \to X$ which satisfies $|Jac\ F(y)| \leq \left(\frac{h_Y}{h_X}\right)^n$ and $d_y F$ is isometric in the equality case.

Sketch of the proof : We already proved the theorem 5.1 and the remark 5.2 (2) when (X, g_X) is (locally) real hyperbolic (see section 4). In the other locally symmetric cases, the proof is exactly the same, except for the fact that, expliciting the new expression of $Dd\rho_{\tilde{X}}$, we have to prove that the function $A \to \frac{det(I-A-\sum_i J_i A J_i)}{(detA)^{1/2}}$ still achieves its minimum at the unique point $A_o = \frac{1}{n} I$. This comes from the log-concavity of the determinant which reduces the problem to minimizing the previous function δ (see [B-C-G 1]). $\diamond$

5.3. COROLLARY (G.D. Mostow).- *Let (X, g_X) and (Y, g_Y) be two compact negatively curved locally symmetric manifolds such that dim$X$ = dim$Y \geq 3$, then any homotopy-equivalence $f : Y \to X$ is homotopic to an isometry.*

Proof : Let $g : X \to Y$ such that $g \circ f \sim id_Y$. By the remark 5.2 (2), there exist $F \sim f$ and $G \sim g$ such that $|Jac(G \circ F)| \leq \left(\frac{h_X}{h_Y}\right)^n \left(\frac{h_Y}{h_X}\right)^n$. As the degree of $G \circ F$ is equal to 1, this inequality is an equality and we are in the equality case of the remark 5.2 (2), thus F is an isometry. $\diamond$

This provides a unified proof for the Mostow's rigidity theorem. Moreover, the isometry F is explicitely constructed (see section 4)

6. APPLICATION TO DYNAMICS AND LICHNEROWICZ'S CONJECTURE :
Let $\phi_t^Y : \dot{c}(0) \to \dot{c}(t)$ (for any geodesic c) be the geodesic flow of Y. Two riemannian manifolds Y and$X$ are said to *have the same dynamics* iff there exists a

C^1-diffeomorphism Φ between their unitary tangent bundles UY and UX which exchanges their geodesic flows, i. e. $\Phi \circ \phi_t^Y = \phi_t^X \circ \Phi$. The fundamental question is : *two riemannian manifolds having the same dynamics are they isometric?* This is generally false, for there exists non isometric manifolds all of whose geodesic are closed with the same period (see [Bes 1]). C. B. Croke and J. P. Otal proved this conjecture to be true for negatively curved surfaces. In any dimension, we get the

6.1. Theorem ([B-C-G 1]).- *Any Riemannian manifold which has the same dynamics as a negatively curved locally symmetric one is isometric to it.*

Proof : As $UY \approx UX$ and $n \geq 3$, the manifolds under consideration Y and X are homotopically equivalent. As the volume and the entropy are invariants of the dynamics, the assumption implies that $h_Y = h_X$ and $Vol(Y) = Vol(X)$; we thus are in the equality case of the theorem 5.1 and Y and X are isometric. $\diamond$

A riemannian manifold is said to be *locally harmonic* when all geodesic spheres of its universal covering have constant mean curvature. Any locally symmetric manifold of rank one is locally harmonic. A. Lichnerowicz asked for the converse question : *Consider any locally harmonic manifold, is it locally symmetric of rank one?.*

When the universal covering $\tilde{X}$ is compact, this conjecture was proved by Z. Szabo ([Sz]). In the case where $\tilde{X}$ is noncompact, the geodesics have no conjugate points ([Bes 1]), and the conjecture is not significantly changed when assuming the sectional curvature to be negative. A counter-example (admitting no compact quotient) was given by E. Damek and F. Ricci ([D-R]). Assuming that $\tilde{X}$ admits a compact quotient, we get the

6.2. Corollary ([B-C-G 1]).- *Any compact negatively curved locally harmonic manifold is locally symmetric of rank one.*

Proof : Under these assumptions, P. Foulon and F. Labourie ([F-L]) proved that the manifold has the same dynamics as a negatively curved locally symmetric manifold. We conclude by applying the theorem 6.1. $\diamond$

References

[B-P] R. Benedetti, C. Petronio, *Lectures on Hyperbolic Geometry*, Universitext, Springer 1993.

[Bes 1] A. L. Besse, *Manifolds all of whose geodesics are closed*, Ergebnisse der Math., Springer 1978.

[Bes 2] A. L. Besse, *Einstein Manifolds*, Ergebnisse der Math., Springer 1987.

[Bes3] A. L. Besse, *Géométrie riemannienne en dimension 4*, Séminaire Arthur Besse, Texte Mathématiques, Cedic-Nathan.

[B-C-G 1] G. Besson, G. Courtois, S. Gallot, *Entropies et rigidités des espaces localement symétriques de courbure négative*, Geom. And Funct. Anal.5 (1995), pp. 731-799.

[B-C-G 2] G. Besson, G. Courtois, S. Gallot, *Minimal Entropy and Mostow's rigidity theorems*, Ergod. Theory Dynam. Syst. 16 (1996), pp. 623-649.

[B-C-G 3] G. Besson, G. Courtois, S. Gallot, *Lemme de Schwarz réel et applications géométriques*, prépublications Centre Math. Ec. Polytechnique (98-7).

[D-R] E. Damek, F. Ricci, *A class of non-symmetric harmonic riemannian spaces*, Bull. Amer. Math Soc. 27 (1992), pp. 139-142.

[D-E] E. Douady, C. Earle, *Conformally natural extension of homeomorphisms of the circle*, Acta Math. 157 (1986), pp. 23-48.

[E-S] J. Eells, J. H. Sampson, *Harmonic mappings of Riemannian manifolds*, Amer. Journ. Math. 86 (1964), pp. 109-160.

[F-L] P. Foulon, F. Labourie, *Sur les variétés compactes asymptotiquement harmoniques*, Invent. Math. 109 (1992), pp. 97-111.

[Fu] H. Furstenberg, *A Poisson formula for semi-simple Lie groups*, Annals of Maths 77 (1963), pp. 335-386.

[Gr1] M. Gromov, *Volume and Bounded Cohomology*, Publ. Math. I. H. E. S. 56 (1981), pp : 213-307.

[Ha] P. Hartman, *On homotopic harmonic maps*, Canad. J. Math. 19 (1967), pp. 673-687.

[Jo] J. Jost, *Riemannian Geometry and Geometric Analysis*, Universitext, Springer 1998.

[LeB 1] C. LeBrun, *Einstein metrics and Mostow rigidity*, Math. Res. Lett. 2 (1995), pp. 1-8.

[LeB 2] C. LeBrun, *Four-manifolds without Einstein metrics*, Math. Res. Lett. 2 (1996), pp. 133-147.

[Mok] N. Mok, *Metric Rigidity Theorems on Hermitian locally symmetric Manifolds*, Series in pure Maths 6, World Scientific.

[Sam] A. Sambusetti, *An obstruction to the existence of Einstein metrics on 4-Manifolds*, to appear in Math. Annalen.

[Sz] Z. Szabo, *The Lichnerowicz conjecture on harmonic manifolds*, J. Diff. Geom. 31 (1990), pp. 1-28.

Sylvestre Gallot
Institut Fourier Maths Pures
UMR 5582 CNRS-UJF
B. P. 74
38402 Saint Martin d'Heres Cedex
France

Evolution of Hypersurfaces by Their Curvature in Riemannian Manifolds

Gerhard Huisken

Abstract. We study hypersurfaces in Riemannian manifolds moving in normal direction with a speed depending on their curvature. The deformation laws considered are motivated by concrete geometrical and physical phenomena and lead to second order nonlinear parabolic systems for the evolving surfaces. For selected examples of such flows the article investigates local and global geometric properties of solutions. In particular, it discusses recent results on the singularity formation in mean curvature flow of meanconvex surfaces (joint with C. Sinestrari) and applications of inverse mean curvature flow to asymptotically flat manifolds used for the modelling of isolated systems in General Relativity (joint with T. Ilmanen).

1991 Mathematics Subject Classification: 53A10, 53A35, 58G11
Keywords and Phrases: Geometric evolution equations, Mean curvature flow, Inverse mean curvature flow, Penrose inequality

1 The evolution equations

Let $F_0 : \mathcal{M}^n \to \mathbb{R}^{n+1}$ be a smooth immersion of an n–dimensional hypersurface $\mathcal{M}_0^n = F_0(\mathcal{M}^n)$ in a smooth Riemannian manifold $(N^{n+1}, \bar{g})$, $\quad n \geq 2$. We study one–parameter families of immersions $F : \mathcal{M}^n \times [0, T[\to (N^{n+1}, \bar{g})$ of hypersurfaces $\mathcal{M}_t^n = F(\cdot, t)(\mathcal{M}^n)$ satisfying an initial value problem

$$\frac{\partial F}{\partial t}(p,t) = -f\nu(p,t), \qquad p \in \mathcal{M}^n,\ t \in [0, T[, \tag{1.1}$$

$$F(\cdot, 0) = F_0, \tag{1.2}$$

where $\nu(p,t)$ is a choice of unit normal at $F(p,t)$ and $f(p,t)$ is some smooth homogeneous symmetric function of the principal curvatures of the hypersurface at $F(p,t)$.

We are interested in the case where $f = f(\lambda_1, \cdots, \lambda_n)$ is monotone with respect to the principal curvatures $\lambda_1, \cdots, \lambda_n$ such that (1.1) is a nonlinear parabolic system of second order. The interaction between geometric properties of the data

$\mathcal{M}_0^n$, $(N^{n+1}, \bar{g})$ and the local and global behaviour of solutions leads to many interesting phenomena and has applications to a number of models in mathematical physics.

Typical examples considered here are the mean curvature flow $f = -H = -(\lambda_1 + \cdots + \lambda_n)$, the inverse mean curvature flow $f = H^{-1}$ and fully nonlinear flows such as the the Gauss curvature flow $f = -K = -(\lambda_1 \cdots \lambda_n)$ or the harmonic mean curvature flow, $f = -(\lambda_1^{-1} + \cdots + \lambda_n^{-1})^{-1}$. We investigate some new developments in the mathematical understanding of these evolution equations and include some applications such as the use of the inverse mean curvature flow for the study of asymptotically flat manifolds in General Relativity.

To fix notation, let $\bar{g} = \{\bar{g}_{\alpha\beta}\}_{0 \le \alpha,\beta \le n}$, $\bar{\nabla}$ and $\bar{\mathrm{Riem}} = \{\bar{R}_{\alpha\beta\gamma\delta}\}$ be the metric, the connection and the Riemann curvature tensor of the target manifold respectively, where the indices sometimes refer to local coordinates $\{y^\alpha\}$ and sometimes to a suitable local orthonormal frame $\{e_\alpha\}$. We write $\bar{g}^{-1} = \{\bar{g}^{\alpha\beta}\}$ for the inverse of the metric and use the Einstein summation convention to sum over repeated indices. The Ricci curvature $\bar{\mathrm{Ric}}$ and scalar curvature $\bar{\mathrm{R}}$ of $(N^{n+1}, \bar{g})$ are then given by

$$\bar{R}_{\alpha\beta} = \bar{g}^{\gamma\delta} \bar{R}_{\alpha\gamma\beta\delta}, \qquad \bar{R} = \bar{g}^{\alpha\beta} \bar{R}_{\alpha\beta},$$

and the sectional curvatures (in an orthonormal frame) are computed as $\bar{\sigma}_{\alpha\beta} = \bar{R}_{\alpha\beta\alpha\beta}$.

If $F : \mathcal{M}^n \to (N^{n+1}, \bar{g})$ is a smooth hypersurface immersion, we denote by $g = \{g_{ij}\}_{1 \le i,j \le n}$, ∇, Riem the induced metric, connection and intrinsic curvature. In an adapted local orthonormal frame $e_1, \cdots, e_n, \nu$ with unit normal ν the second fundamental form $A = \{h_{ij}\}$ is then at each point given by the symmetric matrix

$$h_{ij} = < \bar{\nabla}_{e_i} \nu, e_j > = - < \nu, \bar{\nabla}_{e_i} e_j >,$$

such that the eigenvalues $\lambda_1, \cdots, \lambda_n$ are the principal curvatures of the hypersurface at this point, leading to the classical scalar invariants mentioned earlier.

If the initial hypersurface is smooth and closed, ie compact without boundary, it is wellknown that a smooth solution of the flow exists for a short time, provided on the initial surface the speed function f is elliptic in the sense that $\partial f / \partial \lambda_i < 0$, $1 \le i \le n$. In particular, mean curvature flow always admits shorttime solutions and inverse mean curvature flow admits shorttime solutions for meanconvex initial data, whereas the Gauss curvature flow and harmonic mean curvature flow require the initial data to be convex $\big(\lambda_i > 0\big)$.

Working in the class of surfaces where shorttime existence is guaranteed, the interesting task is to understand the longterm change in the shape of solutions, and to characterise their asymptotic behaviour both for large times and near singularities. For this purpose evolution equations have to be established for all relevant geometric quantities, in particular for the second fundamental form.

THEOREM 1.1 *On any solution $\mathcal{M}_t^n = F(\cdot, t)(\mathcal{M}^n)$ of (1.1) the following equations hold:*

(i) $\frac{\partial}{\partial t} g_{ij} = 2 f h_{ij}$,

(ii) $\frac{\partial}{\partial t}(d\mu) = fH(d\mu)$,

(iii) $\frac{\partial}{\partial t}\nu = -\nabla f$,

(iv) $\frac{\partial}{\partial t}h_{ij} = -\nabla_i\nabla_j f + f\big(h_{ik}h^k_j - \bar{R}_{0i0j}\big)$,

(v) $\frac{\partial}{\partial t}H = -\Delta f - f\big(|A|^2 + \bar{\mathrm{Ric}}(\nu,\nu)\big)$.

Here $d\mu$ is the induced measure on the hypersurface, the index 0 stands for the normal direction and Δ is the Laplace–Beltrami operator with respect to the time-dependent induced metric on the hypersurface.

Notice that $-\Delta f - f\big(|A|^2 + \bar{\mathrm{Ric}}(\nu,\nu)\big) = Jf$ is the Jacobi operator acting on f, as is wellknown from the second variation formula for the area. The relations above are consequences of the definitions and the Gauss–Weingarten relations. To convert the evolution equations for the curvature into parabolic systems on the hypersurface, we introduce for each speed function f the nonlinear operator L_f by setting

$$L_f u = L_f^{ij}\nabla_i\nabla_j u := -\frac{\partial \hat{f}}{\partial h_{ij}}\nabla_i\nabla_j u,$$

where $\hat{f}$ is the symmetric function f considered as a function of the h_{ij}. Note that for mean curvature flow $L_H = \Delta$ is the Laplace–Beltrami operator, for inverse mean curvature flow $f = H^{-1}$ we have $L_f = (1/H^2)\Delta$ and in general L_f is an elliptic operator exactly when f is elliptic. Using then the crucial commutator relations for the second derivatives of the second fundamental form one derives after long but straightforward calculations

COROLLARY 1.2 *On any solution $\mathcal{M}^n_t = F(\cdot,t)(\mathcal{M}^n)$ of (1.1) the second fundamental form h_{ij} and the speed f satisfy*

$$\begin{aligned}
\frac{\partial}{\partial t}h_{ij} &= L_f^{kl}\nabla_k\nabla_l h_{ij} - \frac{\partial^2 f}{\partial h_{kl}\partial h_{pq}}\nabla_i h_{kl}\nabla_j h_{pq} \\
&+ \frac{\partial f}{\partial h_{kl}}\big\{h_{kl}h_{im}h_{mj} - h_{km}h_{il}h_{mj} + h_{kj}h_{im}h_{ml} - h_{km}h_{ij}h_{ml} \\
&+ \bar{R}_{kilm}h_{mj} + \bar{R}_{kijm}h_{ml} + \bar{R}_{mijl}h_{km} + \bar{R}_{0i0j}h_{kl} - \bar{R}_{0k0l}h_{ij} + \bar{R}_{mljk}h_{im} \\
&+ \bar{\nabla}_k\bar{R}_{0jil} + \bar{\nabla}_i\bar{R}_{0ljk}\big\} + f\big(h_{ik}h^k_j - \bar{R}_{0i0j}\big), \\
\frac{\partial}{\partial t}f &= L_f^{ij}\nabla_i\nabla_j f - f\frac{\partial f}{\partial h_{ij}}\big(h_{ik}h^k_j + \bar{R}_{0i0j}\big).
\end{aligned}$$

The curvature terms in this nonlinear reaction-diffusion system provide the key for understanding the interaction between geometric properties of the hypersurface and the ambient manifold. They are the tool to study these geometric phenomena with analytical means. We will now describe some recent developments for selected choices of f: Section 2 discusses the formation of singularities in the mean curvature flow, especially in the case of mean convex surfaces. In section 3 some fully nonlinear equations like Gauss curvature flow are considered.

Finally in section 4 it is explained how the inverse mean curvature flow provides an approach to the Penrose inequality for the total mass of an asymptotically flat manifold.

2 Singularities of the mean curvature flow

In the case of mean curvature flow $f = -H$ it is well known [19] that for closed initial surfaces the solution of (1.1)–(1.2) exists on a maximal time interval $[0, T[$, $0 < T \leq \infty$. If $T < \infty$, as is always the case in Euclidean space, the curvature of the surfaces becomes unbounded for $t \to T$. One would like to understand the singular behaviour for $t \to T$ in detail, having in mind a possible controlled extension of the flow beyond such a singularity. See [22] for a review of earlier results concerning local and global properties of mean curvature flow. We will not discuss singularities in weak formulations of the flow, a good reference in this direction is [32].

Since the shape of possible singularities is a purely local question, we may restrict attention to the case where the target manifold is Euclidean space. Nevertheless, in the light of an abundance even of homothetically shrinking examples with symmetries, the possible limiting behaviour near singularities seems in general beyond classification at this stage.

In recent joint work of C. Sinestrari and the author [26], [27] the additional assumption of nonnegative mean curvature is used to restrict the range of possible phenomena, while still retaining an interestingly large class of surfaces. We derive new a priori estimates from below for all elementary symmetric functions of the principal curvatures, exploiting the one-sided bound on the mean curvature. The estimates turn out to be strong enough to conclude that any rescaled limit of a singularity is (weakly) convex.

Define by

$$S_k(\lambda) = \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} \lambda_{i_1} \lambda_{i_2} \cdots \lambda_{i_k}$$

the elementary symmetric functions of the principal curvatures with $S_1 = H$. Then [27] establishes the estimates

THEOREM 2.1 *(H.-Sinestrari) Suppose $F_0 : \mathcal{M} \to \mathbb{R}^{n+1}$ is a smooth closed hypersurface immersion with nonnegative mean curvature. For each k, $2 \leq k \leq n$, and any $\eta > 0$ there is a constant $C_{\eta,k}$ depending only on n, k, η and the initial data, such that everywhere on $\mathcal{M} \times [0, T[$ the estimate*

$$S_k \geq -\eta H^k - C_{\eta,k} \tag{2.1}$$

holds uniformly in space and time.

The proof proceeds by induction on the degree k of S_k and relies heavily on the algebraic properties of the elementary symmetric functions, the structure of the curvature evolution in this particular flow and the Sobolev inequality for

hypersurfaces. In each step of the iteration an a priori estimate is proved for a quotient

$$Q_k = \frac{S_k}{S_{k-1}}$$

of consecutive elementary symmetric polynmials, making use of the concavity properties of this function. Using techniques in [20] and [29] the result can be extended to starshaped surfaces in $\mathbb{R}^{n+1}$ and to hypersurfaces in Riemannian manifolds.

Similarly as in the theory of minimal surfaces the structure of singularities is then studied by blowup methods, in this case by parabolic rescaling in space and time, compare [15], [21], [26]. Since η is arbitrary in the above estimate and the mean curvature $S_1 = H$ tends to infinity near a singularity, the scaling invariance is broken in inequality (2.1) and implies that near a singularity each S_k becomes nonnegative after appropriate rescaling:

Corollary 2.2 *Let $\mathcal{M}_t$ be a mean convex solution of mean curvature flow on the maximal time interval $[0, T[$ as in Theorem 1.1. Then any smooth rescaling of the singularity for $t \to T$ is convex.*

The structure of the rescaled limit depends on the blowup rate of the singularity: If the quantity $\sup(T-t)|A|^2$ is uniformly bounded (type I singularity), the rescaling will yield a selfsimilar, homothetically shrinking solution of the flow which is completely classified in the case of positive mean curvature, see [21] and [22]. If the quantity $\sup(T-t)|A|^2$ is unbounded (type II singularity), the rescaling of the singularity can be done in such a way that an "eternal solution" (ie defined for all time) of mean curvature flow results where the maximum of the curvature is attained on the surface. In the convex case such solutions were shown by Hamilton to move isometrically by translations, [16]. Hence, combining the classification of type I singularities in [22], the result of Hamilton and the convexity result in Corollary 2.2, one derives a description of all possible singularities (type I and type II) in the mean convex case, compare [27].

Open problems which have to be adressed for the future goal of continuing the flow by surgery concern the classification of convex translating solutions, Harnack estimates for the mean curvature, more precise estimates on the rate of convergence as well as higher order asymptotics near singularities. Some guidance on the possible higher order behaviour near singularities can be taken from the degenerate examples constructed in [6]. A Harnack estimate for the mean curvature has so far only been obtained in the convex case [16], which is too restrictive for many applications. The work of Hamilton on the Ricci flow [15] has a close relation to the mean curvature flow and indicates a strategy for the extension of the flow past singularities once stronger estimates are available [17].

We conclude this section with the one-dimensional case, where an embedded curve is evolving in the plane or on some smooth surface by the curve shortening flow. The remarkable articles of Grayson on this flow [13],[14] show by a number of global arguments that for embedded curves no finite time singularity can occur unless the whole curve contracts to a single point.

The structure of all possible singularities in this case is now well understood: There are no embedded type I singularities except the shrinking circle, which is

the desired outcome, and the only possible rescaling of a type II singularity is a so called grim reaper curve given by $y = \log\cos x$. To prove Graysons result it is therefore sufficient to give an argument excluding this last curve as a possible limiting shape. Such an argument is provided both by Hamilton in [18], where an isoperimetric estimate for the area in subdivisions of the enclosed region is shown, and by the author in [23], where a lower bound for the ratio between extrinsic and intrinsic distances on the evolving curve is proved.

To describe the last result let $F\colon S^1 \times [0,T] \to \mathbb{R}^2$ be a closed embedded curve moving by the curve shortening flow. If $L = L(t)$ is the total length of the curve, the intrinsic distance function l along the curve is smoothly defined only for $0 \le l < L/2$, with conjugate points where $l = L/2$. We therefore define a smooth function $\psi\colon S^1 \times S^1 \times [0,T] \to \mathbb{R}$ by setting

$$\psi := \frac{L}{\pi}\sin(\frac{l\pi}{L}).$$

With this choice of ψ, and with d being the extrinsic distance between two points on the curve, the isoperimetric ratio d/ψ approaches 1 on the diagonal of $S^1 \times S^1$ for any smooth embedding of S^1 in $\mathbb{R}^2$ and the ratio d/ψ is identically one on any round circle.

THEOREM 2.3 *Let $F\colon S^1 \times [0,T] \to \mathbb{R}^2$ be a smooth embedded solution of the curve shortening flow (1.1). Then the minimum of d/ψ on S^1 is nondecreasing; it is strictly increasing unless $d/\psi \equiv 1$ and $F(S^1)$ is a round circle.*

Clearly the estimate prevents a grim reaper type singularity. The proof uses the maximum principle on the cross product of the curve with itself. It is an open problem whether similar lower order estimates can be used for the study or exclusion of certain singularities in higherdimensional flows.

3 FULLY NONLINEAR FLOWS

The Gauss curvature flow, where the speed $f = -K = -\big(\lambda_1 \cdots \lambda_n\big)$ is the product of the principle curvatures, was first introduced by Firey [10] as a model for the changing shape of a tumbling stone being worn from all directions with uniform intensity. The flow is parabolic only in the class of convex surfaces and much more nonlinear in its analytic behaviour than the mean curvature flow. Tso [30] proved existence, uniqueness and convergence of closed convex hypersurfaces to a point for this flow without however determining the limiting shape of the contracting surface. The conjecture of Firey (1974) that the limiting shape is that of a sphere regardless of the initial data, was only recently confirmed by Andrews [2]:

THEOREM 3.1 *(Andrews) Let $\mathcal{M}_0^2$ be a smooth closed strictly convex initial surface in $\mathbb{R}^3$. Then there is a unique smooth solution of (1.1) with $f = -K$ on the time interval $[0,T[$, where $T = V(\mathcal{M}_0^2)/4\pi$ is determined by the enclosed volume of the initial surface, and the surfaces converge to a round sphere after appropriate rescaling.*

The corresponding result for mean curvature flow was obtained earlier by the author in [19] and for a large class of speed functions f including the harmonic mean curvature flow by Andrews in [1]. If the Gauss curvature K is replaced by some power K^α, a whole new range of interesting phenomena appears. If the homogeneity is 1, ie $\alpha = 1/n$, Chow proved contraction to a point and roundness of the limiting shape, [8]. In [5] Andrews shows that in the interval $1/(n+2) < \alpha \leq 1/n$ there is at least some smooth limiting shape at the end of the contraction, while for small values of α a degeneration of the surface near the end of the contraction is expected.

In the special case $\alpha = 1/(n+2)$, the evolution equation (1.1) becomes affine invariant. In line with the results just mentioned Andrews [3] proves by an extension of Calabi's estimate on the cubic ground form that convex initial data contract smoothly to a point in finite time, with ellipsoids as the natural unique limiting shape. As a consequence he derives an elegant proof of the affine isoperimetric inequality. Compare also the work of Sapiro and Tannenbaum [28] on the affine evolution of curves, which has applications in image processing.

For convex hypersurfaces in general Riemannian manifolds speedfunctions f such as the harmonic mean curvature or other quotients of elementary symmetric functions seem to have the best algebraic behaviour. In mean curvature flow the derivatives of the ambient curvature in the evolution equations of Corollary 1.2 are analytically hard to control, compare the dependance of the main result in [20] on these terms. For harmonic mean curvature flow and flows of similar structure Andrews derives an optimal convergence result for hypersurfaces having sufficiently positive principal curvatures in relation to the ambient curvature, [4]. In particular, he shows that such flows contract convex hypersurfaces in manifolds of positive sectional curvature to a point and gives a new argument for the classical 1/4-pinching theorem.

All speedfunctions considered so far were pointing in the same direction as the mean curvature vector, corresponding to contractions in the case of convex surfaces. In the last section we consider an expanding version of the flow.

4 The inverse mean curvature flow

The inverse mean curvature flow $f = H^{-1}$ is well posed for surfaces of positive mean curvature and characterised by its property that the area element is growing exponentially at each point: From Theorem 1.1(i) we have $\partial/\partial t(d\mu) = d\mu$. In particular, the total area of a smooth closed evolving surface is completely determined by its initial area:

$$|M_t^n| = |M_0^n| \exp(t).$$

The standard example is the exponentially expanding sphere of radius $R(t) = R(0)\exp(t/n)$. Further interesting properties of the flow follow from the evolution equation for the mean curvature H, which we derive from the evolution equation for the speed f.

$$\frac{\partial H}{\partial t} = \frac{\Delta H}{H^2} - \frac{2|\nabla H|^2}{H^3} - \frac{|A|^2}{H} - \frac{\bar{R}ic(\nu,\nu)}{H}.$$

Due to the negative sign of the $|A|^2$-term we get from this equation by a simple application of the parabolic maximum principle the remarkable property that the mean curvature H is uniformly bounded in terms of its initial data and the Ricci curvature of the ambient manifold. This is in strong contrast to the mean curvature flow, where the blowup of the mean curvature causes the singularities studied in section 2. For the inverse mean curvature flow the critical behaviour occurs where $H \to 0$ and the speed becomes infinite. In Euclidean space it is clear that the maximum of the mean curvature is decreasing and the same is true for any L^p-norm.

In case $n = 2$ this property of the flow can be extended to closed surfaces in arbitrary three-manifolds of nonnegative scalar curvature: For any two-surface $\Sigma^2 \subset (N^3, \bar{g})$ the so called Hawking quasi-local mass of Σ^2 is defined as the geometric quantity

$$m_H(\Sigma^2) := \frac{|\Sigma^2|^{1/2}}{(16\pi)^{3/2}} \left(16\pi - \int_{\Sigma^2} H^2 \, d\mu\right),$$

and a computation based on the evolution equation for the mean curvature, the area element of the surface and the Gauss-Bonnet formula shows that for a solution M_t^2 of the inverse mean curvature flow

$$\frac{d}{dt} \int_{M_t^2} H^2 \, d\mu = 4\pi\chi(M_t^2) + \int_{M_t^2} -2\frac{|\nabla H|^2}{H^2} - \frac{1}{2}H^2 - \frac{1}{2}(\lambda_1 - \lambda_2)^2 - \bar{R} \, d\mu.$$

Hence, if the surface M_t^2 is connected and the scalar curvature $\bar{R}$ of the three-manifold is nonnegative, we have

$$\frac{d}{dt} \int_{M_t^2} H^2 \, d\mu \leq \frac{1}{2} \left(16\pi - \int_{M_t^2} H^2 \, d\mu\right)$$

and hence the Hawking quasi-local mass is nondecreasing along the inverse mean curvature flow:

$$\frac{d}{dt} m_H(M_t^2) \geq 0.$$

A major reason for the interest in the inverse mean curvature flow comes from the interpretation of this purely geometric fact in General Relativity: The spatial part of the exterior of an isolated gravitating system (like a star, black hole or galaxy) is modelled by the end of an asymptotically flat Riemannian 3–manifold with nonnegative scalar curvature as above. Here an end of a Riemannian 3-manifold $(N^3, \bar{g})$ is called *asymptotically flat* if it is realized by an open set that is diffeomorphic to the complement of a compact set K in $\mathbb{R}^3$, and the metric tensor $\bar{g}$ of M satisfies

$$|\bar{g}_{ij} - \delta_{ij}| \leq \frac{C}{|x|}, \qquad |\bar{g}_{ij,k}| \leq \frac{C}{|x|^2}, \qquad \bar{R}ic \geq -\frac{C\bar{g}}{|x|^2},$$

as $|x| \to \infty$. The derivatives are taken with respect to the Euclidean metric $\delta = \{\delta_{ij}\}$ on $\mathbb{R}^3 \setminus K$. On such asymptotically flat ends a concept of total mass or

energy is defined by a flux integral through the sphere at infinity,

$$m := \lim_{r\to\infty} \frac{1}{16\pi} \int_{\partial B^\delta_r(0)} (\bar{g}_{ii,j} - \bar{g}_{ij,i}) n^j \, d\mu_\delta,$$

which is a geometric invariant, despite being expressed in coordinates. It is finite precisely when the scalar curvature $\bar{R}$ of $\bar{g}$ satisfies

$$\int_{N^3} |\bar{R}| < \infty,$$

and from a physical point of view it is meant to measure both matter content and gravitational energy of the isolated system. Compare the joint papers [24][25] of the author and T. Ilmanen for references to these facts. The Hawking quasi-local mass defined above is used as a geometric concept for the energy of a three-dimensional region contained inside a two-dimensional surface, motivated by the fact that for large approximately round spheres S^2_R it is true that $m_H(S^2_R) \to m$. Furthermore, since in the physically simplest case the outer boundary of a black hole can be represented by a minimal two-surface inside the given three-manifold, the inverse mean curvature flow can provide a relation between the size of the black hole and the total energy m: If there is a smooth connected solution of the inverse mean curvature flow starting from a minimal surface $M^2_0 \subset N^3$, (the apparent horizon of the black hole) and expanding smoothly to large round spheres where $m_H(M^2_t) \to m$, then by the monotonicity result above we have the inequality

$$\frac{1}{4\sqrt{\pi}} |M^2_0|^{1/2} = m_H(M^2_0) \leq m.$$

This relation between the size of the outermost black hole and the total energy of an isolated gravitating system is the Riemannian Penrose inequality, which sharpens the positive mass theorem. The argument just described was first put forward by Geroch, [12].

The crucial question concerns of course the existence of such a solution to the flow by inverse mean curvature. For starshaped surfaces of positive mean curvature in $\mathbb{R}^{n+1}$ Gerhardt [11] and Urbas [31] show that the necessary estimates for complete regularity of the flow can be established and they prove longterm existence as well as asymptotic roundness in this class.

Without an assumption like starshapedness it is quite clear that singularities have to occur in certain situations. For example, the solution evolving from a thin symmetric torus can not exist forever, due to the upper bound on H some blowup in the speed H^{-1} must occur for such initial data. Similar examples can be constructed in the class of two-spheres making it clear that there cannot be a smooth solution for the flow in the general situations that are of natural interest in physics.

To overcome these difficulties [24] introduces a weak concept of solution for the flow which still retains the crucial monotonicity of the Hawking mass. The weak concept is a level-set formulation of (1.1), where the evolving surfaces are given as level-sets of a scalar function u via

$$M^2_t = \partial\{x | u(x) < t\},$$

and (1.1) is replaced by the degenerate elliptic equation

$$\mathrm{div}_N\left(\frac{\nabla u}{|\nabla u|}\right) = |\nabla u|,$$

where the left hand side describes the mean curvature of the level-sets and the right hand side yields the inverse speed. This formulation in divergence form admits locally Lipschitz continuous solutions and is inspired by the work of Evans-Spruck [9] and Chen-Giga-Goto [7] on the mean curvature flow. Using elliptic regularisation and a minimization principle we show existence of a locally Lipschitz-continuous solution with level-sets of nonnegative mean curvature of class $C^{1,\alpha}$, still satisfying monotonicity of the Hawking quasi-local mass, compare [24]. The solution allows the phenomenon of fattening, which corresponds to jumps of the surfaces and is desirable for our main application. We thus succeed in adapting Geroch's original argument and derive the following sharp lower bound for the mass:

THEOREM 4.1 *(H.-Ilmanen) Let N^3 be a complete, connected 3-manifold. Suppose that*

(i) *N^3 has nonnegative scalar curvature,*

(ii) *N^3 is asymptotically flat in the sense above with ADM mass m,*

(iii) *The boundary of N^3 is compact and consists of minimal surfaces, and N^3 contains no other compact minimal surfaces.*

Then $m \geq 0$, and

$$16\pi m^2 \geq |\Sigma^2|,$$

where $|\Sigma^2|$ is the area of any connected component of ∂N^3. Equality holds if and only if N^3 is one-half of the spatial Schwarzschild manifold.

The *spatial Schwarzschild manifold* is the manifold $\mathbb{R}^3 \setminus \{0\}$ equipped with the metric $\bar{g} := (1 + m/2|x|)^4\delta$, representing the spatial exterior region of a single static black hole of mass m.

REFERENCES

[1] B. Andrews, *Contraction of convex hypersurfaces in Euclidean space*, Calc. Var. 2 (1994), 151–171.

[2] B. Andrews, *Gauss curvature flow: The fate of the rolling stones*, preprint ANU Canberra (1998), pp10.

[3] B. Andrews, *Contraction of convex hypersurfaces by their affine normal*, J. Diff. Geom. 43 (1996), 207–230.

[4] B. Andrews, *Contraction of convex hypersurfaces in Riemannian spaces*, J. Diff. Geom. 39 (1994), 407–431.

[5] B. Andrews, *Monotone quantities and unique limits for evolving convex hypersurfaces*, IMRN 20 (1997), 1001–1031.

[6] S.B. Angenent, J.J.L. Velazques, *Degenerate neckpinches in mean curvature flow*, J. Reine Angew. Math. 482 (1997), 15–66.

[7] Y.G. Chen, Y. Giga, and S. Goto, *Uniqueness and Existence of Viscosity Solutions of Generalized Mean Curvature Flow Equations*, J Diff.Geom. 33 (1991), 749–786.

[8] B. Chow, *Deforming convex hypersurfaces by the nth root of the Guassian curvature*, J.Diff.Geom. 23 (1985), 117–138.

[9] L. C. Evans, J. Spruck, *Motion of Level Sets by Mean Curvature I*, J. Diff.Geom. 33 (1991), 635–681.

[10] W.J. Firey, *Shapes of worn stones*, Mathematica 21 (1974), 1–11.

[11] C. Gerhardt, *Flow of nonconvex hypersurfaces into spheres*, J.Diff.Geom. 32 (1990), 299–314.

[12] R. Geroch, *Energy Extraction*, Ann. New York Acad. Sci. 224 (1973), 108–17.

[13] M. Grayson, *The heat equation shrinks embedded plane curves to points*, J. Diff. Geom. 26 (1987), 285–314.

[14] M. Grayson, *Shortening embedded curves*, Annals Math. 129 (1989), 71–111.

[15] R. S. Hamilton, *The formation of singularities in the Ricci Flow*, Surveys in Differential Geometry Vol. II, International Press, Cambridge MA (1993), 7–136.

[16] R. S. Hamilton, *Harnack estimate for the mean curvature flow*, J. Diff. Geom. 41 (1995), 215–226.

[17] R. S. Hamilton, *Four manifolds with positive isotropic curvature*, Comm. Anal. Geom. 5 (1997), 1–92.

[18] R. S. Hamilton, *Isoperimetric estimates for the curve shrinking flow in the plane*, Modern Methods in Compl. Anal., Princeton Univ.Press (1992), 201-222.

[19] G. Huisken, *Flow by mean curvature of convex surfaces into spheres*, J. Diff. Geometry 20 (1984), 237–266.

[20] G. Huisken, *Contracting convex hypersurfaces in Riemannian manifolds by their mean curvature*, Invent. Math. 84 (1986), 463–480.

[21] G. Huisken, *Asymptotic behaviour for singularities of the mean curvature flow*, J. Diff. Geometry 31 (1990), 285–299.

[22] G. Huisken, *Local and global behaviour of hypersurfaces moving by mean curvature*, Proceedings of Symposia in Pure Mathematics 54 (1993), 175–191.

[23] G. Huisken, *A distance comparison principle for evolving curves*, Asian J. Math. 2 (1998), 127–134.

[24] G. Huisken, T. Ilmanen, *The inverse mean curvature flow and the Riemannian Penrose inequality*, preprint http://poincare.mathematik.uni-tuebingen.de, to appear.

[25] G. Huisken, T. Ilmanen, *The Riemannian Penrose inequality*, IMRN 20 (1997), 1045–1058.

[26] G. Huisken, C. Sinestrari, *Mean curvature flow singularities for mean convex surfaces*, Calc. Variations, to appear.

[27] G. Huisken, C. Sinestrari, *Convexity estimates for mean curvature flow and singularities for mean convex surfaces*, preprint, to appear.

[28] G. Sapiro, A. Tannenbaum, *On affine plane curve evolution*, J. Funct. Anal. 119 (1994), 79–120.

[29] K. Smoczyk, *Starshaped hypersurfaces and the mean curvature flow*, Preprint (1997), 13pp.

[30] K. Tso, *Deforming a hypersurface by its Gauss–Kronecker curvature*, Comm. Pure Appl. Math. 38 (1985), 867–882.

[31] J. Urbas, *On the Expansion of Starshaped Hypersurfaces by Symmetric Functions of Their Principal Curvatures*, Math. Z. 205 (1990), 355–372.

[32] B. White, *Partial Regularity of Mean Convex Hypersurfaces Flowing by Mean Curvature*, IMRN 4 (1994), 185–192.

Gerhard Huisken
Mathematisches Institut
Universität Tübingen
Auf der Morgenstelle 10
D-72076 Tübingen
Germany
gerhard.huisken@uni-tuebingen.de

Compact Manifolds with Exceptional Holonomy

Dominic Joyce

Abstract. In the classification of Riemannian holonomy groups, the *exceptional holonomy groups* are G_2 in 7 dimensions, and $Spin(7)$ in 8 dimensions. We outline the construction of the first known examples of compact 7- and 8-manifolds with holonomy G_2 and $Spin(7)$.

In the case of G_2, we first choose a finite group Γ of automorphisms of the torus T^7 and a flat Γ-invariant G_2-structure on T^7, so that T^7/Γ is an *orbifold*. Then we resolve the singularities of T^7/Γ to get a compact 7-manifold M. Finally we use analysis, and an understanding of Calabi-Yau metrics, to construct a family of metrics with holonomy G_2 on M, which converge to the singular metric on T^7/Γ.

1991 Mathematics Subject Classification: 53C15, 53C25, 53C80, 58G30.
Keywords and Phrases: exceptional holonomy, G_2, $Spin(7)$, Ricci-flat.

In the theory of Riemannian holonomy groups, perhaps the most mysterious are the two exceptional cases, the holonomy group G_2 in 7 dimensions and the holonomy group $Spin(7)$ in 8 dimensions. We shall describe the construction of the first known examples of *compact* 7-manifolds with holonomy G_2. There is a very similar construction for compact 8-manifolds with holonomy $Spin(7)$, which we will not discuss because of lack of space. All the details can be found in the author's papers [5], [6], [7] and the forthcoming book [8]. A good reference on Riemannian holonomy groups, and G_2 and $Spin(7)$ in particular, is the book by Salamon [13].

1 Riemannian holonomy groups

Let M be a connected n-dimensional manifold, let g be a Riemannian metric on M, and let ∇ be the Levi-Civita connection of g. Let x, y be points in M joined by a smooth path γ. Then *parallel transport* along γ using ∇ defines an isometry between the tangent spaces T_xM, T_yM at x and y.

Definition 1.1 The *holonomy group* $\mathrm{Hol}(g)$ of g is the group of isometries of T_xM generated by parallel transport around closed loops based at x in M. We consider $\mathrm{Hol}(g)$ to be a subgroup of $O(n)$, defined up to conjugation by elements of $O(n)$. Then $\mathrm{Hol}(g)$ is independent of the base point x in M.

The classification of holonomy groups was achieved by Berger [1] in 1955.

THEOREM 1.2 *Let M be a simply-connected, n-dimensional manifold, and g an irreducible, nonsymmetric Riemannian metric on M. Then either*

(*i*) $\mathrm{Hol}(g) = SO(n)$,
(*ii*) $n = 2m$ *and* $\mathrm{Hol}(g) = SU(m)$ *or* $U(m)$,
(*iii*) $n = 4m$ *and* $\mathrm{Hol}(g) = Sp(m)$ *or* $Sp(m)Sp(1)$,
(*iv*) $n = 7$ *and* $\mathrm{Hol}(g) = G_2$, *or*
(*v*) $n = 8$ *and* $\mathrm{Hol}(g) = Spin(7)$.

Now G_2 and $Spin(7)$ are the exceptional cases in this classification, so they are called the *exceptional holonomy groups*. For some time after Berger's classification, the exceptional holonomy groups remained a mystery. In 1987, Bryant [2] used the theory of exterior differential systems to show that locally there exist many metrics with these holonomy groups, and gave some explicit, incomplete examples. Then in 1989, Bryant and Salamon [3] found explicit, *complete* metrics with holonomy G_2 and $Spin(7)$ on noncompact manifolds. In 1994-5 the author constructed examples of metrics with holonomy G_2 and $Spin(7)$ on *compact* manifolds [5, 6, 7, 8], and these are the subject of this article.

We now introduce the holonomy group G_2. Let $(x_1, \ldots, x_7)$ be coordinates on $\mathbb{R}^7$. Define a metric g_0 and a 3-form φ_0 on $\mathbb{R}^7$ by

$$g_0 = dx_1^2 + \cdots + dx_7^2, \tag{1}$$

$$\begin{aligned}\varphi_0 =\, & dx_1 \wedge dx_2 \wedge dx_7 + dx_1 \wedge dx_3 \wedge dx_6 + dx_1 \wedge dx_4 \wedge dx_5 + dx_2 \wedge dx_3 \wedge dx_5 \\ & - dx_2 \wedge dx_4 \wedge dx_6 + dx_3 \wedge dx_4 \wedge dx_7 + dx_5 \wedge dx_6 \wedge dx_7.\end{aligned} \tag{2}$$

The subgroup of $GL(7, \mathbb{R})$ preserving φ_0 is the *exceptional Lie group* G_2. This group also preserves g_0 and the orientation on $\mathbb{R}^7$. It is a compact, semisimple, 14-dimensional Lie group, a subgroup of $SO(7)$.

A *G_2-structure* on a 7-manifold M is a principal subbundle of the frame bundle of M, with structure group G_2. Each G_2-structure gives rise to a 3-form φ and a metric g on M, such that every tangent space of M admits an isomorphism with $\mathbb{R}^7$ identifying φ and g with φ_0 and g_0 respectively. By an abuse of notation, we will refer to (φ, g) as a G_2-structure.

PROPOSITION 1.3 *Let M be a 7-manifold and (φ, g) a G_2-structure on M. Then the following are equivalent:*

(*i*) $\mathrm{Hol}(g) \subseteq G_2$, *and φ is the induced 3-form,*
(*ii*) $\nabla\varphi = 0$ *on M, where ∇ is the Levi-Civita connection of g, and*
(*iii*) $d\varphi = d^*\varphi = 0$ *on M.*

We call $\nabla\varphi$ the *torsion* of the G_2-structure (φ, g), and when $\nabla\varphi = 0$ the G_2-structure is *torsion-free*. If (φ, g) is torsion-free, then g is Ricci-flat.

PROPOSITION 1.4 *Let M be a compact 7-manifold, and suppose that (φ, g) is a torsion-free G_2-structure on M. Then* $\mathrm{Hol}(g) = G_2$ *if and only if $\pi_1(M)$ is finite. In this case the moduli space of metrics with holonomy G_2 on M, up to diffeomorphisms isotopic to the identity, is a smooth manifold of dimension $b^3(M)$.*

2 A 'Kummer construction' for a 7-manifold

It is well known that metrics with holonomy $SU(2)$ on the $K3$ surface can be obtained by resolving the 16 singularities of the orbifold $T^4/\mathbb{Z}_2$, where $\mathbb{Z}_2$ acts on T^4 with 16 fixed points. This is called the *Kummer construction*. Our construction is motivated by and modelled on this. It can be divided into four steps. Here is a summary of each. For simplicity we will describe the G_2 case only, but the $Spin(7)$ case is very similar.

Step 1. Let T^7 be the 7-torus. Let (φ_0, g_0) be a flat G_2-structure on T^7. Choose a finite group Γ of isometries of T^7 preserving (φ_0, g_0). Then the quotient T^7/Γ is a singular, compact 7-manifold.

Step 2. For certain special groups Γ there is a method to resolve the singularities of T^7/Γ in a natural way, using complex geometry. We get a non-singular, compact 7-manifold M, together with a map $\pi : M \to T^7/\Gamma$, the resolving map.

Step 3. On M, we explicitly write down a 1-parameter family of G_2-structures (φ_t, g_t) depending on a real variable $t \in (0, \epsilon)$. These G_2-structures are not torsion-free, but when t is small, they have small torsion. As $t \to 0$, the G_2-structure (φ_t, g_t) converges to the singular G_2-structure $\pi^*(\varphi_0, g_0)$.

Step 4. We prove using analysis that for all sufficiently small t, the G_2-structure (φ_t, g_t) on M, with small torsion, can be deformed to a G_2-structure $(\tilde{\varphi}_t, \tilde{g}_t)$, with zero torsion. Finally, we show that $\tilde{g}_t$ is a metric with holonomy G_2 on the compact 7-manifold M.

We will now explain the steps in greater detail.

Step 1

Here is an example of a suitable group Γ. Let $(x_1, \ldots, x_7)$ be coordinates on $T^7 = \mathbb{R}^7/\mathbb{Z}^7$, where $x_i \in \mathbb{R}/\mathbb{Z}$. Let (φ_0, g_0) be the flat G_2-structure on T^7 defined by (2). Let α, β and γ be the involutions of T^7 defined by

$$\alpha\big((x_1, \ldots, x_7)\big) = \big(-x_1, -x_2, -x_3, -x_4, x_5, x_6, x_7\big), \tag{3}$$

$$\beta\big((x_1, \ldots, x_7)\big) = \big(-x_1, \tfrac{1}{2} - x_2, x_3, x_4, -x_5, -x_6, x_7\big), \tag{4}$$

$$\gamma\big((x_1, \ldots, x_7)\big) = \big(\tfrac{1}{2} - x_1, x_2, \tfrac{1}{2} - x_3, x_4, -x_5, x_6, -x_7\big). \tag{5}$$

By inspection, α, β and γ preserve (φ_0, g_0), because of the careful choice of exactly which signs to change. Also, $\alpha^2 = \beta^2 = \gamma^2 = 1$, and α, β and γ commute. Thus they generate a group $\Gamma = \langle \alpha, \beta, \gamma \rangle \cong \mathbb{Z}_2^3$ of isometries of T^7 preserving the flat G_2-structure (φ_0, g_0).

Lemma 2.1 *The elements $\beta\gamma$, $\gamma\alpha$, $\alpha\beta$ and $\alpha\beta\gamma$ of Γ have no fixed points on T^7. The fixed points of α, β, γ are each 16 copies of T^3. The singular set S of T^7/Γ is a disjoint union of 12 copies of T^3, 4 copies from each of α, β, γ. Each component of S is a singularity modelled on that of $T^3 \times \mathbb{C}^2/\{\pm 1\}$.*

Thus the singular set splits into a disjoint union of connected components, and each component is very simple. This is helpful because we can desingularize each connected component independently, and simple singularities are easier to resolve.

STEP 2

Our goal is to resolve the singular set S of T^7/Γ to get a compact 7-manifold M with holonomy G_2. How can we do this? In general we cannot, because we have no idea of how to resolve general orbifold singularities with holonomy G_2. However, suppose we can arrange that every connected component of S is locally isomorphic to either

(*a*) $T^3 \times \mathbb{C}^2/G$, for G a finite subgroup of $SU(2)$, or
(*b*) $\mathcal{S}^1 \times \mathbb{C}^3/G$, for G a finite subgroup of $SU(3)$ acting freely on $\mathbb{C}^3 \setminus 0$.

In this case we can use *complex algebraic geometry* to find a natural resolution X of $\mathbb{C}^2/G$ or Y of $\mathbb{C}^3/G$, and then $T^3 \times X$ or $\mathcal{S}^1 \times Y$ gives a local model for how to resolve the corresponding component of S in T^7/Γ.

In case (a), X must have a Kähler metric h with holonomy $SU(2)$ that is asymptotic to the flat Euclidean metric on $\mathbb{C}^2/G$. Such metrics are called *Asymptotically Locally Euclidean* (ALE). They have been classified by Kronheimer [10, 11], and they exist for every finite subgroup $G \subset SU(2)$. The point is that if X has holonomy $SU(2)$, then the product 7-manifold $T^3 \times X$ has holonomy $\{1\} \times SU(2)$. But $\{1\} \times SU(2)$ is a subgroup of G_2, and so $T^3 \times X$ has a torsion-free G_2-structure by Proposition 1.3. Hence, $T^3 \times X$ gives a local model for how to resolve the singularity $T^3 \times \mathbb{C}^2/G$ with holonomy G_2.

In case (b), Y is a *crepant resolution* of $\mathbb{C}^3/G$, and carries an ALE Kähler metric h with holonomy $SU(3)$. Such resolutions and metrics exist for all finite $G \subset SU(3)$, by work of Roan [12] and the author [8]. Since $\{1\} \times SU(3) \subset G_2$, if (Y, h) has holonomy $SU(3)$ then $\mathcal{S}^1 \times Y$ has a torsion-free G_2-structure, and provides a local model for how to resolve the singularity $\mathcal{S}^1 \times \mathbb{C}^3/G$ with holonomy G_2.

Suppose that all the singularities of T^7/Γ are of type (a) or (b). Then we can construct a compact, nonsingular 7-manifold M by resolving each singularity $T^3 \times \mathbb{C}^2/G$ using $T^3 \times X$, and resolving each singularity $\mathcal{S}^1 \times \mathbb{C}^3/G$ using $\mathcal{S}^1 \times Y$, as above. In the example this means gluing 12 copies of $T^3 \times X$ into T^7/Γ, where X is the blow-up of $\mathbb{C}^2/\{\pm 1\}$ at its singular point.

STEP 3

For each resolution X of $\mathbb{C}^2/G$ in case (a), and Y of $\mathbb{C}^3/G$ in case (b), we can find a 1-parameter family $\{h_t : t > 0\}$ of metrics with the properties

(*a*) h_t is a Kähler metric on X with $\mathrm{Hol}(h_t) = SU(2)$. Its injectivity radius satisfies $\delta(h_t) = O(t)$, its Riemann curvature satisfies $\big\|R(h_t)\big\|_{C^0} = O(t^{-2})$, and $h_t = h + O(t^4 r^{-4})$ for large r, where h is the Euclidean metric on $\mathbb{C}^2/G$, and r the distance from the origin.

(*b*) h_t is Kähler on Y with $\mathrm{Hol}(h_t) = SU(3)$, satisfying $\delta(h_t) = O(t)$, $\big\|R(h_t)\big\|_{C^0} = O(t^{-2})$, and $h_t = h + O(t^6 r^{-6})$ for large r.

In fact we can choose h_t to be isometric to $t^2 h_1$, and the properties above are easy to prove.

Suppose one of the components of the singular set S of T^7/Γ is locally modelled on $T^3 \times \mathbb{C}^2/G$. Then T^3 has a natural flat metric h_{T^3}. Let X be the resolution of $\mathbb{C}^2/G$ and let $\{h_t : t > 0\}$ satisfy property (*a*). Then $\hat{g}_t = h_{T^3} + h_t$ is a metric on $T^3 \times X$ with holonomy $\{1\} \times SU(2)$, which is contained in G_2. Thus there is an associated torsion-free G_2-structure $(\hat{\varphi}_t, \hat{g}_t)$ on $T^3 \times X$. Similarly, if a component of S is modelled on $\mathcal{S}^1 \times \mathbb{C}^3/G$, we get a family of torsion-free G_2-structures $(\hat{\varphi}_t, \hat{g}_t)$ on $\mathcal{S}^1 \times Y$.

The idea is to make a G_2-structure (φ_t, g_t) on M by gluing together the torsion-free G_2-structures $(\hat{\varphi}_t, \hat{g}_t)$ on the patches $T^3 \times X$ and $\mathcal{S}^1 \times Y$, and (φ_0, g_0) on T^7/Γ. The gluing is done using a partition of unity. Naturally, the first derivative of the partition of unity introduces 'errors', so that (φ_t, g_t) is not torsion-free. The size of the torsion $\nabla\varphi_t$ depends on the difference $\hat{\varphi}_t - \varphi_0$ in the region where the partition of unity changes. On the patches $T^3 \times X$, since $h_t - h = O(t^4 r^{-4})$ and the partition of unity has nonzero derivative when $r = O(1)$, we find that $\nabla\varphi_t = O(t^4)$. Similarly $\nabla\varphi_t = O(t^6)$ on the patches $\mathcal{S}^1 \times Y$, and so $\nabla\varphi_t = O(t^4)$ on M.

For small t, the dominant contributions to the injectivity radius $\delta(g_t)$ and Riemann curvature $R(g_t)$ are made by those of the metrics h_t on X and Y, so we expect $\delta(g_t) = O(t)$ and $\big\|R(g_t)\big\|_{C^0} = O(t^{-2})$ by properties (*a*) and (*b*) above. In this way we prove the following result, which gives the estimates on (φ_t, g_t) that we need.

Theorem A *On the compact 7-manifold M described above, and on many other 7-manifolds constructed in a similar fashion, one can write down the following data explicitly in coordinates:*

- *Positive constants A_1, A_2, A_3 and ϵ,*
- *A G_2-structure (φ_t, g_t) on M with $d\varphi_t = 0$ for each $t \in (0, \epsilon)$, and*
- *A 3-form ψ_t on M with $d^*\psi_t = d^*\varphi_t$ for each $t \in (0, \epsilon)$.*

These satisfy three conditions:

(*i*) $\|\psi_t\|_{L^2} \leq A_1 t^4$ *and* $\|d^*\psi_t\|_{L^{14}} \leq A_1 t^4$,
(*ii*) *the injectivity radius $\delta(g_t)$ satisfies $\delta(g_t) \geq A_2 t$,*
(*iii*) *the Riemann curvature $R(g_t)$ of g_t satisfies $\big\|R(g_t)\big\|_{C^0} \leq A_3 t^{-2}$.*

Here the operator d^* and the norms $\|\,.\,\|_{L^2}$, $\|\,.\,\|_{L^{14}}$ and $\|\,.\,\|_{C^0}$ depend on g_t.

Here one should regard ψ_t as a *first integral* of the torsion $\nabla\varphi_t$ of (φ_t, g_t). Thus the norms $\|\psi_t\|_{L^2} \leq A_1 t^4$ and $\|d^*\psi_t\|_{L^{14}} \leq A_1 t^4$ are measures of $\nabla\varphi_t$. So parts (*i*)-(*iii*) say that the torsion $\nabla\varphi_t$ must be small compared to the injectivity radius and Riemann curvature of (M, g_t).

STEP 4

We prove the following analysis result.

THEOREM B *In the situation of Theorem A there are constants $\kappa, K > 0$ depending only on A_1, A_2, A_3 and ϵ, such that for each $t \in (0, \kappa]$ there exists a smooth, torsion-free G_2-structure $(\tilde{\varphi}_t, \tilde{g}_t)$ on M with $\|\tilde{\varphi}_t - \varphi_t\|_{C^0} \leq Kt^{1/2}$.*

Basically, this result says that if (φ, g) is a G_2-structure on M, and the torsion $\nabla\varphi$ is sufficiently small, then we can deform to a nearby G_2-structure $(\tilde{\varphi}, \tilde{g})$ that is torsion-free. Here is a sketch of the proof of Theorem B, ignoring several technical points. The proof is that given in [8], which is an improved version of the proof in [5]. For simplicity we omit the subscripts t.

We have a 3-form φ with $d\varphi = 0$ and $d^*\varphi = d^*\psi$ for small ψ, and we wish to construct a nearby 3-form $\tilde{\varphi}$ with $d\tilde{\varphi} = 0$ and $\tilde{d}^*\tilde{\varphi} = 0$. Set $\tilde{\varphi} = \varphi + d\eta$, where η is a small 2-form. Then η must satisfy a nonlinear p.d.e., which we write as

$$d^*d\eta = -d^*\psi + d^*F(d\eta), \tag{6}$$

where F is nonlinear, satisfying $F(d\eta) = O\bigl(|d\eta|^2\bigr)$.

We solve (6) by iteration, introducing a sequence $\{\eta_j\}_{j=0}^{\infty}$ with $\eta_0 = 0$, satisfying the inductive equations

$$d^*d\eta_{j+1} = -d^*\psi + d^*F(d\eta_j), \qquad d^*\eta_{j+1} = 0. \tag{7}$$

If such a sequence exists and converges to η, then taking the limit in (7) shows that η satisfies (6), giving us the solution we want.

The key to proving this is an *inductive estimate* on the sequence $\{\eta_j\}_{j=0}^{\infty}$. The inductive estimate we use has three ingredients, the equations

$$\|d\eta_{j+1}\|_{L^2} \leq \|\psi\|_{L^2} + C_1\|d\eta_j\|_{L^2}\|d\eta_j\|_{C^0}, \tag{8}$$

$$\|\nabla d\eta_{j+1}\|_{L^{14}} \leq C_2\bigl(\|d^*\psi\|_{L^{14}} + \|\nabla d\eta_j\|_{L^{14}}\|d\eta_j\|_{C^0} + t^{-4}\|d\eta_{j+1}\|_{L^2}\bigr), \tag{9}$$

$$\|d\eta_j\|_{C^0} \leq C_3\bigl(t^{1/2}\|\nabla d\eta_j\|_{L^{14}} + t^{-7/2}\|d\eta_j\|_{L^2}\bigr). \tag{10}$$

Here C_1, C_2, C_3 are positive constants independent of t. Equation (8) is obtained from (7) by taking the L^2-inner product with η_{j+1} and integrating by parts. Using the fact that $d^*\varphi = d^*\psi$ and ψ is $O(t^4)$, we get a powerful a priori estimate of the L^2-norm of $d\eta_{j+1}$.

Equation (9) is derived from an *elliptic regularity estimate* for the operator $d + d^*$ acting on 3-forms on M. Equation (10) follows from the *Sobolev embedding theorem*, since $L^{14}_1(M)$ embeds in $C^0(M)$. Both (9) and (10) are proved on small balls of radius $O(t)$ in M, using parts (*ii*) and (*iii*) of Theorem A, and this is where the powers of t come from.

Using (8)-(10) and part (*i*) of Theorem A we show that if

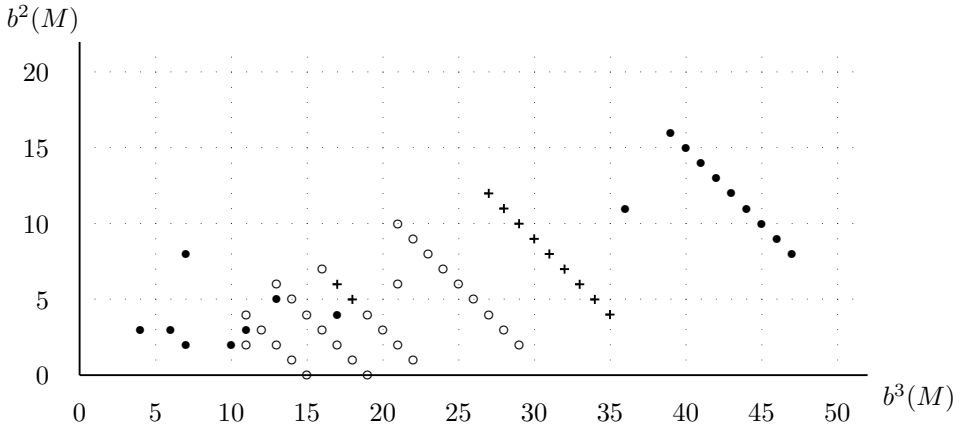
$$\|d\eta_j\|_{L^2} \leq C_4t^4, \quad \|\nabla d\eta_j\|_{L^{14}} \leq C_5, \quad \text{and} \quad \|d\eta_j\|_{C^0} \leq Kt^{1/2}, \tag{11}$$

where C_4, C_5 and K are positive constants depending on C_1, C_2, C_3 and A_1, and if t is sufficiently small, then the same inequalities (11) apply to $d\eta_{j+1}$. Since $\eta_0 = 0$,

by induction (11) applies for all j and the sequence $\{d\eta_j\}_{j=0}^{\infty}$ is bounded in the Banach space $L_1^{14}(\Lambda^3 T^*M)$. One can then use standard techniques in analysis to prove that this sequence converges to a smooth limit $d\eta$. This concludes the sketch proof of Theorem B.

From Theorems A and B we see that the compact 7-manifold M constructed in Step 2 admits torsion-free G_2-structures $(\tilde{\varphi}, \tilde{g})$. Proposition 1.4 then shows that $\mathrm{Hol}(\tilde{g}) = G_2$ if and only if $\pi_1(M)$ is finite. In the example above M is simply-connected, and so $\pi_1(M) = \{1\}$ and M has metrics with holonomy G_2, as we want.

By considering different groups Γ acting on T^7, and also by finding topologically distinct resolutions $M_1, \ldots, M_k$ of the same orbifold T^7/Γ, we can construct many compact Riemannian 7-manifolds with holonomy G_2. Here is a graph of the Betti numbers $b^2(M)$ and $b^3(M)$ of the 68 examples found in [5, 6]. More examples will be given in [8].



Betti numbers of known compact 7-manifolds with holonomy G_2

On this graph the symbol '•' denotes the Betti numbers of a simply-connected 7-manifold, '∘' denotes a non-simply-connected manifold, and '+' denotes both a simply-connected and a non-simply-connected manifold.

So far we have discussed only the holonomy group G_2. There is a very similar construction for compact manifolds with holonomy $Spin(7)$, described in [7] and [8]. Here are some of the similarities and differences in the two cases. The holonomy group $Spin(7)$ is a subgroup of $SO(8)$, a compact 21-dimensional Lie group isomorphic to the double cover of $SO(7)$. It is the subgroup of $GL(8, \mathbb{R})$ preserving a certain 4-form Ω_0 on $\mathbb{R}^8$, and also preserves the Euclidean metric g_0 on $\mathbb{R}^8$.

Thus a $Spin(7)$-structure on an 8-manifold M is equivalent to a pair (Ω, g), where Ω is a 4-form and g a Riemann metric that are pointwise isomorphic to Ω_0 and g_0. Riemannian manifolds with holonomy $Spin(7)$ are Ricci-flat. Compact manifolds M with holonomy $Spin(7)$ are simply-connected spin manifolds, and

computing the index of the Dirac operator shows that their Betti numbers must satisfy $b^3(M) + b^4_+(M) = b^2(M) + b^4_-(M) + 25$.

We can construct compact 8-manifolds with holonomy $Spin(7)$ by resolving the singularities of orbifolds T^8/Γ. The construction is more difficult than the G_2 case in two ways. Firstly, it seems to be more difficult to find suitable orbifolds T^8/Γ, and it is necessary to consider more complicated kinds of orbifold singularities. Secondly, the analysis is more difficult, and one has to try harder to make the sequence converge. In [7] we find at least 95 topologically distinct compact 8-manifolds with holonomy $Spin(7)$, realizing 29 distinct sets of Betti numbers, and [8] will give more examples.

Note that compact manifolds with holonomy G_2 and $Spin(7)$ are examples of compact Ricci-flat Riemannian manifolds. In fact, compact manifolds with holonomy G_2 are the only known source of odd-dimensional examples of compact, simply-connected Ricci-flat Riemannian manifolds.

3 Directions for future research

Here are four areas in which I hope to see interesting developments soon.

- *Other constructions of compact manifolds with exceptional holonomy.* The author has extended the constructions of [5]-[7] to include resolutions of more general quotient singularities, in particular *non-isolated* quotient singularities $\mathbb{C}^m/G$ for G a finite subgroup of $SU(m)$ and $m = 3$ or 4, and the results will be published in [8]. Another promising possibility is to try to replace the orbifold T^7/Γ by $(\mathcal{S}^1 \times W)/\Gamma$, where W is a Calabi-Yau 3-fold.

- Harvey and Lawson's theory of *calibrated geometry* [4] singles out three classes of special submanifolds in manifolds of exceptional holonomy: *associative 3-folds* and *coassociative 4-folds* in G_2-manifolds, and *Cayley 4-folds* in $Spin(7)$-manifolds. They are minimal submanifolds, and have good properties under deformation. Compact examples can be constructed as the fixed point sets of isometries, as in [6].

  It would be interesting to study families of compact manifolds of these types, to understand the way singularities develop in such families, and whether a compact G_2 or $Spin(7)$-manifold can be fibred by coassociative or Cayley 4-manifolds, with some singular fibres.

- *Gauge theory on compact $Spin(7)$-manifolds.* Let M be a compact 8-manifold with holonomy $Spin(7)$, let E be a vector bundle or principal bundle over M, and let A be a connection on E. Then the curvature F_A of A is a 2-form with values in $\mathrm{ad}(E)$. Now the $Spin(7)$-structure induces a splitting $\Lambda^2 T^*M = \Lambda^2_7 \oplus \Lambda^2_{21}$, where $\Lambda^2_7, \Lambda^2_{21}$ are vector bundles over M with fibre $\mathbb{R}^7$, $\mathbb{R}^{21}$ respectively. We call A a $Spin(7)$-*instanton* if the component of F_A in $\mathrm{ad}(E) \otimes \Lambda^2_7$ is zero.

  It turns out that $Spin(7)$-instantons have many properties in common with instantons in 4 dimensions, that are studied in Donaldson theory. Christo-

pher Lewis and the author [9] have proved an existence theorem for $Spin(7)$-instantons with gauge group $SU(2)$ on certain compact 8-manifolds with holonomy $Spin(7)$. In 4 dimensions a sequence of instantons can 'bubble' at a finite number of points. In 8 dimensions we expect 'bubbling' to occur instead around a compact Cayley 4-manifold, and we construct families of instantons in which this happens.

- *Connections with String Theory.* String Theory is a branch of high-energy theoretical physics that aims to unify quantum theory and gravity by modelling particles as 1-dimensional objects called *strings*. One of its features is that it prescribes the dimension of space-time. This depends on the details of the theory, but the most popular model, *supersymmetric string theory*, gives dimension 10. To explain the discrepancy between this and the 4 space-time dimensions that we observe, it is supposed that the universe looks locally like $\mathbb{R}^4 \times M^6$, where M^6 is a compact 6-manifold with very small radius, of order 10^{-33}cm.

  In supersymmetric string theory, M must be a *Calabi-Yau 3-fold*. So string theorists are interested in Calabi-Yau 3-folds, and have contributed many ideas to the subject, including that of *Mirror Symmetry*. However, if instead we consider $\mathbb{R}^3 \times M^7$, corresponding to an observable universe with 3 space-time dimensions, then by work of Vafa and Shatashvili M^7 must be a compact 7-manifold with holonomy G_2. Similarly, if we consider $\mathbb{R}^2 \times M^8$, so that the observable universe has 2 space-time dimensions, then M^8 is a compact 8-manifold with holonomy $Spin(7)$.

  Recently, string theorists have begun to seriously consider the possibility that the universe may have 11 dimensions ('M theory') or even 12 dimensions ('F theory'). To reduce to 4 observable space-time dimensions in these theories will require a manifold of dimension 7 or 8, and it seems likely that compact manifolds with exceptional holonomy will play a rôle in this.

References

[1] M. Berger, 'Sur les groupes d'holonomie homogène des variétés à connexion affines et des variétés Riemanniennes', Bull. Soc. Math. France 83 (1955) 279-330.

[2] R.L. Bryant, 'Metrics with exceptional holonomy', Ann. Math. 126 (1987) 525-576.

[3] R.L. Bryant & S.M. Salamon, 'On the construction of some complete metrics with exceptional holonomy', Duke Math. J. 58 (1989) 829-850.

[4] R. Harvey & H.B. Lawson, 'Calibrated geometries', Acta Math. 148 (1982), 47-157.

[5] D.D. Joyce, 'Compact Riemannian 7-manifolds with holonomy G_2. I', J. Diff. Geom. 43 (1996), 291-328.

[6] D.D. Joyce, 'Compact Riemannian 7-manifolds with holonomy G_2. II', J. Diff. Geom. 43 (1996), 329-375.

[7] D.D. Joyce, 'Compact Riemannian 8-manifolds with holonomy $Spin(7)$', Inventiones mathematicae 123 (1996), 507-552.

[8] D.D. Joyce, 'Compact Riemannian manifolds with special holonomy groups', in preparation, 1998.

[9] C. Lewis, '$Spin(7)$ instantons', Oxford University D.Phil. thesis, in preparation, 1998.

[10] P.B. Kronheimer, 'The construction of ALE spaces as hyperkähler quotients', J. Diff. Geom. 29 (1989), 665-683.

[11] P.B. Kronheimer, 'A Torelli-type theorem for gravitational instantons', J. Diff. Geom. 29 (1989), 685-697.

[12] S.-S. Roan, 'Minimal resolution of Gorenstein orbifolds', Topology 35 (1996), 489-508.

[13] S.M. Salamon, 'Riemannian geometry and holonomy groups', Pitman Res. Notes in Math. 201, Longman, Harlow, 1989.

Dominic Joyce
Lincoln College
Oxford
OX1 3DR
England

Large Groups Actions on Manifolds

François Labourie

Abstract. We shall survey some results concerning large groups actions on manifolds, with an emphasis on rigidity and geometric questions. By large groups actions, we, in short, mean actions of non free groups, with at least a dense orbit. Most of the results will concern lattices, but we shall present results and questions concerning other large groups.

Introduction

In this article, we shall be interested in large group actions on manifolds. Large actions will mean highly non proper actions such as topologically transitive (*i.e.* with a dense orbit), or volume preserving ergodic ones (*i.e.* every invariant subset is either of full or zero volume). Large group is a rather unprecise notion, of which we do not have a definition but examples. They are at least required to be finitely generated non free groups. An important and well studied class since R. Zimmer work [Z] is that of *higher rank lattices* : a *lattice* Γ in a unimodular real Lie group G is a discrete subgroup such that G/Γ has finite volume; it is *cocompact* if G/Γ is compact; using a restricted definition for the sake of simplicity, by *higher rank* we mean that G is simple of real rank greater than 2. A good class of examples of higher rank lattices is $SL(n,\mathbb{Z})$, with $n \geq 3$. However, we shall try not to restrict ourselves to this class and to present results and questions concerning other groups.

Obviously, due to the presence of relations among the elements of our groups, large group actions should be rare and difficult to construct. In particular, given an action of a large group Γ on a manifold M, one would like to answer the local rigidity question, whose answer turns out to be positive in many case.

Local Rigidity Question: *is any smooth action of Γ on M close enough to the original one, conjugate to it within the group of diffeomorphisms ?*

In the case of higher rank lattices, people are even more optimistic. The general belief, supported by Margulis-Zimmer superrigidity, is that every large action is essentially geometric because, in some sense, the lattice carries the geometry determined by its ambient group. This leads to the following precise question

Geometric Question: *does every smooth topologically transitive action of a higher rank lattice preserve a rigid geometric structure (see section* **1** *for definitions) on some open dense set ?*

This expected behaviour is in sharp contrast with actions of $\mathbb{Z}$, *i.e.* action generated by a diffeomorphism. Even for the best understood class, Anosov diffeomophims, the answer for the smooth rigidity question is no. In this context,

a classical question related to the geometric one is the question to decide wether a diffeomorphism is linearizable smoothly in the neighbourhood of a fixed point. Once more the answer is no in the smooth category.

Another measure of the rarity of lattices actions is the following conjecture of Zimmer, which is still open, even for $n = 3$.

ZIMMER'S CONJECTURE: *there is no non trivial smooth volume preserving ergodic action of $SL(n, \mathbb{Z})$, with $n \geq 3$, on a manifold of dimension strictly less than n.*

Here, a non trivial action means an action which does not factor through a finite group.

Although the word survey is written in the abstract, this article has no pretention to be exhaustive. The references quoted here should therefore be considered as starting links to explore the subject rather than the definitive ones on this topic. I also have tried to write it in a way accessible to non experts. It follows that in less than 10 pages, I will have to omit important historical results. Worse than that, most of the results I will present, will only be special cases of the original theorems, thus restricting the generality and beauty of the work of many of my colleagues. Once and for all, I apologize here for all these outrageous omissions and simplifications.

The structure of this article is as follows: the first two sections (1. on rigid geometric structures, 2. on superrigidity) are introductory; we then present known examples of actions of lattices in section 3; section 4 is concerned with hyperbolic (in the dynamical sense) actions; section 5 deals with (non volume preserving) actions on boundary spaces, such as the action of $SL(n, \mathbb{Z})$ on the n-dimensional sphere; in 6, we will discuss analytic actions; section 7 exposes a result of topological nature; finally in section 8, we, at least, quit the realm of lattices and present results and questions on other types of groups.

Unless otherwise specified, all objects and concepts (manifolds, actions, conjugations, *etc*) will be C^∞.

1. RIGID GEOMETRIC STRUCTURES

Let M be a n-dimensional manifold. Its kth-frame bundle, noted $M^{(k)}$, is the bundle over M whose fiber at a point x is the set of all local diffeomorphisms of $\mathbb{R}^n$ into M sending 0 to x, up to the following equivalence relation: having the same derivatives up to order k. The structure group of this bundle is the group $G^{(k)}$ of k-jets of diffeomorphims of $\mathbb{R}^n$, fixing 0. A *geometric A-structure (of type V)* is a section of the bundle associated to an algebraic action of $G^{(k)}$ on an algebraic variety V. Usual geometric structures (*e.g.* affine, riemannian, symplectic, complex, conformal ...) are of this type.

It does make sense for a diffeomorphism to preserve a geometric A-structure, and such a diffeomorphism will be called an *isometry* of the structure. A geometric A-structure is called *rigid* (in the sense of M. Gromov) if there exists some integer l such that all derivatives of an isometry fixing a point is completely determined

by its first l ones. For instance, connections, non degenerate metrics, projective structures are rigid, though complex and symplectic structures are not.

This notion has been introduced by M. Gromov in [Gr] where he also presents a proof (later corrected by Y. Benoist [Be]) of the following important result

Theorem (Gromov [Gr]) *If the pseudogroup of local isometries of a rigid geometric structure has a dense orbit, it has an open dense orbit.*

Basically, this theorem says that, up to a technical point, every geometric structure admitting a topologically transitive group of isometries is modelled on some locally homogeneous space on some open dense set. It follows that to prove an action preserves a rigid geometric structure is an important step (though not the last one) towards an explicit description of this action. This is why I wanted to emphasize the geometric question stated in the introduction. On the other hand, I cannot survey the vast field of actions preserving geometric structures and instead refer to [G], [d'A-G] and [Z2].

2. Lattices and superrigidity

Let Γ be a lattice in a group G. Let's first discuss the interplay between actions of Γ and of G. Obviously an action of G on M restricts to an action of Γ on M. On the other hand, from an ergodic volume preserving action of Γ on M, we get a an ergodic volume preserving action of G on $(M \times G)/\Gamma$. This latter action, called the *suspended action* carries most of the information about the action of Γ. This is why informations about actions of lattices are quite often immediately derived from results about actions of the ambient group and *vice versa*. In particular, one can prove that for topologically transitive actions the fact that the action of Γ preserves a rigid geometric structure on some open dense set is equivalent to the fact the suspended action of G preserves a rigid geometric structure on some open dense set.

The geometric data discussed in Zimmer's version of Margulis superrigidity are the following. First, we have a finite volume preserving ergodic action of Γ, or equivalently of G, on a manifold M. Second, we assume this action lifts to an action on a H-principal bundle and we wish to decribe the lifted action. A basic example, *ρ-twisted action*, is given by a representation ρ of G in H and the action of Γ on $M \times H$ given by $\gamma(m.h) = (\gamma m, \rho(\gamma)m)$.

Let us now fix a class of regularity, *e.g* measurable, continuous, smooth *etc.*, we then suppose that H is a semisimple Lie group and is minimal in the following sense: there is no Γ-invariant section of any associated H/L-bundle, for all semisimple subgroups L of H. Here, the section is required to be of the desired regularity class and defined on a set of full measure (in the case of measurable sections), or on an open dense set (in the continuous case). For the sake of simplicity, assume futhermore that H is simple and non compact.

The superrigidity question is to decide whether or not the action is equivalent (by a bundle isomorphim) to a ρ-twisted action over, maybe, a slightly smaller set (*i.e.* of full measure, or open dense, depending on the category).

Zimmer's version of Margulis superrigidity asserts that the answer is always yes in the measurable category. Topological or smooth superrigidity tries to figure out

a decent extra hypothesis that would make the story work in the topological or smooth context.

In the particular case when the action is the lift of the action of Γ to $M^{(k)}$, one may think of the superrigidity question as a linearized version of the geometric question. A typical application of the ideas of superrigidity in the smooth context is the following result, we shall only state for actions of Lie groups for the sake of simplicity.

THEOREM (R. FERES-F. LABOURIE [F-L]) *Let G be a simple Lie group of real rank greater than 2 (e.g. $SL(n,\mathbb{R})$, $n \geq 3$), acting ergodically preserving a finite volume on a manifold M, assume that some real split element of G (e.g. a real diagonizable matrix different from the identity) preserves a rigid geometric structure on some open dense set, then the whole group preserves a rigid geometric structure on some open dense set.*

The books [Z] and [M] are the standard references on the subject. Notice that [F-L] explains a short and self contained proof of a special case of superrigidity, later expanded in [F]. This latter reference should be recommended for a first approach.

It may be useful to explain an important structural property of simple real Lie groups G of rank greater than 2 which make them very different to those of real rank 1. This property is easily seen in $SL(n,\mathbb{R})$, $n \geq 3$: given two real split matrices B and C there exist a finite sequence of matrices A_i, $i \in 1,\ldots,p$, such that A_i commutes with A_{i+1}, $A_1 = B$ and $A_p = C$. We shall say a group having such a property is *generated by a chain of centralizers*. Although it is not clear that higher rank lattices have this property, it is important that the ambient group have it. One of the major steps of superrigidity is to infer a property of the whole group from a property sastified by one element using chains of centralizers. In the measurable category, we can build, using the Kakutani-Markov theorem, a measurable objet invariant by a single element and from this, build something invariant for the whole group. In the other categories, the existence of an object invariant by a single element is far from granted.

3. EXAMPLES (AND COUNTER-EXAMPLES) OF ACTIONS OF LATTICES

One of the interests of lattices is that they possess many actions, mainly on locally homegeneous spaces.

(a) Isometric actions. This first class may be considered as the trivial case of the theory. Since higher rank lattices admit morphisms into compact groups, it follows that we can construct lots of smooth ergodic actions of a lattice preserving a riemaniann metric. Obviously these examples cannot be classified. They do however exhibit a rigidity property shown by J. Benveniste

THEOREM (J. BENVENISTE [B1]) *Every isometric action of a higher rank cocompact lattice on a compact manifold is locally rigid.*

(b) Volume preserving actions. The following two examples are sometimes called *standard actions*.

(1) This first example generalizes the action of $SL(n,\mathbb{Z})$ on the n-dimensional torus. Let N be a simply connected nilpotent group and Λ a lattice in N, take now a homomorphism of G in $Aut(N)$, such that a lattice Γ normalizes Λ. It follows that Γ acts on N/Λ.
(2) We can also take a morphism of G in a unimodular Lie group H, and take the corresponding left induced left action of Γ on H/Λ where Λ is a cocompact lattice in H. In some sense, this action generalizes the geodesic flow for rank 1 symetric spaces.

(c) Weakly hyperbolic standard actions. It is well known that both the action of $SL(n,\mathbb{Z})$ on the torus and the geodesic flow of negatively curved manifolds have some hyperbolic (or Anosov) properties from the point of view of dynamical system. Here are now subclasses of the above examples which exhibit some hyperbolic behaviour. I will give an algebraic description of these action, refereeing to [M-Q] for the much more useful (but longer) dynamical description.
(3) Start with an example like (1). We then have a natural morphism of Γ in $Aut(\mathcal{N})$ the automorphisms of the Lie algebra $\mathcal{N}$ of N. This morphism essentially comes from a representation π of G and we say the action is *weakly hyperbolic* if π does not contain a trivial representation.
(4) This time, start with a type 2 example. Such a standard action is weakly hyperbolic if the centralizer of $\pi(G)$ in H is discrete.

(d) Actions on boundaries. A compact manifold M will be called a *boundary for* G if $M = G/P$, where P has finetely many components. The groups G and Γ act on boundaries, however these actions will never preserve any measure. Typical examples are te action of $SL(n,\mathbb{R})$ on spheres, projectives spaces, flag manifolds *etc.*

(e) Exotic examples. So far, all the above examples of actions preserve a rigid geometric structure everywhere on the manifold. The first "exotic" example is due to A. Katok and J. Lewis [K-L1]. Start with the action of $SL(n,\mathbb{Z})$ on the n-dimensional torus T^n. This action has a fixed point p_0. We can now blow up this point as algebraic geometers do, that is replace it by the projective space of its tangent plane. This new action will only preserve the original geometric structure on some open dense set. Exploring this idea, J. Benveniste [B2] has contructed a smooth family of ergodic actions of semisimple Lie groups, none of which are conjugate. This in particular implies that the answer to the rigidity question can not be always yes without any extra assumption. However, Beneveniste's examples preserve rigid geometric structures on some open dense set.

4. Hyperbolic actions

Since the original works of S. Hurder [H], A. Katok and J. Lewis [K-L2], actions of lattices with some hyperbolic behaviour have attracted a lot of attention.

Assume a 1-parameter group L_t acts on a space M and suppose its action lifts to a vector bundle E equipped with some metric. We say such an action is *Anosov* on E if we can find a continous splitting $E = E^+ \oplus E^-$, where

$$\exists A,\ B > 0,\ s.t.\ \forall u^{\pm} \in E^{\pm},\ \forall t > 0,\ \|L_{\pm t}(u^{\pm})\| \leq Ae^{-Bt}.$$

Now, we say that the action of a group on a compact manifold is *Anosov* if there exist a 1-parameter group L whose action is Anosov on TM/V, where V is the tangent bundle to the orbits of G. Next the action of a lattice is said to be *Anosov* if the suspended action is Anosov (with a little extra hypothesis, the definition makes sense even for non cocompact lattices). Of course the action of $SL(n,\mathbb{Z})$ on the n-dimensional torus is Anosov. Weak hyperbolicity is a generalisation of this hypothesis.

This is typically a situation where superrigidity will work. Taking the continuous splitting, $E = E^+ \oplus E^-$, as a starting point of the superrigidity procedure, will produce in good cases, after using a chain of centralizers, a rigid geometric structure on M.

In this situation, the following question seems to be within reach

CONJECTURE: *Every Anosov action of a lattice is smoothly conjugate to the standard action on a nilmanifold. More generally, every weakly hyperbolic action of a lattice on a compact manifold should be standard.*

I cannot state all the results on this subject, and I will underline two recent results. The first one is a definitive result on the local rigidity question.

THEOREM (G. MARGULIS - NANTIAN QIAN [M-Q]) *Standard weakly hyperbolic actions of higher rank lattices are smoothly rigid; that is, every smooth action close enough to the original one is smoothly conjugate to it.*

The second is a global result which make very weak topological assumption on the underlying manifold:

THEOREM (R. FERES - F. LABOURIE [F-L]) *Assume we have a Anosov volume preserving action of a lattice in $SL(n,\mathbb{R})$, $n \geq 3$ on some n-dimensional compact manifold M. Then M is a torus and the action preserves the connection of a flat metric on M.*

We should note the preceding results are valid only for higher rank lattices, and make therefore strong use of superrigidity.

Much more strikingly, A. Katok and R. Spatzier have obtained rigidity results for actions of $\mathbb{R}^k$ and $\mathbb{Z}^k$ for $k \geq 2$. The definition of standard actions makes also sense for these groups. Associate to these actions, are representations of $\mathbb{R}^k$ in $Aut(\mathcal{N})$ and $\mathcal{H}$ respectively where $\mathcal{N}$ and $\mathcal{H}$ are the lie algebras of N and H. We say the standard actions of $\mathbb{R}^k$ and $\mathbb{Z}^k$, have *semisimple linear part* if the corresponding representation of $\mathbb{R}^k$ is semisimple. An example of a standard action Anosov action with linear semisimple part is the left action of the group of real diagonalizable $n \times n$ matrices on $SL(n,\mathbb{R})/SL(n,\mathbb{Z})$. The result is then

THEOREM (A. KATOK-R. SPATZIER [K-S]) *Every standard Anosov action of $\mathbb{R}^k$ or $\mathbb{Z}^k$ with semisimple linear part is smoothly rigid.*

5. ACTIONS ON BOUNDARIES

Actions on S^1. Actions of groups on the circle is a subject by itself. I am therefore going to single out for the moment only results concerning lattices. Any

time you have a hyperbolic structure on a compact surface S, this defines on action of $\pi_1(S)$ on the boundary at infinity of the hyperbolic plane, and this action factors through the standard $PSL(2,\mathbb{R})$ action, and in particular preserves a projective structure on the circle. Of course, since we can deform hyperbolic metrics on the surface, such an action is not locally rigid, but nevertheless these actions can be characterized. Let's first remark that to every action of $\pi_1(S)$ on the circle we can associate a number, namely the Euler class of the associated circle bundle on the surface. For the actions I just described, this Euler number is maximal (*i.e.* equal to the euler number of the surface). E. Ghys has proved

Theorem (E. Ghys [Gh]) *Every smooth action of the fundamental group of a compact surface with maximal Euler number factors through an action of $PSL(2,\mathbb{R})$ and in particular preserves a real projective structure on the circle.*

Very recently, E. Ghys and independently, M. Burger and N. Monod have announced results that tend to prove the following conjecture

Conjecture: *There is no non trivial smooth action of a higher rank lattice on the circle.*

This is to compare with the following result of D. Witte

Theorem (D. Witte [W]) *There is no non trivial continuous action of a higher rank lattice of $\mathbb{Q}$-rank greater than 2 on the circle.*

Notice that the smoothness (actually C^1) hypothesis is extremely restrictive, and that the proofs of Ghys, Burger and Monod do not adapt to the continuous case. In both cases, non trivial means actions that do not factor through a finite group.

Higher dimensional boundaries. Since actions on boundaries do not preserve measure superrigidity will not work nicely. However, a standard observation, already used by E. Ghys, is that there is some correspondance between the actions on Γ on G/P and the action of L on G/Γ where L is the reductive part of P. Using this idea and exploiting their results for the rigidity of abelian actions, A. Katok and R. Spatzier have proved the following almost definitive result

Theorem (A. Katok, R. Spatzier [K-S]) *Let Γ be a higher rank cocompact lattice of a simple group G, then the action of Γ on a boundary for G is smoothly rigid.*

This greatly generalizes a previous result of M. Kanai [K] whose completely different proof relied on stochastic calculus.

6. Analytic actions

It is a classical question to determine wether a diffeomorphim is linearizable on the neighbourhood of a fixed point, which, in geometric terms, means to preserve a flat connection. For analytic actions, E. Ghys and G. Cairns have shown

Theorem (G. Cairns, E. Ghys [C-Gh]) *Every higher rank lattice acting analytically on $\mathbb{R}^n$ fixing 0 is linearizable.*

On the other hand, the same authors have shown that there exist smooth actions of $SL(n,\mathbb{R})$ having a fixed point and which are not linearizable. An immediate application of the preceding theorem is the

COROLLARY *Every topologically transitive analytic action of a higher rank lattice having a fixed point preserves a flat connection on some open dense set.*

Quite recently, amongst other results, B. Farb and P. Shalen have shown the following version of Zimmer's conjecture

THEOREM (B. FARB, P. SHALEN [F-S]) *Any analytic action of $SL(n,\mathbb{Z})$, $n \geq 5$ on a compact surface, other than the torus and the Klein bottle, factors through a finite group.*

To prove this, they start with a fixed point for some element of the lattice, then, using the fact the lattice itself is generated by chain of centralizers, they reduce the situation to a 1-dimensional question which is settled by Witte's Theorem.

7. A TOPOLOGICAL RESULT

Even though we cannot for the moment classify actions of higher rank lattices, it is interesting to have some restrictions on the topology of the underlying manifold. Let's start with a definition. Assume a simply connected group acts ergodically on a manifold M preserving a volume form, and notice that the action of G lifts to any finite cover P of M. The action is said to be *totally engaging* is there is no measurable G-invariant section of $P \to M$ for any finite cover P of M. For a simple group of rank greater than 2, the action of G on G/Γ, for Γ a cocompact lattice, is totally engaging. The following result of A. Lubotzky and R. Zimmer sheds some light on the topological structure of the manifold M.

THEOREM (A. LUBOTZKY, R. ZIMMER [L-Z]) *Suppose that the action of a simply connected simple Lie group G of real rank greater than 2 is totally engaging on M, then for all finite dimensional representation σ of $\pi_1(M)$ in $GL(V)$, $\sigma(\pi_1(M))$ is an arithmetic lattice. In fact $\sigma(\pi_1(M))$ is commensurable to $H_{\mathbb{Z}}$, where H is a linear $\mathbb{Q}$-group in which contains a quotient of G.*

Let's try to explain the last sentence. It means first that there exists a group H which is a subgroup of some linear group $Gl(\mathbb{Q}^N)$ and which is defined by polynomial equations with rational coefficients. $H_{\mathbb{Z}}$ consist then of the matrices in H with integer entries. *Being commensurable* is the relation of equivalence generated by the relation *being of finite index in*. This theorem can be thought of as a generalization of Margulis's Arithemeticity Theorems [M]. Again let's notice that Benveniste's examples [B2] are not totally engaging and do not satisfy the conclusion of the above theorem.

8. OTHER GROUPS AND QUESTIONS

So far, we have stayed in the realm of higher rank lattices with a brief excursion in the kingdom of surfaces and abelian groups. Our reasons for that were the following: first, we have lots of examples of actions of lattices; second, by using the supension procedure we can turn questions about lattices into question about

real Lie groups whose structure is quite well known; third the superrigidity method yelds interesting results.

What are now the other candidates for being large groups ? A typical property we would like them to satisfy is that they are generated by chains of centralizers, at least virtually, like lattices. Another important property used in superrigidity is that the centralizers that appear in the chain is non amenable.

Even though we do not have a precise definition, we have examples of groups that are good candidates for being large groups. For instance, E. Ghys and V. Sergiescu advocate the case of the Thomson group [Gh-S] which present a rigidity property. This group is known to be generated by chain of centralizers though its non amenability is not known.

For the moment, the best candidates for being large groups are the mapping class groups $\mathcal{M}(g)$ for surfaces S of genus g greater than 2, which are believed to share many properties of higher rank lattices. E. Ghys has for instance announced

Theorem (E. Ghys) *There is no non trivial actions of the mapping class group on the circle.*

On the other hand, we have lots of examples of actions of the mapping class groups, namely on the space $X(G,g)$ of representations of $\pi_1(S)$ in a compact Lie group G. These actions are known to be volume preserving (actually they are symplectic) and W. Goldman has shown

Theorem (W. Goldman [Go]) *The action of $\mathcal{M}(g)$ on $X(SU(2),g)$ is ergodic.*

Forgetting for a brief moment that $X(G,g)$ are not manifolds, it is tempting to ask whether these actions are rigid. Here is a simpler version of this question. In the case the compact group is S^1, the moduli space is the jacobian torus

$$X(S^1,g) = H^1(S,\mathbb{R})/H^1(S,\mathbb{Z}),$$

and the action factors through a lattice action which is known to be locally rigid. Let's now ask the

Test Question *Is the action of $\mathcal{M}(g)$ locally rigid on $X(S^1,g)$?*

References

[d'A-G] G. d'Ambra, M. Gromov. *Lectures on transformation groups: Geometry and dynamics*, J. Differen. Geom., Suppl. 1, 19-111 (1991).

[Be] Y. Benoist. *Orbites des structures rigides (d'après M. Gromov)*, in *Feuilletages et systèmes intégrables*, Birkhäuser, Prog. Math. 145, 1-17 (1997).

[B1] J. Benveniste. *Deformation rigidity of isometric actions of lattices*, preprint.

[B2] J. Benveniste. *Exotic geometric actions of semisimple groups and their deformations*, preprint.

[C-Gh] G. Cairns, E. Ghys. *The local linearization problem for smooth $SL(n)$-actions*, Enseign. Math., II. Ser. 43, No.1-2, 133-171 (1997).

[F-S] B. Farb, P. Shalen. *Real-analytic actions of lattices,* To appear in Inventiones Math..

[F] R. Feres. *Dynamical Systems and Semisimple Groups, an Introduction,* Cambridge Tracts in Mathematics, vol 126, Cambridge (1998).

[F-L] R. Feres, F. Labourie. *Topological superrigidity and Anosov actions of lattices,* to appear in Ann. Ec. Norm. Sup.

[Gh] E. Ghys. *Rigidité différentiable des groupes fuchsiens,* Publ. Math., Inst. Hautes Etud. Sci. 78, 163-185 (1993).

[Gh-S] E. Ghys, V. Sergiescu. *Sur un groupe remarquable de diffémorphismes du cercle,* Comment. Math. Helv. **62**, 185-239 (1987).

[Go] W. Goldman. *Ergodic Theory on Moduli Spaces,* Annals of Mathematics, 146, 1-33 (1997).

[G] M. Gromov. *Rigid transformations groups,* in *Géométrie différentielle*, Travaux en Cours 33, 65-139, Hermann, Paris (1988).

[H] S. Hurder. *Rigidity for Anosov actions of higher rank lattices,* Ann. of Math, 135, 361-410 (1992).

[K-L1] A. Katok, J. Lewis. *Global rigidity for lattices actions on tori and new examples of volume preserving actions,* Israel J. of Math. 93, 253-281 (1996).

[K-L2] A. Katok, J. Lewis. *Local rigidity for certain groups of toral automorphisms,* Israel J. of Math. 75, 203-241 (1991).

[K-S] A. Katok, R. Spatzier. *Differential rigidity of Anosov actions of higher rank abelian groups and algebraic lattice actions,* preprint.

[L-Z] A. Lubotzky, R. Zimmer. *Arithmetic structure of fundamental groups and actions of semisimple Lie groups,* preprint.

[M] G. Margulis. *Discrete Subgroups of Semisimple Lie Groups,* Springer Verlag, New York (1991).

[M-Q] G. Margulis, Nantian Qian. *Rigidity of weakly hyperbolic actions of higher real rank semisimple Lie groups and lattices*, preprint.

[W] D. Witte. *Arithmetic groups of higher $\mathbb{Q}$-rank cannot act on 1-manifolds*, Proc. Am. Math. Soc. 122, No.2, 333-340 (1994).

[Z1] R. Zimmer. *Ergodic Theory and Semisimple Groups,* Monographs in Mathematics, Birkhäuser, Boston (1984).

[Z2] R. Zimmer. *Automorphism groups and fundamental groups of geometric manifolds,* Differential geometry. Part 3: Riemannian geometry. Proceedings of a summer research institute. Proc. Symp. Pure Math. 54, Part 3, 693-710, Providence (1993).

François Labourie
URA D1169 du CNRS et IUF
Département de Mathématiques
Université Paris Sud
91405 Orsay, France

Curvature Contents of Geometric Spaces

Joachim Lohkamp

Abstract. We discuss curvature relevant deformations of spaces and indicate the existence of some individual capacity of a manifold (and more general spaces) measuring a maximal amount of curvature that could be carried by this space.

1991 Mathematics Subject Classification: 53 and 58
Keywords and Phrases: Curvature, Rigidity, Flexibility

1. Introduction

It turned out recently that the general tendency is that there is a natural upper bound but definitely *no lower bound* for the curvature on a given manifold or at least certain families of such spaces.
Moreover, roughly speaking, as we approach the maximal curvature "amount" we will often reach a particularly rigid geometry or a singular one (if the manifold cannot exhibit the suitable form of symmetry).
On the other hand, by decreasing the curvature (even uniformly) we gain flexibility: For instance we may combine various geometric conditions. That is certain geometric properties which are *coupled* for higher curvature amounts will become more and more independent the more the curvature melts away. Most notably the curvature and the coarse metric geometry, e.g. volumes, systoles or various radii, will "finally" appear entirely unlinked.
Motivated by this sort of observation we are tempted to think of an *individual curvature content* (or capacity) of a given manifold depending however on the curvature problem under consideration.
From this viewpoint it no longer matters whether this content has a particular sign: there is just a maximal amount of positivity that may be carried by the manifold and this may be positive or not.
We have not made any attempt to state a sharp "definition" of our presently still intuitive notion of a "curvature content" as we do not want to destroy the suggestive flavour of this notion. But certainly such a measure will depend on the context (additional constraints etc.), as we will see below.
We will start our short journey in dimension 2 and give an interpretation of some classical results which allows us to proceed directly to higher dimensions where we will treat the three basic notions of scalar, Ricci and sectional curvature and will concentrate entirely on these.

2. Surfaces

In dimension 2 we have a unique notion of curvature (the Gaussian curvature K) and it shares some properties of both extremes: scalar and sectional curvature. For instance it can be conformally deformed into a metric of constant curvature, but its sign also shows the typical implications of sectional curvature.
Furthermore its behaviour when one tries to increase/decrease the curvature amount lies somewhere in between these two extremes.
We start with a reinterpretation of the uniformization theorem:
First note that by attaching a handle to the Euclidean plane one can construct a complete surface N^2 which contains an open bounded set U where $K < 0$ and is isometric to the Euclidean $\mathbb{R}^2 \setminus B_1(0)$ on $N^2 \setminus U$.
Now we take a two-sphere. If we cut a small disc out of S^2 and instead glue a suitably scaled copy of $U \subset N^2$ in its place we get a surface (a torus) which admits a metric with $K = 0$. Iterating this procedure, that is attaching several copies of such an "island", one gets surfaces allowing $K < 0$-metrics. This is obvious due to the *uniformization theorem* because we know that adding $U \subset N^2$ enlarges the genus by 1 for each copy of $U \subset N^2$ while the geometric properties specified above become redundant.
However the point is that we can also succeed by using purely the available geometry: Cover S^2 (or any other surface) by very small discs $D_r(p_i)$ with an upper bounded covering number independent of the radius r. Next substitute the discs $D_{r/10}(p_i)$ (which can be assumed to be disjoint) by suitably rescaled copies of $U \subset N^2$. This new Riemannian manifold can be deformed using a slight and explicit (conformal) deformation to yield a $K < 0$-metric (cf. [L1]).
Of course, this surface will have quite a high genus, but we have not used the uniformization theorem at all. Actually, we will see (in higher dimensions) that it is quite natural to consider this construction a curvature decreasing "deformation" of a given surface.
Now the second interesting point is the following: there is *no counterpart* for getting more *positively* curved surfaces: Of course, this is clear from Gauss-Bonnet, but let us also have a look at what really happens.
The construction of negative curvature can take place in any open set of arbitrarily small metrical size. On the other hand even in the case where we just reverse this construction (obviously gaining some positivity) we have to start by choosing a closed curve which is *not* homotopic to a constant map. This cannot be accomplished locally.

3. Scalar Curvature

The weakest generalization of Gaussian Curvature in dimension 2 to dimensions $n \geq 3$ is the notion of scalar curvature $Scal$.

In this case we can decrease the curvature locally without any topological changes: we can find metrics on $\mathbb{R}^n$ which satisfy $Scal < 0$ on the unit ball in $\mathbb{R}^n$ and are

Euclidean outside it (i.e. the metrical analogue of $U \subset N^2$ above). Here is a generalized version (cf. [L2]):

Theorem 3.1: *Let $U \subset M$ be an open subset and f any smooth function on M with*
$f < \mathrm{Scal}(g)$ on U and $f \equiv \mathrm{Scal}(g)$ on $M \setminus U$.
Then, for each $\varepsilon > 0$, there is a metric g_ε on M and an ε- neighborhood U_ε of U with

$$g \equiv g_\varepsilon \text{ on } M \setminus U_\varepsilon \text{ and } f - \varepsilon \leq \mathrm{Scal}(g_\varepsilon) \leq f \text{ on } U_\varepsilon.$$

This new metric g_ε can be chosen arbitrarily near to the old one in C^0-topology (using also [L5]). Thus we can basically attach some negative curvature without changing the shape of the manifold.

Next we want to know whether there is any corresponding statement for curvature *increasing* deformations (even admitting topological changes). Here one should take a look at a theorem originating from general relativity, the so-called "positive energy theorem" originally proved by Schoen-Yau and Witten (cf. [SY] and [PT]):

Theorem 3.2: *Let (M, g) be an asymptotically flat manifold with $\mathrm{Scal}(g) \geq 0$. Then the energy $E(g)$ is non-negative and $E(g) = 0$ iff (M, g) is flat.*

This already gives a first hint of the existence of a "maximal content" as becomes clear once one realizes that this problem can be solved as follows: In a first step one transforms it to a local one and then one plays this off against some curvature capacity consideration:
(3.2) can be reduced to the (*non*)trivial special case: the *only* complete Riemannian manifold which is Euclidean outside a bounded domain U with $\mathrm{Scal}(g) \geq 0$ on U is the Euclidean space. Now assume the existence of a non-flat manifold of this type. Then one "reverses (3.1)" and constructs manifolds whose positive scalar curvature amount turns out to be actually "too large" for the underlying space thereby proving the non-existence of such a manifold (for details cf. [L2]).
In this context we also meet a recent theorem by Llarull [Ll] for S^n equipped with the round metric g_{round}:

Theorem 3.3: *Let g be any metric on S^n with $g(v, w) \geq g_{round}(v, w)$ for each oriented pair of vectors $v, w \in T_pS^n$, $p \in S^n$ and $Scal(g) \geq 1$. Then $g \equiv g_{round}$.*

Another related subject is the solution of the Yamabe problem (cf. [LP]) claiming that every metric can be conformally deformed into a metric of constant scalar curvature. Recall that the original solution of the Yamabe problem uses the positive energy theorem.

The geometric ingredient of that proof is the following theorem by Aubin and Schoen (cf. [LP]) in dimension $n \geq 3$:

THEOREM 3.4: *For every closed manifold M^n, $n \geq 3$ the infimum of the normalized total curvature functional $\boldsymbol{S}(\varphi)$ within a given conformal class $\varphi^{4/n-2} \cdot g$*

$$\mathbb{S}(M,g) := \inf_{\varphi \not\equiv 0} \boldsymbol{S}(\varphi)$$
$$:= \inf_{\varphi \not\equiv 0} \int_M (\|\nabla\varphi\|^2 + \frac{(n-2)}{4(n-1)} \cdot \mathrm{Scal}(g) \cdot \varphi^2) dV_g / \left(\int_M |\varphi|^{2n/(n-2)} dV \right)^{n-2/n}$$

satisfies: $\mathbb{S} \leq \mathbb{S}(S^n, g_{round})$ with equality iff M is conformal to the round S^n.

This means that taking the supremum over all the conformal classes of metrics on a manifold one gets an individual upper bound for a certain kind of scalar curvature content on the manifold. In the case of a torus for instance this is zero and corresponds to a flat metric (cf. [S]).
This is also a good place to check what happens when we approach the supremum by a sequence of smooth metrics (cf. [A]):
Under some additional assumptions *there is a subsequence that converges either to a smooth Einstein metric or to a singular limit consisting of finitely many smooth noncompact Einstein manifolds with cusps.*

4. RICCI CURVATURE

Now we meet the main candidate for a meaningful notion of curvature content for general smooth manifolds which is the Ricci curvature *Ric*.
We start with a counterpart of (3.1) that is easily derived from [L4] and this time follows from the fact that we can even find metrics which have negative Ricci curvature on a ball in $\mathbb{R}^n$ and are Euclidean outside it:

THEOREM 4.1: *Let $U \subset M$ be an open subset and f any smooth function on the unit tangent bundle SM with $f(\nu) < \mathrm{Ric}(g)(\nu)$ on SU and $f \equiv \mathrm{Ric}(g)$ on $SM \setminus SU$.*
Then there is a smooth metric g_f on M with $Vol(M,g) = Vol(M,g_f)$ and

$$g \equiv g_f \text{ on } M \setminus U \text{ and } \mathrm{Ric}(g_f)(\nu) \leq f(\nu) \text{ on } SU.$$

We supplement this theorem with two examples of the "decoupling" effect of curvature dereasing deformations mentioned in the introduction:
For simplicity take a compact manifold M^n and $U = M$, then there are several extensions of the statement above:
We may also prescribe finitely many Laplace eigenvalues (cf. [L3]) *or* we can choose g_f arbitrarily near to g in various geometric topologies (cf. [L5]).

The counterpart of (4.1) for positive curvature (even allowing topological changes) is excluded already by the scalar curvature argument above, however here we may

also use the standard Bochner argument.
Thus we can immediately proceed to the question of "curvature contents" and their application.
Here one has several very recent results by Cheeger and Colding and by Besson, Courtois and Gallot (cf. [Co],[Ga] and references therein). We state some of them as follows:

Theorem 4.2: *There is an $\varepsilon(n) > 0$ such that if M^n is a closed manifold with $Ric(g) \geq (n-1)g$ and $Vol(M^n) \geq Vol(S^n) - \varepsilon(n)$, then M^n is diffeomorphic to S^n.*

Theorem 4.3: *Let M^n be a closed manifold, and assume there are metrics g_0 on M^n with $Sec \equiv -1$ and g on M^n with $Ric(g) \geq -(n-1)g$. Then $Vol(M^n, g) \geq Vol(M^n, g_0)$. If equality holds and $n \geq 3$, then (M^n, g) and (M^n, g_0) are isometric.*

These two results can be reinterpreted as follows: A manifold M^n (with some normalized volume) admits at most as much (lower bounded) Ricci curvature as the sphere and this extremum is reached if and only if M^n is the sphere. Secondly, if M^n carries some sufficiently symmetric geometry, then the Ricci curvature amount cannot exceed the borderline preassigned by that geometry. Thus even these "local maxima" for the present Ricci curvature content are very distinguished.
There is also another observation: For some Ricci (and sectional) curvature content notions the superlevel sets are very thin: *There are only finitely many homotopy tyes of manifolds whose Ricci curvature capacity exceeds certain bounds* in the respective context (a general reference is [P]).

5. Sectional Curvature

The strongest curvature notion is that of sectional curvature *Sec*. There are natural generalizations of this curvature notion to metric spaces, specifically "Alexandrov spaces", which will also be of interest to us.
We will very briefly remind the reader of these general notions. For details cf. [BN] and [BGP].
A locally complete metric space (X, d) is called *a space of curvature $\geq k$ or $\leq k$ respectively* if the following conditions are satisfied at least locally:

(i) Any two points can be joined by a geodesic, i.e. by a curve whose length equals the distance of its endpoints.

(ii) For any geodesic triangle Δpqr and any point z on an arbitrary side pq, we find a triangle ΔPQR and a point Z on the side PQ in the simply connected smooth surface M_k of constant sectional curvature $= k$ with

$d(p,q) = d_{M_k}(P,Q)$, $d(p,r) = d_{M_k}(P,R)$, $d(q,r) = d_{M_k}(Q,R)$ *and*

$d(z,r) \geq$ *(or* $\leq$*)* $d_{M_k}(Z,R)$, $d(z,p) = d_{M_k}(Z,P)$, $d(z,q) = d_{M_k}(Z,Q)$

Many theorems translate from Riemannian to Alexandrov geometry. For instance, if (X, d) is complete with curvature $\geq k > 0$, then X is compact with $diam(X, d) \leq \pi/\sqrt{k}$. Also, completeness and curvature ≤ 0 imply contractibility of the universal covering.
We will again begin with curvature decreasing deformations. From the preceding remark it is clear that we have to admit "deformations" that alter the topology to some extent. This is what was anticipated in section 2 in the case of surfaces. We will see that we can interpret the concept of "hyperbolization" (cf. [G1], [DJ] and [CD]) as a substitute for adding handles to a given surface locally. A hyperbolization is a process converting a space into a negatively curved one. The known processes work as follows: One starts with some, say PL-manifold M and substitutes (in one or several steps) each simplex by a negatively curved manifold with some smooth boundary. This boundary might be different from the simplex-boundary, but if one carries this out everywhere simultaneously these boundaries fit together and on gets a new PL-manifold $\mathcal{H}(M)$.
This can be achieved in such a way that one obtains a negatively curved Alexandrov space $\mathcal{H}(M)$ and in some obvious sense theses changes are *local* .
Of course, the aim is to find constructions which do not damage the topology too much. This is paraphrased in some kind of axioms in [DJ] and [CD] and the main point is that such processes exist. We state this as a descriptive theorem:

THEOREM 5.1: *There is a process that converts a cell complex K into a new polyhedron $\mathcal{H}(K)$ which admits a metric with curvature ≤ -1 such that*

(i) If K is a PL-manifold, then so is $\mathcal{H}(K)$

(ii) There is a map ϕ: $\mathcal{H}(K) \to K$ which induces a surjection on homology and is such that ϕ pulls back the rational Pontryagin classes from K to those of $\mathcal{H}(K)$.

(iii) $\mathcal{H}$ behaves functorial and preserves the local structure. That is a PL-embedding $f : L \to K$ induces an isometric map $\mathcal{H}(f) : \mathcal{H}(L) \to \mathcal{H}(K)$ such that $\mathcal{H}(L) \subset \mathcal{H}(K)$ is totally geodesic and in the case of $L =$ single simplex $\subset K$ the "ambient angles" (more precisely the link) are mapped PL-isomorphically.

Iterating this process or applying it to sufficiently fine triangulations one gets geometries whose curvatures are arbitrarily strongly negative.

Next we turn to the question of whether one can increase curvature at least in the sense of Alexandrov. But it is easily verified that there is no process that satisfies similar axioms to those above but *increases* the curvature since otherwise one may, for instance, construct a complete *non-compact* Alexandrov space with curvature > 1 which is impossible.
Finally in the case of sectional curvature we meet a lot of new reasonable notions of capacities. We select one example: Gromov's Betti number theorem [G2]

Theorem 5.2: *There is a constant $c(n,k)$ such that for a manifold M^n with $Sec \geq k$ and diameter $= 1$: $\sum_{i=0}^{n} i$-th Betti number $\leq c(n,k)$.*

This gives an obvious form of curvature content: For a normalized diameter and fixed total Betti number there is an upper bound k_0 for the existence of metrics with $Sec \geq k$.
Here we have a nice opportunity to compare Ricci and sectional curvature: For the connected sum of sufficiently many copies of $S^n \times S^m$, $m, n \geq 2$ or $\mathbb{C}P^2$, we find that k_0 becomes arbitrarily strongly negative, while Sha and Yang (cf. [ShY]) resp. Perelman (unpublished) have shown that these manifolds always carry a $Ric > 0$-metric. Thus the maximal amounts of Ricci and sectional curvature on a manifold may differ to any extent.

6. Conclusion

Finally we want to add some general remarks and suggestions.
1. In many cases the curvature capacity is related to lower curvature bounds and correspondingly the "rigid" maximal geometries usually have constant (e.g. Ricci) curvature.
This is not surprising since "nice geometries" should have the "topological" property of not distinguishing between different parts on the manifold. However, when one starts with concrete constructions in topology one frequently breaks up this homogeneity (e.g. in Morse theory).
Then one may still think of curvature contents, this time respecting and/or forcing decompositions of the underlying space which might lead to capacity notions with an own for instance algebraic structure.
2. Motivated by the discussion concerning sectional curvature "deformations" we are led to believe that it will be reasonable in various contexts to include certain types of topological changes in a class of admissable deformations. Sometimes it might even be useful to go one (speculative) step further: consider the space and its geometry as one entity - then such notions of deformations (containing spaces with various topologies) become completely natural.

References

[A] Anderson, M.: Degeneration of metrics with bounded curvature and applications to critical metrics of Riemannian functionals, Proc. Symp. Pure Math. 54, Part 3, AMS (1993), 53-79

[BN] Berestovsii, V.N., Nikolaev, I.G.: Multidimensional generalized Riemannian Spaces, Encycl. Math. Sc. 70, Springer (1993), 165-243

[BGP] Burago, Yu., Gromov, M., Perel'man, G.: Alexandrov spaces with curvature bounded below, Russian Math. Surveys 47:2 (1992), 1-58

[CD] Charney, R. and Davis, M.: Strict Hyperbolization, Topology 34 (1995), 329-350

[C] Chavel, I.: Eigenvalues in Riemannian geometry, Academic Press (1984)

[Co] Colding, T.: These Proceedings

[DJ] Davis, M. and Januszkiewicz, T.: Hyperbolization of polyhedra, J.Diff.Geom. 34 (1991), 347-388

[Ga] Gallot, S.: These Proceedings

[G1] Gromov, M.: Hyperbolic Groups, in Essays in group theory, ed. by Gersten, MSRI Publ. 8, Springer (1987), 75-264

[G2] Gromov, M.: Curvature, diameter and Betti numbers, Comment. Math. Helv. 56 (1982), 179-195

[LP] Lee, J. and Parker, T.: The Yamabe Problem, Bull. of AMS 17 (1987), 37-81

[Ll] Llarull, M.: Sharp estimates and the Dirac Operator, Math. Ann. 310 (1998), 55-72

[L1] Lohkamp, J.: Notes on localized curvatures, Preprint

[L2] Lohkamp, J.: Scalar curvature and hammocks, Math. Ann., to appear

[L3] Lohkamp, J.: Discontinuity of geometric expansions, Comment. Math. Helv. 71 (1996), 213-228

[L4] Lohkamp, J.: Metrics of negative Ricci Curvature, Ann. of Math. 140 (1994), 655-683

[L5] Lohkamp, J.: Curvature h-principles, Ann. of Math. 142, No.3 (1995), 457-498

[PT] Parker, T. and Taubes, C.: On Witten's proof of the positive energy Theorem, Comm. Math. Phys. 84 (1982), 223-238

[P] Petersen, P.: Riemannian Geometry, GTM 171, Springer (1998)

[S] Schoen, R.: Variational theory for the total scalar curvature functional for Riemannian metrics and related, in Topics in calculus of variations, Lect. Notes Math. 1365 (1989), 120-154

[SY] Schoen, R. and Yau, S.T.: On the proof of the positive mass conjecture in general relativity, Comm. Math. Phys. 65 (1979), 45-76

[ShY] Sha, J.P. and Yang, D.G.: Positive Ricci curvature on connected sums of $S^n \times S^m$, J. Differential Geometry 33 (1991), 127-137

Joachim Lohkamp
Institut für Mathematik
Universiät Augsburg
D-86159 Augsburg

Quaternionic Analysis on Riemann Surfaces and Differential Geometry

Franz Pedit[1] and Ulrich Pinkall[2]

Abstract. We present a new approach to the differential geometry of surfaces in $\mathbb{R}^3$ and $\mathbb{R}^4$ that treats this theory as a "quaternionified" version of the complex analysis and algebraic geometry of Riemann surfaces.

1991 Mathematics Subject Classification: 53C42 14H99

1 Introduction

1.1 Meromorphic functions

Let M be a Riemann surface. Thus M is a two-dimensional differentiable manifold equipped with an almost complex structure J, i.e. on each tangent space T_pM we have an endomorphism J satisfying $J^2 = -1$, making T_pM into a one-dimensional complex vector space. J induces an operation $*$ on 1-forms ω defined as

$$*\omega(X) = \omega(JX)\,. \tag{1}$$

A map $f : M \to \mathbb{C}$ is called holomorphic if

$$*df = i\,df\,.$$

A map $f : M \to \mathbb{C} \cup \{\infty\}$ is called meromorphic if at each point either f or f^{-1} is holomorphic. Geometrically, a meromorphic function on M is just an orientation preserving (possibly branched) conformal immersion into the plane $\mathbb{C} = \mathbb{R}^2$ or rather the 2-sphere $\mathbb{CP}^1 = S^2$.

Now consider $\mathbb{C}$ as embedded in the quaternions $\mathbb{H} = \mathbb{R}^4$. Every immersed surface in $\mathbb{R}^4$ can be descibed by a conformal immersion $f : M \to \mathbb{R}^4$, where M is a suitable Riemann surface. In Section 2 we will show that conformality can again be expressed by an equation like the Cauchy-Riemann equations:

$$*df = N df\,, \tag{2}$$

where now $N : M \to S^2$ in $\mathbb{R}^3 = \operatorname{Im}\mathbb{H}$ is a map into the purely imaginary quaternions of norm 1. In the important special case where f takes values in

[1]Research supported by NSF grants DMS 522917, DMS 521088 and SFB 288 at TU-Berlin.
[2]Research supported by SFB 288 at TU-Berlin.

$\mathbb{R}^3 = \operatorname{Im}\mathbb{H}$, N is just the unit normal vector for the surface f. In the differential geometry of surfaces in $\mathbb{R}^4$, N is called the "left normal vector" of f. "Meromorphic functions" $f : M \to \mathbb{R}^4 \cup \{\infty\} = S^4 = \mathbb{HP}^1$ are defined as in the complex case.

To summarize: from the quaternionic viewpoint i is just one special imaginary quaternion of norm one. The transition from complex analysis to surface theory is done by

i. leaving the Riemann surface as it is.

ii. allowing the whole of $\mathbb{H} \cup \{\infty\}$ as the target space of meromorphic functions.

iii. writing the Cauchy Riemann equations with a "variable i".

1.2 Line bundles

A classical method to construct meromorphic functions on a Riemann surface M is to take the quotient of two holomorphic sections of a holomorphic line bundle over M. For example, if M is realized as an algebraic curve in $\mathbb{CP}^n$, then the affine coordinate functions on $\mathbb{CP}^n$ are quotients of holomorphic sections of the inverse of the tautological bundle over M. Another common way to construct meromorphic functions is to take quotients of theta functions, which also can be viewed as sections of certain holomorphic line bundles over M.

In Section 3 we introduce the notion of a holomorphic quaternionic line bundle L over M. Quotients of holomorphic sections of such bundles are meromorphic conformal maps into $\mathbb{H}$ and every conformal map can be obtained as such a quotient in a unique way.

Every complex holomorphic line bundle E gives rise to a certain holomorphic quaternionic bundle $L = E \oplus E$. The deviation of a general holomorphic quaternionic bundle L from just being a doubled complex bundle can be globally measured by a quantity

$$W = \int |Q|^2$$

called the Willmore functional of L. Here Q is a certain tensor field, the Hopf field.

On compact surfaces, W is (up to a constant) the Willmore functional in the usual sense of surface theory of $f : M \to \mathbb{H} = \mathbb{R}^4$, where f is the quotient of any two holomorphic sections of L.

1.3 Abelian differentials

A second classical method to construct meromorphic functions on a Riemann surface M is to use Abelian differentials, i.e. integrals of meromorphic 1-forms. In the quaternionic theory there is no good analog of the canonical bundle K. On the other hand, also in the complex case 1-forms often arise as products of sections of two line bundles E and KE^{-1}. Notably, this is the case in situations where the Riemann-Roch theorem is applied. This setup carries over perfectly to the quaternionic case, including the Riemann-Roch theorem itself.

We show that for each holomorphic quaternionic line bundle L there exists a certain holomorphic quaternionic line bundle KL^{-1} such that any holomorphic section ψ of L can be multiplied with any holomorphic section ϕ of KL^{-1}, the product being a closed $\mathbb{H}$–valued 1-form (ψ, ϕ) that locally integrates to a conformal map f into $\mathbb{H}$:

$$df = (\psi, \phi)\,. \tag{3}$$

In the case where KL^{-1} is isomorphic to L itself, we call L a spin bundle. If ψ is a nowhere vanishing holomorphic section of a spin bundle then

$$df = (\psi, \psi)$$

defines a conformal immersion into $\mathbb{R}^3$. This construction is in fact a more intrinsic version of the "Weierstrass-representation for general surfaces in 3-space" that has received much attention in the recent literature [5], just as (3), when expressed in coordinates, gives a representation for surfaces in $\mathbb{R}^4$. The Hopf field Q mentioned above can be identified as the "Dirac-potential" or "mean curvature half-density" of the surface f:

$$Q = \frac{1}{2} H |df|\,.$$

Here H is the mean curvature, and $|df|$ is the square root of the induced metric.

1.4 Applications

The only geometric application discussed in some detail in this paper is a rigidity theorem for spheres: if $f, g : S^2 \to \mathbb{R}^3$ are two conformal immersions which are not congruent up to scale but have the same mean curvature half-density, then

$$\int H^2 \geq 16\pi\,.$$

This inequality is sharp.

Many other applications, to be discussed in a more elaborate future paper [1], will concern the geometry of Willmore surfaces (critical points of the Willmore functional) both in $\mathbb{R}^3$ and $\mathbb{R}^4$. Moreover, rudiments of an "algebraic geometry of holomorphic curves" in quaternionic projective space $\mathbb{HP}^n$ can be developed.

2 Conformal surfaces: the standard example

Let M be a Riemann surface and $f : M \to \mathbb{R}^3$ a smooth map. The map f is a conformal immersion if

i. $df(v)$ is perpendicular to $df(Jv)$ for any tangent vector v, where J is the complex structure on M, and

ii. $|df(v)| = |df(Jv)| \neq 0$ for $v \neq 0$.

If $N : M \to S^2$ is the oriented unit normal to f, then the conformality condition can be rephrased as

$$df(Jv) = N \times df(v)\,.$$

To see the similarity with complex function theory, we rewrite this condition using quaternions $\mathbb{H} = \mathbb{R} \oplus \mathrm{Im}\,\mathbb{H}$ [4]. We will always think of $\mathbb{R}^3 = \mathrm{Im}\,\mathbb{H}$ as the imaginary quaternions. If $x, y \in \mathbb{R}^3$ then

$$xy = - < x,\, y > + x \times y\,.$$

With the notation (1) the conformality condition for f becomes (2). For the rest of the article we will take this to be the defining equation for conformality, also in the case of maps (not necessarily immersions) into $\mathbb{R}^4$:

DEFINITION 2.1. A map $f : M \to \mathbb{R}^4 = \mathbb{H}$ is *conformal* if there exists a map $N : M \to \mathbb{H}$ such that $N^2 = -1$ and

$$*df = N df\,.$$

At immersed points this is equivalent to the usual notion of conformality, and f determines N uniquely. If f is $\mathbb{R}^3$–valued then N is the oriented unit normal, but otherwise N is not normal to f. We will call N the *left normal* to f. Moreover, if f is conformal so is its Moebius inversion f^{-1}, with left normal $f^{-1}Nf$. Thus, the above definition is Moebius invariant and hence defines conformality of maps $f : M \to S^4 = \mathbb{HP}^1 = \mathbb{H} \cup \{\infty\}$.

3 HOLOMORPHIC QUATERNIONIC LINE BUNDLES

A *quaternionic line bundle* L over a base manifold is a smooth rank 4 real vector bundle whose fibers have the structure of 1-dimensional quaternionic right vector spaces varying smoothly over the base. Two quaternionic line bundles L_1 and L_2 are *isomorphic* if there exists a smooth bundle isomorphism $A : L_1 \to L_2$ that is quaternionic linear on each fiber. We adopt the usual notation $\mathrm{Hom}_{\mathbb{H}}(L_1, L_2)$ and $\mathrm{End}_{\mathbb{H}}(L) = \mathrm{Hom}_{\mathbb{H}}(L, L)$, etc., for the spaces of quaternionic linear maps.

The zero section of a quaternionic line bundle over an oriented surface has codimension 4, so that transverse sections have no zeros. Thus any quaternionic line bundle over a Riemann surface M is smoothly isomorphic to $M \times \mathbb{H}$.

3.1 COMPLEX QUATERNIONIC LINE BUNDLES

EXAMPLE. Given a conformal map $f : M \to \mathbb{H}$ with left normal N, the quaternionic line bundle $L = M \times \mathbb{H}$ also has a complex structure $J : L \to L$ given by $J(\psi) = N\psi$ for $\psi \in L$.

We make this additional complex structure part of our theory:

DEFINITION 3.1. A *complex quaternionic line bundle* over a base manifold is a pair (L, J) where L is a quaternionic line bundle and $J \in \mathrm{End}_{\mathbb{H}}(L)$ is a quaternionic linear endomorphism such that $J^2 = -1$.

Put differently, a complex quaternionic line bundle is a rank two left complex vector bundle whose complex structure is compatible with the right quaternionic structure. Two complex quaternionic line bundles are isomorphic if the quaternionic linear isomorphism is also left complex linear.

The dual of a quaternionic line bundle L,

$$L^{-1} = \{\omega : L \to \mathbb{H}\,;\, \omega \text{ quaternionic linear}\}\,,$$

has a natural structure of a left quaternionic line bundle via $(\lambda\omega)(\psi) = \lambda\omega(\psi)$ for $\lambda \in \mathbb{H}$, $\omega \in L^{-1}$ and $\psi \in L$. Using conjugation, we can regard L^{-1} as a right quaternionic line bundle, $\omega \cdot \lambda = \bar{\lambda}\omega$. If L has a complex structure then the complex structure on L^{-1} is given by

$$J\omega := \omega \circ J\,,$$

so that L^{-1} is also complex quaternionic.

Any complex quaternionic line bundle L can be tensored on the left by a complex line bundle E, yielding the complex quaternionic line bundle EL. On a Riemann surface M we have the canonical and anti-canonical bundles K and $\bar{K}$. It is easy to see that

$$KL = \{\omega : TM \to L\,;\, *\omega = J \circ \omega\}\,,$$

and

$$\bar{K}L = \{\omega : TM \to L\,;\, *\omega = -J \circ \omega\}\,.$$

In this way, we have split the quaternionic rank 2 bundle $\mathrm{Hom}_{\mathbb{R}}(TM, L)$, which has a left complex structure given by $*$, as a direct sum $KL \oplus \bar{K}L$ of two complex quaternionic line bundles.

If E is a complex line bundle, then $L_E := E \oplus E$ becomes a complex quaternionic line bundle with $J(\psi_1, \psi_2) = (i\psi_1, i\psi_2)$ and right quaternionic structure given by

$$(\psi_1, \psi_2)i = (i\psi_1, -i\psi_2)\,, \qquad (\psi_1, \psi_2)j = (-\psi_2, \psi_1)\,.$$

Conversely, for a given complex quaternionic line bundle (L, J) we let $E_L := \{\psi \in L\,;\, J\psi = \psi i\}$ be the $+i$ eigenspace of J. Then $E \subset L$ is a complex line subbundle and $E_L \oplus E_L$ is isomorphic to L. This leads to the following

THEOREM 3.1. *The above correspondences*

$$E \longmapsto L_E\,, \qquad L \longmapsto E_L$$

give a bijection between isomorphism classes of complex line bundles and isomorphism classes of complex quaternionic line bundles. This bijection is equivariant with respect to left tensoring by complex line bundles and respects dualization.

DEFINITION 3.2. The degree of a complex quaternioninc line bundle L over a compact Riemann surface is the degree of the underlying complex line bundle E_L, i.e. $\deg L := \deg E_L$.

On a compact Riemann surface (isomorphism classes of) complex line bundles are characterized by their degrees. Thus, complex quaternionic line bundles also are characterized by their degrees. Given a trivializing section ψ of L we have $J\psi = \psi N$ for some $N : M \to S^2 \subset \operatorname{Im}\mathbb{H}$, $N^2 = -1$, and one easily checks that $\deg L = \deg N$.

3.2 Holomorphic quaternionic line bundles

Definition 3.3. Let (L, J) be a complex quaternionic line bundle over a Riemann surface M and let $\Gamma(L)$ denote the smooth sections of L. A *holomorphic structure* on L is given by a quaternionic linear map

$$D : \Gamma(L) \to \Gamma(\bar{K}L)$$

satisfying

$$D(\psi\lambda) = (D\psi)\lambda + \frac{1}{2}(\psi d\lambda + J\psi * d\lambda) \tag{4}$$

for $\lambda : M \to \mathbb{H}$.

The quaternionic linear subspace $\ker D \subset \Gamma(L)$ is called the space of *holomorphic sections* and is denoted by $H^0(L)$.

One can check that the $\bar{K}L$–part of a quaternionic connection on L gives a holomorphic structure D, which may be used as motivation for the above formula.

Any complex holomorphic structure $\bar{\partial}$ on the underlying complex line bundle E_L is an example of a holomorphic structure $D = \bar{\partial} \oplus \bar{\partial}$. These holomorphic structures on L are characterized by the condition that D and J commute. The failure to commute is measured by

$$Q = \frac{1}{2}(D + JDJ)$$

which is a section of $T^*M \otimes \operatorname{End}_{\mathbb{H}}(L)$. Now

$$\operatorname{End}_{\mathbb{H}}(L) = \operatorname{End}_+(L) \oplus \operatorname{End}_-(L)$$

splits into linear maps commuting and anti-commuting with J. The former is a trivial complex bundle with global sections Id and J. The latter is a non-trivial complex line bundle isomorphic to $\overline{E_L^{-1}} \otimes E_L$. Since Q anti-commutes with J and satisfies $*Q = -J \circ Q$ we see that Q is a section of the complex line bundle $\bar{K}\operatorname{End}_-(L)$. We call Q the *Hopf field* of the holomorphic quaternionic line bundle L. Thus any holomorphic structure D is uniquely decomposed into

$$D = \bar{\partial} + Q$$

with $\bar{\partial}$ commuting with J. Vanishing of the Hopf field Q characterizes the usual complex holomorphic structures. Two quaternionic holomorphic line bundles are isomorphic if there is an isomorphism of complex quaternionic line bundles which intertwines the respective holomorphic structures. On a compact Riemann surface this implies that the underlying complex holomorphic structures are isomorphic

and that the Hopf fields are related up to a constant phase $u \in S^1$. Thus the moduli space of quaternionic holomorphic structures fibers over the Picard group of M. The fiber $\Gamma(\bar{K}\mathrm{End}_-(L_E))/S^1$ over $(E, \bar{\partial})$ is given by the Hopf fields.

A global invariant of the quaternionic holomorphic line bundle L is obtained by integrating the length of the Hopf field Q: we define the density $|Q|^2$ by

$$Q_v \circ Q_v = -|Q|^2(v) Id\,, \qquad v \in TM\,,$$

where we identify J^M–invariant quadratic forms with 2-forms on M. The *Willmore functional* of D is the L^2–norm of the Hopf field

$$||Q||^2 = \int |Q|^2\,.$$

The vanishing of $||Q||$ characterizes the complex holomorphic theory.

EXAMPLE. We have already seen that a conformal map $f : M \to \mathbb{H}$ with left normal $N : M \to S^2$ induces the complex quaternionic bundle $L = M \times \mathbb{H}$ with $J\psi = N\psi$. We define the *canonical* holomorphic structure D on L to be the one for which the constant sections are holomorphic, i.e. D is characterized by $D(1) = 0$. Any other section of L is of the form $\psi = 1\lambda$ for some $\lambda : M \to \mathbb{H}$, and (4) implies that ψ is holomorphic iff

$$*d\lambda = Nd\lambda\,.$$

Thus the holomorphic sections of L are precisely the conformal maps with the same left normal as f. In particular, $\dim H^0(L) \geq 2$.

The Moebius invariance of the holomorhic structure follows since f^{-1} induces an isomorphic holomorphic structure on $M \times \mathbb{H}$. Thus we have assigned to each conformal map into $\mathbb{HP}^1 = S^4$ a quaternionic holomorphic line bundle with at least two holomorphic sections.

The Hopf field for this holomorphic structure is $Q = \frac{1}{4}N(dN + *dN)$, and

$$|Q|^2 = \frac{1}{4}(|H|^2 - K - K^\perp)|df|^2\,,$$

where H is the mean curvature vector of f and $K^\perp$ is the curvature of the normal bundle. We see that $|Q|^2$ is a Moebius invariant density, which is consistent with the Moebius invariance of our setup. Thus, the Willmore energy of our holomorphic structure, $||Q||^2 = \int |Q|^2 = W(f)$, is (up to topological constants) just the Willmore energy of f.

So far we have seen how a conformal map induces a holomorphic line bundle with at least two holomorphic sections. As in the classical complex theory we have the converse construction, i.e. all conformal maps into S^4 arise as quotients of holomorphic sections.

EXAMPLE. Let $L \to M$ be a quaternionic holomorphic line bundle and assume that $\dim H^0(L) \geq 2$ with ψ, ϕ holomorphic sections such that ψ has no zeros. Then $J\psi = \psi N$ for some $N : M \to S^2$. We define $f : M \to \mathbb{H}$ by

$$\phi = \psi f\,,$$

then (4) implies that $*df = Ndf$, i.e. f is conformal with left normal N.

An interesting special case comes from conformal maps with $Q = 0$. It can be shown that they are superconformal in the sense that their curvature ellipse is a circle. Since $Q = 0$, superconformal maps are critical for the Willmore energy and thus Willmore surfaces in S^4. In case f is $\mathbb{R}^3$–valued $Q = 0$ simply means that f is a conformal map into the 2-sphere. The superconformal maps all arise as projections from holomorphic maps into the twistor space $\mathbb{CP}^3$ over S^4 and have been studied by various authors [2, 3]. In our theory these maps arise as quaternionic quotients of holomorphic sections of (doubled) complex holomorphic line bundles.

In the above construction the structure of zeros of quaternionic holomorphic sections becomes important. Applying a result of Aronszajn we can show

THEOREM 3.2. *Let ψ be a non-trivial holomorphic section of a quaternionic holomorphic line bundle L over a Riemann surface M. Then the zeros of ψ are isolated and, if z is a centered local coordinate near a zero $p \in M$,*

$$\psi = z^k \phi + O(|z|^{k+1})$$

where ϕ is a local nowhere vanishing section of L. The integer k and the value $\phi(p) \in L_p$ are well-defined independent of choices. We define the order *of the zero p of ψ by $ord_p \psi = k$.*

We conclude this section with a degree formula:

THEOREM 3.3. *Let ψ be a non-trivial section of a quaternionic holomorphic line bundle L over a compact Riemann surface M. Then*

$$\pi \deg L + ||Q||^2 \geq \pi \sum_{p \in M} ord_p \psi \,. \tag{5}$$

In contrast to the complex holomorphic case, where negative degree bundles do not have holomorphic sections, we see that in the quaternionic theory the Willmore energy of the bundle compensates for this failure and we still can have holomorphic sections.

Equality in (5) is attained by holomorphic bundles L^{-1} where $L = E \oplus E$ is a doubled complex holomorphic bundle E and L has a nowhere vanishing meromorphic section ψ. The holomorphic structure on L^{-1} then is obtained by defining ψ^{-1} to be holomorphic.

We conjecture the following lower bound for the Willmore energy on holomorphic line bundles over the 2-sphere: let $n = \dim H^0(L)$ and $d = \deg L$ then

$$\frac{1}{\pi} ||Q||^2 \geq n^2 - n(d+1) \,. \tag{6}$$

Examples are known where equality holds. Using the degree formula we can prove (6) under certain non-degeneracy assumptions [1]. For $d = -1$, the case of spin bundles (see the next section), this estimate has been conjectured by Taimanov [6].

4 ABELIAN DIFFERENTIALS

DEFINITION 4.1. A *pairing* between two complex quaternionic line bundles L and $\tilde{L}$ over M is a nowhere vanishing real bilinear bundle map $(\ ,\ ) : L \times \tilde{L} \to T^*M \otimes \mathbb{H}$ satisfying

$$(\psi\lambda, \phi\mu) = \bar{\lambda}(\psi, \phi)\mu$$
$$*(\psi, \phi) = (J\psi, \phi) = (\psi, J\phi)$$

for all $\lambda, \mu \in \mathbb{H}$, $\psi \in L$, $\phi \in \tilde{L}$.

A pairing between L and $\tilde{L}$ is actually the same as an isomorphism of complex quaternionic line bundles $\tilde{L} \to KL^{-1}$, given as $\phi \mapsto \alpha$, where

$$\alpha_X(\psi) = (\psi, \phi)(X)\,.$$

If ω is a 1-form on M with values in L and ϕ is a section of L, then we define an L-valued 2-form $(\omega \wedge \phi)$ as

$$(\omega \wedge \phi)(X, Y) = (\omega(X), \phi)(Y) - (\omega(Y), \phi)(X)\,.$$

Similarly, for $\psi \in \Gamma(L)$ and η a 1-form with values in $\tilde{L}$, we set

$$(\psi \wedge \eta)(X, Y) = (\psi, \eta(X))(Y) - (\psi, \eta(Y))(X)\,.$$

LEMMA 4.1. *For each $\omega \in \bar{K}End_-(L)$ there is a unique $\bar{\omega} \in \bar{K}End_-(\tilde{L})$ such that*

$$(\omega\psi \wedge \phi) + (\psi \wedge \bar{\omega}\phi) = 0$$

for all $\psi \in \Gamma(L)$, $\phi \in \Gamma(\tilde{L})$. The map $\omega \mapsto \bar{\omega}$ is complex antilinear:

$$\overline{J\omega} = -J\bar{\omega}\,.$$

THEOREM 4.2. *If two complex quaternionic line bundles L and $\tilde{L}$ are paired, then for any holomorphic structure D on L there is a unique holomorphic structure $\tilde{D}$ on $\tilde{L}$ such that for each $\psi \in \Gamma(L)$, $\phi \in \Gamma(\tilde{L})$ we have*

$$d(\psi, \phi) = (D\psi \wedge \phi) + (\psi \wedge \tilde{D}\phi)\,.$$

The Hopf fields Q and $\tilde{Q}$ of D and $\tilde{D}$ are conjugate:

$$\tilde{Q} = \bar{Q}\,.$$

Thus, a holomorphic structure on L determines a unique holomorphic structure on KL^{-1} such that L and KL^{-1} become paired holomorphic bundles. In this situation, the Riemann-Roch theorem is true in the familiar form of the theory of complex line bundles: on compact Riemann surfaces of genus g we have

$$\dim H^0(L) - \dim H^0(KL^{-1}) = \deg(L) - g + 1\,.$$

Theorem 4.2 suggests a way to construct conformal immersions $f : M \to \mathbb{R}^4 = \mathbb{H}$. If L and $\tilde{L}$ are paired holomorphic bundles and $\psi, \phi \in H^0(L)$ are both nowhere vanishing sections, then (ψ, ϕ) is a closed 1-form that integrates to a conformal immersion into $\mathbb{R}^4$, possibly with translational periods. In fact, this construction is completely general:

THEOREM 4.3. *Let $f : M \to \mathbb{H}$ be a conformal immersion. Then there exist paired holomorphic quaternionic line bundles L, $\tilde{L}$ and nowhere vanishing sections $\psi \in H^0(L)$, $\phi \in H^0(\tilde{L})$ such that*

$$df = (\psi, \phi)\,. \tag{7}$$

L, $\tilde{L}$, ψ and ϕ are uniquely determined by f up to isomorphism.

In the setup of the theorem choose locally non-vanishing sections $\hat{\psi} \in \Gamma(L)$, $\hat{\phi} \in \Gamma(\tilde{L})$ satisfying $\bar{\partial}\hat{\psi} = 0$, $\bar{\partial}\hat{\phi} = 0$, $J\hat{\psi} = -\hat{\psi}i$, $J\hat{\phi} = \hat{\phi}i$. Then there is a $\mathbb{R} \oplus \mathbb{R}i$–valued coordinate chart z on M satisfying

$$dz = (\hat{\psi}, \hat{\phi})\,.$$

We can write

$$\psi = \hat{\psi}(\psi_1 + \psi_2 j) \qquad \phi = \hat{\phi}(\phi_1 + \phi_2 j)$$

with $\mathbb{R} \oplus \mathbb{R}i$–valued functions ψ_α, ϕ_α. Expanding (7) we obtain a generalization of the Weierstrass representation of surfaces in $\mathbb{R}^3$ [5] to surfaces in $\mathbb{R}^4$. The equations $(\bar{\partial} + Q)\psi = 0$ and $(\bar{\partial} + \tilde{Q})\phi = 0$ unravel to Dirac equations for ψ_α and ϕ_α.

DEFINITION 4.2. A holomorphic line bundle Σ over M is called a *spin bundle* if there exists a pairing of Σ with itself such that the second holomorphic structure on Σ provided by Theorem 4.2 coincides with the original one.

As a direct consequence of the definition of a pairing we obtain in the case of spin bundles the relation

$$(\phi, \psi) = -\overline{(\psi, \phi)}\,.$$

Therefore, for any holomorphic section ψ of a spin bundle Σ the equation

$$df = (\psi, \psi)$$

defines a conformal map into $\mathbb{R}^3 = \operatorname{Im}\mathbb{H}$, possibly with translational periods. This is in fact a coordinate-free version of the Weierstrass representation for surfaces in $\mathbb{R}^3$ [5], which could be obtained by a calculation similar to the one given above for $\mathbb{R}^4$. We now show that the "Dirac potential" $H|df|$ featured in this representation can be identified with the Hopf field Q of Σ.

For a spin bundle Σ the map $\bar{K}\mathrm{End}_-(\Sigma) \ni \omega \mapsto \bar{\omega}$ puts a real structure on $\bar{K}\mathrm{End}_-(\Sigma)$ and therefore allows us to define a real line bundle

$$R = \operatorname{Re}(\bar{K}\mathrm{End}_-(\Sigma)) = \{\omega \in \bar{K}\mathrm{End}_-(\Sigma)\,;\, \omega = \bar{\omega}\}\,.$$

We now show that R can be identified with the real line bundle $\mathcal{D}^{-1/2}$ of half densities over M. A half-density U is a function on the tangent bundle TM which is of the form

$$U(X_p) = \rho(p)\sqrt{g(X_p, X_p)}$$

where $\rho \in C^\infty(M)$ and g is a Riemannian metric compatible with the given conformal structure. For each $\psi \in \Gamma(\Sigma)$ the function $X \mapsto |(\psi, \psi)(X)|$ is a half-density.

On the other hand, it can be checked that for each $\psi \in \Gamma(\Sigma)$ we can define a section ω_ψ of R as

$$\omega_\psi(\phi) = \psi(J\psi, \phi)\,.$$

There is a canonical isomorphism $R \to \mathcal{D}^{1/2}$ which takes ω_ψ to $|(\psi, \psi)|$ for all $\psi \in \Sigma$.

Theorem 4.4. *Let ψ be a holomorphic section of a spin bundle Σ over M. Then there is a conformal immersion $f : \tilde{M} \to \mathbb{R}^3$ on the universal cover of M with only translational periods such that*

$$df = (\psi, \psi)\,.$$

Identifying the half-density $|df|$ as explained above with a section of $\bar{K}End_-(\Sigma)$, the mean curvature of f is given in terms of the Hopf field Q of Σ as

$$Q = \frac{1}{2} H |df|\,.$$

We conclude by indicating a proof of the rigidity theorem for spheres stated in Section 1.4. The hypotheses imply that in the situation of the theorem above Σ has a 2–dimensional space of holomorphic sections. Since in the case at hand the conjecture (6) has been proven, we take $n = 2$ and $d = \deg \Sigma = -1$ and obtain

$$\int H^2 |df|^2 = 4 \int |Q|^2 \geq 16\pi\,.$$

References

[1] D. Ferus, F. Pedit, U. Pinkall. *Quaternionic line bundles and differential geometry of surfaces*. In preparation.

[2] T. Friedrich. *The geometry of t-holomorphic surfaces in S^4*. Math. Nachr. 137, 49–62 (1988).

[3] S. Montiel. *Spherical Willmore surfaces in the four-sphere*. Preprint, Granada 1998.

[4] G. Kamberov, F. Pedit, U. Pinkall. *Bonnet pairs and isothermic surfaces*. Duke Math. J., Vol. 92, No. 3, 637–643 (1998).

[5] B. G. Konopolchenko. *Induced surfaces and their integrable dynamics*. Stud. Appl. Math., Vol. 96, 9–51 (1996).

[6] I. A. Taimanov. *The Weierstrass representation of spheres in $\mathbf{R}^3$, the Willmore numbers, and soliton spheres*. Preprint No. 302, SFB 288, TU Berlin 1998.

Franz Pedit
Department of Mathematics
University of Massachusetts
Amherst, MA 01003
USA
franz@gang.umass.edu

Ulrich Pinkall
Fachbereich Mathematik
TU Berlin
Strasse des 17. Juni 136
10623 Berlin
Germany
pinkall@math.tu-berlin.de

Geometry on the Group of Hamiltonian Diffeomorphisms

Leonid Polterovich

Abstract. The group of Hamiltonian diffeomorphisms $\mathrm{Ham}(M,\Omega)$ of a symplectic manifold (M,Ω) plays a fundamental role both in geometry and classical mechanics. For a geometer, at least under some assumptions on the manifold M, this is just the connected component of the identity in the group of all isometries of the symplectic structure Ω. From the point of view of mechanics, $\mathrm{Ham}(M,\Omega)$ is the group of all admissible motions. It was discovered by H. Hofer ([H1], 1990) that this group carries a natural Finsler metric with a non-degenerate distance function. Intuitively speaking, the distance between a given Hamiltonian diffeomorphism f and the identity transformation is equal to the minimal amount of energy required in order to generate f. This new geometry has been intensively studied for the past 8 years in the framework of modern symplectic topology. It serves as a source of refreshing problems and gives rise to new methods and notions. Also, it opens up the intriguing prospect of using an alternative geometric intuition in Hamiltonian dynamics. In the present note we discuss these developments.

1991 Mathematics Subject Classification: 58Dxx (Primary) 58F05 53C15 (Secondary)

1. The group of Hamiltonian diffeomorphisms. Let (M,Ω) be a connected symplectic manifold without boundary. Every smooth compactly supported function F on $M\times[0;1]$ defines a Hamiltonian flow $f_t : M\to M$. This flow is generated by a time-dependent vector field ξ_t on M which satisfies the point-wise linear algebraic equation $\Omega(.,\xi_t) = dF_t(.)$, where $F_t(x)$ stands for $F(x,t)$. Symplectomorphisms f_t arising in this way are called *Hamiltonian diffeomorphisms*. Hamiltonian diffeomorphisms form an infinite-dimensional Lie group $\mathrm{Ham}(M,\Omega)$. When $H^1_{\mathrm{comp}}(M,\mathbf{R}) = 0$ this group coincides with $\mathrm{Symp}_0(M,\Omega)$ - the identity component of the group of all symplectomorphisms in the strong Whitney topology. In general the quotient group $\mathrm{Symp}_0(M,\Omega)/\mathrm{Ham}(M,\Omega)$ is non-trivial but "quite small" [Ba]. The Lie algebra $\mathcal{A}$ of $\mathrm{Ham}(M,\Omega)$ consists of all smooth functions on M which satisfy the following normalization condition. Namely when M is open $F\in\mathcal{A}$ iff F is compactly supported, and when M is closed $F\in\mathcal{A}$ iff F has the zero mean with respect to the canonical measure on M induced by Ω. With this normalization different functions from $\mathcal{A}$ generate different Hamiltonian

vector fields. The Lie bracket on $\mathcal{A}$ is the Poisson bracket, and the adjoint action of $\mathrm{Ham}(M,\Omega)$ on $\mathcal{A}$ is the standard action of diffeomorphisms on functions.

2. HOFER'S METRIC. Consider the L_∞-norm $||F|| = \max_M F - \min_M F$ on $\mathcal{A}$. This norm is invariant under the adjoint action, and thus defines a biinvariant Finsler metric on $\mathrm{Ham}(M,\Omega)$. This Finsler metric determines in the standard way a length structure, and a pseudo-distance ρ on the group. More explicitly, let $\{f_t\},\ t \in [0;1]$ be a path of Hamiltonian diffeomorphisms with $f_0 = \phi$ and $f_1 = \psi$. Let $F(x,t)$ be its *normalized* Hamiltonian function, that is $F(.,t) \in \mathcal{A}$ for all t. Then

$$\mathrm{length}\{f_t\} = \int_0^1 ||F(.,t)||dt,$$

and $\rho(\phi,\psi) = \inf \mathrm{length}\{f_t\}$, where the infimum is taken over all smooth paths $\{f_t\}$ which join ϕ and ψ.

A non-trivial fact is that the pseudo-distance ρ is non-degenerate, that is $\rho(\phi,\psi) \neq 0$ for $\phi \neq \psi$ (this was proved in [H1] for $\mathbf{R}^{2n}$, then extended in [P1] for some other symplectic manifolds, and finally confirmed in [LM1] in full generality). Note that the construction above goes through for any other norm on the Lie algebra which is invariant under the adjoint action, for instance for the L_p-norm. However for all $1 \leq p < \infty$ the corresponding pseudo-distance is degenerate [EP].

Interestingly enough, the quantity $\rho(\mathrm{id},\phi)$ can be interpreted as the distance between a point and a subset in a linear normed space [P7]. Consider the space $\mathcal{F}$ of all smooth compactly supported functions F on $M \times S^1$ such that $F(.,t) \in \mathcal{A}$ for all $t \in S^1 = \mathbf{R}/\mathbf{Z}$. For $F \in \mathcal{F}$ denote by ϕ_F the time-one-map of the Hamiltonian flow generated by F. Every Hamiltonian diffeomorphism can be expressed in this way. Let $\mathcal{H} \subset \mathcal{F}$ be the subset of all functions H which generate loops of Hamiltonian diffeomorphisms, that is $\phi_H = \mathrm{id}$. Introduce a norm on $\mathcal{F}$ by $|||F||| = \max_t ||F(.,t)||$. It is easy to show that

$$(2.A) \qquad \rho(\mathrm{id},\phi_F) = \inf_{H \in \mathcal{H}} |||F - H|||.$$

Thus the set $\mathcal{H}$ carries a lot of information about Hofer's geometry.

We complete this section with the following open problem in the very foundation of Hofer's geometry [EP]. It is quite natural (see section 7 below) to consider the "maximum" and the "minimum" parts of Hofer's length structure separately. Namely set $\mathrm{length}_+\{f_t\} = \int_0^1 \max_x F(x,t)dt$ and $\mathrm{length}_-\{f_t\} = \int_0^1 -\min_x F(x,t)dt$, and define $\rho_+(\phi,\psi)$ and $\rho_-(\phi,\psi)$ as the infimum of positive and negative lengths respectively over all paths $\{f_t\}$ with $f_0 = \phi$ and $f_1 = \psi$. Clearly, $\rho(\phi,\psi) \geq \rho_-(\phi,\psi) + \rho_+(\phi,\psi)$. In fact, in all examples known to me *the equality* holds. It would be interesting either to prove this, or to find a counterexample.

3. DISPLACEMENT ENERGY. Consider any norm on $\mathcal{A}$ invariant under the adjoint action, and denote by ρ' the corresponding pseudo-distance. For a subset U of M denote by G_U the set of all Hamiltonian diffeomorphisms f such that $f(U) \cap U = \emptyset$. Define *the displacement energy* of U as $\rho'(\mathrm{id}, G_U)$. We use the convention that the displacement energy equals $+\infty$ when G_U is empty. Clearly this is a symplectic

invariant. It takes strictly positive values on non-empty open subsets if and only if the pseudo-metric ρ' is non-degenerate [EP]. Denote by $e(U)$ the displacement energy with respect to Hofer's metric.

Example 3.A Every symplectic manifold of dimension $2n$ admits a symplectic embedding of a standard $2n$-dimensional ball of a sufficiently small radius r. The supremum of πr^2 where r runs over such the embeddings is called *Gromov's width of the symplectic manifold*. Hofer showed [H1] that for every open subset U of the standard symplectic vector space $\mathbf{R}^{2n}$ holds $e(U) \geq \text{width}(U)$. Later on it was proved in [LM1] that $e(U) \geq \frac{1}{2}\text{width}(U)$ for every open subset U of an arbitrary symplectic manifold. Conjecturally, in the general case the factor $\frac{1}{2}$ can be removed.

Example 3.B Consider the cotangent bundle $\theta : T^*T^n \to T^n$ with a *twisted* symplectic structure $dp \wedge dq + \theta^*\sigma$, where σ is a closed 2-form on T^n. Such structures arise in the theory of magnetic fields. Denote by $Z \subset T^*T^n$ the zero section. If $\sigma = 0$ then $f(Z) \cap Z \neq \emptyset$ for every Hamiltonian diffeomorphism f (this is the famous Arnold's Lagrangian intersections conjecture proved by Chaperon, Hofer and Laudenbach-Sikorav, see [MS]). Thus $e(Z) = +\infty$. However if $\sigma \neq 0$ then Z admits a nowhere tangent Hamiltonian vector field [P2], and thus $e(Z) = 0$.

4. A paradox of Hofer's geometry. What does the metric space $\text{Ham}(M, \Omega)$ look like? Here we present two results which intuitively contradict one another, and no convincing explanation is known at present. The first one is the following *C^1-flatness phenomenon*.

Theorem 4.A [BP1]. *There exists a C^1-neighbourhood $\mathcal{E}$ of the identity in $\text{Ham}(\mathbf{R}^{2n})$ and a C^2-neighbourhood $\mathcal{C}$ of zero in $\mathcal{A}$ such that $(\mathcal{E}, \rho)$ is isometric to $(\mathcal{C}, ||\ ||)$.*

The isometry takes every C^1-small Hamiltonian diffeomorphism from $\mathcal{E}$ to its classical generating function. Some generalizations can be found in [LM2].

The second result, due to J.-C. Sikorav [S] states that *every one-parameter subgroup of $\text{Ham}(\mathbf{R}^{2n})$ remains a bounded distance from the identity* (see discussion in §6 below). This can be interpreted as a "positive curvature type effect".

It sounds likely that in order to resolve this paradox one should understand properly the interrelation between the topology on $\text{Ham}(M, \Omega)$ which comes from Hofer's metric, and the smooth structure on the group. For instance, paths which are continuous in the metric topology can be non-continuous in the usual sense, and there is no satisfactory way to think about them. In what follows we restrict ourselves to *smooth* paths, homotopies, etc.

5. Geodesics. The C^1-flatness phenomenon above serves as the starting point for the theory of geodesics of Hofer's metric. Indeed, at least on small time intervals the geodesics should behave as the ones in the linear normed space $(\mathcal{A}, ||\ ||)$. This leads to the following definition [BP1]. Consider a smooth path of Hamiltonian diffeomorphisms of (M, Ω) generated by a normalized Hamiltonian function $F(x, t)$. Assume that $||F(., t)|| \neq 0$ for all t. The path is called *quasi-autonomous* if there exist two (time-independent!) points x_+ and x_- on M such that for all t the function $F(., t)$ attains its maximal and minimal values at x_+ and x_- respectively. For instance, every one-parameter subgroup is quasi-autonomous. A path

of Hamiltonian diffeomorphisms is called *a minimal geodesic* if each of its segments minimizes length in the homotopy class of paths with fixed end points. It turns out that every minimal geodesic is quasi-autonomous [LM2]. However the converse is not true in general (see Sikorav's result above). In [H2] Hofer discovered a surprising link between minimality of paths on the group of Hamiltonian diffeomorphisms and closed orbits of corresponding Hamiltonian flows. Numerous further results in this direction (see [HZ],[BP1],[Si1],[LM2] and [Sch]) serve as a motivation for the following conjecture. A closed orbit of period c of a (time-dependent) flow $\{f_t\}$ with $f_0 = \mathrm{id}$ is a piece of the trajectory of a point $x \in M$ on a time interval $[0; c]$ such that $x = f_c x$. A closed orbit is called constant if it corresponds to a fixed point of the flow, that is $f_t x = x$ for all t.

CONJECTURE 5.A. *Let* $\{f_t\}$, $t \in [0; T]$, $f_0 = \mathrm{id}$ *be a quasi-autonomous path of Hamiltonian diffeomorphisms. Assume that the flow* $\{f_t\}$ *has no contractible non-constant closed orbits of period less than* T. *Then this path is a minimal geodesic.*

As an immediate consequence one gets that one-parameter subgroups should be minimal on short time intervals. In 8.A and 9.A below we describe a minimality-breaking mechanism on large time intervals which together with 5.A allows us to detect non-trivial closed orbits. In 9.B we give an example of an infinite minimal geodesic. The study of the breaking of minimality is still far from being completed. Another step in this direction was made in the framework of the theory of conjugate points (see [U],[LM2]) which deals with the local behavior of the length functional under small deformations of quasi-autonomous paths, and where an infinitesimal version of 5.A plays a crucial role.

6. DIAMETER. Here we discuss the following conjecture.

CONJECTURE 6.A. *The diameter of* $\mathrm{Ham}(M, \Omega)$ *with respect to Hofer's metric is infinite.*

The conjecture is established at present for a number of manifolds (see [LM2],[P7],[Sch]). We shall illustrate the methods in the case when (M, Ω) is a closed oriented surface endowed with an area form. In the case when the genus of M is at least 1, the conjecture was proved in [LM2] as follows. One can produce a Hamiltonian flow on M whose lift to the universal cover displaces a disc of an arbitrarily large area. For instance, take a flow which is the standard rotation in a small neighbourhood of a non-contractible curve on M. Inequality 3.A implies that such a flow goes arbitrarily far away from the identity. There is also a different proof [Sch] which is based on the analysis of closed orbits (cf. 5.A).

These methods do not work when M is the 2-sphere. This case was treated in [P7] as follows. Consider the set $\mathcal{H}$ of all 1-periodic normalized Hamiltonians which generate the identity map (see §2). Let L be an equator of S^2.

THEOREM 6.B [P7]. *For every* $H \in \mathcal{H}$ *there exist* $x \in L$ *and* $t \in S^1$ *such that* $H(x, t) = 0$.

Choose now an arbitrary large number c, and a time-independent normalized Hamiltonian function F such that $F(x) \geq c$ for all $x \in L$. It follows from (2.A)

and 6.B above that $\rho(\mathrm{id}, \phi_F) \geq c$, and thus *the diameter of* $\mathrm{Ham}(S^2)$ *is infinite.* In particular, on S^2 (in contrast to $\mathbf{R}^{2n}$, see §4) there are unbounded one-parameter subgroups. As a by-product of this argument we get that *there exists a sequence of Hamiltonian diffeomorphisms of* S^2 *which converges to the identity in the* C^0*-topology but diverges in Hofer's metric.* Indeed, choose the function F above to be equal to a large constant outside a tiny open disc on S^2. Note that ϕ_F acts trivially outside the disc and thus is C^0-small, while $\rho(\mathrm{id}, \phi)$ can be made arbitrary large. Again, in the linear symplectic space $\mathbf{R}^{2n}$ the situation changes drastically. Hofer showed [H2] that if $\phi_i \to \mathrm{id}$ in $\mathrm{Ham}(\mathbf{R}^{2n})$ in the strong C^0-topology then $\rho(\mathrm{id}, \phi_i)$ must converge to 0. The reason is that in $\mathbf{R}^{2n}$ there is *enough room* to shorten "long" paths with "small" supports.

The proof of Theorem 6.B can be reduced to a Lagrangian intersections problem which one solves using a version of Floer Homology developed by Oh (see [O] for a survey). An important ingredient of this reduction is a detailed knowledge about the fundamental group of $\mathrm{Ham}(S^2)$. Our method works also for some four-dimensional manifolds, for instance when $M = \mathbf{CP}^2$.

7. Length spectrum. Let (M, Ω) be a closed symplectic manifold. For an element $\gamma \in \pi_1(\mathrm{Ham}(M, \Omega), \mathrm{id})$ set $\nu(\gamma) = \inf \mathrm{length}\{f_t\}$ where the infimum is taken over all loops $\{f_t\}$ of Hamiltonian diffeomorphisms which represent γ. In principle, Conjecture 5.A above would give a method of computing $\nu(\gamma)$ at least in some examples. The first step in this direction was made in [LM2] for the case $M = S^2$, and recently J. Slimowitz informed me about her work in progress in dimension four. Here we describe a different approach (see [P3-P6]).

The starting observation is that one can develop a sort of Yang-Mills theory for symplectic fibrations over S^2 with the structure group $\mathrm{Ham}(M, \Omega)$. The role of the Yang-Mills functional is played by the L_∞-norm of the curvature of a symplectic connection on such a fibration (see [GLS] for the definition of symplectic curvature). As expected its minimal values correspond to the length spectrum on $\mathrm{Ham}(M, \Omega)$ in the sense of Hofer's geometry. The L_∞-Yang-Mills functional was first introduced in the context of complex vector bundles by Gromov [Gr] , who called its minimal value *the K-area.*

Further, and this seems to be a specific feature of the Hamiltonian situation, the K-area of a symplectic fibration is closely related to the *coupling parameter.* The coupling is a special construction (see [GLS]) which allows one to extend the fiber-wise symplectic structure of a symplectic fibration to a symplectic form defined in the total space of the fibration. The coupling parameter is responsible for an "optimal" cohomology class of such an extension.

The final step of this approach is based on a powerful machinery of Gromov-Witten invariants [R] which provides us with obstructions to deformations of symplectic forms in cohomology. One can use it in order to compute/estimate the value of the coupling parameter in a number of interesting examples. Therefore one gets the desired information about the length spectrum in Hofer's geometry.

Let us give a precise statement relating Hofer's length spectrum to the coupling parameter. Pick up an element $\gamma \in \pi_1(\mathrm{Ham}(M, \Omega), \mathrm{id})$, and let $\{h_t\}$, $t \in S^1$ be a loop which represents γ. Define a fibration $p : P \to S^2$ as follows. Let D_+ and D_- be two copies of the disc D^2 bounded by S^1. Consider a map

$\Psi : M \times S^1 \to M \times S^1$ given by $(z,t) \to (h_t z, t)$. Define now a new manifold $P(\gamma) = (M \times D_-) \cup_\Psi (M \times D_+)$. It is clear that $P(\gamma)$ has the canonical fiber-wise symplectic form, and thus can be considered as a symplectic fibration over S^2. Moreover, homotopic loops $\{h_t\}$ give rise to isomorphic symplectic fibrations. In what follows we assume that the base S^2 is oriented, and the orientation comes from D_+.

The symplectic fibration $P(\gamma)$ carries a remarkable class $u \in H^2(P, \mathbf{R})$ called the *coupling class*. It is defined uniquely by the following two properties. Its restriction to a fiber coincides with the class of the fiber-wise symplectic structure, and its top power vanishes. Denote by a the positive generator of $H^2(S^2, \mathbf{Z})$, and by $p : P(\gamma) \to S^2$ the natural projection. Using the coupling construction one gets that for $E > 0$ large enough the class $u + Ep^*a$ is represented by a canonical (up to isotopy) symplectic form on the total space $P(\gamma)$ which extends the fiber-wise symplectic structure. Define the coupling parameter of γ as the infimum of such E. Finally, consider the positive part of Hofer's norm $\nu_+(\gamma)$ defined as the infimum of $\text{length}_+\{h_t\}$ over all loops $\{h_t\}$ representing γ (compare with the discussion at the end of section 2 above).

THEOREM 7.A [P6]. *The coupling parameter of γ coincides with $\nu_+(\gamma)$.*

Combining this theorem with the theory of Gromov-Witten invariants one gets for instance the following estimate for the length spectrum. Denote by c the first Chern class of the vertical tangent bundle to $P(\gamma)$. In other words, the fiber of this bundle at a point of $P(\gamma)$ is the (symplectic) vector space tangent to the fiber through this point. Assume that M has real dimension $2n$. Define the "characteristic number"

$$I(\gamma) = \int_{P(\gamma)} u^n \cup c.$$

It is easy to see that $I : \pi_1(\text{Ham}(M, \Omega)) \to \mathbf{R}$ is a homomorphism ([P4],[LMP]).

THEOREM 7.B [P4]. *Let (M, Ω) be a monotone symplectic manifold, that is $[\Omega]$ is a positive multiple of $c_1(TM)$. Then there exists a positive constant $C > 0$ such that $\nu(\gamma) \geq C|I(\gamma)|$ for all $\gamma \in \pi_1(\text{Ham}(M, \Omega))$.*

In other words, the homomorphism I calibrates Hofer's norm on the fundamental group. The proof of 7.B uses results from [Se]. Recently Seidel obtained a generalization of this inequality to non-monotone symplectic manifolds.

Let us mention also that there exists a surprising link between Hofer's length spectrum and spectral Riemannian geometry (see [P6]).

8. ASYMPTOTIC GEOMETRIC INVARIANTS. In applications to dynamical systems it is useful to consider asymptotic invariants arising in Hofer's geometry.

8.A. Asymptotic non-minimality [BP2]. Define a function $\mu : \mathcal{A} - \{0\} \to [0; 1]$ as follows. Take a Hamiltonian function F in $\mathcal{A}$ and consider its Hamiltonian flow $\{f_t\}$. Consider all paths on $\text{Ham}(M, \Omega)$ joining the identity with f_s which are homotopic to $\{f_t\}_{t \in [0;s]}$ with fixed end points. Denote by $\mu(F, s)$ the infimum of lengths of these paths. For instance if $\{f_t\}$ is a minimal geodesic then $\mu(F, s) =$

$s||F||$. It is easy to see that the limit

$$\mu(F) = \lim_{s\to+\infty} \frac{\mu(F,s)}{s||F||}$$

exists. This number is called the asymptotic non-minimality of F, and measures the deviation of $\{f_t\}$ from a (semi-infinite) minimal geodesic. If F generates a minimal geodesic then $\mu(F) = 1$. Consider now two subsets of M consisting of all points where the function F attains its maximal and minimal values respectively. One can show [BP2] that *if one of these subsets has finite displacement energy, then* $\mu(F) < 1$, *and in particular* F *does not generate a minimal geodesic.* Note that this method does not allow us to control the length of the time interval on which the curve $\{f_t\}$ can be shortened.

8.B. Asymptotic length spectrum [P4]. For an element $\gamma \in \pi_1(\mathrm{Ham}(M,\Omega))$ set

$$\nu_\infty(\gamma) = \lim_{k\to+\infty} \frac{1}{k}\nu(\gamma^k).$$

This is an analogue of the Gromov-Federer stable norm in Hofer's geometry. Theorem 7.B above implies that for monotone symplectic manifolds $\nu_\infty(\gamma) \geq C|I(\gamma)|$.

Example. Let M be the blow up of the complex projective plane $\mathbf{CP}^2$ at one point. Choose a Kähler symplectic structure Ω on M which integrates to 1 over a general line and to $\frac{1}{3}$ over the exceptional divisor. The periods of the symplectic form are chosen in such a way that its cohomology class is a multiple of the first Chern class of M. One can easily see that (M,Ω) admits an effective Hamiltonian action of the unitary group $U(2)$, in other words there exists a monomorphism $i : U(2) \to \mathrm{Ham}(M,\Omega)$. The fundamental group of $U(2)$ equals $\mathbf{Z}$. Let $\gamma \in \pi_1(\mathrm{Ham}(M,\Omega))$ be the image of the generator of $\pi_1(U(2))$ under i. It turns out (Abreu - McDuff) that $\pi_1(\mathrm{Ham}(M,\Omega))$ equals $\mathbf{Z}$ and is generated by γ. The direct calculation [P4] shows that $I(\gamma) \neq 0$. We conclude that the asymptotic norm ν_∞ is strictly positive for each non-trivial element of the fundamental group of $\mathrm{Ham}(M,\Omega)$.

I do not know the *precise* value of $\nu_\infty(\gamma)$ in any example where this quantity is strictly positive (for instance, in the example above). The difficulty is as follows. In all known examples where Hofer's norm $\nu(\gamma)$ can be computed precisely there exists a closed loop h which minimizes the length in its homotopy class (that is a minimal closed geodesic). It turns out however that every loop loses minimality after a suitable number of iterations. In other words the loop $\{h_{Nt}\}$ can be shortened provided the integer N is large enough [P8].

9. New intuition in Hamiltonian dynamics. A Hamiltonian flow on a symplectic manifold can be considered as a curve on the group of Hamiltonian diffeomorphisms. One may hope that geometric properties of this curve (in the sense of Hofer's metric) are related to dynamics of the flow. In this section we present three examples of such a link, and thus illustrate our thesis that the geometry on the group of Hamiltonian diffeomorphisms gives rise to a different way of thinking about Hamiltonian dynamics.

9.A. Closed orbits of magnetic fields on the torus. This example was born in discussions with V. L. Ginzburg. Consider the cotangent bundle T^*T^n endowed

with a twisted symplectic structure $\Omega_\sigma = dp \wedge dq + \theta^*\sigma$ as in 3.B above. Fix a Riemannian metric g on T^n. The dynamics of a magnetic field is described by the Hamiltonian flow of the function $|p|_g^2$ with respect to Ω_σ. We claim that *if the magnetic field is non-trivial (that is $\sigma \neq 0$) then there exist non-trivial contractible closed orbits of the flow on a sequence of arbitrary small energy levels*. We refer the reader to [Gi] for a survey of related results. Here is a geometric argument. Fix $\epsilon > 0$. Choose a smooth function $r(x)$, $x \in [0; +\infty)$ which equals $x - 2\epsilon$ on $[0; \epsilon]$, vanishes on $[3\epsilon; +\infty)$ and is strictly increasing on $[0; 3\epsilon)$. Consider a normalized Hamiltonian $F(p, q) = r(|p|_g^2)$. Every non-trivial closed orbit of F corresponds to a non-trivial closed orbit of the magnetic field whose energy is less than 3ϵ. The minimum set of F coincides with the zero section and thus its displacement energy vanishes (see 3.B). From 8.A we see that the asymptotic non-minimality of F is strictly less than 1, thus F does not generate a minimal geodesic. Finally, using the assertion of 5.A (which follows in this case from a result in [LM2]) we conclude that the Hamiltonian flow of F has a non-constant contractible closed orbit.

9.B. Invariant Lagrangian tori (along the lines of [BP2], cf. [Si2]). Consider T^*T^n, this time with the standard symplectic structure $dp \wedge dq$. Let $F \in \mathcal{A}$ be a normalized Hamiltonian with $||F|| = 1$. An important problem of classical mechanics is to decide which energy levels $\{F = c\}$ carry invariant Lagrangian tori homotopic to the zero section. Define a "converse KAM" type parameter $K(F)$ as the supremum of $|c|$, where c is as above. One can show that $\mu(F, s) \geq sK(F)$ for all $s > 0$, and thus $\mu(F) \geq K(F)$. The proof is based on an analogue of theorem 6.B above. Suppose now in addition that F is non-negative and its maximum set $L = F^{-1}(1)$ is a section of T^*T^n. The estimate above shows that if L is Lagrangian then F generates a minimal geodesic. If L is not Lagrangian, then its displacement energy vanishes (cf. 3.B) and thus $\mu(F) < 1$ (see 8.A). We conclude that in this case the asymptotic non-minimality of F gives a non-trivial upper bound for the quantity $K(F)$.

9.C. Strictly ergodic Hamiltonian skew products [P8]. Let (M^{2n}, Ω) be a closed symplectic manifold. Given an irrational number α and a smooth loop $h : S^1 \to \mathrm{Ham}(M, \Omega)$, one defines a *skew product diffeomorphism* $T_{h,\alpha}$ of $M \times S^1$ by $T_{h,\alpha}(x, t) = (h_t x, t + \alpha)$. A traditional problem in ergodic theory is to construct skew products with prescribed ergodic properties associated to loops in groups (see e.g. [N] and references therein). The property we are interested in is *the strict ergodicity*. In our situation the skew product $T_{h,\alpha}$ is called strictly ergodic if it has only one invariant Borel probability measure (which is a multiple of $\Omega^n \wedge dt$). One can adjust existing ergodic methods in order to show that for a wide class of symplectic manifolds (say for simply connected ones) there exist α and h such that $T_{h,\alpha}$ is strictly ergodic. It turns out that the loops h arising in this construction are *contractible*. An attempt to understand this phenomenon gives rise to the following definition. An element $\gamma \in \pi_1(\mathrm{Ham}(M, \Omega))$ is called strictly ergodic if if there exist a number α, and a loop h representing γ such that $T_{h,\alpha}$ is strictly ergodic. It turns out that *the asymptotic norm $\nu_\infty(\gamma)$ vanishes for all strictly ergodic classes γ*. Thus the geometry on $\mathrm{Ham}(M, \Omega)$ supplies us with an obstruction to strict ergodicity. For instance, it follows from 8.B above that for the monotone blow up of $\mathbf{CP}^2$ at one point $\gamma = 0$ is the only strictly ergodic class.

10. Does the geometry on $\mathrm{Ham}(M,\Omega)$ determine (M,Ω)? Here is the simplest version of this question. Let (M,Ω) be a closed surface endowed with an area form, and let $c > 1$ be a real number. Are the spaces $\mathrm{Ham}(M,\Omega)$ and $\mathrm{Ham}(M,c\Omega)$ smoothly isometric with respect to their Hofer's metrics ? Here an isometry is smooth if it sends smooth paths, homotopies etc. to the smooth ones. When $M = S^2$ the answer is negative, since these spaces have different length spectra. When the genus of M is at least 1, the length spectrum is trivial, and the answer is unknown. This open problem of 2-dimensional symplectic topology completes our journey.

Acknowledgments. I thank P. Biran, H. Geiges, V.L. Ginzburg, D. McDuff, P. Seidel, K.F. Siburg and E. Zehnder for their help in preparation of this paper.

References

[Ba] A. Banyaga, Sur la structure du groupe des diffeomorphismes qui preservent une forme symplectique, *Comm. Math. Helv.* 53 (1978), 174-227.

[BP1] M. Bialy, L. Polterovich, Geodesics of Hofer's metric on the group of Hamiltonian diffeomorphisms, *Duke Math. J.* 76 (1994), 273-292.

[BP2] M. Bialy, L. Polterovich, Invariant tori and symplectic topology, *Amer. Math. Soc. Transl.* (2) 171 (1996), pp. 23-33.

[EP] Y. Eliashberg, L. Polterovich, Biinvariant metrics on the group of Hamiltonian diffeomorphisms, *International J. of Math.* 4 (1993), 727-738.

[Gi] V. Ginzburg, On closed trajectories of a charge in a magnetic field. An application of symplectic geometry, in *Contact and Symplectic Geometry*, C.B. Thomas editor, Cambridge University Press, 1996, pp. 131- 148.

[Gr] M. Gromov, Positive curvature, macroscopic dimension, spectral gaps and higher signatures, in *Functional Analysis on the eve of the 21st century*, S. Gindikin, J. Lepowsky, R. Wilson eds., Birkhäuser, 1996.

[GLS] V. Guillemin, E. Lerman and S. Sternberg, *Symplectic fibrations and multiplicity diagrams*, Cambridge University Press, 1996.

[H1] H. Hofer, On the topological properties of symplectic maps, *Proc. Royal Soc. Edinburgh* 115A (1990), 25-38.

[H2] H. Hofer, Estimates for the energy of a symplectic map, *Comm. Math. Helv.* 68 (1993), 48-72.

[HZ] H. Hofer and E. Zehnder, *Symplectic invariants and Hamiltonian dynamics*, Birkhäuser, 1994.

[LM1] F. Lalonde and D. McDuff, The geometry of symplectic energy, *Ann. of Math.* 141 (1995), 349-371.

[LM2] F. Lalonde and D. McDuff, Hofer's L^∞-geometry: energy and stability of flows I,II, *Invent. Math.* 122 (1995), 1-69.

[LMP] F. Lalonde, D. McDuff, L. Polterovich, Topological rigidity of Hamiltonian loops and quantum homology, to appear in *Invent. Math.*

[MS] D. McDuff and D. Salamon, *Introduction to Symplectic Topology*, Oxford University Press, 1995.

[N] M. Nerurkar, On the construction of smooth ergodic skew-products, *Ergod. Th. & Dynam. Sys.* 8 (1988), 311-326.

[O] Y.-G. Oh, Relative Floer and quantum cohomology and the symplectic topology of Lagrangian submanifolds, in *Contact and Symplectic Geometry*, C.B. Thomas ed., Cambridge University Press 1996, pp. 201-267.

[P1] L. Polterovich, Symplectic displacement energy for Lagrangian submanifolds. *Ergod. Th. & Dynam. Sys.* 13, 357-367 (1993).

[P2] L. Polterovich, An obstacle to non-Lagrangian intersections, in *The Floer Memorial Volume* (H. Hofer, C. Taubes, A. Weinstein, E. Zehnder, eds.), Progress in Mathematics, Birkhäuser, 1995, pp. 575-586.

[P3] L. Polterovich, Gromov's K-area and symplectic rigidity, *Geom. and Funct. Analysis* 6 (1996), 726-739.

[P4] L. Polterovich, Hamiltonian loops and Arnold's principle, *Amer. Math. Soc. Transl.* (2) 180 (1997), 181-187.

[P5] L. Polterovich, Precise measurements in symplectic topology, to appear in *Proceedings of the 2nd European Congress of Mathematicians*, Birkhäuser.

[P6] L. Polterovich, Symplectic aspects of the first eigenvalue, to appear in *Crelle's Journal*.

[P7] L. Polterovich, Hofer's diameter and Lagrangian intersections, *IMRN* 4(1998), 217-223.

[P8] L. Polterovich, Hamiltonian loops from the ergodic point of view, Preprint DG/9806152, 1998.

[R] Y. Ruan, Topological sigma model and Donaldson type invariants in Gromov theory, *Duke Math. J.* 83 (1996), 461-500.

[S] J.-C. Sikorav, *Systemes Hamiltoniens et topologie symplectique*, ETS, EDITRICE PISA, 1990.

[Sch] M. Schwarz, A capacity for closed symplectically aspherical manifolds, Preprint, 1997.

[Se] P. Seidel, π_1 of symplectic automorphism groups and invertibles in quantum homology rings, *Geom. and Funct. Anal.* 7 (1997), 1046 -1095.

[Si1] K.F. Siburg, New minimal geodesics in the group of symplectic diffeomorphisms, *Calc. Var.* 3 (1995), 299-309.

[Si2] K.F. Siburg, Action minimizing measures and the geometry of the Hamiltonian diffeomorphism group, *Duke Math J.* 92 (1998), 295-319.

[U] I. Ustilovsky, Conjugate points on geodesics of Hofer's metric, *Diff. Geometry and its Appl.* 6 (1996), 327-342.

Leonid Polterovich
School of Mathematical Sciences,
Tel Aviv University,
69978 Tel Aviv, Israel

Quantum Cohomology and its Application

Yongbin Ruan[1]

1 A brief historical reminiscence

In a few years, quantum cohomology has grown to an impressive field in mathematics with relations to different fields such as symplectic topology, algebraic geometry, quantum and string theory, integrable systems and gauge theory. The development in last few years has been explosive. Now, the foundations have been systematically studied and the ground is secure. Many people contributed to the development of quantum cohomology. I am fortunate to be involved in it from the beginning. Quantum cohomology is such a diverse field that it is impossible to make a complete survey in 45 minutes. I will make no attempt to do so. Instead, I will review some of topics where I made some contributions in last several years.

The development of quantum cohomology has roughly two distint periods: an early pioneer period (91-93) and more recent period of technical sophistication (94-present).

First of all, there are two terminologies: Quantum cohomology, Gromov-Witten invariants. Strictly speaking, quantum cohomology is a special case of the theory of Gromov-Witten invariants. However, the terms are commonly used to mean the same thing and we shall use them interchangebly. Roughly, quantum cohomology studies the following Cauchy-Riemann equation. Let V be a $2n$-dimensional smooth manifold and ω be a symplectic form,i.e., ω^n defines a volume form. We can choose a family of ω-tamed almost complex structure J. J is ω-tamed iff $\omega(X, JX) > 0$ for any nonzero tagent vector X. We want to study the solution space (moduli space) of nonlinear elliptic PDE $\bar{\partial}_J f = 0$ and construct topological invariants of the symplectic manifold (V, ω). The motivation of this problem goes back to two great theories in the 80's, Donaldson's gauge theory and Gromov's theory of pseudo-holomorphic curves. In the summer of 91, I visited Bochum with intention to work with A. Floer on gauge theory. After his tragic death, my gauge theory project went nowhere. It was in this summer that my career took a dramatic turn. After Floer's death, H. Hofer was my main contact person. We had some stimulating conversations where he explained to me Gromov-theory of pseudo-holomorphic curves. I was struck by the obvious resemblance between gauge theory and the theory of pseudo-holomorphic curves. I decided to learn more about it and Hofer recommended to me McDuff's survey paper [Mc]. After reading her paper, I immediately saw how to define a

[1]partially supported by a NSF grant and a Sloan fellowship

Donaldson-type invariant using pseudo-holomorphic curves. Since I was primarily motivated by Donaldson theory, I named them Donaldson-type invariants and now these invariants become commonly known as Gromov-Witten invariants or GW-invariants.

Technically speaking, GW-invariants are harder to study than Donaldson invariants since the compactification of moduli space of pseudo-holomorphic curves is more complicated. For any one with a background in Donaldson theory, it is probably not difficult to define such an invariant. But it is much harder to find interesting examples to show that they are nontrivial. I spent many fruitless hours searching algebraic surfaces to find such examples. Back then, a misperception was that the theory of pseudo-holomorphic curves is a theory about lower dimensional manifolds. Luckly, I came across a group of algebraic geometers working on Mori theory in Max Planck institute in the same summer. I was impressed by beautiful relation between these two subjects. It prompted me to abandon 4-manifolds and study symplectic 6-manifolds instead. After this change of strategy, I quickly found the examples of algebraic 3-folds having the same classical invariant with different new invariants [R2]. The same idea leaded to another paper to generalize some of Mori's results to symplectic manifolds [R3]. Later was extended to Calabi-Yau 3-folds by P. Wilson [Wi2].

A short time ago, some of remarkable progress has been made in physics by Witten for topological quantum field theory. One example of his topological quantum field theory is topological sigma model. However, in 91-92, symplectic geometers were unaware of it. The main motivation of studying these invariants was to distinguish symplectic manifolds. The first version of new invariant I defined was very limited due to the technical difficulty of counting multiple-cover maps. In the early 92, I spent several monthes on thinking about how to overcome this difficulty. Finally I realized that the perturbed Cauchy-Riemann equation $\bar{\partial}_J f = \nu$ introduced by Gromov can be used to give an appropriate account of mutilple covered maps. However, the invariants defined by perturbed equation have a different form from previous invariants.

Let Riemann surface be S^2. S^2 has a nontrivial automorphism $SL_2(C)$, which acts on the moduli space. To obtain compactness of moduli space, we need to divide it by SL_2C action. However, if we consider the perturbed equation $\bar{\partial}_J f = \nu$. The group SL_2C no longer acts on the moduli space. One way to deal with this problem is to impose the condition that f maps $0, 1, \infty$ to some codimension 2 submanifolds. In the fall of 92, I met D. Morrison in a conference in southern California, he explained to me Witten's topological sigma model [W1]. I realized that this new version of the invariants is precisely the correlation function of topological sigma model. These results appeared in [R1] in early 93, which contains a construction of genus zero topological sigma model invariants.

The new link to the topological sigma model brought tremendous insight to Gromov-Witten invariants. The general properties of topological quantum field theory predicted that these invariants must satisfy a set of axioms (Quantum cohomology axioms). The next logical step was to establish a mathematical theory of these invariants, namely proving these axioms. It was clear that this is a nontrivial task which needs some new analysis about pseudo-holomorphic curves.

Furthermore, topological sigma model also contains an analgous theory for higher genus pseudo-holomorphic curves. These higher genus invariants had not been studied before. Moreover, there is an evidence that they are different from the enumerative invariants in algebraic geometry. This was very mysterious to me. In the summer of 93, I met Gang Tian in Germany. We soon started the massive task of a systematic study of Gromov-Witten invariants. By December of 93, our work on a mathematical theory of Gromov-Witten invariants on semi-positive symplectic manifolds was virtually completed. Our results was first appeared in an announcement [RT] in January of 1994 and then in two papers [RT1], [RT2].

Up to the end of 93, quantum cohomology was very much a subject of symplectic topology. It was desirable to have an algebro-geometric treatment since most of examples are Kahler manifolds. Algebraic geometry is very sensitive to compactification. On the other hand, symplectic topology is not so sensitive to compactification due to its topological nature. In the early 90, Parker-Wolfson-Ye [PW], [Ye] obtained a delicate compactification of moduli space of pseudo-holomorphic curves as the product of their effort to prove Gromov compactness theorem using bubbling off analysis. Their compactification now is commonly known as the moduli space of stable maps, a name given by Kontsevich, who was the first one to really understand the importance of stable maps. He made an important observation that the moduli space of genus zero stable maps of homogeneous spaces is a smooth orbifold, where classical techniques apply. In early 94, Kontsevich and Mannin [KM] introduced stable maps and quantum cohomology axioms to the algebraic geometry community. [FP] further popularized quantum cohomology among algebraic geometers. Since then, quantum cohomology has attracted an increasing number of young algebraic geometers. Strictly speaking, the algebro-geometric treatment of Gromov-Witten invariants so far was still short to what we had already accomplished using symplectic methods. It was clear that one needed new ideas and techniques to go beyond homogeneous spaces. The next key step was taken by Li and Tian [LT2], where they used a sophisticated excessive intersection technique (normal cone) (See [B] for a different treatment). As a result, they can dispense the semi-positivity condition in the case of algebraic manifolds. Soon after, a new range of techniques were developed by [FO], [LT3], [R4], [S1] to extend GW-invariants to general symplectic manifolds. Recently, Li-Tian [LT4] and Seibert [S2] showed that the algebraic and symplectic definitions of GW-invariants agree. This completed the first stage of the development of quantum cohomology.

2 Theory of Gromov-Witten invariants

To define GW-invariants, we start from a ω-tamed almost complex structure J. Consider the moduli space of pairs (Σ, f), where $\Sigma \in \mathcal{M}_{g,k}$ is a marked Riemann surface of genus g, with k marked points and $f : \Sigma \to V$ satisfies equation $\bar{\partial}_J f = 0$. We call f a J-holomorphic map or a J-map. f carries a fundamental class $[f] \in H_2(V, \mathbf{Z})$. We use $\mathcal{M}_A(g, k, J)$ to denote the moduli space of (Σ, f) with $[f] = A$. The first step is to compactify $\mathcal{M}_A(g, k, J)$. By Parker-Wolfsen-Ye,

we can compactify it by the moduli space of stable maps. Recall that we can compactify $\mathcal{M}_{g,k}$ by adding the stable Riemann surfaces. A stable Riemann surface is a connected (singular) Riemann surface with arithmetic genus g and k-marked points such that each component is stable, i.e., $2g+k \geq 3$. We use $\overline{\mathcal{M}}_{g,k}$ to denote the moduli space of stable Riemann surfaces of genus g and k-marked points.

DEFINITION 2.1: *A J-holomorphic stable map is a pair (Σ, f), where (i) Σ is a connected (possibly singular) Riemann surface with arithmetic genus g and k-marked points; (ii) $f : \Sigma \to V$ is J-holomorphic; (iii) (Σ, f) satisfies the stability condition that any constant component is stable. (A constant component is one where the restriction of f is a constant map.)*

Let $\overline{\mathcal{M}}_A(V, g, k, J)$ be the space of stable maps. By Parker-Wolfson-Ye [PW], [Ye], $\overline{\mathcal{M}}_A(g, k, J)$ is compact. There are two obvious maps

$$\Xi_{g,k} : \overline{\mathcal{M}}_A(V, g, k, J) \to V^k, \tag{2.1}$$

$$\chi_{g,k} : \overline{\mathcal{M}}_A(V, g, k, J) \to \overline{\mathcal{M}}_{g,k}. \tag{2.2}$$

Here $\Xi_{g,k}$ is defined by evaluating f at the marked point and $\chi_{g,k}$ is defined by successively contracting the unstable component of the domain of stable maps. Let $\alpha_i \in H^*(V, \mathbf{R})$ and $K \in H^*(\overline{\mathcal{M}}_{g,k}, \mathbf{R})$ be a differential form. The GW-invariants are intuitively defined as

$$\Psi^V_{(A,g,k,)}(K; \alpha_1, \cdots, \alpha_k) = \int_{\overline{\mathcal{M}}_A(V,g,k,J)} \chi^*_{g,k}(K) \wedge \Xi^*_{g,k} \prod_i \alpha_i. \tag{2.3}$$

Of course, the above formula only makes sense if $\overline{\mathcal{M}}_A(V, g, k, J)$ is a smooth, oriented orbifold, which is almost never the case. The whole development of GW-invariants is to overcome this difficulty.

The initial approach was a homological approach taken in [R1], [RT1], [RT2]. Here, we consider the dual picture, namely the Poincare dual K^*, α^* of K, α. It is a classical fact that intergration corresponds to intersection of homological cycle K^*, α^*. This approach was accomplished for semi-positive symplectic manifolds which includes most of interesting examples like Fano and Calabi-Yau 3-folds. One consequence of this approach is that the genus zero GW-invariants are integral. This property is still difficult to obtain from recent more powerful techniques.

The second approach was using a cohomological approach where we directly make sense of the integration. There are several methods. A conceptually simple method is as follows [R4], [S1]. By omiting the J-holomorphic condition, we obtain an infinite dimensional space $\overline{\mathcal{B}}_A(V, J, g, k)$ (configuration space). One first constructs a finite dimensional vector bundle $\mathcal{E}$ over $\overline{\mathcal{B}}_A(V, J, g, k)$ [S1]. Then we can construct a triple (U, E, S) such that (i) $U \subset \mathcal{E}$ is a finite dimensional smooth open orbifold ; (ii) E is a finite dimensional bundle over U; (iii) S is a proper section of E such that $S^{-1}(0) = \overline{\mathcal{M}}_A(V, g, k, J)$. Let Θ be a Thom form of E, we can replace (2.3) by

$$\Psi^V_{(A,g,k)}(K; \alpha_1, \cdots, \alpha_k) = \int_U S^*\Theta \wedge \chi^*_{g,k}(K) \wedge \Xi^*_{g,k} \prod_i \alpha_i. \tag{2.4}$$

The triple (U, E, S) is called a virtual neighborhood of $\overline{\mathcal{M}}_A(V, g, k, J)$. Ψ is independent of J, virual neighborhood. It depends only on the cohomology classes of K, α_i. A deep fact is that Ψ satisfies a set of quantum cohomology axioms as follows.

Assume $g = g_1 + g_2$ and $k = k_1 + k_2$ with $2g_i + k_i \geq 3$. Fix a decomposition $S = S_1 \cup S_2$ of $\{1, \cdots, k\}$ with $|S_i| = k_i$. Then there is a canonical embedding $\theta_S : \overline{\mathcal{M}}_{g_1,k_1+1} \times \overline{\mathcal{M}}_{g_2,k_2+1} \mapsto \overline{\mathcal{M}}_{g,k}$, which assigns to marked curves $(\Sigma_i; x_1^i, \cdots, x_{k_1+1}^i)$ $(i = 1, 2)$, their union $\Sigma_1 \cup \Sigma_2$ with $x_{k_1+1}^1$ identified to $x_{k_2+1}^2$ and remaining points renumbered by $\{1, \cdots, k\}$ according to S. There is another natural map $\mu : \overline{\mathcal{M}}_{g-1,k+2} \mapsto \overline{\mathcal{M}}_{g,k}$ by gluing together the last two marked points.

Choose a homogeneous basis $\{\beta_b\}_{1 \leq b \leq L}$ of $H_*(Y, \mathbf{Z})$ modulo torsion. Let (η_{ab}) be its intersection matrix. Note that $\eta_{ab} = \beta_a \cdot \beta_b = 0$ if the dimensions of β_a and β_b are not complementary to each other. Put (η^{ab}) to be the inverse of (η_{ab}).

There is a natural map $\pi : \overline{\mathcal{M}}_{g,k} \to \overline{\mathcal{M}}_{g,k-1}$ as follows. For $(\Sigma, x_1, \cdots, x_k) \in \overline{\mathcal{M}}_{g,k}$, if x_k is not in any rational component of Σ which contains only three special points, then we define

$$\pi(\Sigma, x_1, \cdots, x_k) = (\Sigma, x_1, \cdots, x_{k-1}), \tag{2.5}$$

where a distinguished point of Σ is either a singular point or a marked point. If x_k is in one of such rational components, we contract this component and obtain a stable curve $(\Sigma', x_1, \cdots, x_{k-1})$ in $\overline{\mathcal{M}}_{g,k-1}$, and define $\pi(\Sigma, x_1, \cdots, x_k) = (\Sigma', x_1, \cdots, x_{k-1})$.

QUANTUM COHOMOLOGY AXIOMS:

I: Let $[K_i] \in H_*(\overline{\mathcal{M}}_{g_i,k_i+1}, \mathbf{Q})$ $(i = 1, 2)$ *and* $[K_0] \in H_*(\overline{\mathcal{M}}_{g-1,k+2}, \mathbf{Q})$. *For any* $\alpha_1, \cdots, \alpha_k$ *in* $H_*(V, \mathbf{Z})$, *then we have*

(2.6)

$$\Psi^Y_{(A,g,k)}(\theta_{S*}[K_1 \times K_2]; \{\alpha_i\}) =$$
$$\epsilon \sum_{A=A_1+A_2} \sum_{a,b} \Psi^Y_{(A_1,g_1,k_1+1)}([K_1]; \{\alpha_i\}_{i \leq k_1}, \beta_a) \eta^{ab} \Psi^Y_{(A_2,g_2,k_2+1)}([K_2]; \beta_b, \{\alpha_j\}_{j>k_1})$$

with $\epsilon := (-1)^{deg(K_2) \sum_{i=1}^{k_1} deg(\alpha_i)}$,

$$\Psi^Y_{(A,g,k)}(\mu_*[K_0]; \alpha_1, \cdots, \alpha_k) = \sum_{a,b} \Psi^Y_{(A,g-1,k+2)}([K_0]; \alpha_1, \cdots, \alpha_k, \beta_a, \beta_b) \eta^{ab} \tag{2.7}$$

II: Suppose that $(g, k) \neq (0, 3), (1, 1)$.

(1) For any $\alpha_1, \cdots, \alpha_{k-1}$ *in* $H_*(Y, \mathbf{Z})$, *we have*

$$\Psi^Y_{(A,g,k)}(K; \alpha_1, \cdots, \alpha_{k-1}, [V]) = \Psi^Y_{(A,g,k-1)}([\pi_*(K)]; \alpha_1, \cdots, \alpha_{k-1}) \tag{2.8}$$

(2) Let α_k be in $H_{2n-2}(Y, \mathbf{Z})$, then

$$\Psi^Y_{(A,g,k)}(\pi^*(K); \alpha_1, \cdots, \alpha_{k-1}, \alpha_k) \;=\; \alpha_k^*(A)\Psi^Y_{(A,g,k-1)}(K; \alpha_1, \cdots, \alpha_{k-1}) \tag{2.9}$$

where α_k^* is the Poincare dual of α_k.

III: Ψ^V is a symplectic deformation invariant.

Axioms I, II are due to Witten [W1], [W2] and Axiom III is due to Ruan [R2].

The genus zero GW-invariants can be used to define a quantum multiplication as follows. First we define a total 3-point function

$$\Psi^V(\alpha_1, \alpha_2, \alpha_3) = \sum_A \Psi^V_{(A,0,3)}(pt; \alpha_1, \alpha_2, \alpha_3)q^A, \tag{2.10}$$

where $q^A \in \wedge_V$ is an element of ring of formal power series. Then, we define a quantum multiplication $\alpha \times_Q \beta$ over $H^*(V, \wedge_V)$ by the relation

$$(\alpha \times_Q \beta) \cup \gamma[V] = \Psi^V(\alpha_1, \alpha_2, \alpha_3), \tag{?}$$

where $\cup$ represents the ordinary cup product. An important observation is that

$$\alpha \times_Q \beta = \alpha \cup \beta + \text{lower order quantum corrections.}$$

Hence, this quantum product is often called a deformed product. The 3-point function did not use all the genus zero GW-invariant. An extension of previous construction is to define

$$\Psi^V_w(\alpha_1, \alpha_2, \alpha_3) = \sum_A \sum_{k\geq 3} \frac{1}{(k-3)!}\Psi^V_{(A,0,k)}(\overline{\mathcal{M}}_{0,k}; \alpha_1, \alpha_2, \alpha_3, w, \cdots, w). \tag{2.11}$$

Then we can define a family of quantum product

$$(\alpha \times^w_Q \beta) \cup \gamma[V] = \Psi^V_w(\alpha, \beta, \gamma). \tag{2.12}$$

When $w = 0$, we obtain classical quantum product. The Axiom I for g=0 implies that quantum product $\times^w_Q$ is associative. The associativity has far reaching consequences in enumerative geometry, integrable system and mirror symmetry [Ti]. The previous theory can be generalized in a number of directions, for example, for a family of symplectic manifold and symplectic manifold with a group action [R4]. In the later case, the equivariant theory plays an important role in the recent work about mirror symmetry.

3 Surgery and Gluing theory

Many examples of quantum cohomology have been computed. I refer to [QR] for a list of examples. I believe that the most important future research direction is to develop general technique to compute GW-invariant instead of computing specific

examples. Surgery plays a prominent role in geometry and topology. In fact, it is conjectured that one can connect any two Calabi-Yau 3-folds by a sequence of surgeries called flops and extremal transitions. The famous Mori program of birational geometry is basically a surgery theory. On the other hand, surgery has been used in symplectic topology to construct many new examples [Go2], [MW]. Therefore, it is very important to use surgery to study quantum cohomology. This requires a gluing theory of pseudo-holomorphic curves. While we have several choices of surgeries, a particularly useful one in the application to symplectic topology and algebraic geometry is symplectic cutting-symplectic norm sum [L]. Such a gluing technique has been recently established by Li-Ruan [LR] and Ionel-Parker [IP].

Suppose that X admits a local Hamiltonian circle action. Then, we can cut X along a level set and collapse the circle action on the boundary. Then, we obtain a pair of symplectic manifolds X^+, X^- called symplectic cuttings of X. X^+, X^- contains a common codimension 2 symplectic submanifold Z with opposite first Chern class of their normal bundle. Many algebro-geometric surgeries can be interpreted as symplectic cutting, where the Hamiltonian circle action is usually given by complex multiplication. The gluing theory describes the behavior of pseudo-holomorphic curves under stretching the "neck" (the region carring circle action). In the limit, pseudo-holomorphic curves break as pseudo-holomorphic holomorphic curves in X^+, X^- with possibly several components. Moreover, these curves could intersect Z with high tangency condition. Moreover. some component could lie in Z.

To capture these new phenomena from gluing theory, we can introduce a relative Gromov-Witten invariant [LR](see [IP] for a related invariant). Choose a tamed almost complex structure J such that Z is almost complex. Then, one can define *relative stable maps* with prescribed tangency condition on Z. Then one can use the above virtual neighborhood method to define *relative GW-invariants*. There is a natural map from the moduli space of relative stable maps into the moduli space of stable maps. However, this map is not surjective in general. The difference counts the discrepency between relative and absolute invariants, which is caused precisely by the stable maps whose components lie in Z. In favorable circumstances, relative invariants are easy to compute or can be related to regular GW-invariants.

Then, general gluing theory shows that Gromov-Witten invariants of symplectic manifolds can be related to relative invariants of its symplectic cutting. The general formula is complicated and probably not very useful. In applications, we often encounter the situation that most of the relative invariants vanish and it is much easier to count them. Then, we get formula for the GW-invariants. Here are some applications. Recall that a minimal model is an algebraic variety with terminal singularities and nef canonical bundle. In the dimension 3, two different minimal models are connected to each other by flops. By applying gluing theory to the flop, Li-Ruan showed

Theorem 3.1: *Any two smooth three dimensional minimal models have isomorphic quantum cohomology.*

However, it is well-known that they can have different ordinary cohomology. This establishes the first quantum birational invariant. Furthermore, Li-Ruan derived various formulas of quantum cohomology under extremal transition, which are important in mirror symmetry. Moreover, Ionel-Parker use this technique to give an elegant proof of Caparosa-Harris formula of number of curves in $\mathbf{P}^2$ and Bryant-Liang's formula of number of curves in K3-surfaces. I have no doubt that the gluing theory will yield more important applications towards quantum cohomology.

4 Problems and Conjectures

I believe that the future success of quantum cohomology theory depends on its applications. Clearly, the ability to apply quantum cohomology also depends on our understanding of GW-invariants. For quantum cohomology itself, I believe that the biggest problem is our poor understanding of its functorial properties. The reason cohomology is very useful is its naturality. Namely, a continuous map induces a homomorphism on cohomology. Although we have calculated many examples, it help us little on this problem.

Quantum naturality problem: *What are the "morphisms" of symplectic manifolds where quantum cohomology is natural?*

Li-Ruan [LR] suggests that this problem is tied to so called *small transition*, which is the composition of a small contraction and smoothing. Incidentaly, small contractions are the most difficult operations in birational geometry. However, [LR] suggests that they are easiest in quantum cohomology.

I believe that there is a deep relation between quantum cohomology and birational geometry. Theorem 3.1 suggests

Quantum minimal model conjecture: *Theorem 3.1 holds in any dimension.*

This leads to many more questions. For example, one can attempt to find quantum cohomology of a minimal model without knowing minimal model. This problem requires a thorough understanding of blow-up type formula of quantum cohomology. Since quantum cohomology is a deformation invariant, one can try to relax the birational classification by allowing deformation, which we call deformation birational classification. Then, one replace contraction by extremal transition. One can try to construct minimal models using extremal transition. Quantum cohomology should play an important role in this new category. It is even more exciting that such a deformation-birational minimal model program has a natural analogy in symplectic manifolds.

There are many outstanding problems in the quantum cohomology. Let me list several examples, Virasoro conjecture [EHX], quantum hyperplane conjecture [Kim], mirror surgery conjecture [LR], conjectures of characterizations of uniruled varieties and rational connected varieties [KO]. It seems that possible applications are numerous and future is bright for quantum cohomology.

Over the years, I have been benefited from generous help of many people.

Without them, my mathematical career wouldn't be possible. The list is too long to enumerate in this conference. I would like to thank all of them for their help. In particular, I would like to take this special opportunity to thank Liangxi Guo, Haoxuan Zhou and Yingmin Liu for their guidance and help during the early years of my life.

References

[B] K. Behrend, GW-invariants in algebraic geometry, preprint.

[EHX] T. Eguchi, K. Hori, C-S. Xiong, Quantum cohomology and Virasoro algebra, hep-th/9703086

[FO] K. Fukaya and K. Ono, Arnold conjecture and Gromov-Witten invariant, preprint

[FP] W. Fulton and R. Pandharipande, Notes on stable maps and quantum cohomology, preprint alg-geom/9608011

[Go2] R. Gompf, A new construction of symplectic manifolds, preprint.

[Gr] M. Gromov, Pseudo holomorphic curves in symplectic manifolds, Invent. math., 82 (1985), 307-347.

[IP] E-N. Ionel and T. Parker, Gromov-Witten invariants of symplectic sums, preprint, math.SG/9806013

[Kim] B. Kim, Quantum hyperplane section theorem for homogeneous spaces, preprint, alg-geom/9712008

[KO] J. Kollar, Rational curves on algebraic varieties, Springer Verlag.

[KM] M. Kontsevich and Y. Manin, GW classes, Quantum cohomology and enumerative geometry, Comm.Math.Phys., 164 (1994), 525-562.

[L] E. Lerman, Symplectic cuts, Math Research Let 2(1985) 247-258

[LR] An-Min Li and Y. Ruan, Symplectic surgery and GW-invariants of Calabi-Yau 3-folds I, preprint, alg-geom/9803036

[LT2] J. Li and G. Tian, Virtual moduli cycles and GW-invariants, preprint

[LT3] J. Li and G. Tian, Virtual moduli cycles and Gromov-Witten invariants of general symplectic manifolds, preprint.

[LT4] J. Li and G. Tian, Comparison of the algebraic and the symplectic Gromov-Witten invariants, preprint alg-geom/9712035

[Mc] D. McDuff, Elliptic methods in symplectic geometry, Bull. AMS (N.S) 23(1990),311-358

[MW] G. McCarthy and J. Wolfson, Symplectic normal connect sum, Topology, 33(1994), 729-764.

[PW] T. Parker and J. Woflson, A compactness theorem for Gromov's moduli space, J. Geom. Analysis, 3 (1993), 63-98.

[R1] Y. Ruan, Topological Sigma model and Donaldson type invariants in Gromov theory, Math. Duke. Jour. vol 83. no 2(1996), 461-500

[R2] Y. Ruan, Symplectic topology on algebraic 3-folds, J. Diff. Geom. 39(1994) 215-227.

[R3] Y. Ruan, Symplectic topology and extremal rays, Geo, Fun, Anal, Vol 3, no 4 (1993), 395-430.

[R4] Y. Ruan, Virtual neighborhood and pseudo-holomorphic curves, preprint.

[QR] Y. Ruan and Z. Qin, Quantum cohomology of projective bundles over $\mathbf{P}^n$, to appear in Trans of AMS.

[RT] Y. Ruan and G. Tian, A mathematical theory of quantum cohomology, announcement, Math. Res. Let., vol 1, no 1 (1994), 269-278.

[RT1] Y. Ruan and G. Tian, A mathematical theory of quantum cohomology, J. Diff. Geom., vol. 42, no2 (1995) 259-367

[RT2] Y. Ruan and G. Tian, Higher genus symplectic invariants and sigma model coupled with gravity, Invent. Math

[S1] B. Siebert, Gromov-Witten invariants for general symplectic manifolds, preprint.

[S2] B. Siebert, Algebraic and symplectic GW-invariants coincide, preprint math.AG/9804108

[Ti] G. Tian, The quantum cohomology and its associativity, Current Development in Mathmatics.

[Wi2] P.M.H.Wilson, Symplectic deformations of Calabi-Yau threefolds, J. Diff. 45(1997)

[W1] E. Witten, Topological sigma models, Comm. Math. Phys., 118 (1988).

[W2] E. Witten, Two dimensional gravity and intersection theory on moduli space, Surveys in Diff. Geom., 1 (1991), 243-310.

[Ye] R. Ye, Gromov's compactness theorem for pseudo-holomorphic curves, Trans. Amer. math. Soc., 1994.

Yongbin Ruan
Department of Mathematics
University of Wisconsin-Madison
Madison, WI 53706
USA

Section 6

Topology

In case of several authors, Invited Speakers are marked with a *.

Dimension Theory and Large Riemannian Manifolds

A. N. Dranishnikov

Abstract. In this paper we discuss some recent applications of dimension theory to the Novikov and similar conjectures. We consider only geometrically finite groups i.e. groups Γ that have a compact classifying space $B\Gamma$. It is still unknown whether all such groups admit a sphere at infinity [B]. In late 80s old Alexandroff's problem on the coincidence between covering and cohomological dimensions was solved negatively [Dr]. This brought to existence a locally nice homology sphere which is infinite dimensional. In the beginning of 90s S. Ferry conjectured that if such homology sphere can be presented as a sphere at infinity of some group Γ, then the Novikov conjecture is false for Γ. Here we discuss the development of this idea. We outline a reduction of the Novikov conjecture to dimension theoretic problems. The pionering work in this direction was done by G. Yu [Yu]. He found a reduction of the Novikov Conjecture to the problem of finite asymptotic dimensionality of the fundamental group Γ. Our approach is based on the hypothetical equivalence between asymptotical dimension of a group and the covering dimension of its Higson corona. The slogan here is that most of the asymptotic properties of Γ can be expressed in terms of topological properties of the Higson corona $\nu\Gamma$. At the end of the paper we compare existing reductions of the Novikov conjecture in terms of the Higson corona.

§1. Dimension theory of compacta

The *covering dimension* $\dim X$ of a compact metric space X can be defined as the smallest number n such that for any $\epsilon > 0$ there is an ϵ- covering $\{U_1, ..., U_k\}$ of X of order $\leq n+1$. The definition does not depend on the metric on X. There are many equivalent reformulations of this property and not all of them are exactly obvious. Here we give two of them.

Ostrand Theorem. $\dim X \leq n \Leftrightarrow$ *for any positive* ϵ *there exist* $n+1$ *discrete families* $\mathcal{U}_i$ *of mesh* $< \epsilon$ *such that the union* $\cup\mathcal{U}_i$ *forms a cover of* X.

Alexandroff-Hurewicz Theorem. $\dim X \leq n \Leftrightarrow$ *for every map* $\phi : A \to S^n$ *of a closed subset* $A \subset X$ *there is an extension* $\bar{\phi} : X \to S^n$.

The *cohomological dimension* $\dim_{\mathbf{Z}} X$ is the smallest n such that $\check{H}_c^{n+1}(U) = 0$ for all open sets $U \subset X$. The notion of cohomological dimension was introduced by P.S. Alexandroff in late 20s in homology language. Since then until late 80s there was an open problem on the coincidence of dim and $\dim_{\mathbf{Z}}$. In early 30s Alexandroff, collaborating with H. Hopf, proved the following.

ALEXANDROFF THEOREM. *For finite dimensional compacta* $\dim X = \dim_{\mathbf{Z}} X$.

In 70s R.D. Edwards connected the Alexandroff problem with the following more geometric problem: *Can a cell-like map of a manifold raise dimension?* We recall that a map $f : X \to Y$ is called *cell-like* if all fibers $f^{-1}(y)$ have trivial shape. Edwards proved the following

RESOLUTION THEOREM [Wa]. *For every compactum X there is a compactum Y of $\dim Y \leq \dim_{\mathbf{Z}} X$ and a cell-like map $f : Y \to X$.*

In particular the Resolution Theorem allowed to extend the equality $\dim = \dim_{\mathbf{Z}}$ on classes of countable dimensional compacta, ANR-compacta and compacta with C-property [A]. The C-property is a generalization of finite dimensionality in the direction of the Ostrand theorem. A space X has *C-property* if for any sequence $\{\mathcal{U}_i\}$ of covers of X there is a sequence of disjoint families $\{\mathcal{O}_i\}$ such that $\mathcal{O}_i$ is inscribed in $\mathcal{U}_i$ and the union $\cup\mathcal{O}_i$ forms a cover of X.

The Alexandroff problem was solved by a counterexample [Dr]. That counterexample in view of the Resolution Theorem gives a cell-like map $f : S^7 \to X$ with $\dim X = \infty$. The space X is a homology manifold which is locally connected in all dimensions. Every cell-like map of a manifold induces an isomorphism of homotopy groups, homology groups and cohomology groups. It turns out to be that this fails for K-theory.

THEOREM 1 [D-F]. *For any p there is a cell-like map $f : S^7 \to X$ such that $KerK_*(f) \neq 0$ for* $\mod p$ *complex homology K-theory.*

COROLLARY. *The homology sphere X does not admit a map of degree one onto S^7.*

§2. NOVIKOV CONJECTURE

Let G_n^k be the Grassmanian space of k-dimensional oriented vector subspaces in n-space with the natural topology. There is the natural imbedding $G_n^k \subset G_{n+1}^k$. Then one can define the space $G_\infty^k = \lim_\to G_n^k$. The natural imbedding $G_\infty^k \subset G_\infty^{k+1}$ leads to the definition of the space $BO = G_\infty^\infty = \lim_\to G_\infty^k$. The tangent bundle of an n-dimensional manifold N can be obtained as the pull-back from the natural n-bundle over the space G_∞^n. Let $f_\tau : N \to BO$ be a map which induces the tangent bundle on N. The cohomology ring $H^*(BO; \mathbf{Q})$ is a polynomial ring generated by some elements $a_i \in H^{4i}(BO; \mathbf{Q})$. The rational Pontryagin classes of a manifold N are the elements $p_i = f^*(a_i) \in H^{4i}(BO; \mathbf{Q})$. Novikov proved [N] that the rational Pontryagin classes are topological invariants. It was known that they are not homotopy invariants. Hirzebruch found polynomials $L_k(p_1, \ldots, p_k) \in H^{4k}(N; \mathbf{Q})$ which do not depend on N and such that the signature of every closed (oriented) $4k$ manifold N can be defined as the value of L_k on the fundamental class of N. Note that the signature is homotopy invariant and even more, it is bordism invariant. For non-simply connected manifolds Novikov defined the higher signature as follows. Let Γ be the fundamental group of a closed oriented manifold N, let $g : N \to B\Gamma = K(\Gamma, 1)$ be a map classifying the universal cover of N and let $b \in H^*(K(\Gamma, 1); \mathbf{Q})$. Then he defines the b-signature as $sign_b(N) =$

$\langle L_k \cap g^*(b), [N]\rangle$, here $4k + \dim(b) = \dim N$. These rational numbers $sign_b(N)$ are called *higher signatures*. The higher signature are the only possible homotopy invariants [M]. The Novikov conjecture states that they are homotopy invariant.

NOVIKOV CONJECTURE. *Let $h : N \to M$ be an orientation preserving homotopy equivalence between two close oriented manifolds, then $sign_b(N) = sign_b(M)$ for any $b \in H^*(K(\Gamma, 1); \mathbf{Q})$.*

We say that the Novikov conjecture holds for a group Γ if it holds for every manifold with the fundamental group Γ.

A tangent bundle can be defined for a topological manifold N as well. This bundle is classified by a map $f : M \to BTOP$ where $TOP = \lim_{\to} TOP_n$ and TOP_n is a topological group of homeomorphisms $f : \mathbf{R}^n \to \mathbf{R}^n$ with $f(0) = 0$. Since the natural map $BO \to BTOP$ induces an isomorphism of rational cohomology groups, one can define Pontryagin classes and higher signatures for M. In the TOP category there is a functorial 4-periodic surgery exact sequence:

$$\ldots \to \mathcal{S}_n(N) \xrightarrow{\eta} H_n(N; \mathbb{L}) \xrightarrow{\alpha} L_n(\Gamma) \to \mathcal{S}_{n-1}(N) \to \ldots,$$

where Γ is the fundamental group of X, $L_n(\Gamma)$ are Wall's groups, $\mathbb{L}$ is a periodic spectrum generated by G/TOP, and $\mathcal{S}_n(N)$ is the group of manifold structures on N with possible summand $\mathbf{Z}$. The group $\mathcal{S}_n(N)$ can be defined as the group of classes of homotopy equivalences $q : M \to \bar{N}$ with $q \mid_{\partial M} = 1_{\partial \bar{N}}$, here $\bar{N}$ is a regular neighborhood of N in some euclidean space of dimension $n + 4l$ [We]. This sequence is defined for any finite polyhedron. One can consider the lost tribe manifolds [B-F-M-W] to avoid possible extra $\mathbf{Z}$s in the definition of $\mathcal{S}_*(N)$.

The *higher L-genus* of an n-manifold N with the fundamental group Γ is an element $g_*(L(N) \cap [N]) \in H_*(B\Gamma; \mathbf{Q}) = \oplus H_i(B\Gamma; \mathbf{Q})$. This notion is dual to the higher signatures. The Novikov conjecture is equivalent to the statement that for any homotopy equivalence $h : M \to N$ the higher L-genuses of N and M are equal. Note that $H_n(X; \mathbb{L}) \otimes \mathbf{Q} = \oplus_{i=n \mod 4} H_i(X; \mathbf{Q})$. The morphism η takes a homotopy equivalence $q : M \to N$ to the difference $L(M) \cap [M] - L(N) \cap [N]$. Assume that $B\Gamma$ is a finite complex i.e. Γ is geometrically finite, then the map $g : N \to B\Gamma$ and the periodic surgery exact sequence produce the diagram

$$\begin{array}{ccccc} S_*(N) & \xrightarrow{\eta} & H_*(N; \mathbf{Q}) & \xrightarrow{\alpha} & L_*(\Gamma) \otimes \mathbf{Q} \\ & & {\scriptstyle g_*}\downarrow & & {\scriptstyle =}\downarrow \\ & & H_*(B\Gamma; \mathbf{Q}) & \xrightarrow{A} & L_*(\Gamma) \otimes \mathbf{Q} \end{array}$$

So g_* takes the image of the class of a homotopy equivalence q to the difference of the higher signatures of M and N. Thus, the injectivity of the *assembly map* $A : H_*(B\Gamma; \mathbf{Q}) \to L_*(\Gamma) \otimes \mathbf{Q}$ implies the Novikov conjecture. The opposite is also true [K-M].

In the case of geometrically finite Γ it makes sense to ask whether the integral assembly map $A : H_*(B\Gamma; \mathbb{L}) \to L_*(\Gamma)$ is a split monomorphism. This is called the *integral Novikov conjecture*. By Davis' trick with Coxeter groups, it follows

that every finite aspherical complex is a retract of a closed aspherical manifold. A diagram chasing shows that in the class of geometrically finite groups for studying the Novikov conjecture it suffices to consider the case when $B\Gamma$ is a manifold. In that case the universal cover $E\Gamma = X$ is a contractible manifold. Without loss of generality we may assume that X homeomorphic to the euclidean space.

A special case of the Novikov conjecture is the following:

GROMOV-LAWSON CONJECTURE. *An aspherical manifold cannot carry a metric of a positive scalar curvature.*

An open n-dimensional riemannian manifold X is called *hypereuclidean* if there is a Lipschitz map $f : X \to \mathbf{R}^n$ of degree one. The Gromov-Lawson conjecture holds true for hypereuclidean manifolds [G-L]. A metric space X is called *uniformly contractible* if for every $R > o$ there is $S > o$ such that any R-ball $B_R(x)$ centered at x can be contracted to a point in $B_S(x)$ for any $x \in X$. A typical example of a uniformly contractible manifold is a universal cover of a closed aspherical manifold with the lifted metric. A positive answer to the following problem [G2] would imply the Gromov-Lawson conjecture.

Is every uniformly contractible manifold hypereuclidean?

There is also an analytic approach to the Novikov conjecture which reduces the problem to the question of an injectivity of an analytic assembly map $\mathcal{A} : K_*(B\Gamma) \to K_*(C^*(\Gamma))$, where the right part is an algebraic K-theory of some C^*-algebra. This assembly map can be defined in terms of a universal cover $E\Gamma$ [B-C]. Then the assembly map and the conjecture can be to extended to general metric spaces [R1], [H-R].

COARSE BAUM-CONNES CONJECTURE [R1],[R2]. *For every uniformly contractible bounded geometry metric space X the assembly map $A : K_*(X) \to K_*(C^*X)$ is a monomorphism (isomorphism).*

A metric space has a *bounded geometry* if for any $\epsilon > 0$ for every $R > 0$ there is m such that every R-ball contains an ϵ-net consisting of $< m$ points. It is clear that every finitely presented group has a bounded geometry. Without this restriction the coarse Baum-Connes conjecture is not true [D-F-W]. A description of the C^*-algebra $C^*(X)$ can be found in [H-R],[R2]. We note that the coarse Baum-Connes conjecture implies the Gromov-Lawson conjecture [R1] and the isomorphism version of it implies the Novikov conjecture [R2].

A fascinating result in the coarse approach to the Novikov Conjecture was obtained by Yu [Yu]. He proved the following.

THEOREM [Yu]. *If a proper uniformly contractible metric space X has a finite asymptotic dimension, then the coarse Baum-Connes conjecture holds for X.*

The definition of asymptotic dimension is given in the next section where we also sketch the idea how to prove Yu's theorem.

§3. COARSE TOPOLOGY

A metric space (X, d) is called *proper* if every closed ball $B_r(x_0) = \{x \in X \mid d(x, x_0) \leq r\}$ is compact. A map between proper metric spaces $f : (X, d_X) \to$

(Y, d_Y) is called a *coarse morphism* [R2] if it is proper and uniformly expansive i.e. $f^{-1}(C)$ is compact for every compact C and for any $R > 0$ there is $S > 0$ such that $d_Y(f(x), f(x')) < S$ if $d_X(x, x') < R$. Note that every Lipschitz map is a coarse morphism. Vice versa, for a geodesic metric space there are $R > 0$ and $\lambda > 0$ such that $d_Y(f(x), f(x')) < \lambda d_X(x, x')$ for all x, x' with $d_X(x, x') \geq R$. Such maps are called *coarsely Lipschitz*.

In this section we consider a category $\mathcal{C}$ of proper metric spaces with proper coarsely Lipschitz maps as morphisms. The Coarse category is the quotient of $\mathcal{C}$ by the equivalence stating that any two morphisms, which are in a finite distance from each other, are equivalent. We consider only uniformly contractible metric spaces. In the case of general proper metric spaces one should consider morphisms which are not necessarily continuous and the properness should be replaced by the following: $f^{-1}(B)$ is bounded for every bounded set B. In many cases a general type metric space (X, d) admits a uniformly contractible filling $X' \supset X$ with $(X, d' \mid_X)$ coarsely equivalent to (X, d). Thus geometrically finite groups Γ with word metric d have a filling called a universal cover of $B\Gamma$ with lifted metric d'.

Note that a closed subspace $Y \subset X$ of a proper metric space X with the induced metric is an object of $\mathcal{C}$. We define the notion of an absolute extensor in $\mathcal{C}$ as usual: $X \in AE(\mathcal{C})$ if for any $Z \in \mathcal{C}$ and for any closed $A \subset Z$ and a morphism $\phi : A \to X$ there is an extension $\bar{\phi} : Z \to X$.

Let $\mathbf{R}^n_+$ denote the halfspace of dimension n with the induced metric.

THEOREM 2. *$\mathbf{R}^n_+ \in AE(\mathcal{C})$ for all n.*

Note that $\mathbf{R}^n$ is not AE.

We define a *coarse neighborhood* W of $Y \subset X$ as a subset of X with $\lim dist(y, X \setminus W) = \infty$ as $y \in Y$ approaches infinity. Define a finite open cover of (X, d) as a finite cover of X by open coarse neighborhoods with the Lebesgue function $\lambda(x)$ tending to infinity as x approachs infinity.

Note that, $\mathbf{R}^{n+1}$ is obtained from $\mathbf{R}^n$ by the operation analogous to the suspension. By analogy with Alexandroff-Hurewicz theorem we define a coarse dimension $\dim^c(X, d)$ as follows:

$\dim^c(X, d) \leq n$ *if and only if for every closed subspace $A \subset X$ and any coarse morphism $\phi : A \to \mathbf{R}^{n+1}$ there is an extension to a coarse morphism $\bar{\phi} : X \to \mathbf{R}^{n+1}$.*

Here we use $\mathbf{R}^{n+1}$ as an analog of S^n in order to have the equality $\dim^c \mathbf{R}^n = n$. By Pontryagin-Nobeling theorem every n-dimensional compactum can be embedded in the cube I^{2n+1}. Then the following problem is quite natural.

EMBEDDING PROBLEM. *Does a metric space with $\dim^c(X, d) \leq n$ have a coarse embedding in the space $\mathbf{R}^{2n+2}_+$?*

M. Gromov defined [G1] the notion of asymptotic dimension using a coarse analog of the Ostrand theorem. By the definition $\operatorname{asdim}(X, d) \leq n$ *if for any $R > 0$ there are $n + 1$ R-disjoint uniformly bounded families $\mathcal{U}_i$ such that the union forms a cover of X.* The inequality $\operatorname{asdim}(X, d) \leq n$ means that X is coarse equivalent to a simplicial complex K_R of dimension $\leq n$ with all simplices with edges of the length R for an arbitrary large R. This property leads to the

notion of *anti Čech approximation* of X by simplicial complexes. Metric spaces that admit an anti-Čech approximation by finite simplicial complexes are called *spaces of bounded geometry*. We note that universal covers of classifying spaces of geometrically finite groups Γ supplied with a Γ-invariant metric are spaces of bounded geometry. J. Roe defined coarse homology (cohomology) of a metric space using anti-Čech approximation. This leads to the definition of asymptotical cohomological dimension of a metric space. Another approach to the cohomological dimension is the following. Since we are already have n-cells, one can define CW-complexes in the coarse category. Coarse homotopy groups we define below. Then we can construct a coarse Eilenberg-MacLane complexes $K(\mathbf{Z}, n)$ and define $\operatorname{asdim}_{\mathbf{Z}} X \leq n$ if every partial map on X to $K(\mathbf{Z}, n)$ can be extended.

The following is an analog of Kuratowski-Dugundji theorem.

PROPOSITION 1. *Let X be uniformly contractible proper metric space with $\operatorname{asdim} X < \infty$, then $X \in ANE(\mathcal{C})$.*

Following Gromov's idea, we define a homotopy in the coarse category as a morphism of the set $D_X = \{(x,t) \in X \times \mathbf{R} \mid |t| \leq d(x, x_0)\}$ where $x_0 \in X$ is a based point. Note that the subspaces $D_X^+ = \{(x, d(x, x_0)\} \subset D_X$ and $D_X^- = \{(x, -d(x, x_0)\} \subset D_X$ are coarsely isomorphic to X. It is possible to show that coarse homotopic maps induce the same homomorphism of coarse homology (cohomology) groups. The next natural notion is *coarse homotopy type*. Thus, $\mathbf{R}^n$ and $\mathbb{H}^n$ have the same coarse homotopy type. It turns out to be that the coarse Baum-Connes conjecture is invariant under coarse homotopy equivalence [R2]. It is possible to show that the coarse Baum-Connes conjecture holds for coarse polyhedra [R2] and hence for metric spaces which are coarse homotopy equivalent to polyhedra. Now Yu's theorem would follow from Proposition 1 and a coarse analog of the West theorem: ANE-space is homotopy equivalent to a polyhedron. The following straightforward proposition allows to give a simpler approach.

PROPOSITION 2. *Let a metric space X be coarse homotopically dominated by a space Y. Assume that the Baum-Connes conjecture holds for Y, then it holds for X as well.*

Let $f_0 : \mathbf{R}_+ \to X$ be a coarse morphism. A *coarse loop* $\phi : \mathbf{R}_+^2 \to X$ is a morphism such that $\phi \mid_{\mathbf{R}_+} = f_0 = \phi \mid_{-\mathbf{R}_+} \circ(-1)$ where $\mathbf{R}_+$ is naturally imbedded in the first factor of $\mathbf{R}_+^2 = \mathbf{R} \times \mathbf{R}_+$. The product of two coarse loop can be defined by compression of two $\mathbf{R}_+^2$ to quadrants and gluing two quadrants together.

This leads to the definition of the coarse fundamental group and higher dimensional coarse homotopy groups. Since we have the notion of the standard n-simplex in $\mathcal{C}$ we can define singular coarse homology (cohomology) of metric spaces. We expect that all theorems of classical algebraic topology hold here.

§4. HIGSON CORONA

Let (X, d) be a metric space and let $f : X \to \mathbf{R}$ be a function on X. An *r-variation* of f at $x \in X$ is the following number $V_r(f(x)) = \sup\{|f(x) - f(y)| \mid y \in B_r(x)\}$. Let $B(X)$ be the set of all bounded functions $f : X \to \mathbf{R}$ with $\lim_{x \to \infty} V_r(f(x)) = 0$ for ant $r > 0$. We define the Higson compactification of X as the closure $\bar{X}$ of

X embedded in $I^{B(X)}$ by the family $\{f_b \mid b \in B(X)\}$. The remainder $\nu X = \bar{X} \setminus X$ of the Higson compactification is called the *Higson corona* [H],[R1].

The Higson corona is an invariant of a coarse isometry. Hence the Higson corona of a discrete finitely generated group Γ is a group invariant, i.e. $\nu\Gamma$ does not depend on choice of a word metric on Γ. Thus, two metric spaces in a finite distance in the Gromov-Hausdorff metric space have the same Higson coronas. Moreover, the Higson corona is a functor $\nu : \mathcal{C} \to \mathcal{C}omp$ from the coarse category to the category of compact Hausdorff spaces, taking embeddings to embeddings.

There is a partial order on compactifications of a given (locally compact) space X. A compactification cX is dominated by a compactification $c'X$ if there is a continuous map $f : c'X \to cX$ with $f|_X = 1_X$. A compactification, dominated by the Higson compactfication, we call *Higson dominated.*

Many asymptotic properties of metric spaces can be formulated in terms of the Higson corona. We give two examples of such properties. A notion of small action of a discrete group Γ at infinity of a universal cover X of $B\Gamma$ appears naturally in the combinatorial group theory. Thus, Bestvina takes that property as an axiom of his Z-boundary of a group [B]. An action of Γ is *small at infinity* for a given compactification $\bar{X}$ of X if for every $x \in \bar{X} \setminus X$ and a neighborhood U of x in $\bar{X}$, for every compact set $C \subset X$ there is a smaller neighborhood V such that $g(C) \cap V \neq \emptyset$ implies $g(C) \subset U$ for all $g \in \Gamma$. We consider a Γ-invariant metric on X. Since $B\Gamma$ is a finite complex, the Higson corona of X does not depend on choice of metric and coincides with the Higson corona of Γ.

PROPOSITION 3. *The action of Γ on X is small at infinity for a compactification $\bar{X}$ if and only if $\bar{X}$ is Higson dominated.*

Existence of such compactification is crucial in all cases were the Novikov conjecture is proved.

Another property is also related to the Novikov Conjecture.

THEOREM [R1]. *An open n-manifold M is hypereuclidean if and only if there is a map $f : \nu M \to S^{n-1}$ of degree one.*

Since a dimension is an important invariant in the coarse theory we establish the following.

THEOREM 3. $\dim \nu X = \dim^c(X, d)$ *for a proper metric space* (X, d).

THEOREM 4. $\dim \nu X = \operatorname{asdim} X$ *if* $\operatorname{asdim} X < \infty$.

CONJECTURE 1. $\dim \nu X = \operatorname{asdim} X$ *for all* X.

Note that the inequality $\dim \nu X \leq \operatorname{asdim} X$ always holds [D-K-U]. The proof of this inequality makes plausible that $\nu\Gamma$ has the C-property for geometrically finite group Γ. This together with Ancel's theorem (§1), Conjecture 1 and the following conjecture define another approach to the Novikov Conjecture for all geometrically finite groups.

CONJECTURE 2. $\dim_{\mathbf{Z}} \nu X \leq \operatorname{asdim}_Z X$.

The following conjecture is somewhat weaker of the rational Gromov-Lawson conjecture and it is equivalent to Gromov-Lawson's for even dimensional manifolds [D-F].

WEINBERGER CONJECTURE. *For every uniformly contractible n-manifold X the boundary homomorphism $\partial : \check{H}^{n-1}(\nu X; \mathbf{Q}) \to H^n_c(X; \mathbf{Q}) = \mathbf{Q}$ is an epimorphism.*

If X is an universal cover of finite $B\Gamma$ and the homomorphism ∂ in the Weinberger Conjecture is equivariantly split, then the Novikov conjecture for Γ holds true. The following theorem shows that there is a room for a $n-1$-cocycle in νX.

THEOREM 5. *For every uniformly contractible open n-manifold X^n, $\dim \nu X^n \geq n$.*

The exact sequence of pair implies that the Weinberger Conjecture would hold for X^n if the Higson compactification $\bar{X}$ has trivial rational cohomology: $H^n(\bar{X}; \mathbf{Q}) = 0$. The following theorem sets limits to this approach.

THEOREM 6 [D-F]. *$H^n(\overline{\mathbf{R}^n}; \mathbf{Q}) \neq 0$ and $H^n(\overline{\mathbb{H}^n}; \mathbf{Q}) = 0$ for all $n > 1$.*

Note that $H^n(\overline{\mathbf{R}^n}; \mathbf{Q}) \neq H^n(\overline{\mathbb{H}^n}; \mathbf{Q})$ despite on the fact that $\mathbf{R}^n$ and $\mathbb{H}^n$ are coarse homotopy equivalent.

The following example gives a negative answer to the integral version of Gromov's problem.

EXAMPLE [D-F-W]. *There exists a uniformly contractible riemannian metric d on $\mathbf{R}^8$ such that $(\mathbf{R}^8, d)$ is not hypereuclidean.*

This space $(\mathbf{R}^8, d)$ is coarsely isomorphic to an open cone over a homology sphere X from Theorem 1 (§1). We note that in this example $\dim \nu(\mathbf{R}^8, d) = \infty$ and $\dim_{\mathbf{Z}} \nu(\mathbf{R}^8, d) < \infty$ (see [D-K-U]). Although this example is not of bounded geometry, the Weinberger conjecture holds for it.

§5 DESCENT PRINCIPLE

In this section we compare some of the conditions which enable to prove the Novikov conjecture for certain groups. Let Γ be geometrically finite group and let $X = E\Gamma$ be equipped with a Γ-invariant metric. Each of the following four conditions implies the Novikov conjecture:

(CPI) [C-P]. *There is an equivariant rationally acyclic metrizable compactification $\hat{X}$ of X such that the action of Γ is small at infinity.*

(CPII) [C-P2]. *There is an equivariant rationally acyclic (possibly nonmetrizable) compactification $\hat{X}$ of X with a system of covers α of $Y = \hat{X} \setminus X$ by boundedly saturated sets such that the projection to the inverse limits of the nerves of α induces an isomorphism $H_*(Y; \mathbf{Q}) \to H_*(\lim_{\leftarrow} N(\alpha); \mathbf{Q})$.*

(FW) [F-W], [D-F]. *There is an equivariant Higson dominated compactification $\hat{X}$ of X such that the boundary homomorphism $H^{lf}_*(X; \mathbf{Q}) \to H_{*-1}(\hat{X} \setminus X; \mathbf{Q})$ is an equivariant split injection.*

(HR) [R1]. *There is an equivariant rationally acyclic Higson dominated compactification $\hat{X}$ of X.*

Here H_* stands for the Steenrod homology or its extension for nonmetrizable spaces. An open set $U \subset Y = \hat{X} \setminus X$ is called *boundedly saturated* if for for every

closed set $C \subset \hat{X}$ with $C \cap Y \subset U$ the closure of any r-neighborhood $N_r(C \cap X)$ satisfies $\overline{N_r(C \cap X)} \cap Y \subset U$. The homomorphism $H_*(Y; \mathbf{Q}) \to H_*(\lim_{\leftarrow} N(\alpha); \mathbf{Q})$ in CPII is an isomorphism if the system $\{\alpha\}$ is cofinal. We introduce the condition.

$(CPII')$. There is an equivariant rationally acyclic (possibly nonmetrizable) compactification $\hat{X}$ of X with a cofinal system of covers α of $Y = \hat{X} \setminus X$ by boundedly saturated sets.

We denote by CPI' the condition CPI without an assumption of metrizability of $\hat{X}$. Note that the conditions CPI' and $CPII'$ imply the Novikov conjecture as well.

THEOREM 7. *$CPII' \Rightarrow CPI' \Leftrightarrow CPI \Leftrightarrow HR \Rightarrow FW \Leftarrow CPII$.*

Note that $CPI' \Leftrightarrow HR$ by Proposition 3.

In the integral case one should replace the rational homology by the $\mathbb{L}$-homology. The conditions CPI, II remain without changes, in FW and HR we have to add a metrizability of the corona. Then all four would imply the integral Novikov conjecture. It is not clear whether Theorem 7 holds in the integral case. The problem is in the implication $CPI' \Rightarrow CPI$ which can be reduced to the following.

PROBLEM. *Is a $\mathbb{L}_*$-acyclicity equivalent to a $\mathbb{L}^*$-acyclicity for compact Hausdorff spaces?*

REFERENCES

[A] F.D.Ancel, *The role of countable dimensionality in the theory of cell-like relations*, Trans. Amer. Math. Soc. **287** (1985), 1-40.

[B-C] P. Baum and A. Connes, *K-theory of discrete groups. In D. Evans and M. Takesaki, editors, Operator Algebras and Applications*, Cambridge University Press, 1989, pp. 1-20.

[B-F-M-W] J. Bryant, S. Ferry, W. Mio, and S. Weinberger, *Topology of homology manifolds*, Annals of Mathematics **143** (1996), 435-467.

[B] M. Bestvina, *Local homology properties of boundaries of groups*, Michigan Math.J. **43:1** (1996), 123-139.

[C-P] G. Carlsson and E. Pedersen, *Controlled algebra and the Novikov conjecture for K and L theory*, Topology **34** (1995), 731-758.

[C-P2] G. Carlsson and E. Pedersen, *Čech homology and the Novikov conjectures*, Math. Scand. (1997).

[Dr] A. N. Dranishnikov, *On problem of P.S. Alexandrov*, Math. USSR Sbornik **63:2** (1988), 412-426.

[D-F] A. N. Dranishnikov, S. Ferry, *The Higson-Roe Corona*, Uspehi Mat. Nauk (Russian Math Surveys) **52:5** (1996), 133-146.

[D-F2] A. N. Dranishnikov, S. Ferry, *Cell-like images of topological manifolds and limits of manifolds in Gromov-Hausdorff space*, Preprint (1994).

[D-F-W] A. N. Dranishnikov, S. Ferry and S. Weinberger, *Large Riemannian manifolds which are flexible*, Preprint (1994).

[D-K-U] A. N. Dranishnikov, J. E. Keesling and V. V. Uspenskij, *On the Higson corona of uniformly contractible spaces*, Topology **37:4** (1998), 791-803.

[F-H] F. T. Farrell and W.-C. Hsiang, *On Novikov conjecture for nonpositively curved manifolds*, Ann. Math. **113** (1981), 197-209.

[F-W] S. Ferry and S. Weinberger, *A coarse approach to the Novikov Conjecture*, LMS lecture Notes **226** (1995), 147-163.

[G1] M. Gromov, *Asymptotic invariants for infinite groups*, LMS Lecture Notes **182(2)** (1993).

[G2] M. Gromov, *Large Riemannian manifolds*, Lecture Notes in Math. **1201** (1985), 108-122.

[G-L] M. Gromov and H.B. Lawson, *Positive scalar curvature and the Dirac operator*, Publ. I.H.E.S. **58** (1983), 83-196.

[H] N. Higson, *On the relative K-homology theory of Baum and Douglas*, Preprint (1990).

[H-R] N. Higson and J. Roe, *The Baum-Connes conjecture in coarse geometry*, LMS Lecture Notes **227** (1995), 227-254.

[K-M] J. Kaminiker and J.G. Miller, *A comment on the Novikov conjecture*, Proc.Amer. Math. Soc. **83:3** (1981), 656-658.

[M] A.S. Mischenko, *Homotopy invariants of nonsimply connected manifolds. Rational invariants*, Izv. Akad. Nauk SSSR **30:3** (1970), 501-514.

[N] S.P. Novikov, *On manifolds with free abelian fundamental group and applications (Pontryagin classes, smoothings, high-dimensional knots)*, Izv. Akad. Nauk SSSR **30** (1966), 208-246.

[Ra] A. A Ranicki, *Algebraic L-theory and topological manifolds*, Cambridge University Press, 1992.

[R1] J. Roe, *Coarse cohomology and index theory for complete Riemannian manifolds*, Memoirs Amer. Math. Soc. No. 497, 1993.

[R2] J. Roe, *Index theory, coarse geometry, and topology of manifolds*, CBMS Regional Conference Series in Mathematics, Number 90 (1996).

[Ros1] J. Rosenberg, *C^*-algebras, positive scalar curvature and the Novikov conjecture*, Publ. I.H.E.S. **58** (1983), 409-424.

[Ros2] J. Rosenberg, *Analytic Novikov for topologists*, LMS lecture Notes **226** (1995), 338-372.

[W] C.T.C. Wall, *Surgery on compact manifolds*, Academic Press, New York, 1970.

[Wa] J.J.Walsh, *Dimension, cohomological dimension, and cell-like mappings*, Lecture Notes in Math. 870, 1981, pp. 105–118.

[We] S. Weinberger, *The Topological Classification of Stratified Spaces*, University of Chicago Press, 1995.

[Yu] G. Yu, *The Novikov conjecture and groups with finite asymptotic dimensions*, Preprint (1995).

Alexander N. Dranishnikov
Department of Mathematics
University of Florida
358 Little Hall, PO Box 118105
Gainesville, FL 32611-8105
USA
dranish@math.ufl.edu

Lie Groups and p-Compact Groups

W. G. Dwyer

Abstract. A p-compact group is the homotopical ghost of a compact Lie group; it is the residue that remains after the geometry and algebra have been stripped away. This paper sketches the theory of p-compact groups, with the intention of illustrating the fact that many classical structural properties of compact Lie groups depend only on homotopy theoretic considerations.

1 From compact Lie groups to p-compact groups

The concept of *p-compact group* is the culmination of a series of attempts, stretching over a period of decades, to isolate the key homotopical characteristics of compact Lie groups. It has been something of a problem, as it turns out, to determine exactly what these characteristics are. Probably the first ideas along these lines were due to Hopf [10] and Serre [31].

1.1. Definition. A *finite H-space* is a pair (X, m), where X is a finite CW–complex with basepoint $*$ and $m : X \times X \to X$ is a multiplication map with respect to which $*$ functions, up to homotopy, as a two-sided unit.

The notion of compactness in captured here in the requirement that X be a finite CW–complex. To obtain a structure a little closer to group theory, one might also ask that the multiplication on X be associative up to homotopy. Finite H-spaces have been studied extensively; see [18] and its bibliography. Most of the results deal with homological issues. There are a few general classification theorems, notably Hubbuck's theorem [11] that any path-connected homotopy commutative finite H-space is equivalent at the prime 2 to a torus; this is a more or less satisfying analog of the classical result that any connected abelian compact Lie group is a torus. Experience shows, though, that there is little hope of understanding the totality of all finite H-spaces, or even all homotopy associative ones, on anything like the level of detail that is achieved in the theory of compact Lie groups. The problem is that there are too many finite H-spaces; the structure is too lax.

Stasheff pointed out one aspect of this laxity [32] that is particularly striking when it comes to looking at finite H-spaces as models for group theory. He discovered a whole hierarchy of generalized associativity conditions, all of a homotopy

theoretic nature, which are satisfied by a space with an associative multiplication but not necessarily by a finite H-space. These are called *A_n-conditions* ($n \geq 1$); a space is an H-space if it satisfies condition A_2 and homotopy associative if it satisfies condition A_3. Say that a space X is an A_∞-space if it satisfies condition A_n for all n. The following proposition comes from combining [32] with work of Milnor [22] [21] and Kan [17]; it suggests that that A_∞-spaces are very good models for topological groups. From now on we will use the term *equivalence* for spaces to mean *weak homotopy equivalence.*

1.2. Proposition. *If X is a path-connnected CW–complex, the following four conditions imply one another:*

1. *X is an A_∞-space,*
2. *X is equivalent to a topological monoid,*
3. *X is equivalent to a topological group, and*
4. *X is equivalent to the space ΩY of based loops on some 1-connected pointed space Y.*

In fact, there are bijections between homotopy classes of the four structures. There is a similar result for disconnected X, in which conditions 1 and 2 are expanded by requiring that an appropriate multiplication on $\pi_0 X$ make this set into a group.

If X is a topological group as in 1.2(3), then the space Y of 1.2(4) is the ordinary classifying space $\mathrm{B}X$. Proposition 1.2 leads to the following convenient formulation of the notion "finite A_∞-space" or "homotopy finite topological group". This definition appears in a slightly different form in work of Rector [30].

1.3. Definition. A *finite loop space* is a triple $(X, \mathrm{B}X, e)$, where X is a finite CW–complex, $\mathrm{B}X$ is a pointed space, and $e : X \to \Omega\,\mathrm{B}X$ is an equivalence.

Finite loop spaces appear as if they should be very good homotopy theoretic analogs of Lie groups, but one of the very first theorems about them was pretty discouraging. Rector proved in [29] that there are an *uncountable* number of distinct finite loop space structures on the three-sphere S^3. In other words, he showed that there are an uncountable number of homotopically distinct spaces Y with $\Omega Y \simeq S^3$. This is in sharp contrast to the geometric fact that up to isomorphism there is only *one* Lie group structure on S^3. It suggests that the theory of finite loop spaces is unreasonably complicated.

Rector's method was interesting. For any space X, Bousfield and Kan (also Sullivan) had constructed a *rationalization* $X_\mathbb{Q}$ of X, and $\mathbb{F}_p$-completions $X\hat{}_p$ (p a prime); if X is a simply connected space with finitely generated homotopy groups, then $\pi_i X_\mathbb{Q} \cong \mathbb{Q} \otimes \pi_i X$ and $\pi_i X\hat{}_p \cong \mathbb{Z}_p \otimes \pi_i X$. (Here $\mathbb{Z}_p$ is the ring of p-adic integers.) For such spaces there is a homotopy fibre square on the left

$$\begin{array}{ccc} X & \longrightarrow & \prod_p X\hat{}_p \\ \downarrow & & \downarrow \\ X_\mathbb{Q} & \longrightarrow & (\prod_p X\hat{}_p)_\mathbb{Q} \end{array} \qquad\qquad \begin{array}{ccc} \pi_i X & \longrightarrow & \prod_p \mathbb{Z}_p \otimes \pi_i X \\ \downarrow & & \downarrow \\ \mathbb{Q} \otimes \pi_i X & \longrightarrow & \mathbb{Q} \otimes (\prod_p \mathbb{Z}_p \otimes \pi_i X) \end{array}$$

which is a geometric reflection of the algebraic pullback diagrams on the right. This fibre square, called the *arithmetic square* [33] [3], amounts to a recipe for reconstituting X from its $\mathbb{F}_p$-completions by mixing in rational glue. Rector constructed an uncountable number of loop space structures on S^3 by taking the standard Lie group structure on S^3, $\mathbb{F}_p$-completing to get "standard" loop space structures on each of the spaces $(S^3)^\wedge_p$, and then regluing these standard structures over the rationals in an uncountable number of exotic different ways. In particular, all of his loop space structures become standard after $\mathbb{F}_p$-completion at any prime p. Later on [9] it became clear that this last behavior is unavoidable, since up to homotopy there is only *one* loop space structure on the space $(S^3)^\wedge_p$.

Apparently, then, the theory of finite loop spaces simplifies after $\mathbb{F}_p$-completion, and it is exactly this observation that leads to the definition of p-compact group. The definition uses some terminology. We will say that a space Y is $\mathbb{F}_p$-*complete* if the $\mathbb{F}_p$-completion map $Y \to Y^\wedge_p$ is an equivalence; if Y is simply connected and $\mathrm{H}_*(Y;\mathbb{F}_p)$ is of finite type, then Y is $\mathbb{F}_p$-complete if and only if the homotopy groups of Y are finitely generated modules over $\mathbb{Z}_p$. We will say that Y is $\mathbb{F}_p$-*finite* if $\mathrm{H}_i(Y;\mathbb{F}_p)$ is finite-dimensional for each i and vanishes for all but a finite number of i (in other words, if $\mathrm{H}_*(Y;\mathbb{F}_p)$ looks like the $\mathbb{F}_p$-homology of a finite CW–complex).

1.4. Definition. Suppose that p is a fixed prime number. A *p-compact group* is a triple $(X, \mathrm{B}X, e)$, where X is a space which is $\mathbb{F}_p$-finite, $\mathrm{B}X$ is a pointed space which is $\mathbb{F}_p$-complete, and $e : X \to \Omega\,\mathrm{B}X$ is an equivalence.

Here the idea of "compactness" is expressed in the requirement that X be $\mathbb{F}_p$-finite. Assuming in addition that $\mathrm{B}X$ is $\mathbb{F}_p$-complete is equivalent to assuming that X is $\mathbb{F}_p$-complete and that $\pi_0 X$ is a finite p-group.

1.5. Example. If G is a compact Lie group such that $\pi_0 G$ is a p-group, then the $\mathbb{F}_p$-completion of G is a p-compact group.

The definition of p-compact group is a homotopy theoretic compromise between between the inclination to stay as close as possible to the notion of Lie group, and the desire for an interesting and manageable theory. The reader should note that it is the remarkable machinery of Lannes [19] which makes p-compact groups accessible on a technical level. For instance, the machinery of Lannes lies behind the uniqueness result of [9] referred to above.

Organization of the paper. In section 2 we describe a general scheme for translating from group theory to homotopy theory. Sections 3 and 4 describe the main properties of p-compact groups; almost all of these are parallel to classical properties of compact Lie groups [4]. The final section discusses examples and conjectures.

It is impossible to give complete references or precise credit in a short paper like this one. The basic results about p-compact groups are in [6], [7], [8], [23], and [25]. There is a treatment of compact Lie groups based on homotopy theoretic arguments in [4]. The interested reader should look at the survey articles [20], [24], and [26], as well as their bibliographies, for additional information.

1.6. Terminology. There are a few basic topological issues which it is worth pointing out. We assume that all spaces have been replaced if necessary by equivalent CW–complexes. If $f : X \to Y$ is a map of spaces, then $\mathrm{Map}(X, Y)_f$ is the component containing f of the space of maps $X \to Y$. The space $\mathrm{Aut}_f(X)$ is the space of self-equivalences of X over Y; to obtain homotopy invariance, this is constructed by replacing f by an equivalent Serre fibration $f' : X' \to Y$ and forming the space of self homotopy equivalences $X' \xrightarrow{\sim} X'$ which commute with f'.

The notation $\mathrm{H}^*_{\mathbb{Q}_p}(Y)$ stands for $\mathbb{Q} \otimes \mathrm{H}^*(Y; \mathbb{Z}_p)$; this is a variant of rational cohomology which is better-behaved than ordinary rational cohomology for spaces Y which are $\mathbb{F}_p$-complete.

2 A dictionary between group theory and homotopy theory

We now set up a dictionary which will allow us to talk about p-compact groups in ordinary algebraic terms. We begin with concepts that apply to loop spaces in general (a *loop space* is a triple $(X, \mathrm{B}X, e)$ with $e : X \xrightarrow{\sim} \Omega\,\mathrm{B}X$) and then specialize to p-compact groups. From now on we will refer to a loop space or p-compact group $(X, \mathrm{B}X, e)$ as a space X with some (implicit) extra structure.

2.1. Definition. Suppose that X and Y are loop spaces.

- A *homomorphism* $f : X \to Y$ is a pointed map $\mathrm{B}f : \mathrm{B}X \to \mathrm{B}Y$. Two homomorphisms $f, f' : X \to Y$ are *conjugate* if $\mathrm{B}f$ and $\mathrm{B}f'$ are homotopic.
- The *homogeneous space* $Y/f(X)$ (denoted Y/X if f is understood) is the homotopy fibre of $\mathrm{B}f$.
- The *centralizer of* $f(X)$ *in* Y, denoted $\mathcal{C}_Y(f(X))$ or $\mathcal{C}_Y(X)$, is the loop space $\Omega\,\mathrm{Map}(\mathrm{B}X, \mathrm{B}Y)_{\mathrm{B}f}$.
- The *Weyl Space* $\mathcal{W}_Y(X)$ is the space $\mathrm{Aut}_{\mathrm{B}f}(\mathrm{B}X)$; this is in fact a a loop space, essentially because it is an associative monoid under composition (1.2). The *normalizer* $\mathcal{N}_Y(X)$ of X in Y is the loop space of the homotopy orbit space of the action of $\mathcal{W}_Y(X)$ on $\mathrm{B}X$ by composition.
- A *short exact sequence* $X \to Y \to Z$ of loop spaces is a fibration sequence $\mathrm{B}X \to \mathrm{B}Y \to \mathrm{B}Z$; Y is said to be an *extension* of Z by X.

2.2. Remark. If X and Y are discrete groups, treated as loop spaces via 1.2, and $f : X \to Y$ is an ordinary homomorphism, then the above definitions specialize to the usual notions of coset space, centralizer, normalizer, and short exact sequence, at least if $X \to Y$ is injective. It is not hard to see that in general there are natural loop space homomorphisms $\mathcal{C}_Y(X) \to \mathcal{N}_Y(X) \to Y$; the homomorphism $\mathcal{C}_Y(X) \to Y$, for instance, amounts to the map $\mathrm{Map}(\mathrm{B}X, \mathrm{B}Y)_{\mathrm{B}f} \to \mathrm{B}Y$ given by evaluation at the basepoint of $\mathrm{B}X$. There is always a short exact sequence $X \to \mathcal{N}_Y(X) \to \mathcal{W}_Y(X)$.

The key additional definitions for p-compact groups are the following ones.

2.3. Definition. A p-compact group X is *a p-compact torus* if X is the $\mathbb{F}_p$-completion of an ordinary torus, and *a p-compact toral group* if X is an extension of a finite p-group by a p-compact torus. If $f : X \to Y$ is a homomorphism of p-compact groups, then f is a *monomorphism* if $Y/f(X)$ is $\mathbb{F}_p$-finite.

3 Maximal tori and cohomology rings

If X is a p-compact group, a *subgroup Y of X* is a p-compact group Y and a monomorphism $i : Y \to X$ (i is called a *subgroup inclusion*). In general, if $f : Y \to X$ is a homomorphism of p-compact groups, the associated loop space homomorphism $g : \mathcal{C}_X(Y) \to X$ is not obviously a subgroup inclusion; it is not even clear that $\mathcal{C}_X(Y)$ is a p-compact group. For special choices of Y, though, the situation is nicer.

3.1. Proposition. *Suppose that $f : Y \to X$ is a homomorphism of p-compact groups, and that Y is a p-compact toral group. Then $\mathcal{C}_X(Y) \to X$ is a subgroup inclusion.*

A p-compact group is said to be *abelian* if the natural map $\mathcal{C}_X(X) \to X$ is an equivalence.

3.2. Proposition. *A p-compact group is abelian if and only if it is the product of a p-compact torus and a finite abelian p-group. If A is an abelian p-compact group and $f : A \to X$ is a homomorphism, then f naturally lifts over the subgroup inclusion $\mathcal{C}_X(A) \to X$ to a homomorphism $f' : A \to \mathcal{C}_X(A)$.*

A subgroup Y of X is said to be an *abelian subgroup* if Y is abelian, or a *torus in X* if Y is a p-compact torus. If Y' is another subgroup of X, Y' is said to be *contained in Y up to conjugacy* if the homomorphism $Y' \to X$ lifts up to conjugacy to a homomorphism $Y' \to Y$.

3.3. Definition. A torus T in X is said to be a *maximal torus* if any other torus T' in X is contained in T up to conjugacy.

We will say that an abelian subgroup A of X is *self-centralizing* if the map $A \to \mathcal{C}_X(A)$ is an equivalence. If Z is a space which is $\mathbb{F}_p$-finite, the *Euler characteristic* $\chi(Z)$ is the usual alternating sum of the ranks of the $\mathbb{F}_p$ homology groups of Z.

3.4. Proposition. *Suppose that X is a p-compact group and that T is a torus in X. Then T is maximal if and only if $\chi(X/T) \neq 0$. If X is connected, then T is maximal if and only if T is self-centralizing.*

3.5. Proposition. *Any p-compact group X has a maximal torus T, unique up to conjugacy.*

A space is said to be *homotopically discrete* if each of its components is contractible.

3.6. Proposition. *Suppose that X is a p-compact group with maximal torus T. Then the Weyl space $\mathcal{W}_X(T)$ is homotopically discrete, and $\pi_0\mathcal{W}_X(T)$, with the natural composition operation, is a finite group.*

If T is a maximal torus for X, the finitely generated free $\mathbb{Z}_p$-module $\pi_1 T$ is called the *dual weight lattice* L_X of X; its rank as a free module is the *rank* $\mathrm{rk}(X)$ of X. The finite group appearing in 3.6 is called the *Weyl group* of X and denoted W_X; by definition, W_X acts on L_X.

3.7. Definition. If M is a finitely generated free module over a domain R (such as $\mathbb{Z}_p$), an automorphism α of M is said to be a *reflection* (or sometimes a *pseudoreflection* or *generalized reflection*) if the endomorphism $(\alpha - \mathrm{Id})$ of M has rank one. A subgroup of $\mathrm{Aut}(M)$ is said to be *generated by reflections* if it is generated as a group by the reflections it contains.

3.8. Proposition. *Suppose that X is a connected p-compact group of rank r. Then the action of W_X on L_X is faithful and represents W_X as a finite subgroup of $\mathrm{GL}_r(\mathbb{Z}_p)$ generated by reflections.*

3.9. Proposition. *Suppose that X is a connected p-compact group with maximal torus T, Weyl group W, and rank r. Then the cohomology rings $\mathrm{H}^*_{\mathbb{Q}_p}(\mathrm{B}T)$ and $\mathrm{H}^*_{\mathbb{Q}_p}(\mathrm{B}X)$ are polynomial algebras over $\mathbb{Q}_p$ of rank r, and the natural restriction map $\mathrm{H}^*_{\mathbb{Q}_p}(\mathrm{B}X) \to \mathrm{H}^*_{\mathbb{Q}_p}(\mathrm{B}T)^W$ is an isomorphism.*

3.10. Proposition. *If X is a p-compact group, then the cohomology ring $\mathrm{H}^*(\mathrm{B}X; \mathbb{F}_p)$ is finitely generated as an algebra over $\mathbb{F}_p$.*

4 Centers and product decompositions

A *product decomposition* of a p-compact group X is a way of writing X up to homotopy as a product of two p-compact groups, or, equivalently, a way of writing $\mathrm{B}X$ up to homotopy as a product of spaces. The most general product theorem is the following one.

4.1. Proposition. *If X is a connected p-compact group, then there is a natural bijection between product decompositions of X and product decompositions of L_X as a module over W_X.*

In general, connected p-compact groups are constructed from indecomposable factors in much the same way that Lie groups are, by twisting the factors together over a finite central subgroup.

4.2. Definition. A subgroup Y of a p-compact group X is said to be *normal* if the usual map $\mathcal{N}_X(Y) \to X$ is an equivalence. The subgroup Y is *central* if the usual map $\mathcal{C}_X(A) \to X$ is an equivalence.

If the subgroup Y of X is normal, then there is a loop space structure on X/Y (because X/Y is equivalent to the Weyl space $\mathcal{W}_X(Y)$) and a short exact sequence $Y \to X \to X/Y$ of p-compact groups.

4.3. Proposition. *Any central subgroup of a p-compact group X is both abelian and normal; moreover, there exists up to homotopy a unique maximal central subgroup $\mathcal{Z}_X$ of X (called the* center *of X). The center of X can be identified as $\mathcal{C}_X(X)$.*

The center of X is maximal in the sense that up to conjugacy it contains *any* central subgroup A of X.

4.4. Proposition. *If X is a connected p-compact group, the quotient $X/\mathcal{Z}_X$ has trivial center.*

If X is connected, the quotient $X/\mathcal{Z}_X$ is called the *adjoint form* of X. For connected X there is a simple way to compute $\mathcal{Z}_X$ from what amounts to ordinary algebraic data associated to the normalizer $\mathcal{N}_X(T)$ of a maximal torus T in X.

4.5. Definition. A connected p-compact group X is said to be *almost simple* if the action of W_X on $\mathbb{Q}\otimes L_X$ affords an irreducible representation of W_X over $\mathbb{Q}_p$; X is *simple* if X is almost simple and $\mathcal{Z}_X = \{e\}$.

4.6. Proposition. *Any 1-connected p-compact group is equivalent to a product of almost simple p-compact groups. The product decomposition is unique up to permutation of factors.*

4.7. Proposition. *Any connected p-compact group with trivial center is equivalent to a product of simple p-compact groups. The product decomposition is unique up to permutation of factors.*

4.8. Proposition. *Any connected p-compact group is equivalent to a p-compact group of the form*

$$(T \times X_1 \times \cdots \times X_n)/A$$

where T is a p-compact torus, each X_i is a 1-connected almost simple p-compact group, and A is a finite abelian p-subgroup of the center of the indicated product.

5 Examples and conjectures

Call a connected p-compact group *exotic* if it is not equivalent to the $\mathbb{F}_p$-completion of a connected compact Lie group. The reader may well ask whether there are any exotic p-compact groups, or whether on the other hand the study of p-compact groups is just a way of doing ordinary Lie theory under artificially difficult circumstances. In fact, there are many exotic examples: Sullivan constructed loop space structures on the $\mathbb{F}_p$-completions of various odd spheres S^n $(n > 3)$ [33], and it is possible to do more elaborate things along the same lines, see, e.g., [1] and [5].

Conjecturally, the theory splits into two parts.

5.1. Conjecture. *Any connected p-compact group can be written as a product $X_1 \times X_2$, where X_1 is the $\mathbb{F}_p$-completion of a compact Lie group and X_2 is a product of exotic simple p-compact groups.*

In addition, all of the exotic examples are conjecturally known.

5.2. Conjecture. *The exotic simple p-compact groups correspond bijectively, up to equivalence, to the exotic p-adic reflection groups of Clark and Ewing [1].*

Here a p-adic reflection group is said to be *exotic* if it is not derived from the Weyl group of a connected compact Lie group. For example, it would follow from 5.2 there is up to equivalence only *one* exotic simple 2-compact group, the one constructed in [5]. Closely related to the above conjectures is the following one.

5.3. Conjecture. *Let X be a connected p-compact group with maximal torus T. Then X is determined up to equivalence by the loop space $\mathcal{N}_X(T)$.*

This would the analog for p-compact groups of a Lie-theoretic result of Curtis, Wiederhold and Williams [2]. It is easy to see that the loop space $\mathcal{N}_X(T)$ is determined by the Weyl group W_X, the p-adic lattice L_X, and an extension class in $\mathrm{H}^3(W_X; L_X)$. Explicit calculation with examples shows that if p is odd the extension class vanishes. It would be very interesting to find a simple, direct way to construct a connected p-compact group X from $\mathcal{N}_X(T)$. For a connected compact Lie group G, this would (according to [2]) give a direct way to constuct the homotopy type of BG, or at least the $\mathbb{F}_p$-completion of this homotopy type, from combinatorial data associated to the root system of G. All constructions of this type which are known to the author involve building a Lie algebra and then exponentiating it; this kind of procedure does not generalize to p-compact groups.

The strongest results along the lines of 5.3 are due to Notbohm [27] [28].

The theory of homomorphisms between general p-compact groups is relatively undeveloped, though there is a lot of information available if the domain is a p-compact toral group or if the homomorphism is a rational equivalence [12] [14]. The general situation seems complicated [13], but it might be possible to find some analog for p-compact groups of the results of Jackowski and Oliver [15] on "homotopy representations" of compact Lie groups (see for instance [16]).

References

[1] A. Clark and J. Ewing, *The realization of polynomial algebras as cohomology rings*, Pacific J. Math. 50 (1974), 425–434.

[2] M. Curtis, A. Wiederhold, and B. Williams, *Normalizers of maximal tori*, (1974), 31–47. Lecture Notes in Math., Vol. 418.

[3] E. Dror, W. G. Dwyer, and D. M. Kan, *An arithmetic square for virtually nilpotent spaces*, Illinois J. Math. 21 (1977), no. 2, 242–254.

[4] W. G. Dwyer and C. W. Wilkerson, *The elementary geometric structure of compact Lie groups*, Bull. L. M. S., to appear.

[5] ———, *A new finite loop space at the prime two*, J. Amer. Math. Soc. 6 (1993), no. 1, 37–64.

[6] ———, *Homotopy fixed-point methods for Lie groups and finite loop spaces*, Ann. of Math. (2) 139 (1994), no. 2, 395–442.

[7] ———, *The center of a p-compact group*, The Čech centennial (Boston, MA, 1993), Contemp. Math., vol. 181, Amer. Math. Soc., Providence, RI, 1995, pp. 119–157.

[8] ———, *Product splittings for p-compact groups*, Fund. Math. 147 (1995), no. 3, 279–300.

[9] W. G. Dwyer, H. R. Miller, and C. W. Wilkerson, *The homotopic uniqueness of BS^3*, Algebraic topology, Barcelona, 1986, Lecture Notes in Math., vol. 1298, Springer, Berlin, 1987, pp. 90–105.

[10] H. Hopf, *Über die Topologie der Gruppen-Mannigfaltigkeiten und ihre Verallgemeinerungen*, Ann. of Math. (2) 42 (1941), 22–52.

[11] J. R. Hubbuck, *On homotopy commutative H-spaces*, Topology 8 (1969), 119–126.

[12] S. Jackowski, J. McClure, and B. Oliver, *Homotopy classification of self-maps of BG via G-actions. I*, Ann. of Math. (2) 135 (1992), no. 1, 183–226.

[13] ———, *Maps between classifying spaces revisited*, The Čech centennial (Boston, MA, 1993), Contemp. Math., vol. 181, Amer. Math. Soc., Providence, RI, 1995, pp. 263–298.

[14] ———, *Self-homotopy equivalences of classifying spaces of compact connected Lie groups*, Fund. Math. 147 (1995), no. 2, 99–126.

[15] S. Jackowski and B. Oliver, *Vector bundles over classifying spaces of compact Lie groups*, Acta Math. 176 (1996), no. 1, 109–143.

[16] A. Jeanneret and A. Osse, *The K-theory of p-compact groups*, Comment. Math. Helv. 72 (1997), no. 4, 556–581.

[17] D. M. Kan, *A combinatorial definition of homotopy groups*, Ann. of Math. (2) 67 (1958), 282–312.

[18] R. M. Kane, *The homology of Hopf spaces*, North-Holland Mathematical Library, vol. 40, North-Holland Publishing Co., Amsterdam, 1988.

[19] J. Lannes, *Sur les espaces fonctionnels dont la source est le classifiant d'un p-groupe abélien élémentaire*, Inst. Hautes Études Sci. Publ. Math. (1992), no. 75, 135–244, With an appendix by Michel Zisman.

[20] ———, *Théorie homotopique des groupes de Lie (d'après W. G. Dwyer et C. W. Wilkerson)*, Astérisque (1995), no. 227, Exp. No. 776, 3, 21–45, Séminaire Bourbaki, Vol. 1993/94.

[21] J. Milnor, *Construction of universal bundles. I*, Ann. of Math. (2) 63 (1956), 272–284.

[22] ———, *Construction of universal bundles. II*, Ann. of Math. (2) 63 (1956), 430–436.

[23] J. M. Møller and D. Notbohm, *Centers and finite coverings of finite loop spaces*, J. Reine Angew. Math. 456 (1994), 99–133.

[24] J. M. Møller, *Homotopy Lie groups*, Bull. Amer. Math. Soc. (N.S.) 32 (1995), no. 4, 413–428.

[25] D. Notbohm, *Unstable splittings of classifying spaces of p-compact groups*, preprint (Göttingen) 1994.

[26] ———, *Classifying spaces of compact Lie groups and finite loop spaces*, Handbook of algebraic topology, North-Holland, Amsterdam, 1995, pp. 1049–1094.

[27] D. Notbohm, *Homotopy uniqueness of classifying spaces of compact connected Lie groups at primes dividing the order of the Weyl group*, Topology 33 (1994), no. 2, 271–330.

[28] ———, *Topological realization of a family of pseudoreflection groups*, Fund. Math. 155 (1998), no. 1, 1–31.

[29] D. L. Rector, *Loop structures on the homotopy type of S^3*, (1971), 99–105. Lecture Notes in Math., Vol. 249.

[30] ———, *Subgroups of finite dimensional topological groups*, J. Pure Appl. Algebra 1 (1971), no. 3, 253–273.

[31] J.-P. Serre, *Homologie singulière des espaces fibrés. Applications*, Ann. of Math. (2) 54 (1951), 425–505.

[32] J. D. Stasheff, *Homotopy associativity of H-spaces. I, II*, Trans. Amer. Math. Soc. 108 (1963), 275-292; ibid. 108 (1963), 293–312.

[33] D. Sullivan, *Genetics of homotopy theory and the Adams conjecture*, Ann. of Math. (2) 100 (1974), 1–79.

William G. Dwyer
Department of Mathematics
University of Notre Dame
Notre Dame, Indiana 46556
USA
dwyer.1@nd.edu

Constructions of Smooth 4-Manifolds

Ronald Fintushel[1] and Ronald J. Stern[2]

Abstract. We describe a collection of constructions which illustrate a panoply of "exotic" smooth 4-manifolds.

1991 Mathematics Subject Classification: 57R55

1. Introduction

At the time of the previous (1994) International Congress of Mathematicians, steady, but slow, progress was being made on the classification of simply connected closed smooth 4-manifolds. In particular, the Donaldson invariants had begun to take a particularly nice form [13] (also [4]), their computations were becoming more routine [3], and their behavior under blowing up (i.e. taking connected sum with $\overline{\mathbf{CP}}^2$) was well understood [2]. Due to the complexity of the Donaldson invariants, great hope was held out that an even better understanding of these invariants would close the books on the classification of simply connected 4-manifolds.

A few short months after the 1994 ICM, the 4-manifold community was blindsided by the introduction of the now famous Seiberg-Witten equations [28]. Most of the results obtained by using Donaldson theory were found to have quicker, and sometimes more general, counterparts using the Seiberg-Witten technology. The potential applications of the difficult Donaldson technology became much more transparent using these new equations. As of July 1998, there is good news as well as bad news. The good news is that many of the earlier focus problems have been solved. In particular, the Thom conjecture [14] and its natural generalizations have been verified [20, 21]; also the study of symplectic 4-manifolds has taken a more central role [23, 24, 25, 26]. The bad news is that recent constructions and computations indicate that the Seiberg-Witten and Donaldson theories are too weak to distinguish simply connected smooth 4-manifolds [6]. It is these latter constructions and computations that we will discuss at this 1998 International Congress of Mathematicians. It is becoming more apparent that we are seeing only a small constellation of 4-dimensional manifolds. More seriously, we are lacking a reasonable conjectural classification of simply connected closed smooth 4-manifolds.

Current technology has given us many more 4-manifolds than had been expected in 1994. The authors hope that during the 2002 ICM the construction of large classes of new 4-manifolds will be discussed; in particular, they hope that a sufficiently large collection of 4-manifolds will have been discovered so as to allow

[1]Partially supported by NSF Grant DMS9704927
[2]Partially supported by NSF Grant DMS9626330

for some general patterns to emerge and, at least, a conjectural classification to again be on the books.

2. The knot surgery construction

Let X be a simply connected oriented smooth closed 4-manifold. Its most basic invariant is its intersection form

$$Q_X : H_2(X;\mathbf{Z}) \otimes H_2(X;\mathbf{Z}) \to \mathbf{Z}$$

defined by counting signed transverse intersections of embedded oriented surfaces representing given homology classes. It is a famous theorem of M. Freedman [10] that Q_X determines the homeomorphism type of X, and an equally renowned theorem of S.K. Donaldson [1] that Q_X is not sufficient to determine the diffeomorphism type of X. In this section we shall discuss geometric operations on a given smooth 4-manifold which preserve the underlying topological structure and alter its smooth structure. In particular, we shall consider the following construction: Let X be a simply connected smooth 4-manifold which contains a smoothly embedded torus T of self-intersection 0. Given a knot K in S^3, we replace a tubular neighborhood of T with $S^1 \times (S^3 \setminus K)$ to obtain the *knot surgery manifold* X_K.

More formally, this procedure is accomplished by performing 0-framed surgery on K to obtain the 3-manifold M_K. The meridian m of K can be viewed as a circle in M_K; so in $S^1 \times M_K$ we have the smooth torus $T_m = S^1 \times m$ of self-intersection 0. Since a neighborhood of m has a canonical framing in M_K, a neighborhood of the torus T_m in $S^1 \times M_K$ has a canonical identification with $T_m \times D^2$. The knot surgery manifold X_K is given by the fiber sum

$$X_K = X \#_{T=T_m} S^1 \times M_K = (X \setminus T \times D^2) \cup (S^1 \times M_K \setminus T_m \times D^2)$$

where the two pieces are glued together so as to preserve the homology class $[\mathrm{pt} \times \partial D^2]$. This latter condition does not, in general, completely determine the isotopy type of the gluing, and X_K is taken to be any manifold constructed in this fashion.

Because $S^1 \times (S^3 \setminus K)$ has the same homology as a tubular neighborhood of T in X (and because the gluing preserves $[\mathrm{pt} \times \partial D^2]$) the homology and intersection form of X_K will agree with that of X. If it is also assumed that $X \setminus T$ is simply connected, then $\pi_1(X_K) = 1$; so X_K will be homeomorphic to X.

In order to distinguish the diffeomorphism types of the X_K, we rely on Seiberg-Witten invariants. We view the Seiberg-Witten invariant of a smooth 4-manifold as a multivariable (Laurent) polynomial. To do this, recall that the Seiberg-Witten invariant of a smooth closed oriented 4-manifold X with $b_2^+(X) > 1$ is an integer-valued function which is defined on the set of *spin*^{c} structures over X (cf. [28]). In case $H_1(X,\mathbf{Z})$ has no 2-torsion (for example, as here where X is simply connected) there is a natural identification of the *spin*^{c} structures of X with the characteristic elements of $H_2(X,\mathbf{Z})$ (i.e. those elements k whose Poincaré duals $\hat{k}$ reduce mod 2 to $w_2(X)$). In this case we view the Seiberg-Witten invariant as

$$\mathrm{SW}_X : \{k \in H_2(X,\mathbf{Z}) | \hat{k} \equiv w_2(TX) \pmod 2)\} \to \mathbf{Z}.$$

The sign of $\mathcal{SW}_X$ depends on an orientation of $H^0(X,\mathbf{R}) \otimes \det H^2_+(X,\mathbf{R}) \otimes \det H^1(X,\mathbf{R})$. If $\mathrm{SW}_X(\beta) \neq 0$, then β is called a *basic class* of X. It is a fundamental fact that the set of basic classes is finite. Furthermore, if β is a basic class, then so is $-\beta$ with $\mathrm{SW}_X(-\beta) = (-1)^{(\mathrm{e+sign})(X)/4}\,\mathrm{SW}_X(\beta)$ where $\mathrm{e}(X)$ is the Euler number and $\mathrm{sign}(X)$ is the signature of X.

Now let $\{\pm\beta_1,\ldots,\pm\beta_n\}$ be the set of nonzero basic classes for X. Consider variables $t_\beta = \exp(\beta)$ for each $\beta \in H^2(X;\mathbf{Z})$ which satisfy the relations $t_{\alpha+\beta} = t_\alpha t_\beta$. We may then view the Seiberg-Witten invariant of X as the Laurent polynomial

$$\mathcal{SW}_X = \mathrm{SW}_X(0) + \sum_{j=1}^{n} \mathrm{SW}_X(\beta_j)\cdot(t_{\beta_j} + (-1)^{(\mathrm{e+sign})(X)/4}\, t_{\beta_j}^{-1}).$$

As an example of this notational device, consider the simply connected minimally elliptic surface $E(n)$ with holomorphic Euler characteristic n and no multiple fibers. Its Seiberg-Witten invariant is $\mathcal{SW}_{E(n)} = (t-t^{-1})^{n-2}$ where $t = t_F$ for F the fiber class. Thus, $\mathrm{SW}_{E(n)}((n-2m)F) = (-1)^{m-1}\binom{n-2}{m-1}$ for $m = 1,\ldots,n-1$ and $\mathrm{SW}_{E(n)}(\alpha) = 0$ for any other α. When $b^+(X) > 1$, the Laurent polynomial $\mathcal{SW}_X$ is a diffeomorphism invariant of X.

For our theorem, we need to place a mild hypothesis on the embedded torus T. We say that a smoothly embedded torus representing a nontrivial homology class $[T]$ is *c-embedded* if there is a neighborhood N of T in X and a diffeomorphism $\varphi : N \to U$ where U is a neighborhood of a cusp fiber in an elliptic surface and $\varphi(T)$ is a smooth elliptic fiber in U. Equivalently, T is c-embedded if it contains two simple closed curves which generate $\pi_1(T)$ and which bound vanishing cycles in X. Note that a c-embedded torus has self-intersection 0.

THEOREM 2.1 ([6]). *Let X be a simply connected oriented smooth 4-manifold with $b^+ > 1$. Suppose that X contains a c-embedded torus T with $\pi_1(X \setminus T) = 1$, and let K be any knot in S^3. Then the knot surgery manifold X_K is homeomorphic to X and has Seiberg-Witten invariant*

$$\mathcal{SW}_{X_K} = \mathcal{SW}_X \cdot \Delta_K(t)$$

where $\Delta_K(t)$ is the symmetrized Alexander polynomial of K and $t = \exp(2[T])$.

For example, the theorem applies to the K3-surface $E(2)$ where T is a smooth elliptic fiber, and since $\mathcal{SW}_{E(2)} = 1$, we have $\mathcal{SW}_{E(2)_K} = \Delta_K(t)$. It is a theorem of Seifert that any Laurent polynomial of the form $P(t) = a_0 + \sum_{j=1}^{n} a_j(t^j + t^{-j})$ with coefficient sum $P(1) = \pm 1$ is the Alexander polynomial of some knot in S^3. Call such a Laurent polynomial an *A-polynomial*. It follows that if (X,T) satisfies the hypothesis of Theorem 2.1, then for any A-polynomial $P(t)$, there is a smooth simply connected 4-manifold X_P which is homeomorphic to X and has Seiberg-Witten invariant $\mathcal{SW}_{X_P} = \mathcal{SW}_X \cdot P(t)$ where $t = \exp(2[T])$. In particular, for each A-polynomial $P(t)$, there is a manifold homeomorphic to the K3-surface with $\mathcal{SW} = P(t)$.

The relationship between Seiberg-Witten type invariants and the Alexander polynomial was first discovered by Meng and Taubes. In [17] they showed that the 3-manifold Seiberg-Witten invariant is related to Milnor torsion.

If one starts with a fibered knot K, then $S^1 \times M_K$ is a surface bundle over a torus and thus carries a symplectic structure [27] for which T_m is a symplectic submanifold. Thus if X is a symplectic 4-manifold containing a c-embedded symplectic torus T, then $X_K = X\#_{T=T_m} S^1 \times M_K$ is also symplectic [11, 16]. In a fashion similar to the treatment of the Seiberg-Witten invariant as a Laurent polynomial, one can view the Gromov invariant of a symplectic 4-manifold X as a polynomial $\mathcal{G}r_X = \sum \mathrm{Gr}_X(\beta)\, t_\beta$ where $\mathrm{Gr}_X(\beta)$ is the usual Gromov invariant of β. Let $\mathrm{A}_K(t) = t^d \Delta_K(t)$ denote the normalized Alexander polynomial, where d is the degree of $\Delta_K(t)$. As a corollary to Theorem 2.1 and the theorems of Taubes relating the Seiberg-Witten and Gromov invariants of a symplectic 4−manifold [25, 26] we have:

COROLLARY 2.2 ([6]). *Let X be a symplectic 4-manifold with $b^+ > 1$ containing a symplectic c-embedded torus T. If K is a fibered knot, then X_K is a symplectic 4-manifold whose Gromov invariant is $\mathcal{G}r_{X_K} = \mathcal{G}r_X \cdot A_K(\tau)$ where $\tau = \exp([T])$.*

This last calculation can also be made purely within the realm of symplectic topology [12, 15]. Our interest is directed more to the opposite situation. The Alexander polynomial of a fibered knot is monic; i.e. its top coefficient is ± 1. On the other hand:

COROLLARY 2.3 ([6]). *If $\Delta_K(t)$ is not monic, then X_K does not admit a symplectic structure. Furthermore, if X contains a homologically nontrivial surface Σ_g of genus g disjoint from T with $[\Sigma_g]^2 < 2 - 2g$ if $g > 0$ or $[\Sigma_g]^2 < 0$ if $g = 0$, then X_K with the opposite orientation does not admit a symplectic structure.*

Until the summer of 1996, it was still a plausible conjecture (sometimes called the 'minimal conjecture') that each irreducible simply connected 4-manifold should admit a symplectic structure with one of its orientations. The first counterexamples to this conjecture were constructed by Z. Szabo [22]. The knot surgery construction gives a multitude of examples of simply connected irreducible 'nonsymplectic' 4-manifolds. In fact, if X is simply connected with $\mathcal{SW}_X \neq 0$ and if X contains a c-embedded torus T with $\pi_1(X \setminus T) = 1$, then Theorem 2.1 and Corollary 2.3 imply that there are infinitely many distinct nonsymplectic smooth 4-manifolds X_K homeomorphic to X.

If K_1 and K_2 have the same Alexander polynomial, Seiberg-Witten invariants are not able to distinguish X_{K_1} from X_{K_2}. For example, take $X = E(2)$. Then X_K has a self-intersection 0 homology class σ satisfying $\sigma \cdot [T] = 1$ which is represented by an embedded surface of genus $g(K) + 1$ where $g(K)$ is the genus of K. One might hope that these classes could be used to distinguish X_{K_1} from X_{K_2} when $g(K_1) \neq g(K_2)$.

CONJECTURE . *For $X = E(2)$, the manifolds X_{K_1} and X_{K_2} are diffeomorphic if and only if K_1 and K_2 are equivalent knots.*

The proof of Theorem 2.1 proceeds by successively simplifying the manifold X_K in a fashion which mimics the calculation of the Alexander polynomial of K

via skein relations. Recall that $\Delta_K(t)$ can be calculated via the relation

$$\Delta_{K_+}(t) = \Delta_{K_-}(t) + (t^{1/2} - t^{-1/2}) \cdot \Delta_{K_0}(t) \tag{1}$$

where K_+ is an oriented knot or link, K_- is the result of changing a single oriented positive (right-handed) crossing in K_+ to a negative (left-handed) crossing, and K_0 is the result of resolving the crossing as shown in Figure 1.

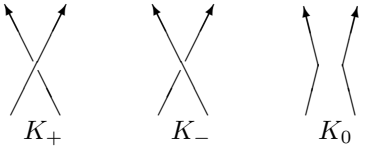


Figure 1

The point of using (1) to calculate Δ_K is that K can be simplified to an unknot via a sequence of crossing changes. One builds a 'resolution tree' starting from K and at each stage adding the bifurcation of Figure 2, where each K_+, K_-, K_0 is a knot or 2-component link, and so that at the bottom of the tree, there are only unknots, and split links. Then, because the Alexander polynomial of an unknot is 1, and is 0 for a split link (of more than one component) one can can work backwards using (1) to calculate $\Delta_K(t)$.

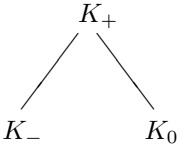


Figure 2

The manifold X_{K_+} can be obtained from X_{K_-} by means of a (+1)-log transform on a nullhomologous torus in X_{K_-}, and then the gluing theorems of [18] show that $\mathcal{SW}_{X_{K_+}}$ can be computed in terms of the Seiberg-Witten invariants of X_{K_-} and a manifold $X_{K_-,0}$ obtained by a 0-log transform on X_{K_-}. With some work, this leads to a related resolution diagram of 4-manifolds where each knot K' corresponds to $X_{K'}$, and this diagram can be used to prove Theorem 2.1.

We conclude this section by pointing out that the knot surgery construction can be generalized to manifolds with $b^+ = 1$ and to links in S^3 of more than one component in a more-or-less obvious way. One glues the complements of c-embedded tori in 4-manifolds to the product of S^1 with the link complement. See [6] for details. For example, if to each boundary component of $S^1 \times (S^3 \setminus N(L))$ we glue $E(1)$ minus the neighborhood of a smooth elliptic fiber, we obtain a manifold with $\mathcal{SW} = \Delta_L(t_1, \ldots, t_n)$, the multivariable Alexander polynomial of the link. Szabo's examples in [22] can be obtained from this construction.

3. Embeddings of surfaces in 4-manifolds

Knot surgery can also be used to change the embedding of a surface in a fixed 4-manifold. To motivate the construction, note that one can tie a knot in the core $\{0\} \times I$ of a cylinder $D^2 \times I$ by removing a tubular neighborhood of a meridian circle

and replacing it with a knot complement $S^3 \setminus K$. We shall perform a parametrized version of this construction in the 4-manifold setting. Consider an oriented surface Σ of genus $g > 0$ which is smoothly embedded in a simply connected 4-manifold X. Let α be a simple closed curve on Σ which is part of a symplectic basis, and let $\alpha \times I$ be an annular neighborhood of α in Σ. In X we see the neighborhood $D^2 \times \alpha \times I$. For a fixed knot K in S^3, we parametrize the above construction so as to perform it on each of the cylinders $D^2 \times \{y\} \times I$, $y \in \alpha$, to obtain an embedded surface Σ_K. This is equivalent to performing knot surgery on the (nullhomologous) rim torus $R = \partial D^2 \times \alpha$. We call this operation *rim surgery*.

THEOREM 3.1 ([7]). *Let X be a simply connected smooth 4-manifold with an embedded surface Σ of positive genus. Suppose that $\pi_1(X \setminus \Sigma) = 1$. Then for each knot K in S^3, rim surgery produces a surface Σ_K, and there is a homeomorphism $(X, \Sigma) \cong (X, \Sigma_K)$.*

The Seiberg-Witten invariant can be used to study these embeddings, but first, an auxilliary construction is needed. For each positive integer g, let Y_g be the union of the Milnor fiber of the $(2, 2g+1, 4g+1)$ Brieskorn singularity and a generalized nucleus consisting of the 4-manifold obtained as the trace of the 0-framed surgery on $(2, 2g+1)$ torus knot in ∂B^4 and a -1 surgery on a meridian. Then Y_g is a Kahler surface and admits a holomorphic fibration over $\mathbf{CP}^1$ with generic fiber a surface S_g of genus g.

Let (X, Σ) be as in Theorem 3.1, and suppose that the self-intersection $\Sigma^2 = 0$. We call (X, Σ) an *SW-pair* if satisfies the property that $\mathcal{SW}_{X \#_{\Sigma = S_g} Y_g} \neq 0$. (In general, if $\Sigma^2 = n > 0$, one makes this definition by first blowing up n times.) For example, if X is symplectic and Σ is a symplectic submanifold (of square 0), then $X \#_{\Sigma = S_g} Y_g$ is symplectic, and it follows that (X, Σ) is an SW-pair. In $X \#_{\Sigma = S_g} Y_g$, the rim torus R becomes homologically essential and is c-embedded. We can use Theorem 2.1 to calculate Seiberg-Witten invariants:

$$\mathcal{SW}_{X \#_{\Sigma_K = S_g} Y_g} = \mathcal{SW}_{(X \#_{\Sigma = S_g} Y_g)_K} = \mathcal{SW}_{X \#_{\Sigma = S_g} Y_g} \cdot \Delta_K(r)$$

where $r = \exp(2[R])$, viewing $[R]$ as a class in the fiber sum. We have:

THEOREM 3.2 ([7]). *Consider any SW-pair (X, Σ) with $\Sigma^2 \geq 0$. If K_1 and K_2 are two knots in S^3 and if there is a diffeomorphism of pairs $(X, \Sigma_{K_1}) \cong (X, \Sigma_{K_2})$, then $\Delta_{K_1}(t) = \Delta_{K_2}(t)$.*

As a special case:

THEOREM 3.3 ([7]). *Let X be a simply connected symplectic 4-manifold and Σ a symplectically embedded surface of positive genus and nonnegative self-intersection. Assume also that $\pi_1(X \setminus \Sigma) = 1$. If K_1 and K_2 are knots in S^3 and if $(X, \Sigma_{K_1}) \cong (X, \Sigma_{K_2})$, then $\Delta_{K_1}(t) = \Delta_{K_2}(t)$. Furthermore, if $\Delta_K(t) \neq 1$, then Σ_K is not smoothly ambient isotopic to a symplectic submanifold of X.*

The second part of the theorem holds because if Σ_K were symplectic, $X \#_{\Sigma_K = S_g} Y_g$ would be a symplectic manifold. The symplectic form ω on this manifold is inherited from the forms on X and Y_g; so $\langle \omega, R \rangle = 0$. But $\mathcal{SW}_{X \#_{\Sigma_K = S_g} Y_g} = \mathcal{SW}_{X \#_{\Sigma = S_g} Y_g} \cdot \Delta_K(r)$, and it follows that the among the basic

classes k of $X\#_{\Sigma_K=S_g}Y_g$, more than one has $\langle\omega,k\rangle$ maximal. This contradicts the fact that, for a symplectic manifold, the maximality of $\langle\omega,K\rangle$ characterizes the canonical class among all basic classes [24].

4. Fiber sums of holomorphic Lefschetz fibrations

In this section we shall construct for every integer $g\geq 3$ a pair (X_g,X_g') of simply connected complex surfaces carrying holomorphic genus g Lefschetz fibrations with the property that their fiber sum (along a regular fiber) is a symplectic 4-manifold Z_g which supports no complex structure; in fact Z_g is not even homeomorphic to a complex manifold.

Let $T(p,q)$ denote the (p,q) torus knot in S^3 and let $N(p,q)$ denote the 4-manifold obtained by attaching a 2-handle to the 4-ball along $T(p,q)$ with 0-framing. It is well known that $N(p,q)$ is a Lefschetz fibration over D^2 with generic fiber a Riemann surface of genus $g(p,q)=(p-1)(q-1)/2$. Let $W(p,q)$ denote the canonical resolution of the Brieskorn singularity $\Sigma(p,q,pq)$, the Seifert-fibered 3-manifold with three exceptional fibers of order p, q, and pq, and with $H_1=\mathbf{Z}$. It is known that $W(p,q)$ also supports the structure of a genus $g(p,q)$ Lefschetz fibration over D^2 with a singular fiber over 0 which is a sequence of 2-spheres plumbed according to the resolution diagram of $\Sigma(p,q,pq)$. Finally, let

$$Z(p,q)=W(p,q)\cup N(p,q).$$

The manifold $Z(p,q)$ is a rational surface which is diffeomorphic to the connected sum of $\mathbf{CP}^2$ and $r(p,q)$ copies of $\overline{\mathbf{CP}}^2$ for some computable integer $r(p,q)$. Furthermore, $Z(p,q)$ supports the structure of a holomorphic Lefschetz fibration whose fiber has genus $g(p,q)$.

Now consider nontrivial torus knots $T(p,q)$ and $T(p',q')$ with the property that $g(p,q)=g(p',q')$. (This is possible for every $g(p,q)\geq 3$.) Let $F(p,q;p',q')$ denote the fiber sum along a regular fiber of $Z(p,q)$ with $Z(p',q')$. Then $F(p,q;p',q')$ is a simply connected symplectic 4-manifold with

$$c_1^2=10+8g(p,q)-r(p,q)-r(p',q'),\quad \chi=(b^+ +1)/2=1+g(p,q).$$

Furthermore, $F(p,q;p',q')$ supports the structure of a Lefschetz fibration with fiber of genus $g(p,q)$. A computation of the Seiberg-Witten invariants of $F(p,q;p',q')$ shows that, up to sign, there is a unique Seiberg-Witten basic class. It follows that $F(p,q;p',q')$ is minimal.

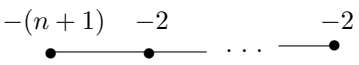
CONJECTURE . *$F(p,q;p',q')$ supports the structure of a complex 4-manifold if and only if $\{p,q\}=\{p',q'\}$.*

As evidence, consider the pairs $(2,2n+1)$ and $(3,n+1)$, $n\not\equiv 2 \mod 3$. For $F(2,2n+1;3,n)$ one can show that $r(2,2n+1)=4n+4$ and $r(3,n+1)=3n+7$ so that

$$c_1^2=n-2,\quad \chi=n+1.$$

Thus, $c_1^2=\chi-3$, which violates the Noether inequality $c_1^2\geq 2\chi-6$. This means that $F(2,2n+1;3,n)$ is a minimal symplectic 4-manifold that is not even homotopy equivalent to a complex manifold. In fact, it can be shown that the fiber sum of $Z(2,2n+1)$ with itself is the elliptic surface $E(n+1)$ and that

the fiber sum of $Z(3, n+1)$ with itself is a Horikawa surface with $\chi = n+1$. Furthermore $F(2, 2n+1; 3, n)$ can be obtained from $E(n+1)$ by removing from $Z(2, 2n+1) \setminus F \subset E(n+1)$, F a regular fiber, the regular neighborhood of the configuration of $(n-2)$ 2-spheres:



whose boundary is the lens space $L((n-1)^2, -n)$ and replacing it with the rational ball that this lens space bounds. (See [3] for all the details concerning this rational blowdown procedure.) Thus $F(2, 2n+1; 3, n)$ is the manifold $Y(n)$ constructed in Lemma 7.5 of [3].

5. Homeomorphic but non-diffeomorphic 4-manifolds with the same Seiberg-Witten invariants

In this section we construct examples of a pair (X_1, X_2) of symplectic 4-manifolds with X_1 homeomorphic to X_2, $\mathcal{SW}_{X_1} = \mathcal{SW}_{X_2}$, but X_1 is not diffeomorphic to X_2. To do this choose a pair of fibered 2-bridge knots $K(\alpha, \beta_1)$ and $K(\alpha, \beta_2)$ with the same Alexander polynomials; for example $K_1 = K(105, 64)$ and $K_2 = K(105, 76)$ with Alexander polynomial

$$\Delta_K(t) = t^{-4} - 5t^{-3} + 13t^{-2} - 21t^{-1} + 25 - 21t + 13t^2 - 5t^3 + t^4.$$

Although these knots have the same Alexander polynomial, they can be distinguished by the fact that their branch covers are the lens spaces $L(\alpha, \beta_1)$ and $L(\alpha, \beta_2)$ which are distinct; in our specific case $L(105, 64)$ is not diffeomorphic to $L(105, 76)$. These knots are also distinguished by their dihedral linking numbers; let S_{K_1} and S_{K_2} denote the 2-fold covers of S^3 branched over K_1 and K_2, with lifted branched loci $\tilde{K}_1$ and $\tilde{K}_2$, respectively. Thus we have knots $\tilde{K}_i$ in $S_{K_i} = L(\alpha, \beta_i)$. Take the α-fold covers of these lens spaces to obtain links $L_i = \{K_1^{(i)}, \ldots, K_\alpha^{(i)}\}$ which are the lifts of the branch loci $\tilde{K}_i$. The linking numbers of the links L_1 and L_2 are known as the 'dihedral linking numbers' of the 2-bridge knot $K(\alpha, \beta)$.

Now perform the knot surgery construction of §2 on the $K3$ surface, replacing $T^2 \times D^2$ with $S^1 \times (S_{K_j} \setminus \tilde{K}_j)$. The resulting 4-manifolds are the manifolds X_i. Either by adapting the arguments of [6] or by using [12] or [15], it can be shown that $\mathcal{SW}_X = \mathcal{SW}_Y = \Delta_K(t) \cdot \Delta_K(-t)$. Unfortunately, the X_i are not simply connected (but are homeomorphic). In particular, $\pi_1(X_1) = \pi_1(X_2) = \mathbf{Z}_\alpha$, and the α-fold covers $\tilde{X}_1$ and $\tilde{X}_2$ of X_1 and X_2 are not diffeomorphic. To see this, observe that $\tilde{X}_i$ is obtained as our link construction in [6] (cf. § 2) by gluing one copy of $E(2)$ minus a neighborhood of a smooth elliptic fiber to every boundary component of $S^1 \times (S^3 \setminus L_i)$. It follows from [6] that

$$\mathrm{SW}_{\tilde{X}_i} = \Delta_{L_i}(t_1, \ldots, t_\alpha) \cdot \prod_{j=1}^{\alpha} (t_j^{1/2} - t_j^{-1/2})$$

Since the linking numbers of the links L_1 and L_2 are different, it can be shown the Hosokawa polynomials of the links L_1 and L_2, when evaluated at 1 are distinct

[9]. Thus their Alexander polynomials are different and $\tilde{X}_1$ is not diffeomorphic to $\tilde{X}_2$.

There is a lesson to be learned from these examples. One must consider the Seiberg-Witten invariants of a 4-manifold X together with those of all of its covers as the appropriate invariant for X.

6. Nonsymplectic 4-manifolds with one basic class

Recall from § 2, that if k is a basic class of X, so is $-k$. Because of this, we say that X *has n basic classes* if the set $\{k \,|\, \mathrm{SW}_X(k) \neq 0\}/\{\pm 1\}$ consists of n elements. There are abundant examples of 4-manifolds with one basic class. Minimal nonsingular algebraic surfaces of general type have one basic class (the canonical class) [28]. The authors and others have constructed many examples of minimal symplectic manifolds with one basic class and $\chi - 3 \leq {c_1}^2 < 2\chi - 6$. (These manifolds cannot admit complex structures due to the geography of complex surfaces.) However, the examples described here are the first nonsymplectic manifolds with one basic class.

Let $X = E(2)$ and T a smooth elliptic fiber. For a knot K of genus g form the knot surgery construction to obtain X_K. In X_K there is a surface Σ of genus $g+1$ with $[\Sigma]^2 = 0$ and $[\Sigma] \cdot [T] = 1$. Let M be the 3-manifold obtained from 0-surgery on the trefoil knot. Then $S^1 \times M$ is a T^2-fiber bundle over T^2. The fiber sum of $g+1$ copies of the fiber bundle gives a 4-manifold Y which is an $F = T^2$-bundle over a surface of genus $g+1$, and it is easily seen that there is a section C. Furthermore, Y is a symplectic 4-manifold with $c_1(Y) = -2g[F]$. Our example, corresponding to the genus g knot K is $Z_K = X_K \#_{\Sigma = C} Y$. We perform this fiber sum so that Z_K is a spin 4-manifold [11]. It can be seen to be simply connected.

Write the symmetrized Alexander polynomial of K as $\Delta_K(t) = a_0 + \sum_{n=1}^{d} a_n(t^n + t^{-n})$, and call d the *degree* of $\Delta_K(t)$. Since the genus of K is g, we have $d \leq g$. If K is an alternating knot, for example, then $d = g$. Say that the Alexander polynomial of K has *maximal degree* if $d = g$. Using techniques of [20] we calculate:

Theorem 6.1 ([8]). *Let K be a knot in S^3 whose Alexander polynomial has maximal degree. Then Z_K has one basic class, k, with $|SW_{Z_K}(k)| = a_d$, the top coefficient of $\Delta_K(t)$. When $|a_d| > 1$, Z_K is nonsymplectic.*

References

[1] S. Donaldson, *Polynomial invariants for smooth 4-manifolds*, Topology **29** (1990), 257–315.

[2] R. Fintushel and R. Stern, *The blowup formula for Donaldson invariants*, Ann. of Math. **143** (1996), 529–546.

[3] R. Fintushel and R. Stern, *Rational blowdowns of smooth 4-manifolds*, Jour. Diff. Geom., **46** (1997), 181–235.

[4] R. Fintushel and R. Stern, *Donaldson invariants of 4-manifolds with simple type*, J. Diff. Geom. **42** (1995), 577–633.

[5] R. Fintushel and R. Stern, *Immersed spheres in 4-manifolds and the immersed Thom conjecture*, Turkish J. Math. **19** (1995), 145–157.

[6] R. Fintushel and R. Stern, *Knots, links, and 4-manifolds*, to appear in Invent. Math.

[7] R. Fintushel and R. Stern, *Surfaces in 4-manifolds*, Math. Res. Letters **4** (1997), 907–914.
[8] R. Fintushel and R. Stern, *Nonsymplectic 4-manifolds with one basic class*, preprint.
[9] R. Fintushel and R. Stern, *Nondiffeomorphic symplectic 4-manifolds with the same Seiberg-Witten invariants*, preprint.
[10] M. Freedman, *The topology of four-dimensional manifolds*, Jour. Diff. Geom. **17** (1982), 357–454.
[11] R. Gompf, *A new construction of symplectic manifolds*, Ann. Math. **142** (1995), 527–595.
[12] E. Ionel and T. Parker, *Gromov invariants and symplectic maps*, preprint.
[13] P. Kronheimer and T. Mrowka, *Embedded surfaces and the structure of Donaldson's polynomial invariants*, J. Diff. Geom. **41** (1995), 573–734.
[14] P. Kronheimer and T. Mrowka, *The genus of embedded surfaces in the projective plane*, Math. Research Letters **1** (1994), 797–808.
[15] W. Lorek, *Lefschetz zeta function and Gromov invariants*, preprint.
[16] J. McCarthy and J. Wolfson, *Symplectic normal connect sum*, Topology **33** (1994), 729–764.
[17] G. Meng and C. Taubes, $\underline{\mathcal{SW}} = $ *Milnor Torsion*, Math. Research Letters **3** (1996), 661–674.
[18] J. Morgan, T. Mrowka and Z. Szabo, *Product formulas along T^3 for Seiberg-Witten invariants*, Math. Res. Letters **4** (1997), 915–930.
[19] J. Morgan, T. Mrowka, Z. Szabo, and C. Taubes, in preparation.
[20] J. Morgan, Z. Szabo and C. Taubes, *A product formula for the Seiberg-Witten invariants and the generalized Thom conjecture*, J. Diff. Geom. **44** (1996), 706-788.
[21] P. Oszvath and Z. Szabo, *The symplectic Thom conjecture*, preprint.
[22] Z. Szabo, *Simply-connected irreducible 4-manifolds with no symplectic structures*, Invent. Math. **132** (1998) 3, 457-466.
[23] C. Taubes, *The Seiberg-Witten invariants and symplectic forms*, Math. Res. Letters **1** (1994), 809–822.
[24] C. Taubes, *More constraints on symplectic manifolds from Seiberg-Witten invariants*, Math. Res. Letters **2** (1995), 9–14.
[25] C. Taubes, *$SW \Rightarrow Gr$, From the Seiberg-Witten equations to pseudo-holomorphic curves*, Jour. Amer. Math. Soc. **9** (1996), 845–918.
[26] C. Taubes, *Counting pseudo-holomorphic submanifolds in dimension 4*, preprint, 1995.
[27] W. Thurston, *Some simple examples of symplectic manifolds*, Proc. Amer. Math. Soc. **55** (1976), 467–468.
[28] E. Witten, *Monopoles and four-manifolds*, Math. Res. Letters **1** (1994), 769–796.

Ronald Fintushel
Department of Mathematics
Michigan State University
East Lansing, Michigan 48824
U.S.A.
ronfint@math.msu.edu

Ronald J. Stern
Department of Mathematics
University of California
Irvine, California 92697
U.S.A.
rstern@uci.edu

Topological Views on Computational Complexity

Michael H. Freedman*

1991 Mathematics Subject Classification: 57-XX Manifolds and cell complexes; 68-XX Computer Science; 81-XX Quantum Theory
Keywords and Phrases: computational complexity, topology, quantum field theory

For the pure mathematician the boundary that Gödel delineated between decidable and undecidable, recursive and nonrecursive, has an attractive sharpness that declares itself as a phenomenon of absolutes. In contrast, the complexity classes of computer science, for example P and NP, require an asymptotic formulation, and like the subject of "coarse geometry", demand a bit of patience before their fundamental character is appreciated.

The heart of the matter is to understand which problems can be solved by an algorithm whose "running time" grows only polynomially with the size of the instance. It is interesting to note that in other areas of mathematics things polynomial tend to have excellent limiting behavior: 1. Any polynomial on cardinals: $x \longmapsto \text{poly}(x)$ is continuous at the first infinite cardinal, whereas the power set function $x \longmapsto 2^x$ is not. 2. In complex analysis, polynomials extend conformally over infinity to yield a branch point, whereas exp is essentially discontinuous at infinity. 3. In coarse geometry, groups with polynomial growth, in common with nilpotent Lie groups, have Carnot manifolds as scaling limits (Gromov [G]) in the Gromov-Hausdorf topology. These examples, particularly the last, suggest that polynomial time algorithms might eventually be understood by constructing a more manageable limiting object as polynomial growth groups are understood via nilpotent Lie groups.

In order to make the discussion of algorithms precise, it is necessary to define a computational model. This is more exciting now than it was ten years ago. The "polynomial Church thesis" is up in the air, and there are two robust computational models to sink one's teeth into: the "Turing model" and "Quantum Computing" (QC). (See http://xxx.lanl.gov/abs/quant-ph and [K] for a suggested solid state implementation based on the hyperfine coupling between electron spin and nuclear spin.) Furthermore it is possible that there will be other, perhaps stronger, computational models based on topological quantum field theory [F1].

The thesis of Alonzo Church, propounded in the mid-1940s, asserts that any two definitions of "computable function" will agree. The "polynomial version" of the Church thesis (although I do not know that it was ever endorsed by Church) says that any two physically *reasonable* models of computation will agree on the

*This work was supported by Microsoft Research.

class of polynomial time functions (but not necessarily on the degree of the polynomial, which may in fact be model dependent). "Reasonableness" implies limited accuracy in preparation and measurement of physical states.

It might seem that if one accepts that the universe is fundamentally quantum mechanical (and I am perfectly prepared to neglect the irreversibility of black hole evaporation) that QC is the ultimate model, and no others need be sought. This argument is not entirely convincing, since a solid state system (perhaps one involving global excitation as occurs in the fractional quantum Hall effect) might be governed to considerable accuracy by an effective field theory whose simulation through local QC gates involves exponential inefficiencies. (Note that a preliminary discussion of simulating local Hamiltonians by gates is given in [L].) Topological field theories, because of their discrete character and their connections to NP-hard (actually $\#P$-hard) combinatorial problems, e.g., the evaluation of the Jones polynomials, are the most interesting candidates for further computational models [F1]. The next section contains definitions, but briefly, the class NP, nondeterministic polynomial time, consists of those decision problems where the time to "check" (rather than find) a proposed solution grows only polynomially in the length of the problem instance.

In pure mathematics, problems of fundamental importance occasionally arrive on our doorstep from physics. The only other cases (i.e., origin outside of physics) I can think of are: probability (gambling), crystallographic groups (chemistry), incompleteness (philosophy), and the P/NP problem (computer science). A proof that $P \neq NP$ would be extraordinarily strong, as it would foreclose the possibility of myriad yet-unimagined theories that might connect, say, the colorings of a graph (which is NP complete) and, say, the cohomology of some associated space (which might well be in P as cohomology is essentially linear algebra). These speculations might suggest that the P/NP problem is undecidable. In a platonic world view, where statements of first order arithmetic, such as "$P = NP$", are either true or false, there are two subcases: the very interesting Case (1): undecidable and true, in which case the NP problems *do* admit P-time algorithms, but there is no *documentation* proving they work; and the less interesting Case(2): undecidable and false: there are no P-time algorithms for the NP-complete problems, but there is no proof of this statement.

The assertion that a problem is important to mathematics is usually supported by sketching its relations to other problems and fields. The P/NP problem enjoys a more interesting status. The practice of mathematics is largely the search for proofs of reasonable length (certainly polynomial in statement length) and so is inside NP. Setting aside the constraints of any particular computational model, the creation of a physical device capable of brutally solving NP problems would have the broadest consequences. Among its minor applications it would supersede intelligent, even artificially intelligent, proof finding with an omniscience not possessing or needing understanding. Whether such a device is possible or even in principle consistent with physical law, is a great problem for the next century.

§1. Preliminaries.

The Turing model of computation consists in a bare formulation of a bi-infinite tape, a head which can read/write symbols from a finite alphabet and which is capable itself of being one of finitely-many *internal* states. Its "program" is a finite set of 5-tuples $\{S, q, S', q', M\}$ which say that if it is in state S and reads q, it will assume state S', overwrite q with q', and move right or left according to the indicated motion M. We can absorb knowledge of the last motion into the state q' and so drop the fifth symbol. Without an applicable instruction the machine halts. The internal state, the head position and the contents of the tape together, form the machine's *complete state*. For convenience, one or more additional tapes may be added to the machine, generally decreasing computation time, but by no more than a square root factor. All conventional computers are implementations of the Turing model.

A next step is to allow probabilistic computation where several 4-tuples may begin "S, q", and these will be assigned positive weights p_i summing to one and will be executed with probability p_i, so that the machine now evolves stochastically through a *mixture* of states. Empirically, it is often easier to find probabilistic algorithms that almost always work, than to find traditional exact algorithms.

A further, more radical, innovation is to allow the weights above, now written $w_{\alpha\beta}$, to be complex numbers satisfying $\sum w_{\alpha\beta}\overline{w}_{\beta\gamma} = \delta_{\alpha\gamma}$, where $w_{\alpha\beta}$ is the transition amplitude for $(S, q) = (S, q)_\alpha \to (S, q)_\beta = (S', q')$. The resulting evolution of the computation is now a unitary evolution $U(t)$ in a vector space of complete states. This, briefly, is the model called *quantum computation* or QC. It is an important consequence of this description that the evolution is *local* at any time t: The t^{th} step, or *gate*, in the time evolution $U(t)$ is the identity except on a tensor factor of bounded dimension (typically $\mathbb{C}^4$ or $\mathbb{C}^8$ in detailed specifications).

In the Turing model P represents the class of decision problems $\{D\}$ (answer $\in$ {yes, no}) so that there is a program F_D and a polynomial P_D with F_D yielding the answer to each instance I of D in time $\leq P_D(\text{length}\, I)$, where length I is the number of bits required to express I. One says D lies in NP (nondeterministic polynomial time) if there is an *existential* program operating on I plus a number of *guess bits* which correctly answer all instances in polynomial time. The existential program is deemed to answer "yes", if for some setting of the guess bits the machine halts on the symbol 1. The fundamental question of computer science is to show that $P \neq NP$, essentially that it is harder to find a solution than to check a guess.

The problem of the existence of a satisfying assignment for a Boolean formula is the canonical *NP-complete* problem, meaning[1] it lies in NP and if a Turing machine were augmented by an *oracle* capable of (quickly) answering that one problem, then all problems in NP could be solved in polynomial time. (A problem is called "hard" rather than complete if only the second assertion is being made.) The class $\#P$ is the counting analog of NP; computing the number of satisfying assignments of a Boolean formula is the canonical $\#P$-complete problem. In oracle notation $P^{NP} \subset P^{\#P}$, meaning a poly-time machine with access to an $\#P$ oracle is at least as powerful as one with access to an NP oracle.

[1] according to Cook [C]

The model QC is not strictly comparable with Turing, since in QC the output, a *measurement* of a final stationary state, is only probabilistic. However it is believed, because of Shor's QC algorithm [Sh] for factoring integers in polytime, that QC is substantially more powerful than P but perhaps *not* powerful enough to solve NP-complete problems in polynomial time. Computational models that are allowed to handle continuous quantities are almost always absurdly strong (e.g., contain NP), if accuracy is not restricted to poly(log) number of bits. ([Sc], [ADH])

On the topological side, the notion of a topological quantum field theory has emerged through Witten's work. A TQFT is usually understood to be a functor Z from (oriented marked surface, a bounding oriented 3-manifold with link; diffeomorphisms)[2] to (finite-dimensional Hilbert spaces over $\mathbb{C}$, vector; linear maps) which satisfies $Z(\Sigma_1 \cup \Sigma_2) = Z(\Sigma_1) \otimes Z(\Sigma_2)$, $Z(\overline{\Sigma}) = Z(\Sigma)^*$, a gluing axiom (gluing bordism corresponds to composing linear maps), and a unitarity axiom. (See [At] for details.) In particular such a theory assigns scalars to closed three-manifolds containing a link L, and Witten identified one such theory W_k, $SU(2)$-Chern-Simons theory at level k, as a value of the Jones [Jo] polynomial V,

$$W_k(L) = V_L(\zeta) \quad , \quad \zeta = e^{\frac{2\pi i}{k+2}} \quad . \tag{1}$$

Since we will be discussing the utility of this TQFT for solving combinatorial problems such as Boolean satisfiability, it is relevant to observe that counting satisfactions, colorings, and many other combinatorial problems provide by far the simplest examples of systems obeying the TQFT axioms; only the source category must be redefined. (It is tempting to look for the corresponding path-integral interpretations.) To see, for example, how the gluing axiom works for the problem of counting vertex colorings of a graph, let $(G_1; H_1, H_2)$ and $(G_2; H_2, H_3)$ be disjoint finite graphs, each with two preferred disjoint subgraphs where $H_2 \subset G_1$ and $H_2 \subset G_2$ are identified by a fixed isomorphism. Let $G = G_1 \cup_{H_2} G_2$. Let i, j and k index the possible legal colorings of H_1, H_2 and H_3 respectively and let $m_{i,j}$ ($n_{j,k}$) be the number of colorings of G_1 restricting i on H_1 and j on H_2 (j on H_2 and k on H_3). Then the number of colorings $g_{i,k}$ of G which restrict to i on H_1 and to k on H_3 satisfies the composition rule:

$$g_{ik} = \sum_j m_{i,j} n_{j,k}$$

§2. What a topologist might think about formal systems.

A. Understanding the class P:

Conjecture: The class of P-time algorithms can be elucidated by constructing some scaling limit as discussed in the introduction. (Also see [F2])

B. General position in formal systems:

There is an empirical connection between computational complexity of a finite decision problem and the undecidability of an infinitary version [F3]. Although

[2] perhaps with additional structures or labelings

oracle separation results [BGS] show that detailed properties of P must enter the proof, P might be distinguished from NP by finding a translation which carries P into decidable statements. Thus it is natural to ask how common decidable statements are. Let X be a formal system subject to Gödel's second incompleteness theorem, such as Peano arithmetic or ZFC. Let $\{S_i \mid i \in Z^+\}$ be the sentences of X enumerated in some syntactically natural way, e.g., alphabetical order. Consider those sentences which are provable (in X or some fixed finite extension X^+ of X) and let p_i denote the number of these with index $\leq i$ which are provable.

Conjecture: "Ubiquity of undecidability" $\operatorname{limsup}(p_i / \sqrt{i}) = 0$. That is, the "number" of provable statements is less than the square root of the number of statements.

According to Kolmogorov and later Chaitin [Ch] at least half of integers fail to admit short descriptions, but particular true instances of the statement "n has no short description" are always undecidable. This provides a fairly large natural family of undecidable statements, but not the conjectured ubiquity of undecidable statements.

Rationale for conjecture: The single most useful principle in geometric topology is that submanifolds $P^p, Q^q \subset M^n$ contained in a manifold, generically satisfy $\dim(P \cap Q) = p + q - n$. For finite sets, if P and Q are drawn randomly from M, the same formula holds:

$$\text{expected value of } \log \operatorname{card}(P \cap Q) = \log \operatorname{card}(P) + \log \operatorname{card}(Q) - \log \operatorname{card}(M). \tag{2}$$

In particular two disjoint subsets P and P' of equal cardinality should satisfy:

$$\operatorname{card}(P) = \operatorname{card}(P') < \sqrt{\operatorname{card} M} \tag{3}$$

if their disjointness is simply a matter of chance.

If X is consistent, then provable statements Q and their negations Q' are disjoint. We believe that in a system complex enough to be incomplete, the global structure of Q inside all statements is essentially random and so expect Q to be asymptotically of less than square root size. This is analogous to thinking that the primes are "randomly" distributed in the integers according to the density $\frac{1}{\log n}$, a model with considerable predictive power.[3]

C. $P \neq NP$ has predictive power in link theory:

In computer science the notions of width arise in identifying subclasses of NP-hard problems which are actually solvable in P-time. Among these are problems of constant width, or even polylog width problems. (Compare page 95 [We].) For the present purpose define the width of a link L to be the $\inf_{\pi} \sup_{r} \left| L \cap \pi^{-1} r \right|$, where π is a smooth product projection $\mathbb{R}^3 \to \mathbb{R}$, $r \in \mathbb{R}$. Let $\mathcal{L}$ be the set of finite links. Call a mapping $i : \mathcal{L} \to \mathcal{L}$ *information preserving*, if some $\#P$-hard data

[3]For example, the density of double primes seems to be correctly predicted, up to a small multiplicative constant, from this assumption.

about $L \in \mathcal{L}$ can be quickly computed from data about $i(L)$, e.g., if $V_L(e^{2\pi i/5})$ is quickly (P-time) calculable from $V_{i(L)}(e^{2\pi i/5})$.

CONJECTURE: The image of an information-preserving map i, $\{i(L)\}$, cannot have constant width (or even width $\leq$ poly (log crossing $\#(L)$).

This conjecture is implied by the conjecture $P \neq NP$ if we also make the modest assumptions that i can be computed on a link L in time $\leq$ poly(#crossing) and that the crossing number obeys: #crossing$(i(L)) <$ poly(#crossing(L)).

The Jones polynomial at $e^{2\pi i/5}$, according to Witten [W], Reshetikhin and Turaev [RT], is the scalar output of a TQFT. Bounded width implies a fixed bound on the dimension of the Hilbert spaces which arises as the link is sliced into elementary bordisms.

Thus the calculation time for the composition of the elementary bordisms in TQFT is linear in the number of compositions. Since the dimension of Witten's Hilbert space grows (only) exponentially with width, poly log width is an adequate assumption for the entire calculation to grow at a polynomial rate.

D. "FINITE TYPE INVARIANTS" IN COMBINATORICS:

Vassiliev's book [V] contained implicitly a notion of "finite type" link invariant, clarified by Birman-Lin [BL] and Bar Natan [BN1], who showed that the perturbative invariants associated to the Witten-Chern-Simons theory are finite type. The fundamental idea of a finite-type invariant can be reproducted in any combinatorial setting where a notion of an (oriented) *elementary difference* can be defined. In oriented link theory the formal difference between two link diagrams, where a positive crossing in the first has been replaced by a negative crossing in the second, is the notion of elementary difference.

We give two examples in graph theory. In both cases the fundamental theorem [BN2] that the finite-type invariants of link theory can be computed in polynomial time continues to hold. One finds for invariants of type $= n$, a bound on computation time $\leq O(\#^n)$, where # is the number of edges in the graph. Analogous to the Witten-Chern-Simon theory where the l.h.s. is a $\#P$-hard nonperturbative invariant and the r.h.s. is an asymptotic expansion with finite-type coefficients, we find that (in the two cases respectively) after suitable change of variables, the chromatic and flow polynomials of a graph, which in their totality are $\#P$-hard to calculate, can be expressed as a polynomial whose k^{th} coefficient is of type $= k$.

EXAMPLE 1: Define an elementary difference on finite graphs modulo isomorphism to be an ordered pair consisting of a finite graph followed by the graph with one edge deleted, $(G, G \smallsetminus e)$. A (real valued) invariant on finite graphs $f : \{\text{graphs}\} \to \mathbb{R}$ is *type* n, if given any collection of $n+1$ edges $\{e_1, \ldots e_{n+1}\}$ of G, all $(n+1)^{\text{st}}$ order differences given by a sum over subsets vanishes:

$$\sum_{S \subset 2^{\{e_1,\ldots,e_{n+1}\}}} (-1)^{|S|} f(G \smallsetminus S) = 0 \tag{4}$$

The chromatic polynomial of a finite graph, $P_G(\lambda)$, has degree $= V$, the number of vertices, and satisfies the "contraction-deletion" recursion relation:

$$P_G(\lambda) - P_{G \smallsetminus e}(\lambda) = -P_{G/e}(\lambda), \tag{5}$$

where $G \smallsetminus e$ is G with e deleted, and G/e is G with e contracted. Let $\overline{P}_G$ be the polynomial with "reversed" coefficients, $\overline{P}_G = \lambda^V P_G(\lambda^{-1})$. The recursion relation on $\overline{P}$ becomes:

$$\overline{P}_G(\lambda) - \overline{P}_{G \smallsetminus e}(\lambda) = -\lambda \overline{P}_{G/e}(\lambda). \tag{6}$$

Line (6) implies the constant term of $\overline{P}(\lambda)$ applied to an elementary difference (l.h.s. (6)) is zero. Inductively it is easy to see that the coefficient of a degree $= n$ term of $\overline{P}$ will vanish any formal differences of order $= n + 1$. Comparing with line (4) we have:

OBSERVATION 1: The coefficient of degree n in $\overline{P}$ is a type $= n$ invariant w.r.t. deletion of finite graphs.

EXAMPLE 2: Dual to the chromatic polynomial is the *flow polynomial* $F_G(\theta)$. It has degree $= E$, the number of edges of G, and satisfies the recursion relation

$$F_G(\theta) - F_{G/e}(\theta) = -F_{G \smallsetminus e}(\theta) \tag{7}$$

Let $\overline{F}_G(\theta) = \theta^E F_G(\theta^{-1})$ so that

$$\overline{F}_G(\theta) - \overline{F}_{G/e}(\theta) = -\theta F_{G \smallsetminus e}(\theta) \tag{8}$$

Now if we define the ordered pair $(G, G/e)$ to be the *elementary difference*, then we obtain a dual notion finite-type graph invariant and have the:

OBSERVATION 2: The coefficient of degree n in $\overline{F}$ is a type $= n$ invariant (w.r.t. contraction) of finite graphs.

A general principle seems to be that if the associated graded objects to the finite type invariants (dual cord diagrams in Vassiliev's theory) span a finite-dimensional space, then calculating finite-type invariants should be polynomial time in the complexity of the instance (eg., link, graph, etc., ...). (Compare with [BN2].)

In the case of graphs, for either of the two preceding notions of elementary difference, the graded object at level n is only 1-dimensional, being spanned by the general "n-singular" graph. In the two cases, the general n-singular graph is a formal signed difference $\sum_{S \subset G_0} (-1)^{|S|}(G \smallsetminus G_0)$ or $\sum_{S \subset G_0} (-1)^{|S|} G/\text{components } G_0$, respectively, where $G_0 \subset G$ is a subgraph of n edges. Thus the only finite-type invariants are polynomials in the coefficients of $\overline{P}$ and $\overline{F}$ respectively.

OBSERVATION 3: For type $= n$ invariants, w.r.t. deletion (or contraction), the time to compute is bounded by $O(E^n)$.

PROOF. Consider deletion; the contraction case is similar. If f is type $= n$, f is zero on graphs with $k + 1-$ singular graph edges and therefore constant on k-singular graphs with isomorphic singular sets. Given, as in the Vassiliev theory, a system of "integration" constants, it takes no more than E steps to evaluate the function f on graphs once f is known on 1-singular graphs. Each of these steps

requires at most E preliminary steps to integrate a function on 2-singular graphs to obtain the evaluation of f on 1-singular graphs. Proceeding in this way, the result follows by induction.

For the chromatic polynomial, there is a subgraph sum formula for the coefficients, which gives the same growth in complexity we just obtained. It is also known that the linear coefficient of P is $\#P$ hard to compute. I presume the same is true for the flow polynomial. It is intriguing that there is a general approach to filtering $\#P$-hard information by polytime "approximations" of increasing degree. The art to finding useful approximations, less trivial than the two examples presented here, seems to be in choosing the "elementary differences". The situation is parallel to the Witten-Chern-Simon theory where there is a $\#P$-hard nonperturbative l.h.s. and an asymptotic expansion on the r.h.s. where the individual coefficients are finite type, and therefore polynomial time invariants.

From group theory we give a final example of an *unoriented* difference motivated by the formal structure of Wertinger presentations.

EXAMPLE 3: An elementary difference (G, G') is defined to be an unordered pair of groups where G and G' admit presentations which agree except for a single relation in which the literals (generators and generator inverses) read *backwards* in G' as compared to G. The consequence of the difference being unordered is that all finite type invariants defined from it are ambiguous up to sign. I have not yet made any investigation of this algebraic version of the "crossing change" in link theory.

§3. THE PHYSICS OF COMPUTATIONAL MODELS.

We should generally be interested in physical systems–even rather hypothetical ones—whose preparation may specify an instance of a problem and whose measurement can be (quickly) deconvolved to give the answer to that instance. A standard pitfall is to expect to make measurements to too great an accuracy, or at too low a temperature, or in some similar way to disregard the presence of some exponentially growing difficulty. At a fundamental level, any device is "analog". The distinction between analog and digital can be expressed as whether the coarse graining occurs later (analog) or earlier (digital). The success of digital over analog in the first 50 years of computers can be explained by realizing that the usual analog representations of a number, e.g., as a voltage, amounts to storing the number in unary and therefore exponentially less efficient than binary notation. On the other hand, it has been known for some time, that physically measurable quantities of some idealized systems are $\#P$-hard to compute. This makes one wonder if it is not worth the price of working in analog long enough to allow nature to make a truly difficult computation, rather than simply executing a gate, before measuring.

The Ising model for vertex spins on a graph with edge interactions has, in the ferromagnetic case, a Hamiltonian $H = -\sum_{\text{edges } e_{ij}} \sigma_i \sigma_j$, $\sigma_i \in \{-1, 1\}$. The partition function $Z(\beta) = \sum_{\text{spin states } \sigma} e^{-\beta H(\sigma)}$, $\beta = \frac{1}{kT}$, when written in a high-

temperature expansion, becomes:

$$Z = e^{\beta|E|} P\left((e^2)^{-\beta}\right) \quad \text{where} \tag{9}$$

$P(x)$ is the generating function $\sum_{k=0}^{|E|} b_k x^k$ with $b_k = \#$ (bipartite subgraphs of k-edges). (See, e.g., [JS].)

For the purposes of this article let us pretend that $Z(\beta)$ is a measurable quantity. To be more realistic one might consider specific heat $= \frac{\partial^2 \log Z}{\partial \beta^2}$, or some correlation function such as $\left(\sum_{\sigma} \sigma_i \sigma_j e^{-\beta H(\sigma)}\right) \Big/ Z(\beta)$, but to illustrate our point we take the partition function as our measurable quantity, since the tie in to the graph theory is most convenient. The following analysis owes much to conversations with Christian Borgs and Jennifer Chayes.

Using the standard methods, the coefficients b_k take time $\leq O(E^k)$ to compute, so the low coefficients are easy; and it is further known [JS] that the highest non-zero b_k is $\#P$-hard. Our goal in building a "statistical mechanical computer" would therefore be to input a graph G and then tease out the leading coefficient $b_{\max}$ from measurements of $Z(\beta)$ at various temperatures. The problem is essentially to recover the coefficients of a polynomial from measurements of its values at $\{e^{-2\beta_i}\}$ for some collection of positive values of temperature T. This is done by inverting the linear system $(e^{-2j\beta_i})$. Since the coefficients are a priori integers, only some threshold accuracy is needed for an exact determination. Unfortunately numerical instabilities are encountered in the inversion. The essential point is that to determine the leading coefficient of a polynomial P, most information is gained by evaluating P at a large number (so that the low-order contribution is negligible). Unfortunately the physical requirement that temperatures be positive restricts $\beta_i > 0$ and therefore $0 < e^{-2\beta_i} < 1$; this forces P to be evaluated only at small values.

One way out of this numerical problem is to study an anti-ferromagnet on graphs with Hamiltonian $H = \sum_{\text{edges}} \sigma_i \sigma_j$; this allows P to be sampled in the range $1 < x < \infty$ where very low (positive) temperatures will be most revealing of $b_{\max}$. This resolves the numerical instability but ushers in a different problem: An antiferromagnet is a highly frustrated system and only approaches its Gibbs measure with exponential slowness: time to equilibrium $\approx O(e^{\frac{1}{kT}})$ as temperature approaches zero. So the "antiferromagnet computer" would take exponentially long to be initialized to the graph G whose $G_{\max}$ it was computing. Essentially, the antiferromagnet is not qualitatively more efficient at finding its equilibrium than the presently available numerical algorithm, the Metropolis method (see page 124 [We]), and might in fact be rather close to a highly parallel implementation of that algorithm.

The joint failure of the ferro- and anti-ferromagnet to lead (even in principle) to an analog computer for $\#P$ problems, suggests a generic weakness of classical statistical mechanical systems for computation. They sample states sequentially

in time and hence have limited information-processing capacity. While a classical system, such as an Ising magnet, explores its state space sequentially in time, quantum mechanics offers another possibility.

The Feynman path integral computes the evolution operator as a coherent superposition of (infinitely many) states. The resulting evolution incorporates a vast amount of information very quickly; to be useful for computing, this evolution must be guided to answer discrete combinatorial problems. The discrete character of topological quantum field theories and their interpretation in terms of the $\#P$-hard Jones polynomial make these an attractive candidate for a new computational model [F1]. In the physics literature the Abelian Chern-Simons functional occurs in the Lagrangian for certain nonclassical surface layer conductivities governed by the integral quantum Hall effect [Ko]. The Abelian CS functional is known to compute linking number [S].

The $SU(2)$-Chern-Simons functional appears to enter into solid state physics via the fractional quantum Hall effect, a phenomenon of "quasi-particle" conductivity [Wi], [TL]. More abstractly, the TQFT with that functional as Lagrangian is known to compute $\#P$-hard values of the Jones polynomial [W].

Here is a notional sketch of how $SU(2)$-Chern-Simons theory might be implemented as a general computational model. A logical problem X, such as satisfiability of a Boolean formula, would be coded as a link $L = \text{code}(X)$. (See [J] and [JVW] for one way in which this may be done). The link would be described as a braid and implemented in a $(2+1)$-dimensional space time by forcing the motion of charge defects, i.e., "quasi particles", in a very cold surface layer of silicon. This is the preparation or "input" phase. If $SU(2)$-Witten-Chern-Simons is really physically important in this situation, one should expect some detectable "observable" consequence of the particular input braid, containing information on its Jones polynomial, as "output". A key point is whether the observable is a real number, e.g., a measured conductivity, in which case it is the analog version of a number expressed in unary. In contrast, if the observable can itself be some configuration or state of an ancillary collection of quasi-particles, then this is the analog version of a binary number, with addressable information, and much more efficient.

The choice of the translation to links raises a topological issue. For a link L of n crossings, an elementary estimate from the skein relations is that if c is a coefficient of the Jones polynomial V_L, then $|c| < (2\sqrt{2})^n$. Very crude statistical considerations—thinking of the coefficient as the result of a random walk as the contributions of various signs accumulate—suggest typically $|c| < (2\sqrt{2})^{n/2}$. On the other hand, in many cases these are overestimates, for torus links $|c| = 0$ or 1. If the observables, say $V_L(e^{2\pi i/p})$, must be read in "unary", it may be essential, given limited accuracy, to have the ensemble of links, image(code), more like the torus links than the generic link. Is it possible to encode the general Boolean formula into links in such a way that (1) from the evaluation of $V_{\text{code}(X)}, \text{sat}(X)$ may be quickly determined, and (2) so that $|c| \leq \text{poly}(\text{length } X)$ for the coefficients c of $V_{\text{code}(X)}$? Here is a separate question, but with the same motivation: Are there TQFTs which yield information about $V_{(p)L}$, the Jones polynomial $\in Z_p[\mathbb{C}]$ with coefficients reduced modulo a prime p? Positive answers to either question would

ease the problem of identifying V_L from observations of limited accuracy, and allow a Chern-Simons theory even with an essentially unary output, to form the basis of a powerful, if still theoretical, model of computation.

Computer science is driving an interaction between logic, physics, and mathematics, which will explore the ability of the physical world to process information. I have tried to convey the excitement and scope of this endeavor and to point to paths that mathematicians, particularly topologists, might penetrate.

References

[ADH] L. Adelman, J. Demarrias and M. Huang *Quantum computability*, Siam J. Computing 26 (1997), 1524–1540.

[At] M. Atiyah, *Geometry and Physics of Knots*, Lezioni Lincee [Lincei Lectures], Cambridge University Press, Cambridge, 1990.

[BGS] T. Baker, J. Gill, and R. Solovay, *Relativizations of the $\mathcal{P} =? \mathcal{NP}$ question*, SIAM J. Comput. 4 (1975), 431-442.

[BL] J. Birman and X.-S. Lin, *Knot polynomials and Vassiliev's invariants*, Invent. Math. 111 (1993), 225–270.

[BN1] D. Bar-Natan, *Perturbative aspects of the Chern-Simons topological quantum field theory*, Ph.D. Thesis, Princeton Univeristy, 1991; *On the Vassiliev knot invariants*, Topology 34 (1995), 423–472.

[BN2] D. Bar-Natan, *Polynomial invariants are polynomial*, Math. Res. Lett. 2 (1995), 239–246.

[C] S. Cook, *The complexity of theorem-proving procedures*, Proc. 3rd ACM Symp. on Theory of Computing, Association for Computing Machinery, New York, 1971, pp. 151–158.

[Ch] G. Chaitin, *Information-theoretic limitations of formal systems*, J. Assoc. Comput. Mach. 21 (1974), 403–424.

[F1] M. H. Freedman, *P/NP, and the quantum field computer*, Proc. Natl. Acad. Sci. USA 95 (1998), 98–101.

[F2] M. H. Freedman [1998] *Limit, logic, and computation*, Proc. Natl. Acad. Sci U.S.A. 95, 95–97.

[F3] M. H. Freedman, k-sat on groups and undecidability, 1998 ACM Symp. on Theory of Comp. (STOC '98) (to appear).

[G] M. Gromov, *Groups of polynomial growth and expanding maps*, Inst. Hautes tudes Sci. Publ. Math. 53 (1981), 53–73; *Carnot-Carathéodory spaces seen from within*, Progr. Math. 144, Sub-Riemannian geometry (1996), 79–323.

[J] F. Jaeger, *Tutte polynomials and link polynomials*, Proc. Amer. Math. Soc. 103 (1988), 647–654.

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

[JS] M. Jerrum and A. Sinclair, *Polynomial-Time Approximation Algorithms for Ising Model*, Proc. 17th ICALP, EATCS, 1990, 462-475.

[JVW] F. Jaeger, D. L. Vertigan, D. J. A. Welsh, *On the computational complexity of the Jones and Tutte polynomials*, Math. Proc. Cambridge Philos. Soc. 108 (1990), 35–53.

[K] B. Kane, *A Silicon-based nuclear spin quantum computer*, Nature 393 (1998), 133-137.

[Ko] M. Kohmoto, *Topological invariants and the quantization of the Hall conductance*, Ann. Physics 160 (1985), 343–354.

[L] S. Lloyd, Universal quantum simulators, Science 273 (1996), 1073–1078.

[P] J. Preskill, *Fault tolerant quantum computation*, http://xxx.lanl.gov/ps/quant-ph/9712048, 19 Dec. 1997 (*Introduction to Quantum Computation* (H.-K. Lo, S. Popescu and T. P. Spiller, eds.) (to appear)).

[RT] N. Reshetikhin and V. Turaev, *Invariants of* 3*-manifolds via link polynomials and quantum groups*, Invent. Math. 103 (1991), 547–597.

[S] A. S. Schwarz, *The partition function of degenerate quadratic functional and Ray-Singer invariants*, Lett. Math. Phys. 2 (1977/78), 247–252.

[Sc] A. Schöhage, Proc. 6th ICALP Lect. Notes in Comp. Sci. (Springer, NY) 71 (1974), 520–529.

[Sh] P. Shor, *Algorithms for quantum computation: discrete logarithms and factoring*, 35th Annual Symposium on Foundations of Computer Science, Santa Fe, NM, 1994, IEEE Comput. Soc. Press, Los Alamitos, CA, 1994, pp. 124–134.

[TL] C. Ting, and C. H. Lai, *Spinning braid-group representation and the fractional quantum Hall effect* Nuclear Phys. B 396 (1993), 429–464.

[V] A. Vassiliev, *Cohomology of knot spaces. Theory of singularities and its applications*, Adv. Soviet Math. 1 (V. I. Arnold, ed.), Amer. Math. Soc. (Providence, RI), 1990.

[W] E. Witten *Quantum field theory and the Jones polynomial*, Comm. Math. Phys. 121 (1989), 351–399.

[We] Welsh, D. J. A. *Complexity: knots, colourings and counting* London Mathematical Society Lecture Note Series 186, Cambridge University Press, Cambridge, 1993.

[Wi] F. Wilczek, *Fractional statistics and anyon superconductivity* World Scientific Publishing Co., Inc., (Teaneck, NJ) 1990.

Michael H. Freedman
Microsoft Research
1 Microsoft Way
Redmond, WA 98052-6399

Toward a Global Understanding of $\pi_*(S^n)$

Mark Mahowald

Abstract. This talk will describe recent advances in getting a global picture of the homotopy groups of spheres. These results begin with the work of Adams on the homotopy determined by K-theory. Substantial new information follows from the nilpotence results of Devinatz, Hopkins and Smith.

1991 Mathematics Subject Classification: 55Q40,55Q45,55Pxx,55Txx
Keywords and Phrases: homotopy groups, spheres, periodicity in homotopy

1 Introduction

Until about 1960, the primary method used to calculate homotopy groups of spheres was the EHP sequence. This was invented by James at the prime 2 and Toda at odd primes. Early steps in this direction were taken by Freudenthal. Basically, the EHP sequence is a consequence of the result that

$$S^n \to \Omega S^{n+1} \to \Omega S^{2n+1}$$

is a 2 local fibration. At odd primes there is a similar result with some twists. Spectral sequences give a way to organize such calculations. We consider the filtration of $\Omega^{n-1}S^n$ given by

$$S^1 \to \Omega S^2 \to \cdots \to \Omega S^{n-2}S^{n-1} \to \Omega^{n-1}S^n.$$

When we apply homotopy to this filtration we get a spectral sequence in the standard fashion. The E_1 is given by

$$E_1^{s,t} = \pi_{t+1}(\Omega^{s-1}S^{2s-1}) = \pi_{t+s}(S^{2s-1}).$$

The key feature here is that the input to this spectral sequence is the output of an earlier calculation. In particular, once $\pi_1(S^1)$ is determined, no other outside calculation is necessary. This seductive feature attracted a lot of attention early on. This feature caused many to miss some obvious additional structure which is the current focus. If we look only at $E_1^{s,t}$ for $s \leq n$, the spectral sequence converges to $\pi_{t+n}S^n$. If we allow all s, the spectral sequences converges to the stable homotopy groups which we write as $\pi_t(S^0)$. The filtration induced on $\pi_t(S^0)$ refers to the

sphere of origin of the class. This means the smallest integer s such that the homotopy class is in the image of the suspension map $\Omega^{s-1}S^s \to \Omega^{\infty-1}S^\infty$. The class in $E_1^{s,t} = \pi_{t+s}(S^{2s-1})$ which projects to a class is called the Hopf invariant of that class. There are a few global results obtained essentially from the EHP sequence approach.

THEOREM 1.1 *(Serre) The groups, $\pi_j S^n$ are finite except if $j = n$ or if $n = 2k$ and $j = 4k - 1$.*

THEOREM 1.2 *(James and Toda) The E_2 term of the EHP spectral sequence is an $\mathbb{F}_p$ vector space.*

James at 2 and Toda at odd primes essentially proved this. This result gives an estimate of the maximum order of the torsion subgroup of $\pi_t S^n$. This result was sharpened to the best possible by the following result.

THEOREM 1.3 *(Cohen, Moore, and Neisendorfer) If $j \neq 2n + 1$ then $p^n \pi_j(S^{2n+1}) = 0$ for p an odd prime. There are classes of order p^n.*

At the prime 2 the sharpest estimate is not known. The result of James implies that $2^{2n}\pi_j(S^{2n+1}) = 0$. The maximum known elements would suggest a more complicated formula but approximately $2^{n+1}\pi_j(S^{2n+1}) = 0$. A precise conjecture is made in the next section.

There is another feature of the EHP spectral sequence which should be noted. Since $E_1^{s,t} = \pi_{t+s}(S^{2s-1})$ it is clear that if $t < 3s - 3$ then $E_1^{s,t}$ depends only on the value of $t - s$. In general, $E_r^{s,t} = E_r^{s+2^{r/2+1},t+2^{r/2+1}}$ provided that $2^{r/2+1} + t < 3(s + 2^{r/2+1}) - 3$. This allows one to describe a stable EHP spectral sequence in which $SE_1^{s,t} = \pi_{t-s-1}(S^0)$. This spectral sequence is defined for all $s \in \mathbb{Z}$. It is a consequence of Lin's theorem that this spectral sequence converges to $\pi_t(S^{-1})$. The paper by Mahowald and Ravenel [8] explores the consequences of this observation and gives complete references.

2 v_1 PERIODICITY

Another global result which does not follow from EHP considerations is the following result.

THEOREM 2.1 *(Nishida) Under composition, any element in a positive stem is nilpotent.*

It is this result which leaves one in a quandary as to how to describe an infinite calculation. Adams was the first to notice how to use Bott periodicity to construct infinite families. This is somewhat easier at odd primes but one can accomplish essentially the same thing by considering the finite complex, $Y^6 = \mathbb{C}P^2 \wedge \mathbb{R}P^2$ at $p = 2$ and $Y^k = S^{k-1} \cup_p e^k$ at p odd. We have the following result.

Proposition 2.2 *Let $q = 2(p-1)$. For each prime p and each $k > 6$ there is a map $Y^{k+q} \to Y^k$ such that all composites*

$$Y^{k+qj} \to Y^{k+q(j-1)} \to \cdots \to Y^k$$

are essential for all j and k. We will call this map v for any j and k.

This means that we can consider the homotopy theory, $[Y^*, \_]$ as a module over $\mathbb{Z}[v]$. We will label this homotopy modlule as $\pi_*(\_, Y)$. We can again ask for freeness and exponents with respect to this module. The question about freeness has been completely answered. The exponent question is completely open. A starting point for understanding freeness in this context is the following result. It is possible to compare the fibers of the single suspension map in the EHP sequence. This gives a new sequence of fibrations

$$W(n) \to S^{2n-1} \to \Omega^2 S^{2n+1}$$

In this context, Serre's theorem is equivalent to the assertion that $W(n)$ is rationally acyclic. To get information about $[Y^*, \_]$ for spheres, it is useful to compare $[Y^*, W(n)]$ for various n. The following result allows this.

Proposition 2.3 *There is a a map $W(n) \to \Omega^{2p} W(n+1)$ which induces an isomorphism in $v^{-1}\pi_*(\_, Y)$ homotopy.*

This proposition is key to determining the homotopy which can be detected in some sense by K-theory. In order to state the result we need to recall a small part of the Snaith splitting theorem. We will state the results for the prime 2. Something similar is true for odd primes.

Theorem 2.4 *(Snaith) There is a map*

$$s_n : \Omega^{2n+1} S^{2n+1} \to \Omega^\infty \Sigma_0^\infty \mathbb{R}P^{2n}$$

which induces a monomorphism in homology.

Using these maps we can prove the following.

Theorem 2.5 *The Snaith maps, s_n induce isomorphisms in the homotopy theory $v^{-1}\pi_*(\_, Y)$.*

All that remains is to compute this homotopy theory and that is an easy calculation. Thus a summand in $\pi_{k+2n+1}(S^{2n+1})$ for each $k \neq 4, 5 \bmod 8$ is determined. For certain values of n there are non-trivial summands for the other values of k. This aspect of homotopy theory is quite well understood. This material appears in several papers, the last one, [3], contains references to earlier work. The computational aspects is being pursued by Bendersky and Davis.

This discussion works in a very similar fashion at odd primes and the result is much easier to state. Let p be an odd prime and $q = 2p-2$. Let $\nu(k)$ be the maximum power of p which divides k.

THEOREM 2.6 *([13]) If $j = kq - 1$ or if $j = kq - 2$ then $j > 2n + 1$, then $\pi_{j+2n+1}(S^{2n+1})$ contains a $\mathbb{Z}/p^{\min(n,\nu(k)}$ summand.*

The homotopy detected by K-theory is special at the prime 2. At all odd primes in behaves in a similar fashion with the summand being defined by number theoretic functions as the above Theorem illustrates. In particular, at odd primes the elements of maximal order are found in the homotopy detected by K-theory. Typically, exponent theorems are proved by showing that the loop space power map has a certain order. In particular, we consider $P(r) : \Omega^{2n+1}S^{2n+1} \to \Omega^{2n+1}S^{2n+1}$ given by multiplication by p^r in the loop variable. Theorem 1.3 is proved by showing that $P(n)$ is null if p is odd. A result this simple at 2 is false. The conjectured result is:

CONJECTURE 2.7 *At the prime 2, where $P(n)$ refers to the 2^n power map we expect:*

- *If $n \equiv -1, 0$ mod 4, then $P(n)$ is null.*
- *If $n \equiv 1, 2$ mod 4 and $n > 1$, then $P(n + 1)$ is null.*
- *Among the torsion classes in $\pi_*(S^{2n+1})$, the element of maximal order is detected by K-theory.*

If all parts of this conjecture are correct, the proof would have to be quite different than the proof of Theorem 1.3 since we have the following result.

THEOREM 2.8 *If $n \equiv -1, 0$ mod 4, then there is a homotopy class of order 2^n detected by K-theory. If $n \equiv 1, 2$ mod 4 and $n > 1$, then the maximum order among the classes detected by K-theory is 2^{n-1}.*

A conjecture of this sort was first made by Barratt. This version is due to Barratt and Mahowald.

3 TELESCOPES AND LOCALIZATIONS

In order to understand the next kind of periodicity I want to introduce some additional notation. The first question which needs to be answered is: "For which finite complexes, F, are there maps, $v : \Sigma^k F \to F$, all of whose iterates are essential?" We will find it easier to suppress the suspension variable in this discussion. We are looking for maps like the map described above for Y. Devinatz, Hopkins and Smith [4]answered this question.

THEOREM 3.1 *Let F be a finite complex and $v : \Sigma^k F \to F$. The composite*

$$\Sigma^{k\cdot j}F \to \Sigma^{k(j-1)}F \to \cdots \to F$$

is essential for all j if and only if $MU_*(v) \neq 0$ where MU_* is complex bordism theory.

MU_* splits into a wedge of theories at a fixed prime. These smaller theories are called Brown-Peterson homology theories, BP_*. Their homotopy is given by $\pi_*(BP) = \mathbb{Z}[v_i, i = 1, \cdots]$. The dimension of v_i is $2(p^i - 1)$. In order to understand a particular periodicity family it is useful to localize BP. Consider the theory defined by $v_n^{-1}BP$. It is possible to get a more efficient theory by first killing v_i, for $i > n$ and inverting v_n in this new theory. Call the resulting spectrum, $E(n)$. We have $\pi_*(E(n)) = \mathbb{Z}_{(p)}[v_1, \cdots, v_n, v_n^{-1}]$. Work of Miller, Ravenel and Wilson, [9], show that this spectrum leads to an important localization. Note that $E(1) = K$ at 2. At other primes $E(1)$ is one of the factors into which K splits.

Bousfield, [2] has introduced a notion of localization at a spectrum. An excellent discussion of this is in the paper by Ravenel, [10]. A particularly important family of localizations is that given by localization with respect to $E(n)$. This gives rise to the chromatic tower. Let $L_n(X)$ be the Bousfield localization of X with respect to $E(n)$. Suppose X is a p-complete spectrum. Then there are commutative diagrams

$$\begin{array}{ccc} L_{i+1}(X) & \to & L_i(X) \\ \uparrow & & \uparrow \\ X & \simeq & X \end{array}$$

such that $X \to \text{homlim } L_i(X)$ is a homotopy equivalence.

The computations in the chromatic tower have been done for the stable sphere if $i = 1$ and all primes or $i = 2$ and the prime is larger than 3. For $i = 1$ the results of the previous section describe the answer. For $i = 2$ the result is very complicated and the reader is referred to the paper by Shimomura and Yabe, [12]. It is quite interesting to note that the Shimomura-Yabe result can be stated in terms of number theory functions for all primes $p > 3$. This is analogous to Theorem 2.6. These results suggest that the answer for the infinite prime might be possible. This would give the homotopy information in terms of functions whose argument is the prime and whose value is the order of a summand. In this sense, Theorem 2.6 and the Shimomura-Yabe result [12] are results for the infinite prime. It seems that if $n > p - 1$, then $L_n(S^0)$ should have such a prime independent description.

The situation for unstable spheres and L_2 localizations is still not clear. Bousfield has also defined localizations of spaces with respect to a spectrum. This seems to be a somewhat harder notion than localization of spectra. For S^{2k+1}, Arone and Mahowald, [1], have constructed a tower which reduces to a finite tower for each L_n. Here are some details. Let X be some space (or spectrum) and let F be some functor. Then Goodwillie [5] constructs a tower of functors

$$\begin{array}{ccccccccc} P_1F(X) & \leftarrow & P_2F(X) & \leftarrow & \cdots & \leftarrow & P_nF(X) & \leftarrow & \cdots \\ \downarrow & & \downarrow & & & & \downarrow & & \\ D_2F(X) & & D_3F(X) & & & & D_{n+1}F(X) & & \end{array}$$

and a collection of maps $F(X) \to P_nF(X)$ such that the inverse limit of the tower is equivalent to $F(X)$ and the fibers at each stage, $D_iF(X)$, are infinite loop spaces. For the example of the identity functor and for $X = S^{2k+1}$, this tower is investigated by Arone and Mahowald in [1]. For our purposes, the key result is the following.

THEOREM 3.2 *For each prime and each n, the L_n localization of S^{2k+1} can be represented by a tower of n fibrations. Each of the fibers is an infinite loop space. The fiber at the stage k, D_k, satisfies $L_{k-1}D_k = pt$.*

The key point is the observation that the Goodwillie tower is constant, except when $n = p^k$. In this case the stable spectrum represented by the fiber at the $n = p^k$ stage has acyclic homology with respect to the homology theory $E(k-1)$. This result can be used to compute the homotopy of L_nS^{2n+1} once one knows the stable theory. This has been done for L_1. It represents an interesting problem for L_2 at primes bigger than 3, in view of the Shimomura-Yabe calculations [12].

If $n > 1$, the homotopy theory defined by a finite complex with a self map detected by v_n seems to detect more homotopy than is present in L_nS^{2n+1}. That this should be the same is called the telescope conjecture. Recent work of Ravenel suggest this conjecture is false. Several proofs of the disproof of the telescope conjecture have been circulated but it is not yet clear if the result is proved.

4 FORMAL GROUPS AND HOMOTOPY THEORY

In addition, the connection of MU_* with formal groups has played an important role in understanding higher periodicities. The starting point is the multiplication map, $\mu : \mathbb{C}P\times\mathbb{C}P \to \mathbb{C}P$. Let $\alpha \in MU^2(\mathbb{C}P)$ represent the cohomology class given by $\mathbb{C}P = MU(1) \to \Sigma^2 MU$. Then $MU^*(\mu)(\alpha)$ is a power series in two variables. This power series, F, satisfies the axioms of a one dimensional commutative formal group over the ring $MU^*(pt)$. The key theorem is due to Quillen.

THEOREM 4.1 *(Quillen) The formal group constructed above induces an isomorphism from the Lazard ring to $MU^*(pt)$. All of the constructions in the theory of one dimensional commutative formal groups carry over to this topological setting.*

Hopkins and Miller have discovered a partial converse to this result. Let $\mathcal{FG}$ denote the category having as objects pairs (k, Γ), where k is a perfect field of characteristic p, and Γ is a formal group of height n over k, and with morphisms $\alpha : (k_1, \Gamma_1) \to (k_2, \Gamma_2)$ consisting of a pair (i, f), where i is a map $i : k_1 \to k_2$ of rings and f is an isomorphism $f : \Gamma_1 \to \Gamma_2$ of formal group laws. Then we have:

THEOREM 4.2 *(Hopkins-Miller) There exists a functor $(k, \Gamma) \to E_{k,\Gamma}$ from $\mathcal{FG}^{op}$ to the category of A_∞ ring spectra, such that,*

1. *$E_{k,\Gamma}$ is a commutative ring spectrum;*
2. *there is a unit in $\pi_2 E_{k,\Gamma}$;*
3. *$\pi_{odd}E_{k,\Gamma} = 0$, from which it follows that $E_{k,\Gamma}$ is complex orientable;*
4. *and such that the corresponding formal group law over $\pi_0 E_{k,\Gamma}$ is the the universal deformation of (k, Γ).*

A discussion of this result and related topics of formal groups and universal deformations is in the course notes prepared by Rezk [11].

Hopkins and Miller apply this result to construct higher K-theories, EO_n, at primes p where $(p-1)|n$. At 2 and 3, EO_2 is very interesting. In particular, this spectrum captures the way in which L_2 differs from the calculations of [12]. There is a connected version of EO_2 which is called eo_2. Various constructions of this spectrum yield various properties. In particular, Hopkins and Miller have constructed a version which makes eo_{2*} into an E_∞ ring spectrum. In [7] the homotopy groups eo_{2*} are computed. That paper also discusses the connection that this spectrum has with elliptic curves over $\mathbb{F}_4$ and height 2 elliptic curves. There will be a sequence of papers by Hopkins, Miller and others which expand on this theory. Without writing down specific groups, we observe that using this spectra we can show that a substantial part of the known calculation of the stable stems fit into periodic families. The basic periodicity of eo_2 at 2 is 192 which represents v_2^{32}. At the prime 3, the period is 72. How all of this should work out on unstable spheres is still not clear.

The connection with elliptic curves should be expanded on. Elliptic curves over a ring R can be co-represented by $\mathbb{Z}[a_1, a_2, a_3, a_4, a_6]$. The coefficients, a_i, are the coefficients in the Weierstrass form of the equation for the curve,

$$x^3 + a_2x^2 + a_4x + a_6 = y^2 + a_1xy + a_3y.$$

This equation represents a curve with a single point on the line at infinity. The discriminant, Δ, is a polynomial in the coefficients. If $\Delta \neq 0$, then the curve is non-singular. Coordinate transformations which preserve the curve are

$$\begin{aligned} x &\mapsto x + r \\ y &\mapsto y + sx + t \end{aligned}$$

These substitutions give transformation formulas for the coefficients. We can use these to construct a Hopf algebroid,

$$\mathbb{Z}[a_1, a_2, a_3, a_4, a_6] \rightrightarrows \mathbb{Z}[a_1, a_2, a_3, a_4, a_6, s, r, t]$$

The homology of this Hopf algebroid is the E_2 term of the Adams-Novikov spectral sequence to calculate $\pi_*(eo_2)$. The homology in dimension 0 is isomorphic to the ring of modular forms. There are differentials in the Adams-Novikov spectral sequence. For more details see [7].

The theories, E_{p-1}, are also related to curves. These curves, of genus $\binom{p-1}{2}$, give rise to formal groups in a more complicated fashion. This is discussed in the paper by Gorbounov and Mahowald, [6]. In this case too, the connection is exploited to give an easy calculation of the E_2 term of the Adams-Novikov spectral sequence.

References

[1] G. Arone and M. Mahowald: The Goodwillie tower of the identity functor and the unstable periodic homotopy of spheres, *Inventiones Mathematicae* (to appear).

[2] A. K. Bousfield: The localization of spectra with respect to homology, *Topology* 18 (1979) 257-281.

[3] D. Davis and M. Mahowald: The image of the stable J-homomorphism, *Topology*, 28 (1989) 39-58.

[4] E. Devinatz, M. Hopkins and J. Smith: Nilpotence and stable homotopy theory, *Ann. Math.* 128 (1988) 207-242.

[5] T. Goodwillie: Calculus II: analytic functors, *K-Theory* 4 (1992) 295-332.

[6] V. Gorbounov and M. Mahowald: Formal completion of the Jacobians of plane curves and higher K-theories, *J. of Pure and Applied Algebra*, (to appear).

[7] M. Hopkins and M. Mahowald: From elliptic curves to homotopy theory, preliminary version on Hopf archive ftp://hopf.math.purdue.edu/pub.

[8] M. E. Mahowald and D. C. Ravenel. The root invariant in homotopy theory, *Topology*, 32:865–898, 1993.

[9] H. Miller, D. Ravenel and W. Wilson: Periodic phenomena in the Adams-Novikov spectral sequence, *Ann. Math.*, 106 (1977) 469-516.

[10] D. Ravenel: Localization with respect to certain periodic homology theories, *Amer. J. Math.* 106 (1984) 351-414.

[11] C. Rezk: Notes on the Hopkins Miller Theorem, *Contemporary Math. Series* (to appear).

[12] K. Shimomura and A. Yabe: The homotopy groups $\pi_*(L_2(S^0))$, *Topology*, 34 (1995) 261-289.

[13] R. D. Thompson: The v_1 periodic homotopy groups of an unstable sphere at odd primes, *Trans. Amer. Math. Soc.* 319(2)(1990) 535-559.

Mark Mahowald
Mathematics Department
Northwestern University
Evanston IL 60208

A Filtration of the Set of Integral Homology 3-Spheres

Tomotada Ohtsuki

Abstract. A filtration on the set of integral homology 3-spheres is introduced, based on the universal perturbative invariant (the LMO invariant) or equivalently based on finite type invariants of integral homology 3-spheres. We also survey on these invariants related to quantum invariants.

1991 Mathematics Subject Classification: 57M
Keywords and Phrases: integral homology 3-spheres, knots, finite type invariant, quantum invariant, Kontsevich invariant, Vassiliev invariant

In 1989, Witten [14] proposed his famous formula of topological invariants of 3-manifolds, based on Chern-Simons gauge theory. The formula is given by using a path integral over all G connections on a 3-manifold M, where G is a fixed Lie group. Following combinatorial properties of the invariants predicted by Witten's formula, the invariants, what we call the *quantum G invariant*, denoted by $\tau_r^G(M)$, have been rigorously reconstructed by many researchers, say by using surgery presentations of 3-manifolds.

Since we have many Lie groups, we have obtained many quantum invariants of 3-manifolds in this decade. To control these many invariants we consider the following two approaches, where, as for the quantum invariants of knots (or links), we had obtained Kontsevich invariant and Vassiliev invariants by the two approaches.

- Unify them into an invariant.
- Characterize them with a common property.

By the first approach, we expect the existence of the universal invariant among quantum invariants, though the universal quantum invariant of 3-manifolds is not found yet. Instead of universal quantum invariants, we have the universal invariant Ω among perturbative invariants of 3-manifolds, where the perturbative G invariant is obtained from the quantum G invariant by "asymptotic expansion". By the second approach, we obtain the notion of finite type such that the d-th coefficient of a perturbative invariant is of finite type of degree d.

Further we consider how fine these invariants distinguish 3-manifolds; for simplicity we consider integral homology 3-spheres (**Z**HS's) instead of 3-manifolds, in this manuscript. To describe it, we obtain a filtration on the set of **Z**HS's by using the invariant Ω, or equivalently by using finite type invariants.

In section 1 we review roles of Kontsevich invariant and Vassiliev invariants to understand quantum invariants of knots (or links). In section 2, as invariants related to quantum invariants of 3-manifolds, we survey on perturbative invariants, the universal perturbative invariant (the LMO invariant) and finite type invariants. Further we introduce a filtration on the set of **Z**HS's based on these invariants in section 3.

1 Invariants of knots related to quantum invariants

We prepare some notations. Let X be a closed (possibly empty) 1-manifold. A *web diagram* on X is a uni-trivalent graph such that each univalent vertex of the graph is on X and a cyclic order of three edges around each trivalent vertex of the graph is fixed; see Figure 1 for examples of web diagrams. We denote by $\mathcal{A}(X;\mathcal{R})$ the quotient vector space over $\mathcal{R}$ spanned by web diagrams on X subject to the AS, IHX and STU relations (see Figure 2), where $\mathcal{R}$ is $\mathbf{C}$ or $\mathbf{Z}$. We define the degree of a web diagram to be half the number of univalent and trivalent vertices of the uni-trivalent graph of the web diagram. We denote by $\mathcal{A}(X;\mathcal{R})^{(d)}$ the subspace spanned by web diagrams of degree d. We denote by $\hat{\mathcal{A}}(X;\mathcal{R})$ the completion of $\mathcal{A}(X;\mathcal{R})$ with respect to the degree. We have a map $\mathcal{A}(X;\mathbf{C}) \to \mathbf{C}$ obtained by "substituting" a Lie algebra $\mathbf{g}$ to dashed lines of web diagrams and a representation R of $\mathbf{g}$ solid lines of web diagrams; we call the map *weight system* and denote it by $W_{\mathbf{g},R}$, and we define $\hat{W}_{\mathbf{g},R} : \hat{\mathcal{A}}(X;\mathbf{C}) \to \mathbf{C}[[h]]$ by $\hat{W}_{\mathbf{g},R}(D) = W_{\mathbf{g},R}(D)h^{\deg(D)}$. For precise definitions of these notations see for example [12].

Figure 1: A web diagram on S^1 and a web diagram on ϕ. For a web diagram on X, the uni-trivalent graph of the web diagram is depicted by dashed lines and X solid lines.

1.1 Kontsevich invariant

The quantum $(\mathbf{g}, R)$ invariant $Q^{\mathbf{g},R}(L)$ of a link L can be constructed by using monodromy of solutions of the Knizhnik-Zamolodchikov equation (the KZ equation) where a Lie algebra $\mathbf{g}$ and its representation R are included in the equation. By considering "universal" version of the KZ equation obtained by replacing $\mathbf{g}$ with a dashed line and R a solid line of web diagrams, we obtain Kontsevich invariant instead of the quantum $(\mathbf{g}, R)$ invariant; see for example [1]. In the following, instead of the original Kontsevich invariant, we use the modified Kontsevich invariant [8], denoted by $\hat{Z}(L) \in \hat{\mathcal{A}}(\coprod^l S^1;\mathbf{C})$ for a framed link L with l components.

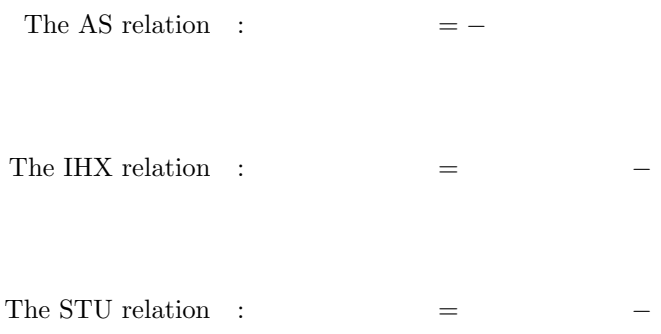
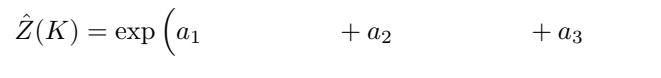


Figure 2: Definition of the AS, IHX and STU relations

For a knot K, the modified Kontsevich invariant $\hat{Z}(K)$ can be expressed as the exponential of a linear sum of web diagrams of connected dashed graphs as



because $\hat{Z}(K)$ is group-like with respect to a Hopf algebra structure of $\hat{\mathcal{A}}(S^1;\mathbf{C})$, and a group-like element can be, in general, expressed as the exponential of a primitive element, which is a linear sum of web diagrams of connected dashed graphs here; see for example [12] for this argument.

By the origin of Kontsevich invariant, any quantum invariant of links recovers from $\hat{Z}$ as

Theorem 1 (see for example [12]) (universality of $\hat{Z}$ among quantum invariants) *For any framed link L, we have*

$$\hat{W}_{\mathbf{g},R}\big(\hat{Z}(L)\big) = Q^{\mathbf{g},R}(L)|_{q=e^h}.$$

1.2 Vassiliev invariants

In this section we review Vassiliev invariants, which characterize quantum invariants in the sense of Corollary 3 below. We introduce Habiro's clasper to describe the weight system of a Vassiliev invariant.

Let $\mathcal{K}$ be the vector space freely spanned by isotopy classes of framed knots. For a knot K and a set C of crossings of K, we put

$$[K,C] = \sum_{C' \subset C} (-1)^{\#C'} K_{C'}$$

where the sum runs over all subset C' of C including the empty set and $K_{C'}$ denotes the knot obtained from K by crossing changes at the crossings of C'. We put $\mathcal{K}_d$ to be the vector subspace of $\mathcal{K}$ spanned by $[K, C]$ such that K is a knot and C is a set of d crossings of K. A linear map $v : \mathcal{K} \to \mathbf{C}$ is called a *Vassiliev invariant* (or a *finite type invariant*) of degree d if $v|_{\mathcal{K}_{d+1}} = 0$.

For a Vassiliev invariant v of degree d, we have a natural linear map $\varphi : \mathcal{A}(S^1; \mathbf{C})^{(d)} \to \mathcal{K}_d/\mathcal{K}_{d+1}$, which is called the *weight system* of v; see for example [1]. Habiro [5] gave a reconstruction of φ as shown in Figure 3 using claspers; see Figure 4 for the definition of Habiro's clasper.

Figure 3: Habiro's reconstruction of the map $\varphi : \mathcal{A}(S^1; \mathbf{C})^{(d)} \to \mathcal{K}_d/\mathcal{K}_{d+1}$ by his claspers. The image of a web diagram by φ in $\mathcal{K}_d/\mathcal{K}_{d+1}$ does not depend on the choice of a knot (the middle picture) and an embedding of the edges of claspers (in the right picture).

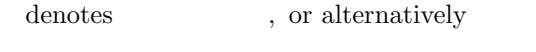


Figure 4: Definition of Habiro's clasper: The right picture implies the result obtained by integral surgery along Hopf link (in the picture) with 0 framings.

THEOREM 2 (KONTSEVICH) (universality of $\hat{Z}$ among Vassiliev invariants) *For any positive integer d, any Vassiliev invariant v of degree d is expressed as the composite map*

$$v : \{knots\} \xrightarrow{\hat{Z}} \hat{\mathcal{A}}(S^1; \mathbf{C}) \overset{projection}{\longrightarrow} \mathcal{A}(S^1; \mathbf{C})^{(\leq d)} \xrightarrow{W} \mathbf{C}$$

with some linear map W. Conversely, for any linear map $W : \mathcal{A}(S^1; \mathbf{C})^{(d)} \to \mathbf{C}$ the above composite map v is a Vassiliev invariant of degree d.

As a corollary of Theorems 1 and 2,[1] we have

COROLLARY 3 (see for example [1, 12]) *The d-th coefficient of $Q^{\mathbf{g},R}(K)|_{q=e^h}$ is a Vassiliev invariant of degree d* (BIRMAN-LIN). *Further the weight system of the Vassiliev invariant is equal to $W_{\mathbf{g},R}$* (PIUNIKHIN).

[1] Before the theorems Corollary 3 had been directly proved by Birman-Lin and Piunikhin.

2 Invariants of 3-manifolds related to quantum invariants

In this section, as invariants related to quantum invariants of 3-manifolds, we survey on perturbative invariants, the universal perturbative invariant (the LMO invariant) and finite type invariants.

2.1 Perturbative invariants of rational homology 3-spheres

By taking asymptotic expansion of Witten's path integral formula [14] of a quantum invariant, we obtain a power series as an invariant of 3-manifolds. As for the quantum G invariant $\tau_r^G(M)$ reconstructed rigorously, we take "asymptotic expansion" at $r \to \infty$ in the sense below.

For simplicity we consider the case $G = SO(3)$ here. It is known that the quantum $SO(3)$ invariant $\tau_r^{SO(3)}(M)$ belongs $\mathbf{Z}[\zeta]$ for any odd prime r,' where $\zeta = \exp(2\pi\sqrt{-1}/r)$. Further it is known [11] that for a rational homology 3-sphere M there exists the series of $\lambda_n \in \mathbf{Z}[1/2, 1/3, \cdots, 1/(2n+1)]$ satisfying

$$\begin{aligned}\tau_r^{SO(3)}(M) &= \left(\frac{|H_1(M;\mathbf{Z})|}{r}\right)\sum_{i=0}^{N}\lambda_n(\zeta-1)^n \\ &+ \Big(\text{terms divisible by } (\zeta-1)^{N+1} \text{ in } \mathbf{Z}[\zeta, \tfrac{1}{2}, \tfrac{1}{3}, \cdots, \tfrac{1}{2N+1}]\Big),\end{aligned}$$

for any positive integer N and any odd prime $r > \max\{2N+1, |H_1(M;\mathbf{Z})|\}$, where $\left(\frac{\cdot}{r}\right)$ denotes the Legendre symbol and $|H_1(M;\mathbf{Z})|$ denotes the order of the first homology group $H_1(M;\mathbf{Z})$. Using the series $\{\lambda_n\}$ we define the *perturbative* $SO(3)$ *invariant* of M by

$$\tau^{SO(3)}(M) = \sum_{n=0}^{\infty}\lambda_n(q-1)^n \in \mathbf{Q}[[q-1]].$$

For example the quantum $SO(3)$ invariant of the lens space $L(5,1)$ is expressed as

$$\begin{aligned}\tau_r^{SO(3)}\Big(L(5,1)\Big) &= \left(\frac{5}{r}\right)\zeta^{-3/5}\frac{\zeta^{1/10}-\zeta^{-1/10}}{\zeta^{1/2}-\zeta^{-1/2}} \\ &= \left(\frac{5}{r}\right)\Big(\frac{1}{5} - \frac{3}{5^2}(\zeta-1) + \frac{11}{5^3}(\zeta-1)^2 + \cdots\Big).\end{aligned}$$

Hence the perturbative $SO(3)$ invariant of $L(5,1)$ is given by

$$\begin{aligned}\tau^{SO(3)}\Big(L(5,1)\Big) &= \frac{1}{5} - \frac{3}{5^2}(q-1) + \frac{11}{5^3}(q-1)^2 + \cdots \\ &= q^{-3/5}\frac{q^{1/10}-q^{-1/10}}{q^{1/2}-q^{-1/2}} \in \mathbf{Q}[[q-1]].\end{aligned}$$

Perturbative invariants might dominate quantum invariants. This can be partially shown as follows, if we assumed Lawrence's conjecture "$\tau^{SO(3)}(M) \stackrel{?}{\in}$

$\mathbf{Z}[[q-1]]$ for any integral homology 3-sphere M". In the following diagram, where $F(q)$ is a cyclotomic polynomial, the value of $\tau^{SO(3)}(M)$ determines the value of $\tau_r^{SO(3)}(M)$ for any odd prime r.[2]

$$\begin{array}{ccccccc} & & & & & & \mathbf{Z}[[q-1]] \ni \tau^{SO(3)}(M) \\ & & & & & & \downarrow \\ \tau_r^{SO(3)}(M) \in & \mathbf{Z}[\zeta] & \overset{\text{isomorphic}}{\cong} & \mathbf{Z}[q]/F(q) & \overset{\text{injective}}{\hookrightarrow} & \mathbf{Z}[[q-1]]/F(q) \end{array}$$

2.2 THE LMO INVARIANT OF 3-MANIFOLDS

The LMO invariant, constructed as below, is universal among perturbative invariants. See also [2] for another approach to construct such a universal invariant among perturbative invariants.

As in [9] there is a series of linear maps $\iota_n : \mathcal{A}(\coprod^l S^1; \mathbf{C}) \to \mathcal{A}(\phi; \mathbf{C})$ obtained by replacing a solid circle including m dashed univalent vertices with some dashed univalent graph with m dashed univalent vertices, such that the degree $\leq n$ part of $\iota_n(\check{Z}(L))$ is invariant under Kirby move KII (handle slide move) on a framed link L, where $\check{Z}(L)$ is a different normalization of the modified Kontsevich invariant $\hat{Z}(L)$. Further, for the 3-manifold obtained from S^3 by integral surgery along L, the degree $\leq n$ part of

$$\Omega_n(M) = \iota_n(\check{Z}(U_+))^{-\sigma_+}\iota_n(\check{Z}(U_-))^{-\sigma_-}\iota_n(\check{Z}(L)) \in \hat{\mathcal{A}}(\phi; \mathbf{C})$$

becomes invariant under Kirby moves KI and KII, where $U_\pm$ is the unknot with ± 1 framing and $\sigma_\pm$ is the number of positive or negative eigenvalues of the linking matrix of L. Hence it becomes a topological invariant of M. Further, since the degree $< n$ part of $\Omega_n(M)$ can be expressed by $\Omega_{n-1}(M)$ as in [9], we put

$$\Omega(M) = \sum_{n=0}^{\infty} \Big(\text{the degree } n \text{ part of } \Omega_n(M)\Big) \in \hat{\mathcal{A}}(\phi; \mathbf{C})$$

and call it the *LMO invariant* (or the *universal perturbative invariant*). Further for a rational homology 3-sphere M we put

$$\hat{\Omega}(M) = \sum_{n=0}^{\infty} |H_1(M; \mathbf{Z})|^{-n} \Big(\text{the degree } n \text{ part of } \Omega_n(M)\Big) \in \hat{\mathcal{A}}(\phi; \mathbf{C}).$$

For example [9], for the 3-manifold $M_{n,k}$ obtained from S^3 by integral surgery along $(2, n)$ torus knot with k framing, the LMO invariant $\Omega(M_{n,k})$ is given by

$$\begin{aligned} \Omega(M_{n,k}) \quad = \quad & \exp\Big(\frac{1}{48}(3n^2 - k^2 + 3k - 5) \\ & + \frac{1}{2^7 \cdot 3^2}(12n^4 - 12kn^3 + 3k^2n^2 - 15n^2 + 12kn - 4k^2 + 4) \end{aligned}$$

[2] This argument was suggested by Thang Le.

$$+(\text{terms of degree} \geq 3)\Big).$$

In general $\Omega(M)$ can be expressed as the exponential of a linear sum of connected web diagrams like the case of $\hat{Z}(K)$ in section 1.1; see [9].

The following theorem is proved for $n = 2$ in [13] and for general n by Lev Rozansky and Thang Le.

Theorem 4 (universality of the LMO invariant among perturbative invariants) *For a rational homology 3-sphere M, the perturbative $PSU(n)$ invariant recovers from the LMO invariant as*

$$\tau^{PSU(n)}(M) = |H_1(M;\mathbf{Z})|^{-n(n-1)/2}\hat{W}_{sl_n}(\hat{\Omega}(M)).$$

2.3 Finite type invariants of integral homology 3-spheres

The notion of finite type invariants of integral homology 3-spheres (**Z**HS's) was introduced in [10] by replacing "crossing change" on knots in the definition of Vassiliev invariants (in section 1.2) with "integral surgery" on **Z**HS's. See also [3] for recent development of finite type invariants of 3-manifolds.

We call a framed link L in a **Z**HS M *algebraically split* if the linking number of any two components of L is zero, and call L *unit-framed* if the framing of each component of L is ± 1. Let $\mathcal{M}$ be the vector space over $\mathbf{C}$ freely spanned by homeomorphism classes of **Z**HS's. For a **Z**HS M and an algebraically split and unit-framed link L in M, we put

$$[M, L] = \sum_{L' \subset L} (-1)^{\#L'} M_{L'} \in \mathcal{M},$$

where the sum runs over all sublinks L' in L including the empty link, and $M_{L'}$ denotes the **Z**HS obtained from M by integral surgery along L'. Further we put $\mathcal{M}_d$ to be the vector subspace of $\mathcal{M}$ spanned by $[M, L]$ such that M is a **Z**HS and L is an algebraically split and unit-framed link with d components in M. Then, as shown in [4], the equalities $\mathcal{M}_{(3d)} = \mathcal{M}_{(3d-1)} = \mathcal{M}_{(3d-2)}$ hold. Hence we put again $\mathcal{M}_d = \mathcal{M}_{(3d)}$. A linear map $v : \mathcal{M} \to \mathbf{C}$ is defined to be of *finite type* of degree d if $v|_{\mathcal{M}_{d+1}} = 0$.

For a finite type invariant v of degree d, we have a linear map $\varphi : \mathcal{A}(\phi;\mathbf{C}) \to \mathcal{M}_d/\mathcal{M}_{d+1}$ associated with v as shown in [4]; the linear map φ is called the *weight system* of v. Habiro [5] gave a reconstruction of φ as shown in Figure 5. By Theorem 5 below, the weight system φ becomes an isomorphic linear map.

Theorem 5 ([7]) (universality of the LMO invariant among finite type invariants) *For any positive integer d, any finite type invariant v of degree d is expressed as the composite map*

$$v : \{\mathbf{Z}HS's\} \xrightarrow{\Omega} \hat{\mathcal{A}}(\phi;\mathbf{C}) \overset{projection}{\longrightarrow} \mathcal{A}(\phi;\mathbf{C})^{(\leq d)} \xrightarrow{W} \mathbf{C}$$

with some linear map W. Conversely, for any linear map $W : \mathcal{A}(\phi)^{(d)} \to \mathbf{C}$ the above composite map v is a finite type invariant of degree d.

As a corollary of Theorems 4 and 5,[3] we have

[3]Before the theorems Corollary 6 for $n = 2$ had beed directly proved by Kricker-Spence [6].

Figure 5: Habiro's reconstruction of the map $\varphi : \mathcal{A}(\phi;\mathbf{C})^{(d)} \to \mathcal{M}_d/\mathcal{M}_{d+1}$ by his claspers. The image of a web diagram by φ in $\mathcal{M}_d/\mathcal{M}_{d+1}$ does not depend on the choice of a **Z**HS (the middle picture) and an embedding of the edges of claspers (in the right picture).

COROLLARY 6 *The d-th coefficient of the perturbative $PSU(n)$ invariant is of finite type of degree d. Moreover the weight system of the finite type invariant is equal to W_{sl_n}.*

3 A FILTRATION OF THE SET OF INTEGRAL HOMOLOGY 3-SPHERES

How fine do quantum invariants distinguish **Z**HS's?[4] Since perturbative invariants might dominate quantum invariants (see section 2.1), this question might be reduced to study of perturbative invariants. Further, by universality of the LMO invariant among perturbative invariants (Theorem 4), the question might be reduced to study of the LMO invariant $\Omega(M)$. Furthermore, since $\Omega(M)$ can be expressed as $\Omega(M) = \exp\Big(\omega(M)\Big)$ (see section 2.2), the question is reduced to the question: how fine does the invariant $\omega(M)$ distinguish **Z**HS's?

On the other hand, noting that $\Omega(M)$ is a universal finite type invariant as in Theorem 5, we also consider the question: how fine do finite type invariants distinguish **Z**HS's? To distinguish them, we define the d-th equivalence relation $\sim_d$ among **Z**HS's as follows. Two **Z**HS's M and M' are *d-th equivalent*, denoted by $M \sim_d M'$, if $v(M) = v(M')$ for any finite type invariant v of degree $< d$. We have the ascending series of the sets of the equivalence classes;

$$\{\mathbf{Z}\text{HS's}\}/\sim_1 \longleftarrow \{\mathbf{Z}\text{HS's}\}/\sim_2 \longleftarrow \{\mathbf{Z}\text{HS's}\}/\sim_3 \longleftarrow \cdots.$$

Such equivalence relations have been originally studied by Habiro by using his claspers; to be precise his definition of the equivalence relations is slightly stronger than ours. Habiro showed that generators of $\{\mathbf{Z}\text{HS's}\}/\sim_d$ are given by web diagrams of degree $< d$ in the sense that the equivalence relation is generated

[4]For simplicity we here discuss for **Z**HS's, not for more general 3-manifolds. To distinguish, say, rational homology 3-spheres in the same way, we might need, not only the invariant ω, but isomorphism classes of cohomology rings.

by the relation obtained by putting the image of the map in Figure 5 to be zero. As a corollary of results of Habiro [5] we have

Theorem 7 (a corollary of results in [5]) *The set* $\{\mathbf{Z}HS\text{'}s\}/\sim_d$ *becomes a commutative group which is isomorphic to* $\mathcal{A}(\phi;\mathbf{Z})^{(<d)}_{conn}$*, where* $\mathcal{A}(\phi;\mathbf{Z})_{conn}$ *denotes the subspace of* $\mathcal{A}(\phi;\mathbf{Z})$ *spanned by connected web diagrams. By taking the direct limit of these isomorphisms we have the following homomorphism*

$$\{\mathbf{Z}HS\text{'}s\} \longrightarrow \varprojlim_d \{\mathbf{Z}HS\text{'}s\}/\sim_d \cong \hat{\mathcal{A}}(\phi;\mathbf{Z})_{conn}. \tag{1}$$

Since the map (1) gives a reconstruction of the invariant ω, Theorem 7 implies

Corollary 8 *The image of* $\omega : \{\mathbf{Z}HS\text{'}s\} \to \hat{\mathcal{A}}(\phi;\mathbf{C})_{conn}$ *is included in a lattice in* $\hat{\mathcal{A}}(\phi;\mathbf{C})_{conn}$ *which is isomorphic to* $\hat{\mathcal{A}}(\phi;\mathbf{Z})_{conn}$*.*

If the map (1) (or the invariant ω) was injective, we would identify $\{\mathbf{Z}\text{HS's}\}$ with a subset of $\hat{\mathcal{A}}(\phi;\mathbf{Z})_{\text{conn}}$ by the map (1), and all invariants related to quantum invariants would be understood as functions of weight systems on the space $\hat{\mathcal{A}}(\phi;\mathbf{Z})_{\text{conn}}$. Further we would expect that there should be structures on the set $\{\mathbf{Z}\text{HS's}\}$ induced by combinatorial structures on the space of web diagrams.

Similar arguments are also available for the set of knots. By using Vassiliev invariants, we define the equivalence relation $\sim_d$ on the set of knots in the same way as above. Then we have $\{\text{knots}\}/\sim_d \cong \mathcal{A}(S^1;\mathbf{Z})^{(<d)}_{\text{conn}}$ and the homomorphism

$$\{\text{knots}\} \longrightarrow \varprojlim_d \{\text{knots}\}/\sim_d \cong \hat{\mathcal{A}}(S^1;\mathbf{Z})_{\text{conn}},$$

which is a reconstruction of the invariant $z(K) \in \hat{\mathcal{A}}(S^1;\mathbf{C})_{\text{conn}}$, where $z(K)$ is the invariant satisfying $\hat{Z}(K) = \exp\Big(z(K)\Big)$; see section 1.1. We would identify the set of knots and would expect structures on the set of knots in the same sense as the above case of $\mathbf{Z}$HS's.

References

[1] Bar-Natan, D., *On the Vassiliev knot invariants*, Topology, 34 (1995) 423–472.

[2] Bar-Natan, D., Garoufalidis, S., Rozansky, L., Thurston, D.P., *The Aarhus invariant of rational homology 3-spheres I: A highly non trivial flat connection on* S^3, preprint, q-alg/9706004.

[3] Cochran, T.D., Melvin, P., *Finite type invariants of 3-manifolds*, preprint, math.GT/9805026.

[4] Garoufalidis, S., Ohtsuki, T., *On finite type 3-manifold invariants III: manifold weight systems*, to appear in Topology.

[5] Habiro, K., *Clasper theory and its applications*, in preparation.

[6] Kricker, A., Spence, B., *Ohtsuki's invariants are of finite type*, preprint, 1996, q-alg/9608007.

[7] Le, T.T.Q., *An invariant of homology 3-sphere which is universal for finite type invariants*, preprint, q-alg/9601002.

[8] Le, T.T.Q., Murakami, J., *The universal Vassiliev-Kontsevich invariant for framed oriented links*, Comp. Math. 102 (1996) 41–64.

[9] Le, T.T.Q., Murakami, J., Ohtsuki, T., *On a universal perturbative invariant of 3-manifolds*, Topology 37 (1998) 539–574.

[10] Ohtsuki, T., *Finite type invariants of integral homology 3-spheres*, J. Knot Theory and Its Rami. 5 (1996) 101-115.

[11] ———, *A polynomial invariant of rational homology 3-spheres*, Invent. Math. 123 (1996) 241–257.

[12] ———, *Combinatorial quantum method in 3-dimensional topology*, lecture note at Oiwake seminar, preprint 1996, available at http://www.is.titech.ac.jp/labs/ohtsukilab.

[13] ———, *The perturbative $SO(3)$ invariant of rational homology 3-spheres recovers from the universal perturbative invariant*, to appear in Topology.

[14] Witten, E., *Quantum field theory and the Jones polynomial*, Commun. Math. Phys. 121 (1989) 360–379.

Tomotada Ohtsuki
Department of Mathematical
and Computing Sciences
Tokyo Institute of Technology
Oh-okayama, Meguro-ku
Tokyo, 152-8552
Japan
tomotada@is.titech.ac.jp

Vector Bundles over Classifying Spaces

Bob Oliver

Abstract. Let $\mathbb{K}(X)$ denote the Grothendieck group of the monoid of (complex) vector bundles over any given space X. This is not in general the same as the K-theory group $K(X)$. When $X = BG$, the classifying space of a compact Lie group G, then $K(BG)$ has already been described by Atiyah and Segal as a certain completion of the representation ring $R(G)$. The main result described here is that the Grothendieck group $\mathbb{K}(BG)$ also can be described explicitly, in terms of the representation rings of certain subgroups of G, and compared with both $R(G)$ and $K(BG)$.

1991 Mathematics Subject Classification: Primary 55R50, secondary 55R35, 55R25
Keywords and Phrases: vector bundles, classifying spaces, K-theory

A vector bundle over a space X can be thought of as a collection of finite dimensional vector spaces (the fibers), one for each point in X, which are combined together in one topological space. A product $X \times \mathbb{R}^n$ or $X \times \mathbb{C}^n$ is a "trivial" vector bundle over X. The simplest example of a nontrivial vector bundle is the Möbius band, regarded as a vector bundle over the circle with fibers $\mathbb{R}^1$. One standard source of vector bundles is the tangent bundle of a smooth manifold; i.e., the union of the tangent planes at all points of the manifold (given an appropriate topology).

We focus attention here on the case of *complex* vector bundles; i.e., vector bundles whose fibers are complex vector spaces. For any topological space X, let $\mathrm{Vect}(X)$ denote the set of isomorphism classes of complex vector bundles over X. This is a commutative monoid under the operation of direct sum. Define $\mathbb{K}(X)$ to be the Grothendieck group of $\mathrm{Vect}(X)$; i.e., the abelian group of all formal differences $x - x'$ for $x, x' \in \mathrm{Vect}(X)$ (with the obvious relations).

When X is compact (e.g., a finite cell complex), then $K(X) \stackrel{\text{def}}{=} \mathbb{K}(X)$ is just the K-theory of X. For such X (in fact, for finite dimensional X), these groups define a cohomology theory. In other words, they have nice properties, such as forming exact sequences and Bott periodicity, which provide useful tools for calculating and applying these groups.

When working with non-compact spaces, and in particular with infinite dimensional spaces, the Grothendieck groups of vector bundles do not define a cohomology theory. So a different definition of the K-theory of X is used, one involving classifying spaces. For each n, there is a classifying space $BU(n)$ for the unitary group $U(n)$ which "classifies" bundles, in the sense that the set $\mathrm{Vect}_n(X)$ of n-dimensional complex vector bundles over X is in one-to-one correspondence with

the set $[X, BU(n)]$ of homotopy classes of maps $X \to BU(n)$. More generally, for any topological group G, there is a space BG which classifies all fiber bundles with "structure group" G. For example, the structure group of an n-dimensional complex vector bundle is $GL_n(\mathbb{C})$, the group of self-transformations of the fiber $\mathbb{C}^n$ which preserve the vector space structure; but any vector bundle can be given an essentially unique (fiberwise) hermitian product which reduces its structure group to the unitary group $U(n)$.

The classifying space construction is functorial, in the sense that any homomorphism $\rho : G \to G'$ induces a map of spaces $B\rho : BG \to BG'$. The space BG is characterized (up to homotopy equivalence) as being the orbit space of a contractible space EG with a free G-action.

Regard $U(n)$ as a subgroup of $U(n+1)$ by identifying an $n \times n$-matrix A with the $(n+1) \times (n+1)$-matrix $\left(\begin{smallmatrix} A & 0 \\ 0 & 1 \end{smallmatrix}\right)$. This allows us to consider $BU(n)$ as a subspace of $BU(n+1)$, and hence to define $BU = \bigcup_{n=1}^{\infty} BU(n)$. For arbitrary X, the K-theory of X is now defined by setting $K(X) \stackrel{\text{def}}{=} [X, \mathbb{Z} \times BU]$. This agrees with the earlier definition $K(X) = \mathbb{K}(X)$ when X is compact, and defines a cohomology theory for general X. There is a natural homomorphism

$$\beta_X : \mathbb{K}(X) \longrightarrow K(X),$$

which is an isomorphism whenever X is compact, but not in general.

The geometrically defined functor $\mathbb{K}(-)$ can behave very differently from $K(-)$. For example, the sequence

$$\widetilde{\mathbb{K}}(\mathbb{C}P^\infty/\mathbb{R}P^\infty) \longrightarrow \mathbb{K}(\mathbb{C}P^\infty) \xrightarrow{\text{restr}} \mathbb{K}(\mathbb{R}P^\infty)$$

is not exact: if ξ_n (for $n \in \mathbb{Z}$) denotes the line bundle over $\mathbb{C}P^\infty$ with Chern class n times some fixed generator of $H^2(\mathbb{C}P^\infty) \cong \mathbb{Z}$, then $[\xi_1] - [\xi_3]$ lies in the kernel of the above restriction map, but not in the image of $\mathbb{K}(\mathbb{C}P^\infty/\mathbb{R}P^\infty)$. One can also show that Bott periodicity fails for the functor $\mathbb{K}(-)$ (see the remarks after Theorem 1.1 in [JO]).

If G is a compact Lie group and BG is its classifying space, then $\mathbb{K}(BG)$ and $K(BG)$ can both be studied by comparing them with the representation ring $R(G)$ of G. Let $\mathrm{Rep}(G)$ be the commutative monoid of isomorphism classes of complex finite dimensional G-representations (with addition defined by direct sum). Define

$$\alpha'_G : \mathrm{Rep}(G) \longrightarrow \mathrm{Vect}(BG)$$

by sending a complex G-representation V to its Borel construction $(EG \times_G V)$, regarded as a vector bundle over BG. Equivalently, if one thinks of a representation of G as a homomorphism $\rho : G \to GL_n(\mathbb{C})$, then α'_G sends ρ to the vector bundle classified by $B\rho : BG \to BGL_n(\mathbb{C})$. This is a homomorphism of monoids, and upon passing to Grothendieck groups we obtain a homomorphism

$$\alpha_G : R(G) \longrightarrow \mathbb{K}(BG)$$

of abelian groups. The *completion theorem* of Atiyah and Segal [AS] says that the composite

$$R(G) \xrightarrow{\alpha_G} \mathbb{K}(BG) \xrightarrow{\beta_{BG}} K(BG)$$

extends to an isomorphism $\widehat{\alpha}_G : R(G)\widehat{\ } \xrightarrow{\cong} K(BG)$, where

$$R(G)\widehat{\ } = \varprojlim_n (R(G)/I^n)$$

is completion with respect to the augmentation ideal $I = \mathrm{Ker}[R(G) \to \mathbb{Z}]$.

The main result described here, which was joint work with Stefan Jackowski [JO], is a description of the Grothendieck group $\mathbb{K}(BG)$ itself, also in terms of representations. This in turn grew out of earlier work by the two of us together with Jim McClure ([JMO], [JMO2], [JMO3]) dealing with maps between classifying spaces of arbitrary pairs of compact Lie groups. Note that the monoid $\mathrm{Vect}(BG)$ is the disjoint union of the sets $\mathrm{Vect}_n(BG) \cong [BG, BU(n)]$.

The starting point when computing $\mathbb{K}(BG)$ was to consider the case where G is a finite p-group, or more generally a p-toral group: an extension of a torus by a finite p-group. By theorems of Dwyer-Zabrodsky [DZ] (when G is a finite p-group) and of Notbohm [Nb] (when G is p-toral),

$$[BG, BG'] \cong \mathrm{Hom}(G, G')/\mathrm{Inn}(G')$$

for any p-toral group G and *any* compact Lie group G'. In particular, when $G' = U(n)$, this says that $\mathrm{Vect}_n(BG) \cong \mathrm{Rep}_n(G)$. In other words,

$$\alpha'_G : \mathrm{Rep}(G) \xrightarrow{\cong} \mathrm{Vect}(BG)$$

is an isomorphism whenever G is p-toral, and hence $\mathbb{K}(BG) \cong R(G)$ for such G.

Now let G be an arbitrary compact Lie group. For each p-toral subgroup P of G, consider the composite

$$\mathrm{Vect}(BG) \xrightarrow{\mathrm{restr}} \mathrm{Vect}(BP) \xrightarrow[\cong]{\alpha_P^{-1}} \mathrm{Rep}(P) \subseteq R(P),$$

where $R(P)$ is the complex representation ring of G. These maps combine to define a homomorphism

$$r_G : \mathrm{Vect}(BG) \longrightarrow R_{\mathcal{P}}(G) \stackrel{\mathrm{def}}{=} \varprojlim_P R(P),$$

where the inverse limit is taken over all p-toral subgroups of G (for all primes p) with respect to inclusion and conjugation of subgroups. The main theorem in [JO] is the following:

THEOREM 1. *For any compact Lie group G, r_G extends to an isomorphism*

$$\bar{r}_G : \mathbb{K}(BG) \xrightarrow{\cong} R_{\mathcal{P}}(G).$$

A more precise version of this theorem is given as Theorem 1′ below.

Of particular interest is the question of when the homomorphism $\alpha_G : R(G) \to \mathbb{K}(BG)$ is surjective; i.e., for which groups all vector bundles over BG are induced (stably, at least) by virtual representations of G. This is the case whenever G is finite, or whenever $\pi_0(G)$ is a p-group for some p (in particular if G is connected). In fact, $\mathbb{K}(BG) \cong R(G)$ (α_G is an isomorphism) whenever $\pi_0(G)$ (the group of connected components of G) has the property that all of its elements have prime power order. Note that this property — all elements have prime power order — is held not only by p-groups, but also by other finite groups such as Σ_4 and A_5. These, and other conditions which imply α_G is onto, are shown in [Ol, Corollary 3.11].

In general, α_G is not surjective, but its cokernel always has finite exponent. The simplest example of a group G for which $\alpha_G : R(G) \to \mathbb{K}(BG)$ is not onto, i.e., not all vector bundles are induced by virtual representations, was given by Adams in [Ad, Example 1.4]. For the group $G = (S^1 \times_{C_2} Q(8)) \times C_3$, he constructed a 2-dimensional complex vector bundle $\xi \to BG$ whose class does not lie in the image of $R(G)$. The cokernel of $\bar{\alpha}_G^{\mathbb{C}}$ for arbitrary G is described in [Ol, Lemma 3.8 and Theorem 3.9].

Another consequence of Theorem 1 is that $\beta_{BG} : \mathbb{K}(BG) \to K(BG)$ is always injective. This result follows upon combining the description of $\mathbb{K}(BG)$ as the inverse limit of the representation rings $R(P)$ for p-toral $P \subseteq G$, with a result of Segal [Se, Proposition 3.10] that $R(P)$ injects into $K(BP) \cong \widehat{R(P)}$ whenever $\pi_0(P)$ is a p-group (and in particular whenever P is p-toral).

The image of $\beta_{BG} : \mathbb{K}(BG) \rightarrowtail K(BG) \cong \widehat{R(G)}$ can be described directly, in terms of the exterior power operations on $K(BG)$. For any space X, homomorphisms $\lambda^k : K(X) \to K(X)$ are defined (for all $k \geq 0$) which send the class of any vector bundle over X to the class of its k-th (fiberwise) exterior power. Adams [Ad] defined and studied the subgroup $FF(BG) \subseteq K(BG)$ generated by the "formally finite dimensional elements"; i.e., those elements $x \in K(BG)$ such that $\lambda^k(x) = 0$ for k sufficiently large. Clearly, the class in $K(BG)$ of any vector bundle over BG satisfies this condition, and hence the image of β_{BG} is contained in $FF(BG)$. The results described here, when combined with those in [Ad], imply that in fact $FF(BG) = \mathrm{Im}(\beta_{BG})$.

The following two examples help describe the difference between the groups $K(BG)$ and $\mathbb{K}(BG)$. Consider first the case $G = T^n$: the n-dimensional torus. Here, $\mathbb{K}(BG) \cong R(G) \cong \mathbb{Z}[t_1, \dots, t_n]$, where the generators t_i all represent one-dimensional representations. The augmentation ideal (i.e., the ideal of virtual zero-dimensional representations) is thus generated as an ideal by the elements $x_i = t_i - 1$. Since $R(G)$ is also generated as a polynomial algebra by the x_i, we see that $K(BG) \cong \widehat{R(G)} \cong \mathbb{Z}[[x_1, \dots, x_n]]$ is a power series algebra.

As a second example, let G be any finite group. Let r_p (for any prime p) denote

the number of conjugacy classes of elements of p-power order. Then

$$\mathbb{K}(BG) \cong \mathbb{Z} \times \prod_p (\mathbb{Z})^{r_p - 1}, \qquad \text{while} \qquad K(BG) \cong \mathbb{Z} \times \prod_p (\widehat{\mathbb{Z}}_p)^{r_p - 1}.$$

(This last statement follows from the proof of [Ja, Theorem 2.2].)

The above discussion has focused on the case of complex bundles, but all of these results are also shown in [JO] to hold for real bundles. In particular, if $\mathbb{KO}(BG)$ denotes the Grothendieck group of the monoid of real vector bundles over BG, and $RO_{\mathcal{P}}(G)$ is the inverse limit over p-toral subgroups $P \subseteq G$ (for all p) of the real representation rings $RO(P)$, then $\mathbb{KO}(BG) \cong RO_{\mathcal{P}}(G)$.

The proof of Theorem 1 is based on the description by Dwyer-Zabrodsky and Notbohm of $\mathbb{K}(BG)$ when G is p-toral (see Proposition 4 below), and a certain decomposition of BG as a homotopy direct limit of classifying spaces of p-toral subgroups. This decomposition is described as follows. Fix a compact Lie group G and a prime p. A p-toral subgroup $P \subseteq G$ is called *p-stubborn* if $N(P)/P$ is finite and contains no nontrivial normal p-subgroup $1 \neq Q \lhd N(P)/P$. Let $\mathcal{R}_p(G)$ denote the category whose objects are the orbits G/P for p-stubborn subgroups $P \subseteq G$, and where $\mathrm{Mor}_{\mathcal{R}_p(G)}(G/P, G/P')$ is the (finite) set of G-maps $G/P \to G/P'$.

PROPOSITION 2. [JMO, Theorem 1.4] *For each prime p, the map*

$$\varinjlim_{G/P \in \mathcal{R}_p(G)} (EG/P) \longrightarrow BG,$$

induced by the projections $EG/P \to EG/G = BG$, is an $\mathbb{F}_p$-homology equivalence.

Here, EG can be any contractible complex with free action of G, and with orbit space BG. Note that $EG/P \simeq BP$ for each P (since the free G-action restricts to a free P-action). Thus, Proposition 2 describes BG, at least p-locally, as a limit of classifying spaces of p-toral subgroups of G.

For any space X, let $X^\wedge_p$ denote the p-adic completion of Bousfield and Kan. This will be used here only when X is 1-connected and its homotopy groups have finite type. So $X^\wedge_p$ can just be thought of as a space together with a map $X \to X^\wedge_p$, which induces isomorphisms $\widehat{\mathbb{Z}}_p \otimes \pi_i(X) \to \pi_i(X^\wedge_p)$ and $\widehat{\mathbb{Z}}_p \otimes H_i(X) \to H_i(X^\wedge_p)$ for all i. Proposition 2 implies that for any such X, the natural maps $EG/P \to BG$ induce a homotopy equivalence

$$\mathrm{Map}(BG, X^\wedge_p) \xrightarrow{\ \simeq\ } \mathrm{Map}\Big(\varinjlim_{G/P \in \mathcal{R}_p(G)} \big(EG/P\big), X^\wedge_p \Big).$$

Thus, in order to study maps to $BU(n)$, it is first necessary to look at maps to the p-completions $BU(n)^\wedge_p$. The following proposition, based on Sullivan's arithmetic pullback square, describes how the information about maps to the p-adic completions can be pieced together to give information about maps to $BU(n)$ itself.

PROPOSITION 3. [JMO3, Proposition 1.2] *Let $T \subseteq G$ be a maximal torus of G, and set $w = |N(T)/T|$. Then the following square is a pullback:*

$$\begin{array}{ccc} [BG, BU(n)] & \longrightarrow & \prod_{p|w}[BG, BU(n)^\wedge_p] \\ {\scriptstyle \text{restr}}\downarrow & & {\scriptstyle \text{restr}}\downarrow \\ [BT, BU(n)] & \longrightarrow & \prod_{p|w}[BT, BU(n)^\wedge_p]. \end{array}$$

The goal now is to describe maps from BG to $BU(n)$ or $BU(n)^\wedge_p$, by replacing BG by the homotopy direct limit of spaces BP described in Proposition 2. To do this, one must understand, not only the sets $[BP, BU(n)^\wedge_p]$ of homotopy classes of maps, but also the higher homotopy groups of the connected components of the mapping spaces $\mathrm{Map}(BP, BU(n)^\wedge_p)$. These can be described with the help of the next proposition, where we also repeat the description of $[BP, BU(n)]$ mentioned earlier.

For any P-representation V (assumed to have a G-invariant hermitian product), $\mathrm{Aut}(V)$ denotes the group of all unitary automorphisms of V, and $\mathrm{Aut}_P(V)$ the subgroup of all P-equivariant unitary automorphisms.

PROPOSITION 4. [DZ],[Nb] *For any prime p and any p-toral group P, the homomorphism*

$$\alpha'_P : \mathrm{Rep}(P) \xrightarrow{\cong} \mathrm{Vect}(BP) \cong \coprod_{n=0}^{\infty}[BP, BU(n)]$$

is an isomorphism of monoids. Also, for any P-representation V, corresponding to a homomorphism $\rho : P \to \mathrm{Aut}(V)$, the homomorphism $P \times \mathrm{Aut}_P(V) \xrightarrow{(\rho,\mathrm{incl})} \mathrm{Aut}(V)$ induces (by adjointness) a homotopy equivalence

$$B\,\mathrm{Aut}_P(V)^\wedge_p \xrightarrow{\simeq} \mathrm{Map}\big(BP, B\,\mathrm{Aut}(V)^\wedge_p\big)_{B\rho}.$$

Here, $\mathrm{Map}(-,-)_{B\rho}$ denotes the connected component of $B\rho : BP \to B\,\mathrm{Aut}(V)$.

Proposition 4 is a special case of more general results, which describe mapping spaces $\mathrm{Map}(BP, BG')$ (where P is p-toral and G' is any compact Lie group), or $\mathrm{Map}(BP, BG'^\wedge_p)$ (where G' must be connected.)

The next proposition, due to Wojtkowiak, describes the obstructions which are encountered when trying to compare the set of homotopy classes of maps $\varinjlim(X_\alpha) \to Y$ defined on a homotopy direct limit, with the inverse limit of the sets $[X_\alpha, Y]$ of maps defined on the "pieces". These obstructions turn out to be higher derived functors of certain inverse limits over the indexing category.

PROPOSITION 5. [Wo] *Fix a discrete category $\mathcal{C}$, and a (covariant) functor $F : \mathcal{C} \to \mathsf{Top}$. Let Y be any space, and fix maps $f_c : F(c) \to Y$ (for all $c \in \mathrm{Ob}(\mathcal{C})$) whose homotopy classes define an element $\widehat{f} = ([f_c])_{c\in\mathcal{C}} \in \varprojlim[F(-), Y]$. Set*

$$\alpha_n(c) = \pi_n\big(\mathrm{Map}(F(c), Y), f_c\big)$$

for all $c \in \mathrm{Ob}(\mathcal{C})$. *Then the obstructions to constructing a map* $f : \underset{\longrightarrow}{\mathrm{hocolim}}(F) \to Y$ *such that* $f|_{F(c)} \simeq f_c$ *for each* c *lie in the groups* $\varprojlim^{n+1}(\alpha_n)$ *for* $n \geq 1$. *Also, given two maps* $f, f' : \underset{\longrightarrow}{\mathrm{hocolim}}(F) \to Y$ *such that* $f|_{F(c)} \simeq f_c \simeq f'|_{F(c)}$ *for each* c, *the obstructions to* f *and* f' *being homotopic lie in the groups* $\varprojlim^{n}(\alpha_n)$ *for* $n \geq 1$.

As usual, these obstructions are iterative, in that the i-th obstruction is defined only if the $(i-1)$-st vanishes (and the i-th obstruction may depend on choices made in earlier constructions). The above formulation avoids certain technical points involving basepoints of mapping spaces and nonabelian fundamental groups; problems which are dealt with in detail in [Wo].

When applying Proposition 5, we will need to deal with the higher limits of homotopy groups of mapping spaces $\mathrm{Map}\big(EG/P, BU(n)^\wedge_p\big)$. The homotopy groups of these spaces are in general unknown or difficult to compute, and this is one of the reasons for the difficulty in describing precisely the sets $\mathrm{Vect}_n(BG) \cong [BG, BU(n)]$. But since we are interested in the Grothendieck group of vector bundles, and not in the vector bundles themselves, it suffices to handle these mapping spaces and groups after stabilizing: more precisely, after taking certain direct limits over all $V \in \mathrm{Rep}(G)$. Very roughly, the following proposition says that while higher limits *can* influence the monoid $\mathrm{Vect}(BG)$ of vector bundles, they have no effect on the Grothendieck group $\mathbb{K}(BG)$.

PROPOSITION 6. [JO, Proposition 1.5] *For each* $i > 0$, *let* $\Pi_i : \mathcal{R}_p(G) \to \widehat{\mathbb{Z}}_p\text{-}\mathtt{mod}$ *be the functor defined by setting*

$$\Pi_i(G/P) = \varinjlim_{V \in \mathrm{Rep}(G)} \pi_i\Big(\mathrm{Map}(EG/P, B\,\mathrm{Aut}(V)^\wedge_p), B\rho_V|_{BP}\Big),$$

where $\rho_V : G \to \mathrm{Aut}(V)$ *is induced by the action of* G *on* V. *Then* $\Pi_i \cong \widehat{\mathbb{Z}}_p \otimes K_G^{-i}(-)$ *as functors on* $\mathcal{R}_p(G)$, *and*

$$\varprojlim^{j}_{\mathcal{R}_p(G)} \Pi_i = 0$$

for all $i, j > 0$.

We are now ready to sketch the proof of Theorem 1. In fact, we prove a somewhat stronger statement. Recall that $r_G : \mathrm{Vect}(BG) \to R_{\mathcal{P}}(G)$ is defined as the inverse limit of the homomorphisms

$$\mathrm{Vect}(BG) \xrightarrow{\ \mathrm{restr}\ } \mathrm{Vect}(BP) \cong \mathrm{Rep}(P) \subseteq R(P)$$

for p-toral subgroups $P \subseteq G$, and that $\bar{r}_G : \mathbb{K}(BG) \to R_{\mathcal{P}}(G)$ is induced by r_G upon passing to Grothendieck groups.

To simplify the notation, when V is a G-representation, η_V will denote the vector bundle induced by the Borel construction on V: $\eta_V = \big(EG \times_G V \to BG\big)$.

THEOREM 1′. [JO] *For any compact Lie group G, $\bar r_G : \mathbb{K}(BG) \xrightarrow{\cong} R_{\mathcal{P}}(G)$ is an isomorphism of groups. More precisely, the following two statements hold.*

(a) For each pair of bundles $\xi, \xi' \to BG$ such that $r_G(\xi) = r_G(\xi')$, there exists a G-representation V such that $\xi \oplus \eta_V \cong \xi' \oplus \eta_V$.

(b) For each $X \in R_{\mathcal{P}}(G)$, there exist a vector bundle $\xi \to BG$ and a G-representation V such that $r_G(\xi) = X + r_G(\eta_V)$.

Any vector bundle over BG can be embedded as a summand of a bundle η_V for some G-representation V; and $\mathbb{K}(BG)$ can thus be obtained from $\mathrm{Vect}(BG)$ *by inverting only those vector bundles coming from G-representations.*

Outline of the proof. The injectivity of $\bar r_G$ follows immediately from point (a), and the surjectivity from (b). The last statement also follows easily from (a) and (b).

For any map $f : BG \to BU(n)$, we write ξ_f for the corresponding vector bundle over BG; i.e., for the pullback via f of the universal vector bundle over $BU(n)$. For any (n-dimensional) G-representation V, $f_V : BG \to BU(n)$ denotes the classifying map of the corresponding homomorphism $G \to U(n)$; or equivalently the classifying map of the vector bundle η_V.

We focus attention on the proof of (a). Fix maps $f, g : BG \to BU(n)$ such that $r_G(\xi_f) = r_G(\xi_g)$. In other words, for each p and each p-toral subgroup $P \subseteq G$, $f|_{BP} \simeq g|_{BP}$. We must show that there is a G-representation W for which

$$f \oplus f_W \simeq g \oplus f_W .$$

By Proposition 3, it suffices to show that $(f \oplus f_W)^\wedge_p \simeq (g \oplus f_W)^\wedge_p$ for each prime p. This is a problem only for primes $p \big| |N(T)/T|$ (where T is a maximal torus in G); hence only for a finite number of primes. It thus suffices to find W_p, for each p, such that

$$(f \oplus f_{W_p})^\wedge_p \simeq (g \oplus f_{W_p})^\wedge_p ; \tag{1}$$

and then set $W = \bigoplus_{p \mid |NT/T|} W_p$.

Fix a prime $p \big| |N(T)/T|$. For each i, let $\Pi_i^{(f)} : \mathcal{R}_p(G) \to \widehat{\mathbb{Z}}_p\text{-}\mathfrak{mod}$ be the functor

$$\Pi_i^{(f)}(G/P) = \pi_i\Big(\mathrm{Map}(EG/P, BU(n)^\wedge_p), f|_{BP}\Big).$$

By Proposition 2,

$$[BG, BU(n)^\wedge_p] \cong \Big[\operatorname*{hocolim}_{G/P \in \mathcal{R}_p(G)} \big(EG/P\big), BU(n)^\wedge_p \Big].$$

So by Proposition 5, the successive obstructions to constructing a homotopy $f \simeq g$ lie in the groups $\varprojlim^i_{\mathcal{R}_p(G)} \Pi_i^{(f)}$ (for all $i \geq 1$). Since $\mathcal{R}_p(G)$ is equivalent to a finite category [JMO, Proposition 1.6], higher derived functors of inverse limits over $\mathcal{R}_p(G)$ can be switched with direct limits over directed categories. Hence for all $i, j \geq 1$,

$$\begin{aligned} \varinjlim_{W \in \mathrm{Rep}(G)} \Big(\varprojlim^j_{\mathcal{R}_p(G)} \Pi_i^{(f \oplus f_W)} \Big) &\cong \varprojlim^j_{\mathcal{R}_p(G)} \Big(\varinjlim_{W \in \mathrm{Rep}(G)} \Pi_i^{(f \oplus f_W)} \Big) \\ &\cong \varprojlim^j_{\mathcal{R}_p(G)} \Big(\varinjlim_{W \in \mathrm{Rep}(G)} \Pi_i^{f_W} \Big) \cong \varprojlim^j_{\mathcal{R}_p(G)} \big(\Pi_i\big) = 0 \end{aligned}$$

by Proposition 6.

In other words, each successive obstruction to showing $f_p^\wedge \simeq g_p^\wedge$ vanishes after replacing f and g by $f \oplus f_W$ and $g \oplus f_W$ for a sufficiently large G-representation W. Also, by [JMO2, Proposition 4.11], the higher limits of any functor $\mathcal{R}_p(G) \to \widehat{\mathbb{Z}}_p\text{-}\mathtt{mod}$ vanish in degrees above some fixed $d(G,p)$ (depending only on G and p); and so there are only finitely many obstructions to constructing the homotopy. And this finishes the proof of (1), and hence the proof of point (a).

The proof of (b) is similar, but slightly more complicated at some points. One first shows that $R_{\mathcal{P}}(G)$ is the Grothendieck group of the monoid $\varprojlim \mathrm{Rep}(P)$ (where the limit is taken over all p-toral subgroups $P \subseteq G$, for all p). In other words, it suffices to show that (b) holds for elements $X = (V_P) \in R_{\mathcal{P}}(G)$, where the $V_P \in \mathrm{Rep}(P)$ are actual representations. Set $n = \dim(V_1)$ (where 1 denotes the trivial subgroup). Then $n = \dim(V_P)$ for all P: just note that $(V_P)|_1 \cong V_1$ since the V_P form an element in the inverse limit.

Fix a prime $p \big| |N(T)/T|$. Let $\Pi_i^{(X)} : \mathcal{R}_p(G) \to \widehat{\mathbb{Z}}_p\text{-}\mathtt{mod}$ be the functor

$$\Pi_i^{(X)}(G/P) = \pi_i\Big(\mathrm{Map}(EG/P, BU(n)_p^\wedge), f_{V_P}\Big).$$

As in the proof of (a), we show that

$$\varinjlim_{W \in \mathrm{Rep}(G)} \Big(\varprojlim_{\mathcal{R}_p(G)}{}^{j}\, \Pi_i^{(X \oplus f_W)} \Big) = 0,$$

with the help of Propositions 4 and 6 and commuting limits. Thus, each successive obstruction to constructing a map $f : BG \to BU(n)$ vanishes after replacing X by $X + r_G(\eta_W)$ for some sufficiently large representation W. Since there are only finitely many nonzero obstructions, we obtain a map $f_p : BG \to BU(n+k)$ such that $f_p|_{BP} \simeq (f_{V_P} \oplus f_W)_p^\wedge$ for some k-dimensional representation W. After stabilizing further, we can arrange that this has been done for all primes $p \big| |N(T)/T|$, and with the same G-representation W. And the pullback square in Proposition 3 can now be used to construct $f : BG \to BU(n+k)$ such that $r_G(\xi_f) = X + r_G(\eta_W)$. $\square$

References:

[Ad] J. F. Adams, Maps between classifying spaces II, Invent. math. 49 (1978), 1–65

[AS] M. Atiyah & G. Segal, Equivariant K-theory and completion, J. Diff. Geometry 3 (1969), 1–18

[DZ] W. Dwyer & A. Zabrodsky, Maps between classifying spaces, Algebraic topology, Barcelona, 1986, Lecture Notes in Math. 1298, Springer-Verlag (1987), 106–119

[Ja] S. Jackowski, Group homomorphisms inducing isomorphisms of cohomology, Topology 17 (1978), 303–307

[JMO] S. Jackowski, J. McClure, & B. Oliver, Homotopy classification of self-maps of BG via G-actions, Annals of Math. 135 (1992), 183–270

[JMO2] S. Jackowski, J. McClure, & B. Oliver, Homotopy theory of classifying spaces of compact Lie groups, Algebraic topology and its applications, M.S.R.I. Publ. 27, Springer-Verlag (1994), 81–123

[JMO3] S. Jackowski, J. McClure, & B. Oliver, Maps between classifying spaces revisited, Čech centennial conference on homotopy theory, Contemp. Math. 181, Amer. Math. Soc. (1995), 263–298

[JO] S. Jackowski & B. Oliver, Vector bundles over classifying spaces of compact Lie groups, Acta math. 176 (1996), 109–143

[Nb] D. Notbohm, Maps between classifying spaces, Math. Z. 207 (1991), 153–168

[Ol] B. Oliver, The representation ring of a compact Lie group revisited, Comment. Math. Helv. (to appear)

[Se] G. Segal, The representation ring of a compact Lie group, Publ. math. I. H. E. S. 34 (1968), 113–128

[Wo] Z. Wojtkowiak, On maps from holim F to Z, Algebraic topology, Barcelona, 1986, Lecture Notes in Math. 1298, Springer-Verlag (1987), 227–236

Bob Oliver
UMR 7539 du CNRS
Institut Galilée
Université Paris Nord
93430 Villetaneuse
France
bob@math.univ-paris13.fr

The Geometry of the Seiberg-Witten Invariants

Clifford Henry Taubes

My purpose in this talk is to describe a curious story about a search for symplectic forms on smooth, compact, 4-dimensional manifolds. However, be aware that at the time of this writing, the story that I relate below has no conclusion.

1 The start of the story

The story starts with the Seiberg-Witten invariants which were introduced just under four years ago by Witten [W1]. These are invariants of compact, smooth, oriented 4-manifolds. (Here, and below, all 4-manifolds will be connected and oriented.) After the choice of orientation for the real line $\det^+ = H^0 \otimes \det(H^1) \otimes \det(H^{2+})$, the Seiberg-Witten invariants constitute a map from the set, $\mathcal{S}$, of $\mathrm{Spin}^{\mathbb{C}}$ structures on the 4-manifold to the integers. There is also an extension of SW in the case where the Betti number b^1 is positive to a map SW: $\mathcal{S} \to \Lambda^* H^1(X;\mathbb{Z})$. (Here and below, $\Lambda^* H^1(X;\mathbb{Z}) = \mathbb{Z} \oplus H^1 \oplus \Lambda^2 H^1 \oplus \cdots \oplus \Lambda^{b_1} H^1$. Note that the projection of the image of SW on the summand $\mathbb{Z}$ reproduces the original map as defined from $\mathcal{S}$ to $\mathbb{Z}$ in [W1].) In either guise, this map, SW, is computed by an algebraic count of solutions to a certain non-linear system of differential equations on the manifold. The equation in question is for a pair of unknowns which consist of a section of a certain $\mathbb{C}^2$ bundle and a connection on this same bundle's determinant line.

The invariant SW and the Seiberg-Witten equations were introduced to the mathematical community by Witten [W1] after his ground breaking work with Seiberg in [SW1], [SW2]. See also [KM], [Mor], [KKM] and [T1]. Few would argue against the assertion that the Seiberg-Witten equations have revolutionized 4-manifold differential topology.

The Seiberg-Witten invariants have proved so useful for questions about compact 4-manifolds because they are at least as powerful as the Donaldson invariants (see, e.g. [DK]) and so much easier to compute. In this regard, note that Witten has conjectured that the two sets of invariants carry the same information about compact 4-manifolds. But, it remains to be seen whether they are equivalent in this context, let alone for their other uses. (An argument for Witten's conjecture is outlined in [PT] and a series of papers by Feehan and Leness begin to address the technical details. See, e.g. [FL].)

However, neither the Seiberg-Witten invariants relation with the Donaldson invariants nor their computability is the subject of this story. Rather, the story I am relating concerns another property of the Seiberg-Witten invariants which is the following: These invariants seem to have a direct, geometric interpretation as an

algebraic count of certain distinguished submanifolds with boundary in the given 4-manifold. Moreover, this geometric interpretation suggests a novel approach to the existence question for symplectic forms. Here, I have used the verbs 'seem' and 'suggests' because, as I said at the outset, the story is not finished. In particular, the geometric interpretation is not yet completely worked out except in some special cases. One of these cases consists of symplectic manifolds, and as symplectic notions are anyway central to this story, they form the subject of the next three chapters.

2 Symplectic manifolds

A 4-dimensional manifold X is symplectic when it carries a closed, non-degenerate 2-form. That is, there is a 2-form ω with $d\omega = 0$ and with $\omega \wedge \omega \neq 0$ at all points. In this regard, the convention will be to orient the manifold in question with $\omega \wedge \omega$. Now, not all 4-manifolds can be symplectic. First of all, the Betti number b^{2+}, which is the dimension of the maximum subspace of $H_2(X; \mathbb{Q})$ on which the intersection pairing is positive definite, must be positive since $\omega \wedge \omega$ is positive. Also, there is a classical, mod 2 obstruction which asserts that an oriented X has a symplectic form which reproduces the given orientation only if the Betti number sum $b^1 + b^{2+}$ is odd. For example, this condition rules out the connect sum of an even number of $\mathbb{CP}^2$.

The Seiberg-Witten invariants give additional obstructions [T2], [T3]. For example, there must be a $\mathrm{Spin}^{\mathbb{C}}$ structure for which the associated Seiberg-Witten invariant is ± 1. The latter rules out the connect sum of an odd number larger than 1 of $\mathbb{CP}^2$.

By the way, it was innocently conjectured that an irreducible, simply connected 4-manifold was always symplectic with some choice of orientation. However, Szabo proved this conjecture false [Sz] and subsequently, Fintushel and Stern (who are speaking in this Congress) found a slew of counter examples as homotopy K3 surfaces [FS]. In both cases, the Seiberg-Witten invariants play a prominent role.

Anyhow, it is important to realize that, at the time of this writing, necessary and sufficient conditions for the existence of a symplectic form are not known.

(Although not relevant for this story, the reader might be interested to know that the problem of classifying symplectic manifolds up to diffeomorphism is unsolved except in some special cases where $b^{2+} = 1$. However, Donaldson has made progress recently towards a classification theory for symplectic 4-manifolds up to deformation of the symplectic form and symplectomorphisms.)

3 The Seiberg-Witten invariants on a symplectic manifold

As remarked in [T1], a symplectic manifold has a natural orientation as does the line $\det^+$. Furthermore, there is a canonical identification of the set $\mathcal{S}$ with $H^2(X; \mathbb{Z})$. Thus, on a symplectic 4-manifold, SW can be viewed as a map from $H^2(X; \mathbb{Z})$ to $\mathbb{Z}$, or, more generally, from $H^2(X; \mathbb{Z})$ to $\Lambda^* H^1(X; \mathbb{Z})$.

Meanwhile, a compact symplectic 4-manifold has a second natural map sending $H^2(X; \mathbb{Z})$ to $\mathbb{Z}$, its Gromov invariant, Gr. The map Gr also extends on a

$b^1 > 0$ symplectic 4-manifold to a map from $H^2(X;\mathbb{Z})$ into $\Lambda^* H^1(X;\mathbb{Z})$; the extension is sometimes called the Gromov-Witten invariant, but it will be denoted here by Gr as well. In either guise, Gr, assigns to a class e a certain weighted count of compact, symplectic submanifolds whose fundamental class is Poincare dual to e. In this regard, a submanifold is symplectic when the restriction of the symplectic form to its tangent space is non-degenerate. (More is said about the count for Gr in the next chapter.)

The Gromov invariant was introduced initially by Gromov in [Gr] and then generalized by Witten [W2] and Ruan [Ru]. See also [T4]. (Note that Gr here does not count maps from a fixed complex curve. It differs in this fundamental sense from the Gromov-Witten invariant introduced in [W2].) The precise definition of Gr is provided in [T4]. Here is the main theorem which relates SW to Gr:

Theorem 1: *Let* X *be a compact, symplectic manifold with* $b^{2+} > 1$. *Use the symplectic structure to orient* X *and the line* $\det^+$*; and use the symplectic structure to define SW as a map from* $H^2(X;\mathbb{Z})$ *to* $\Lambda^* H^1(X;\mathbb{Z})$. *In addition, use the symplectic structure to define Gr:* $H^2(X;\mathbb{Z})$. $\Lambda^* H^1(X;\mathbb{Z})$. *Then SW = Gr.*

Thus, according to Theorem 1, on a symplectic manifold with $b^{2+} > 1$, the smooth invariants of Seiberg-Witten can be interpreted geometrically as a certain count of symplectic submanifolds.

Theorem 1 is proved in [T5]. The equivalence between the Gromov invariant and the original SW map into $\mathbb{Z}$ was announced by the author in [T1]. The proof of Theorem 1 can be divided into three main parts. The first part explains how a non-zero Seiberg-Witten invariant implies the existence of symplectic submanifolds. The second part explains how a symplectic submanifold can be used to construct a solution to a version of the Seiberg-Witten equations. The final part compares the counting procedures for the two invariants. The first and second parts of the proof can be found in [T6] and [T7], respectively and the final part (together with an overview of the whole strategy) is in [T5]. (Some of the early applications of Theorem 1 are also described in [Ko].)

A restricted version of Theorem 1 holds in the case when $b^{2+} = 1$. Here, a fundamental complication is that the Seiberg-Witten invariant depends on more than the differentiable structure. This is to say that there is a dependence on a so called choice of chamber. However, the symplectic form selects out a unique chamber, and with this understood, one has:

Theorem 2: *Let* X *be a compact, oriented 4-manifold with* $b^{2+} = 1$ *and with a symplectic form. Then the symplectic form canonically defines a chamber in which the equivalence SW = Gr holds for classes* $\mathrm{e} \in H^2(X;\mathbb{Z})$ *which obey* $\langle \mathrm{e}, \mathrm{s} \rangle \geq -1$ *when ever the two dimensional homology class* $\mathrm{s} \in H_2(\mathrm{X} : \mathbb{Z})$ *is represented by an embedded, symplectic sphere with self-intersection number -1.*

(Here, $\langle , \rangle$ denotes the pairing between cohomology and homology.) Theorem 2 is also proved in [T5].

By the way, when X is a $b^{2+} = 1$ symplectic manifold and e in $H^2(X;\mathbb{Z})$ is a class for which the conditions of Theorem 2 do not hold, the Seiberg-Witten invariant SW(e) still counts pseudo-holomorphic subspaces [LL]. However, the Gromov invariant as defined in [T4] is not the correct symplectic invariant for such e since the subspaces involved can have singularities. The symplectic invariant in this case

is given by McDuff in [Mc1] (An overview of Seiberg-Witten story on symplectic manifolds is also provided in [T8].)

4 Pseudo-holomorphic subvarieties

When calculating Gr, one should follow Gromov [Gr] and introduce an auxilliary structure on X which consists of an almost complex structure J for TX. By definition, the latter is an endomorphism $J : TX \to TX$ which obeys $J^2 = -1$. Such J's exist precisely when the Betti number sum $b^1 + b^{2+}$ is odd. Moreover, given the symplectic form ω, there exists such J which are *compatible* with ω in the sense that the bilinear form $\omega(\cdot, J(\cdot))$ on TX defines a Riemannian metric. (Moreover, Gromov showed that the space of ω-compatible J's is contractible.)

With an almost complex structure J chosen, certain dimension 2 submanifolds are distinguished, namely those for which J preserves their tangent space in TX. Such submanifolds are called *pseudo-holomorphic*. Note that if $C \subset$ X is a pseudo-holomorphic submanifold, then J orients TC and thus the homology class of C is canonically defined. Moreover, if J is ω-compatible, then ω restricts positively to C and so C is symplectic. Also, the homology class of C is never (rationally) zero when C is pseudo-holomorpic. (Note that the restriction to TC of J endows C with the structure of a complex curve, in which case the inclusion map from C to X is a pseudo-holomorphic map in the original sense defined by Gromov.)

The set of pseudo-holomorphic submanifolds form a geometrically distinguished subset of symplectic submanifolds. This subset is well behaved in as much as the deformation theory for a pseudo-holomorphic submanifold is highly constrained. Indeed, the latter is a Fredholm deformation problem. (Among other things, this last assertion implies that the space of pseudo-holomorphic submanifolds in a given homology class has the structure of a finite dimensional variety.) By the way, there is an important converse to the preceding, which is that every symplectic submanifold is pseudo-holomorphic for some ω-compatible J. (A good, general reference about pseudo-holomorphic geometry is the book by McDuff and Salamon [MS].)

With the pseudo-holomorphic submanifolds understood, the first point of this chapter is simply to remark that the invariant Gr 'counts' symplectic submanifolds in a given homology class by actually counting the pseudo-holomorphic representatives with certain weights. Except for tori with zero self-intersection, these weights are ± 1. The weights for the excepted tori are more involved. Note also that Gr counts disconnected submanifolds. In any event, see [T4] for the full story. By the way, one consequence of Theorems 1 and 2 is an existence theorem for pseudo-holomorphic curves in certain homology classes [T6].

The second point of this chapter is to offer, for use in the subsequent chapters, a reasonable definition of pseudo-holomorphic submanifolds and pseudo-holomorphic varieties inside a non-compact symplectic manifold. Consider:

DEFINITION 3: *Let* X *be a smooth, 4-manifold with symplectic form* ω. *A subset* $C \subset$ X *is a pseudo-holomorphic variety when the following conditions are met:*

- *C is closed.*
- *There is a set* $\Lambda \subset C$ *with at most countably many elements and no accumulation points in* X *such that* $C - \Lambda$ *is a submanifold of* X.
- *J maps* $TC|_{C-\Lambda}$ *to itself in* TX.
- $\int_c \omega < \infty$.

In previous articles, I have sometimes distinguished amongst those $C \subset X$ which satisfy the first three conditions above, but not the final condition. When C also satisfies the final condition, one can say that TC has '*finite energy*'.

Note that the singularities of a pseudo-holomorphic variety (the points of Λ) are essentially those of complex curves in $\mathbb{C}^2$. (See, e.g. [Mc2], [PW], [Ye], [Pa], [MS].)

5 When no symplectic form is handy

Suppose now that X is a compact, oriented 4-manifold which has no known symplectic form. Here is a suggestion for the next best thing: Put a Riemannian metric on X. Among other things, the latter defines a decomposition of the bundle of 2-forms into a direct sum of two three dimensional bundles, $\Lambda^+ \oplus \Lambda^-$. These are the bundles of self-dual and anti-self-dual 2-forms. (Fix an oriented orthonormal frame $\{e^i\}_{1\le i\le 4}$ for T^*X at a given point, and then $\Lambda^+ = \text{Span}\,\{e^1 \wedge e^2 + e^3 \wedge e^4, e^2 \wedge e^3 + e^1 \wedge e^4, e^3 \wedge e^1 + e^2 \wedge e^4\}$.) More to the point, when $b^{2+} \ge 1$, then Hodge-DeRham theory provides a self-dual, closed form ω. That is, ω is a section of Λ^+ and $d\omega = 0$. In particular, this implies that $\omega \wedge \omega = |\omega|^2$ dvol and so ω is symplectic where non-zero.

As this story is about the search for symplectic forms, the preceding suggests an investigation into the zero set of a closed self-dual form. For this purpose, let ω be such a form. In as much as ω is a section of an $\mathbb{R}^3$ bundle, one might expect the zero set to be 1-dimensional in some generic sense. This turns out to be the case. Both Honda [Ho] and LeBrun [Le] offer proofs of the following:

Theorem 4: *Fix a compact, oriented 4-manifold* X *with* $b^{2+} \ge 1$. *If the metric on* X *is suitably generic (chosen from a Baire subset of smooth metrics), then there is a closed, self-dual 2-form* ω *which vanishes transversally as a section of* Λ^+. *In particular, the zero set of w is a finite, disjoint union of embedded circles.*

With the preceding understood, assume from here on that each given closed, self-dual 2-form vanishes as a transversal section of Λ^+.

Work of Carl Luttinger [Lu] demonstrates that, by themselves, the zero circles of any given closed, self-dual ω carry very little in the way of information about the obstruction to finding a symplectic form. In fact, Luttinger shows that there are closed forms which are self-dual for some metric and have arbitrarily many zero set components. Conversely, Luttinger showed how to modify any given ω so that the result is closed and self-dual for some metric, yet has only one component of its zero set. (In both cases, Luttinger's arguments are essentially local in nature. One constructs explicit models in $\mathbb{R}^4$ of 1-parameter families of closed, self-dual

forms for which the topology of the zero set changes either by birth or death of an isolated circle, or by two components melding to one or one component splitting into two. One can then argue using appropriate coordinate charts that these local models can be 'spliced' into any manifold.)

Note however, that the zero circles of ω do carry one small bit of obstruction data. Indeed, Gompf [Go] has shown how data near the zero circle can be used to compute the parity of $b^1 + b^{2+}$. In this regard, it is important to realize that in some neighborhood of a point where ω is zero, there are coordinates $(t, \mathrm{x}) \in \mathbb{R} \times \mathbb{R}^3$ so that $\mathrm{x} = 0$ corresponds to $\omega^{-1}(0)$ and so that

$$\omega = dt \wedge (\mathrm{x}^T)A \cdot d\mathrm{x} + *_3(\mathrm{x}^T A \cdot d\mathrm{x}) + \mathcal{O}(|\mathrm{x}|^2). \tag{1}$$

Here, x^T denotes the transpose of the vector x, while $A = A(t)$ is a 3×3 symmetric (non-degenerate) matrix. Also, $*_3$ denotes the standard Hodge star operator on $\mathbb{R}^3$. Note that the condition $d\omega = 0$ requires that A be both symmetric and traceless.

By changing the sign of the t coordinate, one can then assume that $\det(A) < 0$. That is, A has one negative eigenvector at each t and two positive eigenvectors. In particular, as one moves around any given component of $\omega^{-1}(0)$, the negative eigenspaces of A fit together to yield a line bundle over the circle. The latter can be either oriented or not, and Gompf's observation is that the parity of $b^1 + b^{2+}$ is the opposite of the parity of the number of components of $\omega^{-1}(0)$ for which the aforementioned negative eigenbundle is oriented. (Note: There is no misprint here with the use of 'oriented'.)

6 Pseudo-holomorphic subvarieties in X - Z.

As just seen, $\omega^{-1}(0)$ carries by itself little information about the existence of symplectic forms on X. However, this is not to say that $\omega^{-1}(0)$ is completely irrelevant to the story. Indeed, at least some non-trivial data seems to be stored as configurations of certain kinds of symplectic surfaces in $\mathrm{X} - \omega^{-1}(0)$ which bound $\omega^{-1}(0)$. A digression is required to be more precise in this regard.

To start the digression, introduce as short hand $Z \equiv \omega^{-1}(0)$. By definition, ω restricts to X - Z as a symplectic form. Moreover, if $g : TX \to T^*\mathrm{X}$ denotes the given metric, then the endomorphism $J = \sqrt{2}g^{-1}\omega/|\omega|$ defines an ω-compatible almost complex structure for X - Z. Note that the latter is singular along Z. Indeed, when ω vanishes transversely, then the first Chern class of the associated canonical bundle has degree 2 on all linking 2-spheres of Z. Even so, one can use J to define pseudo-holomorphic subvarieties in X - Z. The pseudo-holomorphic subvarieties in X - Z might be curiosities were it not for the following theorem [T9]:

THEOREM 5: *Suppose that* X *is a compact, oriented, Riemannian manifold with* $b^{2+} \geq 1$ *and a non-zero Seiberg-Witten invariant. Let* ω *be a closed, self-dual 2-form whose zero set, Z, is cut out transversally by* ω*. Then, there exists a pseudo-holomorphic subvariety in* $\mathrm{X} - Z$ *which homologically bounds Z in the sense that it has intersection number 1 with every linking 2-sphere of Z.*

In the preceding, when $b^{2+} = 1$, the Seiberg-Witten invariants in the statement of the theorem are from a certain chamber which is specified by ω. By the way, the statement of the previous theorem can be strengthened in the following direction: Given a specific $\mathrm{Spin}^{\mathbb{C}}$ structure on X with non-zero Seiberg-Witten invariant, there exists a pseudo-holomorphic subvariety in X - Z with homological boundary Z whose relative homology class in X - Z can be foretold in terms of the given $\mathrm{Spin}^{\mathbb{C}}$ structure. Note that Theorem 5 suggests the following likely conjecture:

- The Seiberg-Witten invariants of X can be computed via a specific algebraic count of the pseudo-holomorphic subvarieties in X - Z which bound Z.

Theorem 1 affirms this conjecture in the case where $Z = \emptyset$. Moreover, through work of Hutchings and Lee [HL] and Turaev [Tu], this conjecture has been confirmed also in the case where $Z = S^1 \times M$ where M is a compact, oriented 3-manifold with $b^1 > 0$. (This last case is discussed further in a subsequent chapter.)

Remark that the pseudo-holomorphic subvarieties which arise in the context of Theorem 4 have a well defined Fredholm deformation theory; and this last fact supplies further evidence for the validity of the preceding conjecture.

7 A regularity theorem

Since the almost complex structure J in Theorem 5 is singular along Z, the behavior near Z of a pseudo-holomorphic variety in X - Z is problematic. (As remarked previously, away from Z, such a variety is no more singular than a complex subvariety of $\mathbb{C}^2$.) Indeed, as ω near Z vanishes as the distance to Z, it is not apriori clear that such varieties have finite area. However, it turns out that the fourth condition in Definition 3 is stronger than it looks, and in particular, some (partial) regularity results can be proved, at least under some restrictive hypothesis about the form of ω near its zero set. The results are summarized in Theorem 6, below. However, first comes a digression to explain the restrictions. The restriction on ω is as follows: Near each point in Z, there should exist coordinates (t, x, y, z) such that Z coincides with the set x = y = z = 0 and such that in this coordinate patch,

$$\begin{aligned} &\omega = dt \wedge (\mathrm{x}d\mathrm{x} + ydy - 2zdz) + \mathrm{x}dy \wedge dz + ydz \wedge d\mathrm{x} - 2zd\mathrm{x} \wedge dy, \\ &g = dt^2 + d\mathrm{x}^2 + dy^2 + dz^2. \end{aligned} \tag{2}$$

Note that this is a rather special version of the general form for ω which is given by Eq.(1). However, on any $b^{2+} > 0$ manifold, there are metrics and self-dual forms which satisfy these restrictions. In fact, given any metric and closed self-dual form ω with non-degenerate zeros, both can be modified solely in a given neighborhood of $\omega^{-1}(0)$ so that the resulting form is closed and self-dual for the resulting metric and has the same zero set as the original and obeys the restrictions in Eq.(2). The following summarizes what is presently known about regularity near Z:

THEOREM 6: *Let* X *be a smooth, compact, oriented, Riemannian 4-manifold and let* ω *be a closed, self-dual form which is described near Z by Eq.(2). Let* $C \subset$ X *be a pseudo-holomorphic subvariety. Then, C has finite area. Moreover, except for possibly a finite subset of points on Z, every point on Z has a ball neighborhood which intersects C in a finite number of components. And, the closure of each component in such a ball neighborhood is a real analytically embedded half disk whose straight edge coincides with Z.*

(Note that the behavior of C near each of the singular points can also be described.) Theorem 6 is proved in [T10]. I expect that a very similar theorem holds without the special restriction in Eq.(2). Moreover, I expect that the finite number of singular points are 'removable by perturbation' in the sense that these singularities are, in some well defined sense, codimension one phenomena. (Hofer, Wysocki and Zehnder have an alternative approach to studying X - Z as a symplectic manifold. See, e.g. [HWZ].)

8 An illustrative example

An example with much food for thought has $X = S^1 \times M$ where M is a compact, oriented, 3-manifold with $b^1 > 0$. The set of M for which the corresponding X is symplectic remains (as of this writing) mysterious. However, it is known that X is symplectic when M admits a fibering $f : M \to S^1$, and for all we know at present, these are the only 3-manifolds for which $S^1 \times M$ is symplectic. (The latest results on this question are due to Kronheimer [Kr].) In any event, $X = S^1 \times M$ does have closed, self-dual 2-forms. For example, if one uses a product metric on X (the Euclidean metric on $S^1 = \mathbb{R}/\mathbb{Z}$ plus a metric on M), then all closed, self-dual 2-forms have the form $\omega = dt \wedge \nu + *_3\nu$, where ν is a harmonic 1-form on M and where $*_3$ is the Hodge star operator on $\Lambda^* M$. In particular, one can find harmonic 1-forms which equal df where $f : M \to \mathbb{R}/\mathbb{Z}$ is a non-zero cohomology class. This last case is instructive in as much as one can see that $Z = \omega^{-1}(0)$ is given by $Z = S^1 \times \mathrm{Crit}(f)$, where $\mathrm{Crit}(f)$ is the set of critical points of f. Moreover, for a suitably generic choice of metric on M, the $\mathbb{R}/\mathbb{Z}$-valued function f will have only non-degenerate critical points (see, e.g. [Ho]), and in this case, the corresponding ω will have a transversal zero set in the sense of Theorem 4.

In this last example, subsets of X given by $S^1 \times$ (gradient flow lines of ∇f) are pseudo-holomorphic submanifolds. In fact, when the metric on M is suitably generic, then the pseudo-holomorphic submanifolds promised by Theorem 5 have the form $S^1 \times \Gamma$ where Γ is a finite union of gradient flow lines of ∇f having the following properties: First, each flow line $\gamma \in \Gamma$ is complete and has bounded length. Second, each critical point of f is an end point of one and only one flow line in Γ. (By the way, Hutchings and Lee [HL] have found an intrinsic count of such Γ which computes a certain Alexander polynomial of the associated $\mathbb{Z}$-cover of M. Meanwhile Meng and the author [MT] have a theorem to the effect that such Alexander polynomials essentially determine the Seiberg-Witten invariants of X. See also [Tu].) One lesson from the preceding example is the following: In the S^1 invariant context on $X = S^1 \times M$, self-dual, symplectic geometry on X is nothing more than Morse theory with $\mathbb{R}/\mathbb{Z}$ valued functions on M. This is to

say that the problem of eliminating component circles of the zero set of an S^1 invariant, closed, self-dual 2-form on X is that of eliminating the critical points of a harmonic function on M.

9 A dictionary?

The previous example suggests that there may exist a dictionary which translates Morse theoretic notions in 3-manifold topology to notions which involve closed, self-dual 2-forms on 4-manifolds and their associated pseudo-holomorphic varieties. (Below, I call the second subject 'self-dual symplectic geometry'.) Some of the dictionary has already been established, and some is conjectural. This dictionary is reproduced below. In the dictionary, M is a 3-dimensional Riemannian manifold with $b^1 > 0$ and X is a 4-dimensional Riemannian manifold with $b^{2+} > 0$.

Morse Theory on M	Self-dual symplectic geometry on X
Critical points of an $\mathbb{R}/\mathbb{Z}$-valued harmonic function f	The zero set Z of the closed, self-dual 2- form ω.
Gradient flow lines of ∇f	Pseudo-holomorphic varieties in X - Z with boundary on Z.
Milnor torsion/Alexander polynomial	Seiberg-Witten invariants.
Whitney disk	Lagrangian disk with a boundary piece on a pseudo-holomorphic subvariety.
Self-indexing Morse function	?
Handle sliding	?
Morse-Smale cancellation lemma	?

Here are some comments about the preceding table:

- The appearance of Lagrangian disks as the analog to Whitney disks in 3-dimensional Morse theory is closely related to observations of Donaldson about the appearance of Lagrangian 2-spheres in his study of symplectic Lefschetz pencils. In any event, the point here is that a symplectic submanifold can be symplectically deformed via 'finger moves' along Lagrangian disks which have a part of their boundary on the submanifold in question.

- The Morse-Smale cancellation lemma asserts that a pair of critical points of f can be cancelled (without introducing new critical points or disturbing the configuration of gradient flow lines) if there is a unique, stable minimal energy gradient flow line between them. (The energy of a flow line is simply the drop in f between the start and the finish. A flow line is stable if it persists under perturbation of the gradient flow or the function f.) On the 4-dimensional side, the analogous lemma might be something like the following: Let Z_0 be

a component of Z. If the smallest energy, pseudo-holomorphic variety in X - Z with Z_0 as a boundary component is suitably stable, is unique and is a disk, then the form ω can be altered to produce a new closed form which is self-dual for some metric on X, has zero set with fewer components, and has a less complicated set of bounding, pseudo-holomorphic varieties. Note that the local 'melding' procedure of Luttinger which joins all components of Z into one circle appears to increase the genus of the bounding pseudo-holomorphic varieties unless suitable Lagrangian disks are present.

- In fact, there is a self-dual symplectic analog of handle sliding, but the details are still somewhat obscure (to the author, anyway).

10 SUMMARY

The following two as yet unanswered questions aim at the heart of the matter:

- How much is self-dual symplectic geometry like 3-dimensional Morse theory?
- More to the point, can self-dual symplectic geometry shed any light on 4-manifold differential topology?

REFERENCES

[DK] S. K. Donaldson and P. B. Kronheimer, The Geometry of 4-Manifolds, Oxford University Press, 1990.

[FL] P. Feehan and T. Leness, *PU(2) monopoles, II: Highest level singularities and relations between four-manifold invariants*, preprint 1997.

[FS] R. Fintushel and R. J. Stern, *Knots, links and 4-manifolds*, Inventiones,to appear.

[Go] R. Gompf, private communication.

[Gr] M. Gromov, *Pseudo-holomorphic curves in symplectic manifolds*, Invent. Math.. 82 (1985) 307-347.

[HWZ] H. Hofer, K. Wysocki and E. Zehnder, *Properties of pseudo-holomorphic curves in symplectisations, IV: Asymptotics with degeneracies*, in Contact and Symplectic Geometry, C. Thomas, ed., Cambridge University Press, 1996.

[Ho] K. Honda, *Harmonic forms for generic metrics*, preprint.

[HL] M. Hutchings and Y-J. Lee, *Circle valued Morse theory, R-torsion and Seiberg-Witten invariants of 3-manifolds*, Topology, to appear.

[KKM] D. Kotschick, P. B. Kronheimer and T. S. Mrowka, in preparation.

[Ko] Kotschick, D. *The Seiberg-Witten invariants of symplectic 4-manifolds (after C. H. Taubes)*, Seminaire Bourbaki. 48eme annee (1995-96), no. 812.

[Kr] P. B. Kronheimer, *Minimal genus in* $S^1 \times M^3$, Inventiones, to appear.

[KM] P. B. Kronheimer and T. S. Mrowka, *The genus of embedded surfaces in the projective plane*, Math. Res. Letters 1 (1994) 797-808.

[Le] C. LeBrun, *Yamabe constants and the perturbed Seiberg-Witten equations*, Commun. Anal. and Geom. 5 (1997) 535-553.

[LL] T. J. Li and A. K. Liu, *Family Seiberg-Witten invariants*, preprint.

[Lu] C. Luttinger, unpublished.

[Mc1] D. McDuff, *Lectures on Gromov invariants for symplectic 4-manifolds*, in Gauge Theory and Symplectic Geometry, J. Hurtubise, F. Lalonde and G. Sabidossi, ed., Kluwer Academic Publishers, 1997.

[Mc2] D. McDuff, *Singularities and positivity of intersections of J-holomorphic curves, with Appendix by Gang Liu*, in Proceedings of CIMPA Summer School of Symplectic Topology, Nice 1992, Birkhauser, 1994.

[MS] D. McDuff and D. Salamon, J-Holomorphic Curves and Quantum Cohomology, American Mathematical Society, Providence 1996.

[MT] G. Meng and C. H. Taubes, *SW = Milnor torsion*, Math. Res. Letters 3 (1996) 661-674.

[Mor] J. W. Morgan, The Seiberg-Witten Equations and Applications to the Topology of Smooth Four-Manifolds, Mathematical Notes 44, Princeton University Press, 1996.

[Pa] P. Pansu, *Pseudo-holomorphic curves in symplectic manifolds*, in Holomorphic Curves in Symplectic Geometry, M. Audin and F. Lafontaine, ed. Progress in Math 117, Birkhauser (1994) 233-250.

[PW] T. H. Parker & J. G. Wolfson, *Pseudo-holomorphic maps and bubble trees*, Journ. Geom. Analysis 3 (1993) 63-98.

[PT] V. Y. Pidstrigach and A. N. Tyurin, *Localization of Donaldson invariants along the Seiberg-Witten classes*, preprint.

[Ru] Y. Ruan, *Symplectic topology and complex surfaces*, in Geometry and Topology on Complex Manifolds, T. Mabuchi, J. Noguchi and T. Ochial, eds., World Scientific Publications, Singapore 1994.

[SW1] N. Seiberg and E. Witten, *Electro-magnetic duality, monopole condensation and confinement in N=2 supersymmetric Yang-Mills theory*, Nucl. Phys. B426 (1994) 19-52.

[SW2] N. Seiberg and E. Witten, *Monopoles, duality and chiral symmetry breaking in N=2 supersymmetric Yang-Mills theory*, Nucl. Phys. B426 (1994) 581-640.

[Sz] Z. Szabo, *Simply connected, irreducible 4-manifolds with no symplectic structures*, Inventiones, to appear.

[T1] C. H. Taubes, *The Seiberg-Witten and the Gromov invariants*, Math. Res. Letters 2 (1995) 221-238.

[T2] C. H. Taubes, *The Seiberg-Witten invariants and symplectic forms*, Math. Res. Letters 1 (1994) 809-822.

[T3] C. H. Taubes, *More constraints on symplectic forms from the Seiberg-Witten invariants*, Math. Res. Letters 2 (1995) 9-13.

[T4] C. H. Taubes, *Counting pseudo-holomorphic submanifolds in dimension 4*, Journ. Diff. Geom., 44 (1996) 818-893; and reprinted in Proceedings of the first IP Lecture Series, Vol. II, R. Wentworth ed., International Press, to appear.

[T5] C. H. Taubes, *Gr = SW. Counting curves and connections*, Jour. Diff. Geom., to appear; and reprinted in Proceedings of the first IP Lecture Series, Vol. II, R. Wentworth ed., International Press, to appear.

[T6] C. H. Taubes, *SW ⇒ Gr: From the Seiberg-Witten equations to pseudo-holomorphic curves*, Journ. Amer. Math. Soc., 9 (1996) 845-918; and reprinted with errata in Proceedings of the First IP Lecture Series, Vol. II, R. Wentworth ed., International Press, to appear.

[T7] C. H. Taubes; *Gr ⇒ SW. From pseudo-holomorphic curves to Seiberg-Witten solutions*, Jour. Diff. Geom., to appear; and reprinted in Proceedings of the first IP Lecture Series, Vol. II, R. Wentworth ed., International Press, to appear.

[T8] C. H. Taubes, *The geometry of the Seiberg-Witten invariants*, in Surveys in Diff. Geom., to appear.

[T9] C. H. Taubes, *Seiberg-Witten invariants and pseudo-holomorphic subvarieties for self-dual, harmonic 2-forms*, preprint.

[T10] C. H. Taubes, *The structure of pseudo-holomorphic subvarieties for a degenerate almost complex structure and symplectic form on* $S^1 \times B^3$, preprint.

[Tu] V. Turaev, *A combinatorial formulation for the Seiberg-Witten invariants of 3-manifolds*, preprint.

[W1] E. Witten, *Monopoles and 4-manifolds*, Math. Res. Letters 1 (1994) 769-796.

[W2] E. Witten, *Two dimensional gravity and intersection theory on moduli space*, Surveys in Diff. Geom. 1 (1991) 243-310.

[Ye] R. Ye, *Gromov's compactness theorem for pseudo-holomorphic curves*, Trans. Amer. Math. Soc. 342 (1994) 671-694.

Clifford Henry Taubes
Department of Mathematics
Harvard University
Cambridge, MA 02138
USA

Section 7

Lie Groups and Lie Algebras

In case of several authors, Invited Speakers are marked with a *.

Towards a Stable Trace Formula

James Arthur*

Abstract. The paper is a report on the problem of stabilizing the trace formula. The goal is the construction and analysis of a stable trace formula that can be used to compare automorphic representations on different groups.

1991 Mathematics Subject Classification: Primary 22E55, 11R39.

1. It is an important problem to place the automorphic representation theory of classical groups on an equal footing with that of $GL(n)$. Thirty years after the study of $GL(2)$ by Jacquet-Langlands [12], the theory for $GL(n)$ is now in pretty good shape. It includes an understanding of the relevant L-functions [13], a classification of the discrete spectrum [21] and cyclic base change [10]. One would like to establish similar things for orthogonal, symplectic and unitary groups. A satisfactory solution would have many applications to number theory, the extent of which is hard to even guess at present.

A general strategy has been known for some time. One would like to compare trace formulas for classical groups with a twisted trace formula for $GL(n)$. There is now a trace formula that applies to any group [4]. However, it contains terms that are complicated, and are hard to compare with similar terms for other groups. The general comparison problem has first to be formulated more precisely, as that of stabilizing the trace formula [18]. In this form, the problem is to construct a stable trace formula, a refined trace formula whose individual terms are stable distributions. It includes also the further analysis required to establish identities between terms in the original trace formula and their stable counterparts on other groups. This would allow a cancellation of all the geometric and residual terms from the relevant trace formulas, leaving only terms that describe automorphic spectra. The resulting identity given by these remaining terms would lead to reciprocity laws for automorphic spectra on different groups. In the case of classical groups, such identities would provide the means for attacking the original classification problem.

The purpose of this report is to discuss the construction and deeper analysis of a stable trace formula. I can say nothing about the fundamental lemma (or

* Supported by NSERC Grant A3483.

its analogue for weighted orbital integrals), which is one of the key problems to be solved. The reader can consult [11] and [25] for some special cases that have been resolved. Furthermore, I shall stick to the ordinary trace formula, since the twisted trace formula presents extra difficulties [16]. With these caveats, I believe that the general problem has been essentially solved. Since there are still a number of things to be written out, I shall organize the report conservatively as a series of stabilization problems for the various constituents of the trace formula. The solutions, all being well, will appear in the papers [8] and [9].

2. Let G be a connected, reductive algebraic group over a number field F. To simplify the discussion, we shall actually assume that G is semisimple and simply connected. If V is a finite set of valuations of F, $\mathcal{H}\big(G(F_V)\big)$ will denote the Hecke algebra of functions on $G(F_V)$, the product over $v \in V$ of the groups $G(F_v)$. We shall usually take V to be a large finite set outside of which G is unramified. A function in $\mathcal{H}\big(G(F_V)\big)$ can then be identified with the function on the adèle group $G(\mathbb{A})$ obtained by taking its product with the characteristic function of a maximal compact subgroup K^V of $G(\mathbb{A}^V)$. The trace formula is to be regarded as two different expansions of a certain linear form I on $\mathcal{H}\big(G(F_V)\big)$.

The first expansion

$$I(f) = \sum_{M \in \mathcal{L}} |W_0^M||W_0^G|^{-1} \sum_{\gamma \in \Gamma(M,V)} a^M(\gamma) I_M(\gamma, f) \tag{1}$$

is in terms of geometric data. As usual, $\mathcal{L} = \mathcal{L}^G$ denotes the set of Levi subgroups (over F) that contain a fixed minimal one, and W_0^G is the restricted Weyl group of G. For any $M \in \mathcal{L}$, $\Gamma(M, V)$ is a set of conjugacy classes in $M(F_V)$. The coefficient $a^M(\gamma)$ depends only on M, and is really a global object. It is constructed from rational conjugacy classes in $M(F)$ that project onto γ, and are integral outside of V. The linear form $I_M(\gamma, f)$ on the other hand is a local object. It is an invariant distribution constructed from the weighted orbital integral of f over the induced conjugacy class of γ in $G(F_V)$.

The second expansion

$$I(f) = \sum_{M \in \mathcal{L}} |W_0^M||W_0^G|^{-1} \int_{\Pi(M,V)} a^M(\pi) I_M(\pi, f) d\pi \tag{2}$$

is in terms of spectral data, and is entirely parallel to the first one. For any $M \in \mathcal{L}$, $\Pi(M, V)$ is a certain set of equivalence classes of irreducible unitary representations of $M(F_V)$, equipped with a natural measure $d\pi$. The coefficient $a^M(\pi)$ is again a global object that depends only on M. It is constructed from automorphic representations of $M(\mathbb{A})$ that project onto π, and are integral outside of V. Similarly, the linear form $I_M(\pi, f)$ is a local object. It is an invariant distribution obtained from residues of weighted characters of f at unramified twists π_λ of π. The integral over $\Pi(M, V)$ is actually only known to be conditionally convergent. However, this is sufficient for present purposes, and in any case, could probably be strengthened with the results of Müller [22].

The trace formula is thus the identity obtained by equating the right hand sides of (1) and (2). It is perhaps difficult for a general reader to get a feeling for the situation, since we have not defined the various terms precisely. We would simply like to stress the general structure of the two expansions, and to note that it is the term with $M = G$ in the second expansion that contains the basic information on the automorphic discrete spectrum. For example, if G is anisotropic, this term is just the trace of the right convolution of f on $L^2\big(G(F)\backslash G(\mathbb{A})/K^V\big)$. The term is more complicated for general G, but it includes a discrete part

$$I_{\rm disc}(f) \;=\; \sum_{\pi\in\Pi_{\rm disc}(G,V)} a^G_{\rm disc}(\pi) f_G(\pi)\,, \tag{3}$$

that comes from the discrete spectrum of $L^2\big(G(F)\backslash G(\mathbb{A})/K^V\big)$ as well as induced discrete spectra of proper Levi subgroups [4, (4.3), (4.4)]. The ultimate goal for the trace formula is to deduce information about the multiplicities $a^G_{\rm disc}(\pi)$. In particular, the other terms — those with $M \neq G$ in the spectral expansion and those with any M in the geometric expansion — are to be regarded as objects one would analyze in some fashion to gain information about the discrete part $I_{\rm disc}(f)$ of the first term.

We have actually reformulated somewhat the trace formula from [4]. The invariant distributions $I_M(\gamma, f)$ and $I_M(\pi, f)$ here are defined in terms of the weighted characters of [6], and are independent of the choice of normalizing factors for intertwining operators implicit in [4]. On the geometric side, this modification has the effect of including values of weighted orbital integrals of the characteristic function of $K^V \cap M(\mathbb{A}^V)$ in the global coefficients of [2, (8.1)]. On the spectral side, the effect is to replace the complete automorphic L-functions in the global coefficients of [4, §4] with partial, unramified L-functions.

3. It is hard to extract arithmetic information from the trace formula for G by studying it in isolation. One should try instead to compare it with trace formulas for certain other groups. The groups in question are the *endoscopic* groups for G, a family of quasisplit groups over F attached to G that includes the quasisplit inner form of G. One actually has to work with endoscopic data, which are endoscopic groups with extra structure [18], [19]. We write $\mathcal{E}_{\rm ell}(G, V)$ for the set of isomorphism classes of elliptic endoscopic data for G over F that are unramified outside of V.

Suppose that $G' \in \mathcal{E}_{\rm ell}(G, V)$ and that v belongs to V. In [19], Langlands and Shelstad define a map from functions $f_v \in \mathcal{H}\big(G(F_v)\big)$ to functions $f'_v = f_v^{G'}$ on the strongly G-regular stable conjugacy classes δ'_v of $G'(F_v)$. We recall that stable conjugacy is the equivalence relation on the strongly regular elements in $G'(F_v)$ defined by conjugacy over an algebraic closure of F_v. The map is defined by

$$f'_v(\delta'_v) = \sum_{\gamma v} \Delta_G(\delta'_v, \gamma_v) f_{v,G}(\gamma_v),$$

where γ_v ranges over the ordinary conjugacy classes in $G(F_v)$, $\Delta_G(\delta'_v, \gamma_v)$ is the transfer factor defined in [19], and $f_{v,G}(\gamma_v) = J_G(\gamma_v, f_v)$ is the invariant orbital integral of f_v over the conjugacy class γ_v.

The Langlands-Shelstad transfer conjecture asserts that for any f_v, there is a function $h_v \in \mathcal{H}\big(G'(F_v)\big)$, not necessarily unique, whose stable orbital integral at any δ'_v equals $f'_v(\delta'_v)$. The fundamental lemma is a supplementary conjecture. It asserts that if G and G' are unramified at v, and f_v is the characteristic function of a hyperspecial maximal compact subgroup of $G(F_v)$, then h_v can be chosen to be the characteristic function of a hyperspecial maximal compact subgroup of $G'(F_v)$. Waldspurger [26] has shown, roughly speaking, that the fundamental lemma implies the transfer conjecture. We shall assume from now on that they both hold. A linear form $S' = S^{G'}$ on $\mathcal{H}\big(G'(F_V)\big)$ is said to be *stable* if its value at any $h \in \mathcal{H}\big(G'(F_V)\big)$ depends only on the stable orbital integrals of h. If this is so, there is a linear form $\widehat{S}'$ on the space of stable orbital integrals of functions in $\mathcal{H}\big(G'(F_V)\big)$ such that $S'(h)$ equals $\widehat{S}'(h')$. In particular, we obtain a linear form $f \to \widehat{S}'(f')$ in $f \in \mathcal{H}\big(G(F_V)\big)$.

We can now begin to describe the basic problem. The ultimate goal would be to stabilize the distribution $I_{\rm disc}$ in (3).

PROBLEM 1: *Construct a stable linear form $S^G_{\rm disc}$ on $\mathcal{H}\big(G(F_V)\big)$, for G quasisplit over F, such that for any G at all, $I_{\rm disc}(f)$ equals the endoscopic expression*

$$I^{\mathcal{E}}_{\rm disc}(f) = \sum_{G' \in \mathcal{E}_{\rm ell}(G,V)} \iota(G,G')\widehat{S}'_{\rm disc}(f'), \qquad f \in \mathcal{H}\big(G(F_V)\big).$$

Here $\iota(G,G')$ is a coefficient, introduced by Langlands [18], that can be defined by the formula of [14, Theorem 8.3.1].

The problem has a general structure that is common to many stabilization questions. If we take G to be quasisplit, the required formula amounts to an inductive definition of $S^G_{\rm disc}$. Since G belongs to $\mathcal{E}_{\rm ell}(G,V)$ in this case, and is the endoscopic group of greatest dimension, we can assume inductively that the linear form $S' = S^{G'}$ is defined and stable for any $G' \in \mathcal{E}_{\rm ell}(G,V)$ not equal to G. We can therefore set

$$S^G_{\rm disc}(f) = I_{\rm disc}(f) - \sum_{G' \neq G} \iota(G,G')\widehat{S}'_{\rm disc}(f').$$

The problem then has two parts. If G is quasisplit, one has to show that $S^G_{\rm disc}$ is stable. This is needed to complete the inductive definition. If G is not quasisplit, the summands in the expression $I^{\mathcal{E}}_{\rm disc}(f)$ are all defined inductively in terms of groups G' distinct from G. In this case, it is the identity itself that has to be proved.

The problem has been solved completely only for $SL(2)$ and $U(3)$ (and related groups) [17], [23]. A general solution of Problem 1 would be a milestone. It would relate fundamental global data on different groups by means of a transfer map $f \to f'$ defined in purely local terms. The resulting information would be particularly powerful if it could be combined with a property of strong multiplicity one, either for individual representations, or for packets of representations. For example, a twisted form of the identity in Problem 1 would relate many classical

groups G to $GL(n)$. Together with the identity for G itself, this would provide a powerful tool for dealing with the classification problem discussed earlier.

However, Problem 1 is unlikely to be solved directly. The strategy should be to consider similar problems for the various other terms in the trace formula. Towards this end, we first pose a parallel problem for the entire geometric expansion.

PROBLEM 2: *Construct a stable linear form S^G on $\mathcal{H}\big(G(F_V)\big)$, for G quasisplit over F, such that for any G, $I(f)$ equals the endoscopic expression*

$$I^{\mathcal{E}}(f) \;=\; \sum_{G' \in \mathcal{E}_{\rm ell}(G,V)} \iota(G,G')\widehat{S}'(f'), \qquad\qquad f \in \mathcal{H}\big(G(F_V)\big).$$

4. To deal with Problem 2, we would have to set up a series of stabilization problems for the various terms in the geometric expansion (1). Such problems have been solved for some terms in [18] and [15].

If v is a place of F, we write $\Gamma(G_v)$ for the set of conjugacy classes in $G(F_v)$. Assuming that each such class has been equipped with an invariant measure, we identify $\Gamma(G_v)$ with a set of invariant distributions on $G(F_v)$. In the case of archimedean v, examples of Assem [1, §1.10] suggest that elements in $\Gamma(G_v)$ do not always behave well under endoscopic transfer. We are forced to consider a larger family of distributions. Let us define $\mathcal{D}(G_v)$ to be the space spanned by invariant distributions on $G(F_v)$ of the form

$$\int_{G_c(F)\backslash G(F)} I_c(f_v^x)dx, \qquad\qquad f_v \in C_c^\infty\big(G(F_v)\big),$$

where c is a semisimple element in $G(F_v)$, G_c is the centralizer of c in G, I_c is an invariant distribution on $G_c(F_v)$ that is supported on the unipotent set, and $f_v^x(y) = f_v(x^{-1}cyx)$, for $y \in G_c(F_v)$. We then let $\Gamma_+(G_v)$ be a fixed basis of $\mathcal{D}(G_v)$ that contains $\Gamma(G_v)$. If v is p-adic, $\Gamma(G_v)$ actually equals $\Gamma_+(G_v)$, but $\Gamma(G_v)$ is a proper subset of $\Gamma_+(G_v)$ if G_v is archimedean. We also fix a basis $\Sigma_+(G_v')$ of the stable distributions in $\mathcal{D}(G_v')$, for each endoscopic datum G_v' of G over F_v. Among various compatability conditions, we assume that $\Sigma_+(G_v')$ contains the set of stable strongly G-regular orbital integrals on $G'(F_v)$. Extending the earlier notation, we write $f_v'(\delta_v')$ for the pairing obtained from elements $f_v \in \mathcal{H}(G_v)$ and $\delta_v' \in \Sigma_+(G_v')$. Then we can write $f_v'(\delta_v')$ as a finite linear combination of distributions $f_{v,G}(\gamma_v)$ in $\Gamma_+(G_v)$, with coefficients $\Delta(\delta_v', \gamma_v)$ that reduce to the Langlands-Shelstad transfer factors in the special case that δ_v' is strongly G-regular.

If V is a finite set of valuations as before, we define $\Gamma(G_V)$, $\Gamma_+(G_V)$ etc., by the appropriate products. Thus, if $M_V' = \prod M_v'$ is a product of local endoscopic data for a Levi subgroup M of G, and $\delta' = \prod \delta_v'$ belongs to $\Sigma_+(M_V')$, $\Delta_M(\delta', \gamma)$ equals the product over $v \in V$ of the factors $\Delta_M(\delta_v', \gamma_v)$, for each $\gamma = \prod \gamma_v$ in $\Gamma_+(M_V)$. We shall take M_V' to be the image of a global endoscopic datum $M' \in \mathcal{E}_{\rm ell}(M, V)$ in what follows.

Consider first the local terms $I_M(\gamma, f)$ in the geometric expansion. They are defined at this point only for $\gamma \in \Gamma(M_V)$. However, we shall assume that we can

construct $I_M(\gamma, f)$ for any γ in the larger set $\Gamma_+(M_V)$, by some variant of the techniques in [3, §3-5].

PROBLEM 3: *Construct stable linear forms $S_M^G(\delta)$ on $\mathcal{H}\big(G(F_V)\big)$, for G quasisplit over F and $\delta \in \Sigma_+(M_V)$, such that for any G, M, M' and δ', the linear form*

$$I_M(\delta', f) = \sum_{\gamma \in \Gamma_+(M_V)} \Delta_M(\delta', \gamma) I_M(\gamma, f)$$

equals the endoscopic expression

$$I_M^{\mathcal{E}}(\delta', f) = \sum_{G' \in \mathcal{E}_{M'}(G)} \iota_{M'}(G, G') \widehat{S}_{M'}^{G'}(\delta', f').$$

Here, $\mathcal{E}_{M'}(G)$ is a set of global endoscopic data for G and $\iota_{M'}(G, G')$ is a simple coefficient, both defined as in [7, §3].

Consider now the global coefficients $a^M(\gamma)$. We define a^M on the larger set $\Gamma_+(M_V)$ by setting $a^M(\gamma) = 0$ for any γ in the complement of $\Gamma(M, V)$ in $\Gamma_+(M_V)$.

PROBLEM 4: *Construct coefficients $b^M(\delta)$, for M quasisplit over F and $\delta \in \Sigma_+(M_V)$, such that for any M and γ, $a^M(\gamma)$ equals the endoscopic coefficient*

$$a^{M,\mathcal{E}}(\gamma) = \sum_{M' \in \mathcal{E}_{\mathrm{ell}}(M,V)} \sum_{\delta' \in \Sigma_+(M'_V)} \iota(M, M') b^{M'}(\delta') \Delta_M(\delta', \gamma).$$

We now sketch how to solve Problem 2 in terms of Problems 3 and 4. If G is quasisplit, let us define

$$S^G(f) = \sum_{M \in \mathcal{L}} |W_0^M||W_0^G|^{-1} \sum_{\delta \in \Sigma_+(M_V)} b^M(\delta) S_M^G(\delta, f).$$

According to Problem 3, this is a stable linear form on $\mathcal{H}\big(G(F_V)\big)$, and so satisfies the requirement of Problem 2. It remains to show that with this definition, the endoscopic identity of Problem 2 holds.

Suppose that G is arbitrary. The endoscopic expression of Problem 2 equals

$$I^{\mathcal{E}}(f) = \sum_{G' \in \mathcal{E}_{\mathrm{ell}}(G,V)} \iota(G, G') \sum_{R' \in \mathcal{L}^{G'}} |W_0^{R'}||W_0^{G'}|^{-1} S_{R'}(G'),$$

where $S_{R'}(G')$ is the sum over $\sigma' \in \Sigma_+(R'_V)$ of $b^{R'}(\sigma') \widehat{S}_{R'}^{G'}(\sigma', f')$. By a variant of [7, Lemma 9.2], this in turn equals

$$\begin{aligned} &\sum_{R \in \mathcal{L}^{G^*}} |W_0^R||W_0^{G^*}|^{-1} \sum_{R' \in \mathcal{E}_{\mathrm{ell}}(R,V)} \iota(R, R') \sum_{G' \in \mathcal{E}_{R'}(G^*)} \iota_{R'}(G^*, G') S_{R'}(G^*) \\ &= \sum_R |W_0^R||W_0^{G^*}|^{-1} \sum_{R'} \iota(R, R') \sum_{\sigma' \in \Sigma(R'_V)} b^{R'}(\sigma') I_R^{\mathcal{E}}(\sigma', f), \end{aligned}$$

where G^* is a quasisplit inner form of G, and $I_R^{\mathcal{E}}(\sigma', f)$ is defined as in Problem 3, but with (G, M, M', δ') replaced by (G^*, R, R', σ'). A global analogue of the vanishing property [7, Theorem 8.3] asserts that $I_R^{\mathcal{E}}(\sigma', f)$ vanishes unless R comes from G. If we identify $\mathcal{L}$ with a subset of $\mathcal{L}^{G^*}$, this means that $I_R^{\mathcal{E}}(\sigma', f)$ vanishes unless R is $W_0^{G^*}$-conjugate to a group $M \in \mathcal{L}$. In case R is conjugate to M, there are elements $M' \in \mathcal{E}_{\text{ell}}(M, V)$ and $\delta' \in \Sigma_+(M'_V)$ such that $I_R^{\mathcal{E}}(\sigma', f)$ equals the endoscopic expression $I_M^{\mathcal{E}}(\delta', f)$ of Problem 3. Since we also have $b^{R'}(\sigma') = b^{M'}(\delta')$ and $\iota(R, R') = \iota(M, M')$ in this case, the expression for $I^{\mathcal{E}}(f)$ can be written

$$\sum_{M \in \mathcal{L}} |W_0^M||W_0^G|^{-1} \sum_{M' \in \mathcal{E}_{\text{ell}}(M,V)} \iota(M, M') \sum_{\delta' \in \Sigma_+(M'_V)} b^{M'}(\delta') I_M^{\mathcal{E}}(\delta', f).$$

But the identities of Problems 3 and 4 imply that

$$\sum_{M' \in \mathcal{E}_{\text{ell}}(M,V)} \sum_{\delta' \in \Sigma_+(M'_V)} \iota(M, M') b^{M'}(\delta') I_M^{\mathcal{E}}(\delta', f) = \sum_{\gamma \in \Gamma(M,V)} a^M(\gamma) I_M(\gamma, f).$$

We can therefore conclude that $I^{\mathcal{E}}(f)$ equals

$$\sum_{M \in \mathcal{L}} |W_0^M||W_0^G|^{-1} \sum_{\gamma \in \Gamma(M,V)} a^M(\gamma) I_M(\gamma, f),$$

which is just $I(f)$. This is the required identity of Problem 2.

5. Problems 3 and 4 thus imply Problem 2. To relate Problem 2 to the basic Problem 1, we would need to solve spectral analogues of Problems 3 and 4.

Suppose that v is a place of F. We write $\Pi(G_v)$ for the set of equivalence classes of irreducible representations of $G(F_v)$. If G'_v is a local endoscopic datum for G, we shall write $\Phi(G'_v)$ for a fixed basis of the space of all stable distributions on $G'_v(F_v)$ spanned by irreducible characters. If v is archimedean, we take the elements in $\Phi(G'_v)$ to be analytic continuations (in the appropriate unramified spectral variable) of the stable tempered characters in [24]. In this case, elements in $\Phi(G'_v)$ correspond to Langlands parameters ϕ: $W_{F_v} \to {}^L G'_v$. If v is p-adic, we have to take $\Phi(G'_v)$ to be an abstract basis, obtained by analytic continuation from elements in the basis of tempered stable distributions chosen in [5, Proposition 5.1 and (5.1)]. Extending earlier notation, we write $f'_v(\phi'_v)$ for the pairing obtained from elements $f_v \in \mathcal{H}(G_v)$ and $\phi'_v \in \Phi(G'_v)$. By results in [24] and [5], we can write $f'_v(\phi'_v)$ as a finite linear combination of characters $f_{v,G}(\pi_v)$, with coefficients $\Delta_G(\phi'_v, \pi_v)$ that are spectral analogues of the original transfer factors.

We also extend notation we used earlier for the finite set V of valuations. Thus, if $M'_V = \prod M'_v$ is a product of local endoscopic data for a Levi subgroup M of G, and $\phi' = \prod \phi'_v$ belongs to $\Phi(M'_V) = \prod \Phi(M'_v)$, $\Delta_M(\delta', \gamma)$ equals the product over $v \in V$ of the factors $\Delta_M(\phi'_v, \gamma_v)$, for each $\pi = \prod \pi_v$ in $\Pi(M_V) = \prod \Pi(M_v)$. As before, we shall take M'_V to be the image of a global endoscopic datum $M' \in \mathcal{E}_{\text{ell}}(M, V)$.

PROBLEM 5: *Construct stable linear forms $S_M^G(\phi)$ on $\mathcal{H}\big(G(F_V)\big)$, for G quasisplit over F and $\phi \in \Phi(M_V)$, such that for any G, M, M' and ϕ', the linear form*

$$I_M(\phi', f) = \sum_{\pi \in \Pi(M_V)} \Delta_M(\phi', \pi) I_M(\pi, f)$$

equals the endoscopic expression

$$I_M^{\mathcal{E}}(\phi', f) = \sum_{G' \in \mathcal{E}_{M'}(G)} \iota_{M'}(G, G') \widehat{S}_{M'}^{G'}(\delta', f').$$

PROBLEM 6: *Construct coefficients $b^M(\phi)$, for M quasisplit over F and $\phi \in \Phi(M_V)$, such that for any M and π, $a^M(\pi)$ equals the endoscopic coefficient*

$$a^{M,\mathcal{E}}(\pi) = \sum_{M' \in \mathcal{E}_{\rm ell}(M,V)} \sum_{\phi' \in \Phi(M'_V)} \iota(M, M') b^{M'}(\phi') \Delta_M(\phi', \pi).$$

Set

$$I_{\rm aut}(f) = \int_{\Pi(G,V)} a^G(\pi) f_G(\pi) d\pi, \qquad f \in \mathcal{H}\big(G(F_V)\big).$$

Since $I_G(\pi, f) = \operatorname{tr}\big(\pi(f)\big) = f_G(\pi)$, for any $\pi \in \Pi(G_V)$, this is just the term with $M = G$ in the spectral expansion of $I(f)$. It can be regarded as the purely automorphic part of the trace formula. The discrete part of $I_{\rm aut}(f)$, regarded as a distribution on $\Pi(G, V)$, is just the distribution $I_{\rm disc}(f)$ of Problem 1.

If $M = G$, Problem 5 simply reduces to the expansion of $f'(\phi')$ above. However, Problem 6 is serious in this case, being closely related to Problem 1. It is really the other cases of Problems 5 and 6, those with $M \neq G$, that would be our immediate concern. Assume that these cases have been solved. It is then not hard to show from Problem 6 that if M is quasisplit, b^M is supported on a subset $\Phi(M, V)$ of $\Phi(M_V)$ that has a natural measure $d\phi$. Assuming that Problem 2 has also been solved, we set

$$S_{\rm aut}^G(f) = S^G(f) - \sum_{M \neq G} |W_0^G||W_0^G|^{-1} \int_{\Phi(M,V)} b^M(\phi) S_M^G(\phi, f) d\phi,$$

for any quasisplit group G and any $f \in \mathcal{H}\big(G(F_V)\big)$. According to Problems 2 and 5, this is a stable linear form on $\mathcal{H}\big(G(F_V)\big)$. If G is arbitrary, we consider the endoscopic expression

$$I_{\rm aut}^{\mathcal{E}}(f) = \sum_{G' \in \mathcal{E}_{\rm ell}(G,V)} \iota(G, G') \widehat{S}'_{\rm aut}(f'), \qquad f \in \mathcal{H}\big(G(F_V)\big).$$

Substituting for $\widehat{S}'_{\rm aut}(f')$ in this expression, we obtain a term to which Problem 2 applies, and a spectral expansion that can be treated by the argument we applied

in §4 to the geometric expansion of $I^{\mathcal{E}}(f)$. We arrive in the end at a formula that identifies $I^{\mathcal{E}}_{\rm aut}(f)$ with $I_{\rm aut}(f)$.

We have just sketched a solution of what would be Problem 1 if $I_{\rm disc}$, $S^G_{\rm disc}$ and $I^{\mathcal{E}}_{\rm disc}$ were replaced by $I_{\rm aut}$, $S^G_{\rm aut}$, and $I^{\mathcal{E}}_{\rm aut}$. But $I_{\rm disc}$ is just the discrete part of $I_{\rm aut}$. Using a well known argument that separates a suitable distribution into its continuous and discrete parts, one could obtain a solution of Problem 1 from what we have established.

6. We have not really proved anything. We have tried only to argue that Problems 3–6 are at the heart of stabilizing the trace formula. We shall conclude with a few words on the strategy for attacking these problems.

One begins by fixing G, and assuming inductively that all the problems can be solved if G is replaced by a proper subgroup. Since the coefficients $a^M(\gamma)$ and $a^M(\pi)$ depend only on M, this takes care of the global Problems 4 and 6, except for the case $M = G$. As for Problem 5, the residual distributions $I_M(\pi, f)$ are not independent of the distributions $I_M(\gamma, f)$ of Problem 3. The proof of [10, Theorem II.10.2] can likely be generalized to show that Problem 3 implies Problem 5. Now, the representations $\pi \in \Pi(M, V)$ that occur in the the spectral expansion (2) are unitary. In this case, there are descent and splitting formulas that express $I_M(\pi, f)$ in terms of related distributions on proper Levi subgroups M. Therefore, a solution of Problem 5 for the local terms in the spectral expansion would also follow from our induction assumption. (See [10, p. 145].)

It is Problem 3, then, that becomes the main concern. One has first to state the problem in a more elaborate form, one that generalizes the conjectures in [6, §4] and [7, §3], and clearly separates the inductive definitions from what is to be proved. This entails introducing adjoint transfer factors $\Delta_M(\gamma, \delta')$, that depend only on the image of δ' in a certain set $\Sigma^{\mathcal{E}}_+(M_V)$ attached to M. We cannot go into any detail, but the construction is a generalization of the discussion of [5, §2] and [7, §2] for strongly regular conjugacy classes. To have adjoint relations, and for that matter, the global vanishing theorem mentioned in §4, one has actually to take G to be a certain disjoint union of connected groups — a global K-group, in language suggested in [7]. At any rate, once we have the factor $\Delta_M(\gamma, \delta')$, we can set

$$I^{\mathcal{E}}_M(\gamma, f) = \sum_{\delta' \in R^{\mathcal{E}}(M_V)} \Delta_M(\gamma, \delta') I^{\mathcal{E}}_M(\delta', f), \qquad \gamma \in \Gamma_+(M_V),$$

as in [7, (5.5)]. The required identity of Problem 3 becomes the assertion that $I^{\mathcal{E}}_M(\gamma, f)$ equals $I_M(\gamma, f)$.

The terms in the endoscopic expression $I^{\mathcal{E}}(f)$ of Problem 2 can be defined inductively. An elaboration of the argument sketched in §4, and which is the global analogue [7, Theorem 9.1(a)], then leads to a geometric expansion for $I^{\mathcal{E}}(f)$ that is parallel to the expansion (1) for $I(f)$. The general strategy is to compare these two expansions. In particular, one obtains an explicit geometric expansion for the difference $I^{\mathcal{E}}(f) - I(f)$. On the other hand, similar considerations lead to a spectral expansion of $I^{\mathcal{E}}(f) - I^{\mathcal{E}}_{\rm aut}(f)$. The cases of Problems 5 and 6 implied by the induction hypothesis actually identify the terms in this latter expansion with corresponding terms in the original spectral expansion for $I(f) - I_{\rm aut}(f)$. The

result is a formula

$$I^{\mathcal{E}}(f) - I(f) = I^{\mathcal{E}}_{\rm aut}(f) - I_{\rm aut}(f). \tag{4}$$

To be able to exploit the last formula, one has to extend most of the techniques of Chapter II of [10], (as well as add a few new ones, based on the local trace formula). We mention just one, the problem of descent for the global coefficients. There is a simple descent formula for the coefficients $a^G(\gamma)$ at arbitrary γ in terms of coefficients evaluated at unipotent elements [2, (8.1)]. Using the main theorem of [20], one can establish a parallel descent formula for $a^{G,\mathcal{E}}(\gamma)$. Together with the fundamental lemma, which takes care of the spherical weighted orbital integrals we have built into the definition of these coefficients, this reduces the identity of Problem 4 (with $M = G$) to the case of unipotent γ. It allows one to collapse the terms with $M = G$ in the geometric expansion of the left hand side of (4) to a sum over unipotent elements. Similarly, there is a descent formula for coefficients $a^G(\pi)$ at arbitrary π in terms of discrete parts $a^M_{\rm disc}(\pi)$ and unramified partial L-functions. Using simple combinatorial arguments, one can establish a parallel descent formula for $a^{G,\mathcal{E}}(\pi)$. This reduces the identity of Problem 6 (with $M = G$) to the case of the coefficients $a^G_{\rm disc}(\pi)$, and allows one to replace the right hand side of (4) with the distribution $I^{\mathcal{E}}_{\rm disc}(f) - I_{\rm disc}(f)$. It is this revised form of (4) that should eventually yield the required identities of the various problems.

If G is quasisplit, a global analogue of [7, Theorem 9.1(b)] gives a geometric expansion of the distribution $S^G(f)$ of Problem 2. One has to carry out an analysis of this expansion that is largely parallel to the discussion above. Similar techniques should eventually yield the required stability assertions of the various problems.

REFERENCES

1. M. Assem, *Matching of certain unipotent orbital integrals on p-adic orthogonal groups,* Thesis, University of Washington, 1988.
2. J. Arthur, *On a family of distributions obtained from orbits,* Canad. J. Math. **38** (1986), 179–214.
3. J. Arthur, *The local behaviour of weighted orbital integrals,* Duke Math. J. **56** (1988), 223–293.
4. J. Arthur, *The invariant trace formula II. Global theory,* J. Amer. Math. Soc. **1** (1988), 501–554.
5. J. Arthur, *On local character relations,* Selecta Math. **2** (1996), 501–579.
6. J. Arthur, *Canonical normalization of weighted characters and a transfer conjecture,* to appear in C.R. Math. Rep. Acad. Sci. Canada.
7. J. Arthur, *On the transfer of distributions: weighted orbital integrals,* preprint.
8. J. Arthur, *A stable trace formula,* in preparation.
9. J. Arthur, *On the transfer of distributions: singular orbital integrals and weighted characters,* in preparation.
10. J. Arthur and L. Clozel, *Simple Algebras, Base Change and the Advanced Theory of the Trace Formula,* Annals of Math. Studies **120**, Princeton University Press, 1989.
11. T. Hales, *The fundamental lemma for Sp*(4)*,* Proc. Amer. Math. Soc. **125** (1997), 301–308.

12. H. Jacquet and R.P. Langlands, *Automorphic Forms on $GL(2)$,* Springer Lecture Notes **114**, 1970.
13. H. Jacquet, I.I. Piatetskii-Shapiro, and J.A. Shalika, *Rankin-Selberg convolutions,* Amer. J. Math. **105** (1983), 367–464.
14. R. Kottwitz, *Stable trace formula: cuspidal tempered terms,* Duke Math. J. **51** (1984), 611–650.
15. R. Kottwitz, *Stable trace formula: elliptic singular terms,* Math. Ann. **275** (1986), 365–399.
16. R. Kottwitz and D. Shelstad, *Twisted endoscopy I and II,* preprints.
17. J.-P. Labesse and R.P. Langlands, *L-indistinguishability for $SL(2)$,* Canad. J. Math. **31** (1979), 726–785.
18. R.P. Langlands, *Les Débuts d'une Formule des Traces Stable,* Publications Mathématiques, vol. **13**, Université Paris VII, 1983.
19. R.P. Langlands and D. Shelstad, *On the definition of transfer factors,* Math. Ann. **278** (1987), 219–271.
20. R.P. Langlands and D. Shelstad, *Descent for transfer factors,* The Grothendieck Festschrift, vol. II, Birkhäuser, Boston, 1990, 485–563.
21. C. Moeglin and J.-L. Waldspurger, *Le spectre résidual de $GL(n)$,* Annales de l'ENS **22** (1989), 605–674.
22. W. Müller, *The trace class conjecture in the theory of automorphic forms,* Ann. of Math. **130** (1989), 473–529.
23. J. Rogawski, *Automorphic Representations of Unitary Groups in Three Variables,* Annals of Math. Studies **123**, 1990.
24. D. Shelstad, *L-indistinguishability for real groups,* Math. Ann. **259** (1982), 385–430.
25. J.-L. Waldspurger, *Sur les intégrales orbitales tordues pour les groupes linéaires: une lemme fondamental,* Canad. J. Math. **43** (1991), 852–896.
26. J.-L. Waldspurger, *Le lemme fondamental implique le transfert,* Compositio Math. **105** (1997), 153–236.

James Arthur
Department of Mathematics
University of Toronto
Toronto, Ontario M5S 3G3
Canada
ida@math.toronto.edu

Analytic Structures on Representation Spaces of Reductive Groups

Joseph Bernstein

Abstract. We show that every admissible representation of a real reductive group has a canonical system of Sobolev norms parametrized by positive characters of a minimal parabolic subgroup. These norms are compatible with morphisms of representations. Similar statement also holds for representations of reductive p-adic groups.

1991 Mathematics Subject Classification: 22E46 43A70 46E39
Keywords and Phrases: reductive groups,representations, Sobolev norms, canonical topology

1. Analytic structures on representation spaces

Let us fix a real reductive algebraic group G. We would like to study analytic structures which are naturally defined on a representation (π, G, V) of the group G.

The results which I discuss in this lecture hold for an arbitrary reductive group G but to explain the ideas and motivations I will discuss only the group $G = SL(2, \mathbb{R})$ and later on the more interesting case of the group $G = SL(3, \mathbb{R})$.

Historically, mathematicians first were interested only in unitary representations of G and for such representations analytic structure is rather clear.

But later it was realized that it is convenient to introduce and study also some continuous representations (π, G, V) . Here V is a complex topological vector space (we consider only Banach and Frechet spaces); representation π of the group G in the space V is called continuous if the corresponding map $G \times V \to V$ is continuous.

Here we immediately encounter a problem how we should think about a representation. In order to explain the problem consider the simplest case of the group $G = SL(2, \mathbb{R})$.

The typical representation of this group is described as follows. Fix a complex number λ and consider the space D_λ of even homogeneous functions on the punctured plane $\mathbb{R}^2 \setminus 0$ of homogeneous degree $\lambda - 1$. Then the group $G = SL(2, \mathbb{R})$ naturally acts on the space D_λ and we denote this representation as π_λ.

Now the problem with this description of the representation π_λ is that we have not specified the class of functions which we consider. We can take smooth functions, or functions which are locally L^2 or one of the many other classes of functions (L^p, Sobolev functions, Besov functions and so on).

It is easy to see that all these representations will be non isomorphic as topological representations. However it is intuitively clear that this are just different analytic realizations of the "same" representation or, in other words, all these representations are equivalent.

Harish Chandra proposed the following way to handle this problem. We say that a representation (π, G, V) is ADMISSIBLE if its restriction to a maximal compact subgroup $K \subset G$ has finite multiplicities and it has finite length as a topological representation.

Given such a representation we denote by V_f the space of K-finite vectors (a vector $v \in V$ is called K-FINITE if the subset $\pi(K)v$ lies in a finite dimensional subspace of V). The space V_f is not a G-module, but it has natural actions of the group K and of the Lie algebra $\mathfrak{g}$ of the group G. These two actions are compatible in a natural way.

Now we can abstractly define a purely algebraic notion of a $(\mathfrak{g}, K)$-module as a vector space E equipped with two actions, of the Lie algebra $\mathfrak{g}$ and of the group K, which satisfy these compatibility conditions. We say that a $(\mathfrak{g}, K)$-module E is a HARISH CHANDRA module if it is finitely generated as a $\mathfrak{g}$-module and has finite multiplicities as a K-module (see details in [1]).

Starting with an admissible topological representation (π, G, V) we have constructed a Harish Chandra module V_f. Now, following Harish Chandra, we say that two admissible topological representations are EQUIVALENT if the corresponding Harish Chandra modules are isomorphic.

I propose a slightly different point of view on this problem. Let us agree that AN ADMISSIBLE REPRESENTATION π of the group G is an equivalence class of admissible topological representations of G. Concrete topological representations in this class we consider as different "analytic realizations" of a given representation π.

With such an understanding we see that representations of G are parameterized by Harish Chandra modules (it is not difficult to show that every Harish Chandra module corresponds to some topological representation). In particular, the very difficult problem of classification of irreducible representations of G is reduced to a still difficult, but purely algebraic, problem of classification of irreducible Harish Chandra modules. This classification problem has been solved by several different algebraic methods.

1.1. WHAT WE WANT TO ACHIEVE. Let us come back to our analytic problem. Suppose we are given a representation (for example represented by a Harish Chandra module E). We would like to describe some NATURAL analytic structures which we can define on this representation.

In order to explain what we are after let us consider first a model case. Namely, suppose we are given a C^∞ manifold M and a C^∞ vector bundle E on M and we would like to study possible analytic structures on the space V of the sections of E.

We may consider many different analytic structures on V: smooth sections, C^n-sections, L^2-sections (or, more generally, L^p-sections), different kinds of Sobolev spaces of sections, of Besov spaces of sections and so on.

There is a convenient way to represent all these structures. Namely, we fix the space $V = V^\infty$ of smooth sections and study different topologies $\mathcal{T}$ on this space.

For a given topology $\mathcal{T}$ we denote by $L_\mathcal{T}$ the completion of the space V with respect to $\mathcal{T}$. It is convenient to consider the space $L_\mathcal{T}$ as a subspace of the space of distribution sections of the bundle E.

For example, while the space V does not have a CANONICAL structure of a (pre)Hilbert space it clearly has a canonically defined Hilbert topology (i.e. topology defined by a norm N of Hermitian type , which means that the function $v \mapsto N(v)^2$ is a Hermitian form on V). The completion L of the space V with respect to this topology is a canonically defined Hilbertian space of sections of E.

More generally, for every real number s we can canonically define L^2 type Sobolev topology $\mathcal{T}_s$. It is defined by a Sobolev Hermitian norm S_s on the space V; the completion of the space V with respect to this norm is the Sobolev space of sections L_s.

The explicit description of Sobolev norms S_s is standard, but a little involved. The easiest way to define them is to use Fourier transform - but we are trying to avoid this since we will not be able to generalize this method.

A relatively simple description can be given when $s = k$ is a positive integer. In this case, if E is a trivial bundle and ϕ is a section of E (i.e. a function) supported in a small neighborhood with coordinates (x_i) we can define the Sobolev norm $S_k(\phi)$ to be $(\sum |\partial_\alpha \phi|^2)^{1/2}$, where the sum is over all multiindeces α of degree less or equal to k.

It is important that each analytic structure $\mathcal{T}$ on the space V which we considered has local description. Formally, this means that for any smooth function f on M the operator of multiplication by f is continuous in the corresponding topology; thus using the partition of unit we see that a distribution section v lies in the completion $L_\mathcal{T}$ if and only if this holds locally.

1.2. Analytic structures on representation spaces. Now let us come back to the case of an admissible representation (π, G, V). For every such representation we can consider its smooth part (π, G, V^∞), where V^∞ is the space of smooth vectors $v \in V$ (a vector $v \in V$ is called SMOOTH if the corresponding function $G \to V$, $g \mapsto \pi(g)(v)$, is smooth).

By a remarkable theorem of Casselman and Wallach, for every two realizations of a given admissible representation π their smooth parts are canonically isomorphic as topological representations; in fact they have shown that the functor $V \mapsto V_f$ defines an equivalence of the category of smooth admissible representations of G and the category of Harish Chandra modules (see details in [2]).

In other words, any representation has a CANONICAL "smooth model" (π, G, V) for which $V = V^\infty$.

My aim in this lecture is to discuss different analytic structures (in particular Sobolev structures) which can be canonically constructed on a given representation

π of the group G. As before, we will describe these structures as different topologies on the smooth model (π, G, V) of the given representation π.

2. Case of the group $SL(2, \mathbb{R})$

Consider as a simple example the group $G = SL(2, \mathbb{R})$ and its representation $(\pi_\lambda, G, D_\lambda)$ in the space of homogeneous functions described above. This representation is called a principle series representation; it is induced by some unramified character μ of the Borel subgroup $B \subset G$, $\pi_\lambda = Ind_B^G(\mathbb{C}_\mu)$.

Since the space D_λ is realized as the space of homogeneous functions on $\mathbb{R}^2 \setminus 0$, by restricting to the unit circle $S^1 \subset \mathbb{R}^2 \setminus 0$ we can identify the space V with the space $\mathcal{F}$, where $\mathcal{F} = C^\infty(S^1)_{even}$ is the space of even functions on the unit circle.

Thus, we can define s Sobolev structure on the space V using the norm N_s given in the realization D_λ by the formula $N_s(v) = S_s(v)$.

The problem with this definition is that for generic λ the space V has two natural realizations, as D_λ and as $D_{-\lambda}$, and the norms N_s obtained from these two realizations are not equivalent. Hence this approach does not work.

However, let us analyze these two realizations more carefully. Since the spaces D_λ and $D_{-\lambda}$ are both identified with the space $\mathcal{F}$ the equivalence between them is given by some operator $I_\lambda : \mathcal{F} \to \mathcal{F}$. This operator, which is usually called an INTERTWINING OPERATOR, can be explicitly described. It turns out that it is a pseudo-differential operator of order r, where $r = \operatorname{Re}\lambda$. In fact, it can be realized as a convolution operator with some distribution R_λ on S^1 which is smooth outside of the origin and near the origin is more or less homogeneous of degree $-\lambda - 1$.

So let us try to define the s Sobolev norm N_s on the space V to be $N_s = S_{s+r/2}$, where $r = \operatorname{Re}(\lambda)$ and where the Sobolev norm $S_{s+r/2}$ is computed using the realization D_λ. Then from the description above one can immedeatly see that the corresponding topology on the space V does not depend on the choice of the realization (at least for generic λ). This allows us to define a canonical s-Sobolev structure on the space V (for generic λ).

There are also other representations of principle series. They correspond to characters of the Borel subgroup which lie in a different component, i.e. have different discrete parameters compared with characters λ above. These representations are realized in the space of odd functions on $\mathbb{R}^2 \setminus 0$; since locally on S^1 they are exactly the same as representations D_λ they have the same analytic structure. Thus the Sobolev norm N_s on representations D_λ induces similar norm on these "odd" representations. In other words, discrete parameters do not affect the analytic structure.

We have described our Sobolev norm for generic point λ. For arbitrary λ this construction may not work - for example the operator I_λ may have pole, or, after normalization, it becomes bounded but not invertible. However there are standard algebraic methods which allow to reduce the study of these cases to the study of representations with generic λ.

Now we can use a deep algebraic theorem that every Harish Chandra module E can be imbedded into some generalized representation of principle series, i.e. a

representation induced from some finite dimensional representation of the Borel subgroup B. We can extend our Sobolev norm N_s to these generalized principle series.

Then, again using deep algebraic results about Harish Chandra modules, we can show that the resulting norm N_s on the space E does not depend on the choice of a particular embedding.

Thus using these methods we can extend the definition of the Sobolev norm N_s to all admissible representations (π, G, V^∞) of the group G.

In particular the norm N_0 defines a CANONICAL Hilbertian structure on an admissible representation.

3. Construction of Sobolev norms for a general group G

Let us discuss the case of a general reductive group G. For simplicity assume that G is split (e.g. $G = SL(n, \mathbb{R})$). Then again we can consider a series of representations $(\pi_\lambda, G, D_\lambda)$ parameterized by unramified characters λ of the split Cartan subgroup $A \subset G$; each of these representations can be realized in the space $\mathcal{F}$ of functions on the flag variety $X = G/B$, where B is a Borel subgroup.

In this case it is not clear how to define Sobolev norms on V, since the usual family of Sobolev norms S_s on the space $\mathcal{F}$ depends on one parameter s while representations D_λ depend on several parameters λ.

However, it turns out that the flag variety X has a very special geometric structure. Using this structure we can equip the space $\mathcal{F}$ of functions on X with a canonical system of Sobolev norms S_s parameterized by points s of the $\mathbb{R}$-linear space $\mathfrak{a}^* = \mathrm{Mor}(A, \mathbb{R}^+)$.

This space $\mathfrak{a}^*$ is dual to the Cartan sub algebra $\mathfrak{a} = \mathrm{Lie}(A)$. Using the exponential map we will identify the space $\mathfrak{a}^*$ with the group of positive characters of the Cartan group A, or, equivalently, with the group of positive characters of the Borel group B. We will mostly think about the space $\mathfrak{a}^*$ in this realization; for example, in case of a general reductive group G the space $\mathfrak{a}^*$ is defined as $\mathfrak{a}^* = \mathrm{Mor}(P, \mathbb{R}^+) \simeq \mathrm{Mor}(P, \mathbb{R}^{+*})$, where P is the minimal parabolic subgroup of G.

Now, similarly to the $SL(2)$ case, we can define s-Sobolev norm N_s on the space V as $N_s(v) = S_{s+r/2}(v)$, where the positive character $r = \mathrm{Re}(\lambda) \in \mathfrak{a}^*$ is defined by $r(a) = |\lambda(a)|$ and the norm $S_{s+r/2}$ is computed in D_λ realization.

This construction defines a canonical system of the Sobolev norms N_s on any representation V of the group G which is isomorphic to one of the representations D_λ for generic λ. This system of Sobolev structures is parameterized by points $s \in \mathfrak{a}^*$.

Again, using algebraic methods we can reduce the case of an arbitrary admissible representation V to one of these non degenerate cases and using the definition described above we can define a canonical system of Sobolev topologies $\mathcal{T}_s$ on the space V.

4. Construction of the family of Sobolev norms S_s on the space $\mathcal{F}$

Let us describe how to construct the Sobolev norms S_s on the space $\mathcal{F}$ of functions on the flag variety X. For simplicity we consider only the case of the group $G = SL(3, \mathbb{R})$ – it shows all the ideas and all the difficulties.

In this case the space $\mathfrak{a}^*$ is two dimensional; it is best realized as a quotient space of the space $\{(x_1, x_2, x_3)\}$ modulo the subspace $\{(x, x, x)\}$. Unramified characters of the Borel group are parameterized by the points λ in the complexification $\mathfrak{a}^*_{\mathbb{C}}$. There are also some other characters which differ from the characters λ by some discrete parameters. As before we can ignore these discrete parameters and reduce all constructions to finding the family of Sobolev norms S_s on the space $\mathcal{F}$.

By definition, the space D_λ is unitarily induced from the character λ of the Borel subgroup B (which in this case is the subgroup of upper triangular matrices). This means that D_λ consists of smooth functions ϕ on G satisfying $\phi(bg) = \mu(b)\phi(g)$ for $g \in G$ and $b \in B$; here $\mu = \rho^{-1}\lambda$ is the character of the Borel subgroup B which differs from λ by the standard character ρ.

Restricting these functions to the maximal compact subgroup K we will identify all the spaces D_λ with the space $\mathcal{F}$ of smooth functions on the flag variety X.

We have the natural action of the Weyl group $W = S_3$ on $\mathfrak{a}^*$ and on $\mathfrak{a}^*_{\mathbb{C}}$ given by permutation of coordinates.

It is known that for generic λ all representations $D_{w\lambda}$ are isomorphic. Let us describe this more specifically for the case of a simple reflection σ; for example we consider the simple root $\alpha = (1, -1, 0) \in \mathfrak{a}^*$ and the corresponding permutation $\sigma = \sigma_\alpha \in W$ of indeces 1 and 2. In this case $\sigma\lambda$ has the form $\sigma_\lambda = \lambda - \lambda_\alpha \cdot \alpha$ for some number λ_α and the equivalence between spaces D_λ and $D_{\sigma}\lambda$ can be described using an intertwining operator $I_{\sigma,\lambda_\alpha} : \mathcal{F} \to \mathcal{F}$, which depends only on σ and on the number λ_α.

In fact, in this case the operator I can be described quite explicitly. Namely, consider the natural fibration of the flag variety X over a Grassmannian $X_\alpha = \mathrm{Gr}_{2,3}$. The fibers of this filtration are circles, and on each of these circles we can define an intertwining operator I_α as in the case of $SL(2, \mathbb{R})$. Together these operators represent the operator $I_{\sigma,\lambda_\alpha} : \mathcal{F} \to \mathcal{F}$.

The system of Sobolev spaces L_s for $s \in \mathfrak{a}^*$ should satisfy the following condition:

(*) $\qquad I_{\sigma,\lambda_\alpha} L_s \subset L_{s - r_\alpha \alpha/2}$, where $r_\alpha = \mathrm{Re}\,\lambda_\alpha$.

Since any weight $s \in \mathfrak{a}^*$ is a linear combination of simple weights α and β we see that this condition, together with the similar condition for the root β and with the condition that the space L_0 is the space of L^2-functions, completely determine all the Sobolev norms S_s on $\mathcal{F}$ (up to topological equivalence). Namely if $s = a\alpha + b\beta$ we have to define L_s to be the image $I_{\sigma_\alpha,-2a} \circ I_{\sigma_\beta,-2b}(L^2(X))$ (for generic s). One can check that this definition gives a family of Sobolev norms satisfying (*).

5. Remarks

Remark 1. Let (π, G, V) be a unitary admissible representation of G. Then we have two Hilbertian structures on V - one canonical structure described above and another given by the unitary structure on V. It is natural to assume that these two structures always coincide (and in simplest cases this is true).

If this conjecture holds it may help in the description of unitary representations of the group G.

Remark 2. If we consider representations of a p-adic reductive group G then we will find exactly the same analytic structures. They are parameterized by points s of the real vector space $\mathfrak{a}^* = \mathrm{Mor}(A, \mathbb{R}^+)$, where A is the maximally split Cartan group of G. The proof in this case is different, since there are many representations of G which can not be realized on flag varieties (so called cuspidal representations).

Remark 3. In case of the group $SL(2, \mathbb{R})$ we can use Calderon-Zygmund theorem which implies that the intertwining operator I_r is continuous with respect to L^p Sobolev norms, i.e. it defines a continuous operator $I_r : (\mathcal{F}, S_{p,s}) \to (\mathcal{F}, S_{p,s-r})$. Using this we can canonically define L^p Sobolev structures on representations isomorphic to D_λ.

Thus, it is probable that if we fix a number $p \in [1, \infty)$ then for any group G and any admissible representation (π, G, V) of G we can define a canonical family of L^p type Sobolev norms $N_{p,s}$ on the space V which is parameterized by points $s \in \mathfrak{a}^*$.

Probably the same construction will also yield canonical Besov structures on the space V.

References

[1] N. Wallach, *Real reductive groups II*, Academic Press, 1992.
[2] W. Casselman, *Canonical extension of Harish-Chandra modules to representations of G*, Canad. Journal of Math., vol. 41 (1989), No.3 , 385-438

Joseph Bernstein
Dept. Math.
Tel Aviv University
Ramat Aviv, Israel
bernstei@math.tau.ac.il

From Double Hecke Algebra to Analysis

Ivan Cherednik

Abstract. We discuss q-counterparts of the Gauss integrals, a new type of Gauss-Selberg sums at roots of unity, and q-deformations of Riemann's zeta. The paper contains general results, one-dimensional formulas, and remarks about the current projects involving the double affine Hecke algebras.

Keywords and Phrases: Hecke algebra, Fourier transform, spherical function, Macdonald polynomial, Gauss integral, Gaussian sum, metaplectic representation, Verlinde algebra, braid group, zeta function.

Introduction.

The note is about the role of double affine Hecke algebras in the unification of the classical zonal and p-adic spherical functions and the corresponding Fourier transforms. The new theory contains one more parameter q and, what is important, dramatically improves the properties of the Fourier transform. In contrast to the real and p-adic theories, the q-transform is selfdual and has practically all other important properties of the classical Fourier transform. Here I will mainly discuss the Fourier-invariance of the Gaussian.

There are various applications. In combinatorics, they are via the Macdonald polynomials. As $q \to 1$, we complete the Harish-Chandra theory of the spherical transform. The limit $q \to \infty$ covers the p-adic Iwahori-Matsumoto-Macdonald theory. When q is a root of unity, we generalize the Verlinde algebras, directly related to quantum groups and Kac-Moody algebras, and come to a new class of Gauss-Selberg sums.

However the main applications could be of more analytic nature. The representation of the double affine Hecke algebra generated by the Gaussian and its Fourier transform can be described in full detail. So the next step is to examine the spaces generated by Gaussian-type functions. The Fourier transforms of the simplest examples lead to q-deformations of the classical zeta and L-functions.

Of course there are other projects involving the double Hecke algebras. I will mention at least some of them. The following is far from being complete.

1) *Macdonald's q-conjectures* [M1,M2]. Namely, the norm, duality, and evaluation conjectures [C1,C2]. My proof of the norm-formula is similar to that from [O1] in the differential case (the duality and evaluation conjectures collapse as $q \to 1$). I would add to this list the Pieri rules [C2]. As to the nonsymmetric Macdonald polynomials, see [O2,M3,C3]. See also [M3,DS,Sa] about the $C^\vee C$ (the Koornwinder polynomials), and [I,M4,C4] about the Aomoto conjecture.

2) *K-theoretic interpretation.* I mean the papers [KL1,KK] and more recent [GG,GKV]. Presumably it can lead to the Langlands-type description of irreducible representations of double Hecke algebras, but the answer can be more complicated than in [KL1] (see also recent Lusztig's papers on the representations of affine Hecke algebras with unequal parameters). The Fourier transform is misty in this approach. Let me add here the strong Macdonald conjecture (Hanlon).

3) *Induced and spherical representations.* The classification of the spherical representations is much simpler, as well as the irreducibility of the induced ones. I used the technique of intertwiners in [C4] following a similar theory for the affine Hecke algebras. The nonsymmetric polynomials form the simplest spherical representation. There must be connections with [HO1]. The intertwiners also serve as creation operators for the nonsymmetric Macdonald polynomials (the case of GL is due to [KS]).

4) *Radial parts via Dunkl operators.* The main references are [D1,H,C5]. In the latter it was observed that the trigonometric differential Dunkl operators form the degenerate (graded) affine Hecke algebra [L] ([Dr] for GL_n). The difference, elliptic, and difference-elliptic generalizations were introduced in [C6,C7,C8]. The nonsymmetric Macdonald polynomials are eigenfunctions of the difference Dunkl operators. The connections with the KZ-equation play an important role here. I mean Matsuo's and my theorems from [Ma,C5,C6]. See also [C9].

5) *Harmonic analysis.* In the rational-differential setup, the definition of the generalized Bessel functions is from [O3], the corresponding generalized Hankel transform was considered in [D2,J] (see also [He]). In contrast to the spherical transform, it is selfdual, as well as the difference generalization from [C2,C10]. The Mehta-Macdonald conjecture, directly related to the transform of the Gaussian, was checked in [M1,O1] in the differential case and generalized in [C10]. See [HO2,O2,C11] about applications to the Harish-Chandra theory.

6) *Roots of unity.* The construction from [C2] generalizes and, at the same time, simplifies the Verlinde algebras. The latter are formed by the so-called reduced representations of quantum groups at roots of unity. Another interpretation is via the Kac-Moody algebras [KL2] (due to Finkelberg for roots of unity). A valuable feature is the projective action of $PSL(2,\mathbf{Z})$ (cf. [K, Theorem 13.8]). In [C3] the nonsymmetric polynomials are considered, which establishes connections with the metaplectic (Weil) representations at roots of unity.

7) *Braids.* Concerning $PSL(2,\mathbf{Z})$, it acts projectively on the double Hecke algebra itself. The best known explanation (and proof) is based on the interpretation of this algebra as a quotient of the group algebra of the fundamental group of the elliptic configuration space [C6]. The calculation is mainly due to [B] in the GL-case. For arbitrary root systems, it is similar to that from [Le], but our configuration space is different. Switching to the roots of unity, there may be applications to the framed links including the Reshetikhin-Turaev invariants.

8) *Duality.* The previous discussion was about arbitrary root systems. In the case of GL, the theorem from [VV] establishes the duality between the double Hecke algebras and the q-toroidal (double Kac-Moody) algebras. It generalizes the classical Schur-Weyl duality, Jimbo's q-duality, and the affine analogues from [Dr,C12]. When the center charge is nontrivial it explains the results from [STU],

which were recently extended by Uglov to irreducible representations of the Kac-Moody gl_N of arbitrary positive integral levels.

Let me also mention the relations of the symmetric Macdonald polynomials (mainly of the GL-type) to: a) the spherical functions on q-symmetric spaces (Noumi and others), b) the interpolation polynomials (Macdonald, Lassalle, Knop and Sahi, Okounkov and Olshanski), c) the quantum gl_N (Etingof, Kirillov Jr.), d) the KZB-equation (—, —, Felder, Varchenko). There are connections with the affine Hecke algebra technique in the classical theory of GL_N and S_n. I mean, for instance, [C12], papers of Nazarov and Lascoux, Leclerc, Thibon, and recent results towards the Kazhdan-Lusztig polynomials.

The coefficients of the symmetric GL-polynomials have interesting combinatorial properties (Macdonald, Stanley, Garsia, Haiman, ...). These polynomials appeared in Kadell's work. Their norms are due to Macdonald, the evaluation and duality conjectures were checked by Koornwinder, the Macdonald operators were introduced independently by Ruijsenaars together with elliptic deformations.

Quite a few constructions can be extended to arbitrary finite groups generated by complex reflections. For instance, the Dunkl operators and the KZ-connection exist in this generality (Dunkl, Opdam, Malle). One can try the affine and even the hyperbolic groups (Saito's root systems).

1. ONE-DIMENSIONAL FORMULAS.

The starting point of many mathematical and physical theories is the formula:

$$2\int_0^\infty e^{-x^2}x^{2k}\mathrm{d}x \;=\; \Gamma(k+1/2),\ \Re k>-1/2. \tag{1}$$

Let us give some examples.

(a) Its generalization to the Bessel functions, namely, the invariance of the Gaussian e^{-x^2} with respect to the Hankel transform, is a cornerstone of the Plancherel formula.

(b) The following "perturbation" for the same $\Re k>-1/2$

$$\mathbf{3}(k)\stackrel{def}{=}2\int_0^\infty (e^{x^2}+1)^{-1}x^{2k}\,\mathrm{d}x=(1-2^{1/2-k})\Gamma(k+1/2)\zeta(k+1/2) \tag{2}$$

is fundamental in the analytic number theory.

(c) The multi-dimensional extension due to Mehta with $\prod_{1\le i<j\le n}(x_i-x_j)^{2k}$ instead of x^{2k} gave birth to the theory of matrix models and the Macdonald theory with various applications in mathematics and physics.

(d) Switching to the roots of unity, the Gauss formula

$$\sum_{m=0}^{2N-1} e^{\frac{\pi m^2}{2N}i} \;=\; (1+i)\sqrt{N},\quad N\in\mathbf{N} \tag{3}$$

can be considered as a certain counterpart of (1) at $k=0$.

(e) Replacing x^{2k} by $\sinh(x)^{2k}$, we come to the theory of spherical and hypergeometric functions and to the spherical Fourier transform. The spherical transform of the Gaussian plays an important role in the harmonic analysis on symmetric spaces.

To employ modern mathematics at full potential, we do need to go from Bessel to hypergeometric functions. In contrast to the former, the latter can be studied, interpreted and generalized by a variety of methods in the range from representation theory and algebraic geometry to integrable models and string theory. However the straightforward passage $x^{2k} \to \sinh(x)^{2k}$ creates problems. The spherical transform is not selfdual anymore, the formula (1) has no sinh-counterpart, and the Gaussian looses its Fourier-invariance.

DIFFERENCE SETUP. It was demonstrated recently that these important features of the classical Fourier transform are restored for the kernel

$$\delta_k(x;q) \stackrel{def}{=} \prod_{j=0}^{\infty} \frac{(1-q^{j+2x})(1-q^{j-2x})}{(1-q^{j+k+2x})(1-q^{j+k-2x})},\ 0<q<1,\ k\in\mathbf{C}. \tag{4}$$

Actually the selfduality of the corresponding transform can be expected a priori because the Macdonald truncated theta-function δ is a unification of $\sinh(x)^{2k}$ and the Harish-Chandra function (A_1) serving the inverse spherical transform.

As to (1), setting $q=\exp(-1/a),\ a>0$,

$$(-i)\int_{-\infty i}^{\infty i} q^{-x^2}\delta_k\,\mathrm{d}x = 2\sqrt{a\pi}\prod_{j=0}^{\infty}\frac{1-q^{j+k}}{1-q^{j+2k}},\quad \Re k>0. \tag{5}$$

Here both sides are well-defined for all k except for the poles but coincide only when $\Re k>0$, worse than in (1). This can be fixed as follows:

$$(-i)\int_{1/4-\infty i}^{1/4+\infty i} q^{-x^2}\mu_k\,\mathrm{d}x = \sqrt{a\pi}\prod_{j=1}^{\infty}\frac{1-q^{j+k}}{1-q^{j+2k}},\ \Re k>-1/2 \ \text{ for} \tag{6}$$

$$\mu_k(x;q) \stackrel{def}{=} \prod_{j=0}^{\infty} \frac{(1-q^{j+2x})(1-q^{j+1-2x})}{(1-q^{j+k+2x})(1-q^{j+k+1-2x})},\ 0<q<1,\ k\in\mathbf{C}. \tag{7}$$

The limit of (6) multiplied by $(a/4)^{k-1/2}$ as $a\to\infty$ is (1) in the imaginary variant.

Once we managed Γ, it would be unexcusable not to try (cf. (2))

$$\mathfrak{Z}_q(k) \stackrel{def}{=} (-i)\int_{1/4-\infty i}^{1/4+\infty i}(q^{x^2}+1)^{-1}\mu_k\,\mathrm{d}x \ \text{ for } \Re k>-1/2. \tag{8}$$

It has a meromorphic continuation to all k periodic in the imaginary direction. The limit of $(a/4)^{k-1/2}\mathfrak{Z}_q$ as $a\to\infty$ is $\mathfrak{Z}$ for all k except for the poles. The analytic continuation is based on the shift operator technique. It seems that all zeros of $\mathfrak{Z}_q(k)$ for $a>1, \Re k>-1/2$ are q-deformations of the zeros of $\mathfrak{Z}(k)$.

Jackson and Gauss sums. A most promising feature of special q-functions is a posiblility to replace the integrals by sums, the Jackson integrals.

Let $\int_\sharp$ be the integration for the path which begins at $z = \epsilon i + \infty$, moves to the left till ϵi, then down through the origin to $-\epsilon i$, and then returns down the positive real axis to $-\epsilon i + \infty$ (for small ϵ). Then for $|\Im k| < 2\epsilon, \Re k > 0$,

$$\frac{1}{2i}\int_\sharp q^{x^2}\delta_k\,\mathrm{d}x \;= -\frac{a\pi}{2}\prod_{j=0}^{\infty}\frac{(1-q^{j+k})(1-q^{j-k})}{(1-q^{j+2k})(1-q^{j+1})}\times\mathfrak{g}_q^\sharp,$$

$$\mathfrak{g}_q^\sharp(k) \stackrel{def}{=} \sum_{j=0}^{\infty} q^{\frac{(k-j)^2}{4}}\frac{1-q^{j+k}}{1-q^k}\prod_{l=1}^{j}\frac{1-q^{l+2k-1}}{1-q^l} \;= \tag{9}$$

$$q^{\frac{k^2}{4}}\prod_{j=1}^{\infty}\frac{(1-q^{j/2})(1-q^{j+k})(1+q^{j/2-1/4+k/2})(1+q^{j/2-1/4-k/2})}{(1-q^j)}.$$

The sum $\mathfrak{g}_q^\sharp$ is the Jackson integral for a special choice $(k/2)$ of the starting point. The convergence of the sum (9) is for all k. Similarly,

$$\mathfrak{Z}_q^\sharp(k) \stackrel{def}{=} -\frac{a\pi}{2}\prod_{j=0}^{\infty}\frac{(1-q^{j+k})(1-q^{j-k})}{(1-q^{j+2k})(1-q^{j+1})}\times\mathfrak{z}_q^\sharp,$$

$$\mathfrak{z}_q^\sharp(k) \;=\; \sum_{j=0}^{\infty} q^{-kj}(q^{-\frac{(k+j)^2}{4}}+1)^{-1}\frac{1-q^{j+k}}{1-q^k}\prod_{l=1}^{j}\frac{1-q^{l+2k-1}}{1-q^l}. \tag{10}$$

For all k apart from the poles, $\lim_{a\to\infty}(\frac{a}{4})^{k-1/2}\mathfrak{Z}_q^\sharp(k) = \sin(\pi k)\mathfrak{Z}(k)$.

Numerically, it is likely that all zeros of $\mathfrak{Z}_q^\sharp$ in the strip

$$\{0 \le \Im k < \sqrt{2\pi a} - \epsilon,\ \Re k > -1/2\} \ \text{ for } a > 2/\pi \text{ and small } \epsilon$$

are deformations of the classical ones. Moreover there is a strong tendency for the deformations of the zeros of the $\zeta(k+1/2)$-factor to go to the right (big a). They are not expected in the left half-plane before $k = 1977.2714i$ (see [C13]).

When $q = \exp(2\pi i/N)$ and k is a positive integer $\le N/2$ we come to the Gauss-Selberg-type sums:

$$\sum_{j=0}^{N-2k} q^{\frac{(k-j)^2}{4}}\frac{1-q^{j+k}}{1-q^k}\prod_{l=1}^{j}\frac{1-q^{l+2k-1}}{1-q^l} = \prod_{j=1}^{k}(1-q^j)^{-1}\sum_{m=0}^{2N-1} q^{m^2/4}. \tag{11}$$

They resemble, for instance, [E,(1.2b)]. Substituting $k = [N/2]$ we arrive at (3).

Double Hecke algebras provide justifications and generalizations. In the A_1-case, $\mathcal{H}\mathcal{H} \stackrel{def}{=} \mathbf{C}[\mathcal{B}_q]/((T-t^{1/2})(T+t^{-1/2}))$ for the group algebra of the group $\mathcal{B}_q$ generated by $T, X, Y, q^{1/4}$ with the relations

$$TXT = X^{-1},\ T^{-1}YT^{-1} = Y^{-1},\ Y^{-1}X^{-1}YXT^2 = q^{-1/2} \tag{12}$$

for central $q^{1/4}, t^{1/2}$. Renormalizing $T \to q^{-1/4}T,\ X \to q^{1/4}X,\ Y \to q^{-1/4}Y$,

$$\mathcal{B}_q \cong \mathcal{B}_1 \mod q^{1/4},\ \mathcal{B}_1 \cong \pi_1(\{E \times E \setminus \mathrm{diag}\}/\mathbf{S}_2),\ E = \text{elliptic curve}, \quad (13)$$

a special case of the calculation from [B]. The T is the half-turn about the diagonal, X, Y correspond to the "periods" of E.

Thanks to the topological interpretation, the central extension $PSL_2^c(\mathbf{Z})$ of $PSL_2(\mathbf{Z})$ (Steinberg) acts on $\mathcal{B}_1$ and $\mathcal{HH}$. The automorphisms corresponding to the generators $\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$, $\left(\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right)$ are as follows:

$$\tau_+ : Y \to q^{-1/4}XY,\ X \to X, \quad \tau_- : X \to q^{1/4}YX,\ Y \to Y, \quad (14)$$

fixing T, q, t. When $t = 1$ we get the well-known action of $SL_2(\mathbf{Z})$ on the Weyl and Heisenberg algebras (the latter as $q \to 1$). Formally, τ_+ is the conjugation by q^{x^2} for X represented here and later in the form $X = q^x$.

The Macdonald nonsymmetric polynomials are eigenfunctions of Y in the following $\mathcal{HH}$-representation in the space $\mathcal{P}$ of the Laurent polynomials of q^x :

$$T \to t^{1/2}s + (q^{2x} - 1)^{-1}(t^{1/2} - t^{-1/2})(s - 1),\ Y \to spT \quad (15)$$

for the reflection $sf(x) = f(-x)$ and the translation $pf(x) = f(x + 1/2)$. It is nothing else but the representation of $\mathcal{HH}$ induced from the character $\chi(T) = t^{1/2} = \chi(Y)$. The Fourier transform (on the generalized functions) is associated with the anti-involution $\{\varphi : X \to Y^{-1} \to X\}$ of $\mathcal{HH}$ preserving T, t, q.

Combining τ_+ and φ, we prove that the Macdonald polynomials multiplied by q^{-x^2} are eigenfunctions of the q-Fourier transform and get (6) for $t = q^k$.

When q, k are from (11), let $q^x(m/2) = q^{m/2}$ for $m \in \mathbf{Z},\ -N < m \le N$, and

$$\bowtie \stackrel{def}{=} \{m \mid \mu_k(m/2) \ne 0\} = \{-N + k + 1, \dots, -k,\ k + 1, \dots, N - k\}.$$

The space $V_k = \mathrm{Funct}(\bowtie)$ has a unique structure of an (irreducible) $\mathcal{HH}$-module making the evaluation map $\mathcal{P} \ni f \mapsto f(m/2) \in V_k$ a $\mathcal{HH}$-homomorphism. Setting $V_k = V_k^+ \oplus V_k^-$ where $T = \pm t^{\pm 1/2}$ on $V_k^\pm$, the dimensions for $k < N/2$ are $2(N - 2k) = (N - 2k + 1) + (N - 2k - 1)$. The components $V_k^\pm$ are $PSL_2^c(\mathbf{Z})$-invariant. Calculating its action in V_k^+ (which is a subalgebra of V_k) we come to the formulas from [Ki,C2,C3]; V_k^- is $PSL_2^c(\mathbf{Z})$-isomorphic to V_{k+1}^+. For $k = 1$ it is the Verlinde algebra. Involving the the shift operator, we get (11).

We note that V_k may have applications to the arithmetic theory of coverings of elliptic curves ramified at one point thanks to (13).

2. General results.

Let $R = \{\alpha\} \subset \mathbf{R}^n$ be a root system of type $A, B, ..., F, G$ with respect to a euclidean form (z, z') on $\mathbf{R}^n \ni z, z'$, W the Weyl group generated by the reflections s_α, $\alpha_1, ..., \alpha_n$ simple roots, R_+ the set of positive roots, $\omega_1, ..., \omega_n$ the fundamental weights, $Q = \oplus_{i=1}^n \mathbf{Z}\alpha_i \subset P = \oplus_{i=1}^n \mathbf{Z}\omega_i$. We will also use coroots $\alpha^\vee = 2\alpha/(\alpha, \alpha)$

and the corresponding $Q^\vee$. The form will be normalized by the condition $(\theta,\theta)=2$ for the maximal coroot $\theta \in R_+^\vee$.

The affine Weyl group $\widetilde{W}$ acts on $\tilde{z}=[z,\zeta]\in \mathbf{R}^n\times\mathbf{R}$ and is generated by $s_i=s_{\alpha_i}$ and $s_0(\tilde{z}) = \tilde{z}+(z,\theta)\alpha_0$, for $\alpha_0=[-\theta,1]$. Setting $b(\tilde{z})=[z,\zeta-(z,b)]$ for $b\in P$, $\widetilde{W}=W\ltimes Q\subset \widehat{W} \stackrel{def}{=} W\ltimes P$. We call the latter the *extended affine Weyl group*. It is generated over $\widetilde{W}$ by the group $\pi\in\Pi\cong P/Q$ such that π leave the set $\alpha_0,\alpha_1^\vee,\dots,\alpha_n^\vee$ invariant.

The length $l(\hat{w})$ of $\hat{w}=\pi\tilde{w}\in W^b,\ \pi\in\Pi,\tilde{w}\in W^a$ is by definition the length of the reduced decomposition of $\tilde{w}$ in terms of the simple reflections $s_i, 0\le i\le n$. Given $b\in P$, there is a unique decomposition

$$b=\pi_b w_b \ \text{ such that } \ w_b\in W,\ l(b)=l(\pi_b)+l(w_b) \text{ and } \ l(w_b)=\max. \quad (16)$$

Then $\Pi=\{\pi_{\omega_r}\}$ for the minuscule ω_r: $(\omega_r,\alpha^\vee)\le 1$ for all $\alpha\in R_+$.

DOUBLE HECKE ALGEBRAS. Let $q_\alpha=q^{(\alpha,\alpha)/2}, t_\alpha=q_\alpha^{k_\alpha}$ for $\{k_\alpha\}$ such that $k_{w(\alpha)}=k_\alpha$ (all w), $t_i=t_{\alpha_i}$, $t_0=t_\theta$, $\rho_k=(1/2)\sum_{\alpha\in R_+}k_\alpha\alpha$,

$$X_{\tilde{b}} \ = \ \prod_{i=1}^n X_i^{l_i}q^l \ \text{ if } \ \tilde{b}=[b,l],\ b=\sum_{i=1}^n l_i\omega_i\in P,\ l\in(P,P)=(1/p)\mathbf{Z}$$

for $p\in\mathbf{N}$. By $\mathbf{C}_{q,t}^{\pm}[X]$ we mean the algebra of polynomials in terms of $X_i^{\pm 1}$ over the field $\mathbf{C}_{q,t}$ of rational functions of $q^{1/(2p)}, t_\alpha^{1/2}$. We will also use the evaluation $X_b(q^z)\stackrel{def}{=}q^{(b,z)}$.

The *double affine Hecke algebra* $\mathcal{H}$ is generated over the field $\mathbf{C}_{q,t}$ by the elements $\{T_j,\ 0\le j\le n\}$, pairwise commutative $\{X_i\}$, and the group Π where the following relations are imposed:

(o) $(T_j-t_j^{1/2})(T_j+t_j^{-1/2}) \ = \ 0, \ \ 0 \ \le \ j \ \le \ n$;
(i) $T_iT_jT_i... \ = \ T_jT_iT_j..., \ \ m_{ij}$ factors on each side;
(ii) $\pi T_i\pi^{-1}=T_j,\ \pi X_b\pi^{-1}=X_{\pi(b)}$ if $\pi\in\Pi,\ \pi(\alpha_i^\vee)=\alpha_j^\vee$;
(iii) $T_iX_bT_i \ = \ X_bX_{a_i}^{-1}$ if $(b,\alpha_i^\vee)=1,\ 1\le i\le n$;
(iv) $T_0X_bT_0 \ = \ X_{s_0(b)} \ = \ X_bX_\theta q^{-1}$ if $(b,\theta)=-1$;
(v) $T_iX_b \ = \ X_bT_i$ if $(b,\alpha_i^\vee)=0$ for $0\le i\le n$.

Here m_{ij} are from the corresponding Coxeter relations. Given $\tilde{w}\in\widetilde{W},\ \pi\in\Pi$, the product $T_{\pi\tilde{w}}\stackrel{def}{=}\pi T_{i_1}\cdots T_{i_l}$, where $\tilde{w}=s_{i_1}\cdots s_{i_l}$, $l=l(\tilde{w})$, does not depend on the choice of the reduced decomposition. In particular, we arrive at the pairwise commutative elements

$$Y_b \ = \ \prod_{i=1}^n Y_i^{l_i} \ \text{ if } \ b=\sum_{i=1}^n l_i\omega_i\in P, \ \text{ where } \ Y_i\stackrel{def}{=}T_{\omega_i}, \quad (17)$$

satisfying the relations $T_i^{-1}Y_bT_i^{-1} \ = \ Y_bY_{a_i}^{-1}$ if $(b,\alpha_i^\vee) \ = \ 1$, $\ T_iY_b \ = \ Y_bT_i$ if $(b,\alpha_i^\vee)=0,\ 1\le i\le n$.

The Fourier transform is related to the anti-involution of $\mathcal{HH}$

$$\varphi:\ X_i \to Y_i^{-1},\ \ Y_i \to X_i^{-1},\ \ T_i \to T_i,\ t \to t,\ q \to q,\ 1 \le i \le n. \tag{18}$$

The "unitary" representations are defined for the anti-involution

$$X_i^* = X_i^{-1},\ Y_i^* = Y_i^{-1},\ T_i^* = T_i^{-1},\ t \to t^{-1},\ q \to q^{-1},\ 0 \le i \le n.$$

The next two automorphisms induce a projective action of $PSL_2(\mathbf{Z})$:

$$\begin{aligned}&\tau_+:\ X_b \to X_b,\ \ Y_r \to X_r Y_r q^{-(\omega_r,\omega_r)/2},\ Y_\theta \to X_0^{-1}T_0^{-2}Y_\theta,\\ &\tau_-:\ Y_b \to Y_b,\ \ X_r \to Y_r X_r q^{(\omega_r,\omega_r)/2},\ X_\theta \to T_0 X_0 Y_\theta^{-1} T_0,\end{aligned} \tag{19}$$

where $b \in P$, ω_r are minuscule, $X_0 = qX_\theta^{-1}$. Obviously $\tau_- = \varphi\tau_+\varphi$. The projectivity means that $\tau_+^{-1}\tau_-\tau_+^{-1} = \tau_-\tau_+^{-1}\tau_-$.

POLYNOMIAL REPRESENTATION. Let $\hat{w}(X_{\tilde{b}}) = X_{\hat{w}(\tilde{b})}$ for $\hat{w} \in \widehat{W}$. Combining the action of the group Π, the multiplication by X_b, and the *Demazure-Lusztig operators*

$$T_j = t_j^{1/2}s_j + (t_j^{1/2} - t_j^{-1/2})(X_{\alpha_j} - 1)^{-1}(s_j - 1),\ 0 \le j \le n, \tag{20}$$

we get a representation of $\mathcal{HH}$ in $\mathbf{C}_{q,t}^{\pm}[X]$.

The coefficient of $X^0 = 1$ (*the constant term*) of a polynomilal $f \in \mathbf{C}_{q,t}^{\pm}[X]$ will be denoted by $\langle f \rangle$. Let

$$\mu = \prod_{a \in R_+^\vee} \prod_{i=0}^{\infty} \frac{(1 - X_\alpha q_\alpha^i)(1 - X_\alpha^{-1} q_\alpha^{i+1})}{(1 - X_\alpha t_\alpha q_\alpha^i)(1 - X_\alpha^{-1} t_\alpha q_\alpha^{i+1})}. \tag{21}$$

We will consider μ as a Laurent series with the coefficients in $\mathbf{C}[t][[q]]$. The form $\langle \mu_0 f g^* \rangle$ makes the polynomial representation unitary for

$$X_b^* = X_{-b},\ t^* = t^{-1},\ q^* = q^{-1},\ \mu_0 = \mu_0/\langle\mu\rangle = \mu_0^*.$$

The *Macdonald nonsymmetric polynomials* $\{e_b, b \in P\}$ are eigenvectors of the operators $\{L_f \stackrel{def}{=} f(Y_1, \cdots, Y_n), f \in \mathbf{C}_{q,t}^{\pm}[X]\}$:

$$L_f(e_b) = f(q^{-b_\sharp})e_b, \ \text{ where } \ b_\sharp \stackrel{def}{=} b - w_b^{-1}(\rho_k) \ \text{ for } \ w_b \text{ from (16).} \tag{22}$$

They are pairwise orthogonal with respect to the above pairing and form a basis in $\mathbf{C}_{q,t}^{\pm}[X]$. The normalization $\epsilon_b \stackrel{def}{=} e_b/e_b(q^{-\rho_k})$ is the most convenient in the harmonic analysis. For instance, the duality relations become especially simple: $\epsilon_b(q^{c_\sharp}) = \epsilon_c(q^{b_\sharp})$ for all $b, c \in P$. The next formula establishes that ϵ_c multiplied by the Gaussian are eigenfunctions of the difference Fourier transform:

$$\begin{aligned}&\langle \epsilon_b \epsilon_c^* \tilde{\gamma}^{-1}\mu\rangle = q^{(b_\sharp,b_\sharp)/2+(c_\sharp,c_\sharp)/2-(\rho_k,\rho_k)}\epsilon_c^*(q^{b_\sharp})\times\\ &\prod_{\alpha\in R_+}\prod_{j=1}^{\infty}\frac{1-q_\alpha^{(\rho_k,\alpha^\vee)+j}}{1-t_\alpha q_\alpha^{(\rho_k,\alpha^\vee)+j}} \quad \text{for } \ \tilde{\gamma}^{-1} \stackrel{def}{=} \sum_{b\in P} q^{(b,b)/2}X_b.\end{aligned} \tag{23}$$

When $b = c = 0$ we get (5). Indeed, the series for $\tilde{\gamma}^{-1}$ is nothing else but the expansion of γ^{-1} for $\gamma = q^{x^2/2}$, where we set $X_b = q^{x_b},\ x^2 = \Sigma_{i=1}^n x_{\omega_i} x_{\alpha_i^\vee}$.

JACKSON AND GAUSS SUMS. We fix generic $\xi \in \mathbf{C}^n$ and set $\langle f \rangle_\xi \stackrel{def}{=} |W|^{-1} \sum_{w \in W, b \in B} f(q^{w(\xi)+b})$. Here f is a Laurent polynomial or any function well-defined on $\{q^{w(\xi)+b}\}$. We assume that $|q| < 1$. For instance, $\langle \gamma \rangle_\xi = \tilde{\gamma}^{-1}(q^\xi) q^{(\xi,\xi)/2}$. It is convenient to switch to $\mu^\circ(X,t) \stackrel{def}{=} \mu^{-1}(X, t^{-1})$. Given $b, c \in P$,

$$\langle \epsilon_b\, \epsilon_c^*\, \gamma \mu^\circ \rangle_\xi = q^{-(b_\sharp, b_\sharp)/2 - (c_\sharp, c_\sharp)/2 + (\rho_k, \rho_k)} \epsilon_c(q^{b_\sharp}) \times$$
$$|W|^{-1} \langle \gamma \rangle_\xi \prod_{\alpha \in R_+} \prod_{j=0}^{\infty} \frac{1 - t_\alpha^{-1} q_\alpha^{-(\rho_k, \alpha^\vee)+j}}{1 - q_\alpha^{-(\rho_k, \alpha^\vee)+j}}. \tag{24}$$

For $\xi = -\rho_k$, (24) generalizes (9). If $k \in \mathbf{Z}_+$, then $\mu^\circ = q^{\mathrm{const}} \mu \in \mathbf{C}_q^\pm[X]$, the product in (24) is understood as the limit and becomes finite.

The proof of this formula and the previous one is based on the analysis of the anti-involution (18) in the corresponding representations of $\mathcal{H}\!\!\mathcal{H}$. Here it is the representation in $\mathcal{F} = \mathrm{Funct}(\widehat{W}, \mathbf{C}_{q,t}(q^{(\omega_i, \xi)}))$. For $a, b \in P$, $w \in W, \hat{v} \in \widehat{W}$, we set

$$X_a(bw) = X_a(q^{b+w(\xi)}),\ X_a g(bw) = (X_a g)(bw),\ \hat{v}(g)(bw) = g(\hat{v}^{-1} bw)$$

for $g \in \mathcal{F}$. It provides the action of X, Π. The T act as follows:

$$T_i(g)(\hat{w}) = \frac{t_i^{1/2} q^{(\alpha_i, b+w(\xi))} - t_i^{-1/2}}{q^{(\alpha_i, b+w(\xi))} - 1}\, g(s_i \hat{w})$$
$$- \frac{t_i^{1/2} - t_i^{-1/2}}{q^{(\alpha_i, b+w(\xi))} - 1}\, g(\hat{w}) \quad \text{for} \quad 0 \le i \le n, \tag{25}$$

The formulas are closely connected with (20): the natural evaluation map from $\mathbf{C}_{q,t}^\pm[X]$ to $\mathcal{F}$ is a $\mathcal{H}\!\!\mathcal{H}$-homomorphism. The unitarity is for $\langle \mu_1 f g^* \rangle_\xi$, where the values of $\mu_1 = \mu/\mu(q^\xi) = \mu_1^\circ$ at $\hat{w}$ are $*$-invariant ($\xi^* = \xi$).

Dropping the X-action, we get a deformation of the regular representation of the affine Hecke algebra generated by T, Π. Indeed, taking ξ from the dominant affine Weyl chamber, (25) tend to the p-adic formulas from [Mat] when $q \to \infty$ and t are powers of p. For $\xi = -\rho_k$, the image of the restriction map from $\mathcal{F}$ to functions on the set $\{\pi_b, b \in P\}$, which is a $\mathcal{H}\!\!\mathcal{H}$-homomorphism, generalizes the spherical part of the regular representation. The limit to the Harish-Chandra theory is $q \to 1$ where k is fixed (the root multiplicity). See [He,C11].

Now q will be a primitive N-th root of unity, $P_N = P/(P \cap NQ^\vee)$; the evaluations of Laurent polynomials are functions on this set. Let $\langle f \rangle_N \stackrel{def}{=} \sum_{b \in P_N} f(q^b)$. We assume that $k_\alpha \in \mathbf{Z}_+$ for all $\alpha \in R$ and $\mu(q^{-\rho_k}) \neq 0$. We also pick q to ensure the existence of the Gaussian: $q^{(b,b)/2} = 1$ for all $b \in P \cap NQ^\vee$. It means that when N is odd and the root system is either B or C_{4l+2} one takes $q = \exp(4\pi i m/N)$ for $(m, N) = 1, 0 < 2m < N$. Otherwise it is arbitrary.

We claim that the formula (24) holds for $\langle\ \rangle_N$ instead of $\langle\ \rangle_\xi$ provided the existence of the nonsymmetric polynomials. It readily gives (11) for $b = 0 = c$.

Given $b' \in P_N$ such that $\mu(q^{b'}) \neq 0$, at least one ϵ_b exists with $b_\sharp$ equal to b' in P_N. Denoting the set of all such b' by P'_N, the space $\mathrm{Funct}(P'_N, \mathbf{Q}(q^{1/(2p)}))$ is an algebra and a $\mathcal{H}$-module isomorphic to the quotient of the polynomial representation by the radical of the pairing $\langle \mu f g^* \rangle_N$. The radical also coincides with the set of polynomials f such that $(g(Y)(f))(q^{-\rho_k}) = 0$ for all Laurent polynomials g. The evaluations of ϵ_b depend only on the images of $b_\sharp$ in P_N and form a basis of this module. The evaluations of the symmetric polynomials constitute the *generalized Verlinde algebra*.

Acknowlegments. Partially supported by NSF grant DMS-9622829 and the Guggenheim Fellowship. The paper was completed at the University Paris 7, the author is grateful for the kind invitation.

REFERENCES.

[B] Birman, J.: On braid groups. Communs on Pure and Appl. Math. **22**, 41–72 (1969).
[C1] Cherednik, I.: Double affine Hecke algebras and Macdonald's conjectures. Annals of Math. **141**, 191–216 (1995).
[C2] —: Macdonald's evaluation conjectures and difference Fourier transform. Invent. Math. **122**, 119–145 (1995).
[C3] —: Nonsymmetric Macdonald polynomialas. IMRN **10**, 483–515 (1995).
[C4] —: Intertwining operators of double affine Hecke algebras. Selecta Math. New s. **3**, 459–495 (1997).
[C5] —: Integration of Quantum many-body problems by affine Knizhnik-Zamolodchikov equations. Advances in Math. **106**, 65–95 (1995).
[C6] —: Double affine Hecke algebras, Knizhnik–Zamolodchikov equations, and Macdonald's operators. IMRN (Duke Math. J.) **9**, 171–180 (1992).
[C7] —: Elliptic quantum many-body problem and double affine Knizhnik - Zamolodchikov equation. Communs Math. Phys. **169**:2, 441–461 (1995).
[C8] —: Difference-elliptic operators and root systems, IMRN **1**, 43–59 (1995).
[C9] —: Lectures on Knizhnik-Zamolodchikov equations and Hecke algebras. MSJ Memoirs **1** (1998).
[C10] —: Difference Macdonald-Mehta conjectures. IMRN **10**, 449–467 (1997).
[C11] —: Inverse Harish–Chandra transform and difference operators. IMRN **15**, 733–750 (1997).
[C12] —: A new interpretation of Gelfand-Tzetlin bases. Duke Math. J. **54**:2, 563–577 (1987).
[C13] —: On q-analogues of Riemann's zeta. Preprint (1998).
[DS] Diejen, J.F. van, Stockman, J.V.: Multivariable q-Racah polynomials. Duke Math. J. **91**, 89–136 (1998).
[Dr] Drinfeld, V.G.: Degenerate affine Hecke algebras and Yangians. Funct. Anal. and Appl. **20**, 69–70 (1986).
[D1] Dunkl, C.F.: Diffirential-difference operators associated to reflection groups. Trans. AMS. **311**, 167–183 (1989).
[D2] —: Hankel transforms associated to finite reflection groups. Contemp. Math. **138**, 123–138 (1992).
[E] Evans, R.J.: The evaluation of Selberg character sums. L'Enseignement Math. **37**, 235–248 (1991).
[GG] Garland, H., Grojnowski, I.: Affine Hecke algebras associated to Kac-Moody groups. Preprint (1995).
[GKV] Ginzburg, V., Kapranov, M., Vasserot, E.: Residue construction of Hecke algebras. Preprint (1995).

[H] Heckman, G.J.: An elementary approach to the hypergeometric shift operator of Opdam. Invent. math. **103** 341–350 (1991).
[HO1] Heckman, G.J., Opdam, E.M.: Harmonic analysis for affine Hecke algebras. Preprint (1996).
[HO2] —: Root systems and hypergeometric functions I. Comp. Math. **64**, 329–352 (1987).
[He] Helgason, S.: Groups and geometric analysis. Academic Press, New York (1984).
[I] Ito, M.: On a theta product formula for Jackson integrals associated with root systems of rank two. Preprint (1996).
[J] Jeu, M.F.E. de: The Dunkl transform. Invent. Math. **113**, 147–162 (1993).
[K] Kac, V.G.: Infinite dimensional Lie algebras. Cambridge University Press, Cambridge (1990).
[Ki] Kirillov, A. Jr.: Inner product on conformal blocks and Macdonald's polynomials at roots of unity. Preprint (1995).
[KL1] Kazhdan, D., Lusztig, G.: Proof of the Deligne-Langlands conjecture for Hecke algebras. Invent. Math. **87**, 153–215 (1987).
[KL2] —: Tensor structures arising from affine Lie algebras. III. J. of AMS **7**, 335–381 (1994).
[KS] Knop, F., Sahi, S.: A recursion and a combinatorial formula for Jack polynomials, Preprint (1996), to appear in Invent. Math.
[KK] Kostant, B., Kumar, S.: T-Equivariant K-theory of generalized flag varieties. J. Diff. Geometry **32**, 549–603 (1990).
[Le] Lek, H. van der: Extended Artin groups. Proc. Symp. Pure Math. **40**:2, 117–122 (1981).
[L] Lusztig, G.: Affine Hecke algebras and their graded version. J. of the AMS **2**:3, 599–685 (1989).
[M1] Macdonald, I.G.: Some conjectures for root systems. SIAM J. Math. An. **13**, 988–1007 (1982).
[M2] —: Orthogonal polynomials associated with root systems, Preprint(1988).
[M3] —: Affine Hecke algebras and orthogonal polynomials. Séminaire Bourbaki **47**:797, 01–18 (1995).
[M4] —: A formal identity for affine root systems, Preprint (1996).
[Ma] Matsuo, A.: Knizhnik-Zamolodchikov type equations and zonal spherical functions. Invent. Math. **110**, 95–121 (1992).
[Mat] Matsumoto, H.: Analyse harmonique dans les systemes de Tits bornologiques de type affine. Lecture Notes in Math. **590** (1977).
[O1] Opdam, E.M.: Some applications of hypergeometric shift operators. Invent. Math. **98**, 1–18 (1989).
[O2] —: Harmonic analysis for certain representations of graded Hecke algebras. Acta Math. **175**, 75–121 (1995).
[O3] —: Dunkl operators, Bessel functions and the discriminant of a finite Coxeter group, Comp. Math. **85**, 333–373 (1993).
[Sa] Sahi, S.: Nonsymmetric Koornwinder polynomials and duality. Preprint (1996), to appear in Annals of Math.
[STU] Saito, Y., Takemura, K., Uglov, D.: Toroidal actions on level-1 modules of $U_q(s\hat{l}_n)$. Transformation Groups **3**, 75–102 (1998).
[VV] Varagnolo, M., Vasserot, E.: Double-loop algebras and the Fock space. Preprint (1996), to appear in Invent.Math.

Ivan V. Cherednik
Dept. Math. UNC at Chapel Hill,
Chapel Hill, NC 27599-3250, USA
chered@math.unc.edu

Counting Problems and Semisimple Groups

Alex Eskin[1]

Abstract. Some natural counting problems admit extra symmetries related to actions of Lie groups. For these problems, one can sometimes use ergodic and geometric methods, and in particular the theory of unipotent flows, to obtain asymptotic formulas.

We will present counting problems related to diophantine equations, diophantine inequalities and quantum chaos, and also to the study of billiards on rational polygons.

1991 Mathematics Subject Classification: Primary 11J25, 22E40

1 Counting lattice points on affine homogeneous varieties

In [EMS2], using ergodic properties of subgroup actions on homogeneous spaces of Lie groups, we study asymtotic behaviour of number of lattice points on certain affine varieties. Consider for instance the following:

Let $p(\lambda)$ be a monic polynomial of degree $n \geq 2$ with integer coefficients and irreducible over $\mathbb{Q}$. Let $M_n(\mathbb{Z})$ denote the set of $n \times n$ integer matrices, and put

$$V_p(\mathbb{Z}) = \{A \in M_n(\mathbb{Z}) \ : \ \det(\lambda I - A) = p(\lambda)\}.$$

Hence $V_p(\mathbb{Z})$ is the set of integral matrices with characteristic polynomial $p(\lambda)$. Consider the norm on $n \times n$ real matrices given by $\|(x_{ij})\| = \sqrt{\sum_{ij} x_{ij}^2}$, and let $N(T, V_p)$ denote the number of elements of $V_p(\mathbb{Z})$ with norm less than T.

Theorem 1.1 *Suppose further that $p(\lambda)$ splits over $\mathbb{R}$, and for a root α of $p(\lambda)$ the ring of algebraic integers in $\mathbb{Q}(\alpha)$ is $\mathbb{Z}[\alpha]$. Then, asymptotically as $T \to \infty$,*

$$N(T, V_p) \sim \frac{2^{n-1} h R \omega_n}{\sqrt{D} \cdot \prod_{k=2}^{n} \Lambda(k/2)} T^{n(n-1)/2}$$

where h is the class number of $\mathbb{Z}[\alpha]$, R is the regulator of $\mathbb{Q}(\alpha)$, D is the discriminant of $p(\lambda)$, ω_n is the volume of the unit ball in $\mathbb{R}^{n(n-1)/2}$, and $\Lambda(s) = \pi^{-s}\Gamma(s)\zeta(2s)$.

[1]Supported by the Sloan Foundation and the Packard Foundation

Example 1 is a special case of the following counting problem which was first studied in [DRS] and [EMc].

THE COUNTING PROBLEM: Let W be a real finite dimensional vector space with a $\mathbb{Q}$ structure and V a Zariski closed real subvariety of W defined over $\mathbb{Q}$. Let G be a reductive real algebraic group defined over $\mathbb{Q}$, which acts on W via a $\mathbb{Q}$-representation $\rho : G \to \mathrm{GL}(W)$. Suppose that G acts transitively on V. Let $\|\cdot\|$ denote a Euclidean norm on W. Let B_T denote the ball of radius $T > 0$ in W around the origin, and define

$$N(T, V) = |V \cap B_T \cap \mathbb{Z}^n|,$$

the number of integral points on V with norm less than T. We are interested in the asymptotics of $N(T, V)$ as $T \to \infty$.

We use the rich theory of unipotent flows on homogeneous spaces developed in [Mar2], [DM1], [Rat1], [Rat2], [Rat3], [Rat4], [Sha1] and [DM3] to obtain results in this direction.

Let Γ be a subgroup of finite index in $G(\mathbb{Z})$ such that $\Gamma W(\mathbb{Z}) \subset W(\mathbb{Z})$. By a theorem of Borel and Harish-Chandra [BH-C], $V(\mathbb{Z})$ is a union of finitely many Γ-orbits. Therefore to compute the asymptotics of $N(T, V)$ it is enough to consider each Γ-orbit, say $\mathcal{O}$, separately and compute the asymptotics of

$$N(T, V, \mathcal{O}) = |\mathcal{O} \cap B_T|.$$

Suppose that $\mathcal{O} = \Gamma v_0$ for some $v_0 \in V(\mathbb{Z})$. Then the stabilizer $H = \{g \in G : gv_0 = v_0\}$ is a reductive real algebraic $\mathbb{Q}$-subgroup, and $V \cong G/H$. Define

$$R_T = \{gH \in G/H : gv_0 \in B_T\},$$

the pullback of the ball B_T to G/H.

Assume that G^0 and H^0 do not admit nontrivial $\mathbb{Q}$-characters. Then by the theorem of Borel and Harish-Chandra, G/Γ admits a G-invariant (Borel) probability measure, say μ_G, and $H/(\Gamma \cap H)$ admits an H-invariant probability measure, say μ_H. Now the natural inclusion $H/(\Gamma \cap H) \hookrightarrow G/\Gamma$ is an H-equivariant proper map. Let $\pi : G \to G/\Gamma$ be the natural quotient map. Then the orbit $\pi(H)$ is closed, $H/(\Gamma \cap H) \cong \pi(H)$, and μ_H can be treated as a measure on G/Γ supported on $\pi(H)$. Such finite invariant measures supported on closed orbits of subgroups are called *homogeneous measures*. Let $\lambda_{G/H}$ denote the (unique) G-invariant measure on G/H induced by the normalization of the Haar measures on G and H.

To state our result in the general setting, we need another definition:

DEFINITION 1.2 Let G and H be as in the counting problem. For a sequence $T_n \to \infty$, the sequence $\{R_{T_n}\}$ of open sets in G/H is said to be *focused*, if there exist a proper connected reductive real algebraic $\mathbb{Q}$-subgroup L of G containing H^0 and a compact set $C \subset G$ such that

$$\limsup_{n\to\infty} \frac{\lambda_{G/H}(q_H(CL(Z(H^0) \cap \Gamma)) \cap R_{T_n})}{\lambda_{G/H}(R_{T_n})} > 0,$$

where $q_H : G \to G/H$ is the natural quotient map.

Our main counting result is the following:

Theorem 1.3 *Let G and H be as in the counting problem. Suppose that H^0 is not contained in any proper $\mathbb{Q}$-parabolic subgroup of G^0 (equivalently, $Z(H)/(Z(H)\cap\Gamma)$ is compact), and for some sequence $T_n \to \infty$ with bounded gaps, the sequence $\{R_{T_n}\}$ is not focused. Then asymptotically*

$$N(T, V, \mathcal{O}) \sim \lambda_{G/H}(R_T).$$

For the case when H is an affine symmetric subgroup of G this result was proved previously in [DRS] using harmonic analysis; subsequently a simpler proof using the mixing property of the geodesic flow appeared in [EMc]. We note that focusing cannot occur for the affine symmetric case.

In general, focusing does not seem to occur for most natural examples, even though it does happen; see [EMS2] for an example.

Translates of homogeneous measures. The following theorem is the main ergodic theoretic result which allows us to investigate the counting problems. The result is also of general interest, especially from the view point of ergodic theory on homogeneous spaces of Lie groups.

Theorem 1.4 *Let G be a connected real algebraic group defined over $\mathbb{Q}$, $\Gamma \subset G(\mathbb{Q})$ an arithmetic lattice in G with respect to the $\mathbb{Q}$-structure on G, and $\pi : G \to G/\Gamma$ the natural quotient map. Let $H \subset G$ be a connected real algebraic $\mathbb{Q}$-subgroup admitting no nontrivial $\mathbb{Q}$-characters. Let μ_H denote the H-invariant probability measure on the closed orbit $\pi(H)$. For a sequence $\{g_i\} \subset G$, suppose that the translated measures $g_i\mu_H$ converge to a probability measure μ on G/Γ. Then there exists a connected real algebraic $\mathbb{Q}$-subgroup L of G containing H such that the following holds:*

(i) There exists $c_0 \in G$ such that μ is a $c_0Lc_0^{-1}$-invariant measure supported on $c_0\pi(L)$.

In particular, μ is a homogeneous measure.

(ii) There exist sequences $\{\gamma_i\} \subset \Gamma$ and $c_i \to c_0$ in G such that $\gamma_i H \gamma_i^{-1} \subset L$ and $g_i H = c_i\gamma_i H$ for all but finitely many i.

In order to be able to apply Theorem 1.4 to the problem of counting, we need to know some conditions under which the sequence $\{g_i\mu_H\}$ of probability measures does not escape to infinity. A neccesary and sufficient condition on the reductive $\mathbb{Q}$-group H is given in [EMS1].

The proof of Theorem 1.4 is based on Ratner's measure classification theorem. A key observation is that the limit measure μ is invariant under some unipotent element. Still the result does not follow immediately from Ratner's theorem since we do not know that μ is ergodic. In the case when H is itself generated by unipotent elements, the proof is simpler: see [MS].

Correspondence between counting and translates of measures. We recall some observations from [DRS, Sect. 2]; see also [EMc]. Let the notation be

as in the counting problem stated in the introduction. For $T > 0$, define a function F_T on G by

$$F_T(g) = \sum_{\gamma \in \Gamma/(H \cap \Gamma)} \chi_T(g\gamma v_0),$$

where χ_T is the characteristic function of B_T. By construction F_T is left Γ-invariant, and hence it will be treated as a function on G/Γ. Note that

$$F_T(e) = \sum_{\gamma \in \Gamma/(H \cap \Gamma)} \chi_T(\gamma v_0) = N(T, V, \mathcal{O}).$$

Since we expect, as in Theorem 1.3, that $N(T, V, \mathcal{O}) \sim \lambda_{G/H}(R_T)$, we define $\hat{F}_T(g) = \frac{1}{\lambda_{G/H}(R_T)} F_T(g)$. Thus Theorem 1.3 is the assertion $\hat{F}_T(e) \to 1$ as $T \to \infty$.

The connection between Theorem 1.3 and Theorem 1.4 is via the following formula (see [DRS] and [EMc]):

$$\langle \hat{F_T}, \psi \rangle = \frac{1}{\lambda_{G/H}(R_T)} \int_{R_T} \left(\int_{G/\Gamma} \overline{\psi}\, d(g\mu_H) \right) d\lambda_{G/H}(g),$$

where ψ is any function in $C_0(G/\Gamma)$ and $g\mu_H$ is the translated measure as in Theorem 1.4.

If the non-focusing assumption is satisfied, then by Theorem 1.4, for "most" values of g, the inner integral will approach $\int_{G/\Gamma} \overline{\psi}\, d\mu = \langle 1, \psi \rangle$. Thus, $\hat{F_T} \to 1$ in the weak-star topology on $L^\infty(G/\Gamma, \mu_G)$. In can then be shown that $\hat{F_T} \to 1$ uniformly on compact sets.

2 A quantitative version of the Oppenheim conjecture

Let Q be an indefinite nondegenerate quadratic form in n variables. Let $\mathcal{L}_Q = Q(\mathbb{Z}^n)$ denote the set of values of Q at integral points. The Oppenheim conjecture, proved by Margulis (cf. [Mar2]) states that if $n \geq 3$, and Q is not proportional to a form with rational coefficients, then $\mathcal{L}_Q$ is dense. In joint work with G. Margulis and S. Mozes ([EMM1]) we study some finer questions related to the distribution the values of Q at integral points.

Let ν be a continuous positive function on the sphere $\{v \in \mathbb{R}^n \mid \|v\| = 1\}$, and let $\Omega = \{v \in \mathbb{R}^n \mid \|v\| < \nu(v/\|v\|)\}$. We denote by $T\Omega$ the dilate of Ω by T. Define the following set:

$$V^Q_{(a,b)}(\mathbb{R}) = \{x \in \mathbb{R}^n \mid a < Q(x) < b\}$$

We shall use $V_{(a,b)} = V^Q_{(a,b)}$ when there is no confusion about the form Q. Also let $V_{(a,b)}(\mathbb{Z}) = V^Q_{(a,b)}(\mathbb{Z}) = \{x \in \mathbb{Z}^n \mid a < Q(x) < b\}$. The set $T\Omega \cap \mathbb{Z}^n$ consists of $O(T^n)$ points, $Q(T\Omega \cap \mathbb{Z}^n)$ is contained in an interval of the form $[-\mu T^2, \mu T^2]$,

where $\mu > 0$ is a constant depending on Q and Ω. Thus one might expect that for any interval $[a, b]$, as $T \to \infty$,

$$|V_{(a,b)}(\mathbb{Z}) \cap T\Omega| \sim c_{Q,\Omega}(b-a)T^{n-2} \tag{1}$$

where $c_{Q,\Omega}$ is a constant depending on Q and Ω. This may be interpreted as "uniform distribution" of the sets $Q(\mathbb{Z}^n \cap T\Omega)$ in the real line. Our main result is that (1) holds if Q is not proportional to a rational form, and has signature (p, q) with $p \geq 3$, $q \geq 1$. We also determine the constant $c_{Q,\Omega}$.

If Q is an indefinite quadratic form in n variables, Ω is as above and (a, b) is an interval, we show that there exists a constant $\lambda = \lambda_{Q,\Omega}$ so that as $T \to \infty$,

$$\mathrm{Vol}(V_{(a,b)}(\mathbb{R}) \cap T\Omega) \sim \lambda_{Q,\Omega}(b-a)T^{n-2} \tag{2}$$

Our main result is the following:

THEOREM 2.1 *Let Q be an indefinite quadratic form of signature (p, q), with $p \geq 3$ and $q \geq 1$. Suppose Q is not proportional to a rational form. Then for any interval (a, b), as $T \to \infty$,*

$$|V_{(a,b)}(\mathbb{Z}) \cap T\Omega| \sim \lambda_{Q,\Omega}(b-a)T^{n-2}$$

where $n = p + q$, and $\lambda_{Q,\Omega}$ is as in (2).

Only the upper bound in this formula is new: the asymptotically exact lower bound was proved in [DM3]. Also a lower bound with a smaller constant was obtained independently by M. Ratner, and by S. G. Dani jointly with S. Mozes (both unpublished).

If the signature of Q is $(2, 1)$ or $(2, 2)$ then no universal formula like (1) holds. In fact, we have the following theorem:

THEOREM 2.2 *Let Ω_0 be the unit ball, and let $q = 1$ or 2. Then for every $\epsilon > 0$ and every interval (a, b) there exists a quadratic form Q of signature $(2, q)$ not proportional to a rational form, and a constant $c > 0$ such that for an infinite sequence $T_j \to \infty$,*

$$|V_{(a,b)}(\mathbb{Z}) \cap T\Omega_0| > cT_j^q(\log T_j)^{1-\epsilon}.$$

The case $q = 1$, $b \leq 0$ of Theorem 2.2 was noticed by P. Sarnak and worked out in detail in [Bre]. The quadratic forms constructed are of the form $x_1^2 + x_2^2 - \alpha x_3^2$, or $x_1^2 + x_2^2 - \alpha(x_3^2 + x_4^2)$, where α is extremely well approximated by squares of rational numbers.

However in the $(2, 1)$ and $(2, 2)$ cases, we can still establish an upper bound of the form $cT^q \log T$. This upper bound is effective, and is uniform over compact sets in the set of quadratic forms. We also give an effective uniform upper bound for the case $p \geq 3$.

THEOREM 2.3 *Let $\mathcal{O}(p, q)$ denote the space of quadratic forms of signature (p, q) and discriminant ± 1, let $n = p + q$, (a, b) be an interval, and let $\mathcal{D}$ be a compact*

subset of $\mathcal{O}(p,q)$. Let ν be a continuous positive function on the unit sphere and let $\Omega = \{v \in \mathbb{R}^n \mid \|v\| < \nu(v/\|v\|)\}$. Then, if $p \geq 3$ there exists a constant c depending only on $\mathcal{D}$, (a,b) and Ω such that for any $Q \in \mathcal{D}$ and all $T > 1$,

$$|V_{(a,b)}(\mathbb{Z}) \cap T\Omega| < cT^{n-2}$$

If $p = 2$ and $q = 1$ or $q = 2$, then there exists a constant $c > 0$ depending only on $\mathcal{D}$, (a,b) and Ω such that for any $Q \in \mathcal{D}$ and all $T > 2$,

$$|V_{(a,b)} \cap T\Omega \cap \mathbb{Z}^n| < cT^{n-2} \log T$$

Also, for the $(2,1)$ and $(2,2)$ cases, we have the following "almost everywhere" result:

THEOREM 2.4 *For almost all quadratic forms Q of signature $(p,q) = (2,1)$ or $(2,2)$*

$$|V_{(a,b)}(\mathbb{Z}) \cap T\Omega| \sim \lambda_{Q,\Omega}(b-a)T^{n-2}$$

where $n = p + q$, and $\lambda_{Q,\Omega}$ is as in (2).

CONNECTION WITH QUANTUM CHAOS. It has been suggested by Berry and Tabor that the distribution of the local spacings between eigenvalues of the quantization of a completely integrable Hamiltonian is Poisson. For the Hamiltonian which is the geodesic flow on the flat 2-torus, it was noted by P. Sarnak [Sar] that this problem translates to one of the spacing between the values at integers of a binary quadratic form, and is related to the quantitative Oppenheim problem. We briefly recall the connection following [Sar].

Let β^2 be a positive irrational number, and let M_β denote the rectangular torus with the flat metric and sides π and π/β. Let Λ_β denote the spectrum of the Laplace operator on M_β, i.e.

$$\Lambda_\beta = \{P_\beta(m,n) \ : \ m,n \geq 0, \text{ or } m = 0, n \geq 0, \quad m,n \in \mathbb{Z}\}$$

where $P_\beta(x,y)$ denotes the positive definite quadratic form $x^2 + \beta^2 y^2$. We label the elements of Λ_β (with multiplicity) by

$$0 = \lambda_0(\beta) < \lambda_1(\beta) \leq \lambda_2(\beta) \dots$$

It is easy to see that Weyl's law holds, i.e.

$$|\{j \ : \ \lambda_j(\beta) \leq T\}| \sim c_\beta T$$

where $c_\beta = \pi/(4\beta)$ is related to the area of M_β. We are interested in the distribution of the local spacings $\lambda_j(\beta) - \lambda_k(\beta)$. In particular, set

$$R_\beta(a,b,T) = \frac{|\{(j,k) \ : \ \lambda_j(\beta) \leq T, \lambda_k(\beta) \leq T, j \neq k, a \leq \lambda_j(\beta) - \lambda_k(\beta) \leq b\}|}{T}$$

The quantity R_β is called the pair correlation. The random number (Poisson) model predicts that

$$\lim_{T\to\infty} R_\beta(a,b,T) = c_\beta^2(b-a). \tag{3}$$

Note that the differences $\lambda_j(\beta) - \lambda_k(\beta)$ are precisely the integral values of the quadratic form $Q_\beta(x_1,x_2,x_3,x_4) = x_1^2 - x_3^2 + \beta^2(x_2^2 - x_4^2)$.

P. Sarnak considered in [Sar] a two-parameter family of flat 2-tori and showed that (3) holds on a set of full measure of these tori. Some similar results for forms of higher degree were proved in [Va1] and [Va2].

These methods, however, cannot be used to construct a specific torus for which (3) holds. In [EMM2], using a refinement of the methods of [EMM1] we establish (3) under a mild diophantine condition on β, and in particular for any irrational algebraic β. Our main result is the following:

Theorem 2.5 *Suppose β^2 is* diophantine, *i.e. there exists $N > 0$ such that for all relatively prime pairs of integers (p,q), $|\beta^2 - p/q| > q^{-N}$. Then, for any interval (a,b), (3) holds, i.e.*

$$\lim_{T\to\infty} R_\beta(a,b,T) = c_\beta^2(b-a).$$

In particular, the set of $\beta \subset \mathbb{R}$ for which (3) does not hold has zero Hausdorff dimension.

Thus, if β^2 is diophantine, then M_β has a spectrum whose pair correlation satisfies the Berry-Tabor conjecture.

We note that some diophantine condition in Theorem 2.5 is needed in view of Theorem 2.2.

Quadratic Forms. We now relate the counting problem of Theorem 2.1 to a certain integral expression involving the orthogonal group of the quadratic form and the space of lattices G/Γ, where $G = SL(n,\mathbb{R})$, $\Gamma = SL(n,\mathbb{Z})$. Let f be a bounded function on $\mathbb{R}^n - \{0\}$ vanishing outside a compact subset. For a lattice $\Delta \in SL(n,\mathbb{R})$ let

$$\tilde{f}(\Delta) = \sum_{v\in\Delta} f(v) \tag{4}$$

Let $n \geq 3$, and let $p \geq 2$. We denote $n-p$ by q, and assume $q > 0$. Let $\{e_1, e_2, \dots e_n\}$ be the standard basis of $\mathbb{R}^n$. Let Q_0 be the quadratic form defined by

$$Q_0\left(\sum_{i=1}^n v_i e_i\right) = 2v_1v_n + \sum_{i=2}^p v_i^2 - \sum_{i=p+1}^{n-1} v_i^2 \quad \text{for all } v_1,\dots,v_n \in \mathbb{R}.$$

It is straightforward to verify that Q_0 has signature (p,q). For each quadratic form Q and $g \in G$, let Q^g denote the quadratic form defined by $Q^g(v) = Q(gv)$ for all $v \in \mathbb{R}^n$. By the well known classification of quadratic forms over $\mathbb{R}$, for

each $Q \in \mathcal{O}(p,q)$ there exists $g \in G$ such that $Q = Q_0^g$. For any quadratic form Q let $SO(Q)$ denote the special orthogonal group corresponding to Q; namely $\{g \in G \mid Q^g = Q\}$. Let $H = SO(Q_0)$. Then the map $H\backslash G \to \mathcal{O}(p,q)$ given by $Hg \to Q_0^g$ is a homeomorphism. If $Q = Q_0^g$, let Δ_Q denote the lattice $g\mathbb{Z}^n$.

For $t \in \mathbb{R}$, let a_t be the linear map so that $a_t e_1 = e^{-t} e_1$, $a_t e_n = e^t e_n$, and $a_t e_i = e_i$, $2 \le i \le n-1$. Then the one-parameter group $\{a_t\}$ is contained in H. Let $\hat{K}$ be the subgroup of G consisting of orthogonal matrices, and let $K = H \cap \hat{K}$. It is easy to check that K is a maximal compact subgroup of H, and consists of all $h \in H$ leaving invariant the subspace spanned by $\{e_1 + e_n, e_2, \ldots, e_p\}$. We denote by m the normalized Haar measure on K.

Suppose for simplicity that the set Ω in Theorem 2.1 is invariant under the action of K. Then, it can be shown that for a suitably chosen function f on $\mathbb{R}^n$, $|V_{(a,b)}(\mathbb{Z}) \cap T\Omega|$ can be well approximated by the following expression:

$$T^{n-2} \int_K \tilde{f}(a_t k \Delta_Q)\, dm(k) \tag{5}$$

where $t = \log T$, and $\tilde{f}$ is as in (4). Thus, Theorem 2.1 can be deduced from the following theorem:

THEOREM 2.6 *Suppose $p \ge 3$, $q \ge 1$. Let f be a continuous function on $\mathbb{R}^n$ vanishing outside a compact set. Let $\Delta \in G/\Gamma$ be a unimodular lattice such that $H\Delta$ is not closed. Then*

$$\lim_{t \to +\infty} \int_K \tilde{f}(a_t k \Delta)\, dm(k) = \int_{G/\Gamma} \tilde{f}(y)\, d\mu(y). \tag{6}$$

If in Theorem 2.6, in place of the function $\tilde{f}$ we considered any bounded continuous function ϕ, then (6) would follow easily from [DM3, Theorem 3]). This theorem is a refined version of Ratner's uniform distribution theorem [Rat4]; the proof uses Ratner's measure classification theorem (see [Rat1, Rat2, Rat3]), Dani's theorem on the behavior of unipotent orbits at infinity [Dan1, Dan2], and "linearization" techniques.

Both [DM3, Theorem 3] and Ratner's uniform distribution theorem hold for bounded continuous functions, but not for arbitrary continuous functions from $L^1(G/\Gamma)$. However, for a non-negative bounded continuous function f on $\mathbb{R}^n$, the function $\tilde{f}$ defined in (4) is non-negative, continuous, and L^1 but unbounded (it is in $L^s(G/\Gamma)$ for $1 \le s < n$, where $G = SL(n,\mathbb{R})$, and $\Gamma = SL(n,\mathbb{Z})$). As it was done in [DM3] it is possible to obtain asymptotically exact lower bounds by considering bounded continuous functions $\phi \le \tilde{f}$. However, to carry out the integral in (5) and prove the upper bounds in the theorems stated above we need to examine carefully the situation at the "cusp" of G/Γ, i.e outside of compact sets. Some techniques for handling this were developed in [Mar1], [Dan1], [Dan2]; see also [KM] for a simplified proof and some interesting applications to the metric theory of diophantine approximations. However, these techniques are not sufficient for this problem.

Let Δ be a lattice in $\mathbb{R}^n$. We say that a subspace L of $\mathbb{R}^n$ is Δ-*rational* if $L \cap \Delta$ is a lattice in L. For any Δ-rational subspace L, we denote by $d_\Delta(L)$ or simply

by $d(L)$ the volume of $L/(L \cap \Delta)$. Let us note that $d(L)$ is equal to the norm of $e_1 \wedge \cdots \wedge e_\ell$ in the exterior power $\bigwedge^\ell(\mathbb{R}^n)$ where $\ell = \dim L$ and $(e_1, \cdots, e_\ell)$ is a basis over $\mathbb{Z}$ of $L \cap \Delta$. If $L = \{0\}$ we write $d(L) = 1$. A lattice is Δ *unimodular* if $d_\Delta(\mathbb{R}^n) = 1$. The space of unimodular lattices is isomorphic to $SL(n, \mathbb{R})/SL(n, \mathbb{Z})$.

Let us introduce the following notation:

$$\alpha_i(\Delta) = \sup\left\{ \frac{1}{d(L)} \,\middle|\, L \text{ is a } \Delta\text{-rational subspace of dimension } i \right\}, \quad 0 \le i \le n,$$

$$\alpha(\Delta) = \max_{0 \le i \le n} \alpha_i(\Delta).$$

The following lemma is known as the "Lipshitz Principle":

LEMMA 2.7 ([SCH, LEMMA 2]) *Let f be a bounded function on $\mathbb{R}^n$ vanishing outside a compact subset. Then there exists a positive constant $c = c(f)$ such that*

$$\tilde{f}(\Delta) < c\alpha(\Delta) \tag{7}$$

for any lattice Δ in $\mathbb{R}^n$. Here $\tilde{f}$ is the function on the space of lattices defined in (4).

By (7) the function $\tilde{f}(g)$ on the space of unimodular lattices G/Γ is majorized by the function $\alpha(g)$. The function α is more convenient since it is invariant under the left action of the maximal compact subgroup $\hat{K}$ of G, and its growth rate at infinity is known explicitly. Theorem 2.6 is proved by combining [DM3, Theorem 3] with the following integrability estimate:

THEOREM 2.8 *If $p \ge 3$, $q \ge 1$ and $0 < s < 2$, or if $p = 2$, $q \ge 1$ and $0 < s < 1$, then for any lattice Δ in $\mathbb{R}^n$*

$$\sup_{t>0} \int_K \alpha(a_t k \Delta)^s \, dm(k) < \infty.$$

The upper bound is uniform as Δ varies over compact sets in the space of lattices.

This result can be interpreted as follows. For a lattice Δ in G/Γ and for $h \in H$, let $f(h) = \alpha(h\Delta)$. Since α is left-$\hat{K}$ invariant, f is a function on the symmetric space $X = K\backslash H$. Theorem 2.8 is the statement that if if $p \ge 3$, then the averages of f^s, $0 < s < 2$ over the sets Ka_tK in X remain bounded as $t \to \infty$, and the bound is uniform as one varies the base point Δ over compact sets. We remark that in the case $q = 1$, the rank of X is 1, and the sets Ka_tK are metric spheres of radius t, centered at the origin.

If $(p, q) = (2, 1)$ or $(2, 2)$, Theorem 2.8 does not hold even for $s = 1$. The following result is, in general, best possible:

THEOREM 2.9 *If $p = 2$ and $q = 2$, or if $p = 2$ and $q = 1$, then for any lattice Δ in $\mathbb{R}^n$,*

$$\sup_{t>1} \frac{1}{t} \int_K \alpha(a_t k \Delta) \, dm(k) < \infty,$$

The upper bound is uniform as Δ varies over compact sets in the space of lattices.

We now outline the proof of Theorems 2.8 and 2.9. From its definition, the function $\alpha(g)$ is the maximum over $1 \leq i \leq n$ of left-$\hat{K}$ invariant functions $\alpha_i(g)$. The main idea of the proof is to show that the α_i satisfy a system of integral inequalities which imply the desired bound.

If $p \geq 3$ and $0 < s < 2$, or if $(p,q) = (2,1)$ or $(2,2)$ and $0 < s < 1$, we show that for any $c > 0$ there exist $t > 0$, and $\omega > 1$ so that the the functions α_i^s satisfy the following system of integral inequalities in the space of lattices:

$$A_t \alpha_i^s \leq c_i \alpha_i^s + \omega^2 \max_{0 \leq j \leq n-i,i} \sqrt{\alpha_{i+j}^s \alpha_{i-j}^s} \tag{8}$$

where A_t is the averaging operator $(A_t f)(\Delta) = \int_K f(a_t k \Delta)$, and $c_i \leq c$. If $(p,q) = (2,1)$ or $(2,2)$ and $s = 1$, then (8) also holds (for suitably modified functions α_i), but some of the constants c_i cannot be made smaller than 1.

Let $f_i(h) = \alpha_i(h\Delta)$, so that each f_i is a function on the symmetric space X. When one restricts to an orbit of H, (8) becomes:

$$A_t f_i^s \leq c_i f_i^s + \omega^2 \max_{0 \leq j \leq n-i,i} \sqrt{f_{i+j}^s f_{i-j}^s} \tag{9}$$

If $\operatorname{rank} X = 1$, then $(A_t f)(h)$ can be interpreted as the average of f over the sphere of radius t in X, centered at h. We show that if the f_i satisfy (9) then for any $\epsilon > 0$, the function $f = f_{\epsilon,s} = \sum_{0 \leq i \leq n} \epsilon^{i(n-i)} f_i^s$ satisfies the scalar inequality:

$$A_t f \leq cf + b \tag{10}$$

where t, c and b are constants. We show that if c is sufficiently small, then (10) for a fixed t together with the uniform continuity of $\log f$ imply that $(A_r f)(1)$ is bounded as a function of r, which is the conclusion of Theorem 2.8. If $c = 1$, which will occur in the $SO(2,1)$ and $SO(2,2)$ cases, then (10) implies that $(A_r f)(1)$ is growing at most linearly with the radius, which is the conclusion of Theorem 2.9.

3 Actions on the space of quadratic differentials

Given a Riemann surface structure on a closed surface of genus $g > 1$, recall that a holomorphic quadratic differential ϕ is a tensor of the form $\phi(z)dz^2$ in local coordinates with ϕ holomorphic. Away from the zeroes, a coordinate ζ can be chosen so that $\phi = d\zeta^2$, which determines a Euclidean metric $|d\zeta^2|$ in that chart. The change of coordinates away from the zeroes of ϕ are of the form $\zeta \to \pm\zeta + c$, which preserves the Euclidan metric. Consequently, quadratic differentials are sometimes refered to as translation surfaces or flat structures. If one can always take the + sign in the change of coordinates, the translation surface is orientable. Equivalently, the quadratic differential is the square of an abelian differential. We will henceforth adopt the notation S to refer to the structure of a quadratic differential. A zero of order $k \geq 1$ of S defines a cone angle singularity of the metric: there is a neighborhood of the zero such that the metric is of the form

$$ds^2 = dr^2 + ((k+2)rd\theta/2)^2$$

and the cone angle is $(k+2)\pi$. The number of zeroes counting multiplicity is $4g-4$. Let P be a partition of $4g-4$ (i.e. a representaion of $4g-4$ as a sum of positive integers).

Let $QD(g,P)$ denote the set of flat structures S on a surface of genus g whose zero set is given by P; the space $QD(g,P)$ is called a stratum. The term is justified by the fact that the space of all quadratic differentials on Riemann surfaces of genus g is stratified by the spaces $QD(g,P)$ as P varies over the partitions of $4g-4$. The spaces $QD(g,P)$ have projection maps to the Teichmuller space T_g of genus g.

There is an $SL(2,\mathbb{R})$ action on $QD(g,P)$. In each coordinate patch $\zeta \in \mathbb{R}^2$, and for $A \in SL(2,\mathbb{R})$, the action is the linear action $\zeta \to A\zeta$. The fact that the change coordinates is given by $\zeta \to \pm\zeta + c$ says that this is well-defined. This action preserves a measure μ_0 on $QD(g,P)$ ([Mas1], [Ve1]).

A saddle connection of S is a geodesic segment joining two zeroes of S which has no zeroes in its interior. A saddle connection determines a vector in $\mathbb{R}^2$ since in each coordinate chart it is geodesic with respect to Euclidean metric. A closed geodesic that does not pass through any zeroes determines a cylinder of parallel freely homotopic closed geodesics of the same length. Each boundary component of the cylinder consists of one or more parallel saddle connections.

We note that there is a well known construction which associates a surface with a flat structure to each rational polygon (see [ZK], [Gut], [KMS]). In particular counting families of periodic trajectories of the billiard in the polygon is equivalent to counting cylinders of closed geodesics on the surface constructed from the polygon.

In a recent paper [Ve2] Veech observed that this counting problem is analogous to the counting problem of section §2. The results of this section, which are joint work with H. Masur, are inspired by this paper. We now recall the construction of [Ve2]. In the Teichmuller space situation, for a bounded function f on $\mathbb{R}^2 - \{0\}$, Veech defines a function on the moduli space of quadratic differentials by

$$\tilde{f}(S) = \sum_{v \in V(S))} f(v) \tag{11}$$

where for a point x in the moduli space(i.e a surface and a quadratic differential), $V(S)$ denotes the set of vectors in $\mathbb{R}^2$ corresponding to cylinders of closed geodesics on S.

Let $N(S,T)$ denote the number of cylinders periodic geodesics on S of length at most T. In [Ve2] it is shown that for an appropriate function f, $N(S,T)$ is well approximated by an expression of the form:

$$T^2 \int_K \tilde{f}(a_t k \cdot S)\, dm(k) \tag{12}$$

where a_t and k are as in §2 and $t = \log T$. Thus, this problem is remarkably similar to the problem in §2. However since we are not in the homogeneous space setting, we expect that obtaining results as sharp as in §2 would be very difficult. However, some of the techniques mentioned in §2, in particular, the idea of obtaining upper

bounds via systems of inequalities can be transferred to this situation. This yields a simplified proof of the quadratic upper bound of [Mas2].

One can also obtain an individual ergodic theorem analogous to Theorem 2.4, using an ergodic theorem due to A. Nevo. More precisely, one can prove the following (in weaker form this theorem is proved in [Ve2]):

THEOREM 3.1 *For a (nonorientable) translation surface S and $T > 0$, let $N(S,T)$ denote the number of cylinders of closed geodesics on S such that the norm of the associated vector is at most T. For any P there exists a constant c_P such that for almost all $S \in QD(g,P)$,*

$$N(S,T) \sim c_P T^2$$

In some special examples, one can use the full theory of unipotent flows in a way exactly analogous to that of §2. In particular, one can prove the following:

THEOREM 3.2 *For any rational $0 < p/q < 1$, and any α, $0 < \alpha < 1$, let $P = P_{p/q,\alpha}$ denote the polygon whose boundary is the boundary of the unit square $\partial([0,1] \times [0,1])$ union the segment $\{p/q\} \times [0,\alpha]$. Then, for any irrational α,*

$$N(P,T) \sim c_{p/q} T^2$$

where $c_{p/q}$ is some explicitly computable constant.

We note that the case $p/q = 1/2$ is elementary, and was done previosly by E. Gutkin and C. Judge (private communication).

REFERENCES

[BH-C] A. Borel and Harish-Chandra. Arithmetic subgroups of algebraic groups. *Annals of Math* 75 (1962), 485–535.

[BR] M. Borovoi and Z. Rudnick. Hardy–Littlewood varieties and semisimple groups. *Inv. Math.* (1994)

[Bre] T. Brennan. . Princeton University undergraduate thesis, 1994.

[Dan1] S.G. Dani. On orbits of unipotent flows on homogeneous spaces. *Ergod. Theor. Dynam. Syst.* 4(1984), 25–34.

[Dan2] S.G. Dani. On orbits of unipotent flows on homogenous spaces II. *Ergod. Theor. Dynam. Syst* 6(1986), 167–182.

[Dan3] S.G. Dani. Flows on homogeneous spaces and diophantine approximation. In: *Proc. of ICM (1994)*, Zurich.

[DM1] S.G. Dani and G.A. Margulis. Orbit closures of generic unipotent flows on homogeneous spaces of $\mathrm{SL}(3,\mathbb{R})$. *Math.Ann.* 286 (1990), 101–128.

[DM2] S.G. Dani and G.A. Margulis. Asymptotic behaviour of trajectories of unipotent flows on homogeneous spaces. *Indian. Acad. Sci. J.* 101 (1991), 1–17.

[DM3] S.G. Dani and G.A. Margulis. Limit distributions of orbits of unipotent flows and values of quadratic forms. *Advances in Soviet Math.* 16 (1993), 91–137.

[DRS] W. Duke, Z. Rudnick, and P. Sarnak. Density of integer points on affine homogeneous varieties. *Duke Math. J.* 71(1993), 143–180.

[EMc] A. Eskin and C. McMullen. Mixing, counting and equidistribution in Lie groups. *Duke Math. J.* 71(1993), 143–180.

[EMM1] A. Eskin, G. Margulis, and S. Mozes. Upper bounds and asymptotics in a quantitative version of the Oppenheim conjecture. *Ann. of Math. (2)* 147(1998), no. 1, 93–141

[EMM2] A. Eskin, G. Margulis, and S. Mozes. Quadratic forms of signature $(2,2)$ and eigenvalue spacings on rectangular 2-tori. *Preprint.*

[EMS1] A. Eskin, S. Mozes, and N. Shah. Non-divergence of translates of certain algebraic measures. *Geom. Funct. Anal.* 7(1997), no. 1, 48–80.

[EMS2] Alex Eskin, Shahar Mozes, and Nimish Shah. Unipotent flows and counting lattice points on homogeneous varieties. *Annals of Math.* (1996), 253–299.

[Gut] E. Gutkin. Billiards on almost integrable polyhedral surfaces. *Ergodic Th. Dyn Syst.* 4 (1984) pp. 569–584

[KM] D. Kleinbok and G. Margulis. Flows on homogeneous spaces and Diophantine approximation on manofolds. *To appear in Annals of Math.*

[KMS] S. Kerckhoff, H. Masur and J. Smillie, *Ergodicity of Billiard flows and quadratic differentials*, Annals of Math. 124 (1986), 293–311.

[Marklof] J. Marklof. Spectral form factors of rectangular billiards. *Preprint.*

[Mar1] G. A. Margulis. On the action of unipotent groups in the space of lattices. In *Lie Groups and their representations, (Proc. of Summer School in Group Representations, Bolyai Janos Math. Soc., Akademai Kiado, Budapest, 1971)*, pages 365–370. Halsted, New York, 1975.

[Mar2] G.A. Margulis. Discrete subgroups and ergodic theory. In: *Number Theory, Trace Formulas and Discrete Groups* (symposium in honour of A. Selberg, Oslo). Academic Press, 1989, pp. 377-398.

[Mar3] G.A. Margulis. Dynamical and ergodic properties of subgroups actions on homogeneous spaces with applications to number theory. In:*Proc. of ICM* (Kyoto, 1990). Math. Soc. of Japan and Springer, 1991, pp. 193-215.

[Mar4] G.A. Margulis. Oppenheim Conjecture. In: *Fields Medalists Lectures,* World Scientific (1997), pp. 272–327.

[MS] S. Mozes and N.A. Shah. On the space of ergodic invariant measures of unipotent flows. *Ergodic Th. and Dynam. Syst.* 15 (1995), 149–159.

[Mas1] H. Masur. Interval exchange transformations and measured foliations. *Annals of Math.* 115(1982), 169–200.

[Mas2] H. Masur. The growth rate of trajectories of a quadratic differential. *Ergodic Th. and Dynam. Syst.* 10 (1990), 151–176.

[Rat1] M. Ratner. Strict measure rigidity for nilpotent subgroups of solvable groups. *Invent. Math.* 101 (1990), 449–482.

[Rat2] M. Ratner. On measure rigidity of unipotent subgroups of semisimple groups. *Acta. Math.* 165 (1990), 229–309.

[Rat3] M. Ratner. On Ragunathan's measure conjecture. *Annals of Math.* 134 (1991), 545–607.

[Rat4] M. Ratner. Raghunathan's topological conjecture and distributions of unipotent flows. *Duke Math. J.* 63 (1991), 235–290.

[Rat5] M. Ratner. Interactions between Lie groups, ergodic theory and number theory. *Proc. of ICM (Zurich, 1994).*

[Sar] P. Sarnak. Values at integers of binary quadratic forms. Preprint.

[Sch] W. Schmidt. Asymptotic formulae for point lattices of bounded determinant and subspaces of bounded height. *Duke Math. J.* 35(1968), 327–339.

[Sha1] N. A. Shah. Uniformly distributed orbits of certain flows on homogeneous spaces. *Math. Ann.* 289 (1991), 315–334.

[Ve1] W. Veech. Gauss measures for transformations on the space of interval exchange maps. *Ann. of Math.* (2) 115 (1982), no. 1, 201–242.

[Ve2] W. Veech, *Siegel measures*, preprint.

[Va1] J. Vanderkam. Values at integers of homogeneous polynomials. *To appear in the Duke Math. J.*

[Va2] J. Vanderkam. Pair correlation of four-dimensional flat tori. *To appear in the Duke Math. J.*

[ZK] A.N. Zemlyakov, A.B. Katok. Topological transitivity of billiards in polygons. *Matem. Zametki* 18(2) (1975) pp. 291–300. English translation in *Math. Notes* 18(2) (1976) pp. 760–764.

Alex Eskin
Department of Mathematics
University of Chicago
Chicago
Illinois, USA
eskin@math.uchicago.edu

Harmonic Analysis on Semisimple p-Adic Lie Algebras

Robert E. Kottwitz

Abstract. Certain topics in harmonic analysis on semisimple groups arise naturally when one uses the Arthur-Selberg trace formula to study automorphic representations of adele groups. These topics, which fall under the heading of "comparison of orbital integrals," have been surveyed in Waldspurger's article in the the proceedings of ICM94. As in Harish-Chandra's work, many questions in harmonic analysis on the group can be reduced to analogous questions on its Lie algebra, and in particular this is the case for "comparison of orbital integrals." We will discuss some recent work of this type.

1991 Mathematics Subject Classification: 22E35, 22E50, 11F85
Keywords and Phrases: Orbital integrals, p-adic, Lie algebra

1 Harmonic analysis on $\mathfrak{g}(F)$

Let F be a local field of characteristic 0, let $\bar{F}$ be an algebraic closure of F, and let Γ be the Galois group $\mathrm{Gal}(\bar{F}/F)$. Let G be a connected reductive F-group, and let $\mathfrak{g}$ be its Lie algebra. We begin by recalling Harish-Chandra's viewpoint on the relationship between harmonic analysis on $G(F)$ and harmonic analysis on $\mathfrak{g}(F)$ (see [4] for example).

Harmonic analysis on $G(F)$ is the study of (conjugation) invariant distributions on $G(F)$. Orbital integrals (integrals over conjugacy classes) are such distributions as are the characters of irreducible representations of $G(F)$. Harmonic analysis on $\mathfrak{g}(F)$ is the study of invariant distributions on $\mathfrak{g}(F)$ (invariant under the adjoint action of $G(F)$). Orbital integrals on $\mathfrak{g}(F)$ (integrals over orbits for the adjoint action) are of course the Lie algebra analogs of orbital integrals on the group. One of Harish-Chandra's basic insights was that the Lie algebra analogs of irreducible characters on $G(F)$ are the distributions on $\mathfrak{g}(F)$ obtained as Fourier transforms (in the distribution sense) of orbital integrals.

Harish-Chandra exploited this analogy in two ways. First, he often proved theorems in pairs, one on the group and one on the Lie algebra. For example he showed that both irreducible characters and Fourier transforms of orbital integrals are locally integrable functions, smooth on the regular semisimple set. Second, using the exponential map, he reduced questions in harmonic analysis in a neighborhood of the identity element in $G(F)$ to questions in harmonic analysis in a neighborhood of 0 in $\mathfrak{g}(F)$.

2 ENDOSCOPY FOR $\mathfrak{g}(F)$

In this section we assume the field F is p-adic. In recent years Waldspurger [21, 22, 23] has significantly broadened the scope of the analogy discussed above by developing a Lie algebra analog of the theory of endoscopy. As is the case on the group, the theory of endoscopy on $\mathfrak{g}(F)$ is not yet complete, but Waldspurger's results give extremely convincing evidence that such a theory exists and go a long way towards establishing the theory by reducing everything to the Lie algebra analog of the "fundamental lemma." We now summarize these results.

We begin with the basic definitions. We fix a non-trivial (continuous) additive character $\psi : F \to \mathbf{C}^\times$ and a non-degenerate G-invariant symmetric bilinear form $\langle \cdot, \cdot \rangle$ on $\mathfrak{g}(F)$, and we use $\langle \cdot, \cdot \rangle$, ψ to identify $\mathfrak{g}(F)$ with its Pontryagin dual. We let dX denote the unique self-dual Haar measure on $\mathfrak{g}(F)$ with respect to $\psi\langle \cdot, \cdot \rangle$. Let f belong to $C_c^\infty(\mathfrak{g}(F))$, the space of locally constant, compactly supported functions on $\mathfrak{g}(F)$, and define the Fourier transform $\hat{f}$ of f by $\hat{f}(Y) = \int_{\mathfrak{g}(F)} f(X)\psi\langle X, Y\rangle dX$. A distribution D on $\mathfrak{g}(F)$ is simply a $\mathbf{C}$-linear map $D : C_c^\infty(\mathfrak{g}(F)) \to \mathbf{C}$, and its Fourier transform $\hat{D}$ is the distribution with the defining property that $\hat{D}(f) = D(\hat{f})$ for all $f \in C_c^\infty(\mathfrak{g}(F))$.

We now define normalized orbital integrals. The Haar measure dX on $\mathfrak{g}(F)$ determines a Haar measure dg on $G(F)$ (impose compatibility under the exponential map in a neighborhood of the origin). Now let $X \in \mathfrak{g}(F)$ be a regular semisimple element and let T be its centralizer in G. Choose a Haar measure dt on $T(F)$. Let $D_G(X) = \det(\mathrm{Ad}(X); \mathfrak{g}/\mathfrak{t})$, where $\mathfrak{t}$ is of course the Lie algebra of T. The normalized orbital integral I_X is defined by

$$I_X(f) = |D_G(X)|^{1/2} \int_{T(F)\backslash G(F)} f(\mathrm{Ad}(g^{-1})(X))\, dg/dt \qquad (f \in C_c^\infty(\mathfrak{g}(F)).$$

We also need the stable orbital integral SI_X defined by $SI_X = \sum_{X'} I_{X'}$, the sum being taken over a set of representatives for the $G(F)$-orbits in the set of elements $X' \in \mathfrak{g}(F)$ that are *stably conjugate* to X in the sense that there exists $g \in G$ such that $\mathrm{Ad}(g)(X) = X'$; we use the (canonical) F-isomorphism $\mathrm{Ad}(g)$ from T to the centralizer T' of X' to transport our measure dt over to $T'(F)$.

We let $C_c^\infty(\mathfrak{g}(F))^{\mathrm{unst}}$ denote the subspace of $C_c^\infty(\mathfrak{g}(F))$ consisting of all elements f such that $SI_X(f) = 0$ for all regular semisimple $X \in \mathfrak{g}(F)$. A distribution D on $\mathfrak{g}(F)$ is said to be *stably invariant* if $D(f) = 0$ for all $f \in C_c^\infty(\mathfrak{g}(F))^{\mathrm{unst}}$. Clearly any stably invariant distribution is in fact invariant.

Waldspurger [23] has shown that the Fourier transform of a stably invariant distribution is again stably invariant. It follows that the Fourier transform carries $C_c^\infty(\mathfrak{g}(F))^{\mathrm{unst}}$ onto itself, and hence induces an automorphism of the quotient space $SC_c^\infty(\mathfrak{g}(F)) := C_c^\infty(\mathfrak{g}(F))/C_c^\infty(\mathfrak{g}(F))^{\mathrm{unst}}$.

Now let (H, s, ξ) be an endoscopic triple for G (see [10]). Thus H is a quasi-split F-group, s is a Γ-fixed element in the center of the Langlands dual group $\hat{H}$ of H, and ξ is an L-embedding of ${}^L H$ into ${}^L G$ that identifies $\hat{H}$ with the identity component of the centralizer in $\hat{G}$ of the image of s. If the derived group of G is not simply connected, objects slightly more general than (H, s, ξ) are needed [10], but let us ignore this minor complication.

For any maximal torus T_H in H there is a canonical G-conjugacy class of embeddings $\mathfrak{t}_H \to \mathfrak{g}$. We say that $Y \in \mathfrak{t}_H$ is G-regular if its image under any of these embeddings is regular in $\mathfrak{g}$. Any such embedding that is defined over F is called an *admissible embedding*. If G is not quasi-split, admissible embeddings need not exist. Given G-regular $X_H \in \mathfrak{t}_H(F)$, one says that $X_G \in \mathfrak{g}(F)$ is an *image* of X_H if there is an admissible embedding mapping X_H to X_G. The set of images of X_H in $\mathfrak{g}(F)$ is either empty or a single stable conjugacy class, called the image of the stable class of X_H.

Waldspurger [22], [23] defines transfer factors for $\mathfrak{g}(F)$ analogous to those of Langlands-Shelstad [10]. These are non-zero complex numbers $\Delta(X_H, X_G)$, defined whenever $X_H \in \mathfrak{h}(F)$ is G-regular and X_G is an image of X_H. The transfer factor depends only on the stable conjugacy class of X_H, and for any stable conjugate X'_G of X_G, we have the simple transformation law

$$\Delta(X_H, X'_G) = \Delta(X_H, X_G) \cdot \langle \mathrm{inv}(X_G, X'_G), s_{T_G} \rangle^{-1}, \tag{1}$$

in which the last factor has the following meaning. Let T_G (respectively, T_H) denote the centralizer of X_G (respectively, X_H) in G (respectively, H). We identify T_H and T_G using the unique admissible embedding that carries X_H to X_G. Choose $g \in G$ such that $\mathrm{Ad}(g)(X'_G) = X_G$. Then $\sigma \mapsto g\sigma(g)^{-1}$ is a 1-cocycle of Γ in T_G, whose class we denote by $\mathrm{inv}(X_G, X'_G)$. The element s appearing in our endoscopic data is a Γ-fixed element in the center of $\hat{H}$, and thus can be regarded as a Γ-fixed element s_{T_G} of $\hat{T}_H = \hat{T}_G$. We then pair $\mathrm{inv}(X_G, X'_G)$ with s_{T_G} using the Tate-Nakayama pairing

$$\langle \cdot, \cdot \rangle : H^1(F, T_G) \times \hat{T}_G^{\Gamma} \to \mathbf{C}^{\times},$$

where $\hat{T}_G^{\Gamma}$ denotes the group of fixed points of Γ in $\hat{T}_G$.

Matching Conjecture. See [22]. For every $f \in C_c^{\infty}(\mathfrak{g}(F))$ there exists $f^H \in C_c^{\infty}(\mathfrak{h}(F))$ such that for every G-regular semisimple element $X_H \in \mathfrak{h}(F)$

$$SI_{X_H}(f^H) = \sum_{X_G} \Delta(X_H, X_G) I_{X_G}(f), \tag{2}$$

where the sum ranges over a set of representatives X_G for the $G(F)$-orbits in the set of images of X_H in $\mathfrak{g}(F)$. Here we are using Haar measures on the centralizers T_H, T_G of X_H, X_G that correspond to each other under the unique admissible isomorphism $T_H \simeq T_G$ that carries X_H to X_G.

Since the G-regular semisimple elements in $\mathfrak{h}(F)$ are dense in the set of all regular semisimple elements in $\mathfrak{h}(F)$, the stable regular semisimple orbital integrals of f^H are uniquely determined, and therefore f^H is uniquely determined as an element in $SC_c^{\infty}(\mathfrak{h}(F))$. Assume the matching conjecture is true. Then $f \mapsto f^H$ is a well-defined linear map $C_c^{\infty}(\mathfrak{g}(F)) \to SC_c^{\infty}(\mathfrak{h}(F))$, and dual to this is a linear map (endoscopic induction, generalizing parabolic induction) i_H^G from stably invariant distributions on $\mathfrak{h}(F)$ to invariant distributions on $\mathfrak{g}(F)$, defined

by $i_H^G(D)(f) = D(f^H)$ (where D is a stably invariant distribution on $\mathfrak{h}(F)$ and $f \in C_c^\infty(\mathfrak{g}(F))$). By its very definition endoscopic induction carries SI_{X_H} into $\sum_{X_G} \Delta(X_H, X_G) I_{X_G}$.

Waldspurger [22] observes that the analogous matching conjecture on the group [10] implies the matching conjecture on the Lie algebra (via the exponential map) and that the matching conjecture for all reductive Lie algebras simultaneously implies the matching conjecture for all reductive groups simultaneously (via the theory of descent and local transfer developed by Langlands-Shelstad [11]).

Consider for a moment the case of endoscopy on a real group $G(\mathbf{R})$. Then Shelstad [18] has proved that endoscopic induction carries certain stable combinations of irreducible characters on $H(F)$ to unstable linear combinations of irreducible characters on $G(F)$. The coefficients in these unstable linear combinations can be regarded as spectral analogs of transfer factors. Langlands [12] has conjectured that there is similar theory for p-adic groups as well.

Now we return to endoscopy on $\mathfrak{g}(F)$. Let X be a regular semisimple element in $\mathfrak{g}(F)$. Recall that the Fourier transform $\hat{I}_X$ of the normalized orbital integral I_X is the Lie algebra analog of an irreducible tempered character Θ on $G(F)$. Waldspurger remarks that the Lie algebra analog of the L-packet of Θ is the set of distributions $\hat{I}_{X'}$ where X' ranges through the stable class of X. By Waldspurger's theorem that the Fourier transform preserves stability, the Fourier transform $\widehat{SI}_X = \sum_{X'} \hat{I}_{X'}$ is a stably invariant linear combination of the members in the "L-packet" of $\hat{I}_X$. Waldspurger then makes the following transfer conjecture, analogous to Shelstad's character identities for real groups. The conjecture involves the Fourier transform on $\mathfrak{h}(F)$, which must therefore be normalized properly, using the same additive character ψ as before and using a symmetric bilinear form on $\mathfrak{h}(F)$ deduced from the one of $\mathfrak{g}(F)$. In order to state this conjecture one must assume the truth of the matching conjecture.

WEAK TRANSFER CONJECTURE. See [22]. There is non-zero constant $c \in \mathbf{C}$ (which Waldspurger specifies precisely) such that for all G-regular semisimple elements $X_H \in \mathfrak{h}(F)$ and for all $f \in C_c^\infty(\mathfrak{g}(F))$

$$\widehat{SI}_{X_H}(f^H) = c \sum_{X_G} \Delta(X_H, X_G) \hat{I}_{X_G}(f), \tag{3}$$

where X_G ranges over a set of representatives for the $G(F)$-orbits in the set of images of X_H in $\mathfrak{g}(F)$. Equivalently, the Fourier transform commutes with the map $f \mapsto f^H$ from $C_c^\infty(\mathfrak{g}(F))$ to $SC_c^\infty(\mathfrak{h}(F))$, up to the scalar factor c.

If the matching and weak transfer conjectures are both true, then endoscopic induction commutes with the Fourier transform, up to the scalar c. It should be emphasized that the transfer factors appearing in the weak transfer conjecture are the same transfer factors as before. Thus, for groups there are both "geometric" and "spectral" transfer factors, while on Lie algebras the transfer factors $\Delta(X_H, X_G)$ play a double role.

Waldspurger also reformulates the weak transfer conjecture in such a way that it makes sense without assuming the matching conjecture. Assume for the moment

that the matching conjecture holds, so that endoscopic induction is defined. Let D_H be a stably invariant distribution on $\mathfrak{h}(F)$ and assume that D_H is a locally integrable function Θ_H on $\mathfrak{h}(F)$, locally constant on the regular semisimple set. The stable invariance of D_H is equivalent to the condition that Θ_H be constant on stable conjugacy classes. It follows from the Lie algebra analog of the Weyl integration formula that the invariant distribution $D_G = i_H^G(D_H)$ is a locally integrable function $i_H^G(\Theta_H)$ on $\mathfrak{g}(F)$, locally constant on the regular semisimple set, and that the value of the function $i_H^G(\Theta_H)$ on a regular semisimple element $X_G \in \mathfrak{g}(F)$ is given by

$$|D_G(X_G)|^{1/2} i_H^G(\Theta_H)(X_G) = \sum_{X_H} \Delta(X_H, X_G) \cdot |D_H(X_H)|^{1/2} \Theta_H(X_H), \qquad (4)$$

where the sum ranges over a set of representatives X_H for the stable conjugacy classes in $\mathfrak{h}(F)$ whose image in $\mathfrak{g}(F)$ is the stable conjugacy class of X_G.

By Harish-Chandra's fundamental local integrability theorem these considerations apply to Fourier transforms of orbital integrals. Thus $\hat{I}_{X_G}$ and $\widehat{SI}_{X_G}$ are given by locally integrable functions Θ_{X_G} and $S\Theta_{X_G}$. Then, assuming the matching conjecture holds, the weak transfer conjecture is equivalent to the following conjecture.

Transfer conjecture. See [22], [23]. There is non-zero constant $c \in \mathbf{C}$ (the same as before) such that for all G-regular semisimple elements $X_H \in \mathfrak{h}(F)$

$$i_H^G(S\Theta_{X_H}) = c \sum_{X_G} \Delta(X_H, X_G) \Theta_{X_G}, \qquad (5)$$

where X_G ranges over the $G(F)$-orbits in the set of images of X_H in $\mathfrak{g}(F)$.

Not only can the transfer conjecture be formulated without the matching conjecture, but in fact Waldspurger [22] VIII.7(8) shows that the transfer conjecture implies the matching conjecture. (The proof uses Harish-Chandra's theorem [5] that Shalika germs appear as coefficients in the Lie algebra analog of his local character expansion. Thus the transfer conjecture implies that κ-Shalika germs on $\mathfrak{g}(F)$ can be expressed in terms of stable Shalika germs on $\mathfrak{h}(F)$, and this, by work of Langlands-Shelstad [11], implies matching for both $G(F)$ and $\mathfrak{g}(F)$.)

There is one last conjecture to discuss, the Lie algebra analog of the fundamental lemma (see [12] for a discussion of the fundamental lemma on the group). For this we assume that G and H are unramified, and we let f, f^H denote the characteristic functions of hyperspecial parahoric subalgebras of $\mathfrak{g}(F)$, $\mathfrak{h}(F)$ respectively. The fundamental lemma (a conjecture, not a theorem) asserts that these particular functions satisfy equation (2) (for suitably normalized transfer factors). Using global methods (the Lie algebra analog of the trace formula), Waldspurger has shown that the fundamental lemma implies the transfer conjecture.

Theorem 1 (Waldspurger [23]) *Suppose that the fundamental lemma holds for all p-adic fields and all unramified G and H. Then the transfer conjecture is true, and hence the matching conjecture is true as well.*

3 Transfer factors in the quasi-split case

Once again we allow F to be any local field. We are going to give a simple formula for transfer factors on Lie algebras in the quasi-split case. So suppose that G is quasi-split and fix an F-splitting $(B_0, T, \{X_\alpha\})$ for G. Thus B_0 is a Borel F-subgroup of G, T is a maximal F-torus in B_0, and $\{X_\alpha\}$ is a collection of simple root vectors $X_\alpha \in \mathfrak{g}_\alpha$, one for each simple root α of T in the Lie algebra of B_0, such that $X_{\sigma\alpha} = \sigma(X_\alpha)$ for all $\sigma \in \Gamma$. (As usual, for any root β of T in $\mathfrak{g}$ we write $\mathfrak{g}_\beta$ for the corresponding root subspace of $\mathfrak{g}$.)

Waldspurger's factors are analogous to the transfer factors $\Delta(\gamma_H, \gamma_G; \bar{\gamma}_H, \bar{\gamma}_G)$ [10] with the factor Δ_{IV} removed. On the quasi-split group G Langlands and Shelstad also define transfer factors $\Delta_0(\gamma_H, \gamma_G)$ (see p. 248 of [10]). These depend on the chosen F-splitting. The transfer factors $\Delta_0(X_H, X_G)$ we consider now are complex roots of unity, analogous to $\Delta_0(\gamma_H, \gamma_G)$ with the factor Δ_{IV} removed, and they too depend on the choice of F-splitting.

We write $\mathfrak{b}_0$ for the Lie algebra of B_0. For each simple root α we define $X_{-\alpha} \in \mathfrak{g}_{-\alpha}$ by requiring that $[X_\alpha, X_{-\alpha}]$ be the coroot for α, viewed as element in the Lie algebra of T. We put $X_- := \sum_\alpha X_{-\alpha}$, where α runs over the set of simple roots of T in B_0. Of course X_- lies in $\mathfrak{g}(F)$ and depends on the choice of F-splitting. The following theorem is proved in [9].

THEOREM 2 *The factor $\Delta_0(X_H, X_G)$ is 1 whenever X_G lies in $\mathfrak{b}_0(F) + X_-$.*

Kostant [6] proved that every stable conjugacy class of regular semisimple elements in $\mathfrak{g}(F)$ meets the set $\mathfrak{b}_0(F) + X_-$. Since the values of $\Delta_0(X_H, X_G)$ and $\Delta_0(X_H, X'_G)$ are related by a simple Galois-cohomological factor (see (1)) whenever X_G and X'_G are stably conjugate, this theorem also yields a simple formula for the value of $\Delta_0(X_H, X_G)$ for arbitrary X_G. The methods used to prove the theorem are variants of ones used in [10], [13], [19]. In particular Proposition 5.2 in [13] plays a key role. What is new is the connection with Kostant's set $\mathfrak{b}_0(F) + X_-$.

4 Stability for nilpotent orbital integrals

It is especially interesting to study nilpotent orbital integrals from the point of view of endoscopy. The first question that comes to mind is which linear combinations of nilpotent orbital integrals are stably invariant. One can ask the same question for unipotent orbital integrals on the group. Let u be a unipotent element in $G(F)$. The stable conjugacy class of u is by definition the set of F-rational points on the G-conjugacy class of u. Assume for the moment that F is p-adic and that G is classical. Then Assem [1] made two conjectures concerning the stability of linear combinations of orbital integrals for orbits in the stable class of u. His first conjecture is that there are no non-zero stable combinations unless u is special in Lusztig's sense. Now assume that u is special. The second conjecture asserts that the set of conjugacy classes within the stable class of u can be decomposed as a disjoint union of "stability packets," in such a way that a suitable linear combination of the unipotent orbital integrals for the orbits within a single stability packet is stable, and moreover any stable linear combination is obtained as a sum

of these basic ones. The definition of stability packets involves Lusztig's quotient group of the component group of the centralizer of u. For split classical groups Assem further predicted that the relevant linear combination was simply the sum. Waldspurger has announced a proof of Assem's conjectures and has determined the linear combinations needed in the quasi-split case.

Now consider the real case. We work with nilpotent orbital integrals on a quasi-split Lie algebra. Then there is a formula [7, 8] for the dimension of the space of stable linear combinations of nilpotent orbital integrals for nilpotent orbits within a given stable class. The formula involves constructions of Lusztig and is valid for all quasi-split real groups, even the exceptional ones. We now state the formula precisely for split simple real groups.

Let $\mathbf{O}$ be a nilpotent $G(\mathbf{C})$-orbit in $\mathfrak{g}_{\mathbf{C}}$ and let $r_{\mathbf{O}}$ be the number of $G(\mathbf{R})$-orbits in $\mathfrak{g}(\mathbf{R}) \cap \mathbf{O}$. The linear span of the corresponding orbital integrals is an $r_{\mathbf{O}}$-dimensional space $\mathcal{D}_{\mathbf{O}}$ of (tempered [15]) invariant distributions on $\mathfrak{g}(\mathbf{R})$. The formula we are going to give for the dimension $s_{\mathbf{O}}$ of the subspace $\mathcal{D}_{\mathbf{O}}^{\mathrm{st}}$ of stably invariant elements in $\mathcal{D}_{\mathbf{O}}$ should be thought of as the stable analog of Rossmann's formula for $r_{\mathbf{O}}$, which we now recall.

Let T be a maximal torus in G and let W denote its Weyl group in $G(\mathbf{C})$. Then complex-conjugation, denoted σ, acts on W, and we consider the group W^{σ} of fixed points of σ on W. Inside W^{σ} we have the subgroup $W_{\mathbf{R}}$ consisting of elements in W that can be realized by elements in $G(\mathbf{R})$ that normalize T. Let R_I denote the set of imaginary roots of T (roots α such that $\sigma\alpha = -\alpha$). We define a sign character ϵ_I on W^{σ} in the usual way: $\epsilon_I(w) = (-1)^b$, where b is the number of positive roots $\alpha \in R_I$ such that $w\alpha$ is negative. By restriction we also regard ϵ_I as a character on $W_{\mathbf{R}}$.

Via Springer's correspondence [20] the nilpotent orbit $\mathbf{O}$ determines an irreducible character $\chi_{\mathbf{O}}$ of the abstract Weyl group W_a of $G(\mathbf{C})$. For example the trivial orbit $\mathbf{O} = \{0\}$ corresponds to the sign character ϵ of W_a. Of course we can also think of $\chi_{\mathbf{O}}$ as an irreducible character of the Weyl group W of any T as above, so that we can consider the multiplicity $m_{W_{\mathbf{R}}}(\epsilon_I, \chi_{\mathbf{O}})$ of ϵ_I in the restriction of $\chi_{\mathbf{O}}$ to the subgroup $W_{\mathbf{R}}$. Rossmann [17] proved that

$$r_{\mathbf{O}} = \sum_T m_{W_{\mathbf{R}}}(\epsilon_I, \chi_{\mathbf{O}}),$$

where T runs over the maximal $\mathbf{R}$-tori in G, up to $G(\mathbf{R})$-conjugacy.

THEOREM 3 *The integer $s_{\mathbf{O}}$ is given by*

$$s_{\mathbf{O}} = \sum_T m_{W^{\sigma}}(\epsilon_I, \chi_{\mathbf{O}}),$$

where the index set for the sum is the same as in Rossmann's formula and where $m_{W^{\sigma}}(\epsilon_I, \chi_{\mathbf{O}})$ denotes the multiplicity of ϵ_I in the restriction of $\chi_{\mathbf{O}}$ to W^{σ}.

The proof of this theorem uses the Fourier transform, and one must check that the Fourier transform of a stably invariant tempered distribution on $\mathfrak{g}(\mathbf{R})$ is stably invariant, as in the p-adic case [23], and this can be done using another

theorem of Rossmann [16]. The Fourier transforms of nilpotent orbital integrals are locally integrable functions on $\mathfrak{g}(\mathbf{R})$, and it is easy to recognize when such a locally integrable function represents a stably invariant distribution. Moreover, a description of the image of $\mathcal{D}_{\mathbf{O}}$ under the Fourier transform is implicit in the literature. (I am very much indebted to V. Ginzburg for this remark.) The theorem follows from these observations (see [8]).

In case G is quasi-split the right-hand side of our formula for $s_{\mathbf{O}}$ can be expressed (see [3], [7]) in terms of Lusztig's quotient groups. For simplicity we limit ourselves here to the case of simple split real groups. We define an integer $m(\chi)$ for any irreducible character χ on the abstract Weyl group W_a of $G(\mathbf{C})$ by the right-hand side of our formula for $s_{\mathbf{O}}$, but with $\chi_{\mathbf{O}}$ replaced by χ. Thus by the previous theorem $s_{\mathbf{O}} = m(\chi_{\mathbf{O}})$.

We need to review some results of Lusztig [14]. The set $W_a^\vee$ of isomorphism classes of irreducible representations of W_a is a disjoint union of subsets, called *families*. Associated to a family $\mathcal{F}$ is a finite group $\mathcal{G} = \mathcal{G}_{\mathcal{F}}$. If R is classical, then $\mathcal{G}$ is an elementary abelian 2-group. If R is exceptional, then $\mathcal{G}$ is one of the symmetric groups S_n $(1 \le n \le 5)$.

Associated to any finite group $\mathcal{G}$ is a finite set $\mathcal{M}(\mathcal{G})$, defined as follows (see [14]). Consider pairs (x, ρ), where x is an element of $\mathcal{G}$ and ρ is an irreducible (complex) representation of the centralizer $\mathcal{G}_x$ of x in $\mathcal{G}$. There is an obvious conjugation action of $\mathcal{G}$ on this set of pairs, and $\mathcal{M}(\mathcal{G})$ is by definition the set of orbits for this action. The set $\mathcal{M}(\mathcal{G})$ has an obvious basepoint $(1, 1)$, the first entry being the identity element of $\mathcal{G}$ and the second entry being the trivial representation of $\mathcal{G}$.

We define a function η on $\mathcal{M}(\mathcal{G})$ by

$$\eta(x, \rho) = \sum_{s \in S} d(s, \rho),$$

where S is a set of representatives for the $\mathcal{G}_x$-conjugacy classes of elements $s \in \mathcal{G}_x$ such that $s^2 = x$, and $d(s, \rho)$ denotes the dimension of the space of vectors in ρ fixed by the centralizer $\mathcal{G}_s$ of s in $\mathcal{G}$. (Note that $\mathcal{G}_s$ is a subgroup of $\mathcal{G}_x$, since $s^2 = x$.) In particular $\eta(x, \rho)$ is always a non-negative integer.

If $\mathcal{G}$ is an elementary abelian 2-group, then η is very simple: its value at the basepoint in $\mathcal{M}(\mathcal{G})$ is equal to the order of $\mathcal{G}$ and all remaining values are 0.

Let $\mathcal{F}$ be a family of representations of W_a, and let $\mathcal{G}$ be the associated finite group. Then Lusztig defines (case-by-case) an injection

$$\mathcal{F} \hookrightarrow \mathcal{M}(\mathcal{G}).$$

The image of the unique special representation in $\mathcal{F}$ under this injection is the base point $(1, 1) \in \mathcal{M}(\mathcal{G})$.

THEOREM 4 *Let E be an irreducible representation of the Weyl group W_a, with character χ. Let $x_E \in \mathcal{M}(\mathcal{G})$ denote the image of E under the injection $\mathcal{F} \hookrightarrow \mathcal{M}(\mathcal{G})$. Then $m(\chi)$ is equal to $\eta(x_E)$ if E is non-exceptional and is equal to 1 if E is exceptional. In particular if E is special and non-exceptional, then $m(\chi)$ is the*

number of conjugacy classes of involutions in $\mathcal{G}$. Moreover if R is classical, so that $\mathcal{G}$ is necessarily an elementary abelian 2-group, then $m(\chi) = 0$ if E is non-special, and $m(\chi) = |\mathcal{G}|$ if E is special.

See [2] for the notion of exceptional representations of Weyl groups; they occur only for E_7 and E_8. This theorem was proved for classical groups in [7] and for exceptional groups by Casselman in [3]. Note that if $\mathbf{O}$ is a special non-exceptional nilpotent orbit, then the last two theorems together tell us that the dimension of the space of stable linear combinations of $G(\mathbf{R})$-invariant measures on the $G(\mathbf{R})$-orbits in $\mathbf{O} \cap \mathfrak{g}(\mathbf{R})$ is equal to the cardinality of the Galois cohomology set $H^1(\mathbf{R}, \mathcal{G})$ (where $\mathrm{Gal}(\mathbf{C}/\mathbf{R})$ operates trivially on $\mathcal{G}$); this should be compared with Conjecture C in the introduction to Assem's paper [1].

References

[1] M. Assem. On stability and endoscopic transfer of unipotent orbital integrals on p-adic symplectic groups. Mem. Amer. Math. Soc. (to appear).

[2] W. M. Beynon and G. Lusztig. Some numerical results on the characters of exceptional Weyl groups. *Math. Proc. Cambridge Philos. Soc.*, 84:417–426, 1978.

[3] B. Casselman. Verifying Kottwitz' conjecture by computer. preprint, 1998.

[4] Harish-Chandra. The characters of reductive p-adic groups. In *Contributions to Algebra: A Collection of Papers Dedicated to Ellis Kolchin*, pages 175–182. Academic Press, New York, 1977.

[5] Harish-Chandra. Admissible invariant distributions on reductive p-adic groups. In W. Rossmann, editor, *Proceedings of the 1977 annual seminar of the Canadian Mathematical Congress*, number 48 in Queen's Papers in Pure and Applied Mathematics, pages 281–347, 1978.

[6] B. Kostant. Lie group representations on polynomial rings. *Amer. J. Math.*, 85:327–404, 1963.

[7] R. Kottwitz. Involutions in Weyl groups. preprint, 1998.

[8] R. Kottwitz. Stable nilpotent orbital integrals on real reductive Lie algebras. preprint, 1998.

[9] R. Kottwitz. Transfer factors for Lie algebras. preprint, 1998.

[10] R. Langlands and D. Shelstad. On the definition of transfer factors. *Math. Ann.*, 278:219–271, 1987.

[11] R. Langlands and D. Shelstad. Descent for transfer factors. In *The Grothendieck Festschrift, Vol. II*, Progr. Math., 87, pages 485–563. Birkhäuser, Boston, 1990.

[12] R. P. Langlands. *Les Débuts d'une Formule des Traces Stable.* Publ. Math. Univ. Paris VII, Vol. 13, 1983.

[13] R. P. Langlands. Orbital integrals on forms of $SL(3)$, I. *Amer. J. Math.*, 105:465–506, 1983.

[14] G. Lusztig. *Characters of Reductive Groups over a Finite Field.* Ann. of Math. Studies, 107. Princeton University Press, 1984.

[15] R. Rao. Orbital integrals in reductive groups. *Ann. of Math. (2)*, 96:505–510, 1972.

[16] W. Rossmann. Kirillov's character formula for reductive Lie groups. *Invent. Math.*, 48:207–220, 1978.

[17] W. Rossmann. Nilpotent orbital integrals in a real semisimple Lie algebra and representations of Weyl groups. In *Operator Algebras, Unitary Representations, Enveloping Algebras, and Invariant Theory*, Progr. Math., 92, pages 263–287. Birkhäuser, Boston, 1990.

[18] D. Shelstad. L-indistinguishability for real groups. *Math. Ann.*, 259:385–430, 1982.

[19] D. Shelstad. A formula for regular unipotent germs. *Astérisque*, 171–172:275–277, 1989.

[20] T. Springer. Trigonometric sums, Green functions of finite groups and representations of Weyl groups. *Invent. Math.*, 36:173–207, 1976.

[21] J.-L. Waldspurger. Comparaison d'intégrales orbitales pour des groupes p-adiques. In *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994)*, pages 807–816, Basel, 1995. Birkhäuser.

[22] J.-L. Waldspurger. Une formule des traces locale pour les algèbres de Lie p-adiques. *J. Reine Angew. Math.*, 465:41–99, 1995.

[23] J.-L. Waldspurger. Le lemme fondamental implique le transfert. *Compositio Math.*, 105:153–236, 1997.

Robert Kottwitz
Department of Mathematics
University of Chicago
5734 University Avenue
Chicago, IL 60637
USA
kottwitz@math.uchicago.edu

Chtoucas de Drinfeld et Applications

L. Lafforgue

1. Programme de Drinfeld-Langlands

Drinfeld a introduit un certain nombre de variétés modulaires, d'abord celles classifiant les modules elliptiques puis celles classifiant les chtoucas dans le but de réaliser dans leur cohomologie la correspondance de Langlands sur les corps de fonctions. Rappelons de quoi il s'agit.

On considère X une courbe projective lisse géométriquement connexe sur un corps fini $\mathbb{F}_q$ à q éléments, F son corps des fonctions, $|X|$ l'ensemble des points fermés de X identifiés aux places de F et $\mathbb{A}$ l'anneau des adèles de F. Pour toute place $x \in |X|$, on note $\deg(x)$ le degré sur $\mathbb{F}_q$ de son corps résiduel, F_x le complété de F suivant la valuation associée $\deg_x$ et O_x l'anneau des entiers de F_x. Ainsi, $\mathbb{A}$ est le produit restreint des F_x relativement aux O_x et son anneau des entiers $O_{\mathbb{A}}$ est le produit des O_x.

Fixons un nombre premier ℓ ne divisant pas q et notons $\overline{\mathbb{Q}}_\ell$ une clôture algébrique de $\mathbb{Q}_\ell$. Pour tout entier $r \geq 1$, soit Π_r l'espace des fonctions $\varphi : \mathrm{GL}_r(F) \backslash \mathrm{GL}_r(\mathbb{A}) \to \overline{\mathbb{Q}}_\ell$ localement constantes à support compact modulo $\mathbb{A}^\times$, invariantes par un sous-groupe d'indice fini de $\mathbb{A}^\times$ et telles que pour tout sous-groupe parabolique non trivial P de GL_r on ait

$$\int_{N_P(F)\backslash N_P(\mathbb{A})} \varphi(n_P g) \cdot dn_P = 0 \ , \ \forall g \in \mathrm{GL}_r(\mathbb{A}) \ ,$$

pour N_P le radical unipotent de P et dn_P une mesure de Haar rationnelle sur $N_P(\mathbb{A})$. Le groupe $\mathrm{GL}_r(\mathbb{A})$ agit sur Π_r par translation à droite. La représentation Π_r est somme d'une famille $\{\pi\}_r$ de représentations irréductibles π de $\mathrm{GL}_r(\mathbb{A})$ dites automorphes cuspidales. Chacune s'écrit de manière unique comme le produit restreint de représentations irréductibles π_x des groupes $\mathrm{GL}_r(F_x)$ qui, en toutes les places $x \in |X|$ sauf un nombre fini, sont non ramifiées au sens qu'elles ont un vecteur non nul invariant par $\mathrm{GL}_r(O_x)$; en ces places, il existe r nombres $z_1(\pi_x), \ldots, z_r(\pi_x)$ dans $\overline{\mathbb{Q}}_\ell$, bien définis à permutation près, tels que π_x soit un sous-quotient de l'induite du caractère $z_1(\pi_x)^{\deg_x(\cdot)} \times \cdots \times z_r(\pi_x)^{\deg_x(\cdot)}$ de $\mathrm{GL}_1(F_x) \times \cdots \times \mathrm{GL}_1(F_x)$ (avec la normalisation unitaire après un choix de $\sqrt{q} \in \overline{\mathbb{Q}}_\ell$).

Soit d'autre part $\{\sigma\}_r$ l'ensemble des représentations σ du groupe de Galois G_F de F, continues et irréductibles de dimension r sur $\overline{\mathbb{Q}}_\ell$, dont le déterminant est un caractère d'ordre fini et qui, en toutes les places $x \in |X|$ sauf un nombre fini, sont non ramifiées au sens que le sous-groupe d'inertie de x agit trivialement ; en chaque telle place x, les éléments de Frobenius induisent alors un automorphisme de l'espace sous-jacent à σ dont on note $\lambda_{1,x}(\sigma), \ldots, \lambda_{r,x}(\sigma)$ les valeurs propres.

CONJECTURE 1 (Langlands). — *Pour tout entier $r \geq 1$, on a :*

(i) *A toute représentation automorphe cuspidale $\pi \in \{\pi\}_r$, il est possible d'associer une unique représentation galoisienne $\sigma_\pi \in \{\sigma\}_r$ telle que, en toute place $x \in |X|$ où le facteur π_x est non ramifié, σ_π soit non ramifiée et*

$$\{\lambda_{1,x}(\sigma_\pi), \ldots, \lambda_{r,x}(\sigma_\pi)\} = \{z_1(\pi_x), \ldots, z_r(\pi_x)\} \ .$$

(ii) *Réproquement, pour toute $\sigma \in \{\sigma\}_r$, il existe une unique $\pi \in \{\pi\}_r$, non ramifiée partout où σ est non ramifiée et telle que $\sigma = \sigma_\pi$ au sens de* (i).

CONJECTURE 2 (Ramanujan-Petersson). — *Pour tout entier $r \geq 1$, toute représentation automorphe cuspidale $\pi \in \{\pi\}_r$ et toute place $x \in |X|$ où le facteur π_x est non ramifié, les nombres*

$$z_1(\pi_x), \ldots, z_r(\pi_x)$$

sont algébriques et de module 1 *après tout plongement dans* $\mathbb{C}$.

Quand $r = 1$, la conjecture 1 exprime la loi de réciprocité globale sur les corps de fonctions ; une démonstration géométrique en fut donnée par Lang et Rosenlicht (voir [S]). La conjecture 2 pour $r = 1$ est tautologique.

En rang r arbitraire, l'unicité dans la conjecture 1(i) résulte du théorème de densité de Čebotarev et l'unicité dans la conjecture 1(ii) est connue d'après le "théorème de multiplicité un fort".

On montre d'autre part en utilisant la théorie de Hecke inverse et l'équation fonctionnelle des fonctions L de Grothendieck que si la conjecture 1(i) est vraie en tous rangs $< r$, alors la conjecture 1(ii) l'est en tous rangs $\leq r$ (voir [Lau 1]).

L'étude de la cohomologie ℓ-adique des variétés modulaires de Drinfeld classifiant les modules elliptiques [resp. les chtoucas] de rang r permet d'attaquer les conjectures 1(i) et 2 en rang r arbitraire. Il s'agit d'identifier dans la cohomologie de ces variétés des morceaux de la forme $\pi \otimes \sigma_\pi$ [resp. $\pi \otimes \sigma_\pi \otimes \check{\sigma}_\pi$] avec π décrivant une partie de [resp. tout] l'ensemble $\{\pi\}_r$. Pour ce faire, on combine chaque fois le théorème des points fixes de Grothendieck-Lefschetz, la formule des traces d'Arthur-Selberg et, pour la conjecture 2, le théorème de pureté de Deligne.

En étudiant la cohomologie à coefficients constants [resp. à coefficients dans certains systèmes locaux] des variétés de modules elliptiques de rang 2, Drinfeld a d'abord démontré les conjectures 1(i) et 2 quand $r = 2$ et l'une au moins des composantes π_x de π est la représentation de Steinberg [resp. est supercuspidale]. Ce travail a été généralisé en rang arbitraire par Laumon (voir [Lau 2]) [resp. par Flicker et Kazhdan (voir [FK])].

Après cela, Drinfeld a démontré la conjecture 2 puis la conjecture 1(i) pour $r = 2$ et sans restriction sur π en étudiant la cohomologie de certains ouverts de type fini des variétés de chtoucas de rang 2 (voir [D1] et [D2]) puis de compactifications de ces ouverts qu'il a construites (voir [D3]). D'ailleurs, la théorie des chtoucas de rang 1 reformule de façon particulièrement élégante la démonstration de Lang.

Enfin, Stuhler a défini une notion de D-module elliptique, pour D une algèbre centrale simple sur F, généralisant celle de module elliptique. L'étude des variétés modulaires associées a permis à Laumon, Rapoport et Stuhler de démontrer la correspondance de Langlands locale sur les corps de fonctions en tous rangs (voir [LRS]).

2. CHTOUCAS ET CONJECTURE DE RAMANUJAN-PETERSSON

Rappelons la définition des chtoucas sur la courbe X.
Etant donné S un schéma (sur $\mathbb{F}_q$) et $\mathcal{E}$ un $\mathcal{O}_{X\times S}$-Module localement libre de rang r sur $X\times S$, on appelle MODIFICATION (à droite) de $\mathcal{E}$ tout diagramme $(\mathcal{E} \overset{j}{\hookrightarrow} \mathcal{E}' \overset{t}{\hookleftarrow} \mathcal{E}'')$ où $\mathcal{E}'$, $\mathcal{E}''$ sont aussi des fibrés localement libres de rang r sur $X\times S$ et j,t sont des homomorphismes injectifs dont les conoyaux sont supportés par les graphes de deux morphismes "pôle" $\infty : S\to X$ et "zéro" $0 : S\to X$ et inversibles comme $\mathcal{O}_S$-Modules.
Notant Frob_S l'endomorphisme de Frobenius d'élévation à la puissance q dans tout schéma S sur $\mathbb{F}_q$, on appelle CHTOUCA DE RANG r sur S la donnée d'un $\mathcal{O}_{X\times S}$-Module $\mathcal{E}$ localement libre de rang r sur $X\times S$, d'une modification $(\mathcal{E}\hookrightarrow\mathcal{E}'\hookleftarrow\mathcal{E}'')$ de $\mathcal{E}$ et d'un isomorphisme ${}^{\tau}\mathcal{E} = (\mathrm{Id}_X\times\mathrm{Frob}_S)^*\mathcal{E}\overset{\sim}{\to}\mathcal{E}''$.
En associant à tout schéma S le groupoïde des chtoucas de rang r sur S, on définit un champ Cht^r qui est algébrique au sens de Deligne-Mumford. Comme tout chtouca a par définition un "zéro" et un "pôle", le champ Cht^r est muni d'un morphisme sur $X\times X$ qui s'avère lisse (et en particulier localement de type fini) de dimension relative $2r-2$.
Si maintenant $N=\mathrm{Spec}\,\mathcal{O}_N\hookrightarrow X$ est un sous-schéma fermé fini de la courbe X, on appelle structure de niveau N sur un chtouca $(\mathcal{E}\hookrightarrow\mathcal{E}'\hookleftarrow\mathcal{E}''\cong{}^{\tau}\mathcal{E})$ de rang r sur un schéma S dont le zéro et le pôle évitent N tout isomorphisme $g:\mathcal{E}\otimes_{\mathcal{O}_X}\mathcal{O}_N\overset{\sim}{\to}\mathcal{O}_{N\times S}^r$ tel que $\tau(g)=g\circ u$ où $u:{}^{\tau}\mathcal{E}\otimes_{\mathcal{O}_X}\mathcal{O}_N\overset{\sim}{\to}\mathcal{E}\otimes_{\mathcal{O}_X}\mathcal{O}_N$ est l'isomorphisme induit. Le champ Cht^r_N des chtoucas de rang r avec structure de niveau N est algébrique au sens de Deligne-Mumford. Le morphisme d'oubli $\mathrm{Cht}^r_N\to\mathrm{Cht}^r\times_{X\times X}(X-N)\times(X-N)$ est représentable fini étale galoisien de groupe $\mathrm{GL}_r(\mathcal{O}_N)$.
Notons K_N le noyau de chacun des homomorphismes surjectifs $\mathrm{GL}_r(O_{\mathbb{A}})\to\mathrm{GL}_r(\mathcal{O}_N)$. Prolongeant l'action de $\mathrm{GL}_r(\mathcal{O}_N)$, les doubles classes de $K_N\backslash\mathrm{GL}_r(\mathbb{A})/K_N$ induisent des correspondances dans les Cht^r_N compatibles avec les projections sur $X\times X$. Ce sont les CORRESPONDANCES DE HECKE.
D'autre part, Cht^r et les Cht^r_N sont munis d'endomorphismes dits "de Frobenius partiels" Frob_0 et Frob_∞ qui relèvent $\mathrm{Frob}_X\times\mathrm{Id}_X$ et $\mathrm{Id}_X\times\mathrm{Frob}_X$, dont les composés dans un sens ou dans l'autre sont les endomorphismes de Frobenius et qui commutent avec les correspondances de Hecke. D'après la proposition suivante, leur existence fait que la cohomologie ℓ-adique de $\varprojlim_N\mathrm{Cht}^r_N$ sur le point générique de $X\times X$ est munie d'une action de $\mathrm{GL}_r(\mathbb{A})\times\mathrm{G}_F\times\mathrm{G}_F$:

PROPOSITION (Drinfeld). — *Pour tout ouvert X' de X et notant $\pi(X')$ et $\pi(X'\times X')$ les groupes fondamentaux de X' et $X'\times X'$, il y a équivalencce entre la catégorie des représentations continues de $\pi(X')\times\pi(X')$ et celle des*

représentations continues H de $\pi(X' \times X')$ qui sont munies d'un isomorphisme équivariant $(\mathrm{Frob}_{X'} \times \mathrm{Id}_{X'})^* H \cong H$.

De voir cette cohomologie comme une représentation de $\mathrm{GL}_r(\mathbb{A}) \times \mathrm{G}_F \times \mathrm{G}_F$ donne un sens à la conjecture qu'elle contient comme facteurs irréductibles les $\pi \otimes \sigma_\pi \otimes \check{\sigma}_\pi$ avec π décrivant $\{\pi\}_r$.

Afin de calculer la cohomologie d'une variété définie en termes modulaires, on dispose du théorème des points fixes de Grothendieck-Lefschetz. Mais comme Cht^r n'est pas quasi-compact ni même ses composantes connexes quand $r \geq 2$, ce théorème ne peut s'appliquer dans Chr^r (et les Chr^r_N) qu'à des ouverts de type fini qu'il faut donc définir. On le fait en tronquant par le polygone canonique de Harder-Narasimhan (voir [L1]) :
Si $\widetilde{\mathcal{E}} = (\mathcal{E} \hookrightarrow \mathcal{E}' \hookleftarrow \mathcal{E}'' = {}^\tau\mathcal{E})$ est un chtouca de rang r sur un corps, on appelle sous-objet $\widetilde{\mathcal{F}}$ de $\widetilde{\mathcal{E}}$ tout couple de sous-fibrés $\mathcal{F}$, $\mathcal{F}'$ de $\mathcal{E}$, $\mathcal{E}'$ tels que $\mathcal{E} \hookrightarrow \mathcal{E}'$, ${}^\tau\mathcal{E} \hookrightarrow \mathcal{E}'$ induisent deux homomorphismes $\mathcal{F} \hookrightarrow \mathcal{F}'$, ${}^\tau\mathcal{F} \hookrightarrow \mathcal{F}'$ génériquement bijectifs. Un tel sous-objet a un rang $\mathrm{rg}\widetilde{\mathcal{F}} = \mathrm{rg}\mathcal{F} = \mathrm{rg}\mathcal{F}'$ et il a par exemple le degré $\deg \widetilde{\mathcal{F}} = \deg \mathcal{F}$. A toute filtration $0 = \widetilde{\mathcal{F}}_0 \subsetneq \cdots \subsetneq \widetilde{\mathcal{F}}_i \subsetneq \cdots \subsetneq \widetilde{\mathcal{F}}_k = \widetilde{\mathcal{E}}$ de $\widetilde{\mathcal{E}}$ par des sous-objets, on peut associer son polygone c'est-à-dire l'unique application $[0, r] \to \mathbb{R}$ affine sur les intervalles $[\mathrm{rg}\widetilde{\mathcal{F}}_{i-1}, \mathrm{rg}\widetilde{\mathcal{F}}_i]$ et qui vaut $\deg \widetilde{\mathcal{F}}_i - \frac{\mathrm{rg}\widetilde{\mathcal{F}}_i}{r} \deg \widetilde{\mathcal{E}}$ en chaque $\mathrm{rg}\widetilde{\mathcal{F}}_i$. La famille des polygones de toutes les filtrations de $\widetilde{\mathcal{E}}$ a un plus grand élément qu'on appelle le POLYGONE CANONIQUE DE HARDER-NARASIMHAN de $\widetilde{\mathcal{E}}$.
Pour tout entier $d \in \mathbb{Z}$ et tout polygone $p : [0, r] \to \mathbb{R}$, les chtoucas de rang r, de degré d et dont le polygone canonique est majoré par p sont les points d'un ouvert $\mathrm{Cht}^{r,d,p}$ de Cht^r. Cet ouvert est de type fini et il en est de même des $\mathrm{Cht}^{r,d,p}_N = \mathrm{Cht}^{r,d,p} \times_{\mathrm{Cht}^r} \mathrm{Cht}^r_N$.
Les $\mathrm{Cht}^{r,d,p}_N$ ne sont plus stables par les correspondances de Hecke ni par les endomorphismes de Frobenius partiels mais on peut malgré tout considérer l'action sur leur cohomologie des puissances de l'endomorphisme de Frobenius et leur appliquer le théorème des points fixes de Grothendieck-Lefschetz. Le lien avec la formule des traces d'Arthur-Selberg s'établit en montrant en particulier que lorsqu'on identifie le quotient $\mathrm{GL}_r(F) \backslash \mathrm{GL}_r(\mathbb{A}) / \mathrm{GL}_r(O_\mathbb{A})$ avec le groupoïde des fibrés de rang r sur la courbe X, les troncatures d'Arthur sur les groupes adéliques correspondent exactement aux troncatures par le polygone canonique de Harder-Narasimhan des fibrés. Pour $0, \infty \in |X|$ deux places distinctes, u un multiple commun de $\deg(0)$ et $\deg(\infty)$ et $\pi \in \{\pi\}_r$ une représentation automorphe cuspidale non ramifiée en 0 et ∞, les nombres $q^{(r-1)u} z_i(\pi_0)^{\frac{u}{\deg(0)}} z_j(\pi_\infty)^{\frac{-u}{\deg(\infty)}}$, $1 \leq i, j \leq r$, apparaissent parmi les valeurs propres de Frob^u agissant sur la cohomologie des fibres des ouverts $\mathrm{Cht}^{r,d,p}_N$ au-dessus des points de $X \times X$ supportés par $(0, \infty)$ et on peut leur appliquer le théorème de pureté de Deligne. On obtient (voir [L1]) :

THÉORÈME. — (i) *Pour tout entier $r \geq 1$ et toute représentation automorphe cuspidale $\pi \in \{\pi\}_r$, on a :*

- *Si r est impair, π vérifie la conjecture 2 et on pose $\varepsilon_\pi = 0$.*
- *Si r est pair, ou bien π vérifie la conjecture 2 et on pose $\varepsilon_\pi = 0$, ou bien pour toute place x où π est non ramifiée et tout isomorphisme $\overline{\mathbb{Q}}_\ell \cong \mathbb{C}$, la*

moitié des $z_i(\pi_x)$, $1 \le i \le r$, sont de module $q^{\frac{1}{4}\deg(x)}$ et l'autre moitié de module $q^{-\frac{1}{4}\deg(x)}$ et on pose $\varepsilon_\pi = \frac{1}{4}$.

(ii) *Pour deux telles représentations $\pi \in \{\pi\}_r$, $\pi' \in \{\pi\}_{r'}$, tous les zéros de la fonction de Rankin-Selberg $L(s, \pi \otimes \check{\pi}')$ sont sur la droite $\mathrm{Re}\, s = \frac{1}{2}$ si $\varepsilon_\pi = \varepsilon_{\pi'}$ et sur les droites $\mathrm{Re}\, s = \frac{1}{4}$, $\mathrm{Re}\, s = \frac{3}{4}$ si $\varepsilon_\pi \neq \varepsilon_{\pi'}$.*

3. Compactifications

En passant des Cht_N^r aux ouverts de type fini $\mathrm{Cht}_N^{r,d,p}$, nous avons perdu l'action des correspondances de Hecke et des endomorphismes de Frobenius partiels. Afin de la retrouver, nous construisons des compactifications des $\mathrm{Cht}_N^{r,d,p}$, comme Drinfeld a fait en rang $r = 2$.

a) Dégénérescence des chtoucas sans structures de niveau

L'idée est d'élargir la notion de chtoucas en remplaçant dans leurs diagrammes de définition $(\mathcal{E} \hookrightarrow \mathcal{E}' \hookleftarrow \mathcal{E}'' \cong {}^\tau\mathcal{E})$ les isomorphismes ${}^\tau\mathcal{E} \xrightarrow{\sim} \mathcal{E}''$ par les "homomorphismes complets".

Le schéma Ω^r des HOMOMORPHISMES COMPLETS est celui déduit du schéma des matrices carrées non nulles de rang r en éclatant les $r-1$ fermés des matrices de rang $\le 1, \le 2, \ldots, \le r-1$. Il est muni d'une action de $\mathrm{GL}_r^2/\mathbb{G}_m$ et son quotient $\overline{\Omega}^r$ par l'action du centre $\mathbb{G}_m^2/\mathbb{G}_m$ est le compactifié de De Concini et Procesi de $\mathrm{PGL}_r^2 / \mathrm{PGL}_r$.

Soit $\mathcal{P}^r$ le champ torique quotient de l'espace affine $\mathbb{A}^{r-1}$ par $\mathbb{G}_m^{r-1}$. Ses points correspondent aux orbites de $\mathbb{G}_m^{r-1}$ agissant sur $\mathbb{A}^{r-1}$; celles-ci sont indexées par les partitions $\underline{r} = (0 = r_0 < r_1 < \cdots < r_k = r)$ de l'entier r et chacune contient le point distingué $\alpha_{\underline{r}}$ dont la s-ième coordonnée, $1 \le s < r$, vaut 0 si $s \in \underline{r}$ et 1 si $s \notin \underline{r}$. Le schéma Ω^r est muni d'un morphisme lisse de dimension relative r^2 sur $\mathcal{P}^r$ et sa fibre $\Omega^r_{\underline{r}}$ au-dessus de chaque $\alpha_{\underline{r}}$ classifie les familles constituées d'une filtration croissante (E_s) de $\mathbb{A}^r$ par des sous-espaces de dimension s, $s \in \underline{r}$, d'une filtration décroissante $(\overline{E}_s)$ de $\mathbb{A}^r$ par des sous-espaces de codimension s, $s \in \underline{r}$, et d'isomorphismes $\overline{E}_{s^-}/\overline{E}_s \xrightarrow{\sim} E_s/E_{s^-}$ où s^- désigne le prédécesseur de tout élément $s > 0$ de $\underline{r}$.

On appelle CHTOUCA ITÉRÉ DE RANG r sur un schéma S la donnée d'un $\mathcal{O}_{X\times S}$-Module $\mathcal{E}$ localement libre de rang r, d'une modification $(\mathcal{E} \hookrightarrow \mathcal{E}' \hookleftarrow \mathcal{E}'')$ de celui-ci et d'un homomorphisme complet ${}^\tau\mathcal{E} - - \to \mathcal{E}''$ sur $X \times S$ dont le point image dans $\mathcal{P}^r$ provient d'un point à valeurs dans S et qui est soumis à certaines conditions ouvertes. Le champ $\overline{\mathrm{Cht}^r}$ des chtoucas itérés de rang r est algébrique au sens de Deligne-Mumford (mais non séparé) et localement de type fini sur $X \times X \times \mathcal{P}^r$.

Pour $\underline{r} = (0 = r_0 < r_1 < \cdots < r_k = r)$ une partition de r, soit $\mathrm{Cht}^{\underline{r}}$ le champ des familles de chtoucas de rangs $s - s^-$, $s \in \underline{r}$, $s > 0$, tels que le zéro de chacun soit égal au pôle du suivant ; il est lisse de dimension relative $2r - 2k$ sur $X \times X \times X^{k-1}$. La fibre $\overline{\mathrm{Cht}}^{\underline{r}}_{\underline{r}}$ de $\overline{\mathrm{Cht}^r}$ au-dessus du point $\mathbb{G}_m^{r-1}\backslash\mathbb{G}_m^{r-1}\alpha_{\underline{r}}$ de $\mathcal{P}^r$ est munie d'un morphisme fini, surjectif et radiciel sur $\mathrm{Cht}^{\underline{r}}$ (d'où le nom de "chtoucas itérés").

Disons qu'un polygone $p : [0, r] \to \mathbb{R}$ est T-grand, pour $T \ge 0$ une constante, si $[p(r'+1) - p(r')] - [p(r') - p(r'-1)] \ge T$, $1 \le r' < r$. On a (voir [L2]) :

THÉORÈME. — *A tout entier $d \in \mathbb{Z}$ et tout polygone 2-grand $p : [0, r] \to \mathbb{R}$ on peut associer un ouvert $\overline{\mathrm{Chr}^{r,d,p}}$ de $\overline{\mathrm{Cht}^r}$ dont la trace dans l'ouvert Cht^r est $\mathrm{Chr}^{r,d,p}$ et qui est propre (en particulier séparé et de type fini) sur $X \times X$.*
Si de plus p est T_X-grand, pour $T_X \geq 2$ une constante ne dépendant que de X, $\overline{\mathrm{Chr}^{r,d,p}}$ est lisse de dimension relative $2r-2$ sur $X \times X \times \mathcal{P}^r$.

Cette construction généralise celle de Drinfeld en rang $r = 2$.

b) DÉGÉNÉRESCENCE DES STRUCTURES DE NIVEAU DES CHTOUCAS

En rang $r = 2$, Drinfeld a compactifié les variétés de chtoucas avec structures de niveau N en prenant simplement les normalisations dans le corps des fonctions de Cht^r_N des compactifications sans structures de niveau. En rang arbitraire, nous procédons différemment afin de garder une description modulaire et un contrôle sur les singularités.

COMPACTIFICATION DU CLASSIFIANT DE PGL_r : Considérant de façon générale $N = \mathrm{Spec}\,\mathcal{O}_N$ un schéma fini sur $\mathbb{F}_q$ et $N_0 = \mathrm{Spec}\,\mathcal{O}_{N_0}$ le schéma réduit associé, on note GL_r^N et $\mathbb{G}_m^{N_0}$ les groupes algébriques déduits de GL_r et $\mathbb{G}_m$ par restriction des scalaires à la Weil de $\mathcal{O}_N$ et $\mathcal{O}_{N_0}$ à $\mathbb{F}_q$ et aussi $\overline{\mathrm{GL}}_r^N = \mathrm{GL}_r^N / \mathbb{G}_m^{N_0}$. Pour tout entier $n \geq 1$, nous construisons (voir [L3] et [L4]) une compactification équivariante $\overline{\Omega}^{r,N,n}$ du quotient diagonal $(\overline{\mathrm{GL}}_r^N)^{n+1}/\overline{\mathrm{GL}}_r^N$ qui est lisse au-dessus du champ $\mathcal{P}^{r,N,n}$ quotient d'une variété torique $\mathcal{A}^{r,N,n}$ par son tore $\mathcal{A}^{r,N,n}_\emptyset$. Les points de ce champ torique c'est-à-dire les orbites de $\mathcal{A}^{r,N,n}_\emptyset$ agissant sur $\mathcal{A}^{r,N,n}$ sont naturellement indexées par une famille de pavages d'un polyèdre convexe $S^{r,N,n}$ (qui est le simplexe $\{(x_0, \ldots, x_n) \in \mathbb{R}_+^{n+1} | x_0 + \cdots + x_n = r\}$ quand $\mathcal{O}_N = \mathbb{F}_q$) et ce de telle façon qu'une orbite est dans l'adhérence d'une autre si et seulement si son pavage associé raffine celui de l'autre.
Il y a sur chaque $\overline{\Omega}^{r,N,n}$ un torseur $\Omega^{r,N,n}$ pour le tore $(\mathbb{G}_m^{N_0})^{n+1}/\mathbb{G}_m^{N_0}$ muni d'une action compatible de $(\mathrm{GL}_r^N)^{n+1}$ et dont la restriction à l'ouvert $(\overline{\mathrm{GL}}_r^N)^{n+1}/\overline{\mathrm{GL}}_r^N$ est $(\mathrm{GL}_r^N)^{n+1}/\mathrm{GL}_r^N$. Les fibres de $\Omega^{r,N,n}$ au-dessus des points de $\mathcal{P}^{r,N,n}$ ont une description modulaire en termes des pavages de $S^{r,N,n}$ associés à ces points qui généralise celle des fibres $\Omega^r_{\underline{r}}$ de Ω^r.
Les applications $\iota : \{0, 1, \ldots, m\} \to \{0, 1, \ldots, n\}$ induisent des morphismes $\mathcal{A}^{r,N,n}_\emptyset \to \mathcal{A}^{r,N,m}_\emptyset$, $\mathcal{A}^{r,N,n} \to \mathcal{A}^{r,N,m}$ et $\Omega^{r,N,n} \to \Omega^{r,N,m}$ compatibles entre eux et avec les $(\mathrm{GL}_r^N)^{n+1} \to (\mathrm{GL}_r^N)^{m+1}$ si bien que les $\Omega^{r,N,n}$ constituent un schéma simplicial qui prolonge le classifiant $((\mathrm{GL}_r^N)^{n+1}/\mathrm{GL}_r^N)_{n\geq 1}$ du groupe GL_r^N et que les $\overline{\Omega}^{r,N,n}$ réalisent une sorte de compactification du classifiant de $\overline{\mathrm{GL}}_r^N$. Quand ι est injective, le morphisme induit $\Omega^{r,N,n} \to \Omega^{r,N,m} \times_{\mathcal{P}^{r,N,m}} \mathcal{P}^{r,N,n}$ est lisse.

APPLICATION AUX CHTOUCAS : Faisons le lien avec les chtoucas :

LEMME. — *Etant donné $N \hookrightarrow X$ un sous-schéma fermé fini de X, faisons agir GL_r^N sur $\Omega^{r,N,n}$ via l'action de $(\mathrm{GL}_r^N)^2$ et le plongement $\mathrm{GL}_r^N \hookrightarrow (\mathrm{GL}_r^N)^2$, $g \mapsto (\tau(g), g)$.*
Alors le foncteur $\cdot \otimes_{\mathcal{O}_X} \mathcal{O}_N$ définit un morphisme

$$\overline{\mathrm{Cht}^r} \times_{X\times X} (X-N)\times(X-N) \to \mathrm{GL}_r^N \backslash \Omega^{r,N,1} \times_{\mathcal{P}^{r,N,1}} \mathcal{P}^r$$

et pour tout entier d et tout polygone T_X-grand p, le morphisme induit

$$\overline{\mathrm{Cht}^{r,d,p}} \times_{X\times X} (X-N)\times(X-N) \to \mathrm{GL}_r^N \backslash \Omega^{r,N,1} \times_{\mathcal{P}^{r,N,1}} \mathcal{P}^r \times (X\times X)$$

est lisse.

Notons p_0, p_1, p_2 les morphismes $\Omega^{r,N,2} \to \Omega^{r,N,1}$ ou $\mathcal{A}^{r,N,2} \to \mathcal{A}^{r,N,1}$ induits par les trois injections croissantes $\{0,1\} \to \{0,1,2\}$ d'images $\{1,2\}, \{0,2\}, \{0,1\}$.
Dans la catégorie des variétés toriques et de leurs morphismes équivariants, le diagramme

$$\begin{array}{ccc} \mathbb{A}^{r-1} & \longrightarrow & \mathcal{A}^{r,N,1} \\ & \overset{p_2}{\nearrow} & \\ \mathcal{A}^{r,N,2} & \underset{p_0}{\overset{\tau\circ p_1}{\rightrightarrows}} & \mathcal{A}^{r,N,1} \end{array}$$

a une limite projective $\mathcal{A}^{r,N,\tau}$ de tore $\mathcal{A}_\emptyset^{r,N,\tau}$. Notant $\mathcal{P}^{r,N,\tau}$ le champ torique quotient de $\mathcal{A}^{r,N,\tau}$ par $\mathcal{A}_\emptyset^{r,N,\tau}$, le diagramme

$$\Omega^{r,N,2} \times_{\mathcal{P}^{r,N,2}} \mathcal{P}^{r,N,\tau} \underset{p_0}{\overset{\tau\circ p_1}{\rightrightarrows}} \Omega^{r,N,1}$$

a un noyau $\Omega^{r,N,\tau}$ qui est lisse sur $\mathcal{P}^{r,N,\tau}$ puisque $p_0 : \Omega^{r,N,2} \to \Omega^{r,N,1} \times_{\mathcal{P}^{r,N,1}} \mathcal{P}^{r,N,2}$ est lisse. De plus, $p_2 : \Omega^{r,N,\tau} \to \Omega^{r,N,1} \times_{\mathcal{P}^{r,N,1}} \mathcal{P}^r$ est représentable et projectif et sa restriction au-dessus de $(\mathrm{GL}_r^N)^2/\mathrm{GL}_r^N \cong \mathrm{GL}_r^N$ est l'isogénie de Lang $g \mapsto g^{-1}\circ\tau(g)$.
Pour tout entier d et tout polygone T_X-grand p, le produit fibré

$$\overline{\mathrm{Cht}_N^{r,d,p}} = (\overline{\mathrm{Cht}^{r,d,p}} \times_{X\times X} (X-N)\times(X-N)) \times_{\mathrm{GL}_r^N \backslash \Omega^{r,N,1}\times\mathcal{P}^r} \mathrm{GL}_r^N \backslash \Omega^{r,N,\tau}$$

est donc représentable et projectif sur $\overline{\mathrm{Cht}^{r,d,p}} \times_{X\times X} (X-N)\times(X-N)$. Sa restriction au-dessus de l'ouvert $\mathrm{Cht}^{r,d,p}$ est $\mathrm{Cht}_N^{r,d,p}$. Et d'après le lemme, il est lisse sur $X\times X\times\mathcal{P}^{r,N,\tau}$.

Références bibliographiques

[D1] V.G. Drinfeld, "Varieties of modules of F-sheaves", p. 107-122, Funct. Anal. and its Appl. 21, 1987.

[D2] V.G. Drinfeld, "The proof of Petersson's conjecture for GL(2) over a global field of characteristic p", p. 28-43, Funct. Anal. and its Appl. 22, 1988.

[D3] V.G. Drinfeld, "Cohomology of compactified manifolds of modules of F-sheaves of rank 2", p. 1789-1821, J. of Soviet Math. 46, 1989.

[FK] Y. Flicker et D. Kazhdan, "Geometric Ramanujan conjecture and Drinfeld reciprocity law", p. 201-218, dans "Number theory, trace formula and discrete groups", Academic Press, 1989.

[L1] L. Lafforgue, "Chtoucas de Drinfeld et conjecture de Ramanujan-Petersson", Astérisque 243, SMF, 1997.

[L2] L. Lafforgue, "Une compactification des champs classifiant les chtoucas de Drinfeld", à paraître dans le Journal of the AMS.

[L3] L. Lafforgue, "Pavages des simplexes, schémas de graphes recollés et compactification des $\mathrm{PGL}_r^{n+1} / \mathrm{PGL}_r$", prépublication d'Orsay, 1998.

[L4] L. Lafforgue, "Compactification des $\mathrm{PGL}_r^{n+1} / \mathrm{PGL}_r$ et restriction des scalaires à la Weil", prépublication d'Orsay, 1998.

[Lau 1] G. Laumon, "Transformation de Fourier, constantes d'équations fonctionnelles et conjecture de Weil", p. 131-210, Publ. Math. I.H.E.S. 65, 1987.

[Lau 2] G. Laumon, "Cohomology of Drinfeld Modular Varieties I, II", Cambridge University Press, 1996-97.

[LRS] G. Laumon, M. Rapoport et U. Stuhler "D-elliptic sheaves and the Langlands correspondence", p. 217-338, Inv. Math. 113, 1993.

[S] J.-P. Serre, "Groupes algébriques et corps de classes", Hermannn, 1959.

Laurent Lafforgue
Mathématique
Université Paris-Sud
Bâtiment 425
F-91405 Orsay Cedex
France

Products of Trees, Lattices and Simple Groups

Shahar Mozes

1991 Mathematics Subject Classification: 20E06, 20E08, 20E32, 20F32, 22D40, 22E40

The group of automorphisms of a locally finite tree, denoted $\operatorname{Aut} T$, is a locally compact group which exhibits behavior analogous to that of a rank one simple Lie group. This analogy has motivated many recent works, in particular the study of lattices in $\operatorname{Aut} T$ by Bass, Kulkarni, Lubotzky and others. Recall that in the case of semisimple Lie groups, irreducible lattices in higher rank groups have a very rich structure theory and one encounters many deep and interesting phenomena such as (super)rigidity and arithmeticity. Motivated by this we study in several joint works with Marc Burger and with Robert J. Zimmer cocompact lattices in the group of automorphisms of a product of trees, or rather in groups of the form $\operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ where each of the trees T_i, $i = 1, 2$, is (bi-)regular. The results obtained concerning the structure of lattices in $\operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ enable us to construct the first examples of finitely presented torsion free simple groups, see [BM2].

One Tree. There is a close relation between certain simple Lie groups and groups of tree automorphisms. Let G be a simple algebraic group of rank one over a non-archimedean local field K. Considering the action of G on its associated Bruhat-Tits tree T, we have a continuous embedding of G in $\operatorname{Aut} T$ with cocompact image. In [Tit], Tits has shown that if T is a locally finite tree and its automorphism group $\operatorname{Aut} T$ acts minimally (i.e. without an invariant proper subtree and not fixing an end) on it, then the subgroup $\operatorname{Aut}^+ T$ generated by edge stabilizers is a simple group. In particular the automorphism group of a regular tree is virtually simple. These results motivated the study of $\operatorname{Aut} T$ taking this analogy with rank one Lie groups as a guideline, see [Lu1], [Lu3].

When T is a locally finite tree its automorphism group is locally compact. Recall that a subgroup Γ of a locally compact group G is called a lattice when it is discrete and the quotient $\Gamma\backslash G$ carries a finite invariant measure. In case the quotient is compact the lattice is called uniform. Observe that a subgroup of $\operatorname{Aut} T$ is discrete if and only if it acts with finite stabilizers. One may determine whether a discrete subgroup is a lattice by checking the finiteness of the sum $\sum_{v \in F} 1/|\Gamma_v|$, where Γ_v is the stabilizer of the vertex v and the set F is a fundamental domain for the action of Γ on some $\operatorname{Aut} T$ orbit in T. Of particular interest is the case when $\operatorname{Aut} T$ acts with finitely many orbits on T; in this case a lattice is uniform if and only if the quotient $\Gamma\backslash T$ is finite. Such lattices, called "uniform tree lattices", correspond to finite graphs of groups in which all vertex and edge groups are

finite. These were extensively studied by Bass and Kulkarni, [BK]. By a result of Leighton, cf. [BK], any two uniform tree lattices in $\mathrm{Aut}\, T$ are commensurable up to conjugation. A key role in the study of lattices in semisimple Lie groups is played by their commensurators (cf. [AB]): $\mathrm{Comm}_G(\Gamma) = \{g \in G : g^{-1}\Gamma g \cap \Gamma \text{ is of finite index in both } \Gamma \text{ and } g^{-1}\Gamma g\}$. In particular, Margulis has shown that an irreducible lattice $\Gamma < G$ in a semisimple Lie group is arithmetic if and only if its commensurator is dense in G.

It was shown by Liu [Liu] (confirming a conjecture by Bass and Kulkarni) that the commensurator of a uniform tree lattice is dense. The situation concerning non-uniform lattices is much more involved and not well understood. There are examples, by Bass and Lubotzky [BL2], cf. also [BM1], of non-uniform lattices in the automorphism groups of regular trees whose commensurators are discrete. At the other extreme it was shown by the author that the commensurator of the Nagao lattice $\mathrm{SL}_2(F_p[t])$ in the full automorphism group of the $(p+1)$-regular tree is dense. An example of a cocompact lattice with dense commensurator which is not a uniform tree lattice appears in [BM1].

Note that as all uniform tree lattices of a given tree are commensurable up to conjugation, the isomorphism class of the commensurator of a uniform tree lattice is determined by the tree. In the other direction it is shown in [LMZ] that for regular trees the commensurator determines the tree. In proving this we use a superrigidity theorem for the commensurators of lattices in the automorphism groups of regular trees. In a much more general setting of divergence groups in the isometry group of CAT(-1) spaces we have shown in [BM1] (see also [Bur]) that:

THEOREM 1. *Let X, Y be proper CAT(-1)-spaces, $\Gamma < Is(X)$ a discrete divergence group, $\Lambda < Is(X)$ a subgroup such that $\Gamma < \Lambda < Comm_{\mathrm{Is}(X)}(\Gamma)$ and $\pi : \Lambda \to Is(Y)$ a homomorphism such that $\pi(\Lambda)$ acts convex-minimally and $\pi(\Gamma)$ is not elementary. Then π extends to a continuous homomorphism*

$$\pi_{\mathrm{ext}} : \overline{\Lambda} \to Is(Y) \ .$$

PRODUCTS OF TREES AND LOCALLY PRIMITIVE GROUPS. Among the most striking results concerning lattices in semisimple Lie groups are the arithmeticity and superrigidity theorems established by G.A. Margulis (cf. [Mar], [Zim], [AB]). These assert that:

1. An irreducible lattice in a higher rank (i.e. ≥ 2) semisimple Lie group is arithmetic.

2. Any linear representation of such a lattice with unbounded image essentially extends to a continuous representation of the ambient Lie group.

Recall that a lattice $\Gamma < G$ in a semisimple Lie group is called *reducible* if the following equivalent conditions hold:

1. There exists a decomposition of G (up to isogeny) as a product $G = G_1 \times G_2$ with both G_i non compact semisimple Lie groups and Γ projects discretely on each G_i, $i = 1, 2$.

2. Γ contains a finite index subgroup of the form $\Gamma_1 \times \Gamma_2$ where $\Gamma_i < G_i$ is a lattice and $G = G_1 \times G_2$ a decomposition as above.

Using Borel's density theorem ([Bor], cf. [Fur], [Dan]) we have that a lattice Γ in a semisimple Lie group G is irreducible if it satisfies the following equivalent conditions:

- (Ir1) The projection of Γ on any factor G_i of G with $\ker : G \to G_i$ noncompact has non-discrete image.
- (Ir2) A projection as above has dense image.

Pursuing further the analogy between $\operatorname{Aut} T$ and rank one Lie groups, it is natural to ask for a structure theory for lattices in groups of the form $\operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ with T_i trees. In particular one would like to have "rigidity-" and "arithmeticity-" like results. Some steps in this direction were taken jointly with M. Burger and R.J. Zimmer [BMZ], [BM2], [BM3] Let us assume henceforth (unless explicitly stated otherwise) that our trees are (bi)-regular. A lattice $\Gamma < \operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ is reducible when its projections on each factor are discrete. Restricting our attention to uniform (i.e., cocompact) lattices observe that the projection $\operatorname{pr}_i(\Gamma) < \operatorname{Aut} T_i$ of a lattice $\Gamma < \operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ is never dense. This follows by observing that the compact open subgroup $K = \operatorname{Stab}_{\operatorname{Aut} T_i}(x)$, $x \in T_i$ a vertex, maps onto $(\mathbb{Z}/2\mathbb{Z})^{\mathbb{N}}$ and hence is not topologically finitely generated, whereas the intersection of Γ with the product of K with $\operatorname{Aut} T_{3-i}$, being a uniform lattice in this product, is finitely generated. Thus the intersection of K with the projection of Γ to $\operatorname{Aut} T_i$ cannot be dense in K (see [BM3]). Given a lattice $\Gamma < \operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ denote by $H_i = \overline{\operatorname{pr}_i(\Gamma)}$, $i = 1, 2$. Thus $\Gamma < H_1 \times H_2$. Clearly the representation theory of Γ cannot be more rigid than that of $H_1 \times H_2$. Indeed, one can construct irreducible lattices $\Gamma < \operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ such that the corresponding groups H_i surject onto free groups. Requiring various conditions on the projections of Γ leads to interesting structure theory.

DEFINITION 2. *A subgroup $H < \operatorname{Aut} T$ is called* locally primitive *if for every vertex $x \in T$ its stabilizer in H induces a primitive permutation group, denoted $\underline{H}(x)$, on the set $E(x)$ of neighbouring edges.*

The class of closed locally primitive subgroups of $\operatorname{Aut} T$ has a structure theory reminiscent in some ways to that of semisimple Lie groups. A key role in the study of these groups is played by the following lemma which shows that normal subgroups of a locally primitive group are either free or very large.

LEMMA 3. *Let T be a tree and let $H < \operatorname{Aut} T$ be a closed locally primitive subgroup. Any normal subgroup $N \lhd H$ either acts freely on T, or is cocompact and has a fundamental domain which is either a ball of radius 1 or an edge in T.*

For a locally compact, totally disconnected group H, let $H^{(\infty)} := \bigcap_{L<H} L$, where the intersection is taken over all open subgroup $L < H$ of finite index, and $QZ(H) := \bigcup_{U<H} Z_H(U)$, where the union is taken over all open subgroups $U < H$.

Thus $QZ(H) = \{h \in H : Z_H(h) \text{ is open}\}$. Observe that $H^{(\infty)} = \bigcap_{N \lhd H} N$, where the intersection is taken over all closed, cocompact normal subgroups $N \lhd H$. Note also that any discrete normal subgroup of H is contained in $QZ(H)$. Using Lemma 3 one shows that when H is non-discrete, $H^{(\infty)}$ is cocompact in H.

With the (limited) analogy between closed locally primitive subgroups $H < \operatorname{Aut} T$ and algebraic groups G in mind, we may view $H^{(\infty)} < H$ as playing the role of the subgroup $G^+ < G$ generated by all one parameter unipotent subgroups. We have the following structure theorem:

THEOREM 4. *(Burger-Mozes) Let $H < AutT$ be a closed, non-discrete, locally primitive subgroup. Then $H^{(\infty)}/QZ(H^{(\infty)})$ decomposes as a finite direct product*

$$H^{(\infty)}/QZ(H^{(\infty)}) = M_1 \cdot M_2 \cdot \cdots \cdot M_r$$

Where each M_i, $1 \leq i \leq r$ is a topologically simple group.

Various examples of closed subgroups of $\operatorname{Aut} T$ may be obtained via the following construction: Let $d \geq 3$, and $F < S_d$ be a permutation group. Let $T_d = (X, Y)$ be the d-regular tree and $i : Y \to \{1, 2, \ldots, d\}$ a legal (edge) coloring, that is, a map such that $i(y) = i(\bar{y}), \forall y \in Y$, and $i|_{E(x)} : E(x) \to \{1, 2, \ldots, d\}$ is a bijection, $\forall x \in X$. Define $U(F) = \{g \in \operatorname{Aut} T_d : i|_{E(gx)} g i^{-1}|_{E(x)} \in F,\ \forall x \in X\}$. Observe that $U(F)$ is a closed subgroup of $\operatorname{Aut} T_d$; the group $U(F)$ acts transitively on X; the finite group $\underline{U(F)}(x) < \operatorname{Sym} E(x)$ is permutation isomorphic to $F < S_d$, and hence, when F is a primitive permutation group, $U(F)$ is locally primitive. We notice also the following:

1. Using Tits' theorem, [Tit], it follows that $U(F)^+$ (the subgroup generated by edge stabilizers) is simple.

2. The subgroup $U(F)^+$ is of finite index in $U(F)$ if and only if $F < S_d$ is transitive and F is generated by its subgroups $\operatorname{Stab}_F(j)$, $1 \leq j \leq d$. In this case, $U(F)^+ = U(F) \cap \operatorname{Aut}^+ T_d$ and is of index 2 in $U(F)$.

3. Let $F < S_d$ be a transitive subgroup and $H < \operatorname{Aut} T_d$ be a vertex-transitive subgroup such that, for some $x \in X$, $\underline{H}(x) < \operatorname{Sym} E(x)$ is permutation isomorphic to $F < S_d$. Then, for some suitable legal coloring, we have $H < U(F)$.

We are especially interested at those subgroups $U(F)$ which arise as closures of projections of irreducible uniform lattices. As these must be topologically finitely generated we note:

PROPOSITION 5. *[BM3] Let $F < S_d$ be a transitive permutation group. Then $U(F)(x)$ is topologically finitely generated if and only if $F_1 = \operatorname{Stab}_F(1)$ is perfect and equal to its normalizer in F.*

NOTATION: Denote by $S(x, n)$ the sphere of radius n around a vertex $x \in T$. For $H < \operatorname{Aut} T$, $x \in X$, $n \geq 1$, $H_n(x) = \{h \in H : h|_{S(x,n)} = id\}$, $\underline{H}_n(x) = H_n(x)/H_{n+1}(x)$.

PROPOSITION 6. *Let $F < S_d$ be a 2-transitive permutation group such that $F_1 = \mathrm{Stab}_F(1)$ is non-abelian simple and $H < Aut T_d$ a closed vertex transitive subgroup such that $\underline{H}(x) < \mathrm{Sym}\, E(x)$ is permutation isomorphic to $F < S_d$. Then $\underline{H}_1(x) \simeq F_1^a$ where $a \in \{0, 1, d\}$. Moreover*

$$\begin{array}{ll} H \text{ is discrete} & \Leftrightarrow a \in \{0, 1\} \\ H = U(F) & \Leftrightarrow a = d. \end{array}$$

In the proof of the above proposition one needs to show that when $a = d$ the group H is not discrete. This is established using the Thompson-Wielandt theorem (see [Tho], [Wie], [Fan]).

THEOREM 7. *(Thompson-Wielandt) Let $T = \mathfrak{T}_n$ be the n-regular tree. Let $U < Aut(T)$ be the pointwise stabilizer of a ball of radius 1 around an edge e. (Note that U is an open compact neighborhood of the identity.) Then for every vertex transitive locally primitive lattice $\Gamma < Aut(T)$ the group $\Gamma \cap U$ is an l-group for some prime $l < n$.*

In the context of lattices $\Gamma < \mathrm{Aut}\, T_1 \times \mathrm{Aut}\, T_2$ one would like to verify for a given lattice whether it is reducible or not, namely whether its projections are discrete or not. The Thompson-Wielandt theorem may be used to verify non-discreteness and hence irreducibility in certain cases but we do not know a general algorithm for deciding this question.

RIGIDITY. The following result may be viewed as an analog of the Mostow rigidity theorem:

THEOREM 8. *[BMZ] Let $\Gamma < Aut T_1 \times Aut T_2$ and $\Gamma' < Aut T_1' \times Aut T_2'$ be uniform lattices. Assume that the subgroups $H_i = \overline{pr_i(\Gamma)} < Aut T_i$, $i = 1, 2$ are locally primitive. If we have an isomorphism $\Gamma \cong \Gamma'$, then it is induced by an isometry between $T_1 \times T_2$ and $T_1' \times T_2'$.*

Note that we do not assume that the lattices are irreducible. Bass and Lubotzky [BL1] have shown that a certain class of closed (non-discrete) subgroups of $\mathrm{Aut}\, T$ determines the tree T (up to some natural modifications). We note that any isomorphism between locally primitive lattices $\Lambda < \mathrm{Aut}\, T$ and $\Lambda' < \mathrm{Aut}\, T'$ acting without inversion on the corresponding trees is induced by an isometry between T and T'. (To establish this one notes that the tree structure may be reconstructed from, say, Λ using the correspondence between the vertices and maximal finite subgroups of Λ and between the edges and pairs of such maximal finite groups which generate the group and whose intersection is a maximal subgroup in each.)

THEOREM 9. *[BMZ] Let $\Gamma < Aut T_1 \times Aut T_2$ be a uniform lattice with $H_i = \overline{pr_i(\Gamma)} < Aut T_i$, $i = 1, 2$ locally primitive. Let Y be a proper CAT(-1) space and $\pi : \Gamma \to Is(Y)$ a homomorphism such that $\pi(\Gamma)$ is not elementary and acts convex-minimally on Y. Then π extends to a continuous homomorphism $\pi_{ext} : H_1 \times H_2 \to Is(Y)$ which factors via a proper homomorphism of one of the H_i, $i = 1, 2$.*

In the Lie groups setting, Margulis' superrigidity theorem plays a key role in showing that irreducible lattices in higher rank groups are arithmetic. In the context of lattices in $\operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ we have:

THEOREM 10. *[BMZ] Let $\Gamma < Aut\, T_1 \times Aut\, T_2$ be an irreducible cocompact lattice. Assume that each $H_i = \overline{pr_i(\Gamma)} < Aut\, T_i$ is locally primitive. Then one of the following possibilities holds:*

1. *Every linear image of Γ is finite.*
2. *Γ has an infinite linear image over a field of characteristic 0. Then H_i is a p_i-adic analytic group for some prime p_i. The adjoint map, which we denote by $\varphi = \varphi_1 \times \varphi_2 : H_1 \times H_2 \to Aut(Lie(H_1)) \times Aut(Lie(H_2))$, yields a continuous surjection from $H_1 \times H_2$ onto a semisimple Lie group over some local fields and the image $\varphi(\Gamma)$ is an arithmetic lattice. Moreover, the kernel of this homomorphism is a torsion free discrete subgroup of $H_1 \times H_2$.*
3. *Γ has an infinite linear image over a field of positive characteristic p. Then there is a continuous map with unbounded image from $H_1 \times H_2$ into a simple Lie group over $F_p((t))$.*

Let us remark that:

- It seems reasonable to expect in case 3 of the theorem a result similar to that of 2.
- In case 2:
  - We do not claim that the image $\varphi(\Gamma)$ is an irreducible arithmetic lattice. Indeed, one can construct examples where this lattice is reducible.
  - The algebraic groups $\varphi_i(H_i)$ need not be of rank one. In fact they are of rank one if and only if $\ker \varphi = \{e\}$. Moreover, $\varphi_i(H_i)$ is of rank one and $\ker \varphi_i$ is trivial exactly when the action of H_i on the corresponding tree T_i is locally infinitely transitive, i.e., the stabilizer of each vertex acts transitively on simple paths of arbitrary length starting at the vertex.

An example of an irreducible lattice in $\operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ which is an extension of an arithmetic lattice in a semisimple algebraic group $G = G_1 \times G_2$, where each G_i is a semisimple group over some local field k_i, may be obtained as follows, cf. [BM3]. Associated with each G_i one has an affine building Δ_i on which the group G_i acts. "Draw" on Δ_i a graph $\mathcal{G}_i$ defined in an equivariant way (for example let $G_i = SL_3(\mathbb{Q}_p)$, the associated Bruhat-Tits building is a simplicial complex whose set of vertices has a natural 3-coloring (see [Bro]); consider the graph consisting of the vertices belonging to two fixed colors and the corresponding edges). The group G acts on $\mathcal{G}_1 \times \mathcal{G}_2$ and hence an extension $H_1 \times H_2$ of G by $\pi_1(\mathcal{G}_1 \times \mathcal{G}_2)$ acts on the universal covering space $T_1 \times T_2$ of $\mathcal{G}_1 \times \mathcal{G}_2$. Taking a lattice $\Lambda < G$, its extension by $\pi_1(\mathcal{G}_1 \times \mathcal{G}_2)$ is a lattice in $H_1 \times H_2$.

We examine next the normal subgroups structure of lattices in $\operatorname{Aut} T_1 \times \operatorname{Aut} T_2$.

PROPOSITION 11. *[BM3] Let T_1, T_2 be locally finite trees, $\Gamma < \operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ a discrete subgroup such that $\Gamma \setminus (T_1 \times T_2)$ is finite and $N < \Gamma$ a normal subgroup such that the quotient graphs $pr_i(N) \setminus T_i$, $i = 1, 2$, are finite trees. Then Γ/N has property (T).*

PROPOSITION 12. *[BM3] Let T_1, T_2, Γ be as in Proposition 11 and $H_i := \overline{pr_i(\Gamma)} < \operatorname{Aut} T_i$.*

(a) The homomorphism $\operatorname{Hom}_c(H_1 \times H_2, \mathbb{C}) \to \operatorname{Hom}(\Gamma, \mathbb{C})$ mapping χ to $\chi \mid_\Gamma$ is an isomorphism.

(b) Let (π, V) be an irreducible finite dimensional unitary representation of Γ with $H^1(\Gamma, \pi) \neq 0$.

Then π extends continuously to $H_1 \times H_2$, factoring via one of the projections.

The following result, obtained in [BM3], is an analog of Margulis' normal subgroup theorem:

THEOREM 13. *Let $\Gamma < Aut T_1 \times Aut T_2$ be a cocompact lattice such that $H_i := \overline{pr_i(\Gamma)}$ is locally ∞-transitive and $H_i^{(\infty)}$ is of finite index in H_i, $i = 1, 2$. Then, any non-trivial normal subgroup of Γ has finite index.*

The proof of this theorem follows the lines of the proof by Margulis of the corresponding result in the context of Lie groups. In particular one uses the following analog of the Howe-Moore theorem concerning vanishing of matrix coefficients. In the context of $\operatorname{Aut}^+ T$ with T a regular tree this was shown in [LM], see also [FTN].

THEOREM 14. *[BM3] Let $H < Aut T$ be a closed locally ∞-transitive subgroup and $(\pi, \mathcal{H})$ be a continuous unitary representation of H with no nonzero $H^{(\infty)}$ invariant vectors. Then for every $u, v \in \mathcal{H}$, $\lim \langle \pi(g)u, v \rangle \to 0$ as $g \in H$ tends to ∞.*

However there is an interesting application of Theorem 13 which is based on a fundamental difference between cocompact lattices in $\operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ and cocompact lattices in Lie groups. Whereas any finitely generated subgroup of a linear group is residually finite, finitely generated subgroups, and even cocompact lattices, in $\operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ need not be residually finite. A criterion for establishing that certain lattices are not residually finite is provided by the following:

PROPOSITION 15. *[BM3] Let G_i, $i = 1, 2$ be closed locally compact groups. Let $\Gamma < G_1 \times G_2$ be a discrete subgroup. Assume that for $i = 1, 2$, $G_i^{(\infty)} < \overline{pr_i(\Gamma)} < G_i$. Then*

$$\Gamma^{(\infty)} > [G_1^{(\infty)}, \Lambda_1] \cdot [G_2^{(\infty)}, \ \Lambda_2],$$

where $\Lambda_1 := pr_1((G_1 \times e) \cap \Gamma)$, $\Lambda_2 := pr_2((e \times G_2) \cap \Gamma)$. In particular, if each G_i has trivial centralizer and $\Lambda_1 \times \Lambda_2 \neq e$, then Γ is not residually finite.

In particular, let $\Gamma < \operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ be an irreducible lattice such that each $H_i = \overline{\operatorname{pr}_i(\Gamma)}$ is locally primitive. Each $H_i^{(\infty)}$ acts on T_i with finite quotient and

hence has trivial centralizer. Thus if in addition the projection of Γ to one of the factors $\operatorname{Aut} T_i$ is not injective then Γ is not residually finite. The construction described following Theorem 10 provides a non residually finite lattice.

Combining Proposition 15 and Theorem 13 allows us to construct (see [BM2] and [BM3]) examples of finitely presented torsion free simple groups. One constructs first a non residually finite lattice $\tilde{\Gamma}$ which satisfies the conditions of Theorem 13. Given $\tilde{\Gamma}$ let $\Gamma = \tilde{\Gamma}^{(\infty)}$. It follows that Γ is the minimal finite index subgroup of $\tilde{\Gamma}$. Verifying that Γ satisfies the conditions of Theorem 13, one deduces that Γ is simple. It should be observed that a lattice Γ which satisfies the conditions of Proposition 15 for not being residually finite must have a non trivial normal subgroup of infinite index! namely either Λ_1 or Λ_2. Thus in order to produce a non residually finite lattice which satisfies the conditions of Theorem 13 one uses the geometric description of lattices in $\operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ (see below) to embed a non residually finite lattice obtained using Proposition 15 in a lattice as in Theorem 13.

THEOREM 16. *For every pair (n, m) of sufficiently large even integers there exists a group $\Gamma_{n,m}$ such that:*

1. *The group $\Gamma_{n,m}$ is simple, finitely presented, torsion free and isomorphic to a free amalgam $F *_G F$ where F, G are finitely generated free groups.*
2. *The group $\Gamma_{n,m}$ has cohomological dimension 2.*
3. *$\Gamma_{n,m}$ is automatic.*

The question of existence of simple groups which are amalgams of free groups was raised by P.M. Neumann ([Neu], see also the Kourovka notebook [MK] problem 4.45). M. Bhattacharjee [Bha] constructed examples of amalgams of free groups which do not have any finite quotients. The groups $\Gamma_{n,m}$ are constructed as lattices in $\operatorname{Aut} \mathfrak{T}_n \times \operatorname{Aut} \mathfrak{T}_m$ (where $\mathfrak{T}_k$ denotes the k-regular tree). Considering the action of $\Gamma_{n,m}$ on each of the trees $\mathfrak{T}_k$, $k = n, m$, we obtain two decompositions of $\Gamma_{n,m}$ as amalgams $A *_C B$. The groups A, B and C, being torsion free lattices in $\operatorname{Aut} \mathfrak{T}_k$, are free groups (note that $[A : C] = [B : C] \in \{n, m\}$). Moreover, using the superrigidity theorem 9 it follows that these are the only nontrivial decompositions of $\Gamma_{n,m}$ as amalgamated products. This implies also that the groups $\Gamma_{n,m}$ are mutually non isomorphic. These also form the first examples of finitely presented simple groups of finite cohomological dimension. We refer to [Sco] for a survey and discussion of various families of finitely presented simple groups constructed by R. Thompson, Higman, Brown and Scott.

GEOMETRICAL DESCRIPTION. When a subgroup $D < \operatorname{Aut} T_1 \times \operatorname{Aut} T_2$ acts freely on $T_1 \times T_2$, it may be identified with the fundamental group of the quotient space $\mathcal{Y} = D\backslash(T_1 \times T_2)$. Note that $\mathcal{Y}$ is a square complex whose universal covering space is $T_1 \times T_2$. Square complexes whose universal covering space is a product of trees are characterized as those square complexes in which the link of every vertex is a complete bipartite graph. (Again under the analogy with semisimple groups consider the geometric characterization of locally symmetric spaces.) More generally, when the action is not free, D may be reconstructed as the fundamental group of

a certain complex of groups ([Hae] and see [Ser] and [Bas] for the corresponding theory of graph of groups). This geometric way of considering subgroups and in particular cocompact lattices in $\mathrm{Aut}\, T_1 \times \mathrm{Aut}\, T_2$ allows one to explicitly construct and modify such lattices. Note that any such finite square complex gives a finite presentation of its fundamental group. In [BM3] we give an explicit construction of the complexes associated with the groups $\Gamma_{n,m}$ thus providing an explicit finite presentation of these torsion free simple groups.

MINIMAL VOLUMES. Kazhdan and Margulis [KM] have shown that for any semisimple Lie group there is a positive lower bound on the volume of $\Gamma\backslash G$ for any lattice $\Gamma < G$. In this vein let us mention the works of Lubotzky and Weigel, [Lu2], [LW], who determined the lattices of minimal covolume in the groups of the form $\mathrm{SL}_2(K)$, where K is a non archimedean local field.

In contrast, there is no lower bound for arbitrary lattices in either $\mathrm{Aut}\, T$ or $\mathrm{Aut}\, T_1 \times \mathrm{Aut}\, T_2$. Bass and Kulkarni, [BK], constructed cocompact lattices in $\mathrm{Aut}\, T$, where T is a k-regular tree, $k \geq 3$, with arbitrarily small covolume. I. Levitz determined precisely the (dense) set of (positive) rational numbers appearing as covolumes of uniform lattices in $\mathrm{Aut}\, T$. Moreover, Bass and Lubotzky have shown in [BL2] that given any real number $\alpha > 0$ there exists a non-uniform lattice $\Gamma < \mathrm{Aut}\, T$ such that $\mathrm{Vol}(\Gamma\backslash \mathrm{Aut}\, T) = \alpha$. Considering lattices in $\mathrm{Aut}\, T_1 \times \mathrm{Aut}\, T_2$, Y. Glasner has constructed examples of irreducible lattices $\Gamma < \mathrm{Aut}\, T_1 \times \mathrm{Aut}\, T_2$ of arbitrarily small covolume. In these examples the subgroups $H_i = \overline{\mathrm{pr}_i(\Gamma)} < \mathrm{Aut}\, T_i$ are not locally primitive.

However, a well known conjecture of Goldschmidt and Sims, translated into the language of lattices, asserts that for any given tree T there are only finitely many locally primitive lattices in $\mathrm{Aut}\, T$. Thus in particular there is a lower bound on the covolume of such lattices. The Goldschmidt-Sims conjecture is usually stated as saying that for any $n, m \geq 3$ there are only finitely many effective amalgams $A *_C B$ of finite groups with $[A : C] = n$, $[B : C] = m$ and C is maximal in both A and B. This was established by D. Goldschmidt [Gol] for the case $n = m = 3$. In view of the above results concerning the analogy between semisimple Lie groups and locally primitive groups of tree automorphisms, one is led to ask whether there is a positive lower bound on the covolume of lattices $\Gamma < \mathrm{Aut}\, T_1 \times \mathrm{Aut}\, T_2$ having locally primitive projections (note that for reducible such lattices this would follow from the Goldschmidt-Sims conjecture). Studying this question Y. Glasner [Gla] has proved the following analog of the Goldschmidt-Sims conjecture:

THEOREM 17. *(Glasner) For any fixed primes p, q there are only finitely many effective complexes of groups consisting of a single square whose universal covering space is a product of regular trees $\mathfrak{T}_p \times \mathfrak{T}_q$ and whose fundamental group $\Gamma < Aut\mathfrak{T}_p \times Aut\mathfrak{T}_q$ is an irreducible lattice.*

A central role in the proof of the above theorem is played by the Thompson-Wielandt theorem (Theorem 7). Recall that in establishing the lower bound on the covolume of a lattice in a semisimple Lie group G one may use ([KM], cf. [Rag]) the existence of a "Zassenhaus neighbourhood" $U < G$ such that for every discrete subgroup $\Gamma < G$ the elements of $\Gamma \cap U$ are contained in some connected

nilpotent Lie subgroup of G. The Thompson-Wielandt theorem gives a neighbourhood whose intersection with locally primitive discrete groups is an l-group, and hence, in particular, nilpotent.

ACKNOWLEDGEMENTS. I would like to thank Marc Burger, Yair Glasner, Alex Lubotzky and Bob Zimmer for many helpful discussions.

REFERENCES

[AB] N. A'Campo and M. Burger. Réseaux arithmétiques et commensurateurs d'après G. A. Margulis. *Invent. Math.* 116(1994), 1–25.

[Bas] H. Bass. Covering theory for graphs of groups. *J. Pure and Appl. Algebra* 89(1993), 3–47.

[BK] H. Bass and R. Kulkarni. Uniform tree lattices. *J. Amer. Math. Soc.* 3(1990), 843–902.

[BL1] H. Bass and A. Lubotzky. Rigidity of group actions on locally finite trees. *Proc. London Math. Soc. (3)* 69(1994), 541–575.

[BL2] H. Bass and A. Lubotzky. Tree Lattices. A forthcoming book.

[Bha] M. Bhattacharjee. Constructing finitely presented infinite nearly simple groups. *Comm. Algebra* 22(1994), 4561–4589.

[Bor] A. Borel. Density properties for certain subgroups of semisimple groups without compact components. *Annals of Math.* 72(1960), 179–188.

[Bro] K.S. Brown. *Buildings*. Springer, 1989.

[Bur] M. Burger. Rigidity properties of group actions on CAT(0)-spaces. In *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994)*, pages 761–769, Basel, 1995. Birkhäuser.

[BM1] M. Burger and S. Mozes. CAT(-1)-spaces, divergence groups and their commensurators. *J. Amer. Math. Soc.* 9(1996), 57–93.

[BM2] M. Burger and S. Mozes. Finitely presented simple groups and products of trees. *C. R. Acad. Sci. Paris Sér. I Math.* 324(1997), 747–752.

[BM3] M. Burger and S. Mozes. Products of trees, Lattices and Simple Groups. In preparation.

[BMZ] M. Burger, S. Mozes, and R.J. Zimmer. Irreducible lattices in the automorphism group of a product of trees, Superrigidity and Arithmeticity. In preparation.

[Dan] S.G. Dani. A simple proof of Borel's density theorem. *Math. Z.* 174(1980), 81–94.

[Fan] P.S. Fan. The Thompson-Wielandt Theorem. *Proc. Amer. Math. Soc.* 97(1986).

[FTN] A. Figa-Talamanca and C. Nebbia. *Harmonic Analysis and Representation Theory for Groups Acting on Homogeneous Trees*, volume 162 of *LMS*. Cambridge University Press, 1991.

[Fur] H. Furstenberg. A note on Borel's density theorem. *Proc. Amer. Math. Soc.* 55(1976), 209–212.

[Gla] Y. Glasner. Some Remarks on the Co-volume of Lattices Acting on a Product of Trees. Master's thesis, Hebrew University, 1997.

[Gol] D.M. Goldschmidt. Automorphisms of trivalent graphs. *Annals of Mathematics* 111(1980), 377–406.

[Hae] A. Haefliger. Complexes of groups and orbihedra. In *Group Theory from a Geometrical Viewpoint*, pages 504–540. World Sci. Publishing, River Edge, NJ, 1990.

[KM] D. Kazhdan and G.A. Margulis. A proof of Selberg's hypothesis. *Math. Sbornik* (1968).

[Liu] Y.S. Liu. Density of the commensurability group of uniform tree lattices. *J. of Algebra* 165(1994).

[Lu1] A. Lubotzky. Trees and discrete subgroups of Lie groups over local fields. *Bull. Amer. Math. Soc. (N.S.)* 20(1989), 27–30.

[Lu2] A. Lubotzky. Lattices of minimal covolume in SL_2: A nonarchimedian analogue of Siegel's theorem $\mu \geq \pi/21$. *J. Amer. Math. Soc.* 3(1990), 961–975.

[Lu3] A. Lubotzky. Tree lattices and lattices in Lie groups. *Combinatorial and Geometric Group Theory*, Edinburgh 1993, LMS Lecture Notes Series 204, (1995), 217–232.

[LM] A. Lubotzky and S. Mozes. Asymptotic properties of unitary representations of tree automorphisms. In M.A. Picardello, editor, *Harmonic Analysis and Discrete Potential Theory*. Plenum Press, 1992.

[LMZ] A. Lubotzky, S. Mozes, and R.J. Zimmer. Superrigidity for the commensurability group of tree lattices. *Comment. Math. Helv.* 69(1994), 523–548.

[LW] A. Lubotzky and T. Weigel. Lattices of minimal covolume in SL_2 over nonarchimedean fields. *Proc. London Math. Soc.* (To Appear).

[Mar] G.A. Margulis. *Discrete subgroups of semisimple Lie groups*, volume 17 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991.

[MK] V. D. Mazurov and E. I. Khukhro, editors. *Unsolved problems in group theory. The Kourovka notebook*. Russian Academy of Sciences Siberian Division Institute of Mathematics, Novosibirsk, augmented edition, 1995.

[Moz] S. Mozes. On the congruence subgroup problem for tree lattices. In *Proceedings of workshop on Ergodic theory and Lie groups*. Tata Institute of Fundamental Research, Bombay, India, To appear.

[Neu] P.M. Neumann. The SQ-universality of some finitely presented groups. *J. Austral. Math. Soc.* 16(1973), 1–6. Collection of articles dedicated to the memory of Hanna Neumann, I.

[Rag] M. S. Raghunathan. *Discrete Subgroups of Lie Groups*. Springer-Verlag, Berlin Heidelberg New York, 1972.

[Sco] E.A. Scott. A tour around finitely presented infinite simple groups. In G. Baumslag and C.F. Miller III, editors, *Algorithms and Classification in Combinatorial Group Theory*, volume 23 of *MSRI publications*, pages 83–119. Springer-Verlag, 1989.

[Ser] J-P. Serre. *Trees*. Springer-Verlag, New York, 1980.

[Tho] J.G. Thompson. Bounds for orders of maximal subgroups. *J. of Algebra* 14(1970), 135–138.

[Tit] J. Tits. Sur le groupe des automorphismes d'un arbre. In A. Haefliger and R. Narasimhan, editors, *Essays on Topology and Related Topics (Mémoires dédiés à Georges de Rham)*, pages 188–211. Springer, New York, 1970.

[Wie] H. Wielandt. Subnormal subgroups and permutation groups. Lecture notes. Columbus, Ohio State Univ., 1971.

[Zim] R.J. Zimmer. *Ergodic Theory and Semisimple Groups*, volume 81 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel, 1984.

Shahar Mozes
Institute of Mathematics
Hebrew University
Jerusalem
Israel
mozes@math.huji.ac.il

Characters of Irreducible Representations of Simple Lie Superalgebras

Vera Serganova

1991 Mathematics Subject Classification: 17A70, 17B70
Keywords and Phrases: Lie superalgebra, character, irreducible representation, cohomology

1. Introduction

Simple Lie superalgebras were classified in 1977 by V. Kac [6]. These superalgebras can be divided into three groups

1. *Contragredient* Lie superalgebras, i.e. Lie superalgebras which can be determined by a Cartan matrix. These superalgebras have an invariant symmetric form and Cartan involution. There are two families of such algebras $sl\,(m|n)$ (factored by the center when $m = n$) and $osp\,(m|2n)$. The Lie superalgebra $osp\,(4|2)$ has a one-parameter deformation, called $D\,(\alpha)$. There are also two exceptional Lie superalgebras G_3 and F_4.
2. New "*strange*" superalgebras $Q\,(n)$ and $P\,(n)$, the former consists of operators commuting with an odd nondegenerate operator, the latter consists of operators preserving a non-degenerate odd symmetric form.
3. *Cartan type superalgebras* W_n, S_n, S'_n and SH_n, i.e. superalgebra of vector fields on a supermanifold of pure odd dimension and its simple subalgebras.

In [7] it was shown that all finite-dimensional irreducible representations of a simple Lie superalgebra $\mathfrak{g}$ are enumerated by a highest weight $\lambda \in \mathfrak{h}^*$, satisfying certain conditions of dominance, $\mathfrak{h}$ being a Cartan subalgebra.

The problem of finding the character of an irreducible highest weight module L_λ for a general dominant λ appeared to be unexpectedly difficult. It was solved first for Lie superalgebras W_n and S_n of Cartan type by J. Bernstein and D. Leites in [1] and for SH_n by A. Shapovalov [20]. They considered a module M_λ of tensor fields on a supermanifold of purely odd dimension and described completely its Jordan–Hölder series. Since $\operatorname{ch} M_\lambda$ and the multiplicity $[M_\lambda : L_\mu]$ are known, one can obtain $\operatorname{ch} L_\lambda$ by solving a simple system of linear equations.

For contragredient and strange Lie superalgebras the problem was solved in [7] in the particular case of a generic (*typical*) highest weight λ. The character is given by a nice Weyl type formula:

$$\operatorname{ch} L_\lambda = D \sum_{w\in W} \operatorname{sgn} w \cdot e^{w(\lambda)},\ D = \frac{\prod_{\alpha\in\Delta_1^+}\left(e^{\alpha/2}+e^{-\alpha/2}\right)}{\prod_{\alpha\in\Delta_0^+}\left(e^{\alpha/2}-e^{-\alpha/2}\right)}, \tag{1.1}$$

where W is the Weyl group of $\mathfrak{g}_0$.

This formula can be obtained from the Borel–Weil–Bott theorem for a flag supermanifold and a typical invertible sheaf $\mathcal{O}_\lambda$. Geometry of flag supermanifolds was studied in 1980*s* by Yu. Manin and his pupils I. Penkov, I. Skornyakov and A. Voronov. It was shown by Penkov and Skornyakov in [16] that the invertible sheaf $\mathcal{O}_\lambda$ with a typical dominant weight λ satisfies the Borel–Weil–Bott theorem, namely $H^0\mathcal{O}_\lambda = L_\lambda$ and $H^i\mathcal{O}_\lambda = 0$ for $i > 0$.

On the other hand, any invertible sheaf $\mathcal{O}_\lambda$ on G/B can be considered as a sheaf $\mathcal{L}_\lambda$ on the underlying flag manifold G_0/B_0. $\mathcal{L}_\lambda$ has a filtration by invertible sheaves $\mathcal{O}_{\lambda+\nu}\,(G_0/B_0)$, where ν runs over the set of sums of odd negative roots $\sum_{\alpha_i\in\Delta_1^-}\alpha_i$. Therefore one can write down the Euler characteristic of $\mathcal{L}_\lambda$

$$E_\lambda = \sum_i (-1)^i \operatorname{ch} H^i\mathcal{O}_\lambda = \sum_\nu \sum_i (-1)^i \operatorname{ch} H^i_{G_0/B_0}\mathcal{O}_{\lambda+\nu}, \tag{1.2}$$

using additivity of Euler characteristic. It happens to be exactly the Kac character formula (1.1).

If a highest weight λ is *atypical* the Borel–Weil–Bott theorem fails. Depending on *the degree of atypicality* of λ the $\mathfrak{g}$-module structure of $H^i\mathcal{O}_\lambda$ becomes more and more complicated. There were several conjectures and partial results about a character formula for a general dominant weight (see [2, 13, 22, 21, 9, 14, 4]). For the case $\mathfrak{g} = gl\,(m|n)$ two different formulae were conjectured in [17] and in [5]. The first conjecture was proven in [18]. The second one is believed to be equivalent to the first one.[1] For $\mathfrak{g} = Q\,(n)$ the problem was solved in [15]. For $\mathfrak{g} = osp\,(m|2n)$ it is solved just recently, we announce the results here.

We use the same method to solve the problem for *gl*, *osp* and Q. Namely we calculate the "Euler-multiplicities" $a_{\lambda,\mu} = \sum (-1)^i \left[H^i_{G/B}\mathcal{O}_\lambda : L_\mu\right]$ (in some cases we have to use a suitable parabolic subgroup P instead of Borel subgroup B). Since a lot of numbers $a_{\lambda,\mu}$ vanish, one can find $\operatorname{ch} L_\lambda$ from the system $\sum_\mu a_{\lambda,\mu}\operatorname{ch} L_\mu = E_\lambda$ and (1.2). To calculate the coefficients $a_{\lambda,\mu}$, we represent B as the end of a flag of parabolic subgroups $G = P^{(1)} \supset P^{(2)} \supset \cdots \supset P^{(n)} = B$. Composing push down of sheaves, the coefficients $a_{\lambda,\mu}$ can be expressed in terms of similar coefficients $a^{(j)}_{\lambda,\mu}$ for the supermanifold $P^{(j)}/P^{(j+1)}$.

In fact cohomology of "dominant sheaves" on $P^{(j)}/P^{(j+1)}$ can be completely described by using a certain analogue of *translation* and *reflection* functors. The space of dominant weights can be stratified by the degree of atypicality. A translation functor allows us to move inside a stratum (compare Theorem 2.5), while a reflection functor increases the degree of atypicality by one. This together with the fact that cohomology of irreducible vector bundles on $P^{(j)}/P^{(j+1)}$ are almost semi-simple $P^{(j)}$-modules (see Lemma 4.2) helps to write down the crucial recurrence relations on $a^{(j)}_{\lambda,\mu}$ (compare Theorems 4.1 and 4.3).

In this paper we outline the schematic of the character formulae for *gl* and *osp* omitting all the details needed for the proof only. We also omit the case of Q_n containing additional complications, since irreducible representations of $\mathfrak{h}$ are not 1-dimensional.

[1] Note that in [18] references to [5] and [22] were corrupted.

Currently the problem of finding irreducible characters remains open for exceptional Lie superalgebras and $P(n)$. While similar methods should work for exceptional superalgebras, in the case of $P(n)$ it is unclear how to define translation and reflection functors: the center of the universal enveloping algebra of $P(n)$ is too small and central characters do not separate blocks.

2. Dominant weights, central characters and blocks

Throughout this paper $\mathfrak{g}$ stands for one of the Lie superalgebras $gl(m|n)$, $osp(2m|2n)$ or $osp(2m+1|2n)$. The Lie superalgebra $\mathfrak{g}$ has a root decomposition

$$\mathfrak{g} = \mathfrak{h} \oplus \bigoplus_{\alpha\in\Delta} \mathfrak{g}_\alpha.$$

The roots are called *even* or *odd* depending on the parity of the root space $\mathfrak{g}_\alpha$. Denote the set of even (correspondingly odd) roots by Δ_0 (correspondingly Δ_1). Clearly, $\Delta = \Delta_0 \cup \Delta_1$. Many odd roots are *isotropic*, put $\Delta^{\text{is}} = \{\alpha \mid (\alpha,\alpha) = 0\}$, where $(,)$ stands for the Killing form.

Describe the set of roots in the standard basis $\{\delta_1,\dots,\delta_n,\varepsilon_1,\dots,\varepsilon_m\}$ of $\mathfrak{h}^*$. Note that $(\delta_i,\delta_j) = \delta_{i,j}$, $(\varepsilon_i,\varepsilon_j) = -\delta_{i,j}$.

Let $\mathfrak{g} = gl(m|n)$. Then

$$\Delta_0 = \{\varepsilon_i - \varepsilon_j \mid i,j = 1,\dots,m\} \cup \{\delta_i - \delta_j \mid i,j = 1,\dots,n\},$$
$$\Delta_1 = \Delta^{\text{is}} = \{\pm(\varepsilon_i - \delta_j) \mid i = 1,\dots,m, j = 1,\dots,n\}.$$

Let $\mathfrak{g} = osp(2m|2n)$. Then $\Delta_1 = \Delta^{\text{is}}$ is the same as for $\mathfrak{g} = gl(m|n)$,

$$\Delta_0 = \{\varepsilon_i \pm \varepsilon_j \mid i,j = 1,\dots,m, i \neq j\} \cup \{\delta_i \pm \delta_j \mid i,j = 1,\dots,n\}.$$

Let $\mathfrak{g} = osp(2m+1|2n)$. Then

$$\Delta_0 = \{\varepsilon_i \pm \varepsilon_j \mid i,j = 1,\dots,m, i \neq j\} \cup \{\pm\varepsilon_i \mid i = 1,\dots,m\} \cup \{\delta_i \pm \delta_j \mid i,j = 1,\dots,n\},$$
$$\Delta^{\text{is}} = \{\pm(\varepsilon_i - \delta_j) \mid i = 1,\dots,m, j = 1,\dots,n\}.$$
$$\Delta_1 = \Delta^{\text{is}} \cup \{\pm\delta_i \mid i = 1,\dots,n\}.$$

Fix a subdivision $\Delta = \Delta^+ \cup \Delta^-$ and a triangular decomposition $\mathfrak{g} = \mathfrak{n}^- \oplus \mathfrak{h} \oplus \mathfrak{n}^+$, defined by $\mathfrak{n}^\pm = \bigoplus_{\alpha\in\Delta^\pm} \mathfrak{g}_\alpha$. A choice of Δ^+ is not unique, here we fix it by enumerating simple roots in each case.

For $\mathfrak{g} = gl(m|n)$ choose simple roots $\sigma_1 = \delta_1 - \delta_2, \dots, \sigma_{n-1} = \delta_{n-1} - \delta_n$, $\sigma_n = \delta_n - \varepsilon_1, \dots, \sigma_{m+n-1} = \varepsilon_{m-1} - \varepsilon_m$.

For $\mathfrak{g} = osp(2m|2n)$ choose simple roots $\sigma_1 = \delta_1 - \delta_2, \dots, \sigma_{n-1} = \delta_{n-1} - \delta_n$, $\sigma_n = \delta_n - \varepsilon_1, \dots, \sigma_{m+n-1} = \varepsilon_{m-1} - \varepsilon_m$, $\sigma_{m+n} = \varepsilon_{m-1} + \varepsilon_m$.

For $\mathfrak{g} = osp(2m+1|2n)$ choose simple roots $\sigma_1 = \delta_1 - \delta_2, \dots, \sigma_{n-1} = \delta_{n-1} - \delta_n$, $\sigma_n = \delta_n - \varepsilon_1, \dots, \sigma_{m+n-1} = \varepsilon_{m-1} - \varepsilon_m$, $\sigma_{m+n} = \varepsilon_m$.

For an even root α put $\alpha^\vee = \alpha/(\alpha,\alpha)$. We say that $\lambda \in \mathfrak{h}^*$ is *integral* if $(\lambda,\alpha^\vee) \in \mathbb{Z}$ for any $\alpha \in \Delta_0$. Denote the set of integral weights by Λ.

Denote by L_λ an irreducible module generated by highest vector v of weight $\lambda-\rho$, i.e., $\mathfrak{n}^+v=0$, $hv=\langle\lambda-\rho,h\rangle\,v$ for $h\in\mathfrak{h}$, and

$$\rho=1/2\sum_{\alpha\in\Delta_0^+}\alpha-1/2\sum_{\alpha\in\Delta_1^+}\alpha.$$

Call a weight $\lambda\in\Lambda$ *dominant* if $\dim L_\lambda<\infty$. Denote the set of dominant weights by Λ^+. The conditions on λ to be dominant were first calculated in [6]. We reproduce them here in our notations.

PROPOSITION 2.1. *Let $\mathfrak{g}=gl\,(m|n)$, $osp\,(2m|2n)$ or $osp\,(2m+1|2n)$. Let $\lambda=a_1\delta_1+\cdots+a_n\delta_n+b_1\varepsilon_1+\cdots+b_m\varepsilon_m\in\Lambda$. Then $\lambda\in\Lambda^+$ iff the following conditions on a_i and b_j hold:*

1. *for $gl\,(m|n)$: a_i-a_{i+1}, $b_j-b_{j+1}\in\mathbb{Z}_{>0}$;*
2. *for $osp\,(2m|2n)$: $a_i,b_j\in\mathbb{Z}$, $a_1>a_2>\cdots>a_n>-m$, $b_1>b_2>\cdots>b_{m-1}>|b_m|$ and for each $l\le 0$, $l\ge a_n$, $b_{m+l}=-l$;*
3. *for $osp\,(2m+1,2n)$, $a_i\in 1/2+\mathbb{Z}$, $b_j\in 1/2+\mathbb{Z}$ or $\mathbb{Z}$, $a_1>a_2>\cdots>a_n\ge 1/2-m$, $b_1>b_2>\cdots>b_m>0$ and for each $l\le 0$, $a_n\le l-1/2$, $b_{m+l}=1/2-l$.*

Remark 2.2. For *osp* type superalgebras one can not choose Δ^+ in such a way that the set of simple roots for Δ_0^+ is the subset of simple roots for Δ^+. That is why in this case the conditions on dominance with respect to $\mathfrak{g}_0$ differ from the conditions on dominance with respect to $\mathfrak{g}$.

Let $\mathfrak{g}=osp\,(2m|2n)$ or $osp\,(2m+1|2n)$. By Proposition 2.1 if $\lambda\in\Lambda^+$ and $a_r>0\ge a_{r+1}$, one can find odd isotropic roots $\delta_{r+1}-\varepsilon_{i_1},\dots,\delta_n-\varepsilon_{i_s}$ orthogonal to λ. Call the set of such roots the *tail* of λ and denote it by T_λ. Call the number $t_\lambda=s=n-r$ *the tail length* of λ. For $\mathfrak{g}=gl\,(m|n)$ put $t_\lambda=0$.

Note also that in the case $\mathfrak{g}=osp\,(2m|2n)$ there is a symmetry of Dynkin diagram which induces an outer automorphism τ such that $\tau\,(\sigma_{m+n-1})=\sigma_{m+n}$. The automorphism τ acts on the set of dominant weights. In what follows we always assume that $b_m\ge 0$. If $b_m<0$ we can obtain all coefficients using the rule $a_{\tau(\lambda),\tau(\mu)}=a_{\lambda,\mu}$, $a_{\tau(\lambda),\mu}=0$ if $\tau\,(\mu)<\mu$, $\tau\,(\lambda)<\lambda$.

Let $\mathcal{F}=\mathcal{F}\,(\mathfrak{g})$ be the category formed by finite-dimensional $\mathfrak{g}$-modules. To describe the structure of $\mathcal{F}$, consider the center $Z\,(\mathfrak{g})$ of the universal enveloping algebra $U\,(\mathfrak{g})$. Recall that a *central character* is a homomorphism $\chi\colon Z\,(\mathfrak{g})\to\mathbb{C}$. We say that a $\mathfrak{g}$-module M has a central character χ if for any $z\in Z\,(\mathfrak{g})\,,x\in M$ there is $N\in\mathbb{Z}_{\ge 0}$ such that $(z-\chi\,(z)\,\mathrm{id})^N\cdot x=0$. Clearly any finite-dimensional indecomposable $\mathfrak{g}$-module has some central character, and any finite-dimensional $\mathfrak{g}$-module decomposes into a direct sum of submodules with central characters.

We use a Harish–Chandra homomorphism $HC\colon Z\,(\mathfrak{g})\hookrightarrow \mathrm{Pol}\,(\mathfrak{h}^*)$. The construction of this homomorphism is the same as for semi-simple Lie algebras (see for example [3]). Thus, any $\lambda\in\mathfrak{h}^*$ defines a central character χ_λ by the rule $\chi_\lambda\,(z)=HC\,(z)\,(\lambda)$. Definition of HC immediately implies that an irreducible module L_λ has a central character χ_λ. A central character χ is *dominant* if $\chi=\chi_\lambda$ at least for one $\lambda\in\Lambda^+$.

The following statement was first formulated in [8] and proved in [19] and [12].

PROPOSITION 2.3. *Let $\lambda, \mu \in \Lambda$, W be the Weyl group of $\mathfrak{g}_0$. Then $\chi_\lambda = \chi_\mu$ iff there is a sequence of isotropic roots $\alpha_1, \ldots, \alpha_s \in \Delta^{is}$ and $w \in W$ such that $\mu = w(\lambda + \alpha_1 + \cdots + \alpha_s)$ and $(\lambda + \alpha_1 + \cdots + \alpha_{i-1}, \alpha_i) = 0$ for $i = 1, \ldots, s$.*

Let $\mathcal{F}_\chi = \mathcal{F}_\chi(\mathfrak{g})$ be the full subcategory of $\mathcal{F}$ consisting of modules with central character χ. Obviously, $\mathcal{F} = \bigoplus \mathcal{F}_\chi$, where the summation is taken over all dominant central characters χ. Different categories $\mathcal{F}_\chi$ may be equivalent. They fall into one of 4 series as we will see in Theorem 2.6.

State more constructive condition for $\chi_\lambda = \chi_\mu$. For $\lambda \in \Lambda^+$ let $A_\lambda = \{\alpha_1, \ldots, \alpha_k\}$ be a maximal set of mutually orthogonal positive isotropic roots such that $(\lambda, \alpha_i) = 0$ for $i = 1, \ldots, k$. If $\mathfrak{g} = gl(m|n)$, then the set A_λ is uniquely defined. For *osp*-type $\mathfrak{g}$ we choose A_λ in such way that $T_\lambda \subseteq A_\lambda$. The number $k = |A_\lambda|$ is called *the degree of atypically* of λ and is denoted by $\#\lambda$. A weight $\lambda \in \Lambda^+$ is *typical* if $\#\lambda = 0$.

Let $\bar{\mathfrak{h}}^*_\lambda$ be the subspace of $\mathfrak{h}^*$ generated by all basis vectors ε_i, δ_j orthogonal to the roots from A_λ. Denote by $\bar{\lambda}$ the image of λ under the orthogonal projection $\mathfrak{h}^* \to \bar{\mathfrak{h}}^*_\lambda$. One can see that if $\#\lambda = \#\mu$, then $w\left(\bar{\mathfrak{h}}^*_\lambda\right) = \bar{\mathfrak{h}}^*_\mu$ for some $w \in W$.

PROPOSITION 2.4. *Let $\lambda, \mu \in \Lambda^+$. Then $\chi_\lambda = \chi_\mu$ iff $\#\lambda = \#\mu$ and $w\left(\bar{\lambda}\right) = \bar{\mu}$ for some $w \in W$.*

By Proposition 2.4 one can correctly define $\#\chi$ for a dominant central character χ by putting $\#\chi \stackrel{\text{def}}{=} \#\lambda$ for any dominant λ with $\chi_\lambda = \chi$. Moreover, one can define $\bar{\chi}$ as a typical central character for Lie superalgebra $\bar{\mathfrak{g}}$, here $\bar{\mathfrak{g}}$ is an appropriate subalgebra of $\mathfrak{g}$ isomorphic to $gl(m-k|n-k)$, $osp(2(m-k)|2(n-k))$ or $osp(2(m-k)+1|2(n-k))$ depending on the type of $\mathfrak{g}$, and $k = \#\chi$.

THEOREM 2.5. *A category $\mathcal{F}_\chi$ is indecomposable for any dominant central character χ. If $\mathfrak{g} = gl(m|n)$ or $osp(2m+1|2n)$, then two categories $\mathcal{F}_{\chi_1}$ and $\mathcal{F}_{\chi_2}$ are equivalent iff $\#\chi_1 = \#\chi_2$.*

Let $\mathfrak{g} = osp(2m|2n)$ and $\bar{\tau}$ be the outer automorphism of $\bar{\mathfrak{g}}$ induced by the symmetry of Dynkin diagram. Then two categories $\mathcal{F}_{\chi_1}$ and $\mathcal{F}_{\chi_2}$ are equivalent iff $\#\chi_1 = \#\chi_2$, and for both $i = 1, 2$ either $\bar{\tau}(\bar{\chi}_i) = \bar{\chi}_i$ or $\bar{\tau}(\bar{\chi}_i) \neq \bar{\chi}_i$.

An indecomposable category $\mathcal{F}_\chi$ is called a *block* of $\mathcal{F}$.

THEOREM 2.6. *Let $\mathfrak{g} = gl(m|n)$, $osp(2m+1|2n)$ or $osp(2m|2n)$. A block $\mathcal{F}_\chi$ with $\#\chi = k$ is equivalent to $\mathcal{F}'_{\chi_{\rho'}} \stackrel{def}{=} \mathcal{F}_{\chi_{\rho'}}(\mathfrak{g}')$, where $\mathfrak{g}' = gl(k|k)$ if $\mathfrak{g} = gl(m|n)$, $\mathfrak{g}' = osp(2k+1|2k)$ if $\mathfrak{g} = osp(2m+1|2n)$, $\mathfrak{g}' = osp(2k+2|2k)$ if $\mathfrak{g} = osp(2m|2n)$ and $\bar{\tau}(\bar{\chi}) = \bar{\chi}$, $\mathfrak{g}' = osp(2k|2k)$ if $\mathfrak{g} = osp(2m|2n)$ and $\bar{\tau}(\bar{\chi}) \neq \bar{\chi}$, here ρ' is the analogue of ρ for $\mathfrak{g}'$.*

Let $\Phi\colon \mathcal{F}_\chi \to \mathcal{F}'_{\chi_{\rho'}}$ be a functor establishing equivalence of categories of theorem 2.6. Then Φ sends irreducible objects to irreducible objects, describe the corresponding mapping of highest weights:

PROPOSITION 2.7. *Let $L_\lambda \in \mathcal{O}\mathrm{b}\,\mathcal{F}_\chi$ and $L_{\lambda'} = \Phi L_\lambda$. Let $\lambda = \sum_{i=1}^n a_i\delta_i + \sum_{j=1}^m b_j\varepsilon_j$.*

1. *If* $\mathfrak{g} = gl\,(m|n)$, $A_\lambda = \{\alpha_1 = \delta_{i_1} - \varepsilon_{j_1}, \ldots, \alpha_k = \delta_{i_k} - \varepsilon_{j_k}\}$, *then*

$$\lambda' = \sum_{p=1}^{k} a'_p\,(\delta_p - \varepsilon_{k-p+1})\,,\ a'_p = a_{i_p} + (\rho, \alpha_p)\,;$$

2. *If* $\mathfrak{g} = osp\,(2m|2n)$ *or* $osp\,(2m+1|2n)$, *and*

$$A_\lambda = \left\{\delta_{i_1} + \varepsilon_{j_1}, \ldots, \delta_{i_r} + \varepsilon_{j_r}, \delta_{i_{r+1}} - \varepsilon_{j_{r+1}}, \ldots, \delta_{i_k} - \varepsilon_{j_k}\right\},$$

then $\lambda' = \sum_{p=1}^{k} \left(a'_p \delta_p + |a'_{s(p)}|\varepsilon_p\right)$, $a'_p = a_{i_p} - x_p \cdot \operatorname{sgn} a_{i_p}$, *where*

$$x_p = \#\left\{q, l \mid \left(\bar{\lambda}, \delta_q\right) < |a_{i_p}|, 0 \neq -\left(\bar{\lambda}, \varepsilon_l\right) < |a_{i_p}|\right\},$$

and s *is a permutation such that* $|a_{s(1)}| > \cdots > |a_{s(k)}|$. *In particular* $t_\lambda = t_{\lambda'}$.

EXAMPLE 2.8. Let $\mathfrak{g} = gl\,(3|3)$, $\lambda = 5\delta_1 + 3\delta_2 - \delta_3 + 4\varepsilon_1 + 3\varepsilon_2 + \varepsilon_3$. Then $A_\lambda = \{\delta_3 - \varepsilon_3\}$, $\#\lambda = 1$, $\bar{\mathfrak{g}} \simeq gl\,(2|2)$, $\mathfrak{g}' \simeq gl\,(1|1)$, $\bar{\lambda} = 5\delta_1 + 3\delta_2 + 3\varepsilon_2 + \varepsilon_3$, $\lambda' = -3\delta_1 + 3\varepsilon_1$.

EXAMPLE 2.9. Let $\mathfrak{g} = osp\,(8|6)$, $\lambda = 4\delta_1 + \delta_2 - 3\delta_3 + 3\varepsilon_1 + 2\varepsilon_2 + \varepsilon_3 + 0\varepsilon_4$. Then $A_\lambda = \{\delta_2 + \varepsilon_3, \delta_3 - \varepsilon_1\}, t_\lambda = 1$, $\#\lambda = 2$, $\bar{\mathfrak{g}} \simeq osp\,(4|2)$, $\mathfrak{g}' \simeq osp\,(6|4)$, $\bar{\lambda} = 4\delta_1 + 2\varepsilon_2 + 0\varepsilon_4$, $\lambda' = \delta_1 - 2\delta_2 + 2\varepsilon_1 + \varepsilon_2 + 0\varepsilon_3$.

3. Borel–Weil–Bott theorem and geometric induction

Let $\mathfrak{b} = \mathfrak{h} \oplus \mathfrak{n}^+$ be a Borel subalgebra. For a *parabolic* subalgebra $\mathfrak{p} \supseteq \mathfrak{b}$ denote by $\Delta_\mathfrak{p}$ the set of roots α such that $\mathfrak{g}_{\pm\alpha} \subseteq \mathfrak{p}$. Denote by $L_\lambda\,(\mathfrak{p})$ an irreducible $\mathfrak{p}$-module with highest weight λ, which of course remains irreducible after restriction to a reductive subalgebra $\mathfrak{g}_\mathfrak{p} = \mathfrak{h} \oplus \bigoplus_{\alpha \in \Delta_\mathfrak{p}} \mathfrak{g}_\alpha$. Due to the geometric origin of the following argument we will need representations of the *supergroup* P corresponding to $\mathfrak{p}$. A *flag supermanifold* G/P is a compact homogeneous supermanifold with the underlying manifold G_0/P_0 (for definitions see [11] or [10]). Any finite-dimensional P-module M induces a vector bundle $\mathcal{O}\,(M)$ on G/P with the fiber M over P. Let[2]

$$\Gamma_i^P\,(M) = \left[H^i_{G/P}\mathcal{O}\,(M^*)\right]^*.$$

Since G/P is compact, $\Gamma_i^P\,(M)$ is a finite-dimensional G-module. One can consider Γ_i^P as a derived functor from the category $\mathcal{F}\,(P)$ of finite dimensional P-modules to the category $\mathcal{F}$.

LEMMA 3.1. *Let* $\mathfrak{n}_\mathfrak{p}$ *be the nilpotent ideal in* $\mathfrak{p}$ *such that* $\mathfrak{p} = \mathfrak{g}_\mathfrak{p} \oplus \mathfrak{n}_\mathfrak{p}$. *If* $X \in \operatorname{Ob}\mathcal{F}$, *then coinvariants* $X_{\mathfrak{n}_\mathfrak{p}}$ *form a* P*-module, there is a natural inclusion* $X \hookrightarrow H^0_{G/P}\mathcal{O}\left(X_{\mathfrak{n}_\mathfrak{p}}\right)$.

If $M \in \operatorname{Ob}\mathcal{F}_\chi$, *then* $\Gamma_i^P\,(M^{\mathfrak{n}_\mathfrak{p}}) \in \operatorname{Ob}\mathcal{F}_\chi$, *and there is a natural projection* $\Gamma_0^P\,(M^{\mathfrak{n}_\mathfrak{p}}) \to M$. *In particular when* $M = L_\lambda$ *there is a natural projection* $\Gamma_0^P\,(L_\lambda\,(P)) \to L_\lambda$.

[2]We use doubled duality to avoid a problematic notion of antidominant weight.

LEMMA 3.2. *Let M be a P-module and $E^P(M) = \sum (-1)^i \operatorname{ch} \Gamma_i^P(M)$. Consider D from* (1.1). *Then*

$$E^P(M) = D \sum_{w \in W} \operatorname{sgn} w \cdot w \left(\frac{e^\rho \operatorname{ch} M}{\prod_{\alpha \in \Delta_\mathfrak{p} \cap \Delta_1^+} (1 + e^{-\alpha})} \right). \tag{3.1}$$

The following theorem is a generalization of a result by Penkov–Skornyakov. We say that $\lambda \in \Lambda^+$ is *P-typical* if $A_\lambda \subseteq \Delta_\mathfrak{p}$.

THEOREM 3.3. *Let λ be a P-typical dominant weight. Then $\Gamma_0^P(L_\lambda(\mathfrak{p})) = L_\lambda$ and $\Gamma_i^P(L_\lambda(\mathfrak{p})) = 0$ for $i > 0$.*

Generalizing this theorem, denote by $[M : L_\mu]$ the multiplicity of an irreducible module L_μ in a $\mathfrak{g}$-module M. Define the *Kazhdan–Lusztig polynomials* and coefficients by:

$$K_{\lambda,\mu}^P(q) = \sum_{i=0}^{\dim G/P} \left[\Gamma_i^P(L_\lambda(\mathfrak{p})) : L_\mu\right] q^i, \qquad a_{\lambda,\mu}^P = K_{\lambda,\mu}^P(-1).$$

Let $E_\lambda^P \stackrel{\text{def}}{=} E^P(L_\lambda(\mathfrak{p}))$. Clearly

$$\sum_\mu a_{\lambda,\mu}^P \operatorname{ch} L_\mu = E_\lambda^P. \tag{3.2}$$

PROPOSITION 3.4. *Let $\mathfrak{p}$ be a parabolic subalgebra, $\lambda \in \Lambda^+$ and $T_\lambda \subseteq \Delta_\mathfrak{p}$. Then $K_{\lambda,\lambda}^P = 1$, and $K_{\lambda,\mu}^P \neq 0$ implies $\mu \leq \lambda$, $\chi_\lambda = \chi_\mu$.*

Let $\mathfrak{g} = gl(m|n)$, then $t_\lambda = 0$. Put $P = B$. By Proposition 3.4 the matrix $\left(a_{\lambda,\mu}^B\right)$ is unipotent, thus easy to invert. Let $(b_{\lambda,\mu}) = \left(a_{\lambda,\mu}^B\right)^{-1}$. Then the equations (3.2) imply $\operatorname{ch} L_\lambda = \sum_{\mu \leq \lambda} b_{\lambda,\mu} E_\mu^B$.

If $\mathfrak{g}$ is an algebra of *osp* type then the matrix $\left(a_{\lambda,\mu}^B\right)$ is not invertible. Let $\Lambda_s^+ = \{\lambda \in \Lambda^+ \mid t_\lambda = s\}$ and $\mathfrak{p}^{(r)}$ be the parabolic subalgebra such that $\Delta_{\mathfrak{p}^{(r)}}$ is generated by the simple roots $\sigma_r, \ldots, \sigma_{n+m}$. One can see that $T_\lambda \subseteq \Delta_{\mathfrak{p}^{(r+1)}}$ for any $\lambda \in \Lambda_{n-r}^+$. By Proposition 3.4 the matrix $\left(a_{\lambda,\mu}^{P^{(r+1)}}\right)_{\lambda,\mu \in \Lambda_{n-r}^+}$ is again unipotent, thus easy to invert. Since $\mu \leq \lambda$ implies $t_\mu \geq t_\lambda$, the equation

$$\sum_{\mu \in \Lambda_{n-r}^+} a_{\lambda,\mu}^{P^{(r+1)}} \operatorname{ch} L_\mu = E_\lambda^{P^{(r+1)}} - \sum_{t_\nu > t_\lambda} a_{\lambda,\nu}^{P^{(r+1)}} \operatorname{ch} L_\nu.$$

This taken together with (3.1) expresses $\operatorname{ch} L_\lambda$ in terms of $\operatorname{ch} L_\mu\left(\mathfrak{p}^{(r+1)}\right) = \operatorname{ch} L_\mu\left(\mathfrak{g}_{\mathfrak{p}^{(r+1)}}\right)$, $a_{\mu,\nu}^{P^{(r+1)}}$ for $r = n - t_\lambda$, $\mu, \nu \in \Lambda^+$ with $t_\mu = t_\lambda$, $t_\nu \geq t_\lambda$, and $\operatorname{ch} L_{\nu'}$ for $t_{\nu'} > t_\lambda$. If $t_\lambda < n$, $\operatorname{rk} \mathfrak{g}_{\mathfrak{p}^{(r+1)}} < \operatorname{rk} \mathfrak{g}$, which gives a recurrence relation for $\operatorname{ch} L_\lambda$.

What remains is the case $t_\lambda = n$. Then $T_\lambda = A_\lambda$, and λ is Q-typical for the parabolic subalgebra $\mathfrak{q}$ with $\Delta_\mathfrak{q}$ generated by $\sigma_1, \ldots, \sigma_{m+n-1}$. By theorem 3.3

$$\operatorname{ch} L_\lambda = E_\lambda^Q = D \sum_{w \in W} \operatorname{sgn} w \cdot w \left(\frac{e^\rho \operatorname{ch} L_\lambda(\mathfrak{q})}{\Pi_{\alpha \in \Delta_\mathfrak{q} \cap \Delta_1^+} (1 + e^{-\alpha})} \right) \ldots$$

On the other hand, $\operatorname{ch} L_\lambda(\mathfrak{q}) = \operatorname{ch} L_\lambda(\mathfrak{g}_\mathfrak{q})$, and $\mathfrak{g}_\mathfrak{q}$ is isomorphic to $gl(m|n)$. Since the case $\mathfrak{g} = gl(m|n)$ is already covered, we can calculate $\operatorname{ch} L_\lambda$.

These arguments reduce the calculation of $\operatorname{ch} L_\lambda$ to the calculation of the matrix $\left(a^P_{\mu,\nu}\right)$. The next statement reduces the latter problem to the case of the most atypical central character.

PROPOSITION 3.5. *Let $\lambda \in \Lambda^+$, λ' and $\mathfrak{g}'$ be as in theorem 2.6 and proposition 2.7. Let $\mathfrak{p} = \mathfrak{b}$ if $\mathfrak{g} = gl(m|n)$, $\mathfrak{p} = \mathfrak{p}^{(r+1)}$ if $\mathfrak{g} = osp(2m|2n)$ or $osp(2m+1|2n)$ and $t_\lambda = n - r$. Let $\mathfrak{p}'$ be the analogous parabolic subalgebra in $\mathfrak{g}'$ determined by λ'. Then $K^P_{\lambda,\mu} = K^{P'}_{\lambda',\mu'}$.*

4. CALCULATION OF COEFFICIENTS $a_{\lambda,\mu}$ IN $\mathcal{F}_{\chi_\rho}$

In this section we concentrate on calculation of coefficients $a^P_{\lambda,\mu}$. By Proposition 3.5 it is sufficient to find these coefficients only for the most atypical block $\mathcal{F}_{\chi_\rho}$ and $\mathfrak{g} = gl(k|k)$, $osp(2k|2k)$, $osp(2k+2|2k)$ or $osp(2k+1|2k)$.

First, consider the case $\mathfrak{g} = gl(k|k)$. Here $t_\lambda = 0$ and $P = B$. Introduce a formal operator A in the Grothendieck ring of $\mathcal{F}_{\chi_\rho}$ by the formula

$$A[L_\lambda] = \sum_{\mu\in\Lambda} a^P_{\lambda,\mu}[L_\mu].$$

Let $\mathfrak{p}^{(i)}$ be the parabolic subalgebra such that $\Delta_{\mathfrak{p}^{(i)}}$ is generated by the simple roots $\sigma_i, \ldots, \sigma_{2k-i}$. Consider the flag of parabolic subalgebras $\mathfrak{g} = \mathfrak{p}^{(1)} \supset \mathfrak{p}^{(2)} \supset \cdots \supset \mathfrak{p}^{(k)} \supset \mathfrak{p}^{(k+1)} = \mathfrak{b}$. Note that $P^{(i)}/P^{(i+1)}$ is isomorphic to the supermanifold of $(1|1)$-dimensional subspaces in $\mathbb{C}^{i|i}$. As before define derived functors

$$\Gamma^{(i)}_j \colon \mathcal{F}\left(\mathfrak{p}^{(i+1)}\right) \to \mathcal{F}\left(\mathfrak{p}^{(i)}\right), \qquad \Gamma^{(i)}_j(M) = \left[H^j_{P^{(i)}/P^{(i+1)}}\mathcal{O}(M^*)\right]^*,$$

generating functions $K^{(i)}$ and coefficients $a^{(i)}$ by

$$K^{(i)}_{\lambda,\mu}(q) = \sum_j \left[\Gamma^{(i)}_j\left(L_\lambda\left(\mathfrak{p}^{(i+1)}\right)\right) : L_\mu\left(\mathfrak{p}^{(i)}\right)\right] q^j, \qquad a^{(i)}_{\lambda,\mu} = K^{(i)}_{\lambda,\mu}(-1).$$

Define the operators $A^{(i)}[L_\lambda] = \sum_\mu a^{(i)}_{\lambda,\mu}[L_\mu]$. Obviously,

$$A = A^{(1)} \circ \cdots \circ A^{(k)}. \tag{4.1}$$

Theorem 4.1 below gives recurrence relations for calculating polynomials $K^{(i)}_{\lambda,\mu}$. It is the most difficult result in the paper. Before stating it let us recall that any $\lambda \in \Lambda^+$ with $\chi_\lambda = \chi_\rho$ can be written as $a_1\alpha_1 + \cdots + a_k\alpha_k$ where $\alpha_i = \delta_i - \varepsilon_{k+1-i} \in A_\lambda$ and $a_1 > a_2 > \cdots > a_k$. For $S(q) \in \mathbb{Z}\left[q, q^{-1}\right]$ denote the polynomial part of S by S_+.

THEOREM 4.1. *Let $\mathfrak{g} = gl(k|k)$. Then the following recurrence relations hold:*

1. $K^{(i)}_{\lambda,\lambda} = 1$;
2. *if $a_i > a_{i+1} + 1$, then $K^{(i)}_{\lambda,\lambda-\alpha_i} = 1$ and $K^{(i)}_{\lambda,\mu} = \left(q^{-1}K^{(i)}_{\lambda-\alpha_i,\mu}\right)_+$ for any $\mu \neq \lambda, \lambda - \alpha$;*
3. *if $a_i = a_{i+1} + 1$, then $K^{(i)}_{\lambda,\mu} = qK^{(i+1)}_{\lambda-\alpha_i,\mu}$ for any $\mu \neq \lambda, \lambda - \alpha_i$;*

4. $K^{(k)}_{\lambda,\lambda-\alpha_k} = 1$ *and* $K^{(k)}_{\lambda,\mu} = 0$ *for any* $\mu \neq \lambda, \lambda - \alpha_k$.

These relations uniquely determine the polynomials $K^{(i)}_{\lambda,\mu}$.

Note that the proof of Theorem 4.1 unravelled the following beautiful geometric

LEMMA 4.2. *Let* $\lambda \in \Lambda^+$. *The cohomology* $\Gamma^{(i)}_j\left(L_\lambda\left(\mathfrak{p}^{(i+1)}\right)\right)$ *for* $j > 0$ *and the kernel of the natural projection* $\Gamma^{(i)}_0\left(L_\lambda\left(\mathfrak{p}^{(i+1)}\right)\right) \to L_\lambda\left(\mathfrak{p}^{(i)}\right)$ *are semisimple* $\mathfrak{p}^{(i)}$*-modules, and any irreducible component of* $\bigoplus_j \Gamma^{(i)}_j\left(L_\lambda\left(\mathfrak{p}^{(i+1)}\right)\right)$ *occurs with multiplicity 1.*

Consider the case $\mathfrak{g} = osp\,(2k+l|2k)$, where $l = 0, 1$ or 2, and $\chi_\lambda = \chi_\mu$. Consider the flag of parabolic subalgebras $\mathfrak{g} = \mathfrak{p}^{(1)} \supset \mathfrak{p}^{(2)} \supset \cdots \supset \mathfrak{p}^{(k+1)}$, where $\Delta_{\mathfrak{p}^{(i)}}$ is generated by the simple roots $\sigma_i, \ldots, \sigma_{k+[l/2]}$. Note that $P^{(i)}/P^{(i+1)}$ is isomorphic to the supermanifold of $(1|0)$-dimensional subspaces in $\mathbb{C}^{2k-2i|2k+l}$. Let $r = k - t_\lambda$. As it was explained in section 3, we are interested in calculating $a^P_{\lambda,\mu} = a^{P^{(r+1)}}_{\lambda,\mu}$. Using the same notations as for the case $\mathfrak{g} = gl\,(k|k)$ one can write $A = A^{(1)} \circ \cdots \circ A^{(r)}$.

Next, we write recurrence relations for polynomials $K^{(i)}_{\lambda,\mu}$. Write $A_\lambda = \{\alpha_1, \ldots, \alpha_k\}$, where $\alpha_i = \delta_i + \varepsilon_{j_i}$ for $i \leq r$ and $\alpha_i = \delta_i - \varepsilon_{j_i}$ if $i > r$. We assume that $b_k \geq 0$, see remark 2.2. By Proposition 2.7 λ can be written as $\lambda = a_1\alpha_1 + \cdots + a_k\alpha_k$, where $a_i \in \mathbb{Z} + l/2$ and $a_1 > a_2 > \cdots > a_r > 0 \geq a_{r+1} > \cdots > a_k$.

THEOREM 4.3. *Let* $\mathfrak{g} = osp\,(2k+l|2k)$. *Then the following recurrence relations hold:*

1. $K^{(i)}_{\lambda,\lambda} = 1$;
2. *if* $a_i > a_{i+1} + 1$ *and* $a_i \neq 1 - a_j$ *for any* $j > r$, *then* $K^{(i)}_{\lambda,\lambda-\alpha_i} = 1$ *and* $K^{(i)}_{\lambda,\mu} = \left(q^{-1}K^{(i)}_{\lambda-\alpha_i,\mu}\right)_+$ *for any* $\mu \neq \lambda, \lambda - \alpha$;
3. *if* $a_i = 1 - a_j$ *for some* $j > r$, *then* $K^{(i)}_{\lambda,\lambda-\delta_i-\delta_j} = 1$ *and* $K^{(i)}_{\lambda,\mu} = \left(q^{-1}K^{(i)}_{\lambda-\delta_i-\delta_j,\mu}\right)_+$ *for any* $\mu \neq \lambda, \lambda - \delta_i - \delta_j$;
4. *if* $a_i = a_{i+1} + 1$, *then* $K^{(i)}_{\lambda,\mu} = qK^{(i+1)}_{\lambda-\delta_i,\mu+\varepsilon_{j_i}}$ *for any* $\mu \neq \lambda$;
5. *if* $l = 1$ *and* $a_r = 1/2$, *then* $K^{(r)}_{\lambda,\lambda-\delta_r} = q^{2s+1}$ *and* $K^{(r)}_{\lambda,\mu} = 0$ *for any* $\mu \neq \lambda, \lambda - \delta_r$;
6. *if* $l = 2$ *and* $a_r = 1$, *then* $K^{(r)}_{\lambda,\lambda-2\delta_r} = q^{2s+1}$ *and* $K^{(r)}_{\lambda,\mu} = 0$ *for any* $\mu \neq \lambda, \lambda - 2\delta_r$;
7. *if* $l = 0$, $a_r = 1$ *and* $r \neq k$, *then* $K^{(r)}_{\lambda,\lambda-\delta_r-\delta_{r+1}} = q^{2s}$ *and* $K^{(r)}_{\lambda,\mu} = 0$ *for any* $\mu \neq \lambda, \lambda - \delta_r - \delta_{r+1}$;
8. *if* $l = 0$, $a_k = 1$, *then* $K^{(k)}_{\lambda,\lambda-\delta_k-\varepsilon_k} = 1$ *and* $K^{(k)}_{\lambda,\mu} = 0$ *for any* $\mu \neq \lambda, \lambda - \delta_k - \varepsilon_k$.

These relations uniquely determine the polynomials $K^{(i)}_{\lambda,\mu}$.

Note that in the case of $gl\,(m|n)$ and $P = B$ one can calculate $K^P_{\lambda,\mu}$ basing on (4.1) and Theorem 4.1, since in this case it happens that $K^P_{\lambda,\mu} = a^P_{\lambda,\mu}$. In the cases $P \neq B$ or $\mathfrak{g} = osp$ formulae for $K^P_{\lambda,\mu}$ are not known.

REFERENCES

[1] I. N. Bernshteĭn and D. A. Leĭtes, *Invariant differential operators and irreducible representations of Lie superalgebras of vector fields*, Serdica 7 (1981), no. 4, 320–334 (1982).

[2] J. Bernstein and D. Leites, *A formula for the characters of the irreducible finite dimensional representations of lie superalgebras of series gl and sl*, C. R. Acad. Bulgare Sci. 33 (1980), 1049–1051.

[3] J. Dixmier, *Enveloping algebras*, Graduate Studies in Mathematics, vol. 11, American Mathematical Society, Providence, RI, 1996, Revised reprint of the 1977 translation.

[4] A. Frumkin, *The irreducible characters of the Lie superalgebras of type $A(n, m)$ and filtrations of their Kac modules*, Israel J. Math. 96 (1996), no. , part A, 267–279.

[5] J. W. B. Hughes, R. C. King, and J. Van der Jeugt, *On the composition factors of Kac modules for the Lie superalgebras* $\mathrm{sl}(m/n)$, J. Math. Phys. 33 (1992), no. 2, 470–491.

[6] V. G. Kac, *Lie superalgebras*, Adv. Math. 26 (1977), 8–96.

[7] ———, *Representations of classical Lie superalgebras*, Lecture Notes in Math. 676 (1978), 597–626.

[8] ———, *Laplace operators of infinite–dimensional Lie algebras and theta functions*, PNAS USA 81 (1984), 645–647.

[9] V. G. Kac and M. Wakimoto, *Integrable highest weight modules over affine superalgebras and number theory*, Progress in Math. 123 (1994), 415–456.

[10] Yu. Manin, *Gauge field theory and complex geometry*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 289, Springer-Verlag, Berlin, 1988, Translated from the Russian by N. Koblitz and J. R. King.

[11] I. Penkov, *Borel-Weil-Bott theory for classical lie supergroups*, Contemporary Problems in Mathematics, vol. 32, VINITI, Moscow, 1988.

[12] ———, *Generic representations of classical Lie superalgebras and their localization*, Monatsh. Math. 118 (1994), no. 3-4, 267–313.

[13] I. Penkov and V. Serganova, *Cohomology of G/P for classical complex Lie supergroups G and characters of some atypical G-modules*, Ann. Inst. Fourier (Grenoble) 39 (1989), no. 4, 845–873.

[14] ———, *Generic irreducible representations of finite-dimensional Lie superalgebras*, Internat. J. Math. 5 (1994), no. 3, 389–419.

[15] ———, *Characters of irreducible G-modules and cohomology of G/P for the Lie supergroup $G = Q(N)$*, J. Math. Sci. (New York) 84 (1997), no. 5, 1382–1412, Algebraic geometry, 7.

[16] I. Penkov and I. Skornyakov, *Cohomologie des $\mathcal{D}$-modules tordus typiques sur les supervariétés de drapeaux*, C. R. Acad. Sci. Paris Sér. I Math. 299 (1984), no. 20, 1005–1008.

[17] V. Serganova, *Kazhdan-Lusztig polynomials for Lie superalgebra $GL(m,n)$*, Advances in Soviet Mathematics 16 (1993), 151–165.

[18] ———, *Kazhdan-Lusztig polynomials and character formula for the Lie superalgebra $\mathfrak{gl}(m|n)$*, Selecta Math. (N.S.) 2 (1996), no. 4, 607–651.

[19] A. Sergeev, *Enveloping algebra centre for Lie superalgebras GL and Q*, Ph.D. thesis, Moscow State University, Moscow, 1987.

[20] A. V. Shapovalov, *Invariant differential operators and irreducible representations of finite-dimensional Hamiltonian and Poisson Lie superalgebras*, Serdica 7 (1981), no. 4, 337–342 (1982).

[21] J. Van der Jeugt, *Character formulae for the Lie superalgebra $C(n)$*, Comm. Algebra 19 (1991), no. 1, 199–222.

[22] J. Van der Jeugt, J. W. B. Hughes, R. C. King, and J. Thierry-Mieg, *A character formula for singly atypical modules of the Lie superalgebra* $\mathrm{sl}(m/n)$, Comm. Algebra 18 (1990), no. 10, 3453–3480.

Vera Serganova
Dept. of Mathematics
University of California at Berkeley
Berkeley, CA 94720, USA
serganov@@math.berkeley.edu

Topological Methods in Representation Theory

Kari Vilonen[1]

1991 Mathematics Subject Classification: 22, 55, 14

A few years ago Beilinson and Bernstein introduced a localization technique to representation theory of semisimple Lie groups. Their method allows one to translate questions in representation theory to questions in complex algebraic geometry. Beilinson reported on this work at the Warsaw congress [B]. Consequently, in [K2], Kashiwara initiated a research program, in the form of a series of conjectures, that expands the Beilinson-Bernstein picture. In this survey we will report on work inspired by this point of view. The resulting geometry is no longer complex algebraic; it rather involves real (semi-)algebraic sets. Thus, the methods used will be largely topological. A crucial technique is supplied by the characteristic cycle construction of Kashiwara [K1], which amounts to a version of Morse theory. The majority of the results presented here constitute joint work with Wilfried Schmid.

1. Introduction.

Let $G_{\mathbb{R}}$ denote a semisimple Lie group which we assume, for simplicity, to be linear and connected. For example, one can take $G_{\mathbb{R}}$ to be any of the classical groups: $SL_n(\mathbb{R})$, $SO(n,\mathbb{R})$, $SO(p,q)$, ... To provide motivation for things to come, let us consider one of the outstanding problems in representation theory: the determination of the unitary dual $\hat{G}_{\mathbb{R}}$, i.e., the determination of the set of isomorphism classes of irreducible unitary representations of $G_{\mathbb{R}}$. Ideally, at least from the geometric point of view, the solution of the problem would have the following form. There should exist a manifold X with a $G_{\mathbb{R}}$-action such that $\hat{G}_{\mathbb{R}}$ is in bijection with a certain set of $G_{\mathbb{R}}$-equivariant "objects" on X (they could be sheaves, for example). Let $\mathcal{F}$ be such a $G_{\mathbb{R}}$-equivariant object. Because the group $G_{\mathbb{R}}$ acts both on X and on $\mathcal{F}$, it also acts on the cohomology groups $\mathrm{H}^*(X,\mathcal{F})$. These groups should have a canonical structure of a Hilbert space such that the group $G_{\mathbb{R}}$ acts continuously and via unitary operators on them. At this time there is not even a precise conjecture as to what the set $\hat{G}_{\mathbb{R}}$ ought to be in general. However, the orbit method of Kirillov-Kostant suggests that we should take as X the space $\mathfrak{g}_{\mathbb{R}}^*$, the dual of the Lie algebra $\mathfrak{g}_{\mathbb{R}}$ of $G_{\mathbb{R}}$. In other words, we should be able to associate unitary representations to the coadjoint orbits (the $G_{\mathbb{R}}$-orbits on $\mathfrak{g}_{\mathbb{R}}^*$), or more precisely, to collections of coadjoint orbits together with some extra data. Given such a set of data, Kirillov has proposed that there is a specific formula – a "universal character formula" – for the character of the representation attached to the data.

[1]Partially supported by NSF, NSA, and a Guggenheim fellowship

As we pointed out at the beginning, the above discussion was included only as motivation. In this survey we will work with the class of admissible (finite length) representations. They include all irreducible unitary representations. For this larger class of representations we will:

a) parametrize and exhibit them geometrically
b) give a geometric character formula in the spirit of Kirillov's formula
c) analyze the nilpotent invariants attached to them

2. Geometric parametrization of representations.

Let G denote the complexification of $G_{\mathbb{R}}$ and let X be the flag manifold of G. The group $G_{\mathbb{R}}$ acts on X with finitely many orbits. Let us assume, for the moment, that $G_{\mathbb{R}}$ is compact. Then $G_{\mathbb{R}}$ acts transitively on X and there is only one orbit. All the irreducible representations of $G_{\mathbb{R}}$ are finite dimensional. As is well know, they can be classified and exhibited explicitly as follows. To each $\lambda \in \mathrm{H}^2(X,\mathbb{Z})$ corresponds a complex line bundle $\mathcal{O}(\lambda)$ on X. This line bundle is holomorphic and $G_{\mathbb{R}}$-homogenous, i.e., the action of $G_{\mathbb{R}}$ on X lifts to an action of $G_{\mathbb{R}}$ on $\mathcal{O}(\lambda)$ (strictly speaking, this it true only if $G_{\mathbb{R}}$ is simply connected; if this is not the case, then λ must lie in a sublattice of $\mathrm{H}^2(X,\mathbb{Z})$). Thus we get a representation of $G_{\mathbb{R}}$ on the vector spaces $\mathrm{H}^k(X,\mathcal{O}(\lambda))$. All irreducible representations of $G_{\mathbb{R}}$ arise in this fashion. Furthermore, if we restrict the parameter λ to lie in the dominant cone in $\mathrm{H}^2(X,\mathbb{Z})$ then each irreducible representation occurs exactly once among the representations $\mathrm{H}^0(X,\mathcal{O}(\lambda))$.

When the group $G_{\mathbb{R}}$ is not assumed to be compact, the situation is more complicated. First of all, we have to allow the "twisting" parameter λ to lie in $\mathrm{H}^2(X,\mathbb{C})$, not just in the lattice $\mathrm{H}^2(X,\mathbb{Z})$. To each such $\lambda \in \mathrm{H}^2(X,\mathbb{C})$ we associate the "twisted" G-equivariant sheaf $\mathcal{O}_X^{\mathrm{an}}(\lambda)$ of holomorphic functions on X. This is an "ordinary" sheaf on X only if λ is integral. The second complication arises because the action of $G_{\mathbb{R}}$ on X is not transitive. As a first approximation, we can construct representations $\mathrm{H}^k(S,\mathcal{O}_X^{\mathrm{an}}(\lambda))$ associated to each $G_{\mathbb{R}}$-orbit S and the parameter $\lambda \in \mathrm{H}^2(X,\mathbb{C})$. This construction yields all the "standard representations" but not all the irreducible (admissible) representations of $G_{\mathbb{R}}$.

To get all the representations, we have to allow combinations of $G_{\mathbb{R}}$-orbits and we have to allow the orbits to "interact" with each other. This can be accomplished purely topologically: we consider $G_{\mathbb{R}}$-equivariant (complexes of) $\mathbb{C}$-sheaves on X whose stalks are finite dimensional over $\mathbb{C}$. Note that the category of $G_{\mathbb{R}}$-sheaves on a $G_{\mathbb{R}}$-orbit S is equivalent to the category of (finite dimensional) complex representations of the component group $(G_{\mathbb{R}})_x/(G_{\mathbb{R}})_x^0$ of $(G_{\mathbb{R}})_x$. Here $(G_{\mathbb{R}})_x$ denotes the stabilizer group of any particular $x \in S$. A general $G_{\mathbb{R}}$-equivariant sheaf is "glued" together from such local systems on the various $G_{\mathbb{R}}$-orbits. Technically, these sheaves should be twisted, with twist $-\lambda$, and we should consider them in the context of derived categories, i.e., we should view them as elements in the $G_{\mathbb{R}}$-equivariant derived category of $\mathbb{C}$-sheaves with twist $-\lambda$. For the purposes of this survey, this technical point can be ignored and one can think of them just as sheaves with a $G_{\mathbb{R}}$-action. In particular, one can assume that $\lambda = 0$, in which case $\mathcal{O}(\lambda)$ is the trivial line bundle on X. We define functors

$$\{G_{\mathbb{R}}\text{-equivariant sheaves on } X\} \longrightarrow \{G_{\mathbb{R}}\text{-representations}\} \tag{1.1a}$$

by

$$\mathcal{F} \quad \longmapsto \quad \mathrm{H}^k(X, \mathcal{F} \otimes_{\mathbb{C}} \mathcal{O}_X^{\mathrm{an}}(\lambda))\,. \tag{1.1b}$$

In [KSd] it is shown that the cohomology groups $\mathrm{H}^k(X, \mathcal{F} \otimes_{\mathbb{C}} \mathcal{O}_X^{\mathrm{an}}(\lambda))$ carry a natural Fréchet topology such that the action of $G_{\mathbb{R}}$ is continuous. The topology is induced from the natural topology on $\mathcal{O}_X^{\mathrm{an}}(\lambda)$. In representation theoretic terms the choice of the parameter λ amounts to fixing the infinitesimal character of the representations in (1.1): the space $\mathrm{H}^2(X, \mathbb{C})$ can be identified[2] with the dual of a Cartan $\mathfrak{t}$ in $\mathfrak{g}$.

The representations produced by the functor (1.1) are admissible. In the rest of this paper, the term $G_{\mathbb{R}}$-representation stands for an admissible representation (of finite length). Recall that a $G_{\mathbb{R}}$-representation V is called admissible if, when viewed as a representation of a maximal compact subgroup $K_{\mathbb{R}}$ of $G_{\mathbb{R}}$, each irreducible representation of $K_{\mathbb{R}}$ appears in it with finite multiplicity. We consider admissible representation modulo infinitesimal equivalence. In other words, we identify representations if they are "the same" except for the topology that we put on the representation space. When we work up to infinitesimal equivalence the functor (1.1) is onto. The infinitesimal equivalence class of a $G_{\mathbb{R}}$-representation V is captured by its Harish-Chandra module. Recall that the Harish-Chandra module M of the representation V consists of all vectors $v \in V$ such that $K_{\mathbb{R}} \cdot v$ generates a finite dimensional subspace of V. Both the the lie algebra $\mathfrak{g}_{\mathbb{R}}$ and the group $K_{\mathbb{R}}$, and hence their complexifications $\mathfrak{g}$ and K, act compatibly on M. Harish-Chandra modules are algebraic objects and are not equipped with a topology.

Let us continue to consider a particular $G_{\mathbb{R}}$-representation which is associated to the parameter $\lambda \in \mathrm{H}^2(X, \mathbb{C})$ and a $G_{\mathbb{R}}$-sheaf $\mathcal{F}$. To construct the Harish-Chandra module associated to this representation geometrically, we appeal to the work of Beilinson-Bernstein. Slightly paraphrased, they constructed functors

$$\{K\text{-equivariant sheaves on } X\} \longrightarrow \{\text{H-C-modules}\} \tag{1.2a}$$

by

$$\mathcal{F} \quad \longmapsto \quad \mathrm{H}^k(X, \mathcal{F} \otimes_{\mathbb{C}} \mathcal{O}_X^{\mathrm{alg}}(\lambda))\,. \tag{1.2b}$$

Here, just as in our previous discussion, the sheaf $\mathcal{F}$ is properly viewed as an element in the K-equivariant derived category of $\mathbb{C}$-sheaves with twist $-\lambda$. The symbol $\mathcal{O}_X^{\mathrm{alg}}(\lambda)$ stands for the twisted sheaf of complex algebraic functions on X. It is a subsheaf of $\mathcal{O}_X^{\mathrm{an}}(\lambda)$. On the other hand, in [MUV] we construct an equivalence of categories

$$\{G_{\mathbb{R}}\text{-equivariant sheaves on } X\} \xrightarrow{\Gamma} \{K\text{-equivariant sheaves on } X\}\,. \tag{1.3}$$

[2]Under this identification the value $\lambda = 0$ corresponds to the element "ρ = half the sum of positive roots" in $\mathfrak{t}^*$.

This equivalence is constructed via an averaging procedure. Loosely speaking, we average a $G_{\mathbb{R}}$-sheaf over the orbits of $K/K_{\mathbb{R}}$. The Harish-Chandra module of the representation associated to the $G_{\mathbb{R}}$-sheaf $\mathcal{F}$ is gotten by applying the Beilinson-Bernstein functor (1.2) to the sheaf $\Gamma\mathcal{F}$. The commutative diagram below summarizes our discussion:

$$\begin{array}{ccc} \{G_{\mathbb{R}}\text{-representations}\} & \longrightarrow & \{\text{H-C-modules}\} \\ {\scriptstyle (1.1)}\uparrow & & \uparrow{\scriptstyle (1.2)} \\ \{G_{\mathbb{R}}\text{-equivariant sheaves on } X\} & \xrightarrow{\Gamma} & \{K\text{-equivariant sheaves on } X\}\,. \end{array} \tag{1.4}$$

We have have not specified the degree k that we should use for the cohomology groups in formulas (1.1b) and (1.2b). If we restrict λ to lie in the dominant cone then there is a natural choice of a subcategory of complexes of $G_{\mathbb{R}}$-sheaves and a subcategory of complexes of K-sheaves such that the functors (1.1b) and (1.2b) are nonzero on these subcategories only for the value $k = 0$. Furthermore, restricted to these subcategories the functors (1.1b) and (1.2b) are equivalences, provided that the parameter λ is regular[3]. On the K-side the subcategory has a useful characterization. It consists of K-equivariant perverse sheaves. On the $G_{\mathbb{R}}$-side no direct characterization is known.

The motivation for the original work of Beilinson-Bernstein was to understand how standard representations decompose into irreducibles (Kazhdan-Lusztig conjectures). Via the functor (1.2) this problem translates into the problem of understanding how standard perverse sheaves decompose into irreducible perverse sheaves. This problem, in turn, can be solved by using the theory of mixed sheaves. For a survey, see Beilinson's talk at the Warsaw congress [B]. In the same vain other questions in representation theory can be translated to questions about the geometry of (closures of) K-orbits. Because K is a complex algebraic group, we are in the context of (complex) algebraic geometry. The situation on the $G_{\mathbb{R}}$-side is different. The $G_{\mathbb{R}}$-orbits are only semi-algebraic sets and hence appear to be more difficult to work with. In the rest of this paper we use topological techniques that allow one to work on the $G_{\mathbb{R}}$-side. Although the categories of $G_{\mathbb{R}}$-equivariant sheaves and K-equivariant sheaves are equivalent certain things appear to be easier to extract from one side than the other. For example, it appears impossible at this time to give a proof of the Kazhdan-Lusztig conjectures on the $G_{\mathbb{R}}$-side. On the other hand, the character formula that we explain in the next section crucially depends on the $G_{\mathbb{R}}$-side.

3. A geometric character formula.

In this section we will explain a character formula which can be viewed as a generalization of Kirillov's "universal character formula", valid for all admissible representations. Recall that to any representation we can associate its character which is a conjugation invariant, locally L^1-function on the group $G_{\mathbb{R}}$. The function on $\mathfrak{g}_{\mathbb{R}}$ gotten by pulling back the character under the exponential map (and multiplied by the square root of the Jacobian) is called the Lie algebra character.

[3]If we identify $\mathrm{H}^2(X, \mathbb{C})$ with a Cartan $\mathfrak{t}$ this amounts to the regularity of $\lambda + \rho$.

Let us recall the formula proposed by Kirillov:

(2.1) The Lie algebra character of the representation associated to the coadjoint orbit $\mathcal{O}_{\mathbb{R}}$ of $\mathfrak{g}_{\mathbb{R}}^*$ is the Fourier transform of the canonical measure on $\mathcal{O}_{\mathbb{R}}$.

As a coadjoint orbit, $\mathcal{O}_{\mathbb{R}}$ has a canonical symplectic form and hence a canonical measure. In [R1] Rossmann gave a proof of Kirillov's formula for tempered representations, i.e., for the irreducible unitary representations that "appear" in the regular representation $L^2(G_{\mathbb{R}})$.

In [R2], Rossmann made the following proposal to obtain a Kirillov type character formula in general. Let us fix the parameter $\lambda \in \mathrm{H}^2(X, \mathbb{C})$ and let us consider $G_{\mathbb{R}}$-representations associated to this parameter. Recall that the dual $\mathfrak{t}^*$ of any Cartan $\mathfrak{t} \subset \mathfrak{g}$ can be identified with $\mathrm{H}^2(X, \mathbb{C})$ (see footnote 1). Hence, the element λ specifies a coadjoint G-orbit $\Omega_\lambda \subset \mathfrak{g}^*$. If λ is regular, as we will assume from now on, there is an isomorphism $\mu_\lambda : T^*X \to \Omega_\lambda$, due to Rossmann, which he calls the twisted moment map. To have some feel for this map, we describe it loosely. First of all, it is the twisted version of the moment map $\mu : T^*X \to \mathfrak{g}^*$ for the G-action. The moment map μ has its image in the nilpotent cone $\mathcal{N}^*$ in $\mathfrak{g}^* \cong \mathfrak{g}$. Note that, under the identification $\mathfrak{g}^* \cong \mathfrak{g}$, the cotangent space T_x^*X is identified with $\mathfrak{n}_x$, where $\mathfrak{n}_x$ is the nilpotent radical of the Borel subalgebra corresponding to x. With these identifications the map μ is the identity on T_x^*X. To describe μ_λ, let $U_{\mathbb{R}}$ be the compact form of G which is "compatible" with $K_{\mathbb{R}}$ and $G_{\mathbb{R}}$. Because $U_{\mathbb{R}}$ acts transitively on X, the flag manifold X can be identified with a canonical $U_{\mathbb{R}}$-orbit inside Ω_λ. The map μ_λ is obtained by translating the moment map μ by the $U_{\mathbb{R}}$-embedding of X in Ω_λ. On the zero section of T^*X the map μ_λ reduces to the $U_{\mathbb{R}}$-embedding of X in Ω_λ. The twisted moment map is $U_{\mathbb{R}}$-equivariant and only real algebraic, not complex algebraic.

Let us consider the complex vector space spanned by the Lie algebra characters of all the representations associated to the parameter λ. This is the space of invariant eigendistributions (associated to the parameter λ, which we have assumed to be regular). Rossmann shows that any invariant eigendistribution on $\mathfrak{g}_{\mathbb{R}}$ can be uniquely written in the following form. We set

$$T_{G_{\mathbb{R}}}^*X \;=\; \bigcup_{S \text{ a } G_{\mathbb{R}}\text{-orbit}} T_S^*X \;\subset\; T^*X\,.$$

Here T_S^*X denotes the conormal bundle of the orbit S in X; by definition T_S^*X is a subspace of T^*X. If we let n denote the complex dimension of X, then the space $T_{G_{\mathbb{R}}}^*X$ has real dimension $2n$. Let us denote by $\mathrm{H}_{2n}^{inf}(T_{G_{\mathbb{R}}}^*X, \mathbb{C})$ the space of $2n$-cycles with closed (possibly infinite dimensional) support in $T_{G_{\mathbb{R}}}^*X$ with coefficients in $\mathbb{C}$. Rossmann shows that for any invariant eigendistribution θ on $\mathfrak{g}_{\mathbb{R}}$ associated to λ there exists a unique cycle $C \in \mathrm{H}_{2n}^{inf}(T_{G_{\mathbb{R}}}^*X, \mathbb{C})$ such that

$$\theta(\phi) \;=\; \frac{1}{(2\pi i)^n} \int_{\mu_\lambda(C)} \widehat{\phi}\, \sigma_\lambda^n\,.$$

Here ϕ is any smooth compactly supported function on $\mathfrak{g}_{\mathbb{R}}$ and σ_λ is the canonical complex algebraic symplectic form on Ω_λ. In other words, we can view the construction of the character as a map

$$\{G_{\mathbb{R}}\text{-representations}\} \longrightarrow \mathrm{H}^{inf}_{2n}(T^*_{G_{\mathbb{R}}}X, \mathbb{C}). \tag{2.2}$$

To understand the map (2.2) geometrically, the right hand side immediately suggests that we should parametrize the $G_{\mathbb{R}}$-representations, by $G_{\mathbb{R}}$-sheaves (rather than by K-sheaves). Then, as is shown in [SV2], the map (2.2) coincides with the characteristic cycle construction of Kashiwara

$$\mathrm{CC} : \{G_{\mathbb{R}}\text{-sheaves on } X\} \longrightarrow \mathrm{H}^{inf}_{2n}(T^*_{G_{\mathbb{R}}}X, \mathbb{Z}). \tag{2.3}$$

We discuss this construction briefly in the next section. Note that (2.3) shows, in particular, that the map (2.2) factors through $\mathrm{H}^{inf}_{2n}(T^*_{G_{\mathbb{R}}}X, \mathbb{Z})$. To summarize:

THEOREM. *The Lie algebra character of the representation associated to $G_{\mathbb{R}}$-sheaf $\mathcal{F}$ is given by*

$$\theta(\mathcal{F})(\phi) = \frac{1}{(2\pi i)^n} \int_{\mu_\lambda(\mathrm{CC}(\mathcal{F}))} \widehat{\phi}\, \sigma_\lambda^n\,, \qquad (\phi \in C_c^\infty(\mathfrak{g}_{\mathbb{R}}))\,.$$

When $\mathcal{F}$ gives rise to a discrete series representation or, more generally, to a tempered representation, our formula reduces to the original formula of Rossmann: one shows that the cycle $\mu_\lambda(\mathrm{CC}(\mathcal{F}))$ is homologous to the appropriate coadjoint orbit.

Remark. As we explained in the first section, we can, completely equivalently, parametrize representations either by K-sheaves or by $G_{\mathbb{R}}$-sheaves. The K-side seems, at least at the first sight, more appealing and simpler as it allows one to work entirely in the realm of complex algebraic geometry. However, as the theorem above shows, from the point of view of understanding characters of representations the $G_{\mathbb{R}}$-side seems indispensable.

4. The Characteristic Cycle Construction.

Let X be a real algebraic manifold of dimension n which we assume, for simplicity, to be oriented. We consider constructible sheaves on X, i.e., sheaves of $\mathbb{C}$-vector spaces with the following property: there exists a (semi-)algebraic decomposition of X such that the sheaf restricted to any constituent of the decomposition is constant of finite rank. As before we consider complexes of constructible sheaves and we should be working in the context of derived categories. Given a (complex of) constructible sheaves $\mathcal{F}$ on X, Kashiwara in [K1] shows how to associate to it a Lagrangian, $\mathbb{R}^+$-invariant cycle $\mathrm{CC}(\mathcal{F})$ in T^*X. Recall that T^*X has a canonical symplectic structure and that conormal bundles of smooth submanifolds are prototypes of Lagrangian, $\mathbb{R}^+$-invariant submanifolds of T^*X. The construction of $\mathrm{CC}(\mathcal{F})$ is Morse-theoretic. The cycle $\mathrm{CC}(\mathcal{F})$ measures how the local Euler characteristic (=the Euler characteristic of the stalks) of $\mathcal{F}$ changes

as we move to a particular direction from a point on X. From this description it is apparent that CC satisfies the following properties:

(a) $\mathrm{CC}(\mathbb{C}_X) = [X]$,
(b) CC is additive in short exact sequences,
(c) CC is locally defined on X.

Here the symbol $[X]$ stands for the zero section viewed as a cycle on T^*X with its given orientation. The index theorem of Kashiwara [K1] states that the global Euler characteristic of $\mathcal{F}$ coincides with the intersection product of the zero section $[X]$ and $\mathrm{CC}(\mathcal{F})$. By property a) above, this amounts to a generalization to sheaves of the classical index theorem of Hopf: the Euler characteristic of a compact manifold X is given by the self intersection number of the zero section in T^*X. Kashiwara's index theorem can be generalized to the relative case: for a proper map $f : X \to Y$ and a sheaf $\mathcal{F}$ on X we can describe the characteristic cycle of the push-forward of $\mathcal{F}$ in terms of $\mathrm{CC}(\mathcal{F})$ and an intersection product [KSa].

To be able to calculate the the effect of CC under all the operations on sheaves it is necessary and sufficient to have a formula for the characteristic cycle of a pushforward under an open embedding. As this is our most important tool, we will give the statement. To this end, let $j : U \hookrightarrow X$ be an open embedding and let f be a defining equation for the boundary of U. Then, for a sheaf $\mathcal{F}$ on U, we have

$$\text{(d)} \qquad \mathrm{CC}(Rj_*\mathcal{F}) \;=\; \lim_{s\to 0^+} \left(\mathrm{CC}(\mathcal{F}) + s\frac{df}{f} \right) .$$

This formula is proved in [SV1]. It is modeled after a similar formula proved by Ginzburg in the complex analytic case. The properties (a)-(d) completely determine the operation CC, i.e., they could be taken as axioms. The construction CC amounts to a (weak) but very workable form of microlocalization.

5. Nilpotent invariants.

In this section, as an application of our techniques, we will identify two rather different invariants of representations. Both of these invariants involve nilpotent orbits. Invariants that involve nilpotent orbits are particularly interesting because, as was explained in §1, it is generally believed that unitary representations are best parametrized using such data. One of the invariants, due to Vogan, is purely algebraic and the other, due to Barbasch-Vogan [BV], is analytic. The statement that these invariants coincide has become known as the Barbasch-Vogan conjecture.

Let us consider an irreducible representation V of $G_{\mathbb{R}}$. The analytic invariant is defined as follows. Let θ denote the Lie algebra character of the representation V. Take the Fourier transform of the leading term of the asymptotic expansion of θ at the origin. Barbasch and Vogan show that this Fourier transform is a $\mathbb{C}$-linear combination of canonical measures on nilpotent coadjoint orbits in $i\mathfrak{g}_{\mathbb{R}}{}^*$. In other words, this Fourier transform can be written as

$$\mathrm{WF}(V) \;=\; \sum a_j [\mathcal{O}_j^{\mathbb{R}}] \,,$$

where the $\mathcal{O}_j^{\mathbb{R}}$ are $G_{\mathbb{R}}$-orbits in $i\mathfrak{g}_{\mathbb{R}}^* \cap \mathcal{N}^*$ and $a_j \in \mathbb{C}$. Recall that $\mathcal{N}^*$ denotes the "nilpotent cone" in $\mathfrak{g}^* \cong \mathfrak{g}$. The cycle $\mathrm{WF}(V)$ is called the wave front cycle of V.

The algebraic invariant is defined via the Harish-Chandra module M of V. We choose a K-invariant good filtration M_j of M with respect to the canonical filtration of the universal enveloping algebra $U(\mathfrak{g})$. The associated graded $\mathrm{gr}(M)$ is a module over the symmetric algebra $S(\mathfrak{g})$. As such, it determines a well defined algebraic cycle on $\mathfrak{g}^*$. The support of this cycle coincides with the support of the module $\mathrm{gr}(M)$. Vogan [V] shows that the algebraic cycle is K-invariant and is supported on $\mathfrak{p}^* \cap \mathcal{N}^*$. The space $\mathfrak{p}$ is given by the Cartan decomposition $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$. Hence we have associated to V a cycle

$$\mathrm{Ass}(V) \;=\; \sum b_j [\mathcal{O}_j^K]\,,$$

where the $\mathcal{O}_j^K$ stand for K-orbits in $\mathfrak{p}^* \cap \mathcal{N}^*$ and b_j are non-negative integers. The cycle $\mathrm{Ass}(V)$ is called the associated cycle of V. In [Se] Sekiguchi constructs a bijection between $G_{\mathbb{R}}$-orbits on $i\mathfrak{g}_{\mathbb{R}}^* \cap \mathcal{N}^*$ and K-orbits on $\mathfrak{p} \cap \mathcal{N}^*$. The following result is proved in [SV3]:

THEOREM. *The wave front cycle and the associated cycle coincide under the Kostant-Sekiguchi correspondence. In particular, the constants a_j are non-negative integers.*

Let us briefly discuss the general structure of the argument. It can by summarized in the form of the following commutative diagram:

$$\begin{array}{ccc}
\{G_{\mathbb{R}}\text{-representations}\} & \longrightarrow & \{\text{H-C-modules}\} \\
\wr\big\downarrow & & \big\downarrow\wr \\
\{G_{\mathbb{R}}\text{-equivariant sheaves on } X\} & \xrightarrow{\Gamma} & \{K\text{-equivariant sheaves on } X\} \\
\mathrm{CC}\big\downarrow & & \big\downarrow\mathrm{CC} \\
\{\text{Lagrangian cycles on } T^*_{G_{\mathbb{R}}}X\} & \xrightarrow{\Psi} & \{\text{Lagrangian cycles on } T^*_K X\} \\
\mu_*\big\downarrow & & \big\downarrow\mu_* \\
\{G_{\mathbb{R}}\text{-orbits in } \mathcal{N}^* \cap i\mathfrak{g}_{\mathbb{R}}^*\} & \xrightarrow{\psi} & \{K\text{-orbits in } \mathcal{N}^* \cap \mathfrak{p}^*\}\,.
\end{array} \tag{5.1}$$

The vertical arrows from the top to bottom can be identified with the wave front cycle and the associated cycle constructions, respectively. The crux of the argument is the explicit computation of the map Ψ induced by Γ. This computation takes us outside of the realm of semi-algebraic and subanalytic sets. We make essential use of the geometric categories of [DM]. In particular, we work in the context of the geometric category associated to the o-minimal structure $\mathbb{R}_{\mathrm{an,exp}}$. The last step is the identification of the map ψ, induced by Ψ, with the Kostant-Sekiguchi correspondence.

The fact that the vertical arrows in (5.1) amount to the invariants WF and Ass shows that they can be extracted from the appropriate characteristic cycles. Let us phrase this more precisely. Consider the diagram $X \xleftarrow{\pi} T^*X \xrightarrow{\mu} \mathcal{N}^*$ of spaces where π is the projection and $\mu : T^*X \to \mathcal{N}^*$ is the moment map. If $\mathcal{F}$

is a $G_{\mathbb{R}}$-equivariant sheaf on X then the corresponding wave front cycle is given by "microlocalizing" $\mathcal{F}$ via the CC construction to a cycle on T^*X and then integrating this cycle[4] over the fibers of μ. The analogous process on the K-side produces the associated cycle (this fact is due to J.-T. Chang). The characteristic cycles carry much more information than the wave front cycle and the associated cycle and it is conceivable that some of this extra information is crucial in understanding the unitary representations attached to nilpotent orbits. Furthermore, the construction CC is a bit too crude at least in one respect. The diagram (5.1) should be at least extended so that the objects in the last two rows are $G_{\mathbb{R}}$ and K-equivariant, respectively.

References

[BV] D. Barbasch and D. Vogan, *The Local Structure of Characters*, Jour. Func. Anal. 37 (1980), 27–55.

[B] A. Beilinson, *Localization of representations of reductive Lie algebras.*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Warsaw, 1983), PWN, Warsaw, 1984, pp. 699–710.

[DM] L. van den Dries and C. Miller, *Geometric categories and o-minimal structures*, Duke Math. Jour. 84 (1996), 497 – 540.

[K1] M. Kashiwara, *Index theorem for constructible sheaves*, Astérisque 130 (1985), 193 – 209.

[K2] M. Kashiwara, *Open problems in group representation theory*, Proceedings of Taniguchi Symposium held in 1996, RIMS preprint 569, Kyoto University, 1987.

[KSa] M. Kashiwara and P. Schapira, *Sheaves on manifolds*, Springer, 1990.

[KSd] M. Kashiwara and W. Schmid, *Quasi-equivariant $\mathcal{D}$-modules, equivariant derived category, and representations of reductive Lie groups*, Lie Theory and Geometry, in Honor of Bertram Kostant, Progress in Mathematics, Birkhäuser, 1994, pp. 457–488.

[MUV] I. Mirković, T. Uzawa, and K. Vilonen, *Matsuki Correspondence for Sheaves*, Inventiones Math. 109 (1992), 231–245.

[R1] W. Rossmann, *Kirillov's Character Formula for Reductive Lie Groups*, Inventiones Math. 48 (1978), 207–220.

[R2] W. Rossmann, *Invariant Eigendistributions on a Semisimple Lie Algebra and Homology Classes on the Conormal Variety I, II*, Jour. Func. Anal. 96 (1991), 130–193.

[SV1] W. Schmid and K. Vilonen, *Characteristic cycles of constructible sheaves*, Inventiones Math. 124 (1996), 451–502.

[4]Before integrating the cycle, we multiply it with the cohomology class $e^{\lambda+\rho}$

[SV2] W. Schmid and K. Vilonen, *Two geometric character formulas for reductive Lie groups*, to appear in Jour. AMS.

[SV3] W. Schmid and K. Vilonen, *Characteristic cycles and wave front cycles of representations of reductive Lie groups*, preprint.

[SE] J. Sekiguchi, *Remarks on nilpotent orbits of a symmetric pair*, Jour. Math. Soc. Japan 39 (1987), 127–138.

[V] D. Vogan, *Gelfand-Kirillov dimension for Harish-Chandra modules*, Inventiones Math. 48 (1978), 75–98.

Kari Vilonen
Department of Mathematics
Brandeis University
Waltham, MA 02254
USA
vilonen@math.brandeis.edu

Representation Theory of Affine Superalgebras at the Critical Level

Dedicated to the memory of my father Shoji Wakimoto

Minoru Wakimoto

1991 Mathematics Subject Classification: Primary 17B65; Secondary 17B67, 81R10.
Keywords and Phrases: Superalgebra, character, atypical representation, free field construction, mock theta function, modular transformation

0. Introduction.

As is known well, the representation theory of affine Lie algebras has a number of important connection with various area of mathematics and physics, while the representation of superalgebras still remains mysterious, and only a few has been known about it because of serious technical and intrinsic difficulties arising in its analysis.

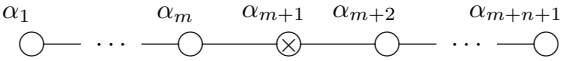
The representation theory of superalgebras, however, seems quite interesting and fascinating, since the associated Macdonald identities are of different kinds from those of the usual affine Lie algebras, and also the Ramanujan's famous mock theta functions take place closely related to the denominator identities for certain affine superalgebras.

The effective theory for superalgebras are not well established, and some of the results exposed in this note are obtained with the help of a computer and Reduce 3.6.

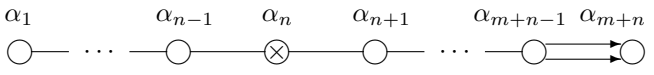
1. Characters of affine superalgebras.

Our interest in this note is a finite-dimensional or affine simple Lie superalgebra, with a non-degenerate even super-invariant super-symmetric bilinear form $(\ |\ )$. The finite-dimensional ones other than Lie algebras are listed in the following table:

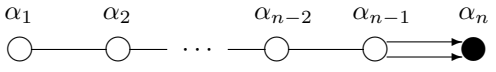
$A(m,n)$ $(m,n \geq 0)$ $(m+n \geq 1)$:



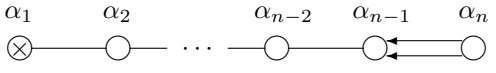
$B(m,n)$ $(m,n \geq 1)$:



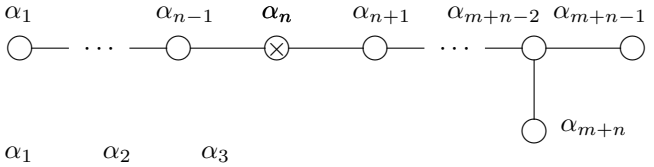
$B(0,n)$ $(n \geq 1)$:



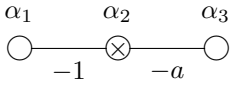
$C(n)$ $(n \geq 3)$:



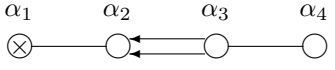
$D(m,n)$ $(m \geq 2)$ $(n \geq 1)$:



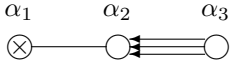
$D(2,1;a)$ $(-1, 0 \neq a \in \mathbb{C})$:



$F(4)$:



$G(3)$:



We note that, for a superalgebra with isotropic odd roots, its Dynkin diagram is not determined uniquely, and the diagrams in the above list are *standard* ones. But *non-standard* ones are never less important, and a suitable choice of Dynkin diagrams has sometimes a crucial importance in the representation theory. For example, Dynkin diagrams in Fig 1.1 $\sim$ 1.3 are useful diagrams of $A(1,0) = \mathfrak{sl}(2,1)$, $A(1,1) = \mathfrak{sl}(2,2)/\mathfrak{z}$ and $A(2,2) = \mathfrak{sl}(3,3)/\mathfrak{z}$ respectively (here and henceforth $\mathfrak{z}$ stands for the center of the superalgebra in question) with their affinizations shown in Fig 1.1′ $\sim$ 1.3′ :

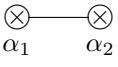


Fig.1.1

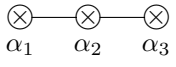


Fig.1.2

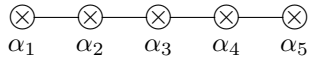


Fig.1.3

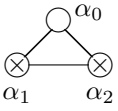


Fig.1.1′

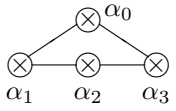


Fig.1.2′

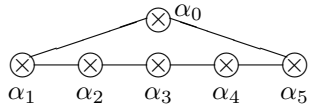


Fig.1.3′

Let $\mathfrak{g}$ be a finite-dimensional or affine Lie superalgebra and $\mathfrak{h}$ its Cartan subalgebra. The inner product $(\ |\ )$ is normalized so that the dual Coxeter number $h^\vee$ is a non-negative rational number and the square length of the longest roots is equal to 2 or -2. For usual notations and terminologies of superalgebras, we refer to [5] and [9]. In particular the concept of *integrable* weight is never obvious for superalgebras and given in Chapter 6 of [9].

The character of a *suitable* irreducible integrable highest weight $\mathfrak{g}$-module

$L(\Lambda)$ is given in [9]:

$$\mathrm{ch}L(\Lambda) = \frac{c_\Lambda}{e^\rho R} \sum_{w \in W^\sharp} \varepsilon(w) w \left(\frac{e^{\Lambda+\rho}}{\prod_{i=1}^k (1+e^{-\beta_i})} \right) \tag{1}$$

with a rational number c_Λ, where $\{\beta_1, \cdots, \beta_k\}$ is a maximal set of mutually orthogonal isotropic positive odd roots satisfying $(\Lambda+\rho|\beta_i) = 0$ for all i, and R is the denominator:

$$R := \frac{\prod_{\alpha \in \Delta_{0+}} (1-e^{-\alpha})^{\mathrm{mult}(\alpha)}}{\prod_{\alpha \in \Delta_{1+}} (1+e^{-\alpha})^{\mathrm{mult}(\alpha)}} \tag{2}$$

In the above, $W^\sharp$ is not always the *full* Weyl group W but its subgroup (cf. [9]), and Δ_+ stands, as usual, for the set of all positive roots, and the suffix "0" or "1" implies the set of "even roots" or "odd roots" respectively.

The highest weight Λ is called *typical* when $k = 0$, and *k-atypical* or simply *atypical* when $k > 0$. The number k is the *atypicality* of a highest weight Λ.

A particular feature and difficulty of the representation theory of superalgebras is that the formula (1) is not necessarily valid for *all* representations, but fails to hold in some cases. The formula (1) is true in case when (i) Λ is typical ([6]) or (ii) $\beta_1, \cdots, \beta_k$ are chosen from simple roots with a suitable choice of the Dynkin diagram ([9]).

The trivial representation $L(0)$ for an affine superalgebra with non-zero dual Coxeter number satisfies the condition (ii) above, and the formula (1) applied to this, called the denominator identity, provides a Lie superalgebraic interpretation to some of number theoretic formulas as is discussed in [9].

For affine superalgebras with $h^\vee = 0$, whose complete list is $\widehat{A}(n,n)$(=the affinization of $\mathfrak{sl}(n+1|n+1)/\mathfrak{z}$) and $\widehat{D}(n+1,n)$(= $\widehat{\mathfrak{osp}}(2n+2,2n)$) and $\widehat{D}(2,1;a)$, the trivial representation is a representation of critical level. The simplest among them is $\widehat{A}(1,1)$ with the Dynkin diagram Fig 1.2′. In this case, we claim the following denominator identity, which is obtained from the analysis of the super Boson-Fermion correspondence ([10]) :

$$R = \sum_{w \in \langle r_{\alpha_1+\alpha_2},\, t_{\alpha_1+\alpha_2} \rangle} \varepsilon(w) w \left(\frac{1}{(1+e^{-\alpha_0}) \prod_{n=1}^\infty (1+e^{\alpha_2-n\delta})(1+e^{-\alpha_2-(n-1)\delta})} \right), \tag{3}$$

where r_α is the reflection with respect to α, and t_α is the linear transformation on $\mathfrak{h}$ defined by

$$t_\alpha(\lambda) := \lambda + (\lambda|\delta)\alpha - \left(\frac{|\alpha|^2}{2}(\lambda|\delta) + (\lambda|\alpha) \right) \delta.$$

We note the following. For $\widehat{A}(1,1)$ with the Dynkin diagram Fig 1.2′, one can choose

$$\{\alpha_0 + n\delta,\ \ \alpha_2 + n\delta\}_{n \geq 0} \cup \{-\alpha_0 + n\delta,\ \ -\alpha_2 + n\delta\}_{n>0} \tag{4}$$

as a maximal system of mutually orthogonal isotropic positive odd roots which are orthogonal to $\Lambda+\rho$ since $\Lambda=\rho=0$, where $\delta:=\sum_{i=0}^{3}\alpha_i$ is the canonical imaginary root as usual. So $\Lambda=0$ is "infinitely atypical" in this case. A very remarkable fact is that not all of β_i's in this maximal system (4) does take a part in the denominator identity (3).

The formula becomes more complicated in higher rank cases. For example, the denominator identity of $\widehat{A}(2,2)$ is given as follows with respect to the Dynkin diagram Fig 1.3′ :

$$\begin{aligned} & R\cdot\prod_{n=1}^{\infty}\frac{(1-e^{\alpha_1+\alpha_3+\alpha_5-2n\delta})(1-e^{-\alpha_1-\alpha_3-\alpha_5-2(n-1)\delta})}{(1-e^{-n\delta})(1-e^{-(2n-1)\delta})} \\ = & \sum_{w\in W^{\sharp}}\varepsilon(w)w\left(\frac{1}{(1+e^{-\alpha_0})\prod\limits_{i=2,4}\prod\limits_{n=1}^{\infty}(1+e^{\alpha_i-n\delta})(1+e^{-\alpha_i-(n-1)\delta})}\right) \end{aligned} \tag{5}$$

where $W^{\sharp}:=\langle r_{\alpha_1+\alpha_2},\, r_{\alpha_3+\alpha_4},\, t_{\alpha_1+\alpha_2},\, t_{\alpha_3+\alpha_4}\rangle$.

We now look at other simplest examples of representations at the critical level, namely *integrable* $\widehat{\mathfrak{sl}}(2,1)$-modules of level -1. For a linear form $\Lambda:=k\Lambda_0-(k+1)\Lambda_1$ $(k\in\mathbb{Z})$ which is atypical with respect to α_2, we claim the following:

$$\begin{aligned} \mathrm{ch}L(\Lambda) \;=\; & \frac{\prod\limits_{n=1}^{\infty}(1-e^{-n\delta})}{e^{\rho}R}\times \\ & \begin{cases} \displaystyle\sum_{w\in\langle r_{\alpha_0}\rangle}\sum_{j=0}^{\infty}\varepsilon(w)wt_{j\alpha_0}\left(\frac{e^{\Lambda+\rho}}{\prod\limits_{n=1}^{\infty}(1+e^{\alpha_2-n\delta})(1+e^{-\alpha_2-(n-1)\delta})}\right) & (\text{if } k\geq 0) \\ \displaystyle\sum_{w\in\langle r_{\alpha_1+\alpha_2}\rangle}\sum_{j=0}^{\infty}\varepsilon(w)wt_{j(\alpha_1+\alpha_2)}\left(\frac{e^{\Lambda+\rho}}{\prod\limits_{n=1}^{\infty}(1+e^{\alpha_2-n\delta})(1+e^{-\alpha_2-(n-1)\delta})}\right) & (\text{if } k<0). \end{cases} \end{aligned} \tag{6}$$

Note that the sum in the right-hand side is *not* taken over a subgroup of the affine Weyl group. From this formula, $\mathrm{ch}L(-\rho)$ is obtained as follows:

$$\begin{aligned} & \mathrm{ch}L(-\rho) \\ = \; & \frac{\prod\limits_{n=1}^{\infty}(1-e^{-n\delta})}{e^{\rho}R}\left\{\sum_{k=1,2}\frac{\sum\limits_{j=0}^{\infty}e^{j\alpha_k-\frac{j(j+1)}{2}\delta}}{\prod\limits_{n=1}^{\infty}(1+e^{\alpha_k-n\delta})(1+e^{-\alpha_k-(n-1)\delta})}-\prod_{n=1}^{\infty}(1-e^{-n\delta})\right\} \\ = \; & \frac{\prod\limits_{n=1}^{\infty}(1-e^{-n\delta})}{e^{\rho}R}\times \\ & \left\{\frac{\sum\limits_{j\geq 0}e^{j\alpha_1-\frac{j(j+1)}{2}\delta}}{\prod\limits_{n=1}^{\infty}(1+e^{\alpha_1-n\delta})(1+e^{-\alpha_1-(n-1)\delta})}-\frac{\sum\limits_{j<0}e^{j\alpha_2-\frac{j(j+1)}{2}\delta}}{\prod\limits_{n=1}^{\infty}(1+e^{\alpha_2-n\delta})(1+e^{-\alpha_2-(n-1)\delta})}\right\}. \end{aligned}$$

The second equality in the above is due to the Jacobi triple product identity

$$\prod_{n=1}^{\infty}(1-e^{-n\delta}) = \frac{\sum\limits_{j\in\mathbb{Z}} e^{j\alpha}e^{-\frac{j(j+1)}{2}\delta}}{\prod\limits_{n=1}^{\infty}(1+e^{\alpha-n\delta})(1+e^{-\alpha-(n-1)\delta})}. \tag{7}$$

2. Free field construction of atypical $\widehat{\mathfrak{sl}}(2,1)$-modules at the critical level.

It is known in [2], [3] and [7] that, in case of affine algebras, irreducible highest weight representations and their characters display a remarkable behavior at the critical level. This may also be expected for superalgebras. To get an information on the structure of these representations, we give an explicit construction of atypical highest weight modules at the critical level for the simplest affine superalgebra $\widehat{\mathfrak{sl}}(2,1)$.

Let $\{h_i,\ e_i,\ f_i\}_{i=1,2}$ be a system of Chevalley generators of $\mathfrak{sl}(2,1)$ with respect to its Dynkin diagram Fig 1.1. Then these elements together with $e_{12} := -[e_1,\ e_2]$ and $f_{12} := [f_1,\ f_2]$, satisfying $[e_{12},\ f_{12}] = h_1+h_2$, form a basis of $\mathfrak{sl}(2,1)$. Choosing the super-invariant super-symmetric bilinear form $(\ |\ )$ such that $(e_i|f_i) = 1$ $(i = 1, 2)$, we consider its affinization

$$\widehat{\mathfrak{sl}}(2,1) := \mathfrak{sl}(2,1)\otimes\mathbb{C}[t,t^{-1}]\oplus\mathbb{C}\cdot K\oplus\mathbb{C}\cdot t\frac{\partial}{\partial t}$$

with bracket

$$[X\otimes t^m,\ Y\otimes t^n] := [X,\ Y]\otimes t^{m+n} + m(X|Y)K\delta_{m+n,0},$$

which is written as follows

$$X(z)Y(w) = \frac{[X,\ Y](w)}{z-w} + \frac{(X|Y)K}{(z-w)^2},$$

in terms of operator products of fields $X(z) := \sum\limits_{n\in\mathbb{Z}} X\otimes t^n\cdot z^{-n-1}$.

Let $a_j,\ a_j^*,\ b_j,\ \psi_j,\ \psi_j^*$ $(j\in\mathbb{Z})$ be linear operators on a linear space

$$V := \mathbb{C}[x_n;\ n\in\mathbb{Z}]\otimes\mathbb{C}[y_n;\ n\in\mathbb{Z}_{>0}]\otimes\wedge[\xi_n;\ n\in\mathbb{Z}],$$

defined by

$$a_j := \begin{cases}\frac{\partial}{\partial x_j} & \text{if } j\geq 0\\ x_j & \text{if } j<0,\end{cases} \qquad a_j^* := \begin{cases}-x_j & \text{if } j\geq 0\\ \frac{\partial}{\partial x_j} & \text{if } j<0,\end{cases}$$

$$b_j := \begin{cases}\frac{\partial}{\partial y_j} & \text{if } j>0\\ -jy_{-j} & \text{if } j<0,\end{cases}$$

$$\psi_j := \begin{cases} \frac{\partial}{\partial \xi_j} & \text{if } j \geq 0 \\ \xi_j \wedge & \text{if } j < 0, \end{cases} \qquad \psi_j^* := \begin{cases} \xi_j \wedge & \text{if } j \geq 0 \\ \frac{\partial}{\partial \xi_j} & \text{if } j < 0, \end{cases}$$

b_0 being the scalar operator with a complex number b_0. As usual, we introduce the following fields :

$$a(z) := \sum_{j \in \mathbb{Z}} a_j z^{-j-1}, \qquad a^*(z) := \sum_{j \in \mathbb{Z}} a_j^* z^j, \qquad b(z) := \sum_{j \in \mathbb{Z}} b_j z^{-j-1},$$
$$\psi(z) := \sum_{j \in \mathbb{Z}} \psi_j z^{-j-1}, \qquad \psi^*(z) := \sum_{j \in \mathbb{Z}} \psi_j^* z^j.$$

PROPOSITION 2.1. *The space V is an $\widehat{\mathfrak{sl}}(2,1)$-module by the following action:*

$$\begin{aligned}
h_1(z) &:= -:a(z)a^*(z):+:\psi(z)\psi^*(z): \\
h_2(z) &:= -:a(z)a^*(z):-b(z) \\
e_1(z) &:= a(z)\psi^*(z) \\
e_2(z) &:= -\psi(z) \\
e_{12}(z) &:= a(z) \\
f_1(z) &:= -a^*(z)\psi(z) \\
f_2(z) &:= \partial\psi^*(z)+:a(z)a^*(z)\psi^*(z):+b(z)\psi^*(z) \\
f_{12}(z) &:= -\partial a^*(z)-:a(z)a^*(z)a^*(z):+:a^*(z)\psi(z)\psi^*(z):-a^*(z)b(z).
\end{aligned}$$

And the constant function 1 *is a highest weight vector in V of weight* $(b_0-1)\Lambda_0 - b_0\Lambda_2$.

In this case, the weights of variables are given by

$$\begin{aligned}
\mathrm{wt}(x_j) &= \begin{cases} -j\delta - (\alpha_1+\alpha_2) & \text{if } j \geq 0 \\ j\delta + (\alpha_1+\alpha_2) & \text{if } j < 0, \end{cases} \\
\mathrm{wt}(y_j) &= -j\delta, \\
\mathrm{wt}(\xi_j) &= \begin{cases} -j\delta - \alpha_2 & \text{if } j \geq 0 \\ j\delta + \alpha_2 & \text{if } j < 0, \end{cases}
\end{aligned}$$

and so, by counting the character of V, one obtains the following:

COROLLARY 2.1. *Let $\mathfrak{g} = \widehat{\mathfrak{sl}}(2,1)$ with Dynkin diagram Fig* 1.1′, *and $\Lambda \in \mathfrak{h}^*$ be a linear form of level -1, atypical with respect to α_i $(i = 1 \text{ or } 2)$. Then*

$$\begin{aligned}
\mathrm{ch}L(\Lambda) &\leq \frac{e^\Lambda}{R} \prod_{n=1}^{\infty} \frac{(1-e^{-n\delta})}{(1+e^{-n\delta+\alpha_i})(1+e^{-(n-1)\delta-\alpha_i})} \\
&\leq \frac{e^\Lambda}{R} \prod_{n=1}^{\infty} (1-e^{-n\delta})^2.
\end{aligned}$$

The second inequality in the above is due to (7).

3. Mock theta identities and the associated modular functions.

The formula (1) applied to the trivial representation gives rise to some kinds of identities of Lambert series. The simplest and most remarkable ones among them are mock theta identities obtained from the denominator identities of affine superalgebras $\widehat{\mathfrak{sl}}(2,1)$ and $\widehat{B}(1,1)$ (cf. [9]). The mock theta function associated to $\widehat{\mathfrak{sl}}(2,1)$, already appearing in the classical book [12], has a crucial importance in conformal field theory since it gives rise to the formula of the modular transformation of the characters of the minimal series representations of N=2 superconformal algebras (cf. [9]). In this section, we give an exposition of the modular transformation of another simplest mock theta functions associated to $\widehat{B}(1,1)$.

First we look at the formula (3). Putting $u := e^{-\alpha_1}, w := e^{-\alpha_2}, v := e^{-\alpha_3}$ and $q := e^{-\delta}$, this is written as follows:

$$\begin{aligned}
&\prod_{n=1}^{\infty} \frac{(1-q^n)^2(1-uwq^{n-1})(1-(uw)^{-1}q^n)(1-vwq^{n-1})(1-(vw)^{-1}q^n)}{(1+uq^{n-1})(1+u^{-1}q^n)(1+vq^{n-1})(1+v^{-1}q^n)} \\
\times\ &\prod_{n=1}^{\infty} \frac{1}{(1+wq^{n-1})(1+w^{-1}q^n)(1+uvwq^{n-1})(1+(uvw)^{-1}q^n)} \\
= \quad & \frac{1}{\prod_{n=1}^{\infty}(1+wq^{n-1})(1+w^{-1}q^n)} \cdot \sum_{k\in\mathbb{Z}} \frac{w^{-k}q^{k(k+1)/2}}{1+(uvw)^{-1}q^{k+1}} \\
- \quad & \frac{1}{\prod_{n=1}^{\infty}(1+uq^{n-1})(1+u^{-1}q^n)} \cdot \sum_{k\in\mathbb{Z}} \frac{u^{k+1}q^{k(k+1)/2}}{1+v^{-1}q^{k+1}}. \qquad (8)
\end{aligned}$$

Letting w=u, we obtain the following formula, which coincides with the mock theta identity associated to $\widehat{B}(1,1)$ (cf. [1] and [4]):

$$\begin{aligned}
&\prod_{n=1}^{\infty} \frac{(1-q^n)^2(1-u^2q^{n-1})(1-u^{-2}q^n)(1-uvq^{n-1})(1-u^{-1}v^{-1}q^n)}{(1+uq^{n-1})(1+u^{-1}q^n)(1+vq^{n-1})(1+v^{-1}q^n)(1+u^2vq^{n-1})(1+u^{-2}v^{-1}q^n)} \\
= \quad & \sum_{k\in\mathbb{Z}} \frac{u^{-k}q^{k(k+1)/2}}{1+u^{-2}v^{-1}q^{k+1}} - \sum_{k\in\mathbb{Z}} \frac{u^{k+1}q^{k(k+1)/2}}{1+v^{-1}q^{k+1}} \qquad (9) \\
= \quad & \left\{ \sum_{j,k\geq 0} - \sum_{j,k<0} \right\} (-1)^j \left(u^{-2j-k}v^{-j} - u^{1+k}v^{-j}\right) q^{(k+1)(k+2j)/2} \\
= \quad & \left\{ \sum_{\substack{m,n\geq 0 \\ \text{s.t. } m\equiv n \bmod 2}} - \sum_{\substack{m,n<0 \\ \text{s.t. } m\equiv n \bmod 2}} \right\} (-1)^{\frac{m-n}{2}} v^{\frac{m}{2}} (u^{-2}v^{-1})^{\frac{n}{2}} q^{\frac{(m+1)n}{2}}. \qquad (10)
\end{aligned}$$

We consider the theta function

$$f(\tau,z) := e^{\frac{\pi i\tau}{4}} e^{-\pi iz} \prod_{n=1}^{\infty} (1-q^n)(1-e^{2\pi iz}q^{n-1})(1-e^{-2\pi iz}q^n),$$

defined for $\tau \in \mathbb{C}_+ := \{\tau \in \mathbb{C};\ \mathrm{Im}\tau > 0\}$ and $z \in \mathbb{C}$, where $q = e^{2\pi i\tau}$ as usual. This function satisfies the modular transformation:

$$f\left(-\frac{1}{\tau}, \frac{z}{\tau}\right) = -i(-i\tau)^{\frac{1}{2}} e^{\frac{\pi i z^2}{\tau}} f(\tau, z). \tag{11}$$

We put

$$F(\tau, z_1, z_2) := \frac{\eta(\tau)^3 f(\tau, z_1 + z_2) f(\tau, \frac{z_1 - z_2}{2})}{f(\tau, z_1) f(\tau, z_2) f(\tau, \frac{z_1 + z_2}{2})}. \tag{12}$$

Then by (11), we have

$$F\left(-\frac{1}{\tau}, \frac{z_1}{\tau}, \frac{z_2}{\tau}\right) = \tau e^{\frac{\pi i z_1 z_2}{\tau}} F(\tau, z_1, z_2). \tag{13}$$

We now fix a positive even integer M, and consider the set Ω of all equivalence classes in $\mathbb{Z} \times \mathbb{Z}$ with respect to the equivalence relation

$$(j,k) \sim (j',k') \Longleftrightarrow \begin{cases} j-k,\ \ j'-k' & \in M\mathbb{Z}, \\ (j+k)-(j'+k') & \in 2M\mathbb{Z}. \end{cases}$$

For $(j,k) \in \Omega$, we put

$$\begin{aligned} G_{j,k}(\tau, z_1, z_2) &:= e^{\frac{\pi i}{M\tau}\{(z_1 + j\tau)(z_2 + k\tau) - z_1 z_2\}} F(M\tau, z_1 + 1 + j\tau, z_2 + k\tau), \\ H_{j,k}(\tau, z_1, z_2) &:= G_{j,k}(\tau, z_1 - 1, z_2), \end{aligned}$$

whose Lambert series expressions and power series expansions are easily calculated from (9) and (10).

We note the following important lemma which is deduced from (13) and the power series expansion (10):

LEMMA 3.1. *The function F satisfies the following transformation property:*

$$\begin{aligned} F\left(-\frac{M}{\tau}, \frac{z_1}{\tau}, \frac{z_2}{\tau}\right) &= \frac{\tau}{M} e^{\frac{\pi i z_1 z_2}{\tau M}} \\ \times \sum_{\substack{a,b \in \mathbb{Z}/M\mathbb{Z} \\ \text{s.t. } a \equiv b+1 \mod 2}} (-1)^a \Big\{ & e^{\frac{\pi i \tau a b}{M}} e^{\frac{\pi i (a z_1 + b z_2)}{M}} F(M\tau, z_1 + 1 + b\tau, z_2 + a\tau) \\ & + e^{\frac{\pi i \tau a (b+M)}{M}} e^{\frac{\pi i (a z_1 + (b+M) z_2)}{M}} F(M\tau, z_1 + 1 + (b+M)\tau, z_2 + a\tau) \Big\}. \end{aligned}$$

By this lemma, we obtain the modular transformation of $G_{j,k}$ and $H_{j,k}$ $(j,k \in \Omega)$ as follows:

THEOREM 3.1.

$$G_{j,k}\left(-\frac{1}{\tau}, \frac{z_1}{\tau}, \frac{z_2}{\tau}\right) = (-1)^k \frac{\tau}{M} e^{\frac{\pi i z_1 z_2}{M\tau}} \times$$

$$\begin{cases} \sum\limits_{\substack{(a,b)\in\Omega \\ \text{s.t. } a\equiv b \mod 2}} e^{-\frac{\pi i(jb+ka)}{M}} G_{a,b}(\tau,z_1,z_2) & \text{if } j\equiv k \mod 2 \\ \sum\limits_{\substack{(a,b)\in\Omega \\ \text{s.t. } a\equiv b \mod 2}} e^{-\frac{\pi i((j-1)b+ka)}{M}} H_{a,b}(\tau,z_1,z_2) & \text{if } j\equiv k+1 \mod 2, \end{cases}$$

$$H_{j,k}\left(-\frac{1}{\tau},\frac{z_1}{\tau},\frac{z_2}{\tau}\right) = (-1)^k \frac{\tau}{M} e^{\frac{\pi i z_1 z_2}{M\tau}} \times$$

$$\begin{cases} \sum\limits_{\substack{(a,b)\in\Omega \\ \text{s.t. } a\equiv b+1 \mod 2}} e^{-\frac{\pi i(jb+k(a+1))}{M}} G_{a,b}(\tau,z_1,z_2) & \text{if } j\equiv k \mod 2 \\ \sum\limits_{\substack{(a,b)\in\Omega \\ \text{s.t. } a\equiv b+1 \mod 2}} e^{-\frac{\pi i((j-1)b+k(a+1))}{M}} H_{a,b}(\tau,z_1,z_2) & \text{if } j\equiv k+1 \mod 2. \end{cases}$$

The transformation under $\tau \longrightarrow \tau+1$ is easily obtained:

THEOREM 3.2.

$$G_{j,k}(\tau+1,z_1,z_2) = (-1)^k e^{-\frac{\pi i M}{4}} \times \begin{cases} e^{\frac{\pi i j k}{M}} G_{j,k}(\tau,z_1,z_2) \\ \qquad \text{if } j\equiv k \mod 2 \\ e^{\frac{\pi i(j+1)k}{M}} H_{j,k}(\tau,z_1,z_2) \\ \qquad \text{if } j\equiv k+1 \mod 2, \end{cases}$$

$$H_{j,k}(\tau+1,z_1,z_2) = (-1)^k e^{-\frac{\pi i M}{4}} \times \begin{cases} e^{\frac{\pi i j k}{M}} H_{j,k}(\tau,z_1,z_2) \\ \qquad \text{if } j\equiv k \mod 2 \\ e^{\frac{\pi i(j-1)k}{M}} G_{j,k}(\tau,z_1,z_2) \\ \qquad \text{if } j\equiv k+1 \mod 2. \end{cases}$$

Furthermore, an interesting family of modular functions is obtained by putting

$$g_{j,k}(\tau,z) := G_{j,k}(\tau,z,-z) \qquad \text{and} \qquad h_{j,k}(\tau,z) := H_{j,k}(\tau,z,-z).$$

Since $g_{j,k} = -g_{-k,-j}$ and $h_{j,k} = -e^{-\frac{\pi i}{M}(j+k)} h_{-k,-j}$, the explicit matrices of modular transformation of these functions are written in terms of the trigonometric function "sin". They may look similar to at a glance but turn to be quite different from those of the characters of the N=2 superconformal algebra.

It is known in [9] that a specialization of the denominator identity of $\widehat{\mathfrak{sl}}(2,1)$ gives the formula

$$\prod_{n=1}^{\infty} \frac{(1-q^n)^2}{(1+zq^{n-1})(1+z^{-1}q^n)} = \sum_{n\in\mathbb{Z}} (-1)^n \frac{q^{n(n+1)/2}}{1+zq^n}, \tag{14}$$

which appeared in [11] and was rediscovered by [8] in connection with Hecke indefinite modular forms. A similar identity

$$\prod_{n=1}^{\infty} \frac{(1-q^{\frac{n}{2}})(1-q^{2n})(1-zq^{2n-1})(1-z^{-1}q^{2n-1})}{(1+zq^{n-\frac{1}{2}})(1+z^{-1}q^{n-\frac{1}{2}})} = \sum_{n\in\mathbb{Z}} (-1)^n \frac{q^{n(n+1)/4}}{1+zq^{n+\frac{1}{2}}} \tag{15}$$

is deduced from (10) by letting $u = q^{\frac{1}{4}}$ and $v = z^{-1}q^{\frac{1}{4}}$.

REFERENCES

[1] B. C. Berndt : Ramanujan's Notebooks Part III, Springer-Verlag, 1991.

[2] B. L. Feigin and E. Frenkel : Representations of affine Kac-Moody algebras and bosonization, in "Physics and Mathematics of Strings ∼ memorial volume of Vadim Knizhnik" ed. by L. Brink, D. Friedan and A. M. Polyakov, World Scientific, 1990, 271-316.

[3] B. L. Feigin and E. Frenkel : Affine Kac-Moody algebras at the critical level and Gelfand-Dikii algebras, in "Proceedings of the RIMS Research Project 1991 on Infinite Analysis" ed. by A. Tsuchiya, T. Eguchi and M. Jimbo, Advanced Series in Math. Phys. Vol. 16, World Scientific, 1992, 197-215.

[4] D. Hickerson : A proof of the mock theta conjectures, Invent. math. 94 (1988), 639-660.

[5] V. G. Kac : Lie superalgebras, Advances in Math. 26 (1977), 8-96.

[6] V. G. Kac : Contravariant form for infinite-dimensional Lie algebras and superalgebras, in Lecture Notes in Phys. 94, Springer-Verlag, 1979, 441-445.

[7] V. G. Kac and D. A. Kazhdan : Structure of representations with highest weight of infinite-dimensional Lie algebras, Advances in Math. 34 (1979), 97-108.

[8] V. G. Kac and D. Peterson : Infinite-dimensional Lie algebras, theta functions and modular forms, Advances in Math. 53 (1984), 125-264.

[9] V. G. Kac and M. Wakimoto : Integrable highest weight modules over affine superalgebras and number theory, in "Lie Theory and Geometry ∼ in honor of Bertram Kostant" ed. by J. L. Brylinski, R. Brylinski, V. G. Guillemin and V. Kac, Progress in Math. Phys. Vol. 123, Birkhäuser, 1994, 415-456.

[10] V. G. Kac and M. Wakimoto : in preparation.

[11] J. Tannery and J. Molk : Élements de la théorie des fonctions elliptiques, Paris, 1898.

[12] E. T. Whittaker and G. N. Watson : A Course of Modern Analysis, fourth edition, Cambridge Univ. Press, 1927.

Minoru Wakimoto
Graduate School of Mathematics
Kyushu University
Fukuoka, 812-8581, Japan

Section 8

Analysis

In case of several authors, Invited Speakers are marked with a *.

Analytic Aspects of Quasiconformality

Kari Astala

Abstract. We discuss recent advances in quasiconformal mappings.

1991 Mathematics Subject Classification: 30, 35
Keywords and Phrases: Quasiconformal mappings, elliptic PDE's, singular integrals.

1. Quasiconformal mappings

Quasiconformal mappings are homeomorphisms which on the infinitesimal scale preserve, up to uniform bounds, relative sizes and shapes of nearby objects. These local bounds then lead to strong global constraints. The need to understand such quantities arises in a variety of different areas of geometric analysis such as hyperbolic geometry, complex dynamics, differential equations, analysis on manifolds, Gromov hyperbolic groups and so on. Hence in many respects these mappings live naturally in geometric settings.

On the other hand, the development of basic properties of quasiconformal mappings themselves usually requires considerations that are analytic in nature. In this talk we shall discuss recent advances in understanding of the fundamentals of quasiconformal mappings. In particular, we shall see how these reflect and yield new information on other topics in analysis.

There are several possible ways to give a precise meaning to the intuitive notion of quasiconformality, i.e. that infinitesimal distortion is uniform in all directions. The most "elementary" is the *metric definition*: We say that a homeomorphism $f : D \mapsto D'$, where D, D' are domains in $\mathbf{R}^n$, is quasiconformal if there exists a constant $H < \infty$ such that

$$H_f(x) \equiv \limsup_{r \to 0} \frac{\max\{|f(x) - f(y)| : |x - y| = r\}}{\min\{|f(x) - f(z)| : |x - z| = r\}} \leq H, \quad x \in D. \tag{1}$$

According to the *analytic definition*, the homeomorphism f is quasiconformal if $f \in W^{1,n}_{loc}(D)$ and the directional derivatives satisfy

$$\max_\alpha |\partial_\alpha f(x)| \leq K \min_\alpha |\partial_\alpha f(x)| \quad \text{a.e. } x \in D \tag{2}$$

for a constant $K < \infty$. Quantifying this we speak of K-*quasiconformal* mappings if (2) holds. The equivalence of the analytic and metric definitions follows essentially from the Rademacher-Stepanoff theorem; for details in n dimensions see the work of Gehring [G1].

The essential feature of quasiconformality is that the infinitesimally bounded distortions (1), (2) give strong global constraints. This leads to the notion of quasisymmetry: a mapping $f : A \to B$, $A, B \subset \mathbf{R}^n$, is called *quasisymmetric* if

$$\frac{|f(x) - f(y)|}{|f(x) - f(z)|} \leq \eta(\frac{|x - y|}{|x - z|}) \tag{3}$$

for all points $x, y, z \in A$ and for some continuous strictly increasing function $\eta : \mathbf{R}_+ \to \mathbf{R}_+$ with $\eta(0) = 0$. It is clear that (3) implies (1); for mappings of the whole $\mathbf{R}^n$ (and in general locally) the converse is also true [G1], [TV].

A recent surprising result of Heinonen and Koskela shows that in fact the assumption (1) can be considerably weakened

THEOREM 1.1. [HK1] *Suppose $H < \infty$ and $f : \mathbf{R}^n \mapsto \mathbf{R}^n$ is a homeomorphism for which*

$$\liminf_{r \to 0} \frac{\max\{|f(x) - f(y)| : |x - y| = r\}}{\min\{|f(x) - f(z)| : |x - z| = r\}} \leq H, \quad x \in \mathbf{R}^n.$$

Then f is quasisymmetric. In particular f is quasiconformal.

The fact that one can replace *lim sup* by *lim inf* is very useful; the result has immediate applications e.g. in rigidity questions in holomorphic dynamics [PR]. Furthermore, the notions (1), (3) are well defined in all metric spaces and the argument of Theorem 1.1 is based on the notion of discrete modulus combined with considerations of a general nature. Consequently, a version of the result extends to a large family of spaces, such as the length metric spaces that for some $q > 1$ satisfy a general $(1, q)$ Poincare-inequality and posses a q-regular measure, see [HK1], [HK2], [BK].

In the Euclidean two dimensional situation, a special flavour is added by Beltrami differential equation

$$\overline{\partial} f(x) = \mu(x)\, \partial f(x), \quad a.e.\ x \in D \tag{4}$$

which in $\mathbf{R}^2$ is equivalent to the inequality (2). Here μ is the complex dilatation with $|\mu(x)| \leq \frac{K-1}{K+1} < 1$ a.e. $x \in D$. In particular, in two dimensions quasiconformal considerations interact strongly with the theory of linear elliptic PDE's.

Naturally one can consider also non-homeomorphic functions satisfying (2): We say that a function f is *K-quasiregular* if firstly $f \in W^{1,n}_{loc}(D)$ and secondly the condition (2) holds at a.e. $x \in D$. In particular, the n-integrability of the derivatives guarantees that the Jacobian determinant J_f is locally integrable.

According to the fundamental theorems of Reshetnyak [Re], all quasiregular mappings are open and discrete. In many respects quasiregular mappings form in $\mathbf{R}^n$ the natural geometric counterpart of the theory of analytic functions, c.f. [Ri1].

Note also that in dimension two each quasiregular mapping factors as a composition of an analytic function and a quasiconformal homeomorphism.

2. Regularity

One of the cornerstones in the quasiconformal theory is that from the weak assumptions (1), (2) one gains improved regularity, i.e. improved integrability properties of the derivatives. In plane this fact was shown by Bojarski [Bj] and in higher dimension by Gehring [G2].

It is natural to search here for the best possible degrees of regularity. This is particularly rewarding since such bounds will lead for instance to optimal results on metric distortion properties. It turns out that they will also have consequences on different topics outside the field.

Conversely, in a dual manner one is led to ask how much can the regularity assumption $f \in W^{1,n}_{loc}$ be weakened. For quasiconformal mappings it is in fact enough to assume $f \in W^{1,1}_{loc}$, see [LV], [IKM]. However, for the noninjective quasiregular mappings one needs certain degrees of higher integrability. To state the problem more precisely, let us call a mapping $f \in W^{1,q}_{loc}(D)$ *weakly K-quasiregular* if (2) is satisfied at *a.e.* $x \in D$. The question is then to decide how small can we take q in order to still deduce that f is (strongly) quasiregular, in particular open and discrete. Optimal bounds for the q's yield then e.g. sharp quasiregular removability results.

In the case of two dimensions one has now an essentially complete understanding of these topics. To a large degree such properties are reduced to the following recent work of Astala on the distortion of area.

Theorem 2.1. [As2] *For each $K-$quasiconformal mapping f of $\mathbf{R}^2$ fixing $0, 1$ and ∞, we have*

$$|f(E)| \leq M_K |E|^{1/K}, \quad E \subset \mathbf{R}^2, \tag{5}$$

where M_K depends only on K.

The implications to the regularity of quasiconformal mappings are then as follows.

Corollary 2.2. *If f is a $K-$quasiconformal mapping in a domain $D \subset \mathbf{R}^2$ then $f \in W^{1,p}_{loc}(D)$ for all $p < \frac{2K}{K-1}$.*

In fact, since $|\partial_\alpha f|^2 \leq K J_f$ a.e, the bound of Theorem 2.1 is equivalent to $J_f \in L^{K/(K-1)}_{weak}$. Locally we have also the reverse Hölder estimates

$$(\frac{1}{|B|}\int\limits_B J_f^p dx)^{1/p} \leq C\,(\frac{1}{|B|}\int\limits_B J_f dx), \quad p < \frac{K}{K-1}, \tag{6}$$

for the Jacobian of a quasiconformal mapping f in a domain $D \subset \mathbf{R}^2$. The constant C depends only on p, K and $\mathrm{dist}(B, \partial D)/\mathrm{diam}(B)$.

As an example, the radial mapping

$$f_0(x) = x|x|^{\frac{1}{K}-1} \tag{7}$$

is $K-$quasiconformal in $\mathbf{R}^2$ but $f_0 \notin W^{1,p_o}_{loc}$ for $p_o = \frac{2K}{K-1}$. Therefore the regularity given by Corollary 2.2 is the best possible.

As mentioned above, the optimal regularity results yield also quantitative bounds on metric distortion properties. According to Ahlfors [Ah2] and Mori [Mo] $K-$quasiconformal mappings are $1/K$-Hölder continuous. This follows from (5) and (3) when one chooses $E = B(x, |x-y|)$. More importantly, we can control the distortion of Hausdorff-dimension under quasiconformal deformations.

COROLLARY 2.3. [As2] *If f is K-quasiconformal in $\mathbf{R}^2$, then for any set $E \subset \mathbf{R}^2$*

$$\frac{1}{K}\Big(\frac{1}{dim(E)} - \frac{1}{2}\Big) \leq \frac{1}{dim(fE)} - \frac{1}{2} \leq K\Big(\frac{1}{dim(E)} - \frac{1}{2}\Big). \tag{8}$$

Moreover, for any $0 < t < 2$ and any $K \geq 1$, there are sets E with $dim(E) = t$ and K-quasiconformal mappings f such that the equality holds in the above left (or respectively, right) estimate.

Let us then consider the regularity properties of weakly quasiregular mappings. In the plane the case of weakly 1-quasiregular mappings is simple; for higher dimensions the problem is more subtle and we return to it later. In two dimensions each such mapping is a weak solution of $\overline{\partial} f = 0$ and if $f \in W^{1,1}_{loc}$ then by Weyl's lemma f is holomorphic. However, for $K > 1$ the $W^{1,1}_{loc}$-regularity is not enough [IM]. Such examples combined with Corollary 2.2 and the measurable Riemann mapping theorem give

COROLLARY 2.4. *Let $1 < K < \infty$ and $D \subset \mathbf{R}^2$. Then every weakly K-quasiregular mapping, contained in a Sobolev space $W^{1,q}_{loc}(D)$ with $\frac{2K}{K+1} < q \leq 2$, is quasiregular in D.*

For each $q < \frac{2K}{K+1}$ there are weakly K-quasiregular mappings $f \in W^{1,q}_{loc}(\mathbf{R}^2)$ which are not quasiregular.

Thus only the borderline case $q = \frac{2K}{K+1}$ remains open; it is conjectured that we obtain the strong quasiregularity also in this situation. See [AIS] where the conjecture is reduced to open questions on the Beurling transform.

By the factorization properties in $\mathbf{R}^2$, the higher integrability estimates of quasiconformal mappings are also the basis for the removability results of bounded quasiregular functions. A refinement of Corollary 2.3 gives the following counterpart of the classical Painlevé-theorem.

THEOREM 2.5. [As3] *If $E \subset \mathbf{R}^2$ has Hausdorff $\frac{2}{K+1}$-measure zero, then the set E is removable for all bounded $K-$quasiregular functions.*

Moreover, for each $t > \frac{2}{K+1}$ there are sets E of dimension $dim_H(E) = t$ not removable for some bounded $K-$quasiregular functions.

3. Elliptic equations

Quasiconformal mappings are well-known to be closely connected, in many different ways, to elliptic differential equations. In two dimensions this connection

is especially effective since the governing equations (4) are linear. Indeed, the measurable Riemann mapping theorem, providing homeomorphic solutions to all Beltrami equations (4) with $\|\mu\|_\infty < 1$, is the basis of the theory of two-dimensional quasiconformal mappings.

Similarly the results of the previous sections have consequences on elliptic equations. For instance, by results of Bers, Lavrentiev and others, the solutions to $\nabla \cdot \sigma \nabla u = 0$ can be interpreted as components of quasiregular mappings, yielding sharp smoothness and removability estimates. Furthermore, let us consider in more details another example, the nonlinear systems in $\mathbf{R}^2$. Identifying $\mathbf{R}^2$ with $\mathbf{C}$, take a measurable function $H : \mathbf{C}\, x\, \mathbf{C} \to \mathbf{C}$ such that for all z, a, b

$$H(z,0) \equiv 0 \quad \textit{and} \quad |H(z,a) - H(z,b)| \le k|a-b| \tag{9}$$

with a constant $0 \le k < 1$. Then the equation

$$\overline{\partial} w(z) = H(z, \partial w(z)) + h(z), \quad z \in D, \tag{10}$$

covers all uniformly elliptic linear first order systems for $w = u + iv$ as well as general nonlinear systems $\Phi(z, \overline{\partial} w(z), \partial w(z)) = 0$ that are elliptic in the sense of Lavrentiev; c.f. [BI1].

Assuming in (10) that $h \in L^p(\mathbf{C})$, let us study the existence and uniqueness of solutions w such that $\nabla w \in L^p(\mathbf{C})$. Here we need the Beurling transform $S : L^p(\mathbf{C}) \to L^p(\mathbf{C})$, $1 < p < \infty$,

$$(Sf)(z) = -\frac{1}{2\pi i} \int_{\mathbf{C}} \frac{f(w) dw \wedge d\overline{w}}{(z-w)^2} \tag{11}$$

which intertwines the $\overline{\partial}$ and ∂ derivatives, $S(\overline{\partial} w) = \partial w$ for $\nabla w \in L^p(\mathbf{C})$.

The recent work of Astala, Iwaniec and Saksman, which applies quasiconformal coordinate changes and the reverse Hölder inequalities (6) shows

Theorem 3.1. [AIS] *Under the assumption (9) the nonlinear singular integral operator $\mathcal{B} : L^p(\mathbf{C}) \to L^p(\mathbf{C})$, $\mathcal{B}g = g - H(\,\cdot\,, Sg)$, is invertible, and in fact a bi-Lipschitz homeomorphism on $L^p(\mathbf{C})$, whenever $1 + k < p < 1 + \frac{1}{k}$.*

These bound on p are optimal, even for smooth linear equations. For instance, for each $p \ge 1 + 1/k$ there are $h \in L^p(\mathbf{C})$ and $\mu \in C^\infty(\mathbf{C})$ with $\|\mu\|_\infty = k$ which oscillate at ∞ so that the non-homogeneous Beltrami equation $\overline{\partial} w - \mu \partial w = h$ admits no solutions with $\nabla w \in L^p(\mathbf{C})$.

Another application to elliptic equations of a completely different nature was established by Nesi [N] who proved that the quasiconformal area distortion can be used to determine the optimal bounds in certain G-closure problems.

4. holomorphic motions

A picture of planar quasiconformal mappings would not be complete without mentioning the holomorphic motions. In studying the stability phenomena in complex dynamics Mañé, Sad and Sullivan coined the following effective and elegant notion.

DEFINITION 4.1. *Let $\Delta = \{z \in \mathbf{C} : |z| < 1\}$. A function $\Phi : \Delta \times A \to \overline{\mathbf{C}}$ is called a holomorphic motion of a set $A \subset \overline{\mathbf{C}}$ if*

(i) for any fixed $z \in A$, the map $\lambda \mapsto \Phi(\lambda, z)$ is holomorphic in Δ,
(ii) for any fixed $\lambda \in \Delta$, the map $z \mapsto \Phi_\lambda(z) = \Phi(\lambda, z)$ is injective and
(iii) the mapping Φ_0 is the identity on A.

Typical examples of holomorphic motions arise in deformations of Kleinian groups and dynamical systems of rational functions. The "$\lambda-$lemmas" of Mañé, Sad and Sullivan [MSS] and Slodkowski [Sl] give them strong and unexpected rigidity properties. In fact, any holomorphic motion is of the form

$$\Phi(\lambda, z) = f^{\mu_\lambda}(z), \quad z \in A. \tag{12}$$

where f^{μ_λ} is a homeomorphic solution of the equation $\overline{\partial} f = \mu_\lambda \, \partial f$ in $\mathbf{C}$ and the coefficient $\mu = \mu_\lambda \in L^\infty$ depends holomorphically on the parameter λ.

The converse is also true, as solutions to (4) depend holomorphically on μ, c.f. (14). We see that general holomorphic motions are precisely the same as (holomorphic families of) quasiconformal mappings, one is just a different representation of the other. As an immediate application this relation note that [MSS], [Sl] and (8) give

COROLLARY 4.2. . *Given a holomorphic motion $\Phi : \Delta \times E \to \overline{\mathbf{C}}$ of a subset $E \subset \overline{\mathbf{C}}$ write $E_\lambda = \Phi_\lambda(E)$. Then*

$$\frac{1-|\lambda|}{1+|\lambda|}\Big(\frac{1}{dim_H(E)} - \frac{1}{2}\Big) \le \frac{1}{dim_H(E_\lambda)} - \frac{1}{2} \le \frac{1+|\lambda|}{1-|\lambda|}\Big(\frac{1}{dim_H(E)} - \frac{1}{2}\Big). \tag{13}$$

For some sets E and motions Φ either one of the bounds holds as an equality.

5. Singular integrals and higher dimensional regularity

Beltrami equation $\overline{\partial} f = \mu \partial f$ connects quasiconformal mappings to the singular integrals and, in particular, to the Beurling transform (11). If μ has compact support and the quasiconformal mapping f is properly normalized, then we have

$$\overline{\partial} f = (I - \mu S)^{-1}\mu, \qquad \partial f(z) = 1 + (I - S\mu)^{-1} S(\mu). \tag{14}$$

The expressions are well defined since S is an isometry on $L^2(\mathbf{C})$ and $\|\mu\|_\infty < 1$.

Consequently, quasiconformal distortion properties are equivalent to bounds on the Beurling transform. For instance, an approach to Theorem 2.1 by Eremenko and Hamilton [EH] yields the following optimal estimate: Let B be a disk in $\mathbf{R}^2$ and suppose $E \subset B$. Then

$$\int\limits_{B \setminus E} |S(\chi_E)| dx \le |E| \log(\frac{|B|}{|E|}) \tag{15}$$

The equality holds here when E is a subdisk with the same center as B. Duality gives also sharp exponential integrability for the Beurling transform of bounded functions. If $|\omega(z)| \le \chi_B(z)$ *a.e.* then $|\{z \in B : |\Re\, S\omega(z)| > t\}| \le Ce^{-t}$.

However, the important question of the precise value of the L^p-norm of the Beurling transform remains still open; the best estimate so far is due to Banuelos and Wang [BW], based on probabilistic methods. It has been conjectured that $\|S\|_p = \max\{p-1, 1/(p-1)\}$. Combined with (14) this would give a new proof the regularity results 2.2-2.5.

Recently Iwaniec and Martin [IM] achieved a breakthrough in applying the theory of singular integrals in higher dimensional quasiconformality. The approach applies and developes the work of Donaldson and Sullivan [DS] on quasiconformal structures on 4-manifolds.

The starting point here is to use the differential forms. Let $\Lambda^l = \Lambda^l(\mathbf{R}^n)$ be the l'th exterior power of $\mathbf{R}^n$. Then the Hodge star operator $* : \Lambda^l \to \Lambda^{n-l}$ with respect to the standard innerproduct of $\mathbf{R}^n$ is given by $\alpha \wedge *\beta = (\alpha, \beta)$. Let d be the exterior derivative $d : C^\infty(\Lambda^l) \to C^\infty(\Lambda^{l+1})$ on (smooth) l-forms of $\mathbf{R}^n$. Its formal adjoint d^* is given by $d^* = (-1)^{nl+n+1} * d * : C^\infty(\Lambda^l) \to C^\infty(\Lambda^{l-1})$.

Next, each linear operator A on $\mathbf{R}^n$ extends naturally to an operator $A_\# : \Lambda^l \to \Lambda^l$. In particular, this is true for the (formal) derivative $Df(x)$ at *a.e.* $x \in \mathbf{R}^n$ of a weakly quasiregular mapping f. If $G_f(x) = Df(x)^t Df(x) J_f(x)^{-n/2}$ is the dilatation matrix of f at x, linear algebraic considerations show that

$$\big(G_f(x)\big)_\# * Df(x)^t_\# = J_f(x)^{(2l-n)/n} Df(x)^t_\# * . \tag{16}$$

Furthermore, [IM] proves that if $\alpha \in C^\infty(\Lambda^{l-1})$ has linear coefficients and $f \in W^{1,lp}_{loc}$, $p \geq 1$, then as distributions

$$d(f^*\alpha) = f^*(d\alpha). \tag{17}$$

As a first consequence let us see how this machinery can be applied to the regularity theory in even dimensions. For weakly 1-quasiregular mappings Iwaniec and Martin prove the following precise form of the Liouville theorem.

THEOREM 5.1. [IM] *Suppose $n > 2$ is even. Let $f \in W^{1,n/2}_{loc}(D)$, $D \subset \mathbf{R}^n$, be weakly 1-quasiregular. Then f is the restriction of a Möbius transformation.*

Moreover, for all $p < n/2$ there are non-continuous weakly 1-quasiregular mappings in $f \in W^{1,p}_{loc}(\mathbf{R}^n)$.

Indeed, for f in Theorem 5.1 the matrix dilatation $G \equiv Id$. If $l = n/2$ and $\alpha \in C^\infty(\Lambda^{l-1})$ has linear coefficients, then $f^*d\alpha = Df(x)^t_\# d\alpha$ and from (16), (17) we deduce that $f^*d\alpha$ has vanishing d and d^* derivatives. The assumption $f \in W^{1,n/2}_{loc}(D)$ justifies the use of Weyl's lemma and hence as a harmonic function $f^*d\alpha$ is C^∞-smooth. It follows that the same is true for the Jacobian derivative J_f. Earlier proofs of the Liouville theorem [BI2] complete then the argument.

The connection to singular integrals comes from the Hodge theory. Denote by $L^p(\mathbf{R}^n, \Lambda^l)$ the space of l forms with p-integrable coefficients. Each such form w admits the decomposition $w = d\alpha + d^*\beta$ where $d^*\alpha = d\beta = 0$. Therefore we can define

$$S : L^p(\mathbf{R}^n, \Lambda^l) \to L^p(\mathbf{R}^n, \Lambda^l), \quad S(w) = d\alpha - d^*\beta. \tag{18}$$

It turns out that (18) defines a singular integral operator resembling in many ways the two dimensional Beurling transform, for details see [IM]. In particular, S is an isometry on L^2 and bounded on L^p, $1 < p < \infty$. In fact, if f is weakly quasiregular and $G_f(x)$ is its dilatation matrix as above, one may define also the counterpart of the complex dilatation $\mu : L^p(\mathbf{R}^n, \Lambda^l) \to L^p(\mathbf{R}^n, \Lambda^l)$ by

$$\mu_f = \frac{(G_f)_\# - Id}{(G_f)_\# + Id}$$

If α is an l-form with linear coefficients, $l = n/2$, multiply $f^*\alpha$ by a test function $\phi \in C_0^\infty(\mathbf{R}^n)$. Then for forms α such that the "conformal part" $d^+\alpha \equiv \frac{1}{2}(Id + (-i)^l *)da = 0$, one obtains [IM] a representation similar to (14),

$$d(\phi f^*\alpha) = (Id + S)(Id - \mu S)^{-1}\omega, \tag{19}$$

where one can control the L^p-properties of ω. In consequence, a following estimate of Caccioppoli type is obtained; crucial here is that the integrability exponent r can be also be below n.

THEOREM 5.1. [IM] *Suppose n is even and $D \subset \mathbf{R}^n$. Then there are exponents $p_0 < n < p_1$, both depending only on n and K, such that if $f \in W^{1,p}_{loc}(D)$ is weakly K-quasiregular with $p_0 < p < p_1$, then*

$$\int_D |\phi Df|^p \le C(n, K) \int_D |f|^p |\nabla \phi|^p \tag{20}$$

for all test functions $\phi \in C_0^\infty(D)$

In fact, (20) follows for those p's for which $\|\mu\|_\infty \|S\|_{L^{2p/n}(\Lambda^l)} < 1$.

Essential in the above argument is that for $l = n/2$ the matrix dilatation operates linearly on $Df_\#$, c.f. (16). Hence for odd dimensions one necessarily needs nonlinear arguments. Iwaniec [I1] resolved the problem with the help of a nonlinear Hodge theory. In a subsequent work [I2] he obtained the following beautiful refinement.

THEOREM 5.3. [I2] *For each $n \ge 2$ there is an exponent $p_0(n) < n$ such that for all $F \in W^{1,p}(\mathbf{R}^n, \mathbf{R}^n)$ with $p > p_0(n)$ we have*

$$\left| \int_{\mathbf{R}^n} |DF|^{p-n} J_f(x) dx \right| \le \lambda_p(n) \int_{\mathbf{R}^n} |DF|^p dx \tag{21}$$

where $\lambda_p(n) < 1$. Moreover, for n even this holds for $p_0(n) = \frac{n}{2}$.

In general dimensions $n \ge 2$ we obtain then the Cacciopoli type estimates (20) for all weakly K-quasiregular mappings, for exponents p with $\lambda_p(n)K < 1$, by choosing $F = \phi f$ in (21).

As a consequence we obtain removability and regularity results for quasiregular mappings, complementing the higher integrability theorems of Gehring [G2].

COROLLARY 5.4. *Let $1 < K < \infty$ and $D \subset \mathbf{R}^n$. Then there is a number $q_1 < n$ such that every weakly K-quasiregular mapping, contained in a Sobolev space $W^{1,q}_{loc}(D)$ with $q_1 < q$, is quasiregular in D.*

Corollary 5.5. *For all $K \geq 1$ there is a $\delta = \delta(n, K) > 0$ such that all sets $E \subset \mathbf{R}^n$ of dimension $dim(E) < \delta$ are removable for bounded $K-$quasiregular mappings.*

In the converse direction Rickman [Ri2] shows that there are Cantor sets $E \subset \mathbf{R}^3$ of arbitrarily small Hausdorff dimension that are not removable for some bounded quasiregular mappings. Very recently Bishop [Bi] extended the result to quasiconformal mappings.

In conclusion, for $n > 2$ the optimal bounds for q_1, δ in Corollaries 5.4 and 5.5 are still open. However, Iwaniec [I2] connects this with problems in nonlinear elasticity and, in particular, with convexity questions. Recall that a function of matrices $\mathcal{F} : M^{nxm} \to \mathbf{R}$ is *quasiconvex* if $\mathcal{F}(A)|D| \leq \int_D \mathcal{F}(A + D\psi)$ for all $A \in M^{nxm}$ and $\psi \in C_0^\infty(\mathbf{R}^n, \mathbf{R}^m)$. Quasiconvexity governs the lower semicontinuity of the functionals $I(u) = \int_D \mathcal{F}\big(Du(x)\big)dx$ in the appropriate Sobolev spaces and hence understanding the notion is a fundamental problem in higher dimensional calculus of variations. An explicit necessary condition is that of rank-one convexity, i.e. that $t \mapsto \mathcal{F}(A + tB)$ is convex for all rank-one matrixes B. However, Sverak [Sv] found examples showing that in general rank-one convexity is not sufficient for quasiconvexity when $n \geq 2$ and $m \geq 3$.

Developing methods towards finding the precise bounds [I2] proves that the functions

$$\mathcal{F}_p(A) = |1 - \frac{n}{p}||A|^p - |A|^{p-n} detA, \quad p > \frac{n}{2},$$

are rank-one convex in all dimensions $n \geq 2$. This gives support to the conjecture that the optimal bound in (21) is $\lambda_r(n) = |1 - \frac{n}{p}|$, in other words that $\mathcal{F}_p$ is quasiconvex at $A = 0$. If that is indeed the case, then the optimal regularity bounds of Corollaries 2.2 - 2.5 generalize to all dimensions n, i.e. 5.4, 5.5 hold with $q = \frac{nK}{K+1}$, $\delta = \frac{n}{K+1}$ and K-quasiconformal mappings have locally $p-$integrable derivatives for $p < \frac{nK}{K-1}$. Combined with arguments originally due to Burkholder [Bu] this would also prove the above mentioned conjecture for the L^p-norms of the Beurling transform.

It seems evident that further advances in quasiconformal regularity require a deeper understanding of the notion of quasiconvexity in the plane and as well as under special symmetries in higher dimensions.

References

[Ah] Ahlfors, L., *On quasiconformal mappings*, J. Analyse Math. **3** (1954), 1-58.

[AIS] Astala, K., Iwaniec, T. and Saksman, E., *Beltrami operators*, Preprint.

[As1] Astala, K., *Area distortion of quasiconformal mappings*, Acta Math. **173** (1994), 37-60.

[As2] Astala, K., *Painleve's theorem and removability properties of planar quasiregular mappings*, Preprint.

[BK] Balogh, Z. and Koskela P., *Quasiconformality, quasisymmetry and removability in Loewner spaces*, Preprint.

[BW] Banuelos, R. and Wang, G., *Sharp inequalities for martingales with applications to the Beurling-Ahlfors and Riesz transforms*, Duke Math. J **80** (1995), 575-600.

[Bi] Bishop, C., *Non-removable sets for quasiconformal and bilipschitz mappings in* $\mathbf{R}^n$, Preprint.

[Bo] Bojarski, B., *Generalized solutions of a system of differential equations of first order and elliptic type with discontinuous coefficients*, Math. Sb. **85** (1957), 451-503.

[BI1] Bojarski, B. and Iwaniec, T., *Quasiconformal mappings and non-linear elliptic equations in two variables I, II.*, Bull. Acad. Pol. Sci. **XXII** (1974), 473-484.

[BI2] Bojarski, B. and Iwaniec, T., *Another approach to Liouville theorem*, Math. Nahr. **107** (1982), 253-262.

[Bu] Burkholder, D., *Sharp inequalities for martingales and stochastic integrals*, Asterisque **157/158** (1988), 75-94.

[DS] Donaldson, S.K. and Sullivan D.P., *Quasiconformal 4-manifolds*, Acta Math **163** (1989), 181-252.

[EH] Eremenko, A. and Hamilton, D., *On the area distortion by quasiconformal mappings*, Proc. Amer. Math. Soc. **123** (1995), 2793–2797.

[G1] Gehring, F.W., *Rings and quasiconformal mappings in space*, Trans. Amer.Math.Soc **103** (1962), 353-393.

[G2] Gehring, F.W., *The L^p-integrability of the partial derivatives of a quasiconformal mapping*, Acta Math. **130** (1973), 265-277.

[HK1] Heinonen, J. and Koskela P., *Definitions of quasiconformality*, Invent. Math **120** (1995), 61-79.

[HK2] Heinonen, J. and Koskela P., *Quasiconformal maps in metric spaces with controlled geometry*, Acta Math. **To appear**.

[IM] Iwaniec, T. and Martin G., *Quasiregular mappings in even dimensions*, Acta Math. **170** (1992), 29-81.

[IKM] Iwaniec, T. Koskela, P. and Martin, G., *Mappings of BMO–bounded distortion and Beltrami type operators*, Preprint.

[I1] Iwaniec, T., *p-Harmonic tensors and quasiregular mappings*, Annals of Math **136** (1992), 589-624.

[I2] Iwaniec, T., *An approach to Cauchy-Riemann operators in* $\mathbf{R}^n$, Preprint.

[LV] Lehto O. and Virtanen K., *Quasiconformal mappings in the plane*, Second edition. Springer-Verlag, 1973.

[MSS] Mañé R., Sad P. and Sullivan D., *On the dynamics of rational maps*, Ann. Sci. École Norm. Sup. **16** (1983), 193-217.

[Mo] Mori, A., *On an absolute constant in the theory of quasiconformal mappings*, J. Math. Soc. Japan **8** (1956), 156-166.

[N] Nesi, V., *Quasiconformal mappings as a tool to study certain two-dimensional G-closure problems*, Arch. Rational Mech. Anal **134** (1996), 17-51.

[PR] Przytycki, F. and Rohde S., *Rigidity of Holomorphic Collet-Eckmann repellers*, Fund. Math. **To appear**.

[Re] Reshetnyak, Yu. G., *Space mappings with bounded distortion*, Sibirsk. Math. Zh **8** (1967), 629-658.

[Ri1] Rickman S., *Quasiregular mappings*, Springer-Verlag, 1993.

[Ri2] Rickman S., *Nonremovable Cantor sets for bounded quasiregular mappings*, Ann. Acad. Sci. Fenn. Ser. A I Math. **20** (1995), 155-165.

[Sl] Slodkowski, Z., *Holomorphic motions and polynomial hulls*, Proc. Amer. Math. Soc **111** (1991), 347-355.

[Sv] Šverák, V., *Rank-one convexity does not imply quasiconvexity*, Proc. Roy. Soc. Edinburgh Sect. A **120** (1992), 185-189.

[TV] Tukia P. and Väisälä J., *Quasisymmetric embeddings of metric spaces*, Ann. Acad. Sci. Fenn. Ser. A I Math. **5** (1980), 97-114.

Kari Astala
Department of Mathematics
University of Jyväskylä
P.O. Box 35
FIN-40351 Jyväskylä, Finland
astala@math.jyu.fi

Singularity and Regularity — Local and Global

Michael Christ

Abstract. There exists a smoothly bounded, pseudoconvex domain in $\mathbb{C}^2$ for which the Bergman projection fails to preserve the class of functions which are globally smooth up to the boundary. The counterexample is explained and placed in a wider context through a broader discussion of the local and global regularity of solutions to subelliptic and more degenerate partial differential equations in various function spaces.

1991 Mathematics Subject Classification: 32F20, 35N15, 35P30, 42B99
Keywords and Phrases: Hypoellipticity, global regularity, Bergman projection, $\bar\partial$–Neumann problem.

1 Introduction

Consider a bounded open set $\Omega \subset \mathbb{C}^n$, assumed always to have C^∞ boundary. The Bergman projection B is the orthogonal projection from $L^2(\Omega)$ (with respect to Lebesgue measure) onto the closed subspace consisting of all L^2 holomorphic functions. Our purpose is to explain and to place in a wider context the following counterexample.

Theorem 1. [8] *There exists a smoothly bounded, pseudoconvex domain $\Omega \subset \mathbb{C}^2$ for which the Bergman projection fails to preserve $C^\infty(\overline{\Omega})$.*

Barrett [1] had given a nonpseudoconvex example, but the issue is most natural for pseudoconvex domains. The first motivation was Bell and Ligocka's discovery that if $C^\infty(\overline{\Omega})$ were always preserved then any biholomorphic mapping between two (smoothly bounded) pseudoconvex domains would extend smoothly to a diffeomorphism of their closures; this in turn would have implications for the classification of domains up to biholomorphism by means of boundary invariants.[1] Secondly, it is one of many problems about the regularity of solutions of the $\bar\partial$–Neumann problem and related PDE.

This paper stresses the author's own work. Because of rigid limitations on the lengths of text and bibliography, the important contributions of many authors are slighted, including S. Baouendi, E. Bernardi, A. Bove, D. Catlin, S.-C. Chen, D. Geller, C. Goulaouic, N. Hanges, B. Helffer, A. A. Himonas, M. Derridj, V. Grušin, G. Komatsu, J. J. Kohn, G. Métivier, Pham The Lai, D. Robert, N. Sibony, D. Tartakoff, and C.-C. Yu. A more complete bibliography and discussion are in [10].

[1] The question of boundary extendibility of biholomorphic mappings remains open.

2 Some Background

Except for very symmetric domains, the best method known for analyzing the Bergman projection is by means of the $\bar\partial$–Neumann problem. This is a boundary value problem[2] $\Box u = f$ on Ω, with boundary conditions $u \lrcorner \bar\partial\rho = 0$ and $\bar\partial u \lrcorner \bar\partial\rho = 0$ on $\partial\Omega$, where u, f are $(0,1)$ forms, ρ is any defining function for Ω, $\Box = \bar\partial\bar\partial^* + \bar\partial^*\bar\partial$ and $\lrcorner$ denotes the interior product of forms.

$\Box$ is simply the Laplacian times the identity matrix, but the boundary conditions are noncoercive. In $\mathbb{C}^n$ a Dirichlet condition is imposed on one of the n components of u; on each of the other components is imposed a complex Neumann condition; however the problem does not decouple into separate scalar problems, instead there is an interaction between the good (Dirichlet) component and bad (complex Neumann) components. This interaction, and consequently the regularity of solutions, depend heavily on the complex geometry of the boundary.

For any pseudoconvex, bounded, smoothly bounded domain Ω there exists for each $f \in L^2$ a unique solution $u \in L^2$ which satisfies the boundary conditions in an appropriate sense; the bounded linear operator N mapping f to u is called the Neumann operator. The Bergman projection is related to N by Kohn's formula $B = I - \bar\partial^* N \bar\partial$. In particular, if the $\bar\partial$–Neumann problem is *globally regular* in the sense that N preserves $C^\infty(\overline{\Omega})$, then B also preserves $C^\infty(\overline{\Omega})$. In $\mathbb{C}^2$ these properties are actually equivalent; there is a less simply formulated generalization in higher dimensions.

More commonly studied is hypoellipticity. The $\bar\partial$–Neumann problem is said to be hypoelliptic (in C^∞) if for every $p \in \overline{\Omega}$ and $f \in L^2(\Omega)$ which is C^∞ near p, the solution u likewise is C^∞ near p. A partial differential operator L is said to be hypoelliptic (in C^∞) in an open set U if for any distribution, $u \in C^\infty$ in any open subset of U where $Lu \in C^\infty$. These notions can be modified by replacing C^∞ by other function classes such as C^ω, the real analytic functions, or G^s, the Gevrey classes. Hypoellipticity implies global regularity.

The issue in hypoellipticity is whether N *transports* singularities in f from one place to another, while in global regularity the issue is whether N *creates* singularities out of nothing. We will argue in §5 that this point of view, though literally correct, is misleading.

Global regularity is a very weak property. A standard example is $L = \partial_{x_1} + \alpha\partial_{x_2}$ on a two-torus, where $\alpha \in \mathbb{R}$ is constant; L is globally regular, unless α has exceptional Diophantine properties, yet is never hypoelliptic. Similarly, on any compact Lie group, convolution with any distribution preserves $C^\infty(G)$.

The $\bar\partial$–Neumann problem is hypoelliptic if Ω is strictly pseudoconvex or more generally is of finite type. The latter condition is necessary for subellipticity, but not for hypoellipticity; see for instance [13].

For $\mathbb{C}^2$, the $\bar\partial$–Neumann problem is closely related to sums of squares of (two) real vector fields in a three real dimensional space.[3] Indeed, the general method of reduction to the boundary reduces matters to an equation $\Box^+ v = g$ on $\partial\Omega$, where the pseudodifferential Calderón operator $\Box^+$, near its characteristic variety

[2] In this paper only the $\bar\partial$–Neumann problem for forms of bidegree $(0,1)$ will be discussed.
[3] In higher dimensions matters are more subtle.

$\Sigma \subset T^*\partial\Omega$ (and modulo an elliptic factor), takes the form $\bar\partial_b \circ \bar\partial_b^*$, modulo certain lower order terms which are omitted here to simplify the exposition. Here $\bar\partial_b$ is a Cauchy-Riemann operator associated to the CR structure on $\partial\Omega$; thus the complex geometry of $\partial\Omega$ enters the problem quite directly. Locally $\bar\partial_b = X + iY$ where X, Y are everywhere linearly independent, smooth real vector fields. Thus $\bar\partial_b \circ \bar\partial_b^* = -X^2 - Y^2 + i[X,Y]$ modulo relatively harmless lower order terms. Pseudoconvexity guarantees that the principal symbol of $i[X,Y]$ is nonnegative near Σ, so it does not substantially alter the character of $-X^2 - Y^2$. Henceforth we assume always that $n = 2$.

The $\bar\partial$–Neumann problem is said to be *compact* if N is a compact mapping from L^2 to L^2. It is *exactly regular* in the Sobolev space H^s if N maps $H^s(\Omega)$ to itself, and is simply said to be exactly regular if it is exactly regular in H^s for every $s \geq 0$. A simple perturbation argument shows that for any Ω there exists $\delta > 0$ such that exact regularity holds in H^s for all $0 \leq s < \delta$. Subellipticity implies compactness, which implies exact regularity, which implies global regularity in C^∞. All existing proofs of global regularity proceed by establishing exact regularity. The other two implications just stated are not reversible; nor does compactness imply hypoellipticity.

Compactness is easily shown to fail for domains in $\mathbb{C}^2$ whose boundaries contain one-dimensional complex disks. No satisfactory characterization is known; Matheos [19] has constructed Hartogs domains in $\mathbb{C}^2$ whose boundaries contain no complex disks, yet for which N is noncompact.

Global regularity can hold without compactness. It holds in the presence of sufficient symmetry, no matter how degenerate the domain. A related but deeper theorem of Boas and Straube [3] requires only an approximate symmetry: it suffices to have a smooth real vector field T on $\partial\Omega$ which is everywhere transverse to the complex tangent space, and for which $[T,X]$ and $[T,Y]$ belong everywhere to the span of X, Y (where X, Y denote the real and imaginary parts of $\bar\partial_b$ in local coordinates). Moreover a weaker approximate version of this condition still suffices [3], and is quite important.

An interesting special class of domains consists of those for which the set W of all weakly pseudoconvex boundary points is a smooth one-dimensional complex manifold with boundary. To any such domain is associated [4] a cohomology class $\alpha \in H^1(W)$, which vanishes if and only if there exists a vector field T having the required weaker version of the above commutation property. α also admits complex geometric descriptions. Consequently global regularity holds (a) whenever W is *simply connected*, and (b) (paradoxically) whenever the CR structure is *sufficiently degenerate* near W.

In the negative direction, Kiselman [18] proved that for certain nonsmooth domains with corners, both exact and global regularity fail. Barrett [2] extended the analysis to show that for the famous worm domains, exact regularity cannot hold for large s; this left open the possibility that N might map H^s to $H^{s-\varepsilon}$ for all $s \geq 0$ and $\varepsilon > 0$. Roughly speaking, he produced Kiselman's domains as limits of blowups of worm domains and used the common scaling of the two sides in the inequality $\|Nu\|_{H^s} \leq C\|u\|_{H^s}$ to pass from exact regularity for worm domains to the same for Kiselman's domains.

The worm domains were originally invented by Diederich and Fornæss [14] as examples of smoothly bounded, pseudoconvex domains whose closures lack[4]. arbitrarily small pseudoconvex neighborhoods. A worm domain $\mathcal{W} \subset \mathbb{C}^2$ takes the form

$$\mathcal{W} = \{z : |z_1 + e^{i \log |z_2|^2}|^2 < 1 - \phi(\log |z_2|^2)\} \tag{1}$$

with the following properties: (i) $\mathcal{W}$ has smooth boundary and is pseudoconvex; (ii) $\phi \in C^\infty$ takes values in $[0,1]$, vanishes identically on $[-r,r]$ for some $r > 0$, and vanishes nowhere else; and (iii) $\mathcal{W}$ is strictly pseudoconvex at every boundary point where $|\log |z_2|^2| > r$. There do exist ϕ for which these properties hold [14]. The two caps, where $|\log |z_2|| > r$, serve to make $\mathcal{W}$ be bounded. Properties of worm domains include: (iv) The set of all weakly pseudoconvex points of $\partial\mathcal{W}$ is the annular complex manifold with boundary $\mathcal{A}_r = \{z : z_1 = 0 \text{ and } |\log |z_2|^2| \le r\}$. (v) The cohomology class $\alpha \in H^1(\mathcal{A}_r)$ is nonzero. (vi) There is a one-parameter global symmetry group, $\rho_\theta(z) = (z_1, e^{i\theta} z_2)$ for $\theta \in \mathbb{R}$.

3 Comments on the proof

The proof of Theorem 1 demonstrates that global regularity fails for all worm domains; moreover N and B fail to map $C^\infty(\overline{\mathcal{W}})$ to H^s, where $s(r)$ tends to zero as $r \to \infty$. Siu [24] has given an alternative proof that there exist worm domains for which B fails to map $C^\infty(\overline{\mathcal{W}})$ to a Hölder class $\Lambda_s(\overline{\mathcal{W}})$; he obtains good control on the dependence of s on r. *Grosso modo* he shows that the two caps can be chosen so that their effects on $B(f)$ cancel for a certain f, reducing matters to Kiselman's analysis. Both proofs exploit special features of worm domains and appear quite limited in scope. Only the original proof will be discussed here.

Boundary reduction leads to a global regularity problem for a pseudodifferential equation on the real three-dimensional manifold $\partial\mathcal{W}$; the pseudodifferential operator is closely analogous to $-X^2 - Y^2$ for certain real vector fields. With respect to the symmetries ρ_θ, $L^2(\partial\mathcal{W})$ decomposes by Fourier analysis into orthogonal subspaces $\mathcal{H}_j$. The equation respects this decomposition. Fixing any such j, one may identify functions in $\mathcal{H}_j$ with functions of two real variables.

A model captures the essence of the situation. Fix an open neighborhood V of $A = [-r,r] \times \{0\} \subset \mathbb{R}^2$, with coordinates (x,t). Let $L = -X^2 - Y^2 + b(x,t)$ where $X = \partial_x$, Y is a real vector field which in the region $|x| \le r$ takes the form $[a(x)t + O(t^2)]\partial_t$ with a nowhere vanishing, and $X, Y, [X,Y]$ span the tangent space everywhere on $V \backslash A$. Suppose moreover that $\text{Re}\,\langle Lu, u\rangle \ge c\|u\|_{L^2}^2$ for all $u \in C^2$ supported in V, and likewise for the transpose of L.

The last hypothesis mimics the $L^2(\mathcal{W})$ boundedness of N. A corresponds to the set of all weakly pseudoconvex points in $\partial\mathcal{W}$; L is hypoelliptic on $V \backslash A$. The condition $a(x) \ne 0$ corresponds to the nonvanishing of $\alpha \in H^1(\mathcal{W})$; if $a(x,t) \equiv 0$ for $|x| \le r$, then L is more degenerate but paradoxically becomes globally regular, as follows from the method of [3]. Under the hypotheses stated, there exists $u \notin C^\infty(V)$ such that $Lu \in C^\infty(V)$. The proof is quite indirect; no construction of singular solutions is known to me. Of its three steps, the principal one is:

[4]Provided that the parameter r below is $\ge \pi$.

Proposition 2. *There exists a discrete set $\Sigma \subset [0,\infty)$, with $0 \notin \Sigma$, so that for every $s \notin \Sigma$, one has $\|u\|_{H^s} \leq C_s \|Lu\|_{H^s}$ for every $u \in C_0^\infty(V)$.*

The hypotheses ensure that $L^{-1} : L^2 \mapsto L^2$ is well defined and bounded. Step 2 is to show[5] that L^{-1} cannot map H_0^s to H^s for large s. Supposing the contrary, scaling $(x,t) \mapsto (x,\lambda t)$, and letting $\lambda \to \infty$ as in [2], one deduces that the limiting operator $\mathcal{L} = -\partial_x^2 - (a(x)t\partial_t)^2 + b(x,0)$ on $[-r,r]\times(0,\infty)$, with Dirichlet boundary conditions at $x = \pm r$, must be exactly regular in a (homogeneous) Sobolev space of the same order s.

Applying the Mellin transform with respect to t and conjugating by ∂_t^s, one arrives at ODEs $-\partial_x^2 - a^2(x)(s+i\tau)^2 + b(x,0)$, for $\tau \in \mathbb{R}$, with Dirichlet boundary conditions on $[-r,r]$. There must exist nonlinear eigenvalues $\sigma + i\tau$ for which there are nonzero solutions f (with $f(\pm r) = 0$); $f(x)t^{\sigma+i\tau}$ is then a solution of the two variable Dirichlet problem, and is singular at $t = 0$. Consequently L^{-1} cannot preserve H^s for $s > \sigma + \frac{1}{2}$.

Step 3 is merely to observe that if L^{-1} did map $C_0^\infty(V)$ to $C^\infty(V)$ then because L^{-1} is bounded on L^2, a density argument combined with step 1 would imply that L^{-1} maps $H_0^s(V)$ to $H^s(V)$, for all $s \notin \Sigma$, contradicting step 2.

The more intricate analysis for Step 1 divides naturally into three overlapping regions: (i) the complement of A, where L is subelliptic; (ii) the Cartesian product of $[-r,r]$ with an arbitrarily small neighborhood of 0, where the natural tool is Mellin analysis in the t coordinate and reduction to properties of the family of one dimensional Dirichlet problems described above, modulo certain error terms; and (iii) an arbitrarily narrow transitional region $r \leq |x| < r + \delta$, for which little information is available; $\|u\|_{L^2} \leq C\delta^{+1}\|\partial_x u\|_{L^2}$ for functions supported there. The final ingredient is an *a priori* inequality $\|\partial_x u\|_{H^s} \leq C_s\|Lu\|_{H^s} + C_s\|u\|_{H^s}$. This combined with the three region analysis yields the proof.

4 Other regularity problems

Henceforth we discuss the regularity of solutions of $Lu = f$ where $L = \sum_j X_j^2$ and the X_j are real vector fields in some open set or compact manifold without boundary, denoted in either case by V. Their coefficients are assumed to belong to whichever function space we are working in. (Many of the results do however have analogues for the $\bar\partial$–Neumann problem.) Regularity in C^∞, C^ω and to a lesser extent G^s, will be discussed, in both the global and local (that is, hypoellipticity) senses. There are two types of positive results for each function space $\mathcal{F}$: (a) If L is sufficiently strong then it is hypoelliptic in $\mathcal{F}$. (b) If L is arbitrarily weak but satisfies an appropriate commutation condition then it is still hypoelliptic in $\mathcal{F}$. We assume an inequality valid for all $u \in C_0^2$:

$$\int |\hat u|^2(\xi) w^2(\xi)\, d\xi \leq C \sum_j \|X_j u\|_{L^2}^2, \tag{2}$$

where $w(\xi) \to \infty$ as $|\xi| \to \infty$, suitably interpreted in the manifold case.

[5] Certain points are slurred over in the discussion for the sake of brevity.

C^∞, global. (a) The validity of (2) with some $w \to \infty$ is equivalent to compactness, which implies global regularity. A type (b) result is that of Boas and Straube [3]; here it is not required that $w \to \infty$.

C^∞, local. (a) If $w(\xi)/\log|\xi| \to \infty$ as $|\xi| \to \infty$ then L is C^∞ hypoelliptic [20]. This is sharp in general. A consequence [13] is hypoellipticity of the $\bar\partial$–Neumann problem for any domain in $\mathbb{C}^2$ for which the set of weakly pseudoconvex points is a real hypersurface $M \subset \partial\Omega$ transverse to the complex tangent space, for which the Levi form λ is $\gg \exp(-c\,\text{distance}\,(z,M)^{-1})$ for all $c > 0$. A result of type (b) is roughly as follows; for more precise statements see [17],[21] and the many references therein.

Suppose that for any ray $R \subset T^*V$ and any small conic neighborhood Γ of R there exists a scalar valued symbol $0 \le \psi \in S^0_{1,0}$ such that $\psi \equiv 0$ in some smaller conic neighborhood of R, $\psi \ge 1$ on $T^*V\backslash\Gamma$, and such that for each $\delta > 0$ there exists $C_\delta < \infty$ such that for any relatively compact open subset $U \Subset V$ and for all $u \in C^2_0(U)$ and each index i,

$$\|\operatorname{Op}\big[\log\langle\xi\rangle\{\psi,\sigma(X_i)\}\big]u\|^2 \le \delta\sum_j \|X_j u\|^2 + C_\delta\|u\|^2 \tag{3}$$

Then L is hypoelliptic, indeed microhypoelliptic, in V. Here $\operatorname{Op}(\cdot)$ denotes the pseudodifferential operator with the indicated symbol, and $\{\cdot\}$ the Poisson bracket.

C^ω, local. (a) $w(\xi) \ge c|\xi|$ is equivalent to ellipticity, which by a theorem of Petrowsky, implies analytic hypoellipticity. (b) Denote by $\Sigma \subset T^*V$ the characteristic variety of L. By assumption, Σ is conic. Assume that Σ is a manifold, and that the symbol of L vanishes to order exactly two at each point of Σ. Suppose that for each $p \in T^*V$ and each small neighborhood W of p, there exists $\psi \in C^\omega(W)$ such that $\psi(p) = 0$, $\psi > 0$ near the boundary of W, and $H_{\sigma_j}(\psi) \equiv 0$ in W, where H_{σ_j}, here and below, denotes the Hamiltonian vector field associated to the principal symbol of X_j. Then L is analytic hypoelliptic, by a theorem of Grigis and Sjöstrand[6] [16]. A closely related commutation condition appears in the work of Tartakoff.

G^s, local. (a) If (2) holds with $w(\xi) = |\xi|^{1/s}$ then L is hypoelliptic in the Gevrey class G^s by a theorem of Derridj and Zuily; this is the optimal condition on w. A type (b) result is in [17]. Two examples indicate the intricacy of the problem. (i) [9] In $\mathbb{R}^3$ with coordinates (x, y_1, y_2), the operator $\partial_x^2 + x^{2(m-1)1}\partial_{y_1}^2 + x^{2(n-1)}\partial_{y_2}^2$ is hypoelliptic in G^s if and only if $s \ge \max(n/m, m/n)$; however it satisfies (2) only with $w(\xi) \sim |\xi|^{1/\max(n,m)}$. (ii) [11] In $\mathbb{R}^2$ with coordinates (x,t), for $p \ge 1$, $\partial_x^2 + x^{2(m-1)}\partial_t^2 + x^{2(m-1-k)}t^{2p}\partial_t^2$ is hypoelliptic in G^s if[7] $s^{-1} \le 1 - \tilde p^{-1}(1 - m^{-1})$ where $\tilde p = p(m-1)/k$. The optimal w here is $\sim |\xi|^{1/m}$. In the positive direction these results were obtained independently and in greater generality by Matsuzawa, and were also proved by Bernardi, Bove and Tartakoff. The negative result for (ii) for $m = 2, k = 1, p = 1$ is due to Métivier. An intriguing conjecture of Treves [26] proposes to relate analytic hypoellipticity to the fine symplectic geometry of the

[6]The theorem is not formulated explicitly but does seem to be proved in [16].

[7]I am confident that this exponent can be proved to be optimal for many parameters m, p, k, by the method used in [7] to disprove analytic hypoellipticity, but have not verified the details.

characteristic variety Σ of L; these examples illustrate that at least for $s > 1$, G^s hypoellipticity is not controlled by Σ alone.

C^ω, global. The result and method in [7] show that there is no better result of type (a) than for local C^ω regularity. I know of no really satisfactory general result of type (b), although there are many particular results of that flavor.

We turn to results in the negative direction, concentrating on the C^ω case. The theory here is fragmentary, with a large gap between counterexamples and the results above. A common structure underlies the proofs. To L one associates a one-parameter family of simpler operators, $\mathcal{L}_z$; in all the results below, these are ordinary differential operators.[8] In simple cases, solutions to the ODE lead to solutions of $Lu = 0$, by separation of variables. One proves the existence of at least one nonlinear eigenvalue $\zeta \in \mathbb{C}$ for which $\mathcal{L}_\zeta$ has a nonzero solution f_ζ in the Schwartz class on $\mathbb{R}^1$.

Analytic hypoellipticity implies that all solutions of $Lu = f$ satisfy certain *uniform* Cauchy-type inequalities in terms of f. When separation of variables applies, scaling and f_ζ lead to a one-parameter family of solutions of L which violate any such Cauchy inequalities as $\lambda \to \infty$. For instance, for the Baouendi-Goulaouic example $\partial_x^2 + x^2\partial_t^2 + \partial_y^2$, one has solutions $u = \exp(i\lambda t + i\zeta\lambda^{1/2}y)f(\lambda^{1/2}x)$ where $-\zeta^2, f$ are a Hermite eigenvalue and corresponding eigenfunction. This method was pioneered by Oleĭnik and Radkevič [22], and developed much further, to situations where separation of variables does not apply directly, by G. Métivier.

Theorem 3 is a bit more complicated, and the proofs of Theorems 4 and 5 are even more intricate, because separation of variables does not apply directly. The latter two theorems rely on reasoning by contradiction. Assuming the Cauchy inequalities, the structure of the equation is used to deduce stronger *a priori* bounds on solutions. Exact solutions of $Lu_\lambda = f_\lambda$ for precisely chosen f_λ are then proved to be well controlled by solutions of a simpler related partial differential equation, which in turn can be analyzed by separation of variables. Eventually solutions which are supposed to be holomorphic in certain regions are proved to have poles, a contradiction. This reasoning has elements in common with the proof of global C^∞ irregularity for the worm domains.

Theorem 3. [5] *Consider $L = X^2 + Y^2$ in $\mathbb{R}^3$, where X, Y are linearly independent at each point. Suppose there exists a nonconstant curve $\gamma \subset \mathbb{R}^3$ such that at each point $p \in \gamma$, the tangent vector $\dot\gamma(p)$ is in the span of X, Y, and moreover $X, Y, [X, Y]$ fail to span the tangent space to $\mathbb{R}^3$ at p. Then L is not analytic hypoelliptic.*

This is a very special case of an older conjecture of Treves [25].

Next consider $L = X^2 + Y^2$ in an open subset of $\mathbb{R}^2$, and $\tilde L = (X + iY)(X - iY)$, where X, Y do not simultaneously vanish at any point. Assume the bracket hypothesis; for $\tilde L$ we also impose a certain natural pseudoconvexity hypothesis (see [6]). The positive parts of the following theorem are special cases of an old theorem of Grušin.

[8] Barrett has studied nonlinear eigenvalue problems for elliptic PDE on smoothly bounded Riemann surfaces, which are relevant to global regularity for the $\bar\partial$–Neumann problem.

THEOREM 4. [6] *$\tilde{L}$ is microlocally analytic hypoelliptic if and only if there exist coordinates (x,t) in which* $\operatorname{span}\{X,Y\} = \operatorname{span}\{\partial_x, x^{m-1}\partial_t\}$, *as $C^\omega(\mathbb{R}^3)$-modules, for some $m \geq 1$. For generic*[9] *pairs X, Y, L is analytic hypoelliptic if and only if the same condition holds.*

The generalization to more than two vector fields (for L) is straightforward, but matters are much subtler in $\mathbb{R}^n$ for $n > 2$.

THEOREM 5. [7] *There exists a bounded, pseudoconvex domain $\Omega \subset \mathbb{C}^2$ with C^ω boundary, for which the Szegö projection fails to preserve $C^\omega(\partial\Omega)$.*

F. Tolli has shown that, in contrast to the C^∞ case, there exists such a domain which is strictly pseudoconvex except at a single isolated point.

5 A metric in phase space

For definiteness let $L = \sum X_j^2$ be a sum of squares of vector fields, in an open subset of $\mathbb{R}^n$. Let $\sigma_j(x,\xi)$ be the principal symbol of X_j and H_{σ_j} the associated Hamiltonian vector field in $T^*\mathbb{R}^n$. Assume the bracket hypothesis of Hörmander to hold to some order $\leq m$; define the effective symbol $\tilde{\sigma}(x,\xi)$ to be the square root of $\sum_I |\sigma_I(x,\xi)|^{2/|I|}$, where each σ_I is an iterated Poisson bracket of the functions σ_j, $I = (j_1, \dots j_{|I|})$, $1 \leq |I| \leq m$.

All the positive results above are consistent with a vague and partly conjectural principle: "energy" propagates in phase space along the integral curves of H_{σ_j}, while decaying at a rate dictated by $\tilde{\sigma}$. An analogue is the Feynman-Kac formula for $-\Delta + V$ with potential $V \geq 0$; heat propagates along Brownian paths, decaying at a relative rate proportional to V. From this point of view, global and local regularity are similar notions; the former fails when too much energy is transported from small $|\xi|$ to large $|\xi|$, whereas (micro)local regularity fails whenever too much energy is transported from any one place to another in phase space.

To make this more precise we define [12] a metric ρ_L on $T^*\mathbb{R}^n$: $\rho_L(p,q)$ is the supremum of $|\psi(p) - \psi(q)|$, over all C^1 functions $\psi : T^*\mathbb{R}^n \mapsto \mathbb{R}$ satisfying (i) $|H_{\sigma_j}\psi| \leq \tilde{\sigma}$ and (ii) $|\xi|^{-1}|\nabla_x\psi| + |\nabla_\xi\psi| \leq 1$. This definition is distinct from a phase space metric introduced by Fefferman [15] and Parmeggiani [23]; ρ_L is unchanged if L is multiplied by a constant.

Points $(x,\xi), (x',\xi')$ are said to be δ-separated if $|x-x'| + (|\xi| + |\xi'|)^{-1}|\xi - \xi'| \geq \delta$. Denote by ρ_Δ the metric associated by the above definition to the Laplacian; essentially $d\rho_\Delta^2 = |\xi|^2 dx^2 + d\xi^2$.

The results concerning C^ω/G^s hypoellipticity discussed in this paper are consistent [12] with the requirement that for each $\delta > 0$ there exists $c_\delta < \infty$ such that for all δ-separated pairs p, q, $\rho_L(p,q) \geq c_\delta \rho_\Delta^{1/s}(p,q)$ (with $s = 1$ for $C^\omega = G^1$). For example, the exponent $[1 - \tilde{p}^{-1}(1 - m^{-1})]^{-1}$ encountered above is exactly predicted by this comparison inequality. The same is roughly true for C^∞ hypoellipticity, with the condition $\rho_L(p,q)/\log \rho_\Delta(p,q) \to \infty$ as $\rho_\Delta(p,q) \to \infty$, for δ-separated points p, q, provided that an effective symbol $\tilde{\sigma}$ is defined in an *ad hoc* way on a

[9] See [6]. The genericity hypothesis is needed at present solely because the underlying nonlinear eigenvalue problem is not completely solved.

case by case basis. For global regularity the same remarks apply, provided merely that δ-separatedness is replaced by the assumption that $\big||\xi| - |\xi'|\big| \geq \delta|\xi| + \delta|\xi'|$.

A fundamental question, then, is to what extent ρ_L controls the hypoellipticity and global regularity of L. Skepticism is in order because only $\nabla\psi$, rather than higher-order derivatives, is taken into account. In existing proofs of hypoellipticity, ψ belongs to an appropriate symbol class; in [16], for instance, it must be analytic, with appropriate bounds as $|\xi| \to \infty$.

A delicate example is $X^2 + Y^2$ in $\mathbb{R}^3$, with coordinates (x, y, t), where $X = \partial_x + b(x, y)\partial_t$, $Y = \partial_y + a(x, y)\partial_t$, $a, b \in C^\omega$ are real, and $\partial_x a - \partial_y b \equiv x^6 + y^6 + x^2y^2$. It is shown in [12] that (i) $\rho_L(p, q) \geq c\rho_\Delta(p, q)$ for δ-separated points, but (ii) if ψ is additionally required to belong to the standard class $S^1_{1,0}$, then the modified metric which results no longer satisfies the inequality. To determine whether or not this operator is analytic hypoelliptic might well represent a substantial advance. It would also be desirable to have proofs of negative results based on the same point of view as ρ_L, rather than the nonlinear eigenvalue method.

References

[1] D. Barrett, *Irregularity of the Bergman projection on a smooth bounded domain in* $\mathbb{C}^2$, Ann. of Math. 119 (1984), 431-436.

[2] D. Barrett, *Behavior of the Bergman projection on the Diederich-Fornæss worm,* Acta Math. 168 (1992), 1-10.

[3] H. Boas and E. Straube, *Sobolev estimates for the $\bar\partial$–Neumann operator on domains in $\mathbb{C}^n$ admitting a defining function that is plurisubharmonic on the boundary*, Math. Zeitschrift 206 (1991), 81-88.

[4] H. Boas and E. Straube, *de Rham cohomology of manifolds containing the points of infinite type, and Sobolev estimates for the $\bar\partial$-Neumann problem,* J. Geom. Anal. 3 (1993), 225–235.

[5] M. Christ, *A necessary condition for analytic hypoellipticity*, Math. Research Letters 1 (1994), 241-248.

[6] M. Christ, *Analytic hypoellipticity in dimension two*, preprint.

[7] M. Christ, *The Szegö projection need not preserve global analyticity*, Ann. of Math. 143 (1996), 301-330.

[8] M. Christ, *Global C^∞ irregularity of the $\bar\partial$–Neumann problem for worm domains*, J. Amer. Math. Soc. 9 (1996), 1171-1185.

[9] M. Christ, *Intermediate optimal Gevrey exponents occur*, Comm. Partial Differential Equations 22 (1997), 359-379.

[10] M. Christ, *Remarks on global irregularity in the $\bar\partial$–Neumann problem*, to appear, proceedings of MSRI special year in several complex variables.

[11] M. Christ, *Examples pertaining to Gevrey hypoellipticity*, Math. Research Letters 4 (1997), 725-733.

[12] M. Christ, *Hypoellipticity: Geometrization and speculation*, Proceedings of Conference on Complex Analysis and Geometry in honor of P. Lelong, to appear.

[13] M. Christ, *Hypoellipticity in the infinitely degenerate regime*, preprint.

[14] K. Diederich and J. E. Fornæss, *Pseudoconvex domains: an example with nontrivial Nebenhülle*, Math. Ann. 225 (1977), 275-292.

[15] C. Fefferman, personal communication.

[16] A. Grigis and J. Sjöstrand, *Front d'onde analytique et sommes de carrés de champs de vecteurs*, Duke Math. J. 52 (1985), 35-51.

[17] K. Kajitani and S. Wakabayashi, *Propagation of singularities for several classes of pseudodifferential operators*, Bull. Sc. Math. 2^e série 115 (1991), 397-449.

[18] C. Kiselman, *A study of the Bergman projection in certain Hartogs domains*, Proc. Symp. Pure Math. 52 (1991), Part 3, 219-231.

[19] P. Matheos, *Failure of compactness for the d-bar Neumann problem for two complex dimensional Hartogs domains with no analytic disks in the boundary*, UCLA Ph.D. dissertation, June 1998. To appear in J. Geom. Analysis.

[20] Y. Morimoto, *A criterion for hypoellipticity of second order differential operators*, Osaka J. Math. 24 (1987), 651-675.

[21] Y. Morimoto and T. Morioka, *The positivity of Schrödinger operators and the hypoellipticity of second order degenerate elliptic operators*, Bull. Sc. Math. 121 (1997), 507-547.

[22] O. A. Oleĭnik, On the analyticity of solutions of partial differential equations and systems, *Astérisque* 2,3 (1973), 272–285.

[23] A. Parmeggiani, *Subunit balls for symbols of pseudodifferential operators*, Adv. Math. 131 (1997), 357-452.

[24] Y.-T. Siu, preprint in preparation.

[25] F. Treves, *Analytic hypo-ellipticity of a class of pseudodifferential operators with double characteristics and applications to the $\bar{\partial}$–Neumann problem*, Comm. Partial Differential Equations 3 (1978), 475-642.

[26] F. Treves, *Symplectic geometry and analytic hypo-ellipticity*, preprint.

Michael Christ
Department of Mathematics
University of California
Berkeley, CA 94720-3840
USA
mchrist@math.berkeley.edu

The Baum-Connes Conjecture

Nigel Higson

Abstract. The report below is a short account of past and recent work on a conjecture of P. Baum and A. Connes about the K-theory of group C^*-algebras.

1991 Mathematics Subject Classification: 96K56, 46L80
Keywords and Phrases: Baum-Connes conjecture, group C^*-algebras, K-theory.

1. Introduction. Let G be a second countable, locally compact group. The Baum-Connes conjecture [6,7] proposes a means of calculating K-theory for the reduced C^*-algebra of G using group homology and the representation theory of compact subgroups. It originates in work of Kasparov [19] and Mishchenko [25] on the Novikov higher signature conjecture, ideas of Connes in foliation theory [10], and Baum's geometric description of K-homology theory [8]. The validity of the conjecture has implications in geometry and topology, most notably the Novikov conjecture [14] and the 'stable' Gromov-Lawson-Rosenberg conjecture [30] about positive scalar curvature manifolds. In addition there appear to be close connections to issues in harmonic analysis, for instance the problem of finding explicit realizations of discrete series representations [5]. Indeed a very striking feature of the conjecture is its generality, and the breadth of mathematics with which it makes contact.

The conjecture was first set forth in a 1982 article of Baum and Connes [6], which was unfortunately never published;[1] its current formulation was given in [7]. The last several years have seen a considerable clarification and development of the relationship between the Baum-Connes conjecture and topology, thanks largely to insights of Weinberger [33] linking index theory to surgery theory. A recent monograph of Roe [28] describes the current state of affairs here. Recent progress on the conjecture itself will be described in Section 7 below. A major obstacle to further progress is the lack of a full understanding of the relationship between harmonic analysis and the Baum-Connes conjecture. It seems likely that underlying the conjecture is an as yet unknown governing principle of harmonic analysis. But the conjecture has not drawn the attention of harmonic analysts the way it has the topologists, and this issue remains largely unexamined. This report ends in Section 8 with a few very tentative remarks on the problem.

[1] In fact it *will* be published in the near future, after a 16 year delay.

2. GROUP C*-ALGEBRAS. Denote by $L^1(G)$ the convolution algebra of integrable, complex-valued functions on the locally compact group G. There is a natural involution on $L^1(G)$, making it into a Banach $*$-algebra, and it is easy to see that the non-degenerate $*$-representations of $L^1(G)$ on Hilbert space are in one-to-one correspondence with the unitary representations of G. The group C^*-algebra of G, denoted $C^*(G)$, is the enveloping C^*-algebra of the $L^1(G)$. Its representations are also in one-to-one correspondence with the unitary representations of G, and so both $L^1(G)$ and $C^*(G)$ offer the possibility of a functional-analytic approach to the unitary representation theory and harmonic analysis of G. See [13, Chapter 13].

The group C^*-algebra is particularly well adapted to problems in which the unitary dual $\hat{G}$ [13] is viewed not as a set but as a topological space. Kazhdan's property T [23] offers a good illustration of this. It is equivalent to the assertion that there exists in $C^*(G)$ a projection p whose image in any unitary representation of G is the orthogonal projection onto the G-fixed vectors. In effect p is the continuous function on $\hat{G}$ which is 1 on the trivial representation and zero on its complement. It does not belong to $L^1(G)$ unless G is compact, so generally the disconnectedness of $\hat{G}$ is not simply reflected in the L^1-algebra.

If G is abelian then the Fourier transform provides an isomorphism from $C^*(G)$ to the commutative C^*-algebra of continuous functions on the Pontrjagin dual which vanish at infinity. So in this case $C^*(G)$ precisely captures the topological structure of $\hat{G}$. If G is non-abelian then the ordinary topological structure on the unitary dual is typically very poor (for instance if G is discrete then $\hat{G}$ is a T_0 space only when G is virtually abelian). But following a point of view emphasized by Connes [11] it is now standard in operator algebra theory to think of the noncommutative C^*-algebra $C^*(G)$ as an algebra of continuous functions on $\hat{G}$ which amplifies the classical topological structure of $\hat{G}$.

3. C*-ALGEBRA K-THEORY AND INDEX THEORY. The K-theory groups of a C^*-algebra A are defined in such a way that if A is the C^*-algebra $C_0(X)$ of continuous, complex-valued functions, vanishing at infinity, on a locally compact space then $K_j(A)$ is the Atiyah-Hirzebruch K-theory group $K^{-j}(X)$. See [11] and [19] for overviews of the subject, and the references therein for more details. Following the principle that $C^*(G)$ substitutes for $\hat{G}$, operator algebraists view $K_j(C^*(G))$ as a substitute for $K^{-j}(\hat{G})$. Of course, if G is abelian then thanks to the Fourier isomorphism the formula $K_j(C^*(G)) \cong K^{-j}(\hat{G})$ is not only a point of view but actually a theorem.

There is a direct link between the K-theory of $\hat{G}$ and the index theory of elliptic operators. Suppose that M is a smooth closed manifold and that D is an elliptic partial differential operator on M. It has an integer-valued Fredhom index, but if $\pi_1(M)$ is provided with a homomorphism into a discrete group G then a more refined index, valued in $K_0(C^*(G))$, can be defined as follows. The quotient of $\tilde{M} \times C^*(G)$ by the diagonal action of $\pi_1(M)$ is a flat bundle over M whose fibers are finitely-generated projective modules over $C^*(G)$. If D_G denotes the canonical lifting of D to act on sections of this flat bundle then both kernel(D_G) and cokernel(D_G) are $C^*(G)$-modules. In favorable circumstances they are finitely

generated and projective, and one defines

$$\mathrm{Index}_G(D) = [\mathrm{kernel}(D_G)] - [\mathrm{cokernel}(D_G)] \in K_0(C^*(G)).$$

In general $\mathrm{kernel}(D_G)$ and $\mathrm{cokernel}(D_G)$ may be perturbed so as to become finitely generated and projective, and $\mathrm{Index}_G(D)$ is defined by means of such a perturbation. For abelian groups this construction is due to Lusztig [24]. It was generalized to arbitrary G by Mischenko and, independently, Kasparov.

The ordinary Fredholm index of D can be recovered in an interesting way from $\mathrm{Index}_G(D)$ by noting first that any trace on a C^*-algebra A defines a functional on $K_0(A)$ [11], and then that $C^*(G)$ has a natural trace τ associated to the regular representation of G. It may be shown that $\tau[\mathrm{Index}_G(D)] = \mathrm{Index}(D)$; this is essentially a reformulation of Atiyah's index theorem for covering spaces [3]. A less interesting, but simpler, method is to note that the trivial representation of G also defines a trace τ_0 on $C^*(G)$. It is more or less a tautology that $\tau_0[\mathrm{Index}_G(D)] = \mathrm{Index}(D)$.

If G is a finite group then $\mathrm{Index}(D)$ is the only information within $\mathrm{Index}_G(D)$, but if G is infinite then $\mathrm{Index}_G(D)$ can contain a good deal more. The question of just what it contains is important for the following reason:

3.1. Proposition. [19, 29] *The G-index of the Dirac operator on a closed spin-manifold vanishes if the manifold has positive scalar curvature. The G-index of the signature operator on a oriented manifold is an oriented homotopy invariant.*

4. The Assembly Map. It is well known that a Dirac operator on a closed manifold M^n, combined with a map $M^n \to BG$, determines a class in the K-homology group $K_n(BG)$. This point was first emphasized by Atiyah [2], and an elegant development of it was given by Baum, who *realized* $K_n(X)$ as equivalence classes of triples (M, E, f), where M is a closed spin^c-n-manifold, E is a complex vector bundle on M, and $f\colon M \to X$ is a continuous map. Baum's equivalence relation involves a simple direct sum-disjoint union relation, bordism, and another relation related to the multiplicativity of the index of elliptic operators on fiber bundles. If $X = BG$ then a triple (M^{2n}, E, f) has an index in $K_0(C^*(G))$: form the Dirac operator on M with coefficients in E, and take its G-index along the map $\pi_1(M) \to G$ induced from f. The index depends only on the equivalence class of (M, E, f) and together with a related construction for odd-dimensional manifolds it defines a map

$$\mu\colon K_*(BG) \to K_*(C^*(G)).$$

This *assembly map,* so-called because of its connection with the assembly map of surgery theory [14], was first defined by Kasparov (*c.f.* [19]), although using Kasparov's own realization of K-homology rather than Baum's. He also formulated the following:

4.1. Strong Novikov Conjecture. *The assembly map $\mu\colon K_*(BG) \to K_*(C^*(G))$ is rationally injective.*

Thanks to Proposition 3,1, the Strong Novikov Conjecture implies the Novikov higher signature conjecture (hence its name) [14], which asserts that the class in

$K_n(BG) \otimes \mathbb{Q}$ of the signature operator on a closed, oriented manifold M^n is an oriented homotopy invariant. The Strong Novikov Conjecture also implies the 'stable' Gromov-Lawson-Rosenberg conjecture [30] on positive scalar curvature.

5. THE BAUM-CONNES CONJECTURE. From the point of view of applications to geometry and topology the Strong Novikov conjecture is the most important issue in C^*-algebra K-theory, but the problem nonetheless remains to calculate the K-theory of group C^*-algebras. The Baum-Connes conjecture expresses the idea that every class in the K-theory of a group C^*-algebra is an index, and that the only relations among elements are the natural relations (like bordism, and so on) among index theory problems. Actually the conjecture concerns the quotient of $C^*(G)$ corresponding to the closed subset $\hat{G}_{\mathrm{red}} \subset \hat{G}$ comprised of those unitary representations which are weakly contained in the regular representation. This *reduced* C^*-algebra of G, denoted $C^*_{\mathrm{red}}(G)$, is the completion of $L^1(G)$ in its regular representation as bounded operators on $L^2(G)$. The C^*-algebra $C^*_{\mathrm{red}}(G)$ coincides with $C^*(G)$ if and only if G is amenable [13, Chapter 18].

The Baum-Connes conjecture is most easily formulated for discrete groups without torsion. It has already been noted that if G is a finite group then the assembly map is essentially trivial. Furthermore a finite, nontrivial subgroup H in any discrete group G contributes a projection $1/|H| \sum_{h \in H}[h]$ to $C^*(G)$ whose K-theory class is not in the image of the assembly map. So the restriction to torsion-free groups in the following is certainly necessary:

5.1. BAUM-CONNES CONJECTURE FOR TORSION-FREE GROUPS. *If G is a discrete and torsion-free group then the assembly map*

$$\mu_{\mathrm{red}} \colon K_*(BG) \to K_*(C^*_{\mathrm{red}}(G)),$$

obtained from the assembly map μ of 4.1 *using the regular representation $C^*(G) \to C^*_{\mathrm{red}}(G)$, is an isomorphism.*

It is usual to cite Kazhdan's property T as the reason why $C^*_{\mathrm{red}}(G)$ is used in place of $C^*(G)$: the Kazhdan projection $p \in C^*(G)$, if it exists, defines a class in $K_0(C^*(G))$ which is not in the image of the assembly map μ (if G is infinite), since the traces τ and τ_0 of the last section disagree on it, whereas $\tau(x) = \mathrm{Index}(D) = \tau_0(x)$ for every class $x \in K_0(C^*(G))$ which is the G-index of an elliptic operator D.

In fact if G is infinite and has property T then *every* finite-dimensional, unitary representation of G determines a Kazhdan-type projection in $C^*(G)$ and in this way the entire character ring of finite-dimensional unitary representations of G embeds as a direct summand of $K_0(C^*(G))$. This ring is typically very large and very complicated, and it is not within the range of the assembly map. So the idea that μ (as opposed to μ_{red}) is an isomorphism is not only incorrect, it is hopelessly wrong.

The Kazhdan projection maps to zero in $C^*_{\mathrm{red}}(G)$ so these problems vanish for the reduced C^*-algebra and for μ_{red}. Of course this is not in itself a very powerful reason to believe in 5.1, which could fail for any number of reasons unrelated to property T. More compelling evidence will be presented in the next section.

The statement of the Baum-Connes conjecture for general (second-countable) locally compact groups uses Kasparov's equivariant KK-theory [21]. Associated to any G there is a proper G-space $\mathcal{E}G$, which is universal in the sense that any other proper G-space maps into it in a way which is unique up to equivariant homotopy [7]. Using Kasparov's KK-theory the equivariant K-homology $K_*^G(\mathcal{E}G)$ may be defined. If G is discrete and torsion free then $\mathcal{E}G$ is the universal principal space EG and $K_*^G(\mathcal{E}G) = K_*(BG)$. For general G there is an assembly map

$$\mu_{\mathrm{red}}\colon K_*^G(\mathcal{E}G) \to K_*(C^*_{\mathrm{red}}(G))$$

very similar to the one already considered: a cycle for $K_*^G(\mathcal{E}G)$ is an 'abstract' elliptic operator D on a proper G-space and μ_{red} associates to D its equivariant index.

5.2. Baum-Connes Conjecture. [7] *If G is any second countable, locally compact group then the assembly map μ_{red} is an isomorphism.*

6. Lie Groups. What is currently known about the Baum-Connes conjecture? Progress has been made in two ways: by representation-theoretic arguments which calculate both $K_*^G(\mathcal{E}G)$ and $K_*(C^*_{\mathrm{red}}(G))$ explicitly as abelian groups, for certain G, and verify that μ_{red} is an isomorphism; and by K-theoretic arguments which construct an inverse to μ_{red} in certain other cases.

Perhaps the best evidence in favour of the Baum-Connes conjecture comes from the former method. If G is a connected Lie group and K is its maximal compact subgroup then the homogeneous space G/K is a universal proper G-space $\mathcal{E}G$. If, for simplicity, G/K is even-dimensional and admits a G-equivariant spin-structure then the Baum-Connes conjecture is equivalent to the assertion that the map

$$\tilde{\mu}_{\mathrm{red}}\colon R(K) \to K_0(C^*_{\mathrm{red}}(G)),$$

which associates to each representation $[V] \in R(K)$ the G-index of the twisted Dirac operator D_V on G/K, is an isomorphism of abelian groups, and that in addition, $K_1(C^*_{\mathrm{red}}(G)) = 0$. See [7]. This is also known as the *Connes-Kasparov conjecture* for G [19,11].

If G is semisimple then a unitary representation of G is weakly contained in the regular representation if and only if it is tempered. If G is a *complex* semisimple group then there is a Morita equivalence $C^*_{\mathrm{red}}(G) \sim C_0(\hat{G}_{\mathrm{red}})$ [26] connecting $C^*_{\mathrm{red}}(G)$ to the tempered dual $\hat{G}_{\mathrm{red}}$, which in the complex case is a Hausdorff locally compact space. For general semisimple groups $\hat{G}_{\mathrm{red}}$ is 'almost' Hausdorff and $C^*_{\mathrm{red}}(G)$ is Morita equivalent to a C^*-algebra which is 'almost' $C_0(\hat{G}_{\mathrm{red}})$. And while phenomena such as the reducibility of non-generic principal series representations and the non-vanishing of associated R-groups complicate matters, it is nonetheless possible (at least in the linear case) to explicitly compute $C^*_{\mathrm{red}}(G)$ and the groups $K_*(C^*_{\mathrm{red}}(G))$ [32]. It is further possible to calculate $\tilde{\mu}_{\mathrm{red}}$ and verify that it is indeed an isomorphism. Each discrete series representation of G contributes a generator to K-theory, and following [5] these are accounted for as equivariant indices of twisted Dirac operators on G/K—in other words by elements in $R(K)$.

The proof [32] that μ_{red} is an isomorphism is a careful extension of these discrete series arguments. It covers a great deal of the territory of tempered representation theory: as yet there is no simple, conceptual proof.[2]

For certain Lie groups G, for instance $G = SO(n,1)$, it is known by the other methods that the Baum-Connes conjecture is true not only for G but for any discrete subgroup (see the next section). Since the representation-theoretic analysis of $\tilde{\mu}_{\mathrm{red}}$ reveals nothing special about $SO(n,1)$, an optimistic extrapolation suggests that if a counterexample to the Baum-Connes conjecture exists, it ought not to be found among the discrete subgroups of semisimple groups.

7. THE DIRAC-DUAL DIRAC ARGUMENT. For most groups it is impossible to determine $\hat{G}_{\mathrm{red}}$, or to otherwise directly compute $K_*(C^*_{\mathrm{red}}(G))$. Kasparov has however devised a beautiful, indirect route to the Baum-Connes conjecture using his equivariant KK-theory [21].

If A and B are G-C^*-algebras then the Kasparov's $KK^G(A,B)$ is the group of morphisms from A to B in additive category which broadens in a certain sense the homotopy category of G-C^*-algebras and equivariant $*$-homomorphisms. The composition law

$$KK^G(A,B) \otimes KK^G(B,C) \to KK^G(A,C)$$

is called the *Kasparov product.* Key to Kasparov's approach to the Baum-Connes conjecture is the notion of a *proper* G-C^*-algebra [21,15], which is a mildly noncommutative generalization of the notion of a proper G-space (the algebra of continuous functions, vanishing at infinity, on a locally compact, proper G-space is the prototypical example of a proper G-C^*-algebra).

7.1. PROPOSITION. *Let G be a second-countable, locally compact group. If the elementary G-C^*-algebra $\mathbb{C}$ is KK^G-equivalent to a proper G-C^*-algebra then the Baum-Connes assembly map for G is an isomorphism.*

This result of Tu [31], which condenses Kasparov's method to its essence, may be interpreted as follows.[3] Using KK-theory a generalized assembly map

$$\mu_{\mathrm{red},A}\colon KK^G_*(\mathcal{E}G, A) \to K_*(C^*_{\mathrm{red}}(G,A))$$

can be defined, involving a 'coefficient' G-C^*-algebra A (here $C^*_{\mathrm{red}}(G,A)$ denotes the crossed product C^*-algebra). If $A = \mathbb{C}$ then $\mu_{\mathrm{red},A}$ is the assembly map of 5.2. If on the other hand A is proper then $\mu_{\mathrm{red},A}$ should be an isomorphism, because proper actions are locally modeled by actions of compact groups, while the Baum-Connes conjecture is readily verified for compact groups. But if $\mathbb{C}$ is KK^G-equivalent to a proper G-C^*-algebra then as far as the assembly map is concerned $\mathbb{C}$ *is* a proper G-C^*-algebra, and the proposition follows.

[2] Similar remarks apply to p-adic groups. The Baum-Connes conjecture has been checked for p-adic $GL(n)$ by a detailed determination of $\hat{G}_{\mathrm{red}}$, of $C^*_{\mathrm{red}}(G)$, of $K^G_*(\mathcal{E}G)$, and of the assembly map for $GL(n)$ [9]. But the verification does not offer much insight into what underlies the Baum-Connes isomorphism.

[3] Tu's proof is somewhat different, but see [15] for a similar result which is proved along these lines.

The approach to the Baum-Connes conjecture through 7.1 is hereditary, in the sense that it proves the Baum-Connes conjecture not only for G but for any closed subgroup of G at the same time. The method has been successfully applied to connected, amenable Lie groups [21], to $SO(n,1)$ [20], to $SU(n,1)$ [18], and most recently to groups which admit continuous, isometric *affine* actions on Hilbert space which are proper in the sense that $\|g \cdot v\| \to \infty$ as $g \to \infty$, for every vector v [16,17]. This last class includes all the previous ones, *all* (second countable) amenable groups, and all Coxeter groups. It seems to push Kasparov's method almost as far as it can go.

Thanks to important work of Pimsner concerning group actions on trees [27] it is also known that the groups to which 7.1 applies are closed under operations like amalgamated free products [31]. As far as discrete groups are concerned, nothing more is known, and in particular the conjecture is known for no infinite, discrete, property T group. Indeed, a feature of Kasparov's method, as summarized in 7.1, is that it treats $C^*(G)$ and $C^*_{\mathrm{red}}(G)$ equally, and proves that $K_*(C^*(G)) \cong K_*(C^*_{\mathrm{red}}(G))$ (in the language of KK-theory, G is K-amenable).

The 'Dirac-dual Dirac' terminology comes from Kasparov's original work on connected Lie groups. If $M = G/K$ then the G-C^*-algebra $A = C_0(TM)$ is proper. The Dirac operator on TM defines an element $\alpha \in KK^G(A, \mathbb{C})$, while a class $\beta \in KK^G(\mathbb{C}, A)$ may be defined which is, in the case $G = \mathbb{R}^n$, closely related to the Fourier transform of the Dirac operator. Kasparov was able to show that, for any G, $\beta \circ \alpha = 1 \in KK^G(A, A)$ (the argument is closely related to Atiyah's elliptic operator proof of Bott periodicity [1]), but not in general that $\alpha \circ \beta = 1 \in KK^G(\mathbb{C}, \mathbb{C})$. Indeed, thanks to property T the latter is false in general. But the former is enough to imply the split injectivity of the assembly map for G and its discrete subgroups, a result which was subsequently extended by Kasparov and Skandalis to p-adic groups [22], and others. Injectivity of μ_{red} for a discrete group G implies the Novikov higher signature conjecture.

8. Representation Rings. Let G be a compact group and denote by $R(G)$ the its character ring. Atiyah, Hirzebruch and Segal [4] defined and studied an interesting ring homomorphism

$$\nu \colon R(G) \to K(BG),$$

connecting $R(G)$ to the representable K-theory of the classifying space BG. In his work on the Novikov conjecture [19,21] Kasparov defined a similar homomorphism for any locally compact group. It is in some sense dual to the assembly map.

A (unitary) *Fredholm representation* of a locally compact group G consists of two separable Hilbert spaces H_0 and H_1, equipped with unitary representations of G, and a bounded Fredholm operator $F \colon H_0 \to H_1$ for which $g(F) - F$ is a compact operator-valued and norm-continuous function of $g \in G$. The *Kasparov representation ring* of G, denoted $R(G)$, is the abelian group of homotopy equivalence classes of Fredholm representations of G. As its name suggests, $R(G)$ is a ring. In fact $R(G) = KK_G(\mathbb{C}, \mathbb{C})$ and the ring structure is a special case of the Kasparov product.

The ring of finite-dimensional unitary representations of G maps into Kasparov's $R(G)$, and if G is compact then this map is an isomorphism: its inverse is defined

by averaging over G any Fredholm representation to make the operator F exactly equivariant, and then taking the G-index of F, which is a formal difference of finite-dimensional unitary representations.

Kasparov's generalized Atiyah-Hirzebruch-Segal map $\nu\colon R(G) \to K(BG)$ takes a Fredholm representation $F\colon H_0 \to H_1$ and associates to it an equivariant family of Fredolm operators $F_x\colon H_0 \to H_1$, parametrized by the universal space EG, characterized up to equivariant homotopy by the property that each F_x is a compact perturbation of F. It follows from Kuiper's theorem that $K(BG)$ may be described as equivariant homotopy classes of equivariant Fredholm families on EG. Actually, for the present purposes it is better to consider the equivariant K-theory group $K_G(\mathcal{E}G)$, defined from equivariant homotopy classes of equivariant Fredholm families on $\mathcal{E}G$. Kasparov's prescription then defines a map

$$\nu\colon R(G) \to K_G(\mathcal{E}G).$$

If G is compact then, unlike the Atiyah-Hirzebruch-Segal map, this is tautologically an isomorphism; if G is a connected Lie group then ν identifies with the restriction map $R(G) \to R(K)$.

If G is any group to which 7.1 applies then the map $\nu\colon R(G) \to K_G(\mathcal{E}G)$ is a ring isomorphism. If G is a Lie group and if $\gamma \in R(G)$ is the Kasparov product $\alpha \circ \beta \in KK^G(\mathbb{C}, \mathbb{C})$ considered in the last section then γ is an idempotent and Kasparov showed that the map ν takes $\gamma \cdot R(G)$ isomorphically to $K_G(\mathcal{E}G)$. So if ν is an isomorphism then $\gamma = 1$ and the Baum-Connes conjecture follows. On the other hand, if G has property T then certainly $\gamma \neq 1$ and so ν is not an isomorphism.

To explore this issue further it is worth contemplating a *non-unitary* Kasparov representation ring $R_{\text{n.u.}}(G)$, comprised of homotopy classes of *non-unitary* Fredholm representations (they are defined more or less as before, but the representations of G on H_0 and H_1 are required only to be continuous, not unitary). Considering non-unitary representations makes no change to $K_G(\mathcal{E}G)$, and one can proceed to construct and study the natural homomorphism $\nu\colon R_{\text{n.u.}}(G) \to K_G(\mathcal{E}G)$.

8.1. PROPOSITION. *If $G \subset SL(n, \mathbb{R})$ then the homomorphism $\nu\colon R_{\text{n.u.}}(G) \to K_G(\mathcal{E}G)$ is a ring isomorphism.*

The proof hinges on showing that the image of $\gamma \in R(G)$ in $R_{\text{n.u.}}(G)$ is 1. It can likely be generalized to any Lie group.

Suppose now that $C^*_{\text{red}}(G)$ is put to one side for a moment and in its place is considered a Banach, or Frechet, algebra tailored to the full (not necessarily unitary) representation theory of G. For instance for a finitely generated discrete group one might consider

$$\mathcal{S}(G) = \{\, f\colon G \to \mathbb{C} \,:\, \sum |f(g)| A^{|g|} < \infty,\ \forall A > 0 \,\},$$

where $|g|$ denotes the word length of $g \in G$. A Baum-Connes type assembly map

$$\mu_{\text{n.u.}}\colon K^G_*(\mathcal{E}G) \to K_*(\mathcal{S}(G))$$

may be defined and it is reasonable to guess that 8.1 will imply that $\mu_{\text{n.u.}}$ is an isomorphism (for Lie groups). The proof should be an adaptation of Kasparov's Dirac-dual Dirac argument to the context of Frechet algebras and non-unitary representations. The tools for carrying out such an argument—most important among them a serviceable KK-theory for non-C^*-algebras—are now coming into being [12], and it is likely that this guess about $\mu_{\text{n.u.}}$ is not far from a theorem. Granted this, the Baum-Connes isomorphism for $C^*_{\text{red}}(G)$ reduces to a sort of restriction isomorphism

$$K_*(\mathcal{S}(G)) \underset{\cong}{\longrightarrow} K_*(C^*_{\text{red}}(G))$$

identifying K-theoretic invariants of the 'non-unitary dual' of G with the same for the reduced, or tempered, dual.

It might be thought that the inclusion of the Frechet algebra $\mathcal{S}(G)$ as a dense subalgebra of $C^*_{\text{red}}(G)$ induces an isomorphism in K-theory by an elementary approximation argument like the one which shows that K-theory for manifolds, defined using smooth vector bundles, is the same as topological K-theory. Unfortunately the approximation arguments which are known do not apply: even at the cohomological level of K-theory a good deal of analysis $\mathcal{S}(G)$ seems to separate from $C^*_{\text{red}}(G)$. On the other hand the idea that the non-unitary representation theory of G is describable, or parametrizable, in terms of the tempered dual is not foreign to harmonic analysis. At the present time a closer analysis of this point and its relation to K-theory seems to offer the best chance of progress on the Baum-Connes conjecture.

References

1. M. F. Atiyah, *Bott periodicity and the index of elliptic operators*, Oxford Quarterly J. Math. **19** (1968), 113–140.
2. M.F. Atiyah, *Global theory of elliptic operators*, Proc. Int. Symp. on Functional Analysis, University of Tokyo Press, Tokyo, 1969, pp. 21–30.
3. M.F. Atiyah, *Elliptic operators, discrete groups and von Neumann algebras*, Astérisque **32–3** (1976), 43–72.
4. M.F. Atiyah and G. Segal, *Equivariant K-theory and completion*, J. Diff. Geom. **3** (1969), 1–18.
5. M.F. Atiyah and W. Schmid, *A geometric construction of the discrete series for semisimple Lie groups*, Inventiones Math. **42** (1977), 1–62.
6. P. Baum and A. Connes, *Geometric K-theory for Lie groups and foliations*, Preprint (1982).
7. P. Baum, A. Connes and N. Higson, *Classifying space for proper G-actions and K-theory of group C^*-algebras*, Contemp. Math. **167** (1994), 241–291.
8. P. Baum and R. Douglas, *K-homology and index theory*, Proc. Symp. Pure Math **38**, Part 1, (Operator Algebras and Applications, R. Kadison, ed.), American Mathematical Society, 1982, pp. 117–173.
9. P. Baum, N. Higson and R. Plymen, *A proof of the Baum-Connes conjecture for p-adic $GL(n)$*, Comptes Rendus Acad. Sci. Paris, Serie I **325** (1997), 171-176.
10. A. Connes, *A survey of foliations and operator algebras*, Operator algebras and applications, Part 1, Amer. Math. Soc., Providence, 1982, pp. 521–628.
11. A. Connes, *Noncommutative geometry*, Academic Press, 1994.
12. J. Cuntz, *Bivariante K-theorie für lokalkonvexe Algebren und der Chern-Connes-Charakter*, Doc. Math. **2** (1997), 139–182.

13. J. Dixmier, *C^*-algebras*, English translation, revised edition, North Holland, Amsterdam New York Oxford, 1982.
14. S. Ferry, A. Ranicki and J. Rosenberg, *A history and survey of the Novikov conjecture*, Novikov conjectures, Index theorems and rigidity, vol. 1, S. Ferry, A. Ranicki and J. Rosenberg, editors, Cambridge University Press, Cambridge, 1995, pp. 7–66.
15. E. Guentner, N. Higson and J. Trout, *Equivariant E-theory*, Preprint (1997).
16. N. Higson and G. Kasparov, *Operator K-theory for groups which act properly and isometrically on Hilbert space*, E.R.A. Amer. Math. Soc. **3** (1997), 131–142.
17. P. Julg, *Travaux de N. Higson et G. Kasparov sur la conjecture de Baum-Connes*, Séminaire Bourbaki **841** (1997-98).
18. P. Julg and G.G. Kasparov, *Operator K-theory for the group $SU(n,1)$*, J. Reine Angew. Math. **463** (1995), 99–152.
19. G.G. Kasparov, *Operator K-theory and its applications: elliptic operators, group representations, higher signatures, C^*-extensions*, Proc. Internat. Congress of Mathematicians, vol 2, Warsaw, 1983, pp. 987–1000.
20. G.G. Kasparov, *Lorentz groups: K-theory of unitary representations and crossed products*, (English Transl.), Sov. Math. Dokl. **29** (1984), 256–260.
21. G.G. Kasparov, *Equivariant KK-theory and the Novikov conjecture*, Inventiones Math. **91** (1988), 147-201.
22. G.G. Kasparov and G. Skandalis, *Groups acting on buildings, operator K-theory and the Novikov conjecture*, K-Theory **4** (1991), 303–338.
23. D. Kazhdan, *Connection of the dual space of a group with the structure of its closed subgroups*, Funct. Anal. and its Appl. **1** (1967), 63–65.
24. G. Lusztig, *Novikov's higher signature and families of elliptic operators*, Invent. Math. **7** (1972), 229–256.
25. A. S. Mishchenko, *C^*-algebras and K-theory*, Springer Lecture Notes in Math. **763** (1979), 262–274.
26. M. Penington and R. Plymen, *The Dirac operator and the principal series for complex semisimple Lie groups*, J. Functional Anal. **53** (1983), 269-286.
27. M. Pimsner, *KK-groups of crossed products by groups acting on trees*, Invent. Math. **86** (1986), 603–634.
28. J. Roe, *Index theory, coarse geometry and topology of manifolds*, CBMS conference series **90** (1996).
29. J. Rosenberg, *C^*-algebras, positive scalar curvature and the Novikov conjecture*, Publ. Math. IHES **58** (1983), 197–212.
30. S. Stolz, *Positive scalar curvature and the Baum-Connes conjecture*, In preparation.
31. J.-L. Tu, *The Baum-Connes conjecture and discrete group actions on trees*, Preprint (1997).
32. A. Wassermann, *Une démonstration de la conjecture de Connes-Kasparov pour les groupes de Lie linéaires connexes réductifs*, C.R. Acad. Sci. Paris, Sér 1 **304** (1987), 559–562.
33. S. Weinberger, *Aspects of the Novikov conjecture*, Contemporary Math. **105** (1990), 281–297.

Nigel Higson
Department of Mathematics
Pennsylvania State University
University Park
PA 16802 USA
higson@math.psu.edu

On the Bilinear Hilbert Transform

Michael T. Lacey[1]

Abstract.

In joint work with C. Thiele, the author has shown that A. Calderón's bilinear Hilbert transform extends to a bounded operator on certain products of L^p spaces. This article illustrates the method of proof by giving a complete proof of an inequality which is slightly weaker than the original conjecture of Calderón concerning this transform.

1991 Mathematics Subject Classification: 42A50
Keywords and Phrases: hilbert transform, time–frequency analysis

1 Introduction

This note discusses a recently developed theory for the bilinear Hilbert transform, defined by

$$Hfg(x) := \lim_{\epsilon \to 0} \int_{|y|>\epsilon} f(x+y)g(x-y)\ \frac{dy}{y}.$$

A conjecture of A. Calderón concerned possible extensions of this transform to a bounded operator on products of L^p spaces. In this regard, note that the term dy/y is dimensionless, so that any inequalities satisfied by Hfg are those of Hölder's inequality. In collaboration with C. Thiele, the author has established

Theorem 1.1. *H extends to a bounded operator on $L^p \times L^q$ into L^r if*

$$1/p + 1/q = 1/r, \qquad 1 < p, q \leq \infty, \qquad 2/3 < r < \infty.$$

In particular, H maps $L^2 \times L^2$ into L^1, which is notable as the linear Hilbert transform does not preserve L^1. This was the form of Calderón's origonal conjecture dating from 1964.

The interest in the Theorem lies in the method of proof, as it can be seen as an outgrowth and rëexamination of a group of sophisticated and subtle techniques invented first by L. Carleson [1] and later by C. Fefferman [3]—those in the celebrated proof of the pointwise convergence of Fourier series. The central point is to

[1]The author has been supported by an NSF Grant, DMS—9706884.

exploit orthogonality in a situation which is highly sensitive to the temporal and frequency aspects of the functions under consideration.

To explain a significant part of the method of proof, we limit our discussion to a single instance of the Theorem above, one that is free of some of the technicalities of the general case treated in [4, 5, 6, 7]. The next section provides background for the rest of the paper. Then we present the geometric–combinatorial model of the bilinear Hilbert transform and prove that it maps $L^2 \times L^2$ into weak L^1.

I am grateful to L. Grafakos for taking the time to carefully read this manuscript and providing me with a long list of improvements.

2 Preliminaries

The Fourier transform is taken to be

$$\mathcal{F}f(\xi) = \hat{f}(\xi) := \int e^{-2\pi i x\xi} f(x)\, dx.$$

We refer to x as the time variable and ξ as the frequency variable. The inner product will be denoted by $\langle f, g\rangle = \int f\bar{g}\, dx$.

The linear Hilbert transform is given by the principal value of convolution with $1/y$.

$$Hf(x) := \lim_{\epsilon\to 0} \frac{1}{\pi} \int_{|y|>\epsilon} f(x-y)\, \frac{dy}{y}.$$

Context will distinguish linear and bilinear forms of the transform. We recall that

$$\mathcal{F}Hf(\xi) = \mathcal{F}f(\xi) \lim_{\epsilon\to 0} \frac{1}{\pi} \int_{\epsilon<|y|<1/\epsilon} e^{-2\pi i y\xi}\, \frac{dy}{y} = i\mathcal{F}f(\xi)\ \mathrm{sign}(\xi).$$

But $\mathcal{F}$ is unitary, so that H is bounded on L^2. This calculation also shows that the singularity of $1/y$ has the effect of distinguishing the origin in frequency—$\mathrm{sign}(\xi)$ has a jump discontinuity at 0.

Deeper properties of H are addressed by decomposing the non–local kernel $1/y$ into a sum of localized kernels. Fix a symmetric Schwartz function ψ which resolves the identity in that

$$\sum_{j=-\infty}^{\infty} \psi(2^j y) \equiv 1, \qquad y \neq 0. \tag{2.2}$$

Then, let $\psi_j(y) = y^{-1}\psi(2^j y)$, which is the same as $\psi_j(y) = 2^j\psi_1(2^j y)$. Each ψ_j gives rise to an operation

$$H_j f(x) = \int f(x-y)\psi_j(y)\, dy. \tag{2.3}$$

This is a local operation in that ψ_j is no longer a singular kernel. In fact, $|\psi_j(y)| \le C2^j$, and it decays rapidly for $|y| > 2^{-j}$. Each H_j trivially maps L^p into itself for $1 \le p \le \infty$. We say that H_j as *scale* 2^j.

Carleson's Theorem can now be recalled in a form specific to our purposes. To prove the pointwise convergence of Fourier series, the maximum partial Fourier sums must be controlled. On the real line, this is equivalent to providing a control on the maximal operator

$$\mathcal{C}f(x) := \sup_{\lambda} |H[e^{2\pi i\lambda\cdot} f](x)|.$$

Carleson proved that $\mathcal{C}$ maps L^2 into itself. See [1].

The aspect of $\mathcal{C}$ that distinguishes it from other operators of harmonic analysis is its invariance under conjugation of f by exponentials. Conjugation being dual to translation, we see that $\mathcal{C}$ has no distinguished point in frequency, which is a feature shared with the bilinear Hilbert transform. Denote, for the moment, $f^\lambda(x) = e^{2\pi i\lambda x} f(x)$. Then one readily sees that $Hf^\lambda g^\lambda(x) = e^{4\pi i\lambda x} Hfg(x)$. That is, Hfg commutes with conjugation, and so again it has no distinguished points in frequency. This suggests that the analysis of the two operators should be intimately related.

It also suggests that such an analysis cannot distinguish points in the frequency variable. We follow such a path. Each scale of the $1/y$ kernel requires a separate decomposition of f, g that is sensitive to both temporal and frequency aspects of the functions. It is convienent to introduce this with elements of the geometry of the phase plane.

Let φ be a Schwartz function on L^2 norm one, with Fourier transform supported in $[-\frac{1}{2}, \frac{1}{2}]$, so that

$$|\hat\varphi(\xi)|^2 + |\hat\varphi(\xi + \tfrac{1}{2})|^2 \equiv \text{constant} \quad -\tfrac{1}{2} \le \xi < \tfrac{1}{2}.$$

Let $s = I_s \times \omega_s$ be a rectangle of area 1 and set

$$\varphi_s(x) := e^{2\pi i c(\omega_s)x} |I_s|^{-1/2} \varphi\Big(\frac{s - c(I_s)}{|I_s|}\Big), \tag{2.4}$$

where $c(J)$ denotes the center of the interval J. With this definition, I_s plays the role of the temporal variable and ω_s of the frequency. The initial set of rectangles we are interested in are

$$\mathbf{S}_1 := \{(n, n+1) \times (m/2, m/2+1) \mid m, n \in \mathbb{Z}\},$$

which have scale 1. To make our considerations independent of scale, for each integer j, let $\mathbf{S}_j$ be the rectangles which are images of those in $\mathbf{S}_1$ under the area preserving dilation $(x, \xi) \to (2^{-j}x, 2^j\xi)$. And set $I_j f = \frac{1}{2} \sum_{s \in \mathbf{S}_j} \langle f, \varphi_s\rangle \varphi_s$. Each I_j is just a change of scale applied to I_1. And I_1 is in fact the identity on L^2. See [2, Section 3.4.4] for a full discussion of this fact.

Now, for the transform, take a particular scale of the transform, as given in (2.3). We have

$$H_j fg = I_j[\, H_j(I_j f, I_j g)\,] = \sum_{s1, s2, s3 \in \mathbf{S}_j} \langle f, \varphi_{s1}\rangle \langle g, \varphi_{s2}\rangle \langle H_j \varphi_{s1} \varphi_{s2}, \varphi_{s3}\rangle. \tag{2.5}$$

The point underlying this expression is that the triple sum over $\mathbf{S}_j$ diagonalizes, and the "diagonal terms," collected over 2^j, have useful combinatorial structures.

The significant aspect of the diagonalization takes place on the Fourier side, which is made explicit by choosing the kernel ψ in (2.2) so that $\mathcal{F}\psi$ is supported on $(-10,-4]\cup[4,10)$. Recall that φ is supported on $[-\frac{1}{2},\frac{1}{2}]$ and therefore φ_s is supported on ω_s.

Then the final inner product in (2.5) is determined in part by the Fourier support of $H_j\varphi_{s1}\varphi_{s2}$. We claim that this function is zero unless $\pm(\omega_{s1}-\omega_{s2})\cap[4\cdot 2^j, 10\cdot 2^j)\neq 0$, and then the Fourier transform is supported in $\omega_{s1}+\omega_{s2}$. That is, the three frequency intervals ω_{sk} depend in fact on only one parameter. A formal calculation verifies the claim. Expand φ_{s1} in a dual variable τ_1 and φ_{s2} in τ_2, then

$$H_j\varphi_{s1}\varphi_{s2}(x)=\iint e^{2\pi i(\tau_1+\tau_2)x}\widehat{\varphi_{s1}}(\tau_1)\widehat{\varphi_{s2}}(\tau_2)\widehat{\psi}(2^{-j}(\tau_1-\tau_2))\,d\tau_1 d\tau_2.$$

In this expression, recall that $\mathcal{F}\varphi_s$ is supported on ω_s. Note that $\tau_1+\tau_2$ is the frequency variable for the left hand side, and then our claims follow.

The second diagonalization is of a more trivial nature. Recall that φ is a Schwartz function, so that φ_s is highly localized around I_s. In addition, ψ has rapid decay away from the origin. From these considerations, it follows that for all $s1, s2, s3$,

$$|\langle H_j\varphi_{s1}\varphi_{s2},\varphi_{s3}\rangle|\le C_N|I_{s1}|^{-1/2}[1+|I_{s1}|^{-1}\max_{j\neq k}\operatorname{dist}(I_{sj},I_{sk})]^{-N},\quad N\ge 1.$$

A final diagonalization procedure, which we do not present here, reduces the sum over j in (2.5) to a sum of "model sums" as defined in the next section. This is worth doing because the model sums can be analysized using only natural geometric–combinatorial considerations: There are no *ad hoc* features of the argument for the model sums.

3 Model Sums

In this section we state and prove a theorem which is general enough to prove the most important single inequality for the bilinear Hilbert transform. But this will require some definitions—we give them and illustrate with special cases which contain all of the difficulty of the general case.

A collection of intervals $\mathcal{G}$ is a *grid* if for all $I, I'\in\mathcal{G}$, we have $I\cap I'$ equal to $\emptyset$, I or I', and $I\subset_{\neq} I'$ implies $2|I|\le|I'|$. The special cases of interest are collections of dyadic and triadic intervals.

A collection $\mathbf{S}$ of rectangles $s=R_s\times\rho_s$ are called *tiles* if for all $s,s'\in\mathbf{S}$ we have $|s|\le 4|s'|$, and in addition, $\{R_s\mid s\in\mathbf{S}\}$ and $\{\rho_s\mid s\in\mathbf{S}\}$ are grids. For an example consider triadic rectangles $s=R_s\times\rho_s$ of area one. Each ρ_s is a union of three triadic intervals of equal length, $\rho_{s1},\rho_{s2},\rho_{s3}$. Then all of the rectangles $R_s\times\rho_s, R_s\times\rho_{sj}$, $1\le j\le k$ form a collection of tiles.

This is most relevant, because of the connection to the decomposition in (2.5), and the diagonalization that was discussed there. For that reason, we make

a further definition along these lines: A collection of tiles $\mathbf{S}$ are called *tri–tiles* if to each $s = R_s \times \rho_s \in \mathbf{S}$ there are three tiles $sj = R_s \times \rho_{sj}$, $1 \le j \le 3$, with these properties. For all $s \in \mathbf{S}$ and all j, (a) $\rho_{sj} \subset \rho_s$, (b) $\xi_1 < \xi_2 < \xi_3$, for all $\xi_j \in \rho_{sj}$, and (c) $\{s, s1, s2, s3 \mid s \in \mathbf{S}\}$ is a collection of tiles.

The tri–tiles will describe the location of functions in the time–frequency plane. A collection of functions $\{\phi_{sj} \mid s \in \mathbf{S},\ 1 \le j \le 3\}$ are *adapted* to a collection of tri–tiles $\mathbf{S}$ if for all $s \in \mathbf{S}$ and all j, $\|\phi_{sj}\|_2 \le 1$, there is an affine map α_j on $\mathbb{R}$ so that $\mathcal{F}\phi_{sj}$ is suported on $\alpha_j(\rho_{sj})$, and we have

$$|\phi_{sj}(x)| \le \frac{C_0}{\sqrt{|R_s|}} \Big(1 + \frac{|x - c(I_s)|}{|R_s|}\Big)^{-10}, \tag{3.6}$$

$$\langle \phi_{sj}, \phi_{s'j} \rangle = 0 \quad \text{if } I_s \ne I_{s'},\ \omega_{sj} = \omega_{s'j}. \tag{3.7}$$

One should compare these conditions to the definition of φ_s in (2.4).

The discrete combinatorial model of the bilinear Hilbert transform is

$$\mathcal{H}^{\mathbf{S}} f_1 f_2(x) = \sum_{s \in \mathbf{S}} |R_s|^{-1/2} \phi_{s3}(x) \prod_{j=1}^{2} \langle f_j, \phi_{sj} \rangle.$$

Here, $\mathbf{S}$ is a set of tri–tiles and the ϕ_{sj} are adapted to $\mathbf{S}$. Compare this sum to (2.5).

These operators obey the same inequalities as in Theorem 1.1. But for this exposition, we restrict our attention to that case which follows from purely L^2 arguments, that is

Lemma 3.8. *The operator $\mathcal{H}^{\mathbf{S}}$ maps $L^2 \times L^2$ into weak L^1. The norm of the operator depends only on the constant C_0 in (3.6).*

3.1 The Key Lemma

We state and prove the key Lemma in the proof of Lemma 3.8. But first we shall have to delineate the structures with which the Lemma must be stated. The tri–tiles admit a partial order. Thus we write $s < s'$ if $\rho_s \supset \rho_{s'}$ and $I_{s'} \supset I_s$. We note that s and s' intersect as rectangles if and only if they are comparable under '$<$.'

We say that a collection of tri–tiles $\mathbf{T} \subset \mathbf{S}$ is a *tree with top t* if for every $s \in \mathbf{T}$, $s < t$. (The top need not be in the tree and tops are not unique.) It is easy to see that this partial order does not admit a cycle, so our use of the phrase tree conforms to common usage. We note that any collection $\mathbf{S}$ of tri–tiles is a union of trees. Simply let $\mathbf{S}^*$ denote those elements of $\mathbf{S}$ that are maximal under '$<$,' and for each $t \in \mathbf{S}^*$, let $\mathbf{T}_t$ be the maximal tree in $\mathbf{S}$ with top t. $\mathbf{S}$ is the union of the $\mathbf{T}_t$, $t \in \mathbf{S}^*$.

We refine the notion of a tree by saying that $\mathbf{T}$ is a j–tree if $\mathbf{T}$ is a tree with top t and $\rho_{sj} \cap \rho_t = \emptyset$ for all $s \ne t \in \mathbf{T}$. Under this condition, the functions $\{\phi_{sj} \mid s \in \mathbf{T}\}$ are orthogonal. If on the other hand we were to assume that $\rho_{sk} \cap \rho_t \ne \emptyset$ for all $s \in \mathbf{T}$, it follows from the grid structure that $\mathbf{T}$ is a j–tree for $j \ne k$. Now, for $s \ne t \in \mathbf{T}$, at most one of the three intervals ρ_{sk} can intersect ρ_t,

so an arbitrary tree is a union of at most three subtrees which are j–trees for two choices of j.

Trees have important analytic properties. The time intervals in a tree refine that of the top. The frequency intervals increase in length, but only at the rate dictated by uncertainty. And all the frequency intervals contain that of the top, hence the trees have localized the frequency variables in the transform. For a k–tree set

$$\Delta(T,k) := \Big[|R_t|^{-1} \sum_{s\in\mathbf{T}} |\langle f_k, \phi_{sk}\rangle|^2\Big]^{1/2},$$

Here, we specifically include the case of $k = 3$, as well as $k = 1, 2$, for ultimately we will form the inner product of $\mathcal{H}f_1f_2$ with a well–chosen third function f_3. For an arbitrary collection of tiles $\mathbf{S}$, we set

$$k\text{–size}(\mathbf{S}) = \sup\{\Delta(\mathbf{T},k) : \mathbf{T} \subset \mathbf{S} \text{ is a } k\text{–tree.}\}.$$

We note that a singleton $\mathbf{T} = \{t\}$ is a k–tree for all k. So k–size($\mathbf{S}$) dominates the terms $|R_s|^{-1/2}|\langle f_k, \phi_{sk}\rangle|$ for all $s \in \mathbf{S}$.

It should be noted, for we will relie upon this fact latter, that for any collection of tri–tiles $\mathbf{S}$, we have k–size($\mathbf{S}$) $\le K\|f_k\|_\infty$.

Now, fix $k = 1, 2$ or 3 and let $\mathbf{T} \subset \mathbf{S}$ be a k–tree for $j \neq k$ and let t denote the top of the tree. Then, by applying Cauchy–Schwarz,

$$|\langle \mathcal{H}^{\mathbf{T}} f_1 f_2, f_3\rangle| \le \sum_{s\in\mathbf{T}} \frac{|\langle f_j, \phi_{sj}\rangle|}{\sqrt{|R_s|}} \prod_{j\neq k} |\langle f_k, \phi_{sk}\rangle| \le |R_t| \prod_{j=1}^{3} k\text{–size}(\mathbf{S}). \tag{3.9}$$

This is the central estimate on a tree.

LEMMA 3.10. *Let $k = 1, 2$ or 3. Let f_k be a Schwartz function and $\mathbf{S}$ any collection of tri–tiles. Then $\mathbf{S}$ is a union of $\mathbf{S}_1$ and $\mathbf{S}_2$ which have these two properties. Let $\mathbf{S}_1^*$ denote the elements in $\mathbf{S}_1$ that are maximal under $<$. We have*

$$\sum_{t\in\mathbf{S}_1^*} |R_t| \le C_1 k\text{–size}(\mathbf{S})^{-2} \|f_k\|_2^2, \tag{3.11}$$

$$k\text{–size}(\mathbf{S}_2) \le \tfrac{1}{2}[k\text{–size}(\mathbf{S})]. \tag{3.12}$$

The constant C_1 depends only on C_0 in (3.6).

Note that the second collection $\mathbf{S}_2$ is better in that it has smaller size. But we have controlled the number of trees that are in $\mathbf{S}_1$ the first collection—this is the critical condition, the one that orthogonality gives us.

For the proof of the Lemma it suffices to consider the case of k–size($\mathbf{S}$) = 1. We construct $\mathbf{S}_1$ in two similar steps, with the construction being motivated by the particulars of the issues of orthogonality with which we conclude this argument.

It is required to distinguish between two types of k–trees. A k–tree $\mathbf{T}$ with top t will be called a *left tree* if for all $s \neq t \in \mathbf{T}$, ρ_{sk} lies to the left of ρ_t. Note

that for s, s' in a left tree $\mathbf{T}$, the relation $s < s'$ implies that ρ_{sk} lies to the left of $\rho_{s'k}$. A *right tree* has a corresponding definition.

We construct $\mathbf{S}_{1\ell} \subset \mathbf{S}$ as a union of trees $\tilde{\mathbf{T}}_l$ with tops $t(l)$, for integers $l \geq 1$. Each $\tilde{\mathbf{T}}_l$ will be associated to a left tree $\mathbf{T}_l$. The construction is inductive. We take $\mathbf{T}_1 \subset \mathbf{S}^1$ to be a left tree which satisfies several conditions. First, $\Delta(\mathbf{T}_1, k) \geq 1/4$. Second, $\mathbf{T}_1$ is maximal with respect to inclusion. Third, the top $t(1)$ is to be maximal with respect to '$<$.' Finally, $\rho_{t(1)}$ is to be left most—that is $\min\{\xi \mid \xi \in \rho_{t(1)}\}$ is minimal among the maximal tops. After selecting $\mathbf{T}_1$, we take $\tilde{\mathbf{T}}_1$ to be the maximal tree with top $t(1)$ in $\mathbf{S}$. We remove $\tilde{\mathbf{T}}_1$ from $\mathbf{S}$ and repeat this procedure until there is no left tree $\mathbf{T}_l \subset \mathbf{S}$ meeting these criteria. The union of the $\tilde{\mathbf{T}}_l$ is then $\mathbf{S}_{1\ell}$.

Under this construction, it is obvious that $\Delta(\mathbf{T}, k) < 1/4$ for all left trees $\mathbf{T} \subset \mathbf{S} - \mathbf{S}_{1\ell}$. Moreover—and this is the essential combinatorial observation—the $\mathbf{T}_l$ satisfy this disjointness condition.

$$\text{If } s \in \mathbf{T}_l,\ s' \in \mathbf{T}_{l'} \text{ and } \rho_{sk} \subset_{\neq} \rho_{s'k} \text{ then } R_{t(l)} \cap R_{s'} = \emptyset. \tag{3.13}$$

Indeed, the grid structure implies that $\rho_{t(l)} \subset \rho_s \subset \rho_{sk'}$, and so $\rho_{t(l)}$ lies to the left of $\rho_{t(l')}$. Thus, the tree $\mathbf{T}_l$ was constructed first. But then $s' \notin \tilde{\mathbf{T}}_l$, so that we must have $s' \not< t(l)$, which is to say $R_{t(l)} \cap R_{s'} = \emptyset$.

We verify that $\mathbf{S}_{1\ell}$ satisfies (3.11) momentarily.

We finally construct $\mathbf{S}_{1r}$ as a union of right trees in $\mathbf{S} - \mathbf{S}_{1\ell}$, with the obvious changes in the argument above. We conclude that this collection satisfies (3.11) as well. Then $\mathbf{S}_1$ is the union $\mathbf{S}_{1\ell} \cup \mathbf{S}_{1r}$ and $\mathbf{S}_2 := \mathbf{S} - \mathbf{S}_1$. For any left or right tree $\mathbf{T} \subset \mathbf{S}_2$, we have $\Delta(\mathbf{T}, k) < 1/4$ so that (3.12) follows.

We have still to establish (3.11), for the collections $\mathbf{S}_{1\ell}$ and $\mathbf{S}_{1r}$ defined above. This is the point that orthogonality enters the proof, and in fact we need only consider $\mathbf{S}_{1\ell}$. For the purposes of this argument, we suppose that $\mathbf{S} = \bigcup_l \mathbf{T}_l$, as constructed above. The properties of the $\mathbf{T}_l$ that we need are

$$\begin{aligned} \Delta(\mathbf{T}_l, k) &\geq 1/4 \quad \text{for all } l, \\ \frac{|\langle f_k, \phi_{sk}\rangle|}{\sqrt{|R_s|}} &\leq 1 \quad \text{for all } s \in \mathbf{S}, \end{aligned}$$

and the trees satisfy the disjointness condition (3.13). We demonstrate the inequality

$$\sum_{l=1}^{\infty} |R_{t(l)}| \leq K \|f_k\|_2^2. \tag{3.14}$$

For the proof of (3.14), it suffices to assume that $\mathbf{S}$ is finite, so that the number below is finite.

$$B := \Big\| \sum_{s \in \mathbf{S}} \langle f_k, \phi_{sk}\rangle \phi_{sk} \Big\|_2.$$

We show that $B \le K\|f_k\|_2$, which proves the Lemma as the next inequality shows.

$$\begin{aligned}\sum_{l=1}^{\infty} |R_{t(l)}| &\le \sum_{s\in\mathbf{S}} |\langle f_k, \phi_{sk}\rangle|^2 \\ &= \Big\langle f_k, \sum_{s\in\mathbf{S}} \langle f_k, \phi_{sk}\rangle \phi_{sk} \Big\rangle \\ &\le \|f_k\|_2 B.\end{aligned}$$

Note that we are exploiting the self–dual nature of the problem.

Now, we expand B^2, to get a diagonal term $\mathcal{D}$ and off–diagonal term $\mathcal{O}$. The diagonal term is $\mathcal{D} = \sum_{s\in\mathbf{S}} |\langle f_k, \phi_{sk}\rangle|^2$, which is no more than $B\|f_k\|_2$ as we have just seen. The off–diagonal term is $\mathcal{O} := 2\sum_{s\in\mathbf{S}} |\langle f_k, \phi_{sk}\rangle| \mathcal{O}(s)$, where

$$\mathcal{O}(s) := \sum_{s'\in\mathbf{S}(s)} |\langle \phi_{sk}, \phi_{s'k}\rangle| |\langle f_k, \phi_{s'k}\rangle|,$$

and we use the notation $\mathbf{S}(s) := \{s' \in \mathbf{S} \mid \rho_{sk} \subset_{\neq} \rho_{s'k}\}$. This is justified by (3.7) and the fact that the Fourier transform of ϕ_{sk} is supported on an affine image of ρ_{sk}. Hence the inner product of ϕ_{sk} and $\phi_{s'k}$ is zero unless ρ_{sk} and $\rho_{s'k}$ intersect. But then we can assume that one interval is contained in another due to the grid structure.

Note that by Cauchy–Schwarz, $\mathcal{O} \le 2B^{1/2}[\sum_{s\in\mathbf{S}} \mathcal{O}(s)^2]^{1/2}$ and we claim that for each tree $\mathbf{T}_l$

$$\sum_{s\in\mathbf{T}_l} \mathcal{O}(s)^2 \le K|R_{t(l)}|. \tag{3.15}$$

This will complete the proof of the Lemma, for we will then have $B^2 \le KB\|f\|_2$.

To see the claim, fix a tree top $t(l)$ and consider $s \in \mathbf{T}_l$. We observe that (3.6) implies that

$$|\langle \phi_{sk}, \phi_{s'k}\rangle| \le K \frac{\sqrt{|R_{s'}|}}{\sqrt{|R_s|}} (1 + \operatorname{dist}(R_s, R_{s'})|R_s|^{-1})^{-5}.$$

A detailed proof of the estimate is left to the reader, but note that the right hand side is only slightly bigger than $\|\phi_{s'k}\|_{L^1(R_{s'})} \|\phi_{sk}\|_{L^\infty(R_{s'})}$. But also note that (3.13) implies that every $s' \in \mathbf{S}(s)$ must satisfy $R_{t(l)} \cap R_{s'} = \emptyset$. Hence,

$$\begin{aligned}\mathcal{O}(s) &\le K|R_s|^{-1/2} \sum_{s'\in\mathbf{S}(s)} (1 + \operatorname{dist}(R_s, R_{s'})|R_s|^{-1})^{-5} |R_s| \\ &\le K|R_s|^{1/2} \int_{R_{t(l)}^c} (1 + \operatorname{dist}(R_s, x)|R_s|^{-1})^{-5} \frac{dx}{|R_s|} \\ &\le K|R_s|^{1/2} (1 + \operatorname{dist}(R_s, R_{t(l)}^c)|R_s|^{-1})^{-4}.\end{aligned}$$

But the tree structure imposes restrictions on the intervals R_s: They are in relation to $R_{t(l)}$ as the dyadic intervals $[j2^{-n}, (j+1)2^{-n}]$, $0 \le j < 2^n$ are to $[0,1]$. Thus, one sees that (3.15) holds. The proof is done.

3.2 Application of the Key Lemma

The Key Lemma, together with some considerations of a more familiar nature, give Lemma 3.8. It suffices to prove the inequality

$$|\{x \mid \mathcal{H}^{\mathbf{S}} f_1 f_2(x) \geq 2\}| \leq K, \tag{3.16}$$

for all f_1, f_2 of L^2 norm one and any collection of tri–tiles $\mathbf{S}$. K depends only on C_0 in (3.6). This is due to an invariance of the model operators under dilation. As this is a commonplace reduction, we omit the easy argument to pass from the inequality above to Lemma 3.8.

As f_1 and f_2 are Schwartz functions, $\mathbf{S}$ has finite 1 and 2 size, say no more than 2^{-n_0}. If $n_0 \geq 0$, there is nothing for us to do at this point. So assume that $n_0 < 0$. We iteratively apply Lemma 3.10 in the following fashion. Apply Lemma 3.10 for those $k = 1, 2$ for which k–size$(\mathbf{S}) \geq 2^{-n_0-1}$. We see that $\mathbf{S} = \mathbf{S}_{n_0} \cup \mathbf{S}^{n_0+1}$ where

$$\sum_{t \in \mathbf{S}^*_{n_0}} |R_t| \leq C 2^{2n_0} \quad \text{and} \quad k\text{–size}(\mathbf{S}^{n_0+1}) \leq 2^{-n_0-1},\ k = 1, 2.$$

The point that distinguishes the case $n_0 < 0$ is that the sum of the tops of the trees is less than a constant.

Continuing this procedure, we may write $\mathbf{S} = \bigcup_{n=-\infty}^{0} \mathbf{S}_n$, where k–size$(\mathbf{S}_n) \leq 2^{-n}$ for all $-\infty < n \leq 0$ and $k = 1, 2$. And for n strictly less than zero,

$$\sum_{t \in \mathbf{S}^*_n} |R_t| \leq C 2^{2n}, \quad n < 0.$$

There is no such control for $\mathbf{S}_0$, and we shall return to this collection momentarily.

Now for a fixed tree $\mathbf{T}$ with top t in $\mathbf{S}_n$, away from the interval $2R_t$, we have

$$\begin{aligned} |\mathcal{H}^{\mathbf{T}} f_1 f_2(x)| &\leq \sup_{s \in \mathbf{T}} \prod_{1}^{2} \frac{|\langle f_j, \phi_{sj} \rangle|}{\sqrt{|R_s|}} \sum_{s \in \mathbf{T}} \sqrt{|R_s|} |\phi_{s3}(x)| \\ &\leq K 2^{-2n} \Big(1 + \frac{\operatorname{dist}(x, R_t)}{|R_t|}\Big)^{-10}, \quad \text{if } x \notin 2R_t, \end{aligned}$$

as (3.6) and the tree structure easily imply. Thus, if we set $E_n = \bigcup_{t \in \mathbf{S}^*_n} 2^{-n} R_t$, then we have $|E_n| \leq K 2^n$. Thus, the union of these sets has bounded measure, so we need only estimate $\mathcal{H}^{\mathbf{S}_n}$ off of the set E_n, but then

$$\|\mathcal{H}^{\mathbf{S}_n} f_1 f_2\|_{L^1(E_n^c)} \leq K 2^{3n}.$$

Consequently, the collection $\mathbf{S}^0 = \bigcup_{n=-\infty}^{-1} \mathbf{S}_n$ satisfies (3.16).

Therefore, we can assume that the 1 and 2 size of $\mathbf{S}$ is at most 1. This is the case in which the Key Lemma is decisive. Set $G := \{x \mid \mathcal{H}^{\mathbf{S}} f_1 f_2(x) > 1/2\}$, which is a set of finite measure as we take f_1 and f_2 to be Schwartz functions. We can assume that $|G| > K'$ for otherwise there is nothing to prove. Then take f_3 to be a Schwartz functions approximating $|G|^{-1/2} \mathbb{I}_G$ so well that

$$|G|^{1/2} \leq 2 |\langle \mathcal{H}^{\mathbf{S}} f_1 f_2, f_3 \rangle|,$$

but in addition $|f_3(x)| \le 2|G|^{-1/2}$ for all x. It follows that the 3–size($\mathbf{S}$) can be assumed to be at most one, provided K' is large enough.

By applying Lemma 3.10 inductively, we can write $\mathbf{S}$ as $\bigcup_{n=0}^{\infty} \mathbf{S}_n$ where each $\mathbf{S}_n$ has k–size at most 2^{-n}, for $1 \le k \le 3$, and

$$\sum_{t \in \mathbf{S}_n^*} |R_t| \le C 2^{2n}.$$

But then by (3.9) it follows that

$$|\langle \mathcal{H}^{\mathbf{S}_n} f_1 f_2, f_3 \rangle| \le K 2^{-3n} \sum_{t \in \mathbf{S}_n^*} |R_t| \le K 2^{-n}.$$

And this is summable over $n > 0$, from which we conclude that $|G|^{1/2} \le K$. The proof is done.

References

[1] L. Carleson. "On convergence and growth of partial sums of Fourier series." *Acta Math.* 116 (1966) pp. 135-157.

[2] I. Daubechies. *Ten Lectures on Wavelets.* CBMS–NSF regional conference series in applied mathematics. SIAM (1992) Philadelphia.

[3] C. Fefferman. "Pointwise convergence of Fourier series." *Ann. Math.* 98 (1973) 551–571.

[4] M.T. Lacey, C.M. Thiele. "Bounds for the bilinear Hilbert transform on L^p." *Proc. Nat. Acad. Sci.* 94 (1997) 33—35.

[5] M.T. Lacey, C.M. Thiele. "On Calderòn's Conjecture." To appear in *Ann. Math.*

[6] M.T. Lacey, C.M. Thiele. "On Calderòn's Conjecture for the bilinear Hilbert transform." *Proc. Nat. Acad. Sci.* 95 (1998) 4828–4830.

[7] M.T. Lacey, C.M. Thiele. "L^p Bounds for the bilinear Hilbert transform, $p > 2$." Ann. Math. 146 (1997) 693—724.

Michael T. Lacey
School of Mathematics
Georgia Institute of Technology
Atlanta GA 30332
USA
email: lacey@math.gatech.edu
web: www.math.gatech.edu/~lacey

Rectifiability, Analytic Capacity, and Singular Integrals

Pertti Mattila

Abstract. This is a survey of some interplay between geometric measure theory (rectifiability), complex analysis (analytic capacity) and harmonic analysis (singular integrals). Vaguely, it deals with the following three principles:

1. The analytic capacity of a 1-dimensional compact subset of the complex plane $\mathbf{C}$ is zero if and only if E is purely unrectifiable.
2. The analytic capacity of a 1-dimensional compact subset E of $\mathbf{C}$ is positive if and only if the Cauchy singular integral operator is L^2-bounded on a large part of E.
3. Singular integrals behave nicely on an m-dimensional subset E of $\mathbf{R}^n$ if and only if E is in some sense rectifiable.

1991 Mathematics Subject Classification: Primary 28A75; Secondary 31A05, 42B20.
Keywords and Phrases: Analytic capacity, Cauchy integral, rectifiable set, Menger curvature.

1. Analytic capacity; finite length. First a general remark: since the list of complete references would be very long I have omitted many which the reader can find in [C1], [D1], [G] or [M2]. The analytic capacity of a compact subset E of $\mathbf{C}$ was defined by Ahlfors in 1947 as

$$\gamma(E) = \sup_f \lim_{z\to\infty} |zf(z)|$$

where the supremum is taken over all analytic functions $f\colon \mathbf{C}\setminus E \to \mathbf{C}$ such that $|f(z)| \le 1$ and $f(\infty) = 0$. Ahlfors showed that (see [G]) $\gamma(E) = 0$ if and only if E is removable for bounded analytic functions. That is, whenever U is an open set containing E, any bounded analytic function in $U\setminus E$ has an analytic extension to U. Or equivalently, the only bounded analytic functions in $\mathbf{C}\setminus E$ are constants. This is all very easy (but Ahlfors proved deep results about the existence of an extremal and its properties) and this characterization of removability is quite complex analytic. One would wish to find a geometric characterization. This is often called the Painlevé problem, since Painlevé started to study it about 100 years ago.

There are two very easy results, see [G] or [M2]: If the 1-dimensional Hausdorff measure $\mathcal{H}^1(E) = 0$, then $\gamma(E) = 0$. If the Hausdorff dimension of E $\dim E > 1$, then $\gamma(E) > 0$. Thus the following recent theorem of David [D2] leaves the question open only for sets E with $\mathcal{H}^1(E) = \infty$ and $\dim E = 1$ (but there is quite a variety of them).

1.1. THEOREM. *Let $E \subset \mathbf{C}$ be compact with $\mathcal{H}^1(E) < \infty$. Then $\gamma(E) = 0$ if and only if $\mathcal{H}^1(E \cap \Gamma) = 0$ for every rectifiable curve Γ.*

Sets E such that $\mathcal{H}^1(E \cap \Gamma) = 0$ for every rectifiable curve Γ are called purely unrectifiable according to Federer's terminology. Besicovitch studied their properties extensively in the 20's and 30's and called them irregular. They and their rectifiable (regular) counterparts and higher dimensional generalizations are quite basic in geometric measure theory.

I discuss the proof of Theorem 1.1, which also brings forth clearly the role of singular integrals. Suppose first that E is not purely unrectifiable. Then it meets some rectifiable curve in positive length. Some rather easy arguments show that it meets also some Lipschitz graph Γ with small Lipschitz constant in positive length. Calderòn showed in 1977 that the Cauchy singular integral operator C_Γ,

$$C_\Gamma g(z) = \lim_{\varepsilon \to 0} \int_{\Gamma \setminus B(z,\varepsilon)} \frac{g(\zeta)}{\zeta - z}\, d\mathcal{H}^1\zeta,$$

is bounded in $L^2(\Gamma)$ for such a Γ. (Later Coifman, McIntosh and Meyer showed that this is true for all Lipschitz graphs.) By that time it was already known, see [C1], for example, that then there is some bounded non-negative function h on Γ such that $f = C_\Gamma h$ is bounded in $\mathbf{C} \setminus E$. Thus $\gamma(E) > 0$.

The last step is based on a duality argument using the Hahn–Banach theorem and no constructive method of finding a non-constant bounded analytic function in $\mathbf{C} \setminus E$ is known even if Γ is C^1. If it is $C^{1+\varepsilon}$, then such a method exists, see [G].

Suppose then that $\gamma(E) > 0$. We should find a rectifiable curve Γ such that $\mathcal{H}^1(E \cap \Gamma) > 0$. First, there is a non-constant bounded analytic function f in $\mathbf{C} \setminus E$ vanishing at infinity, and an easy argument, see, e.g., [M2], using the Cauchy integral formula yields a bounded Borel function $\varphi\colon E \to \mathbf{C}$ such that $f = C_E\varphi$.

Let us assume that $\mathcal{H}^1\big(E \cap B(z,r)\big) \le Cr$ for $z \in \mathbf{C}$ and $r > 0$; this is not a really serious restriction. Then it is still easy to see that even the maximal function $C_E^*\varphi$,

$$C_E^*\varphi(z) = \sup_{\varepsilon > 0} \left| \int_{E \setminus B(z,\varepsilon)} \frac{\varphi(\zeta)}{\zeta - z}\, d\mathcal{H}^1\zeta \right|$$

is bounded in $\mathbf{C}$. Suppose we would be lucky enough to find φ so that it is also non-negative. Set $\mu = \varphi\mathcal{H}^1|E$. Using Fubini's theorem as Melnikov and Verdera

did in [MV] we get for all $\varepsilon > 0$,

$$\begin{aligned}\infty > C &\geq \int \left| \int_{\mathbf{C}\setminus B(z,\varepsilon)} \frac{1}{\zeta - z}\, d\mu\zeta \right|^2 d\mu z \\ &= \iiint\limits_{A_\varepsilon} \frac{1}{(z_1 - z_3)\,\overline{(z_2 - z_3)}}\, d\mu z_1\, d\mu z_2\, d\mu z_3 + O(1) \\ &= \frac{1}{6} \iiint\limits_{A_\varepsilon} \sum_\sigma \frac{1}{(z_{\sigma(1)} - z_{\sigma(3)})\,\overline{(z_{\sigma(2)} - z_{\sigma(3)})}}\, d\mu z_1\, d\mu z_2\, d\mu z_3 + O(1).\end{aligned}$$

Here σ runs through all six permutations of $\{1,2,3\}$ and $A_\varepsilon = \{(z_1,z_2,z_3) : |z_i - z_j| > \varepsilon \text{ for } i \neq j\}$. To get that the error term is bounded is an easy estimate using $\mu\big(B(z,r)\big) \leq Cr$ for all z, r. Now a remarkable identity found by Melnikov in [M] says that

$$\sum_\sigma \frac{1}{(z_{\sigma(1)} - z_{\sigma(3)})\,\overline{(z_{\sigma(2)} - z_{\sigma(3)})}} = c(z_1,z_2,z_3)^2$$

where $c(z_1,z_2,z_3)$ is the reciprocal of the radius of the circle passing through $z_1, z_2, z_3 \in \mathbf{C}$. This is 0 if and only if these points are collinear. The number $c(z_1,z_2,z_3)$ is called the Menger curvature of the triple (z_1,z_2,z_3). Menger introduced it in the early 30's to define the curvature for continua in compact, convex metric spaces, see [K]. Using the above formulas and letting $\varepsilon \to 0$, we obtain

$$c^2(\mu) \equiv \iiint c(x,y,z)^2\, d\mu x\, d\mu y\, d\mu z < \infty.$$

So now we have some geometric information about μ, and looking at it more closely we find that "most" (in μ-sense) triples of points which lie close to each other must be nearly collinear. This gives good hopes for a construction of rectifiable curves which carry positive μ measure.

The following theorem was first proved by David and then Legér [L] gave a different proof which also allows a higher dimensional version. Note that the situation is somewhat similar to that in Jones's traveling salesman result in [J].

1.2. THEOREM. *If $\mu = \varphi\mathcal{H}^1|E$, $\varphi \in L^\infty(E)$, $\varphi \geq 0$, $\mathcal{H}^1(E) < \infty$ and $c^2(\mu) < \infty$, then there are rectifiable curves Γ_i such that*

$$\mu\left(\mathbf{C} \setminus \bigcup_{i=1}^\infty \Gamma_i\right) = 0.$$

We have then that $\mathcal{H}^1(E \cap \Gamma_i) > 0$ for some i, and so E is not purely unrectifiable.

This would end the proof of Theorem 1.1 except that we have made the unjustified assumption that $\varphi \geq 0$. Note that in the above argument to get $c^2(\mu) < \infty$ we did not need the uniform boundedness of the Cauchy transform; we only needed the boundedness in L^2. Hence Theorem 1.1 follows if we can show:

1.3. THEOREM. *If there is a non-zero $\varphi \in L^\infty(E)$ such that $C_E^*\varphi$ is bounded, then there is $F \subset E$ such that $\mathcal{H}^1(F) > 0$ and the truncated operators $C_{F,\varepsilon}$;*

$$C_{F,\varepsilon}g(z) = \int_{F\setminus B(z,\varepsilon)} \frac{g(\zeta)}{\zeta - z}\, d\mathcal{H}^1\zeta,$$

are uniformly bounded in $L^2(\mathcal{H}^1|F)$.

Then we can use the constant function 1 to get $c^2(\mathcal{H}^1|F) < \infty$.

Theorem 1.3 follows from [DM] and [D2]. First φ was transformed to an accretive function ψ (i.e., $\operatorname{Re}\psi \geq \delta > 0$) with L^2-estimates for the Cauchy transform in [DM] with a construction relying on ideas of Christ from [C2], where Theorem 1.3 was proved for AD-regular sets (see Section 3). Then David proved a general $T(b)$-theorem in [D2] yielding Theorem 1.3. A little later Nazarov, Treil and Volberg gave in [NTV3] a different simpler proof for Theorem 1.3 also obtaining a general $T(b)$-theorem.

The problem of removable sets for Lipschitz harmonic functions is very much like that for bounded analytic functions. Theorem 1.1 is valid also in this case, but we don't know if the two classes of removable sets are exactly the same. The reason for this similarity is that rather than studying bounded harmonic functions we are studying harmonic functions with bounded gradient and the gradient of the fundamental solution $c\log|z|$ is essentially the Cauchy kernel.

This problem is interesting also in $\mathbf{R}^n$. There are several partial results, see [MP], but nothing like the analog of Theorem 1.1, even for $(n-1)$-dimensional AD-regular sets. Now the kernel is $|x|^{-n}x$, $x \in \mathbf{R}^n$, but we don't know anything useful to replace Melnikov's identity with.

2. ANALYTIC CAPACITY; INFINITE LENGTH. If $\mathcal{H}^1(E) < \infty$, then by a result of Besicovitch E is purely unrectifiable if and only if

$$\mathcal{H}^1\big(p_\theta(E)\big) = 0 \quad \text{for almost all } \theta \in [0,\pi), \tag{2.1}$$

where p_θ is the orthogonal projection onto the line making angle θ with the real axis. Vitushkin conjectured in the 60's that (2.1) would be equivalent to $\gamma(E) = 0$ for all compact sets $E \subset \mathbf{C}$. Thus Theorem 1.1 says that he was right when $\mathcal{H}^1(E) < \infty$. The general conjecture was shown to be false in [M1] where it was shown that (2.1) is not conformally invariant, but this did not say which of the two implications is false. In [JM] Jones and Murai gave a concrete example where (2.1) holds but $\gamma(E) > 0$. It is not known if $\gamma(E) = 0$ implies (2.1). Now Melnikov has a new conjecture:

2.2. CONJECTURE. *For any compact $E \subset \mathbf{C}$, $\gamma(E) > 0$ if and only if there is a (non-negative) Radon measure μ on E such that $\mu(E) > 0$, $\mu\big(B(z,r)\big) \leq r$ for all $z \in \mathbf{C}$, $r > 0$, and $c^2(\mu) < \infty$.*

Melnikov proved in [M] that the "if part" of this conjecture is true. In fact, he proved the quantitative estimate

$$\gamma(E) \geq C\frac{\mu(\mathbf{C})^{3/2}}{\big(\mu(\mathbf{C}) + c^2(\mu)\big)^{1/2}},$$

if μ is as in Conjecture 2.2.

Using this result of Melnikov, Joyce and Mörtes have given in [JoM] another example with simpler arguments where (2.1) holds but $\gamma(E) > 0$.

Conjecture 2.2 is not known even for non-degenerate continua; then the analytic capacity is positive. Another test case, which is open, is given by the Cantor sets of Garnett in [G, p. 87]. There is a mistake in [G] and the characterization for $\gamma(E) > 0$ given in Theorem 2.2 is not correct, see [M3] and [E] for this and some related results.

3. Singular integrals on regular sets. A compact subset E of $\mathbf{C}$ is called AD-regular (Ahlfors–David) if there exists a positive number C such that

$$r/C \le \mathcal{H}^1\big(E \cap B(z,r)\big) \le Cr \quad \text{for } z \in E,\ 0 < r < 1.$$

The following theorem was proved in [MMV] using the above relations between the Cauchy kernel and Menger curvature. This also gave Theorem 1.1 for AD-regular sets. Some generalizations, but still partial results of Theorem 1.1, were given by Lin [Li] (doubling condition) and Pajot [P] (positive lower density).

3.1. Theorem. *Let $E \subset \mathbf{C}$ be AD-regular. The truncated operators $C_{E,\varepsilon}$ are uniformly bounded in $L^2(\mathcal{H}^1|E)$ if and only if E is uniformly rectifiable, that is, there is an AD-regular curve containing E.*

The uniform L^2-boundedness of $C_{E,\varepsilon}$ is equivalent to the boundedness of the principal value operator C_E, if we know that the principal values exist almost everywhere for a dense set of functions. But we don't know this a priori, hence the above formulation. This is also equivalent to the boundedness of the maximal operator C_E^*.

It is obvious what the AD-regularity means for m-dimensional subsets of $\mathbf{R}^n$; we just replace r by r^m and $\mathcal{H}^1$ by the m-dimensional Hausdorff measure $\mathcal{H}^m$. It is less obvious what the uniform rectifiability should mean if $m > 1$, but David and Semmes have shown that there exist several natural equivalent definitions and they have developed an extensive theory of such sets, see [DS]. They have also studied singular integrals on them and shown that they are bounded in $L^2(\mathcal{H}^m|E)$ for a large class of Calderón–Zygmund kernels. The converse is also valid, i.e., L^2-boundedness implies uniform rectifiability, if one assumes the L^2-boundedness for the operators related to all kernels of the type $\varphi(|x|)\,|x|^{-m-1}x$, $x \in \mathbf{R}^n$, where φ is a smooth non-negative function. However, it is not known if the converse is valid if one only uses one single kernel, for example, the Riesz kernel $K_m(x) = |x|^{-m-1}x$. The problem is again that we don't have anything like the curvature identity. Farag [F] has looked at different ways of forming sums of permutations starting from K_m, but all of them take both positive and negative values and are thus difficult to use.

We can also ask if results like Theorem 3.1 hold for other kernels in the plane. The same method works for the real and imaginary parts of the Cauchy kernel, for example, but I don't know any other essentially different kernel for which this, or some other, method would work. Joyce has looked at the kernels $|z|^{-2k}\,z^{2k-1}$, $k = 1, 2, \ldots$, and again found for the sum of permutations both positive and negative values, when $k > 1$.

One can also study m-regular sets E for non-integral m, but Vihtilä showed in [Vi] that then the singular integral operator related to K_m is never bounded in $L^2(\mathcal{H}^m|E)$.

4. EXISTENCE OF PRINCIPAL VALUES. In the previous section we saw that the L^2-boundedness of the singular integral operators is often equivalent to uniform rectifiability. For non-uniform rectifiability there are characterizations with the existence of principal values. A subset E of $\mathbf{R}^n$ is called m-rectifiable if there are C^1 (or, equivalently, Lipschitz) m-dimensional surfaces S_i such that

$$\mathcal{H}^m\left(E \setminus \bigcup_{i=1}^{\infty} S_i\right) = 0.$$

4.1. THEOREM. *Let $E \subset \mathbf{C}$ be $\mathcal{H}^1$ measurable with $\mathcal{H}^1(E) < \infty$. Then E is 1-rectifiable if and only if*

$$\lim_{\varepsilon \to 0} \int_{E \setminus B(z,\varepsilon)} \frac{1}{\zeta - z}\, d\mathcal{H}^1\zeta$$

exists for $\mathcal{H}^1$ almost all $z \in E$.

The fact that the existence of principal values implies rectifiability was proved by Tolsa in [T3] using results of Nazarov, Treil and Volberg from [NTV3] and the curvature method. Hence this is restricted to the Cauchy kernel and 1-dimensional sets. It is not known if the analogue of Theorem 4.1 holds for m-dimensional sets if $m \geq 2$. The existence of principal values was proved in [MM]. Verdera gave a different proof in [V] which also works in general dimensions (see also [M2]). With an extra condition on positive lower density we have the following, see [MPr] or [M2].

4.2. THEOREM. *Let $E \subset \mathbf{R}^n$ be $\mathcal{H}^m$ measurable with $\mathcal{H}^m(E) < \infty$. Then E is m-rectifiable if and only if for $\mathcal{H}^m$ almost all $x \in E$,*

$$\liminf_{r \to 0} r^{-m} \mathcal{H}^m\big(E \cap B(x,r)\big) > 0$$

and

$$\lim_{\varepsilon \to 0} \int_{E \setminus B(x,\varepsilon)} \frac{x - y}{|x - y|^{m+1}}\, d\mathcal{H}^m y$$

exists.

Huovinen proved in [H] a result analogous to Theorem 4.2 for some other kernels in $\mathbf{C}$.

Tolsa has given in [T2] a complete geometric characterization, involving curvature, of those Radon measures μ on $\mathbf{C}$ for which

$$C_\nu(z) = \lim_{\varepsilon \to 0} \int_{\mathbf{C} \setminus B(z,\varepsilon)} \frac{1}{\zeta - z}\, d\nu\zeta$$

exists for μ almost all z for all Radon measures ν in $\mathbf{C}$.

There is another very nice result in [T2]: if the Cauchy operator $g \mapsto (1/z) * (g d\mu)$ is bounded in $L^2(\mu)$ (meaning again the uniform boundedness of the truncated operators), then the principal values $C_\mu(z)$ exist for μ almost all $z \in \mathbf{C}$. This is again known (essentially) only for the Cauchy kernel, since the proof uses curvature.

5. Calderón–Zygmund theory in non-homogeneous spaces. We have already mentioned several times the works of Nazarov, Treil and Volberg [NTV1–3] and Tolsa [T1–3]. Their starting point was the following question. Let μ be a Radon measure in $\mathbf{R}^n$. If μ is doubling; $\mu\big(B(x,2r)\big) \leq C\mu\big(B(x,r)\big)$ for all $x \in \operatorname{spt}\mu$, $r > 0$ (or, in other words, $(\operatorname{spt}\mu, \mu)$ is a space of homogeneous type), most of the Calderón–Zygmund theory of singular integrals is valid. Surprisingly, the works mentioned above show that almost always the doubling condition is not needed at all. Tolsa uses the curvature method, and this is again restricted to the Cauchy kernel. Nazarov, Treil and Volberg have developed a beautiful method using random lattices of dyadic cubes and showing that with a large probability such a lattice is in a good position in order that useful estimates can be established. This works for general Calderón–Zygmund kernels in $\mathbf{R}^n$. Then one obtains in great generality such basic results as the equivalence of the L^2-boundedness to the L^p-boundedness for $1 < p < \infty$ and to the weak L^1-boundedness, Cotlar's inequality, $T(1)$- and $T(b)$-theorems.

The $T(b)$-theorem for singular integral operators T such as the Cauchy operator says that if there exists $b \in L^\infty(\mu)$ such that $\operatorname{Re} b \geq \delta > 0$ (this can be replaced with weaker conditions) and $T(b) \in \mathrm{BMO}(\mu)$, then T is bounded in $L^2(\mu)$. The first such theorem without any doubling condition was proved by David in [D2]; this was the last missing piece in the proof of Theorem 1.1. There is some difference with David's $T(b)$-theorem and that of Nazarov, Treil and Volberg since David is defining BMO with generalized "dyadic cubes" which depend on the measure μ.

References

[C1] M. Christ, *Lectures on Singular Integral Operators*, Regional Conference Series in Mathematics 77, Amer. Math. Soc., 1990.

[C2] M. Christ, *A $T(b)$ theorem with remarks on analytic capacity and the Cauchy integral*, Colloq. Math. **60/61** (1990), 601–628.

[D1] G. David, *Wavelets and Singular Integrals on Curves and Surfaces*, Lecture Notes in Math. 1465, Springer-Verlag, 1991.

[D2] G. David, *Unrectifiable 1-sets have vanishing analytic capacity*, to appear in Rev. Mat. Iberoamericana.

[DM] G. David and P. Mattila, *Removable sets for Lipschitz harmonic functions in the plane*, preprint.

[DS] G. David and S. Semmes, *Analysis of and on Uniformly Rectifiable Sets*, Mathematical Surveys and Monographs 38, Amer. Math. Soc., 1993.

[E] V. Eiderman, *Hausdorff measure and capacity associated with Cauchy potentials*, Mat. Zametki **6** (1998).

[F] H. Farag, *The Riesz kernels do not give rise to higher dimensional analogues of the Menger–Melnikov curvature*, preprint.

[G] J. Garnett, *Analytic Capacity and Measure*, Lecture Notes in Math. 297, Springer-Verlag, 1972.

[H] P. Huovinen, *Existence of singular integrals and rectifiability of measures in the plane*, Ann. Acad. Sci. Fenn. Ser. A I Math. Dissertationes 109, 1997.

[J] P. W. Jones, *Rectifiable sets and the traveling salesman problem*, Invent. Math. **102** (1990), 1–15.

[JM] P. W. Jones and T. Murai, *Positive analytic capacity but zero Buffon needle probability*, Pacific J. Math. **133** (1988), 99–114.

[JoM] H. Joyce and P. Mörters, *A set with finite curvature and projections of zero length*, preprint.

[K] S. Kass, *Karl Menger*, Notices Amer. Math. Soc. **43:5** (1996), 558–561.

[L] J.-C. Léger, *Menger curvature and rectifiability*, preprint.

[Li] Y. Lin, *Menger curvature, singular integrals and analytic capacity*, Ann. Acad. Sci. Fenn. Ser. A I Math. Dissertationes 111, 1997.

[M1] P. Mattila, *Smooth maps, null-sets for integralgeometric measure and analytic capacity*, Ann. of Math. **123** (1986), 303–309.

[M2] P. Mattila, *Geometry of Sets and Measures in Euclidean Spaces*, Cambridge Studies in Advanced Mathematics 44, Cambridge University Press, 1995.

[M3] P. Mattila, *On the analytic capacity and curvature of some Cantor sets with non-σ-finite length*, Publ. Mat. **40** (1996), 195–204.

[MM] P. Mattila and M. S. Melnikov, *Existence and weak type inequalities for Cauchy integrals of general measures on rectifiable curves and sets*, Proc. Amer. Math. Soc. **120** (1994), 143–149.

[MMV] P. Mattila, M. S. Melnikov and J. Verdera, *The Cauchy integral, analytic capacity, and uniform rectifiability*, Ann. of Math. **144** (1996), 127–136.

[MP] P. Mattila and P. V. Paramonov, *On geometric properties of harmonic* Lip_1*-capacity*, Pacific J. Math. **171** (1995), 469–491.

[MPr] P. Mattila and D. Preiss, *Rectifiable measures in* $\mathbf{R}^n$ *and existence of principal values of singular integrals*, J. London Math. Soc. (2) **52** (1995), 482–496.

[M] M. S. Melnikov, *Analytic capacity: discrete approach and curvature of measure*, Sbornik Mathematics **186** (1995), 827–846.

[MV] M. S. Melnikov and J. Verdera, *A geometric proof of the L^2 boundedness of the Cauchy integral on Lipschitz graphs*, Internat. Math. Res. Notices **7** (1995), 325–331.

[NTV1] F. Nazarov, S. Treil and A. Volberg, *Cauchy integral and Calderón–Zygmund operators on nonhomogeneous spaces*, Internat. Math. Res. Notices **15** (1997), 703–726.

[NTV2] F. Nazarov, S. Treil and A. Volberg, *Weak type estimates and Cotlar's inequality for Calderón–Zygmund operators on nonhomogeneous spaces*, to appear in Internat. Math. Res. Notices.

[NTV3] F. Nazarov, S. Treil and A. Volberg, *Perfect hair*, preprint.

[P] H. Pajot, *Conditions quantitatives de rectifiabilité*, Bull. Soc. Math. France **125** (1997), 1–39.

[T1] X. Tolsa, *L^2-boundedness of the Cauchy integral operator for continuous measures*, to appear in Duke Math. J.

[T2] X. Tolsa, *Cotlar's inequality and existence of principal values for the Cauchy integral without the doubling condition*, to appear in J. Reine Angew. Math.

[T3] X. Tolsa, *Curvature of measures, Cauchy singular integral and analytic capacity*, Ph.D. Thesis, Universitat Autonoma de Barcelona (1998).

[V] J. Verdera, *A weak type inequality for Cauchy transforms of measures*, Publ. Mat. **36** (1992), 1029–1034.

[Vi] M. Vihtilä, *The boundedness of Riesz s-transforms of measures in* $\mathbf{R}^n$, Proc. Amer. Math. Soc. **124** (1996), 3797–3804.

Pertti Mattila
University of Jyväskylä
Department of Mathematics
P.O. Box 35
FIN-40351 Jyväskylä, Finland
pmattila@jylk.jyu.fi

Randomness and Pattern in Convex Geometric Analysis

Vitali Milman[1]

1991 Mathematics Subject Classification: 46B, 52A

0. This text can be complemented by the survey [M96] where surprising geometric phenomena observed in high dimensional spaces are described. The presentation there is more geometric with the emphasis on convex asymptotic geometry. In this talk we try to understand the reasons behind these very unusual geometric phenomena. A perceived random nature of high dimensional spaces we observe is at the root of the reasons I will discuss in the talk and the patterns it produces create the unusual phenomena we observe.

A more technical description of the results of the Asymptotic Theory of Finite Dimensional Normed Spaces up to 1986 can be found in [M86]. The following surveys and books may complete the picture in the direction of Local Theory: [MS86], [P89], [TJ88], [LM93], [M92]. For a description of the Concentration Phenomenon technique and its applications to Functional Analysis, Probability and Discrete Mathematics, see [MS86], [M88a], [T95], [T96], [LT91], [AlSp92].

In the dictionary, "randomness" is exactly the opposite of "pattern". Randomness means "no pattern". But, in fact, objects created by independent identically distributed random processes, being different, are in a sense, most undistinguishable and similar in the statistical sense. It is a challenge to discover these similarities, a pattern, in very different looking objects. We will do this on the example of convex bodies and normed spaces of high dimension. In fact, when we discover very similar patterns in arbitrary, and apparently very diverse convex bodies or normed spaces of high dimension we interpret them as a manifestation of the randomness principle mentioned above.

1. We demonstrate one such pattern through the following theorem. We first put it in a non-precise "meta" form: *for every convex compact body $K \subset \mathbb{R}^n$ there corresponds an ellipsoid $\mathcal{E}_K$ of the same volume* ($\operatorname{vol} K = \operatorname{vol} \mathcal{E}_K$) *and with the same barycenter – "a pattern" – which represents K in many respects.*

To put this in an exact form we will need some notation.

Let $X = (\mathbb{R}^n, \|\cdot\|, |\cdot|)$ be a normed space equipped with a norm $\|\cdot\|$ and the standard euclidean norm $|\cdot|$. Let D be the standard euclidean ball and $K(= K_X)$ be the unit ball of the normed space $(X, \|\cdot\|)$. We write $|A|$ for the volume of the set A. We call the family of convex bodies $\{uK \mid u \in SL_n\}$ associated with K the family of its *positions*. We have two parallel languages to describe the same

[1]Partially supported by a Binational US-Israel Science Foundation Grant.

results. On one hand, we construct some special ellipsoid, say $\mathcal{E}$, which represents the body K (in a sense which will be specified later), but on the other hand, we may change the position of K and consider $\widehat{K} = uK$, $u \in SL_n$, where u is chosen such that $u\mathcal{E} = \lambda D$ ($\lambda :=$vol.rad. $\mathcal{E} = (|\mathcal{E}|/|D|)^{1/n}$ is the volume radius of $\mathcal{E}$). Then the euclidean ball λD now represents $\widehat{K}$; however this position of K is a specially chosen position and our "pattern" is shifted from a "special ellipsoid" to a "special position". Below, we prefer the language of *positions*.

THEOREM 1. *$\exists C$ s.t. $\forall n$ and any four convex bodies K_i, $i = 1, \dots, 4$, of volume radius 1, i.e. $|K_i| = |D|$, and with 0 being the centroid of K_i, the following is true: there are positions $\widehat{K}_i = u_i K$, $u_i \in SL_n$, for every i, and a couple of orthogonal operators $\{v_1, v_2\} \subset O(n)$ so that the body*

$$Q = \mathrm{Conv}\big[(\widehat{K}_1 \cap v_1\widehat{K}_2) \cup v_2(\widehat{K}_3 \cap v_1\widehat{K}_4)\big]$$

is C-close to the euclidean ball D, i.e. $D/\sqrt{C} \subset Q \subset \sqrt{C}D$. Moreover, (i) the probability that a randomly chosen couple $\{v_1; v_2\} \subset O(n) \times O(n)$ satisfies the theorem is very high; it is larger than $1 - 1/2^n$ (this is the reason we will call such a couple "a random couple"); (ii) for any $v \in O(n)$

$$\text{vol. rad.}(\widehat{K}_1 \cap v\widehat{K}_2) \geq \tfrac{1}{\sqrt{C}} \quad \text{and vol. rad. } \mathrm{Conv}(\widehat{K}_1 \cup v\widehat{K}_2) \leq \sqrt{C}\,.$$

(We may say that ellipsoids $\mathcal{E}_i = u_i^{-1}D$ represent "essential" symmetries of K, but only in an "isomorphic" sense, and not in the "isometric" one as it is usual in geometry.)

(This Theorem was proved by the author in the centrally-symmetric case; see [M96] for references. For an extension to the general case, see [MP98].)

2. To continue with examples of very "regular" asymptotic behavior of an arbitrary high dimensional space we need more notation. As before, let a normed space $X = (\mathbb{R}^n, \|\cdot\|)$ be equipped with the euclidean norm $|\cdot|$. Denote $b = \|Id : (\mathbb{R}^n, |\cdot|) \to (\mathbb{R}^n, \|\cdot\|)\|$ and $a = \frac{1}{2}\operatorname{Diam} K_X$. So, $\frac{1}{a}|x| \leq \|x\| \leq b|x|$. The dual norm $\|x\|^* = \sup_{y \neq 0} \frac{|(x,y)|}{\|y\|}$ is naturally defined and then $b = \frac{1}{2}\operatorname{Diam} K^0$ where the polar body $K^0 = K_{X^*}$, X^* is the dual space to X. Let $M \equiv \int_{S^{n-1}} \|x\| d\mu(x)$, $S^{n-1} = \partial D$ be the unit euclidean sphere and $\mu(x)$ be the probability rotation invariant measure on S^{n-1}. Similarly, M^* is the expectation of $\|x\|^*$ on the sphere S^{n-1}, i.e. $M^* = \int_{S^{n-1}} \|x\|^* d\mu(x)$. There is the natural geometric meaning of M^* as being half of the mean width of K_X.

We will show below that these four numbers: a, b, M and M^*, uniquely describe (but again in an "isomorphic" sense) many geometric and analytic properties of the space X (and its unit ball K_X). Some of these properties are quantitively described by the following parameters:

$$k(X) = \max\big\{k \mid \mu_{G_{n,k}}\big\{E \in G_{n,k} \;\big|\; \tfrac{1}{2}M|x| \leq \|x\| \leq 2M|x|, \text{ for } \forall x \in E\big\} > 1 - \tfrac{k}{n+k}\big\},$$

where $\mu_{G_{n,k}}$ in the formula is the Haar probability measure on the Grassmannian manifold $G_{n,k}$ of all k-dimensional subspaces of n-dimensional space $\mathbb{R}^n$,

$$t(X) = \min\big\{t \mid \exists u_i \in O(n) \quad \text{and} \quad \tfrac{1}{2}M \cdot |x| \leq \tfrac{1}{t}\textstyle\sum_1^t \|u_i x\| \leq 2M|x|\big\}\,.$$

So, $k(X)$ is a "local" parameter, meaning it describes the behavior of the subspaces of a space which belongs to a set of properties we call "the local structure", and $t(X)$ is a "global" parameter because it relates to a property of the whole space. Let us also agree to write $f \sim \varphi$ when there are two universal constants (independent of anything) c_1 and c_2 and $c_1\varphi \le f \le c_2\varphi$. So the two quantities φ and f are *uniformly* (universally) equivalent.

Theorem 2. (i) ([M71]; [MS97]) $k(X) \sim n(\frac{M}{b})^2$; (ii) [(BLM88]; [MS97]) $t(X) \sim (\frac{b}{M})^2$. *Therefore, these local and global parameters are related in a very precise form:* $k(X) \cdot t(X) \sim n$ *([MS97]).*

A few comments and interpretations:

(i) For any operator $A : \ell_2^n \to X$ we may similarly introduce $M(A) = \int_{S^{n-1}} \|Ax\| d\mu(x)$ and $k(A)$ (putting $\|Ax\|$ instead of $\|x\|$ in the definition of $k(X)$). Then (i) may be rewritten in the form $\|A\| \sim M(A)\sqrt{n/k(A)}$. Here $\|A\|$ is the standard operator norm of operator A and this gives an asymptotic formula for the operator norm through the average and some geometric parameters related to the operator A.

(ii) Considering the dual space X^* we have, of course, $k^* \equiv k(X^*) \sim n(M^*/a)^2$, meaning that a "random" orthogonal projection $P_E K$ onto a subspace E of dimension k^*, looks, up to a factor 4, like a euclidean ball: $\frac{1}{2}M^* \cdot D(E) \subset P_E K \subset 2M^* \cdot D(E)$. Furthermore, for any integer $n \ge k \ge k^*$ and for a "random" subspace E, dim $E = k$,

$$\text{Diam } P_E K \sim \text{Diam } K \cdot \sqrt{k/n}$$

and $P_E K \sim M^* D(E)$ for $k \le k^*$ (in particular, Diam $P_E K$ is stabilized on $2M^*$).

So, we observe the regular decay (by a factor $\sqrt{k/n}$) of the diameter of a "random" k-dimensional projection of K till stabilization when this projection becomes almost a euclidean ball itself, and this fact is true for *any* convex centrally symmetric body – another pattern of behavior. It also provides us with an example of "phase transition" - a typical asymptotic phenomenon as we will see also later.

For quite a long time, we have known how to write very precise estimates, reflecting different asymptotic behavior of high dimensional normed spaces. Usually, we knew that these estimates are exact on some important subclasses of spaces. However, the new "message", based on many recent results, indicates that, in fact, available estimates are exact for every sequence of spaces of increasing dimension (we can say, "for every individual space"). We call such exact estimates "asymptotic formulas".

In the next three sections we will demonstrate more asymptotic formulas, each of which represents a specific pattern of behavior of an arbitrary high dimensional normed space. We would like to emphasize that it is less important in this presentation how these formulas look. The central issue is that such asymptotic formulas do exist and are applicable to any norm, that very little information on a norm (or a convex body) implies deep understanding of a complicated behavior of these normed spaces.

3. Can we also describe how the ball $M^* D(E)$ is "filled" by random projections from the inside? A clear pattern of behavior is seen again in asymptotic formulas

for the radius of the largest ball inscribed into the random projection $P_E K$ for $\dim E = k,\ k \gg k^*$. We compute it in the dual form. This means that we compute (estimate) the diameter of a random k-dimensional *section* of the polar body K^0. There is a well-known and useful fact, the so-called Low M^*-estimate (see [M85], [PT86], [Gor88]), which gives a simply formulated upper bound for such sections. However, it is not exact and is far from being the asymptotic formula we are interested in. To perceive the kind of result that should be expected here, I will mention one particular fact from [GM97a]: *Let $k = [n/2]$ and r be the solution of the following equation: $M^*(K \cap rD) = \frac{1}{2}r$ (the unique solution always exists); then the diameter of a random k-dimensional central section of K is less than $2r$. On the other hand, solve the equation $M^*(K \cap r_1 D) = (1 - \frac{1}{48.36})r_1$; a random k-dimensional section of K has diameter greater than $\frac{1}{60}r_1$.*

There is a more precise form of answer which requires deeper information on the body K but is still easily computable (I am now taking a Computational Geometry point of view). Define the following functions: for $k = \lambda n$, $0 < \lambda < 1$,

$$S_K^*(\lambda) = \int_{E \in G_{n,k}} M^*(K \cap E) d\mu(E)\,, \text{ and } D_K(\lambda) = \tfrac{1}{2} \int_{E \in G_{n,k}} \operatorname{diam}(K \cap E) d\mu(E)\,.$$

THEOREM 3 ([GM98a]). *Let $\frac{1}{b}D \subset K \subset aD$ and $ab \leq n^t$ (the non-degeneracy condition). Then $\forall \lambda \in (0,1)$*

$$S_K^*(\lambda) \leq D_K(\lambda) \leq c' S_K^*(\lambda_1) / \sqrt{1 - \lambda_2}$$

for $\lambda = \lambda_1 \lambda_2$ (and $\lambda_1 - \lambda \geq c'' t \log n/n$) and c', c'' two universal constants.

4. We will return to these asymptotic formulas but let us now continue our search for patterns of asymptotically "similar behavior" of any convex set in $\mathbb{R}^n$. We will now study (following [LMS98]) the geometric structure of the level sets $K \cap rS^{n-1} = A(r)$ and will see that, from a point of view we put forward below, these sets in some interval of values of "r" appear very similar. Define

$$r_t = \min\left\{ \tfrac{1}{2} \operatorname{Diam} \bigcap_1^t u_i K \mid u_i \in O(n) \right\},$$

and also the inverse function $T(r) = \min\left\{ t \mid \exists u_i \in O(n) \text{ and } \bigcap_1^t u_i K \subset rD \right\}$. (So, $T(r_t) = t$.) Of course, the meaning of $T(r)$ is that there is a covering of rS^{n-1} by $T(r)$ rotations of $rS^{n-1} \backslash A(r)$ and there is no covering with a smaller number of rotations. Again, in the interval $2/b \leq r \leq 1/2M$ the function $T(r)$ is exactly described [LMS98]: $\log T(r) \simeq n/r^2 b^2$, although under some kind of non-degeneracy condition: $br \lesssim \sqrt{n/\log n}$ (just note that always $br \leq b/2M \lesssim \sqrt{n}$).

The exponential behavior of $T(r)$ for $r \leq 1/2M$ (and any fixed number $\lambda > 1$ may be substituted for 2) changes to "polynomial" around level $1/M$: Let $T \sim (\frac{b}{M})^2 \cdot \frac{1}{\varepsilon^2}$; then, (i) for a random choice of $\{u_i\}_1^T \subset O(n)$, $\cap u_i K \subset \frac{1+\varepsilon}{M} D$, but, (ii) for *any* choice $\{u_i\}_1^T \subset O(n)$, the intersection $\cap u_i K \not\subseteq \frac{1}{(1+\varepsilon)M} D$.

Of course, not for every spherical level r do different convex bodies look similar. Consider, for example, the unit balls of ℓ_∞^n and ℓ_1^n (the cube and the

cross-polytope) normalized so that they are inscribed in the euclidean ball D of the same radius (say 1). Then the contact points with the sphere are in the first case 2^n and in the second, only $2n$. Naturally, for $r < 1$ but close to 1, the level sets are completely different. So, on what level does this phenomenon of similarity of spherical level sets start? Naturally, in this language the maximal such expected level cannot be above r_2. So, can r_2 be described by very little "statistical" information about K? The answer is "Yes":

Theorem 4 ([GM97b]). (i) $r_2 \le \sqrt{2} D_K(1/2)$ *(we introduced the average diameter* $D_K(\lambda)$ *above);* (ii) *there are universal numbers* $C > 1$ *and* $0 < c < 1$ *such that* $D_K(c) \le C \cdot r_2$.

I would like to recall that we also saw that $D_K(\lambda)$ is well described by the well computable function $S_K^*(\lambda)$.

5. Much more delicate analytic information about the level sets for $r < 1/M$ (and even slightly above this level) may, in fact, be provided in another language.

Let $M_q = \left(\int_{S^{n-1}} \|x\|^q d\mu(x) \right)^{1/q}$, $q \ge 1$, and let

$$t_q(X) = \min \left\{ t \mid \exists \{u_i\}_1^t \subset O(n) \text{ such that } \frac{1}{2} M_q |x| \le \left(\frac{1}{t} \sum_1^t \|u_i x\|^q \right)^{1/q} \le 2 M_q |x| \right\}.$$

(Note, that the information on the level sets is obtained by choosing q such that $r = 1/M_q$.) Then we again have asymptotic formulas describing the behavior of M_q and t_q.

Theorem 5 ([LMS98]). (i) $M_q \sim M_1$ *for* $1 \le q \le k(X) \sim n(M/b)^2$, $M_q \sim b\sqrt{q/n}$ *for* $k(X) \le q \le n$ *and* $M_q \sim b$ *for* $q \ge n$. *(Note again a "phase transition").*

(ii) $t_q \sim t_1$ $(= t(X) \sim (b/M)^2)$ *for* $1 \le q \le 2$, $t_q^{2/q} \sim t_1 (M_1/M_q)^2$ *for* $2 \le q$; *again a phase transition. However, because also* M_q *has its phase transition, we have two phase transitions for the function* t_q *on the interval* $1 \le q \le n$: $t_q \sim (b/M)^2$ *for* $1 \le q \le 2$, $t_q^{2/q} \sim (b/M)^2$ *for* $2 \le q \le k(X)$ *and* $t_q^{2/q} \sim n/q$ *for* $k(X) \le q \le n$.

6. As another example of pattern-type behavior of any convex body in $\mathbb{R}^n$, let us mention the following recent fact, proved in [ABV98]:

Theorem 6. *Let* K *be a convex body in* $\mathbb{R}^n$ *with 0 in its interior. For any* $\varepsilon > 0$ *the probability (measured by the standard Lebesgue measure on* K*) of two points, say* x *and* y, *in* K *having* K*-distance of at most* $t = \sqrt{2}(1-\varepsilon)$, *i.e.* $x - y \in tK$, *is at most* $\exp\{-\varepsilon^2 n/2\}$. *(Therefore there are exponentially many points in* K *such that their pairwise differences do not belong to* tK *for* $t < \sqrt{2}(1-\varepsilon)$*).*

So, we again see that the number "$\sqrt{2}$" which is natural for the euclidean ball is also the crucial bound for any other convex body K.

7. Let us return to the study of the special position of the body K (or, equivalently, the special ellipsoid) which we already encountered in Theorem 1. It is usually called the M-position of K. Its formal definition is the following. Let $N(K,T)$ denote the covering number of K by T (i.e. the minimum number of

shifts of T which cover K). Then K is in an M-position (with parameter $\sigma > 0$) if, for $\lambda = (|K|/|D|)^{1/n}$

$$(*) \qquad N(K, \lambda D) \cdot N(\lambda D, K) \cdot N(K^0, \lambda D) \cdot N(\lambda D, K^0) \leq e^{\sigma n} \ .$$

(It is enough to assume $N(K, \lambda D) \leq e^{\sigma n}$ and $(*)$ will follow with a different $\sigma_1 = C \cdot \sigma$, where C is a universal number – see [MS97], [MP98].)

THEOREM 7. *There is a universal number $\sigma > 0$ such that any convex body K with barycenter 0 has an M-position with parameter σ (i.e. $\exists u \in SL_n$ such that uK is in this M-position).*

(For centrally symmetric K, see [M88b] or [M96] for references or the book, [P89]; extension for general convex bodies, [MP98]; generalization to centrally-symmetric p-convex bodies, [BBP95]).

This position of K gives the "correct balance" between the body K (in such a position) and the euclidean ball (or, between the norm and the euclidean structure). Let us explain this by some facts. First, we already demonstrated the use of M-position of a body K in Theorem 1. A few more facts:

THEOREM 8 ([MS97]). *Assume that the unit ball K of a space $X = (\mathbb{R}^n, \|\cdot\|, |\cdot|)$ is in an M-position. Assume further that there are $\{u_i\}_1^t \subset O(n)$ and $0 < r, C < \infty$ such that*

$$r|x| \leq \frac{1}{t} \sum_1^t \|u_i x\| \leq Cr|x| \quad (\textit{for all } x \in \mathbb{R}^n) \, .$$

Then there is a C', depending on t, C and the σ-constant of the M-position only, and $v \in O(n)$ such that, for some r',

$$r'|x| \leq \|x\| + \|vx\| \leq C'r'|x| \quad (\textit{for all } x \in \mathbb{R}^n) \, .$$

Note that the assumption that K is in an M-position (i.e. that the euclidean structure is specially chosen for our norm $\|\cdot\|$) is absolutely essential. Without this assumption, for any $t \ll n/\log n$, and any $\lambda < 1$, one may construct a family of norms (for spaces of dimensions increasing to infinity) such that some average of t-rotations will be uniformly isomorphic to the euclidean norm, but *no* averages of λt rotations can be uniformly equivalent to any euclidean norm (for other such facts, see [MS97]).

Also we observe a remarkable "restructuring" of volume distribution over K under "random" projections where "randomness" is understood in an M-euclidean structure:

THEOREM 9 ([M90]-symmetric case; [MP98]-general case). *Let a convex set K with barycenter at 0 be in an M-position. Then for any $0 < \lambda < 1$ a random orthogonal projection $P_E K \subset E \in G_{n,[\lambda n]}$ has volume ratio bounded by a constant $C(\lambda, \sigma)$ depending only on the proportion of the space λ ($\dim E = [\lambda n]$) and the constant σ of the M-position. (The volume ratio of a body T is the $\frac{1}{n}$th power of the ratio of $|T|$ and the volume of the maximal volume ellipsoid inscribed in T, called John's ellipsoid of T; see [P89] for the importance of this notion in Local Theory).*

8. Additional results. In this section I would like to give a brief review of a few recent developments in Local Theory/Convexity.

(i) *Brascamp-Lieb inequalities and their applications.* In 1989 Keith Ball [Bal89] discovered the relevance of the Brascamp-Lieb [BL76] inequalities to convex geometry. He put these inequalities in the following form:

Theorem 10. *Let $m \geq n$, $(u_i)_{i=1}^m$ be unit vectors in $\mathbb{R}^n$ and let $(c_i)_{i=1}^m$ be positive real numbers such that $\sum_{i=1}^m c_i u_i \otimes u_i = I_n$. Then for all non-negative functions $f_i \in L_1(\mathbb{R})$, $i = 1, \ldots, m$ one has*

$$\int_{\mathbb{R}^n} \prod_{i=1}^m f_i^{c_i}(\langle x, u_i \rangle) dx \leq \prod_{i=1}^m \Big(\int f \Big)^{c_i}.$$

The additional condition which relates the u_i's and the c_i's is often available in convexity and describes, for example, the isotropicity of the John ellipsoid of a given body K. The Brascamp-Lieb inequalities provide sharp upper estimates for volumes. As an application K. Ball obtained sharp upper bounds for the volumes of central linear sections of the unit cube. He also proved that the volume ratio of any symmetric convex body in $\mathbb{R}^n$ is less than that of the cube [Bal89], and that the simplex has maximal volume ratio [Bal91]. This article also contains a reverse isoperimetric inequality: for every convex body K there exists an affine image TK of K such that the ratio $|\partial(TK)|/|TK|^{\frac{n-1}{n}}$ is less than the same quantity computed for the simplex (in the symmetric case, the cube is extremal). For other applications, see [SSc95], [Sc98].

A general reverse Brascamp-Lieb inequality conjectured earlier by Ball [Bal91] was proved by F. Barthe [Bar98b]. His proof uses measure transportation, a new tool started by the result of Brenier [Br91] and developed by McCann [MC95]. It provides Lieb's general inequality and its converse altogether. This new proof allows one to settle the problem of equality cases in the applications of the Brascamp-Lieb inequality to convexity. The reverse inequality may be viewed as a generalization of the Prekopa-Leindler inequality. In particular, it provides lower estimates of volumes of convex hulls and new Brunn-Minkowski type estimates for sum of convex sets sitted in subspaces ([Bar98b]). The particular case of these inequalities which corresponds to Ball's formulation of the Brascamp-Lieb inequalities says:

Theorem 11 ([Bar97]; [Bar98b]). *Let $m \geq n$, let $(u_i)_{i=1}^m$ be unit vectors in $\mathbb{R}^n$ and let $(c_i)_{i=1}^m$ be positive real numbers such that $\sum_{i=1}^m c_i u_i \otimes u_i = I_n$. Then for all non-negative functions $f_i \in L_1(\mathbb{R})$, $i = 1, \ldots, m$ one has*

$$\int_{\mathbb{R}^n}^* \sup_{x = \sum c_i \theta_i u_i} \prod_{i=1}^m f_i^{c_i}(\theta_i) dx \geq \prod_{i=1}^m \Big(\int f \Big)^{c_i}.$$

This result allows Barthe ([Bar98b]) to find the convex bodies of extremal exterior volume ratio and to prove that among the bodies whose John ellipsoid is the Euclidean unit ball, the regular n-simplex has maximal mean width [Bar98a] (this is dual to [Sc98]).

Returning to measure transportation type results let us emphasize that thay are used together with regularity results by Caffarelli [Ca92]. Another curious and useful consequence of this combination of results is the following statement [ADM98]: Let K and T be convex open sets of the same (finite) volume; then there is a smooth measure preserving onto map $\varphi : K \to T$ such that $K + T = \{x + \varphi(x) | x \in K\}$.

(ii) *Economic embedding of n-dimensional subspaces of L_q to ℓ_p^N.* Let us mention here a few new groups of results on embedding some classical spaces to other classical spaces which is a more traditional direction in Local Theory. First, the problem of embedding euclidean subspaces (up to a $(1+\varepsilon)$-isomorphism) into different classes of normed spaces was well understood in the earlier stages of the theory (see [MS86]). Interesting additions in *isometric* embeddings of ℓ_2^n into ℓ_p^N were done in [M88c], [L70], [R92], [LV93], [K95].

Also, an "isomorphic form" of Dvoretzky Theorem was proved in [MS95] and [MS98] showing that ℓ_∞^n gives essentially the worst embedding of ℓ_2^k for any $k > \log n$. More precisely, for some absolute constant $K > 0$ and for every n and every $\log n \leq k < n$, any n-dimensional normed space, X, contains a k-dimensional subspace, Y, satisfying $d(Y, \ell_2^k) \leq K\sqrt{\frac{k}{\log(1+n/k)}}$, and this is exact for all the range of k for ℓ_∞^n spaces ([CP88], [Gl89]).

However, the main interest was directed to non-euclidean embeddings. First, an extremely surprising result by Johnson-Schechtman [JS82] stated that ℓ_q^n may be $(1+\varepsilon)$-embedded into ℓ_p^N for $p < q < 2$ and $N \sim c(\varepsilon; p; q)n$ (for some function $c(\varepsilon; p; q)$). Then Schechtman [S85], [S87] discovered another simple approach to deal with the problem of economic embedding of subspaces of L_q into another ℓ_p (the so-called "empirical method"). This method is not connected with a euclidean structure and the standard use of the Concentration Phenomenon through euclidean spaces, and is equally well applied to the search for large subspaces in a given space without special consideration to the structure of the norm we are working with. It was then used in [BLM89] and [T90] and the question of economic "random" embedding of a subspace $E_n \subset L_q$ of dimension n into ℓ_p^N with exact bounds on $N(n)$ is well understood although some "residual" $\log n$ factors are still distorting the picture.

The question of "natural" embedding (as opposed to "random" embedding) of some subspaces of L_p in low dimensional ℓ_p-spaces happened to be completely different. The whole theory of such embeddings arose in [FJS91]. A few sample results follow:

THEOREM 12. (i) *Let R_n be the span of the first n Rademacher functions in L_1; if X is a subspace of L_1 containing R_n and 2-isomorphic to ℓ_1^m then $m > c^n$ for some universal $c > 1$ (and the same is true for n Gaussian functions).*

(ii) *Every norm one operator from a $C(K)$ space which is a good isomorphism when restricted to a k-dimensional well isomorphic to euclidean subspace also preserves a subspace of dimension c^k (for some $c > 1$) which is well isomorphic to an ℓ_∞-space.*

Another important type of embedding is a complemented embedding (i.e. embedding of a space to another space with a well bounded projection on it). The

empirical method mentioned before provides good estimates for complemented embeddings as well. However additional remarkable results were achieved in [JS91] using some kind of "discrete homothety". For example,

Theorem 13 [JS91]. *If ℓ_p^n is decomposed into a direct sum $X+Y$ with X well isomorphic to a Hilbert space, then Y is well isomorphic to an ℓ_p^m-space.*

The final result given by the theorem is in a direction where some hard work was also done previously (see [BTz87]).

(iii) *Extension of the Dvoretzky-Rogers Lemma and corresponding factorization results.* In 1988, Bourgain and Szarek [BS88] strongly improved the classical Dvoretzky-Rogers Lemma. In the form of a "proportional factorization" their result states: If X is an n-dimensional normed space, then for every $\delta \in (0,1)$ one can find $m \geq (1-\delta)n$ and two operators $\alpha : \ell_2^m \to X$, $\beta : X \to \ell_\infty^m$, such that $id_{2,\infty} = \beta \circ \alpha$ and $\|\alpha\| \cdot \|\beta\| \leq C(\delta)$ for some constant $C(\delta)$ depending on δ only. The dependence on δ was improved to $C(\delta) \lesssim \delta^{-2}$ in [ST89]. It is now known (see [G96], [Ru97]) that the best possible exponent on δ in the proportional Dvoretzky-Rogers factorization must lie between 1 and 1/2. (All these results have immediate application for estimating the maximal Banach-Mazur distance of ℓ_∞^n to any other n-dimensional normed space.)

It was observed in [GM97c] that the factorization result from [G96] is a consequence of a coordinate version of the Low M^*-estimate. The following "coordinate" result was proved: If $\mathcal{E}$ is an ellipsoid then for every $\delta \in (0,1)$ we can find a coordinate subspace $\mathbb{R}^\sigma (= F)$ where $\sigma \subseteq \{1, \dots, n\}$, $|\sigma| \geq (1-\delta)n$, such that for the orthogonal (coordinate) projection $P_F(\mathcal{E})$,

$$P_F(\mathcal{E}) \supseteq \frac{c\sqrt{\delta}}{\sqrt{\log 2/\delta}\; M(\mathcal{E})} D \cap F$$

(for the definition of the expectation $M(\mathcal{E})$, see Sect.2). Note that the factorization discussed above is a consequence of such a coordinate estimate. There is also an extension of this fact to some general classes of bodies (instead of to an ellipsoid).

9. Isotropic positions in convex geometry. In all previous results an isomorphic view on the theory was one of the main messages. Even some definitions were done in an isomorphic form (say, a universal constant σ in the definition of an M-position or M-ellipsoid). However, it is not impossible that a more traditional isometric approach exists which would describe our isomorphic results. (K. Ball suggested such a possibility to me some time ago based on, I believe, his results which I described in 8(i); the "isotropic" view presented below is based on our joint work with Giannopoulos [GM98b].)

Let us start with the isotropic position of a centrally symmetric convex body $K \subset \mathbb{R}^n$ equipped with an inner product $(\cdot,\cdot)$. So, K is isotropic iff $|K| = 1$ and there is a constant L such that

$$\int_K (f,x)(x,\varphi)dx = L(f,\varphi)$$

for any f and φ in $\mathbb{R}^n$. Many remarkable properties of such a position are known and well studied (see, e.g. [MP89]). But our interest is in the following remark

(from the same source): Consider $\min_{u\in SL_n} \int_{uK} |x|^2 dx$ (where $|x|^2 = (x,x)$). Then min. is achieved on the isotropic position.

We understand now that it is a very general fact and for many natural functionals $f(uK)$ considered as functions defined on SL_n (i.e. $u \in SL_n$), the minimum is achieved on some kind of isotropic position (but for a measure which should be found and properly described). For example, the result of F. John about the maximal volume ellipsoid in K provides such an isotropic measure supported on contact points of K and the maximal volume ellipsoid (and the theorem is a consequence of such a general view [GM98b]). But our interest in the framework of this paper has resulted in the fact that some positions used in Asymptotic Convex Geometry (and, in fact, all important used positions we know) have an isometric description as isotropic positions which we derive by minimizing a correctly chosen functional. In such a way the very important ℓ-position, after slight modification becomes an isotropic position for some measure on the sphere. We will mention in addition only an M-position which is also an isotropic position. Indeed, let $|K| = |D|$, and consider the problem

$$\min\{|uK + D| \ \mid u \in SL_n\} \ .$$

The minimum is achieved for some u_0 such that the body $u_0K + D$ has minimal surface area ([GM98b]) and u_0K is in an M-position. At the same time it is known ([Pe61], [GP98]) that a convex body T has minimal surface area iff its surface area measure (supported on S^{n-1}) is isotropic. So, an originally isomorphically defined position also has a purely isometric description.

CONCLUDING REMARK. I see the results of this theory as "a window" to the World of very high degree of freedom, just examples of organized behavior we should expect in the study of that World; not a chaotic diversity, exponentially increasing with increasing degree of freedom (=dimension in the presented Theory), but on the contrary, an asymptotically well organised World with "residual freedom" reflected in our Theory in a "uniformly isomorphic" view on the results.

REFERENCES

[ABV98] J. Arias-de-Reyna, K. Ball, R. Villa, Concentration of the distance in finite dimensional normed spaces, to appear in *Mathematika*.

[ADM98] S. Alesker, S. Dar, V.D. Milman, A remarkable measure preserving diffeomorphism between two convex bodies in $\mathbf{R}^n$, to appear in *Geom. Ded.*

[AlSp92] N. Alon, J.H. Spencer, *The Probabilistic Method*, Wiley Interscience, 1992.

[Bal89] K.M. Ball, Volumes of sections of cubes and related problems. In J. Lindenstrauss, V.D. Milman, eds., *GAFA Israel Seminar* 1376, Springer LNM 1989.

[Bal91] ———, Volume ratio and a reverse isoperimetric inequality, *J. London Math. Soc.* 44:2 (1991),351–359.

[Bar98a] F. Barthe, An extremal property of the mean width of the simplex, to appear in *Math. Annalen.*

[Bar98b] ———, On a reverse form of the Brascamp-Lieb inequality. To appear in *Invent. Math.*

[Bar97] ———, Inégalités de Brascamp-Lieb et convexité. *C.R. Acad. Sci. Paris* 324 (1997), 885–888.

[BBP95] J. Bastero, J. Bernués, A. Peña, An extension of Milman's reverse Brunn-Minkowski inequality, *GAFA* 5:3 (1995), 572–581.

[BS88] J. Bourgain, S.J. Szarek, The Banach-Mazur distance to the cube and the Dvoretzky-Rogers factorization, *Israel J. Math.* 62 (1988), 169–180.

[BLM88] J. Bourgain, J. Lindenstrauss, V.D. Milman, Minkowski sums and symmetrizations, *Springer LNM* 1317 (1988), 44–66.

[BLM89] ———, Approximation of zonoids by zonotopes, *Acta Math.* 162:1-2 (1989), 73–141.

[BL76] H.J. Brascamp, E.H. Lieb, Best constants in Young's inequality, its converse and its generalization to more than three functions, *Adv. Math.* 20 (1976), 151–173.

[BTz87] J. Bourgain, L. Tzafriri, Complements of subspaces of l_p^n, $p \geq 1$, which are uniquely determined, *GAFA Israel Seminar (1985/86)*, Springer LNM 1267 (1987), 39–52.

[Br91] Y. Brenier, Polar factorization and monotone rearrangement of vector-valued functions, *Comm. Pure Appl. Math.* 44 (1991), 375–417.

[CP] B. Carl, A. Pajor, Gelfand numbers of operators with values in Hilbert spaces, *Invent. Math. 94* (1988), 479-504.

[Ca92] L.A. Caffarelli, *A-priori Estimates and the Geometry of the Monge-Ampère Equation*, Park City/IAS Mathematics Series II, 1992.

[FJS88] T. Figiel, W.B. Johnson, G. Schechtman, Factorizations of natural embeddings of l_p^n into L_r, I, *Studia Math.* 89:1 (1988), 79–103.

[FJS91] ———, Factorizations of natural embeddings of l_p^n into L_r, II, *Pacific J. Math.* 150:2 (1991), 261–277.

[G96] A.A. Giannopoulos, A proportional Dvoretzky-Rogers factorization result, *Proc. Amer. Math. Soc.* 124 (1996), 233–241.

[GM97a] A.A. Giannopoulos, V.D. Milman, On the diameter of proportional sections of a symmetric convex body, *International Math. Res. Notices* 1 (1997), 5–19.

[GM97b] ———, How small can the intersection of a few rotations of a symmetric convex body be?, *C.R. Acad. Sci. Paris* 325 (1997), 389–394.

[GM97c] ———, Low M^*-estimates on coordinate subspaces, Journal of Funct. Analysis 147 (1997), 457–484.

[GM98a] ———, Mean width and diameter of proportional sections of a symmetric convex body, *Journal für die reine und angewandte Mathematik* 497 (1998), 113–140.

[GM98b] ———, Extremal problems and isotropic positions of convex bodies, preprint.

[GPa] A.A. Giannopoulos, M. Papadimitrakis, Isotropic surface area measures, to appear in *Mathematika*.

[Gl89] E.D. Gluskin, Extremal properties of orthogonal parallelepipeds and their applications to the geometry of Banach spaces, *Math. USSR Sbornik 64* (1989), 85-96.

[Gor88] Y. Gordon, On Milman's inequality and random subspaces which escape through a mesh in $\mathbf{R}^n$, Springer Lecture Notes in Mathematics 1317 (1988), 84–106.

[JS82] W.B. Johnson, G. Schechtman, Embedding l_p^m into l_1^n, *Acta Math.* 149:1-2 (1982), 71–85.

[JS91] ———, On the distance of subspaces of l_p^n to l_p^k, *Trans. Amer. Math. Soc.* 324:1 (1991), 319–329.

[K95] H. König, Isometric imbeddings of euclidean spaces into finite dimensional l_p-spaces, *Banach Center Publ.* 34 (1995), 79–87.

[LM93] J. Lindenstrauss, V.D. Milman, The local theory of normed spaces and its

applications to convexity. *Handbook of Convex Geometry, Vol. A,B*, 1149–1220, North-Holland, Amsterdam, 1993.

[LMS98] A.E. Litvak, V.D. Milman, G. Schechtman, Averages of norms and quasi-norms, to appear in *Math. Ann.*; see also A.E. Litvak, V.D. Milman, G. Schechtman, Averages of norms and behavior of families of projective caps on the sphere, *C.R. Acad. Sci. Paris Ser. I Math.* 325:3 (1997), 289–294.

[LT91] M. Ledoux, M. Talagrand, *Probability in Banach Spaces, Ergeb. Math. Grenzgeb.* 3. Folge, Vol. 23 Springer, Berlin (1991).

[L70] Yu. Lyubich, On the boundary spectrum of a contraction in Minkowski spaces, *Siberian Math. J.* 11 (1970), 271–279.

[LV93] Yu. Lyubich, L. Vaserstein, Isometric imbeddings between classical Banach spaces, cubature formulas, and spherical designs, *Geom. Ded.* 47 (1993), 327–362.

[McC95] R.J. McCann, Existence and uniqueness of monotone measure preserving maps, *Duke Math. J.* 80 (1995), 309-323.

[M71] V.D. Milman, New proof of the theorem of Dvoretzky on sections of convex bodies, *Funct. Anal. Appl.* 5 (1971), 28–37.

[M85] ———, Random subspaces of proportional dimension of finite dimensional normed spaces: Approach through the isoperimetric inequality, *Springer Lecture Notes in Math.* 1166 (1985), 106–115.

[M86] ———, The concentration phenomenon and linear structure of finite-dimensional normed spaces, *Proceedings of the ICM, Berkeley, California, USA* (1986), 961-975.

[M88a] ———, The heritage of P. Lévy in geometrical functional analysis, Colloque Paul Lévy sur les Processus Stochastiques (Palaiseau, 1987), *Asterisque* 157-158 (1988), 273–301.

[M88b] ———, Isomorphic symmetrization and geometric inequalities, GAFA Sem. (1986/87), *Springer Lecture Notes in Math.* 1314 (1988), 107–131.

[M88c] ———, A few observations on the connections between local theory and some other fields, GAFA Sem. (1986/87), *Springer LNM* 1317, (1988), 283–289.

[M90] ———, Some geometric duality relations, *C.R. Acad. Sci. Paris Ser. I Math.* 310:4 (1990), 183–187.

[M92] ———,Dvoretzky's theorem–thirty years later, *GAFA* 2:4 (1992), 455–479.

[M96] ———, Surprising geometric phenomena in high-dimensional convexity theory, to appear in *Proceedings of ECM2*.

[MP89] V.D. Milman, A. Pajor, Isotropic position and inertia ellipsoids and zonoids of the unit ball of a normed n-dimensional space, *Springer LNM* 1376 (1989), 64–104.

[MP98] ———, Entropy and asymptotic geometry of non-symmetric convex bodies, preprint.

[MS86] V.D. Milman, G. Schechtman, *Asymptotic Theory of Finite Dimensional Normed Spaces*, Springer LNM 1200 (1986).

[MS95] ———, An "isomorphic" version of Dvoretzky's theorem, *C.R. Acad. Sci. Paris Ser. I Math.* 321:5 (1995), 541–544.

[MS97] ———, Global versus Local asymptotic theories of finite-dimensional normed spaces, *Duke Math. J.* 90 (1997), 73–93.

[MS98] ———, An "Isomorphic" Version of Dvoretzky's Theorem, II, to appear in *Convex Geometry, MSRI Publications* 34 (1998).

[PT86] A. Pajor, N. Tomczak-Jaegermann, Subspaces of small codimension of finite dimensional Banach spaces, *Proc. Amer. Math. Soc.* 97 (1986), 637–642.

[Pe61] C.M. Petty, Surface area of a convex body under affine transformations, Proc. Amer. Math. Soc. 12 (1961), 824–828.

[P89] G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry*, Cambridge Tracts in Mathematics 94 (1989).

[R92] B. Reznick, Sums of even powers of real linear forms, *Memoirs AMS* 96 (1992), no. 463.

[Ru97] M. Rudelson, Contact points of convex bodies, *Israel J. Math.* 101 (1997), 93–124.

[S85] G. Schechtman, Fine embeddings of finite-dimensional subspaces of L_p, $1 \leq p < 2$, into l_1^m, *Proc. Amer. Math. Soc.* 94:4 (1985), 617–623.

[S87] ———— More on embedding subspaces of L_p in l_r^n, *Compositio Math.* 61:2 (1987), 159–169.

[SSc95] G. Schechtman, M. Schmuckenschläger, A concentration inequality for harmonic measures on the sphere, GAFA Sem. (1992-1994), *Oper. Theory Adv. Appl.* 77, Birkhauser (1995), 255–273.

[Sc98] M. Schmuckenschläger, An Extremal Property of the Regular Simplex, to appear.

[ST89] S.J. Szarek, M. Talagrand, An "isomorphic" version of the Sauer-Shelah lemma and the Banach-Mazur distance to the cube, GAFA Sem. (1987–88), *Springer Lecture Notes in Math.* 1376 (1989), 105–112.

[T90] M. Talagrand, Embedding subspaces of L_1 into l_1^N, *Proc. Amer. Math. Soc.* 108:2 (1990), 363–369.

[T95] ———, Concentration of measure and isoperimetric inequalities in product spaces. *IHES Publ. Math.* 81 (1995), 73–205.

[T96] ———, A new look at independence. *Ann. Probab.* 24:1 (1996), 1–34.

[TJ88] N. Tomczak-Jaegermann, *Banach-Mazur Distances and Finite Dimensional Operator Ideals*, Pitman Monographs 38 (1989), Pitman, London.

V. D. Milman
School of Math. Sci.,
Sackler Faculty of Exact Sciences,
Tel Aviv Univ.
Tel Aviv 69978, Israel

Functional Calculus on Lie Groups and Wave Propagation

Detlef Müller

1. Introduction

Consider a "sum of squares" operator

$$L = -\sum_{j=1}^{k} X_j^2 + X_{k+1} \tag{1.1}$$

on a smooth manifold M of dimension d, where $X_1, \dots, X_{k+1}$ are smooth, real vector fields on M satisfying the following bracket condition: $X_1, \dots, X_{k+1}$, together with the iterated commutators $[X_{j_1}, [X_{j_2}, [\dots [X_{j_{\ell-1}}, X_{j_\ell}] \dots]]]$, span the tangent space of M at every point of M. If $k = d$, then L might for example be a Laplace-Beltrami operator. If $k < d$, then L is not elliptic, but, according to a celebrated theorem of L. Hörmander [14], it is still hypoelliptic. Operators of this type arize in various contexts, for instance in higher dimensional complex analysis (see e.g. [32]). Assume in addition that L is essentially selfadjoint on $C_0^\infty(M)$ with respect to some volume element dx. Then the closure of L, again denoted by L, admits a spectral resolution $L = \int_0^\infty \lambda \, dE_\lambda$ on $L^2(M)$, and any function $m \in L^\infty(\mathbb{R}^+)$ gives rise to an L^2-bounded operator

$$m(L) := \int_0^\infty m(\lambda) \, dE_\lambda.$$

An important question is then under which additional conditions on the spectral multiplier m the operator $m(L)$ extends from $L^2 \cap L^p(M)$ to an L^p-bounded operator, for a given $p \neq 2$. If so, m is called an *L^p-multiplier for L*, and we write $m \in \mathcal{M}^p(L)$.

Since, without additional properties of M and L, there is little hope in finding answers to this questions, we shall assume that M is a connected Lie group G, with right-invariant Haar measure dx, and that $X_1, \dots, X_k$ are left-invariant vector fields which generate the Lie algebra $\mathfrak{g}$ of G. Moreover, for simplicity, we shall assume $X_{k+1} = 0$, so that L is a so-called *sub-Laplacian.* The choice of a right-invariant Haar measure and left-invariant vector fields ensures that the formal transposed ${}^tX$ of $X \in \mathfrak{g}$ is given by $-X$, so that, by a straight-forward extension of a well-known theorem by E. Nelson and F. Stinespring, L is selfadjoint.

The objective of the talk will be to survey some of the relevant developments concerning this question, and moreover to link it to questions concerning estimates for the associated *wave equation*, more precisely the following Cauchy-problem:

$$\begin{aligned} &(\frac{\partial^2}{\partial t^2} + L)\, u(x,t) = 0 \quad \text{on} \quad G \times \mathbb{R}, \\ &u(x,0) = f(x), \qquad \frac{\partial u}{\partial t}(x,0) = 0, \end{aligned} \tag{1.2}$$

whose solution is given by $u(\cdot,t) = \cos(t\sqrt{L})f$.

The classical model case is the Laplacian $L = -\Delta = -\sum_{j=1}^d \frac{\partial^2}{\partial x_j^2}$ on $\mathbb{R}^d$. Since, on $\mathbb{R}^d$, the spectral decomposition of the Laplacian is induced by the the Fourier transformation, $m(L)$ is here a *Fourier multiplier operator*

$$\widehat{m(L)f}(\xi) = \mu(\xi)\hat{f}(\xi)\ ,$$

with a radial Fourier multiplier $\mu(\xi) = m(|\xi|^2)$, and a sufficient condition for m to be an L^p-multiplier for $-\Delta$ follows from a well-known Fourier multiplier theorem going back to J. Marcinkiewicz, S. Mikhlin and L. Hörmander (see [18][13]):

Fix a non-trivial cut-off function $\chi \in C_0^\infty(\mathbb{R})$ supported in the interval $[1,2]$, and define for $\alpha > 0$

$$\|m\|_{sloc,\alpha} := \sup_{r>0} ||\chi\, m(r\cdot)||_{H^\alpha},$$

where $H^\alpha = H^\alpha(\mathbb{R})$ denotes the Sobolev-space of order α. Thus, $\|m\|_{sloc,\alpha} < \infty$, if m is locally in H^α, uniformly on every scale. Notice also that, up to equivalence, $\|\cdot\|_{sloc,\alpha}$ is independent of the choice of the cut-off function χ.

"MMH-THEOREM". *Suppose that $\|m\|_{sloc,\alpha} < \infty$ for some $\alpha > d/2$. Then $m \in \mathcal{M}^p(-\Delta)$ for every $p \in]1,\infty[$. Moreover, $m(-\Delta)$ is of weak-type (1,1).*

This result is sharp with respect to the critical degree of smoothness $d/2$ for the multiplier for $p = 1$; for $1 < p < \infty$, less restrictive conditions follow by suitable interpolation with the trivial L^2-estimate (see [16], [27]).

The proof of this theorem is based on the following weighted L^2- estimate, which can be deduced from Plancherel's theorem: Let $K_m \in \mathcal{S}'(\mathbb{R}^d)$ be such that $\widehat{K_m} = \mu$, so that $m(L)f = f * K_m$. Then, for m supported in $]0,1]$,

$$\int_{\mathbb{R}^d} |(1+|x|)^\alpha\, K_m(x)|^2\, dx \le C||m||_{H^\alpha}^2\ . \tag{1.3}$$

Now, if G is an arbitrary Lie group, then it has been shown by Y. Guivarc'h and J. Jenkins (see e.g. [35]) that G is either of polynomial growth, or of exponential growth, in the following sense: Fix a compact neighborhood U of the identity element e in G. Then G has *polynomial growth*, if there exists a constant $c > 0$ such that $|U^n| \le cn^c$ for every $n \in \mathbb{N}$, where $|A|$ denotes the Haar measure of a Borel subset A of G. In that case, it is known that there is in fact an integer Q and $C > 0$ such that

$$C^{-1}n^Q \le |U^n| \le Cn^Q \quad \text{for every } n \ge 1. \tag{1.4}$$

G is said to have *exponential growth,* if

$$|U^n| \geq Ce^{\kappa n} \quad \text{for every } n \geq 1, \tag{1.5}$$

for some $\kappa > 0, C > 0$. In this case, there does in fact also hold a similar estimate from above.

Clearly, Euclidean groups are of polynomial growth, and, more generally, the same is true for nilpotent groups.

From an analytic point of view, there is a strong difference between both types of groups: Whereas groups of polynomial growth are *spaces of homogeneous type* in the sense of R. Coifman and G. Weiss (compare [32]), so that standard methods from the Calderón-Zygmund theory of singular integrals do apply, this is not true of groups of exponential growth. I shall mainly concentrate on groups of polynomial growth, and only briefly report on some phenomena discovered in the recent study of a few examples of groups of exponential growth.

2. Polynomial volume growth

Beginning with some early work by A. Hulanicki and E.M. Stein, various analogous of the MMH-Theorem for groups of polynomial growth have been proved in the course of the past two decades. A main objective of these works by various authors, among them L. De Michele, G. Mauceri, J. Jenkins, M. Christ, S. Meda, A. Sikora and G. Alexopoulos (see e.g. [1], also for further references), was the quest for the sharp critical exponent of smoothness in the corresponding theorems on such groups, which is in fact not concluded yet.

Let us look at the important special case of a *stratified Lie group* G, whose Lie algebra $\mathfrak{g}$ admits a decomposition into subspaces

$$\mathfrak{g} = \mathfrak{g}_1 \oplus \cdots \oplus \mathfrak{g}_p,$$

such that $[\mathfrak{g}_i, \mathfrak{g}_k] \subset \mathfrak{g}_{i+k}$ for all i, k, and where $\mathfrak{g}_1$ generates $\mathfrak{g}$ as a Lie algebra. We then form L in (1.1) from a basis $X_1, \ldots, X_k$ of $\mathfrak{g}_1$, with $X_{k+1} = 0$. Such a group is clearly nilpotent and admits a one-parameter group of automorphisms $\{\delta_r\}_{r>0}$, called *dilations*, given by $\delta_r\Big|_{\mathfrak{g}_j} := r^j \mathrm{Id}_{\mathfrak{g}_j}$. Then L is homogeneous of degree 2 with respect to these dilations, and the bi-invariant Haar measure transforms under δ_r as follows:

$$\delta_r^*(dx) = r^Q dx, \tag{2.1}$$

where

$$Q := \sum_{j=1}^{p} j \dim \mathfrak{g}_j$$

is the so-called *homogeneous dimension* of G. It agrees with the growth exponent Q in (1.4). Notice that for groups which are nilpotent of step $p > 1$, the homogeneous dimension is greater than the Euclidean dimension $d = \dim G$; only for abelian groups, both are the same. The following theorem is due to M. Christ ([5], see also [17]):

THEOREM 1. *If G is a stratified Lie group of homogeneous dimension Q, and if $\|m\|_{sloc,\alpha} < \infty$ for some $\alpha > Q/2$, then $m(L)$ is bounded on $L^p(G)$ for $1 < p < \infty$, and of weak type (1,1).*

If $m \in L^\infty(\mathbb{R}^+)$, then, by left–invariance and the Schwartz kernel theorem, it is easy to see that $m(L)$ is a convolution operator $m(L)f = f * K_m = \int_G f(y)K_m(y^{-1}\cdot)dy$, where a priori the convolution kernel K_m is a tempered distribution. We also write $K_m = m(L)\delta_e$. The main problem in proving Theorem 1 is to draw information on K_m, namely to show that K_m is a Calderón-Zygmund kernel, from relatively abstract information on the multiplier m. This is usually done by appealing to estimates for the *heat kernels* $p_t = e^{-tL}\delta_e$, $t > 0$, by some method of subordination. In fact, through work by D. Jerison and A. Sanchez-Calle for the case of stratified groups, and N. Varopoulos and his collaborators for more general Lie groups [35], one knows that p_t, say on a stratified group, satisfies estimates of the following form:

$$p_t(x) \le C_\varepsilon t^{-Q/2} e^{-\frac{d(x,e)^2}{4(1+\varepsilon)t}}\ , \tag{2.2}$$

which are essentially optimal. Here, d denotes the so-called *optimal control* or *Carnot–Carathéodory distance* associated to the Hörmander system of vector fields $X_1, \dots, X_k$ (see [35]), defined as follows: An absolutely continuous path $\gamma : [0,1] \to G$ is called *admissible*, if

$$\dot\gamma(t) = \sum_{j=1}^k a_j(t) X_j(\gamma(t)) \ \text{ for a.e. } \ t \in [0,1].$$

The *length* of γ is then given by $|\gamma| := \int_0^1 \Big(\sum_j a_j(t)^2\Big)^{1/2} dt$, and the associated distance function is defined by $d(x,y) := \inf\{|\gamma| : \gamma \text{ is admissible, and } \gamma(0) = x, \gamma(1) = y\}$, where $\inf\emptyset := \infty$. Hörmander's bracket condition ensures that $d(x,y) < \infty$ for every $x, y \in G$.

Observe that in (2.1) and (2.2), in comparizon to the classical case $G = \mathbb{R}^d$, the homogeneous dimension Q takes over the role of the Euclidian dimension d. Because of this fact, which is an outgrowth of the homogeneity of L with respect to the automorphic dilations, the condition $\alpha > Q/2$ in Theorem 1 appeared natural and was expected to be sharp. The following result, which was found in joint work with E.M. Stein [25], and independently also by W. Hebisch [11], came therefore as a surprise:

Fix $n \in \mathbb{N}$, and let $\mathbb{H}_n$ denote the *Heisenberg group* of Euclidean dimension $d = 2n+1$, for which the group law, expressed in coordinates $(x,y,u) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$, is

$$(x,y,u)\cdot(x',y',u') = (x+x',\, y+y',\, u+u'+\tfrac{1}{2}(x\cdot y' - x'\cdot y)),$$

where $x \cdot y$ denotes the Euclidean inner product. A basis of the Lie algebra of $\mathbb{H}_n$ is then given by the left-invariant vector fields

$$U = \frac{\partial}{\partial u}, \quad X_j = \frac{\partial}{\partial x_j} - \tfrac{1}{2} y_j \frac{\partial}{\partial u}, \quad Y_j = \frac{\partial}{\partial y_j} + \tfrac{1}{2} x_j \frac{\partial}{\partial u}, \quad j = 1, \dots, n.$$

The only nontrivial commutation relations among these elements are the Heisenberg relations $[X_j, Y_j] = U$, $j = 1, \dots, n$. The corresponding sub-Laplacian $L := -\sum_{j=1}^{n} (X_j^2 + Y_j^2)$ is then homogeneous with respect to the automorphic dilations $\delta_r(x, u) := (rx, r^2 u)$, and the homogeneous dimension is $Q = 2n + 2$.

THEOREM 2. *For the sub-Laplacian L on $\mathbb{H}_n$, the statement in Theorem 1 remains valid under the weaker condition $\alpha > d/2$ instead of $\alpha > Q/2$.*

Even though the proofs in [11] and [25] are somewhat different in nature, both draw heavily on the fact that the Heisenberg group has a large group of symmetries. The approach in [25] rests on the following estimate which, surprisingly, is better then what the Euclidean analogue (1.3) would predict:

Let $m \in H^{3/2}$ be supported in the interval $]0, 1]$. Moreover let a "homogeneous norm" on $\mathbb{H}_n$ be given by $|(x, y, u)| := (|x|^4 + |y|^4 + u^2)^{1/4}$, so that in particular $|\delta_r g| = r|g|$ for every $g \in \mathbb{H}_n$ and $r > 0$. Then

$$\int_{\mathbb{H}_n} |(1 + |g|)^2 K_m(g)|^2 \, dg \le C ||m||^2_{H^{3/2}}.$$

For extensions of these results to groups of "Heisenberg type" and "Marcinkiewicz-type" multiplier theorems, see [21],[22].

It is an open questions whether Theorem 1 does hold under the weaker condition $\alpha > d/2$ for arbitrary groups of polynomial growth.

3. SUBORDINATION UNDER THE WAVE EQUATION AND THE CASE OF THE HEISENBERG GROUP

It does not seem possible to derive Theorem 2 from estimates for heat kernels alone. Some approaches to multiplier theorems on polynomially growing groups also make use of information on the associated wave equation (1.2), namely the finite propagation speed for these waves (see.eg. [1]), an idea apparently going back to M. Taylor (see e.g. [34]). However, also these approaches do not yield the sharp result in Theorem 2.

In this section we shall show how, on the other hand, stronger information on wave propagation, namely sharp Sobolev estimates for solutions to (1.2), might in fact lead to sharp multiplier theorems for such groups. For the case of the Heisenberg group, such estimates have been established very recently in joint work with E.M. Stein [26].

Consider the Cauchy problem (1.2). It is natural here to introduce Sobolev norms of the form

$$||f||_{L^p_\alpha} := ||(1 + L)^{\alpha/2} f||_{L^p}.$$

Estimates for $u(\cdot,t)$, for fixed time t, in terms of Sobolev norms of the initial datum f, then reduce essentially to corresponding estimates for the operator $e^{it\sqrt{L}}$. For $p=2$, this operator is unitary, hence bounded on $L^2(G)$, but for $p\neq 2$ this operator will lead to some loss of regularity.

For the classical case of the Laplacian on $\mathbb{R}^d$, such estimates have been established by A. Miyachi [19] and J. Peral [28]. Extensions to the setting of Fourier integral operators, and in particular to elliptic Laplacians, have been given by A. Seeger, C. Sogge and E.M. Stein [29].

However, for non-elliptic sub-Laplacians, the methods in the latter article, which rely on a representation of $e^{it\sqrt{L}}$ as a Fourier integral operator, break down – already the first step, namely to identify $\sqrt{L}$ as a pseudodifferential operator in a "good" symbol class, fails.

Nevertheless, making use of the detailed representation theory of $\mathbb{H}_n$, and in particular of some explicit formulas for certain projection operators due to R. Strichartz, the following analogue of Miyachi-Peral's result has been proved in [26]:

THEOREM 3. *Let L denote the sub-Laplacian on the Heisenberg group, and let $p\in[1,\infty]$. Then, for $\alpha>(d-1)\left|\frac{1}{p}-\frac{1}{2}\right|$, one has*

$$||e^{i\sqrt{L}}f||_{L^p}\leq C_{p,\alpha}||(1+L)^{\alpha/2}f||_{L^p}.$$

By a simple scaling argument, based on the homogeneity of L, one obtains the following estimate for arbitrary time t (we concentrate here on the most important case $p=1$):

$$||e^{it\sqrt{L}}f||_{L^1}\leq C(1+|t|)^{\alpha}||(1+L)^{\alpha/2}f||_{L^1},\quad \text{if } \alpha>(d-1)/2. \tag{3.1}$$

The multiplier theorem in Theorem 2 can be deduced from this result by means of the following principle, respectively variants of it:

SUBORDINATION PRINCIPLE. *Assume that L is a sub-Laplacian on a Lie group G satisfying (3.1) for some $\alpha>0$ and every $t\in\mathbb{R}$. Let $\beta>\alpha+1/2$. Then there is a constant $C>0$, such that for any multiplier $\varphi\in H^\beta(\mathbb{R})$ supported in $[1,2]$, the corresponding convolution kernel K_φ is in $L^1(G)$, and*

$$||K_\varphi||_{L^1(G)}\leq C||\varphi||_{H^\beta(\mathbb{R})}.$$

Proof. Observe first that $(1+L)^{-\varepsilon}\delta_e\in L^1(G)$ for any $\varepsilon>0$. This follows from the formula

$$(1+L)^{-\varepsilon}\delta_e=\frac{1}{\Gamma(\varepsilon)}\int_0^\infty t^{\varepsilon-1}e^{-t(1+L)}\delta_e\,dt=\frac{1}{\Gamma(\varepsilon)}\int_0^\infty t^{\varepsilon-1}e^{-t}p_t\,dt$$

and the fact that the heat kernel p_t is a probability measure on G.

Write

$$\varphi(\lambda)=\psi(\lambda)(1+\lambda^2)^{-\gamma},\quad \text{with } \gamma>\alpha/2,$$

and put $k := (1+L)^{-\gamma}\delta_e \in L^1(G)$.

Then $||\psi||_{H^\beta} \simeq ||\varphi||_{H^\beta}$, and

$$K_\varphi = \psi(\sqrt{L})((1+L)^{-\gamma}\delta_e) = \psi(\sqrt{L})k = \int_{-\infty}^{\infty} \hat{\psi}(t)e^{it\sqrt{L}}k \; dt.$$

Estimate (3.1) then implies

$$||K_\varphi||_{L^1} \leq \int_{-\infty}^{\infty} |\hat{\psi}(t)|(1+|t|)^\alpha||(1+L)^{\alpha/2}k||_{L^1}dt.$$

But, $(1+L)^{\alpha/2}k = (1+L)^{\alpha/2-\gamma}\delta_e \in L^1$, hence, by Hölder's inequality and Plancherel's theorem,

$$||K_\varphi||_{L^1} \leq C(\int_{-\infty}^{\infty} |\hat{\psi}(t)(1+|t|)^\beta|^2 dt)^{1/2} \cdot (\int_{-\infty}^{\infty} (1+|t|)^{2(\alpha-\beta)}dt)^{1/2} \leq C'||\psi||_{H^\beta}.$$

Q.E.D.

In case of the Heisenberg group, it suffices to choose $\beta > \frac{d-1}{2} + \frac{1}{2} = \frac{d}{2}$ in this subordination principle. This is just the required regularity of the multiplier in Theorem 2, and one can in fact deduce Theorem 2 from Theorem 3 by a refinement of the above subordination principle and standard arguments from Calderón-Zygmund theory.

In view of the above considerations, it would be desirable to extend Theorem 3 to larger classes of polynomially growing groups. I do have some hope that such extensions may be achievable by linking the estimates more directly to the underlying geometry through methods from geometrical optics.

4. Exponential volume growth

Comparatively little is yet known in the case of groups with exponential volume growth, even if one deals with full Laplacians.

There are basically two, partially complementary, multiplier theorems of general nature available in this context, both requiring the multiplier to be holomorphic in some neighborhood of the L^2-spectrum of L for $p \neq 2$.

The first, applying to multipliers of so-called Laplace transform type, is due to E. M. Stein [31] and is based on the theory of heat diffusion semigroups and Littlewood-Paley-Stein theory. The second, initiated by M. Taylor (see e.g. [34]), applies to Laplace-Beltrami operators on Riemannian manifolds with "bounded geometry" and lower bound on the Ricci curvature, and makes use of the finite propagation speed of waves on these manifolds.

Let us say that a sub-Laplacian L is of *holomorphic L^p-type*, if there exist a point λ_0 in the L^2-spectrum and an open neighborhood U of λ_0 in $\mathbb{C}$, such that every multiplier $m \in M^p(L)$ extends holomorphically to U.

It is well-known that Riemannian symmetric spaces of the non-compact type are of holomorphic L^p-type for $p \neq 2$, see e.g. [7], [34].

In contrast, we say that L admits a *differentiable L^p-functional calculus*, if there is some integer $k \in \mathbb{N}$ such that $C_0^k(\mathbb{R}_+) \subset \mathcal{M}^p(L)$.

In 1991 W. Hebisch [10] showed that certain distinguished Laplacians L on a particular class of solvable Lie groups G with exponential volume growth, namely the "Iwasawa AN components" of complex semisimple Lie groups, do admit a differentiable L^1- functional calculus, and not only this: m lies in $\mathcal{M}^1(L)$ if and only if $m \in \mathcal{M}^1(-\Delta)$, where Δ denotes the Laplacian on the Euclidian space of the same dimension as G. For variants and extensions of these results, see e.g. [9], [12].

This surprising result does, however, not extend to arbitrary solvable Lie groups, as has recently been shown in joint work with M. Christ [6]. Consider the following group G_1, whose Lie algebra $\mathfrak{g}_1$ has a basis T, X, Y, U such that the only non-trivial commutation relations are

$$[T, X] = X,\ [T, Y] = -Y,\ [X, Y] = U,$$

and the associated Laplacian $L = -(T^2 + X^2 + Y^2 + U^2)$.

G_1 is a semidirect product of the Heisenberg group $\mathbb{H}_1$ with $\mathbb{R}$ (analogues do exist also for higher dimensional Heisenberg groups). Then L is of holomorphic L^p-type for every $p \neq 2$.

As has been proved by H. Leptin and D. Poguntke [15], G_1 is in fact the lowest dimensional solvable Lie group whose group algebra $L^1(G)$ is non-symmetric, and the existence of differentiable L^p- functional calculi for Laplacians on Lie groups seems to be related to the symmetry of the corresponding group algebras.

The few results known so far raise two major questions: Suppose G is a, say, solvable Lie group of exponential growth, and let L be a sub-Laplacian on G. Under which conditions is L of holomorphic L^p-type for $p \neq 2$, respectively, when does it admit differentiable L^p- functional calculi? In the latter case, do theorems of MMH-type hold? The last question would require a good understanding of the integral kernels $m(L)\delta_e$ "at infinity", and is still completely open.

5. Local smoothing for the wave equation

Let us turn back to the Laplacian $L = -\Delta$ on $\mathbb{R}^d$. Then $m(L)$ corresponds to the *radial* Fourier multiplier $\xi \mapsto m(|\xi|^2)$. For such radial multipliers and $1 < p < \infty$, $p \neq 2$, better L^p-estimates can be proved than those obtained from interpolating the MMH-estimate with the trivial L^2-estimate, by making use of the curvature of the sphere $|\xi| = 1$. Let us look at the important model case of the *Bochner-Riesz multipliers*

$$m_\alpha(\lambda) := (1 - \lambda)_+^\alpha.$$

By interpolation, the MMH-Theorem implies

$$m_\alpha \in \mathcal{M}^p(-\Delta),\ \text{if } \alpha > (d-1)\,|\frac{1}{p} - \frac{1}{2}|\ . \tag{5.1}$$

However, the famous *Bochner-Riesz-conjecture* states that

$$m_\alpha \in \mathcal{M}^p(-\Delta),\ \text{if } \alpha > \max\,(d\,|\frac{1}{p} - \frac{1}{2}| - \frac{1}{2}, 0). \tag{5.1}$$

Put $p_d := \frac{2d}{d-1}$. The conjecture reduces to proving that $m_\alpha \in \mathcal{M}^{p_d}(-\Delta)$ for every $\alpha > 0$, whereas (5.1) requires $\alpha > \frac{d-1}{2d}$ for the critical exponent $p = p_d$.

In two dimensions, the conjecture has been proved by L. Carleson and P. Sjölin [3] by means of a more general theorem on oscillatory integral operators (for a variant of their proof by L.Hörmander, and another approach due to C. Fefferman and A. Cordoba, see e.g. [32]). In higher dimensions, only partial results are known hitherto, see e.g. [2], [36].

The Bochner-Riesz-conjecture is again linked to the wave equation via the stronger *local smoothing conjecture*, due to C. Sogge, according to which the solution $u(x,t)$ to the Cauchy problem (1.2) for the wave equation satisfies space-time estimates

$$||u||_{L^{p_d}(\mathbb{R}^d \times [1,2])} \leq C_\varepsilon ||(1-\Delta)^\varepsilon f||_{L^{p_d}(\mathbb{R}^d)}, \tag{5.3}$$

for all $\varepsilon > 0$, again with $p_d = 2d/(d-1)$.

This conjecture is still open even in two dimensions; for interesting partial results, see e.g. [30], [20].

It is known that the validity of the local smoothing conjecture would imply that of other outstanding conjectures in Fourier analysis, like the "restriction conjecture" or the "Kakeya conjecture".

A common theme underlying all these conjectures is the interplay between curvature properties (here, the curvature of the Euclidian sphere) and Fourier analysis. For a comprehensive account of the state of these conjectures and the correlations between them, see [33].

Let me finish by describing a recent joint result with A. Seeger [24] concerning the local smoothing conjecture.

Introduce polar coordinates $x = r\theta$, $r > 0$, $\theta \in S^{d-1}$, where S^{d-1} denotes the unit sphere in $\mathbb{R}^d$. Correspondingly, define mixed norms

$$||f||_{L^p(\mathbb{R}_+, L^2(S^{d-1}))} := \left(\int_0^\infty \Big(\int_{S^{d-1}} |f(r\theta)|^2 d\theta \Big)^{p/2} r^{d-1} dr \right)^{1/p}.$$

Moreover, denote by L_θ the Laplace-Beltrami-operator on the sphere S^{d-1}.

THEOREM 4. *If u is the solution of the Cauchy problem (1.2), and if $2 \leq p < p_d$, then*

$$||u||_{L^p(\mathbb{R}_+ \times [1,2], L^2(S^{d-1}))} \leq C_{p,\varepsilon} \, ||(1-L_\theta)^\varepsilon f||_{L^p(\mathbb{R}_+, L^2(S^{d-1}))},$$

for all $\varepsilon > 0$.

The mixed norm of u in this estimate has to be taken with respect to the variables $(r,t;\theta)$. Slightly sharper endpoint estimates will be contained in [24]. For the case of radial initial data f, endpoint results for $p = p_d$ had been obtained before in [23], [8].

The proof of Theorem 4 makes use of the development of $f(r\theta)$ with respect to θ into spherical harmonics and the corresponding Plancherel theorem on the sphere, and some explicit formulas for the integral kernel of $\cos t\sqrt{-\Delta}$ in polar coordinates,

obtainable through the Hankel inversion formula (see [4]). The integral kernel is decomposed into suitable dyadic pieces which, after applying suitable coordinate changes and re-scalings, finally can be estimated by means of the following vector-valued variant of Carleson-Sjölin's theorem for oscillatory integral operators [24], whose proof does not simply follow from an extension of one of the existing proofs for the scalar valued case:

VECTOR-VALUED CARLESON-SJÖLIN THEOREM. *Let $\phi \in C^\infty(\mathbb{R}^2 \times \mathbb{R})$ be a smooth, real phase function and $a \in C_0^\infty(\mathbb{R}^2 \times \mathbb{R})$ be a compactly supported amplitude. Consider the oscillatory integral operator T_λ given by*

$$T_\lambda f(z) := \int e^{i\lambda\phi(z,y)} a(z,y) f(y)\, dy.$$

Suppose that the Carleson-Sjölin determinant $\det \begin{pmatrix} \phi''_{z_1 y} & \phi''_{z_2 y} \\ \phi'''_{z_1 yy} & \phi'''_{z_2 yy} \end{pmatrix}$ *does not vanish on the support of a (it is in this condition where some curvature condition is hidden). Assume that $2 \le p \le 4$, and put*

$$w_p(\lambda) := \tfrac{(\log(2+\lambda))^{1/2-1/p}}{(1+\lambda)^{1/2}}.$$

Then

$$\Big\| \Big(\sum_j |T_{\lambda_j} f_j|^2 \Big)^{1/2} \Big\|_{L^p(\mathbb{R}^2)} \le C \Big\| \Big(\sum_j w_p^2(|\lambda_j|)\, |f_j|^2 \Big)^{1/2} \Big\|_{L^p(\mathbb{R})}$$

for every sequence of functions $f_j \in L^p(\mathbb{R})$ and every sequence of real numbers λ_j.

REFERENCES

1. G. Alexopoulos, *Spectral multipliers on Lie groups of polynomial growth*, Proc. Amer. Math. Soc. **120** (1994), 897–910.
2. J. Bourgain, *Some new estimates on oscillatory integrals*, Essays in Fourier Analysis in honor of E.M. Stein, Princeton Univ. Press (1995), 83–112.
3. L. Carleson, P. Sjölin, *Oscillatory integrals and a mutiplier problem for the disc*, Studia Math. **44** (1972), 287–299.
4. J. Cheeger, M. Taylor, *On the diffraction of waves by conical singularities I*, Comm. Pure Appl. Math. **25** (1982), 275–331.
5. M.Christ, *L^p bounds for spectral multipliers on nilpotent Lie groups*, Trans. Amer. Math. Soc. **328 no. 1** (1991), 73–81.
6. M. Christ, D. Müller, *On L^p spectral multipliers for a solvable Lie group*, Geom. and Funct. Anal. **6** (1996), 860–876.
7. J.-L. Clerc, E.M. Stein, *L^p multipliers for noncompact symmetric spaces*, Proc. Nat. Acad. Sci. U.S.A. **71** (1974), 3911–3912.
8. L. Colzani, A. Cominardi, K. Stempak, *Radial solutions of the wave equation*, preprint.
9. M. Cowling, S. Guilini, A. Hulanicki, G. Mauceri, *Spectral multipliers for a distinguished Laplacian on certain groups of exponential growth*, Studia Math. **111 (2)** (1994), 103–121.
10. W. Hebisch, *The subalgebra of $L^1(AN)$ generated by the Laplacian*, Proc. Amer. Math. Soc. **117** (1993), 547–549.
11. W. Hebisch, *Multiplier theorem on generalized Heisenberg groups*, Coll. Math. **LXV** (1993), 231-239.
12. W. Hebisch, *Spectral multipliers on exponential growth solvable Lie groups*, preprint.

13. L. Hörmander, *Estimates for translation invariant operators in L^p spaces*, Acta Math. **104** (1960), 93–140.
14. L. Hörmander, *Hypoelliptic second-order differential equations*, Acta Math. **119** (1967), 147–171.
15. H. Leptin, D. Poguntke, *Symmetry and nonsymmetry for locally compact groups*, J. Funct. Anal. **33** (1979), 119–134.
16. W. Littman, *Multipliers in L^p and interpolation*, Bull. Amer. Math. Soc. **71** (1965), 764–755.
17. G. Mauceri, S. Meda, *Vector valued multipliers on stratified groups*, Revista Mat. Iberoamer. **6** (1990), 141–154.
18. S.G. Mikhlin, *On the multipliers of Fourier integrals*, (Russian) Dokl. Akad. Nauk **109** (1956), 701–703.
19. A. Miyachi, *On some estimates for the wave equation in L^p and H^p*, Journ. Fac. Sci. Tokyo, Sci.I.A. **27** (1980), 331–354.
20. G. Mockenhaupt, A. Seeger, C. Sogge, *Local smoothing of Fourier integral operators and Carleson-Sjölin estimates*, J. Math. Soc. **6** (1993), 65–131.
21. D. Müller, F. Ricci, E.M. Stein, *Marcinkiewicz multipliers and multi-parameter structure on Heisenberg (-type) groups I*, Invent. Math. **119** (1995), 199–233.
22. D. Müller, F. Ricci, E.M. Stein, *Marcinkiewicz multipliers and multi-parameter structure on Heisenberg (-type) groups II*, Math. Z. **221** (1996), 267–291.
23. D. Müller, A. Seeger, *Inequalities for spherically symmetric solutions of the wave equation*, Math.Z. **218** (1995), 417–426.
24. D. Müller, A. Seeger, *Regularity properties of wave propagation on conic manifolds*, preprint.
25. D. Müller, E.M. Stein, *On spectral multipliers for Heisenberg and related groups*, J. Math. Pure Appl. **73** (1994), 413–440.
26. D. Müller, E.M. Stein, *L^p- estimates for the wave equation on the Heisenberg group*, preprint.
27. J. Peetre, *Applications de la théorie des espaces d'interpolation dans l'analyse harmonique*, Ricerche -Mat. **15** (1966), 3–36.
28. J. Peral, *L^p estimates for the wave equation*, J. Funct. Anal. **36** (1980), 114–145.
29. A. Seeger, C. Sogge, E.M. Stein, *Regularity properties of Fourier integral operators*, Ann. of Math. **134** (1991), 231–251.
30. C. Sogge, *Propagation of singularities and maximal functions in the plane*, Invent. Math. **104** (1991), 349–367.
31. E.M. Stein, *Topics in harmonic analysis*, Princeton Univ. Press, N.J., 1970.
32. E.M. Stein, *Harmonic Analysis – Real-variable methods, orthogonality, and oscillatory integrals*, Princeton Univ. Press, 1993.
33. T. Tao, *The Bochner-Riesz conjecture implies the restriction conjecture*, preprint.
34. M.E. Taylor, *L^p estimates for functions of the Laplace operator*, Duke Math. J. **58** (1989), 773–793.
35. N.Th. Varopoulos, L. Saloff-Coste, T. Coulhon, *Analysis and geometry on groups*, Cambridge Univ. Press, 1992.
36. T.H. Wolff, *An improved bound for Kakeya type maximal functions*, Revista Mat. Iberoamericana **11** (1995), 651–674.

Detlef Müller
Christian-Albrechts Universität Kiel
Mathematische Seminar
Ludewig-Meyn Str. 4
24098 Kiel, Germany
mueller@math.uni-kiel.de

Author Index for Volumes II, III

Unexpected Solutions of First and Second Order Partial Differential Equations

Stefan Müller and Vladimir Šverák

Abstract. This note discusses a general approach to construct Lipschitz solutions of $Du \in K$, where $u : \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$ and where K is a given set of $m \times n$ matrices. The approach is an extension of Gromov's method of convex integration. One application concerns variational problems that arise in models of microstructure in solid-solid phase transitions. Another application is the systematic construction of singular solutions of elliptic systems. In particular, there exists a 2×2 (variational) second order strongly elliptic system $\operatorname{div} \sigma(Du) = 0$ that admits a Lipschitz solution which is nowhere C^1.

1991 Mathematics Subject Classification: 35F30, 35J55, 73G05
Keywords and Phrases: Partial differential equations, elliptic systems, regularity, variational problems, microstructure, convex integration

1 Introduction and examples

In this note we discuss a general method to construct solutions to a large class of nonlinear first and second order partial differential equations. The method makes strong use of work by Gromov (who substantially generalized earlier results of Nash and Kuiper) and is especially suitable for nonconvex problems where standard compactness arguments fail. One application concerns the (unexpected) existence of solutions in mathematical models of solid-solid phase transformations (see Example c) below). Another application is the recent resolution of the regularity question for weak solutions to the Euler-Lagrange equations of multiple integrals.

Theorem 1.1 *There exists a smooth, strongly elliptic 2×2 system*

$$-\operatorname{div} \sigma(Dv) = 0, \quad v : \mathbb{R}^2 \to \mathbb{R}^2 \tag{1.1}$$

that admits

(i) nontrivial Lipschitz solutions with compact support;

(ii) Lipschitz solutions that are nowhere C^1.

Moreover σ can be chosen such that (1.1) is the Euler-Lagrange equation of a variational integral $\int f(Dv)\,dx$, where f is smooth and uniformly quasiconvex in the sense of Morrey.

The existence of irregular solutions of the Euler-Langrange equations (1.1) is in sharp contrast with the (partial) regularity theory for minimizers of quasiconvex integrals developed by Evans [Ev 86], Acerbi-Fusco, Giaquinta-Modica, Fusco-Hutchinson and many others (due to space constraints I keep references to the minimum; the slightly enlarged version [MS 98] contains a more detailed list of references).

This raises the question which structure conditions on σ are needed to ensure a good regularity theory for quasilinear systems (Tartar has raised this issue in connection with the closely related question of compactness and stability of solutions, see e.g. [Ta 79], p. 160) For a scalar equation the De Giorgi-Moser-Nash Theorem shows that ellipticity is the natural condition. For systems, there is a large literature for monotone σ (see [Gi 83] for a summary, further references and a sketch of the history) and many results can be extended to so-called quasimonotone σ, but these conditions are too restrictive for applications e.g. to nonlinear elasticity. Similar issues arise for problems in nondivergence form as recent counterexamples by Nadirashvili ([Na 97]) show. In the theory of harmonic maps there are also striking differences between minimizers, weak solutions of the Euler-Langrange equations and the intermediate class of so-called stationary harmonic maps (see [He 97], [Si 96] for recent overviews).

We remark in passing that our counterexamples to regularity are quite different from the famous examples of De Giorgi, Bombieri-De Giorgi-Giusti and many subsequent works. The latter are based on finding equations that admit certain point singularities like $x/|x|$ (or certain cones), while our approach uses the fact that the equation is compatible with certain large oscillations of Du (small oscillations must be smooth by ellipticity). The construction of counterexamples is thus reduced to certain algebraic calculations in the space of matrices (see Section 4 below). Here our point of view is strongly influenced by Tartar's work [Ta 79], [Ta 98].

Before we return to the case of 2×2 systems let us review the general setting and some illustrative examples. Given a subset of the $m \times n$ matrices $M^{m\times n}$, a (bounded) domain $\Omega \subset \mathbb{R}^n$ and a map $u_0 : \Omega \to \mathbb{R}^m$ we seek to find Lipschitz maps $u : \Omega \to \mathbb{R}^m$ that satisfy

$$Du(x) \in K \quad \text{for a.e.} \quad x \in \Omega, \tag{1.2}$$

$$u = u_0 \quad \text{on} \quad \partial\Omega. \tag{1.3}$$

Generalizations to problems of the form $F(x, u(x), Du(x)) = 0$ a.e., to maps between manifolds and to higher order derivatives are possible. In order to avoid technicalities as much as possible I focus on (1.2) and (1.3) in the following. This setting already includes a number of interesting examples.

EXAMPLE A) (Scalar u, Hamilton-Jacobi equations) Let $m = 1$. It follows from Theorem 2.4 below that (1.2), (1.3) has a solution if u_0 is C^1 (or piecewise C^1 and continuous) and

$$Du_0 \in \text{int conv}K.$$

For affine u_0 the condition $Du_0 \in \overline{\text{conv}}K$ is clearly necessary. On the other hand, the examples $K = \{a, b\}$ or $K = \{(-1,-1),(-1,1),(1,-1),(1,1)\}$ show that the condition $Du_0 \in \overline{\text{conv}}K$ is in general not sufficient, even if u_0 is affine. If K is the level set of a convex coercive function there is also a good theory of viscosity solutions as developed by Kružkov, Crandall-Lions and many others. In general, Theorem 2.4 below yields existence of solutions in many cases where no viscosity solution exists, but the solutions have much weaker properties (no uniqueness, no comparison principle), a more detailed comparision appears in recent work of Cardaliaguet-Dacorogna-Gangbo-Georgy.

B) (Isometries) If $K = O(n)$ or

$$K = O(n,m) = \{F \in M^{m\times n} :\ F^T F = id_{\mathbb{R}^n}\}$$

then (1.2), (1.3) admit a solution if u_0 is a 'short' map, i.e.

$$Du_0 \in \text{int conv}K = \{F \in M^{m\times n} : \lambda_{max}(F^T F) < 1\},$$

see [Gr 86], Chapter 2.4.11, p. 216. In fact for $m > n$ (and $u_0 \in C^1$) one can obtain C^1 solutions, see [Gr 86], Chapter 2.4.9, Thm. (A), p. 203 .

C) (Two-well problem) In the study of phase transitions in crystals ([BJ 87], [CK 88]; see [Mu 98] for further references) the set

$$K = SO(2)A \cup SO(2)B \subset M^{2\times 2},$$

with A, B symmetric, positive definite, $\det A = \det B = 1$ arises. Theorem 3.2. below shows that solutions exist if $u_0 \in C^{1,\alpha}$(for $0 < \alpha < 1$) and

$$Du_0 \in \text{int}\,\overline{\text{conv}}K \cap \{\det = 1\}.$$

D) $m \times 2$ (Elliptic systems) Let $\sigma : M^{m\times 2} \to M^{m\times 2}$ be a C^1 map and consider the second order system

$$-\text{div}\,\sigma(Dv) = 0 \quad \text{in}\,\Omega, \tag{1.4}$$

i.e. $-\sum_{\alpha=1}^{2} \partial_\alpha \sigma_{i\alpha}(Dv) = 0$, for $i = 1, \cdots m$. If Ω is simply connected then (1.4) can be expressed as

$$\sigma(Dv)J = Dw, \quad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

and if we let $u = \begin{pmatrix} v \\ w \end{pmatrix}$, then (1.4) can be rewritten as

$$Du \in K, \quad K = \left\{ \begin{pmatrix} F \\ G \end{pmatrix} \in M^{2m\times 2} : \sigma(F)J = G \right\}.$$

E) (Four-point configuration). The following example played an important rôle in clarifying different convexity notions in the calculus of variations and was discovered independently (in different contexts) by several authors ([AH 86], [CT 93], [Ta 93]). It will be crucial in the construction of nontrivial solutions to 2×2 elliptic systems. Let (see Figure 1 in section 4)

$$K = \{A_1, A_2, A_3, A_4\},\ -A_1 = A_3 = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix},\ -A_2 = A_4 = \begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}.$$

One easily checks that all solutions of $Du \in K$ are trivial. Corollary 4.1 below shows that there is a large number of nontrivial maps whose gradient stays in an arbitrarily small neighbourhood of K.

2 Convex integration

The first striking results on solutions of relations like (1.2) appeared in the fundamental work of Nash [Na 54] and Kuiper [Ku 55] on isometric immersions. Specifically, Kuiper showed that for any $\varepsilon > 0$ there exist an isometric C^1 immersion $u : S^2 \to \mathbb{R}^3$ that maps S^2 in a ball of radius ε, while a classical theorem of Hilbert states that C^2 isometric immersions are rigid motions (Borisov studied rigidity and non-rigidity in $C^{1,\alpha}$). Extending these ideas Gromov [Gr 86] developed a general method, called 'convex integration' to treat (1.2) and much more general partial differential relations (Spring's recent book [Sp 98] gives a detailed exposition). The main emphasis in [Gr 86] is on the construction of C^1 solutions. In the context of equidimensional isometric immersions Gromov also studies the Lipschitz case in detail and later states a general result for Lipschitz solutions, see Chapter 2.4.11, p. 218. The setting is that of jet bundles and thus the result covers in particular systems of the form $F(x, u(x), \dots, D^{(m)}u(x)) = 0$ a.e. in Ω subject to $D^{(l)}u = v^{(l)}$ on $\partial\Omega, 0 \le l \le m-1$.

A short self-contained proof for the special case (1.2), (1.3) appeared in [MS 96]. Following work of Cellina on ordinary differential inclusions, a slightly different approach based on Baire's theorem was pursued in a series of papers by Dacorogna and Marcellini, beginning with [DM 97], [DM 98]. As we shall see, Gromov's setting (or that of Dacorogna and Marcellini) suffices to discuss Examples a) and b), while for c)–e) additional ideas are needed.

The basic idea of convex integration is that nontrivial solutions of (1.2), (1.3) exist if Du_0 takes values in (the interior of) a suitable convex hull, called the P-convex hull. For sets $K \subset M^{m\times n}$ the notion of P-convexity reduces to what is called lamination convexity in [MS 96] ([MP 98] use the term set-theoretic rank-1 convexity). A set $E \subset M^{m\times n}$ is lamination convex if for all matrices $A, B \in E$ that satisfy $\mathrm{rk}(B - A) = 1$, the whole segment $[A, B]$ is contained in E. The lamination convex hull E^{lc} is the smallest lamination convex set containing E. The relevance of rank-1 matrices stems from the fact that they arise exactly as gradients of maps $x \mapsto u(x \cdot n)$ which only depends on one variable. These maps (and slight modifications thereof; see Lemma 2.2 below) are the basic building blocks in Gromov's construction. In the scalar case $m = 1$ lamination convexity of course reduces to ordinary convexity.

The construction of solutions now proceeds in two steps. First one considers open sets $U \subset M^{m\times n}$, and this case is easily reduced to an open neighbourhood of two matrices A, B with $\mathrm{rk}(B-A)=1$. Secondly one passes to general sets K by approximating them from the inside by open sets contained in K^{lc}. In the following we say that a map $u : \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$ is piecewise linear if it is continuous, if there exist finitely or countably many disjoint sets Ω_i whose union has full measure such that $u_{|_{\Omega_i}}$ is affine and if Du is (essentially) bounded.

Lemma 2.1 *Suppose that $U \subset M^{m\times n}$ is open and bounded and that $u_0 : \Omega \to \mathbb{R}^m$ is piecewise linear and satisfies*

$$Du_0 \in U^{lc} \text{ a.e.} \tag{2.1}$$

Then there exists, for any $\delta > 0$, a piecewise linear u such that

$$Du \in U \text{ a.e.}, \quad u = u_0 \text{ on } \partial\Omega\,, \tag{2.2}$$

$$\sup |u - u_0| < \delta\,. \tag{2.3}$$

For the proof it clearly suffices to consider the case $u_0(x) = Fx$. Using the fact that U^{lc} can be defined inductively by successive addition of rank-1 segments one easily reduces the proof of Lemma 2.1 to the following special case (see [MS 96]).

Lemma 2.2 *Suppose that $\mathrm{rk}(B-A)=1$, i.e. $B-A = a\otimes n$, and let $F = \lambda A + (1-\lambda)B$. If U is an open neighbourhood of $\{A, B\}$ then there exists a piecewise linear u such that*

$$Du \in U \text{ a.e.}, \quad u = Fx \text{ on } \partial\Omega\,, \tag{2.4}$$

$$\sup |u(x) - Fx| < \delta\,. \tag{2.5}$$

To construct u, assume without loss of generality $F = 0, n = e_n$ and consider the set $M = (-1,1)^{n-1} \times \varepsilon(-\lambda, 1-\lambda)$ and the one-dimensional map $v(x) = ah(x_n)$, where $h'(x_n) = 1-\lambda$ for $x_n < 0$, $h'(x_n) = -\lambda$ for $x_n > 0, h(0) = \varepsilon\lambda(1-\lambda)$. Then $Dv \in \{A, B\}$ and $h > 0$ in A. The function $u(x) = ag(x)$, with $g(x) = h(x) - \varepsilon\sum_{i=1}^{n-1} |x_i|$ has the desired properties on the diamond-shaped set $\tilde{M} = M \cap \{g > 0\}$. For general sets $\Omega \subset \mathbb{R}^n$ one can use Vitali's theorem to exhaust Ω by disjoint scaled copies of $\tilde{M}$. Choosing the scaled copies sufficiently small one obtains (2.5).

This finishes the argument for open sets U. For general sets K one needs an suitable approximation by open sets.

Definition 2.3 (Gromov) *A sequence of open sets $U_i \subset M^{m\times n}$ is an in-approximation of a set $E \subset M^{m\times n}$ if*

(i) the U_i are uniformly bounded;

(ii) $U_i \subset U_{i+1}^{lc}$;

(iii) $U_i \to E$ in the following sense: if $F_i \in U_i$ and $F_i \to F$ then $F \in E$.

EXAMPLE For $m = 1$ the shells $U_i = \{x : 1 - 2^{-i+2} < |x| < 1\}$ are an in-approximation of S^{n-1} and $U_1 = B^n = \text{int conv} S^{n-1}$.

THEOREM 2.4 ([GR 86, p. 218; [MS 96]) *Suppose that $\{U_i\}$ is an in-approximation of $K \subset M^{m\times n}$ and that $u_0 : \Omega \to \mathbb{R}^m$ is C^1 (or piecewise C^1) and satisfies*

$$Du_0 \in U_1 \text{ a.e.}$$

Then there exists $u \in W^{1,\infty}(\Omega;\mathbb{R}^m)$ such that

$$Du \in K \text{ a.e.}, \quad u = u_0 \text{ on } \partial\Omega$$

For the proof one uses Lemma 2.1 to inductively construct approximations $u^{(i)}$ with $Du^{(i)} \in U_i$. The key point is to assure that the $Du^{(i)}$ converge strongly. At first glance it is surprising that this can be achieved since the construction in Lemma 2.1 yields solutions with highly oscillatory gradients. Nonetheless by a judicious choice of the C^0 error δ in Lemma 2.1 one can ensure that the oscillations added in each iteration step are essentially independent of the previous ones and only effect a set of small measure. This construction, which is reminiscent of the construction of continuous, nowhere differentiable functions is one of the key ideas of convex integration (in [DM 97] it is replaced by a very elegant, but slightly less flexible, Baire category argument); see [MS 96] for the details.

3 CONSTRAINTS AND SETS WITHOUT RANK-1 CONNECTIONS

The theory explained so far applies to Example a) and b) but not to c) - e). As regards c), the constraint $\det F = 1$ is stable under lamination convexity since $F \mapsto \det F$ is affine in rank-1 directions. Hence the set K in c) does not admit an in-approximation by open sets. The set K in e) contains no rank-1 connections and hence $K^{lc} = K$, and similarly U^{lc} contains only points near K for small neighbourhoods U of K. As regards d), Ball [Ba 80] showed that for strongly elliptic systems that arise as Euler-Lagrange equations (i.e. $\sigma = Df, f : M^{m\times n} \to \mathbb{R}$) again $K^{lc} = K$. It turns out, however, that the previous results can be extended to a slightly larger hull than K^{lc}, namely the rank-1 convex hull K^{rc} (called the functional rank-1 convex hull in [MP 98]), and that this hull can be nontrivial in Examples d) and e).

We say that a function $f : E \to \mathbb{R}$ is rank-one convex on a set E if it is convex on each rank-one line $t \mapsto F + ta \otimes n$. For a compact set K we define the rank-one convex hull relative to E by

$$K^{rc,E} = \left\{ F \in M^{m\times n} : f(F) \le \inf_K f, \forall f : E \to \mathbb{R} \text{ rank-1 convex} \right\},$$

i.e. $K^{rc,E}$ consists of those points that cannot be separated from K by rank-1 convex functions. For a (relatively) open set U the set $U^{rc,E}$ is defined as the union of all $K^{rc,E}$ where $K \subset U$ is compact. If $E = M^{m\times n}$ we simply write K^{rc}

and U^{rc}. The main result is the following variant of Lemma 2.1. Given an $r \times r$ minor M ($r \geq 2$) and a real number $t \neq 0$ we let

$$\Sigma = \left\{ F \in M^{m \times n} : M(F) = t \right\}.$$

LEMMA 3.1 *(i) Let $U \subset M^{m \times n}$ be open, let $F \in U^{rc}$ and let $\varepsilon > 0$. Then there exists a piecewise linear map $u : \Omega \to \mathbb{R}^m$ that satisfies*

$$\begin{aligned} Du &\in U^{rc} \text{ a.e.}, \quad u(x) = Fx \text{ on } \partial\Omega, \\ \text{meas}\,\{Du &\notin U\} < \varepsilon|\Omega|. \end{aligned}$$

(ii) If U is relatively open in Σ and $F \in U^{rc,\Sigma}$, then u can be chosen such that in addition $Du \in U^{rc,\Sigma} \subset \Sigma$ a.e.

By a simple iteration one obtains the counterpart of Lemma 2.1 with U^{lc} replaced by U^{rc} (or $U^{rc,E}$ if a constraint is imposed). The proof of part(i) uses three facts. First, for a compact set K, the rank-1 convex hull K^{rc} consists of the barycentres of a certain class $\mathcal{M}^{rc}(K)$ ('laminates') of probability measures supported on K. Precisely, a probability measure belongs to $\mathcal{M}^{rc}(K)$ if and only if $\langle \nu, f \rangle \geq f(\langle \nu, id \rangle)$ for all rank-1 convex f. Secondly, we use a result of Pedregal [Pe 93] that laminates can be approximated (in the weak$*$ topology of measures) by simpler measures, the so-called laminates of finite order, that are supported on U^{rc}, where U is a (small) neighbourhood of K. The class $\mathcal{L}(U^{rc})$ of laminates of finite order is defined inductively as follows: all Dirac masses δ_F with $F \in U^{rc}$ belong to $\mathcal{L}(U^{rc})$. If $\sum_{i=1}^{k} \lambda_i \delta_{F_i}$ belongs to $\mathcal{L}(U^{rc})$ and if $F_k = \mu A + (1-\mu)B$, $A, B \in U^{rc}$, $\text{rk}(B - A) = 1$, $\mu \in (0,1)$, then $\sum_{i=1}^{k-1} \lambda_i \delta_{F_i} + \lambda_k \mu \delta_A + \lambda_k (1-\mu)\delta_B$ belongs to $\mathcal{L}(U^{rc})$. Third, we inductively use Lemma 2.2 to associate to each $\nu \in \mathcal{L}(U^{rc})$ a map $u : \Omega \to \mathbb{R}^m$ such that $Du \in U^{rc}$ and $|\text{meas}\,\{Du = F_i\} - \lambda_i \text{meas}\,\Omega| < 2^{-i}\varepsilon$. Then u has the desired properties.

To treat the case with constraint one first has to extend Pedregal's result to this situation. Secondly one has to prove a version of Lemma 2.2 which includes the constraint $Du \in \Sigma$ (one can relax (2.4) to the conditions $Du \in U^{rc}$ and meas $\{Du \notin U\} < \varepsilon$.) To this end one first obtains a C^∞ approximation by smoothing and considering a flow of a suitable (divergence-free) vectorfield. Then one constructs a piecewise linear approximation using, among other facts, a result of Dacorogna and Moser on the solvability of $\det Du = f$ in $C^{k,\alpha}$ spaces.

By iteration and the same approximation argument as in the proof of Theorem 2.4 one finally obtains the following result. We say that $\{U_i\}$ is an *rc-in-approximation* of K if the conditions in Definition 2.3 hold with U_i^{lc} replaced by U_i^{rc}.

THEOREM 3.2 *(i) Suppose that $K \in M^{m \times n}$ admits an rc-in-approximation by open sets U_i. Suppose that $u_0 \in C^1(\Omega; \mathbb{R}^m)$ (or u_0 piecewise C^1) and*

$$Du_0 \in U_1^{rc}.$$

Then there exists a map u that satisfies

$$Du \in K \text{ a.e. }, \quad u = u_0 \text{ on } \partial\Omega.$$

(ii) If $K \subset \Sigma$ and $u_0 \in C^{2,\alpha}(\Omega;\mathbb{R}^m)$ then the same assertion holds if the U_i are only relatively open in Σ. If $m = n = r$ then the condition $u_0 \in C^{1,\alpha}(\Omega;\mathbb{R}^m)$ suffices.

Remark. 1. If K is open (or relatively open in Σ) one can take the trivial in-approximation $U_i \equiv K$.
2. As in Gromov's work one can achieve the boundary condition in the stronger sense of fine approximation: for each function $\eta \in C^0(\Omega)$ with $\eta > 0$ there exists a solution that satisfies $|u - u_0| < \eta$. In particular, if $u_0 = 0$ one can find solutions with $Du = 0$ on $\partial\Omega$.

4 Applications

Example a) Lamination convexity reduces to ordinary convexity and an in-approximation with $U_1 \supset \operatorname{int}\operatorname{conv} K$ is given by $U_i = \{F \in \operatorname{conv} K : 0 < \operatorname{dist}(F, K) < 2^{-i+2}\operatorname{diam} K\}$. Hence the assertions in Section 1 follow from Theorem 2.4.

Example b) One easily checks that

$$K^{lc} = K^{rc} = \operatorname{conv} K = \left\{F \in M^{m\times n} : \lambda_{max}(F^T F) \le 1\right\}.$$

It follows that

$$U_i = \left\{F \in M^{m\times n} : 1 - 2^{-i+2} < \lambda_{max}(F^T F) < 1\right\}$$

provides an in-approximation with $U_1 = \operatorname{int}\operatorname{conv} K$, and Theorem 2.4 applies.

Example c) Let $\Sigma = \left\{F \in M^{2\times 2} : \det F = 1\right\}$. By a result of Šverák [Sv 93]

$$K^{lc} = K^{rc} = \Sigma \cap \operatorname{conv} K.$$

Thus $U_i = \Sigma \cap \left\{F \in \operatorname{int}\operatorname{conv} K : 0 < \operatorname{dist}(F, K) < 2^{-i+2}\operatorname{diam} K\right\}$ provides an in-approximation (relative to Σ) with $U_1 \supset \Sigma \cap \operatorname{int}\operatorname{conv} K$.

Example e) Let

$$J_1 = -J_2 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad J_2 = -J_4 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then $K^{lc} = K$ but $K^{rc} = [-1,1]^2 \cup \bigcup_{i=1}^4 [J_i, A_{i+1}]$ (see Figure 1).

As an immediate consequence of Theorem 3.2 we obtain:

Corollary 4.1 *Let $U \supset K$ be open, and suppose that $F \in U^{rc} \supset K^{rc}$. Then there exist $u : \Omega \to \mathbb{R}^2$ such that*

$$Du \in U \text{ a.e.}, \quad u = Fx \text{ on } \partial\Omega.$$

As regards Example d) we recall the result mentioned in the introduction.

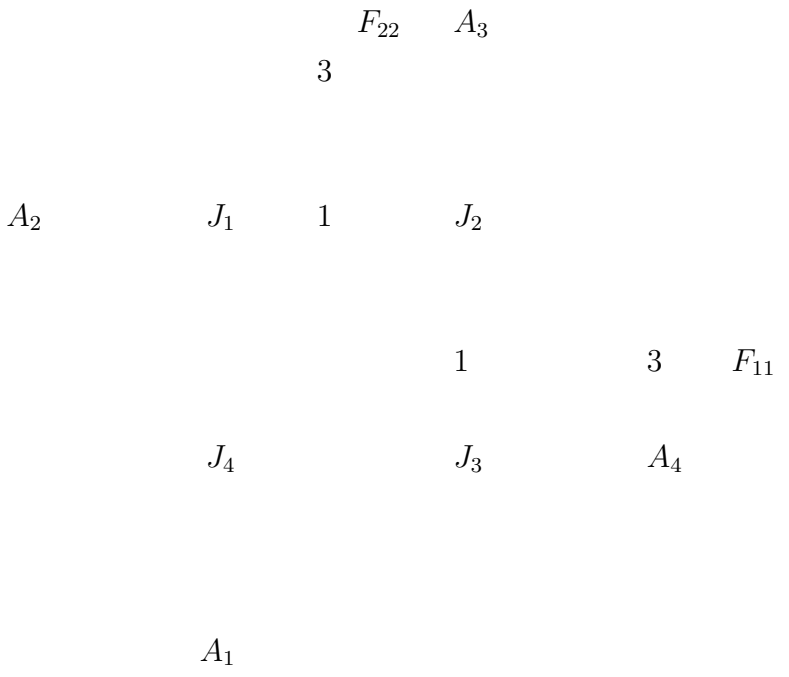


Figure 1: The set $\{A_1, A_2, A_3, A_4\}$ is lamination convex but the rank-one convex hull contains the shaded square and the line segments $[J_i, A_{i+1}]$.

THEOREM 4.2 *There exists a smooth strongly elliptic* 2×2 *system*

$$-\mathrm{div}\sigma(Dv) = 0, \quad v : \mathbb{R}^2 \to \mathbb{R}^2 \tag{4.1}$$

that admits

(i) nontrivial Lipschitz solutions with compact support;

(ii) Lipschitz solutions that are nowhere C^1.

Moreover σ *can be chosen such that (4.1) is the Euler-Lagrange equation of a variational integral* $\int f(Dv)\,dx$, *where* f *is smooth and uniformly quasiconvex in the sense of Morrey.*

Sketch of proof. Our interest lies mainly in the variational case but the main idea can already be seen in the simpler non-variational situation. The key idea is to embed the four-point configuration in Figure 1 in the set

$$K = \left\{ \begin{pmatrix} F \\ G \end{pmatrix} : F, G \in M^{2\times 2},\, \sigma(F)J = G \right\}.$$

This turns out to be surprisingly simple. Consider first the restriction of σ to diagonal matrices and let

$$\sigma_{11}(F_{11}, F_{22}) = F_{11} - g(F_{22}), \quad \sigma_{22}(F_{11}, F_{22}) = F_{22} - h(F_{11}).$$

Strong ellipticity on diagonal matrices reduces to the conditions

$$\frac{\partial \sigma_{11}}{\partial F_{11}} \geq c > 0, \quad \frac{\partial \sigma_{22}}{\partial F_{22}} \geq c > 0,$$

and is clearly satisfied. Moreover g and h can be chosen such that the set $\{\sigma_{11} = \sigma_{22} = 0\}$ includes the points A_1, A_2, A_3, A_4 in Figure 1 and $(0,0)$. If we extend σ to nondiagonal matrices by $\sigma_{12} = kF_{12}$, $\sigma_{21} = kF_{21}$ then σ is elliptic for sufficiently large k, and a careful analysis shows that K^{rc} typically contains a neighbourhood U of $0 \in M^{4\times 2}$, and K admits an rc-in-approximation $\{U_i\}$ with $U_1 = U$.

Let Ω be a smooth and bounded domain in $\mathbb{R}^n$. By Theorem 3.2 there exist a solution

$$Du \in K \text{ a.e.}, \quad u = 0 \text{ on } \partial\Omega.$$

Writing $u = \begin{pmatrix} v \\ w \end{pmatrix}$ we obtain

$$-\mathrm{div}\sigma(Dv) = 0 \quad \text{in } \Omega, \quad v = 0 \quad \text{on } \partial\Omega.$$

Since $Dw = \sigma(Dv)J$, the trace theorem yields $\sigma(Dv)n = 0$ on $\partial\Omega$.

Now extend v by zero to $\mathbb{R}^n$. Since $\sigma(0) = 0$ the map v is a solution of (4.1) with compact support. Regarding (ii) one can use (i) and scaling to construct solutions that can only be regular on a set of arbitrarily small measure. To obtain the full strength of (ii) one has to slightly modify the construction in the proof of Theorem 2.4.

5 Some open problems

A necessary condition for the solvability of

$$Du \subset K \text{ a.e. in } \Omega, \quad u(x) = Fx \text{ on } \partial\Omega$$

is that F belongs to the so-called quasiconvex hull K^{qc} of K which in general is bigger than the rank-1 convex hull K^{rc} (see [Sv 95] or [Mu 98] for definitions and further references). This raises the following questions

- Does Theorem 3.2 hold if one replaces U_i^{lc} by U_i^{qc} in the definition of in-approximation?
- Can one compute (or estimate) K^{qc} for the set K in Example d)?
- Can one find manageable conditions on σ that guarantee $K^{qc} = K$?

Even checking whether $K^{rc} = K$ is in general not obvious. The following Theorem gives a recent example.

THEOREM 5.1 *Let $f(F) = \det F$, for $F \in M^{2\times 2}$, and let $\sigma(F) = Df(F) = \det F \,\mathrm{cof}\, F$. Then the set*

$$K = \left\{ \begin{pmatrix} F \\ G \end{pmatrix} : \sigma(F)J = G \right\}$$

satisfies $K^{rc} = K$.

Acknowledgements

The work reported here is the result of a long collaboration which was supported by the Alexander von Humboldt foundation through the Max Planck research award. We also greatly benefited from discussions with L. Tartar and B. Kirchheim. The Institute for Mathematics and its Applications at the University of Minneapolis kindly provided office space and logistic support during several visits of the first named author.

References

[AH 86] R. Aumann and S. Hart, Bi-convexity and bi-martingales, *Israel J. Math.* 54 (1986), 159 – 180.

[Ba 80] J.M. Ball, Strict convexity, strong ellipticity and regularity in the calculus of variations, *Math. Proc. Cambridge Phil. Soc.* 87 (1980), 501 – 513.

[BJ 87] J.M. Ball and R.D. James, Fine phase mixtures as minimizers of energy, *Arch. Rat. Mech. Anal.* 100 (1987), 13 – 52.

[CT 93] E. Casadio-Tarabusi, An algebraic characterization of quasi-convex functions, *Ricerche Mat.* 42 (1993), 11 – 24.

[CK 88] M. Chipot and D. Kinderlehrer, Equilibrium configurations of crystals, *Arch. Rat. Mech. Anal.* 103 (1988), 237 – 277.

[DM 97] B. Dacorogna and P. Marcellini, General existence theorems for Hamilton-Jacobi equations in the scalar and vectorial cases, *Acta Math.* 178 (1997), 1 – 37.

[DM 98] B. Dacorogna and P. Marcellini, Cauchy-Dirichlet problem for first order nonlinear systems, *J. Funct. Anal.* 152 (1998), 404 – 446.

[Ev 86] L.C. Evans, Quasiconvexity and partial regularity in the calculus of variations, *Arch. Rat. Mech. Anal.* 95 (1986), 227–252.

[Gi 83] M. Giaquinta, *Multiple integrals in the calculus of variations and nonlinear elliptic systems*, Princeton UP, 1983.

[Gr 86] M. Gromov, *Partial differential relations*, Springer, 1986.

[He 97] F. Hélein, *Harmonic maps, conservation laws and moving frames*, Diderot, 1997.

[Ku 55] N.H. Kuiper, On C^1 isometric embeddings, I., *Nederl. Akad. Wetensch. Proc.* A 58 (1955), 545 – 556.

[MP 98] J. Matoušek and P. Plecháč, On functional separately convex hulls, *J. Discrete Comput. Geom.* 19 (1998), 105 – 130.

[MS 96] S. Müller and V. Šverák, Attainment results for the two-well problem by convex integration, *Geometric analysis and the calculus of variations*, (J. Jost, ed.), International Press, 1996, 239 – 251.

[Mu 98] S. Müller, Variational models for microstructure and phase transitions, to appear in: Proc. C.I.M.E summer school 'Calculus of variations and geometric evolution problems' (S. Hildebrandt and M. Struwe, eds.), http://www.mis.mpg.de/

[MS 98] S. Müller and V. Šverák, Unexpected solutions of first and second order partial differential equations, http://www.mis.mpg.de/

[Na 97] N. Nadirashvili, Nonuniqueness in the martingale problem and the Dirichlet problem for uniformly elliptic operators, *Ann. SNS Pisa* Ser. IV 24 (1997), 537 – 550.

[Na 54] J. Nash, C^1 isometric embeddings, *Ann. Math.* 60 (1954), 383–396.

[Pe 93] P. Pedregal, Laminates and microstructure, *Europ. J. Appl. Math.* 4 (1993), 121–149.

[Si 96] L. Simon, *Theorems on regularity and singularity of energy minimizing maps*, Birkhäuser, 1996.

[Sp 98] D. Spring, *Convex integration theory*, Birkhäuser, 1998.

[Sv 93] V. Šverák, On the problem of two wells, in: *Microstructure and phase transitions*, IMA Vol. Appl. Math. 54 (D. Kinderlehrer, R.D. James, M. Luskin and J. Ericksen, eds.), Springer, 1993, 183 – 189.

[Sv 95] V. Šverák, Lower semicontinuity of variational integrals and compensated compactness, in: Proc. ICM 1994 (S.D. Chatterji, ed.), vol. 2, Birkhäuser, 1995, 1153–1158.

[Ta 79] L. Tartar, Compensated compactness and partial differential equations, in: *Nonlinear Analysis and Mechanics: Heriot-Watt Symposium* Vol. IV, (R. Knops, ed.), Pitman, 1979, 136–212.

[Ta 93] L. Tartar, Some remarks on separately convex functions, in: *Microstructure and phase transitions*, IMA Vol. Math. Appl. 54, (D. Kinderlehrer, R.D. James, M. Luskin and J.L. Ericksen, eds.), Springer, 1993, 191–204.

[Ta 98] L. Tartar, *Homogenisation, compensated compactness and H-measures*, CBMS-NSF conference, Santa Cruz, June 1993, lecture notes in preparation.

Stefan Müller
Max Planck Institute for
Mathematics in the Sciences
Inselstr. 22-26
D-04103 Leipzig, Germany

Vladimir Šverák
Department of Mathematics
University of Minnesota
206 Church Street SE
Minneapolis, MN55455, USA

Reflection Principle in Higher Dimensions

Klas Diederich and Sergey Pinchuk

Abstract. The article discusses the use of the reflection principle in studying the following conjecture: Let $D, D' \subset \mathbb{C}^n$ be domains with smooth real-analytic boundaries and $f : D \to D'$ a proper holomorphic map. Then f extends holomorphically to a neighborhood of the closure of D.

1991 Mathematics Subject Classification: 32H40, 32H99, 32D15
Keywords and Phrases: Proper holomorphic maps, reflection principle, Segre variety, holomorphic correspondences

1 Boundary regularity of proper holomorphic maps

Let $D, D' \subset\subset \mathbb{C}^n$, $n \geq 2$, be domains and $f : D \to D'$ a proper holomorphic map. The following two questions are very natural to ask:

1) Suppose, the boundaries ∂D and $\partial D'$ are both $\mathcal{C}^\infty$-smooth. Does f always admit a $\mathcal{C}^\infty$ extension $\hat{f} : \overline{D} \to \overline{D'}$?

2) Suppose, the boundaries ∂D and $\partial D'$ are both $\mathcal{C}^\omega$-smooth. Does f always admit a holomorphic extension $\hat{f}$ to a neighborhood of $\overline{D}$?

Both questions in full generality are open. However, a lot has been found out about them since the early 70's. The emphasis of this article is on question 2). For a survey until 1989 see [13].

The modern development for question 1 started with the article by Ch. Fefferman [12], showing that there is a $\mathcal{C}^\infty$-extension of f, if ∂D and $\partial D'$ are both strictly pseudoconvex and f is biholomorphic. For question 2) the positive answer for strictly pseudoconvex domains was obtained by H. Lewy [15] and S. Pinchuk [16] independently (again f biholomorphic).

Concerning question 1, important further progress was made using methods by S. Webster, E. Ligocka and S. Bell. With them the positive answer was obtained in the case of pseudoconvex domains D, D' of finite type (see [7] and [3]). (A local version needed in section 3 is contained in [4].) After M. Christ discovered in [6], that on the so-called worm domains the $\overline{\partial}$-Neumann problem is not globally hypoelliptic, it has become clear, that these methods do not carry over to the general case of pseudoconvex domains.

Concerning question 2) again the case of pseudoconvex domains has been successfully treated independently in [2] and [8] (both articles also contain local

versions). The case of question 2 for D and D' not necessarily pseudoconvex, has been positively solved for $n = 2$ in [10] building on previous work [11], [8] and [9].

The main methods used in treating question 2 (the real-analytic case) are variations of a reflection principle in several complex variables. There are two major forms, an analytic and a geometric one. Let us at first briefly look at the analytic variant. We choose real-analytic defining functions $\rho(z, \overline{z})$ and $\rho'(z', \overline{z'})$ for the domains D resp. D'. The properness of the map f implies, that we have $\rho'(f(z), \overline{f(z)}) \equiv 0$ on ∂D. In the case $n = 1$, by the implicit function theorem, this equation can be solved in the form $f(z) = \overline{\lambda'(f(z))}$ with a holomorphic function λ'. This gives the extension. In dimension $n > 1$, we need at least n independent equations giving the separation into holomorphic and antiholomorphic parts. Under suitable conditions on the boundaries, they can be obtained by applying tangental CR-operators to the equation $\rho'(f(z), \overline{f(z)}) \equiv 0$. In the strictly pseudoconvex case (see [15] and [16]) one differentiation is enough. However, for boundaries of finite type the number of differentiations is a-priori undetermined. Hence this method, in general, applies only if it is known in advance, that the map f extends in a $\mathcal{C}^\infty$ way up to ∂D.

The geometric version of the reflection principle uses the complexification of the defining functions and the so-called Segre varieties given by them. It will be explained in the next section. For $n = 1$ Segre varieties are just points such that this reflection principle is the well-known Schwarz principle. For $n = 2$ this version was successfully applied in [10] and the articles on which this was built. We point out, that [10] also includes many relevant results for arbitrary $n \geq 2$. A new general result is contained here in section 3.

2 Segre varieties and the geometric reflection principle

Let $D \subset\subset \mathbb{C}^n$ be a domain, such that ∂D is real-analytic smooth near $z^0 \in \partial D$. We may assume $z^0 = 0$. On a suitable open neighborhood W of 0 we can choose a real-analytic defining function $\rho(z, \overline{z})$ for D. After shrinking W the complexification $\rho(z, \overline{w})$ of ρ, which is holomorphic in z and antiholomorphic in w, is well-defined and has a power series convergent on $W \times W$. We now can associate to any point $w \in W$ its so-called "Segre variety" defined as

$$Q_w := \{ z \in W : \rho(z, \overline{w}) = 0 \} \tag{2.1}$$

It is a closed complex submanifold of W not depending on the choice of the defining function ρ. It easily follows, that these Segre varieties are also invariant under biholomorphic changes of coordinate systems and, hence, under local biholomorphisms. The geometric reflection principle makes systematic use of these local invariants and their behavior (A complete list of basic properties needed can be found in Prop. 2.2 of [10]). We will now explain its main ideas and some more technical details needed in section 3 for the proof of Theorem 3.1.

For convenience we will use for $z \in \mathbb{C}^n$ the notation $z = ('z, z_n)$. We can choose so-called normal coordinates associated to ∂D at 0 (see [5]). With respect

to them, ρ has the form

$$\rho(z,\overline{z}) = 2x_n + \sum_{j=0} \rho_j('z,'\overline{z})(2y_n)^j \tag{2.2}$$

with real-analytic functions ρ_j vanishing at 0 and without purely holomorphic or antiholomorphic terms. The complexification of ρ then can be written as

$$\rho(z,\overline{w}) = z_n + \overline{w}_n + \sum_{j=0}^{\infty} \rho_j('z,'\overline{w})(-i)^j(z_n - \overline{w}_n)^j \tag{2.3}$$

It follows, that one has

$$\rho(z,\overline{w}) = 0 \quad \Leftrightarrow \quad z_n + \overline{w}_n + \sum_{|k|>0} \overline{\lambda_k(w)}'z^k = 0 \tag{2.4}$$

where the summation is over multiindices $k = (k_1, \ldots, k_{n-1})$ with $k_j \geq 0$ and each λ_k is a holomorphic function on W. It follows from (2.4) for later use

$$\rho(z,\overline{w}) = (1 + \alpha(z,\overline{w}))\left(z_n + \overline{w}_n + \sum_k \overline{\lambda_k(w)}'z^k\right) \tag{2.5}$$

with a $\mathcal{C}^\omega-$function $\alpha(z,\overline{w})$, holomorphic in z, antiholomorphic in w, vanishing at 0.

For convenience we write $\lambda_0(w) := w_n$. The holomorphic map

$$W \ni w \longmapsto \hat{\lambda}(w) := \left(\lambda_k(w) : k \in \mathbb{N}_0^{n-1}\right)$$

is called the "Segre map". Because of the Noether property, there is an integer $L > 0$ associated to ∂D, such that the terms up to total order L in $\hat{\lambda}$ completely determine $\hat{\lambda}$. If L is chosen with this property we also call Segre map the part

$$W \ni w \longmapsto \lambda(w) := (\lambda_k(w) : |k| \leq L) \in \mathbb{C}^N \tag{2.6}$$

It is important to observe, that the Segre map is often not injective. Therefore, the size of the complex-analytic sets

$$A_w := \{\, z : Q_z = Q_w\} \tag{2.7}$$

is decisive for the geometric reflection principle. We say

Definition 2.1 *The domain D is called essentially finite at $0 \in \partial D$, if A_0 (and, hence, A_w for all w close to 0) is finite (see [11] and [1]).*

Real-analytic smooth hypersurfaces of finite type are always essentially finite. Furthermore, if ∂D is essentially finite at 0, the Segre map λ is finite and, hence, proper on W sufficiently small, . In this case, the set $\mathcal{S} := \lambda(W) \subset \mathbb{C}^N$ is closed complex analytic in a suitable neighborhood of $\lambda(0)$.

Let now $D, D' \subset\subset \mathbb{C}^n$ be real-analytic smooth domains. According to [14] they are of finite type. We will apply the above considerations both to D and D'. We introduce the notational convention, that the objects associated to D' will be denoted by the same letters as for D with a prime added (for instance, $Q'_{w'}$ is the Segre variety associated to $\partial D'$ at w').

Suppose now a proper holomorphic map $f : D \to D'$ is given. The program of using the Segre varieties for constructing a holomorphic extension of f to a neighborhood of $\overline{D}$ consists of the following two major steps:

1) Let $0 \in \partial D$ be an (arbitrary) point. Show, that there is a neighborhood W of 0, an open set $W' \subset \mathbb{C}^n$ and a proper holomorphic correspondence $F : W \to W'$ extending f from $W \cap D$ to W.

2) Let W be an open neighborhood of $0 \in \partial D$, $W' \subset \mathbb{C}^n$ open, and suppose, that a proper holomorphic correspondence $F : W \to W'$ extends f from $W \cap D$ to W. Show, that this implies the extendability of f as a holomorphic map to a neighborhood of 0.

This program has been carried out in full detail for $n = 2$ in [10]. However, many considerations of [10] are valid for general n and the step 2) for general n will be completed in this article in Theorem 3.1.

For the first step of the above-mentioned program we proceed essentially as follows: We choose the neighborhood W of 0 suitably and denote by W' a small open neighborhood of $\partial D'$. For $w' \in W'$, we denote by ${}^s w'$ the point on $Q'_{w'}$ on the complex normal through w' to $\partial D'$ and by ${}_{s_{w'}}Q'_{w'}$ the germ of $Q'_{w'}$ at ${}^s w'$. We put

$$V := \left\{ (w, w') \in \left(W \setminus \overline{D}\right) \times \left(W' \setminus \overline{D'}\right) : f(Q_w \cap D) \supset {}_{s_{w'}}Q'_{w'} \right\} \tag{2.8}$$

Notice, that, since the Segre map λ' is, in general, not injective, the set V will usually contain several points (w, w') lying over one point w.

After now showing at first by totally different techniques, that f extends as a holomorphic map to a neighborhood of a dense subset of ∂D, a long chain of steps distinguishing between boundary points of different CR-nature allows to show, that the set V can be extended across $\partial D \cap W$ in such a way, that a proper holomorphic correspondence $F : W \to W'$ is obtained extending f.

In step 2) an extending proper holomorphic correspondence $F : W \to W'$ is given. It induces a continuous extension of f to $W \cap \overline{D}$. Again one uses the fact, that f extends holomorphically across a dense subset of ∂D to deduce from the invariance of the Segre varieties under biholomorphisms the following much stronger invariance property with respect to F (Corollary 4.2 and 5.5 of [10]):

THEOREM 2.2 *If neighborhoods W of 0 and W' of $0' := f(0)$ are chosen suitably, then there is a bijective holomorphic map $\varphi : \mathcal{S} \to \mathcal{S}'$, such that the diagram*

$$\begin{array}{ccc} \mathcal{S} & \overrightarrow{\varphi} & \mathcal{S}' \\ \lambda \uparrow & & \uparrow \lambda' \\ W & \overrightarrow{\hat{F}} & W' \end{array}$$

is commutative. (Here we denoted by $\hat{F}$ the set-valued map induced by F.)

We mention the following immediate consequence needed in the proof of Theorem 3.1. It concerns the functions λ_k from (2.4):

Lemma 2.3 *Under the hypothesis of Theorem (2.2) and with W, W' chosen as there, for every multiindex $k = (k_1, \ldots, k_{n-1})$, $|k| > 0$, the set $\lambda'_k(\hat{F}(z))$ consists of a unique complex number for every $z \in W$ and this defines a holomorphic function on W.*

3 Extending correspondences are maps in all dimensions

We will show in this section:

Theorem 3.1 *Let $D, D' \subset\subset \mathbb{C}^n$ be domains, $n \geq 2$. Suppose that $z^0 \in \partial D$ and $z'^0 \in \partial D'$ have open neighborhoods W resp. W' such that $\partial D \cap W$ and $\partial D' \cap W'$ are smooth real-analytic, essentially finite hypersurfaces and let $f : D \to D'$ be a proper holomorphic map. Furthermore, suppose, that the given map f extends as a proper holomorphic correspondence F to a neighborhood of z^0 such that $\hat{F}(z^0) = z'^0$. Then the map f extends holomorphically to a neighborhood of z^0.*

Proof: We may assume, that $z^0 = 0 = z'^0$, that the given correspondence F extending f is defined over W and the coordinates z, z' have been chosen to be normal at 0. Hence, a suitable defining function $\rho \in \mathcal{C}^\omega(W)$ can be written as in (2.2), similarly for D' near 0. We apply all notions of section 2. After rescaling the coordinates, we have polydiscs $U \subset W$ and $U' \subset W'$ around 0 of radius 2, such that $\hat{F}(U) \subset U'$ and the following property holds:

All functions $\rho(z, w)$, $\rho_j('z, 'w)$, $\lambda_k(w)$, $\sum_k \lambda_k(w)' z^k$ and the corresponding functions for the image are holomorphic in polydiscs around 0 of radius 2 in the corresponding dimensions. In particular, the series $\sum_k |\lambda_k(w)|$, $\sum_k |\frac{\partial \lambda_k}{\partial w_n}(w)|$ and the corresponding series for the image converge uniformly on compact subsets of U (resp. U'). Because of the normality of the coordinates we also have $\lambda'_k(0) = 0$ and $\frac{\partial \lambda'_k}{\partial w'_n}(0) = 0$ for all k. Therefore we have

$$\sum_k |\lambda'_k(0)| = 0 \quad \text{and} \quad \sum_k \left| \frac{\partial \lambda'_k}{\partial w'_n}(0) \right| = 0 \tag{3.1}$$

Since, as explained in [10], Prop.7.2, $f_n(z) = z_n h(z)$ on U with h holomorphic and $h(0) \neq 0$, we can make a biholomorphic coordinate change by replacing z_n by $z_n h(z)$. However, we have to be aware of the fact that the new coordinates are no longer normal for ∂D at 0.

Now the series $\sum_k \lambda'_k(w') \overline{\lambda'_k(\overline{\zeta'})}$ converges on $U' \times U'$ and represents a holomorphic function there. Putting $\zeta' := \overline{w'}$ and $w' \in \hat{F}(z)$, we get because of Lemma 2.3 $\sum_k |\lambda'_k(w')|^2 \in \mathcal{C}^\omega(U')$ and $\sum_k |\lambda'_k(\hat{F}(z))|^2 \in \mathcal{C}^\omega(U)$.

Since $\partial D'$ is supposed to be essentially finite, we may assume, that for all $a' \in U'$ the set

$$A'_a := \{ ('w', w'_n) \in U' : w'_n = a'_n, \lambda'_k(w') = \lambda'_k(a') \ \forall |k| > 0 \} \tag{3.2}$$

is finite.

We now introduce the following two decisive auxiliary open sets depending on a sufficiently large number $M \gg 1$ and $\varepsilon \in (-\frac{1}{M}, 0]$:

$$\mathcal{D}'(M,\varepsilon) := \left\{ w' \in U' : 2\,\mathrm{Re}\, w'_n + M|w'_n|^2 + M\sum_k |\lambda'_k(w')|^2 < \varepsilon \right\} \tag{3.3}$$

$$\mathcal{D}(M,\varepsilon) := \left\{ z \in U : 2x_n + M|z_n|^2 + M\sum_k \left|\lambda'_k\Big(\hat{F}(z)\Big)\right|^2 < \varepsilon \right\} \tag{3.4}$$

We have

LEMMA 3.2 *The open sets $\mathcal{D}'(M,\varepsilon)$ and $\mathcal{D}(M,\varepsilon)$ are pseudoconvex and their boundaries are of finite type at all points in U resp. U' where they are smooth.*

Proof: Both open sets are obviously inside the polydisk U resp. U' as sublevelsets of plurisubharmonic functions (for (3.4) we know from Lemma 2.3, that the $\lambda'_k(\hat{F}(z))$ are holomorphic functions on U). Hence they are pseudoconvex.

Next we observe, that, the defining function for $\mathcal{D}'(M,\varepsilon)$ from (3.3) can be rewritten

$$\rho'_{M,\varepsilon}(w',\overline{w}') := M\left|w'_n + \frac{1}{M}\right|^2 + M\sum_k |\lambda'_k(w')|^2 - \varepsilon - \frac{1}{M} \tag{3.5}$$

If $\partial\mathcal{D}'(M,\varepsilon)$ is smooth near a point $w'^0 \in U'$ and $h : \Delta \to \partial\mathcal{D}'(M,\varepsilon)$ is a holomorphic map with $h(0) = w'^0$ and $h(\Delta) \subset \partial\mathcal{D}'(M,\varepsilon)$, then because of (3.5), h_n and $\lambda'_k \circ h(t)$ have to be constant for all k. Since A'_a is finite, h itself has to be constant showing that $\partial\mathcal{D}'(M,\varepsilon)$ is of finite type at w'^0. The reasoning for $\mathcal{D}(M,\varepsilon)$ goes the same way. □

In general, it is not true that $\mathcal{D}'(M,\varepsilon) \subset D'$ (resp. $\mathcal{D}(M,\varepsilon) \subset D$). However, we have the following crucial

LEMMA 3.3 *If $M \gg 1$ is sufficiently large, then one has for any $\varepsilon \in (-\frac{1}{M}, 0]$*
a) the non-smooth part of $\partial\mathcal{D}'(M,\varepsilon)$ is contained in D';
b) the non-smooth part of $\partial\mathcal{D}(M,\varepsilon)$ is contained in D.

Proof: We show at first a). Since $\varepsilon \le 0$, $w' \in \partial\mathcal{D}'(M,\varepsilon)$ implies because of (3.5)

$$M\left|w'_n + \frac{1}{M}\right|^2 + M\sum_k |\lambda'_k(w')|^2 \le \frac{1}{M} \tag{3.6}$$

Hence, the next three estimates follow directly

$$-\frac{2}{M} \le \mathrm{Re}\, w'_n \le 0, \quad |w'_n|^2 \le \frac{4}{M^2}, \quad \sum_k |\lambda'_k(w')|^2 \le \frac{1}{M^2} \tag{3.7}$$

Since 0 is the only solution of the system

$$w'_n = 0,\ \lambda'_k(w') = 0\ \forall |k| > 0$$

shrink to the origin for $M \to \infty$. In particular, necessarily also $w' \to 0$ as $M \to \infty$.

Let now $\partial\mathcal{D}'(M,\varepsilon)$ be non-smooth at w'. Then grad $\rho'_{M,\varepsilon}(w',\overline{w}') = 0$. Hence

$$\frac{\partial \rho'_{M,\varepsilon}}{\partial w'_n}(w',\overline{w}') = 0$$

implying

$$\frac{1}{M} + \operatorname{Re} w'_n + \operatorname{Re} \sum_k \frac{\partial \lambda'_k}{\partial w'_n}(w')\overline{\lambda_k(w')} = 0 \tag{3.8}$$

By (3.7) we have $|\lambda'_k(w')| \le \frac{1}{M}$ and, therefore,

$$\left| \sum_k \frac{\partial \lambda'_k}{\partial w'_n}(w')\overline{\lambda'_k(w')} \right| \le \frac{1}{M} \sum_k \left| \frac{\partial \lambda'_k}{\partial w'_n}(w') \right|$$

Because of (3.1) the sum on the right side is $o(1)$ for $w' \to 0$ uniformly in ε (this uniformity in $\varepsilon \in (-\frac{1}{M}, 0]$ holds in all the following estimates). Hence we get

$$\sum_k \frac{\partial \lambda'_k}{\partial w'_n}(w')\overline{\lambda'_k(w')} = o\left(\frac{1}{M}\right) \text{ for } M \to \infty \tag{3.9}$$

Together with (3.8) we get

$$\operatorname{Re} w'_n = -\frac{1}{M} + o\left(\frac{1}{M}\right) \tag{3.10}$$

Using again $|\lambda'_k(w')| \le \frac{1}{M}$ we also obtain

$$\sum_k \overline{\lambda'_k(w')}' w'^k = o\left(\frac{1}{M}\right) \tag{3.11}$$

Putting (3.10) and (3.11) into (2.5), we deduce

$$\rho'(w',\overline{w}') = (1 - \alpha'(w',\overline{w}'))\left(2 \operatorname{Re} w'_n + \sum_k \overline{\lambda'_k(w')}' w'^k\right) = -\frac{2}{M} + o\left(\frac{1}{M}\right) < 0$$

for large M uniformly in ε. Hence $w' \in D'$ finishing part a).

For showing b) we consider the defining function

$$\rho_{M,\varepsilon}(z,\overline{z}) := 2 \operatorname{Re} z_n + M|z_n|^2 + M \sum_k \left| \lambda'_k\left(\hat{F}(z)\right) \right|^2 - \varepsilon \tag{3.12}$$

of $\mathcal{D}(M,\varepsilon)$ and keep in mind that $f_n(z_n) = z_n$. Let now $z \in \partial\mathcal{D}(M,\varepsilon)$ be a non-smooth boundary point. In complete analogy to a) we get

$$2\,\mathrm{Re}\, z_n + M|z_n|^2 + M\sum_k \left|\lambda_k'\Big(\hat F(z)\Big)\right|^2 \le 0 \tag{3.13}$$

and the three inequalities

$$-\frac{2}{M} \le x_n \le 0, \quad |z_n|^2 \le \frac{4}{M^2}, \quad \sum_k \left|\lambda_k'\Big(\hat F(z)\Big)\right|^2 \le \frac{1}{M^2} \tag{3.14}$$

and since, again, $\hat F(z) \to \{0\}$ as $z \to 0$, we have

$$\sum_k \overline{\lambda_k'\Big(\hat F(z)\Big)}\Big[{}'\hat F(z)\Big]^k = o\Big(\frac{1}{M}\Big) \tag{3.15}$$

However, since $\frac{\partial\lambda_k'(\hat F(z))}{\partial z_n}$ does not necessarily vanish at 0, the analogue of (3.10) might not hold. But there exists at least a $c > 0$ such that for large $M \gg 1$

$$x_n \le -\frac{c}{M} \tag{3.16}$$

Namely, if (at least on a suitable subsequence) $x_n = o(\frac{1}{M})$, then we get from (3.13) $|\lambda_k'(\hat F(z))| = o(\frac{1}{M})$ and, therefore,

$$\sum_k \frac{\partial\lambda_k'\Big(\hat F(z)\Big)}{\partial z_n}\overline{\lambda_k'\Big(\hat F(z)\Big)} = o\Big(\frac{1}{M}\Big)$$

This, however, is a contradiction to

$$\frac{1}{M} + x_n + \mathrm{Re}\sum_k \frac{\partial\lambda_k'}{\partial z_n}\Big(\hat F(z)\Big)\overline{\lambda_k'\Big(\hat F(z)\Big)} = 0 \tag{3.17}$$

which holds in analogy to (3.8). This shows (3.16).

In order to show that $z \in D$ we will use the multivalued "function" $\rho'(\hat F(z), \hat F(z))$, observing at first, that according to Prop. 7.1 from [10], for any fixed z, all its values have the same sign, namely, D always goes to D' under $\hat F(z)$ and the exterior goes to the exterior.

Because of (2.5) we have

$$\rho'\Big(\hat F(z), \overline{\hat F(z)}\Big) = \Big(1 + \alpha'\Big(\hat F(z), \overline{\hat F(z)}\Big)\Big)\left(2x_n + \sum_k \overline{\lambda_k'\Big(\hat F(z)\Big)}\Big[{}'\hat F(z)\Big]^k\right)$$

with $\alpha'(0,0) = 0$. Hence, we obtain from (3.15) and (3.16)

$$\rho'\Big(\hat F(z), \overline{\hat F(z)}\Big) \le -\frac{2c}{M} + o\Big(\frac{1}{M}\Big) < 0 \tag{3.18}$$

Therefore, $z \in D$. □

The next essential step in proving Theorem 3.1 is to show

Lemma 3.4 *If $M \gg 1$ is chosen as in Lemma 3.3, then f extends holomorphically to a proper map $\hat{f} : \mathcal{D}(M,0) \to \mathcal{D}'(M,0)$.*

Proof: For such M and ε close to $-\frac{1}{M}$, $\mathcal{D}(M,\varepsilon)$ is a small neighborhood of the set

$$A := \left\{ z_n = -\frac{1}{M}, \lambda'_k\Big(\hat{F}(z)\Big) = 0 \ \forall |k| > 0 \right\}$$

Because of (2.5) we have for any $z \in A$ (notice, that $z_n = -\frac{1}{M}$)

$$\rho'\Big(\hat{F}(z), \overline{\hat{F}(z)}\Big) = \Big(1 + \alpha'\Big(\hat{F}(z), \overline{\hat{F}(z)}\Big)\Big) \cdot 2x_n < 0$$

Hence $\mathcal{D}(M,\varepsilon) \subset D$ and $\mathcal{D}'(M,\varepsilon) \subset D'$ if in addition $\varepsilon \in (-\frac{1}{M}, 0]$ is close to $-\frac{1}{M}$. Therefore, by the definition of $\mathcal{D}(M,\varepsilon)$, $f : \mathcal{D}(M,\varepsilon) \to \mathcal{D}'(M,\varepsilon)$ is proper holomorphic.

Now let $\varepsilon \in (-\frac{1}{M}, 0]$ be maximal such that f extends to a proper holomorphic map $\hat{f} : \mathcal{D}(M,\varepsilon) \to \mathcal{D}'(M,\varepsilon)$. Notice at first, that this map is extended as a proper holomorphic correspondence to a neighborhood of $\overline{\mathcal{D}(M,\varepsilon)} \cap U$ by F. Therefore, $\hat{f}$ is continuous up to the boundary. Because of Lemma 3.3 and known results about holomorphic extension as mentioned in section 2 this map extends as a proper holomorphic map to $\mathcal{D}(M,\tilde{\varepsilon})$ with $\tilde{\varepsilon} > \varepsilon$ unless $\varepsilon = 0$. □

End of the proof of Theorem 3.1: By applying the same arguments as at the end of the last proof and using that $0 \in \partial\mathcal{D}(M,0)$, we see, that $\hat{f}$ extends holomorphically to a neighborhood of 0. □

References

1. Baouendi, M. S., Jacobowitz, H., Treves, F.: *On the analyticity of CR mappings*, Ann. Math. 122 (1985), 365–400.

2. Baouendi, M. S., Rothschild, L. P.: *Germs of CR maps between real analytic hyperfaces*, Invent. Math. 93 (1988), 481–500.

3. Bell, S., Catlin, D.: *Boundary regularity of proper holomorphic mappings*, Duke Math. J. 49 (1982), 385–396.

4. Bell, S., Catlin, D.: *Regularity of CR mappings*, Math. Z. 199 (1988), 357–368.

5. Chern, S. Y., Moser, J.: *Real hypersurfaces in complex manifolds*, Acta Math. 133 (1974), 219–271.

6. Christ, M.: *Regularity properties of the $\overline{\partial}_b$ equation on weakly pseudoconvex CR manifolds of dimension 3*, J. Amer. Math. Soc. 1 (1988), 587–646.

7. Diederich, K., Fornæss, J. E.: *Boundary regularity of proper holomorphic mappings*, Inventiones Math. 67 (1982), 363–384.

8. Diederich, K., Fornæss, J. E.: *Proper holomorphic mappings between real-analytic pseudoconvex domains in $\mathbb{C}^n$*, Math. Ann. 282 (1988), 681–700.

9. Diederich, K., Fornæss, J. E., Ye, Z.: *Biholomorphisms in dimension 2*, J. Geom. Analysis 4 (1994), 539–552.

10. Diederich, K., Pinchuk, S. I.: *Proper holomorphic maps in dimension 2 extend*, Indiana Univ. Math. J. 44 (1995), 1089–1126.

11. Diederich, K., Webster, S.: *A reflection principle for degenerate real hypersurfaces*, Duke Math. J. 47 (1980), 835–845.

12. Fefferman, C.: *The Bergman kernel and biholomorphic mappings of pseudoconvex domains*, Inventiones Math. 26 (1974), 1–65.

13. Forstnerič, F.: *A survey on proper holomorphic mappings*, Proceedings of the Special Year in SCV's at the Mittag-Leffler Institute (Princeton, N. J.) (Fornæss, J. E., ed.), Math. Notes, vol. 38, Princeton University Press.

14. Lempert, L.: *On the boundary behavior of holomorphic mappings*, Contributions to Several Complex Variables (in honour of Wilhelm Stoll) (Braunschweig - Wiesbaden) (Howard, A., Wong, P.M., eds.), Vieweg and Sons, pp. 193–215.

15. Lewy, H.: *On the boundary behavior of holomorphic mappings*, Acad. Naz. Linc. 35 (1977), 1–8.

16. Pinchuk, S.: *On the analytic continuation of holomorphic mappings*, Math. USSR-Sb. 27 (1975), 375–392.

Klas Diederich
Mathematik, Univ. Wuppertal
Gausstr. 20
D-42097 Wuppertal, GERMANY
klas@math.uni-wuppertal.de

Sergey Pinchuk
Department of Mathematics
Indiana University
Bloomington, IN 47405, USA
pinchuk@indiana.edu

Developments from Nonharmonic Fourier Series

Kristian Seip

Abstract. We begin this survey by showing that Paley and Wiener's unconditional basis problem for nonharmonic Fourier series can be understood as a problem about weighted norm inequalities for Hilbert operators. Then we reformulate the basis problem in a more general setting, and discuss Beurling-type density theorems for sampling and interpolation. Next, we state some multiplier theorems, of a similar nature as the famous Beurling-Malliavin theorem, and sketch their role in the subject. Finally, we discuss extensions of nonharmonic Fourier series to weighted Paley-Wiener spaces, and indicate how these spaces are explored via de Branges' Hilbert spaces of entire functions.

1991 Mathematics Subject Classification: 30, 42, 46

1. From Paley-Wiener to Hunt-Muckenhoupt-Wheeden

The theory of nonharmonic Fourier series begins with Paley and Wiener [18], who discovered that the trigonometric system $\{e^{ikx}\}$ remains an unconditional basis for $L^2(-\pi,\pi)$ when the integer frequencies k are replaced by "nonharmonic" frequencies λ_k satisfying $|\lambda_k - k| \le d$ for some $d < 1/\pi^2$. This result led to quite extensive activity around the problem of describing all unconditional bases of the form $\{e^{i\lambda_k x}\}$ for $L^2(-\pi,\pi)$. A decisive breaktrough was made by Pavlov [19], and a complete solution to the problem as just stated is now available [9,12,15].

We shall present below a survey of recent developments which are closely related to the problem of Paley and Wiener. Let us therefore begin by clarifying how the unconditional basis problem can be understood: It can be recast as a question concerning boundedness of Hilbert operators in certain weighted L^2 (or more generally L^p) spaces of functions and sequences, and thus leads us to the Hunt-Muckenhoupt-Wheeden theorem [7]. We will follow [12], in which this shift from Hilbert space geometry to weighted norm inequalities is made.

We restate the Paley-Wiener problem in terms of entire functions. Denote by PW^p $(0 < p \le \infty)$ the classical Paley-Wiener spaces, which consist of all entire functions of exponential type at most π whose restrictions to the real line are in L^p. We endow PW^p with the natural $L^p(\mathbb{R})$-norms, and note that they are Banach spaces when $1 \le p \le \infty$ and complete metric spaces when $0 < p < 1$. For $1 < p < \infty$, we say that a sequence of complex numbers $\Lambda = \{\lambda_k\}$, $\lambda_k = \xi_k + i\eta_k$ is a *complete interpolating sequence* for PW^p if the interpolation problem $f(\lambda_k) = a_k$ has a unique solution $f \in PW^p$ for every sequence $\{a_k\}$ satisfying

$$\sum_k |a_k|^p e^{-p\pi|\eta_k|}(1 + |\eta_k|) < \infty.$$

Via the Paley-Wiener theorem, it is found that Λ is a complete interpolating sequence for PW^2 if and only if the system $\{e^{i\lambda_k x}\}$ is an unconditional basis for $L^2(-\pi,\pi)$.

Let us see how the Hilbert operator comes into play when we seek to describe complete interpolating sequences. Suppose Λ is a complete interpolating sequence for PW^p, $1<p<\infty$. Let us assume for simplicity that all the points of Λ lie in a horizontal strip, that $0\notin\Lambda$, and $\xi_k\le\xi_{k+1}$ for all k. It is easy to show that Λ has to be a separated sequence, i.e., $\inf_{j\neq k}|\lambda_j-\lambda_k|>0$, and also that it must be uniformly dense, i.e., that $\sup_k(\xi_{k+1}-\xi_k)<\infty$. In what follows, one should think of Λ roughly as an arithmetic progression.

If the function $f_0\in PW^p$ solves the interpolation problem $f_0(\lambda_k)=\delta_{0,k}$, $k\in\mathbb{Z}$, then $f_0(\mu)\neq 0$ for $\mu\in\mathbb{C}\setminus\Lambda$, since otherwise the function $(z-\lambda_0)(z-\mu)^{-1}f_0(z)$ belongs to PW^p and vanishes on Λ, contradicting the uniqueness of the solution of the interpolation problem. It is a short step from this observation to conclude that the limit

$$S(z)=\lim_{R\to\infty}\prod_{|\lambda_k|<R}(1-z/\lambda_k) \tag{1}$$

exists and defines an entire function of exponential type π. This function is called the *generating function* of the sequence Λ. It follows that if $a=\{a_j\}$ is a sequence such that $a_j=0$ except for finitely many j's, the unique solution of the interpolation problem $f(\lambda_j)=a_j$, has the form

$$f(z)=\sum_j \frac{a_j}{S'(\lambda_j)}\frac{S(z)}{(z-\lambda_j)}.$$

Now if $\Gamma=\{\gamma_j\}$ is any other separated and uniformly dense sequence lying in a horizontal strip, a classical inequality of Plancherel and Pólya [11, pp. 50–51] shows that

$$\sum_j |f(\gamma_j)|^p \lesssim \int_{\mathbb{R}}|f(x)|^p dx.$$

(We write $g\lesssim h$ whenever there is a positive constant C such that $g\le Ch$, and $g\simeq h$ if both $g\lesssim h$ and $h\lesssim g$.) Because the solution of the interpolation problem is unique, the open mapping theorem implies that $\sum|f(\lambda_j)|^p\simeq\int|f(x)|^p dx$, and so

$$\sum_j |f(\gamma_j)|^p \lesssim \sum_j |a_j|^p. \tag{2}$$

We claim that this inequality is just a weighted norm inequality for a discrete Hilbert operator. To see this, let ℓ^p_w be the space of all sequences $b=\{b_k\}$ satisfying $\|b\|^p_{w,p}:=\sum|b_k|^p w_k<\infty$ for some positive weight sequence $w=\{w_j\}$. If we put $u=\{|S'(\lambda_j)|^p\}$ and $v=\{|S(\gamma_j)|^p\}$, (2) says that the Hilbert operator $H_{\Lambda,\Gamma}:\ell^p_u\to\ell^p_v$ defined as

$$(\mathcal{H}_{\Lambda,\Gamma}b)_j=\sum_k\frac{b_k}{\gamma_j-\lambda_k},$$

is a bounded operator.

So far we have not assumed anything about Γ, except that it is separated and uniformly dense. We may in fact tailor it specifically to Λ in such a way that the weights u and v become identical, apart from a multiplicative constant. To see how this can be done, set $\varepsilon = \inf_{j\neq k} |\lambda_j - \lambda_k|/3$, and observe that since S has no zeros in the disk $|z - \lambda_j| \leq \varepsilon$, we can find a point γ_j with $|\gamma_j - \lambda_j| = \varepsilon$ and

$$|S(\gamma_j)| = \varepsilon |S'(\lambda_j)|.$$

We are now in a familiar situation, and obtain in accordance with the celebrated Hunt-Muckenhoupt-Wheeden theorem [7] that the weight $w = \{|S'(\lambda_j)|^p\}$ must satisfy a discrete Muckenhoupt (A_p) condition:

$$\sup_{k\in\mathbb{Z},n>0} \left(\frac{1}{n}\sum_{j=k+1}^{k+n} w_j\right)\left(\frac{1}{n}\sum_{j=k+1}^{k+n} w_j^{-\frac{1}{p-1}}\right)^{p-1} < \infty. \tag{3}$$

The analogy is clear: The classical continuous (A_p) condition for a positive weight $v(x) > 0,\ x \in \mathbb{R}$ is

$$\sup_{I} \left\{\left(\frac{1}{|I|}\int_I v dx\right)\left(\frac{1}{|I|}\int_I v^{-\frac{1}{p-1}} dx\right)^{p-1}\right\} < \infty, \tag{4}$$

where I ranges over all intervals in $\mathbb{R}$, and the Hunt-Muckenhoupt-Wheeden theorem [7] says that (3) is necessary and sufficient for boundedness of the classical Hilbert operator on the weighted space of functions $L^p(\mathbb{R}; vdt)$. It is clear that (3) is essentially a special case of (4). In fact, in our case, we may use either of the conditions, because it may be proved that (3) with $w = \{|S'(\lambda_j)|^p\}$ is equivalent to (4) with $v = |S(x)/\mathrm{dist}(x, \Lambda)|^p$.

The above reasoning has provided an essential piece of evidence for the main theorem of [12], which we will now state. We remove the assumption that Λ be located in a horizontal strip. It is then convenient to introduce the distance function

$$\delta(z, \zeta) = \frac{|z - \zeta|}{1 + |z - \overline{\zeta}|},$$

which expresses that we deal with Euclidean geometry close to the real axis and hyperbolic geometry far away from the real axis. We say that Λ is *δ-separated* if $\inf_{j\neq k} \delta(\lambda_j, \lambda_k) > 0$. Moreover, Λ is said to satisfy the *two-sided Carleson condition* if for any square Q of side-length $l(Q)$ and with one of its sides sitting on the real axis, we have

$$\sum_{\lambda_k\in Q\cap\Lambda} |\Im\lambda_k| \leq Cl(Q),$$

with C independent Q.

The main theorem of [12] is:

THEOREM 1. *A sequence $\Lambda = \{\lambda_k\}$ of complex numbers is a complete interpolating sequence for PW^p $(1 < p < \infty)$ if and only if the following three conditions hold.*

(i) *The sequence Λ is δ-separated and satisfies the two-sided Carleson condition.*

(ii) *The limit $S(z)$ in* (1) *exists and represents an entire function of exponential type π.*

(iii) *The weight $(|S(x)|/\mathrm{dist}(x,\Lambda))^p$ $(x \in \mathbb{R})$ satisfies the (A_p) condition* (4).

This theorem should be read in the following way: Condition (i) is a separation condition in which the Carleson condition is present because we solve in particular an interpolation problem in H^p; (ii) is mainly a density condition, as it gives the type of S; (iii) is a condition on the "balance" of the sequence. It is in fact a working condition, if one makes use of the equivalence between the (A_2) and Helson-Szegö conditions. For instance, the so-called Kadets 1/4 theorem, which says that $|\lambda_k - k| \le d < 1/4$ is the best possible inequality in the Paley-Wiener condition, is a direct consequence (see [9]). A similar perturbation result can be proved for PW^p, as shown in [12].

2. Beurling-type density theorems for sampling and interpolation

Stated as an interpolation problem for entire functions, the Paley-Wiener basis problem makes sense for a large class of holomorphic spaces. In this section, we shall extend the setting, and then consider the complementary situation that complete interpolating sequences are nonexistent. Building on a basic contribution by Beurling [2], who considered a problem of balayage of Fourier-Stieltjes transforms and a corresponding interpolation problem, we reformulate the Paley-Wiener problem by seeking to describe separately so-called sampling and interpolating sequences. Again, problems of this type can be traced back to classical work on nonharmonic Fourier series [5,8]; for the modern state of research on such nonharmonic Fourier series, see [21].

Assume we are given a weighted L^p space of holomorphic functions defined on some domain Ω in the complex plane. We denote this space by $\mathcal{B}$ and assume that the functional of point evaluation $f \mapsto f(z)$ is bounded for each $z \in \Omega$. The norm of this functional is called the *majorant* of $\mathcal{B}$, and it is denoted by $M(z)$. If $p = 2$, then $\mathcal{B}$ is a Hilbert space and $M(z) = \sqrt{K(z,z)}$, where $K(z,\zeta)$ is the reproducing kernel of the space. We say that a sequence of distinct points $\Lambda = \{\lambda_k\}$ in Ω is a *sampling sequence* for $\mathcal{B}$ if $\|f\|_{\mathcal{B}} \simeq \|\{f(\lambda_k)/M(\lambda_k)\}\|_{\ell^p}$ for $f \in \mathcal{B}$. We say that Λ is an *interpolating sequence* for $\mathcal{B}$ if the interpolation problem $f(\lambda_k) = a_k$ has a solution $f \in \mathcal{B}$ whenever $\{a_k/M(\lambda_k)\} \in \ell^p$. Finally, we say that Λ is a *complete interpolating sequence* for $\mathcal{B}$ if it is both sampling and interpolating. It is not difficult to check (using the open mapping theorem) that this definition is in line with the one given in the previous section.

Saying that a complete interpolating sequence is both a sampling and an interpolating sequence is a way of expressing that it exists as a compromise between two competing density conditions: A sampling sequence should be uniformly "dense", while an interpolating sequence should be uniformly "sparse". However, the reasoning of the previous section shows that there is more to it than only competing

density conditions: Existence of complete interpolating sequences is tied to norm inequalities for Hilbert operators between weighted L^p spaces. This means that we can expect to find such sequences only when $1 < p < \infty$ and in spaces with a special underlying geometry.

In this section, we shall present an aspect of the following striking dichotomy: *Geometric density conditions characterize sampling and interpolating sequences if and only if there are no complete interpolating sequences.* Of course, we are not able to claim that the truth of this statement is universal, but it covers at least three wide classes of model spaces: weighted Paley-Wiener spaces PW^p_ψ (to be considered in Section 4), weighted Fock spaces F^p_ψ, and weighted Bergman spaces A^p_ψ (to be defined shortly). In all three cases, the growth of functions is controled by e^ψ, where ψ is a subharmonic function whose Laplacian has an appropriate behavior compared to the underlying geometry: in the Paley-Wiener case, $\Delta\psi$ is supported by the real line and the Riesz measure of ψ is $\mu(x)dx$, with $\mu(x) \simeq 1$; in the Fock case, $\Delta\psi(z) \simeq 1$ for all $z \in \mathbb{C}$; in the Bergman case, $\Delta\psi(z) \simeq (1-|z|^2)^{-2}$ for all z in the unit disk $\mathbb{D}$. A common feature is that density conditions for sampling and interpolation are expressed in terms of $\Delta\psi$. We note that *it is only for Paley-Wiener spaces that we have weighted norm inequalities for the Hilbert operators attached to the possible complete interpolating sequences.*

We comment first on the Fock and Bergman cases, which are similar. We will only present results for Bergman spaces; the Fock case has been treated in the recent paper [17]. The results to be presented here for Bergman spaces are new, and we shall sketch proofs which are quite different from those of [17]. We call these results *Beurling-type density theorems*, because results of this type were first presented by Beurling and because certain parts of Beurling's analysis seem indispensable in whatever setting we consider.

We need to give a precise definition of the weighted Bergman spaces A^p_ψ. Suppose a subharmonic function ψ on the unit disk is given, whose Laplacian satisfies $\Delta\psi(z) \simeq (1-|z|^2)^{-2}$ for all $z \in \mathbb{D}$. Let dm denote Lebesgue area measure on $\mathbb{C}$. Define

$$\|f\|^p_{\psi,p} = \int_{\mathbb{D}} |f(z)|^p e^{-p\psi(z)} \, (1-|z|^2)^{-1} dm(z)$$

for $p < \infty$, and $\|f\|_{\psi,\infty} = \sup_z |f(z)|e^{-\psi(z)}$. We denote by A^p_ψ $(0 < p \le \infty)$ the set of all functions f analytic in $\mathbb{D}$ such that $\|f\|_{\psi,p} < \infty$. A prime example is obtained by setting $\psi(z) = -\beta \log(1-|z|^2)$, with $\beta > 0$.

Now set

$$\rho(z,\zeta) = \left| \frac{z-\zeta}{1-\overline{\zeta}z} \right|,$$

which is the pseudohyperbolic distance between z and ζ. We say that a sequence $\Lambda = \{\lambda_j\}$ is *ρ-separated* if $\inf_{j \neq k} \rho(\lambda_j, \lambda_k) > 0$. For a fixed ρ-separated sequence Λ, we denote by $n(z,r)$ the number of points $\lambda_k \in \Lambda$ which satisfy $\rho(z,\lambda_k) < r$, and set correspondingly

$$a_\psi(z,r) = \int_{\rho(z,\zeta)<r} \Delta\psi(\zeta) dm(\zeta).$$

The *lower uniform density of* Γ *with respect to* ψ is defined as

$$D^-_\psi(\Lambda) = \liminf_{r\to 1^-} \inf_{z\in\mathbb{D}} \frac{\int_0^r n(z,t)dt}{\int_0^r a_\psi(z,t)dt},$$

and the *upper uniform density of* Γ *with respect to* ψ is

$$D^+_\psi(\Lambda) = \limsup_{r\to 1^-} \sup_{z\in\mathbb{D}} \frac{\int_0^r n(z,t)dt}{\int_0^r a_\psi(z,t)dt}.$$

We have then the following two Beurling-type density theorems.

THEOREM 2. *A sequence* Λ *is sampling for* A^p_ψ *if and only if it contains a* ρ*-separated subsequence* Λ' *satisfying* $D^-_\psi(\Lambda') > 1/\pi$ *and in addition, when* $0 < p < \infty$*, it is a finite union of* ρ*-separated sequences.*

THEOREM 3. *A sequence* Λ *is interpolating for* A^p_ψ *if and only if it is* ρ*-separated and satisfies* $D^+_\psi(\Lambda) < 1/\pi$.

For $\psi(z) = -\beta\log(1-|z|^2)$ these are the main results of [20]. In the next section, we will sketch how the general case follows from these special results, via a certain multiplier theorem. Here we restrict ourselves to making two remarks concerning the proof for $\psi(z) = -\beta\log(1-|z|^2)$; in this case, with a slight abuse of notation, we set $A^p_\psi = A^p_\beta$ and $\|\cdot\|_{\psi,p} = \|\cdot\|_{\beta,p}$.

First, we would like to point out what is the core of Beurling's approach as it appears when transferred to $\mathbb{D}$. Namely, A^p_β enjoys the following group invariance: If τ is a Möbius self-map of $\mathbb{D}$, the operator T_τ defined by

$$(T_\tau f)(z) = (\tau'(z))^{\beta+1/p} f(\tau(z))$$

acts isometrically on A^p_β. This implies that sampling and interpolating sequences are Möbius invariant, and in fact, by a normal family argument, any compact-wise limit of a sequence $\tau_n\Lambda$, where τ_n are Möbius self-maps of $\mathbb{D}$, is sampling/interpolating if Λ is sampling/interpolating. An analysis of such compact-wise limits plays an essential role in Beurling's scheme. This part of Beurling's proof is of a general nature and is applicable whenever we have a suitable group invariance; we refer to [16] for a discussion of how the notion of "group invariance" can be extended to spaces with general weights.

Our second remark concerns the proof of the sufficiency of the density condition for interpolation. In [22], this was done by first relating the upper uniform density to a density used by Korenblum for describing the zeros of functions in A^∞_β, and then use this relation to construct a linear operator of interpolation. A less intricate and more direct proof, using Hörmander-type L^2 estimates for $\bar\partial$, has later been given by Berndtsson and Ortega-Cerdà [1]. This approach works also for F^p_ψ.

We end this section with a few words about the original interpolation problem considered by Beurling [2], to illustrate that "Beurling-type" density conditions may be rather subtle. Beurling considered interpolating values only along the real

axis, in which case uniform densities of real sequences yield a complete description. If we permit complex sequences Λ, we are led to combine techniques from entire functions and Hardy spaces in a nontrivial manner, and to solve simultaneously an interpolation problem in H^∞.

Suppose Λ is δ-separated, and let h be a positive number. Denote by $n_h^+(r)$ the maximum number of points from Λ to be found in a rectangle of the form $\{z = x + iy : \ t < x < t + r, \ |y| < h\}$, where t is any real number. The upper uniform density of Λ is defined to be

$$D^+(\Lambda) = \lim_{h\to\infty} \lim_{r\to\infty} \frac{n_h^+(r)}{r}.$$

We have then the following "mixed" Beurling-type and Carleson theorem.

THEOREM 4. *A sequence Λ is interpolating for PW^∞ if and only if it is δ-separated, satisfies the two-sided Carleson condition, and $D^+(\Lambda) < \tau/\pi$.*

This result is proved in [17]. A key ingredient in the proof will be presented in the next section. There is of course a similar result for the sampling problem, but it is more elementary. The result holds also when PW^∞ is replaced by PW^p, $p < 1$, which is an easier case than $p = \infty$.

3. The role of multipliers

The most distinguished example of a *multiplier theorem* is the following deep result of Beurling and Malliavin [3,10]: *If f is an entire function of exponential type with bounded logarithmic integral,*

$$\int_{\mathbb{R}} \frac{\log^+ |f(x)|}{1 + x^2} dx < \infty,$$

then, for every $\varepsilon > 0$ there exists an entire function g of exponential type ε with both $|g|$ and $|fg|$ bounded on the real axis.

In this section, we discuss how certain more modest multiplier theorems fit into our theory. As for the Beurling-Malliavin theorem, proofs are based on atomizing Riesz measures of certain subharmonic functions, but the details are quite straightforward in our case. However, it should be noted that we obtain more precise estimates on what corresponds to the product $|fg|$ above. This is why these multiplier theorems have an interesting role to play in our subject.

For the Paley-Wiener case, we have the following mulitplier theorem:

THEOREM 5. *Suppose Λ is a δ-separated sequence and ω is a subharmonic function of the form*

$$\psi(z) = \int_{-\infty}^{\infty} [\log|1 - z/t| + (1 - \chi_{[-1,1]}(t))\Re z/t]\mu(t)dt, \tag{5}$$

where $\mu(t) \simeq 1$. Then there exists an entire function g with δ-separated zero sequence $Z(g)$ lying in a horizontal strip, with $\delta(\Lambda, Z(g)) > 0$, and such that $|g(z)|e^{-\psi(z)} \simeq \delta(z, \Gamma)$.

The corresponding result for the disk is:

THEOREM 6. *Suppose Λ is a ρ-separated sequence in $\mathbb{D}$, and let ϕ be subharmonic in $\mathbb{D}$ so that its Laplacian $\Delta\phi$ satisfies $\Delta\phi(z) \simeq (1-|z|^2)^{-2}$ for all $z \in \mathbb{D}$. Then there exists a function g analytic in $\mathbb{D}$, with ρ-separated zero sequence $Z(g)$ and $\rho(Z(g),\Lambda) > 0$, and such that $|g(z)| \simeq \rho(z, Z(g))e^{\phi(z)}$.*

There is also an analogous result for the Fock case [14]. The two theorems above are in fact inspired by that result. A proof of Theorem 5 can be found in [16], while Theorem 6 is a slight variant of Theorem 2 of [22].

Theorem 5 is a key ingredient in the proof of Theorem 4. It is used both to transform Beurling's interpolation problem into an H^∞ problem, and to "correct" H^∞ solutions to produce solutions which are entire functions. We give only a hint how the first transformation is done. If we assume Λ satisfies the conditions of Theorem 4 and set $\varepsilon = 1-D^+(\Lambda)$, then Theorem 5 yields the existence of a function h vanishing on Λ, and satisfying the estimate $|h(z)| \simeq e^{\pi(1-\varepsilon/2)|\Im z|}\delta(z, Z(h))$, where $Z(h)$ is the zero sequence of h. (Incidentally, this argument shows that every interpolating sequence for PW^∞ is contained in a sequence which is a complete interpolating sequence for each of the spaces PW^p, $1 < p < \infty$. It is a striking fact that, on the other hand, there exists an interpolating sequence for PW^2 which is not a subsequence of any complete interpolating sequence for PW^2, as shown in [21].)

Next, we sketch how Theorem 6 can be used to prove Theorems 2 and 3 from the case of regular weights $-\beta\log(1-|z|^2)$. To this end, we begin by showing that A^p_ψ can be embedded into A^p_β for a sufficently large β: Choose β so large that $\phi(z) = \beta\log(1/(1-|z|^2)) - \psi(z)$ is a subharmonic function satisfying the condition of Theorem 6. Taking g to be the function of Theorem 6, it is clear that $f \in A^p_\psi$ if and only if $fg \in A^p_\beta$, and that $\|f\|_{\psi,p} \simeq \|f\|_{\beta,p}$. In other words, *we may associate A^p_ψ with the closed subspace of A^p_β which consists of functions vanishing on $Z(g)$.*

We now take Λ to be the ρ-separated sequence of Theorem 6, and claim that then Λ *is sampling/interpolating for A^p_ψ if and only if $Z(g)\cup\Lambda$ is sampling/interpolating for A^p_β.* The sufficiency of the condition $Z(g) \cup \Lambda$ being sampling/interpolating is trivial in view of the observation we just made, while the necessity can be obtained from the fact that $Z(g)$ is interpolating for A^p_β, as follows from Theorem 3 in the case of regular weights. Now Theorems 2 and 3 follow from the regular case by a simple rewriting of the density conditions.

For other applications of Theorem 6, see [6,22].

4. From de Branges to weighted Paley-Wiener spaces

Suppose ψ is a subharmonic function in $\mathbb{C}$ of the form (5) with $\mu(t) \simeq 1$. Set $w = e^{-\psi}$, and define

$$\|f\|^p_{w,p} = \int_{\mathbb{R}} |f(t)w(t)|^p\, dt$$

for $p < \infty$, and $\|f\|_{w,\infty} = \sup_z |f(z)|e^{-\psi(z)}$. We denote by PW^p_ψ $(0 < p \le \infty)$ the set of all entire functions f such that $\|f\|_{w,p} < \infty$ and $\log|f(z)| \le C_\varepsilon + \psi(z) + \varepsilon|z|$ for all $\varepsilon > 0$. The Phragmén-Lindelöf principle ensures that these spaces are complete with respect to their norms.

Following the reasoning at the end of Section 3, we may extend Theorem 1 and Theorem 4 to cover these weighted Paley-Wiener spaces. Thus our choice of weights is natural if we wish to see how far the basic results of nonharmonic Fourier series can be extended. But weighted Paley-Wiener spaces are interesting for other reasons. One particularly interesting point is the connection to de Branges' Hilbert spaces of entire functions [4], and that this link can be used to explore the nature of weighted Paley-Wiener spaces. We shall briefly indicate how this may work. The presentation is based on [14], where a complete treatment can be found. We stick from now on to the Hilbert space case $p = 2$.

A natural question is: Why is our choice of weights $e^{-\psi}$ reasonable? It is quite easy to see that our condition on the weight implies $M(x)w(x) \simeq 1$. By means of de Branges' theory, we can prove that this relation, which is a regularity condition on w, in fact characterizes weighted Paley-Wiener spaces. To be more precise, suppose $\mathcal{H}$ is a Hilbert space of entire functions whose norm is given by $\|\cdot\|_{w,2}$, where w is a positive weight function. We assume the functional of point evaluation is bounded for each $z \in \mathbb{C}$, and further that $\mathcal{H}$ is closed under the operations $f(z) \mapsto f(z)(z-\overline{\zeta})/(z-\zeta)$ (provided $f(\zeta) = 0$) and $f(z) \mapsto f^*(z)$, where $f^*(z) = \overline{f(\overline{z})}$. If $M(x)w(x) \simeq 1$, we say that w is a *majorant weight*. Then the following holds:

THEOREM 7. *A positive function w is a majorant weight for some space $\mathcal{H}$ if and only if there exists a function $\mu(x) \simeq 1$ and a real entire function g such that*

$$\log w(x) + g(x) + \int_{-\infty}^{\infty} [\log|1 - x/t| + (1 - \chi_{[-1,1]}(t))x/t]\mu(t)dt \in L^\infty. \tag{6}$$

The function g represents an inessential part of the weight, because a replacement of w by we^{-g} corresponds to multiplying all functions in $\mathcal{H}$ by a factor $\exp g/2$. Then assuming $g \equiv 0$, it is plain from de Branges' theory that $M(x)w(x) \simeq 1$ and the form of w together force $\mathcal{H}$ to be a weighted Paley-Wiener space.

It is interesting to note that if w has a bounded logaritmic integral, the condition (6) of Theorem 7 says that we have a representation $\log w = u + v$, with $u \in L^\infty$ and $(\tilde{v})' \in L^\infty$, where $\tilde{v}$ denotes the Hilbert transform of v.

The proof of Theorem 7 is based on converting the problem in the following way. By de Branges' theory, $\mathcal{H}$ coincides with a de Branges space $H(E)$; here E is an entire function without zeros in the upper half-plane, $|E(z)| \geq |E(\bar{z})|$ for all $\Im z > 0$, and $f \in H(E)$ if and only if f/E and f^*/E both belong to H^2 of the upper half-plane. That $\mathcal{H} = H(E)$ means in particular that $\|f/E\|_2 = \|f\|_{w,2}$ for all $f \in \mathcal{H}$. With this relation established, the proof becomes a problem of exploring the distribution of the zeros of E. To give a hint about the nature of the problem, we mention the following result. Let $\Lambda = \{\xi_k - i\eta_k\}$ denote the zero sequence of E, and suppose $\xi_k \leq \xi_{k+1}$ for all k. Then: *$H(E)$ equals (up to norm equivalence) a weighted Paley-Wiener space if and only if Λ is uniformly dense and a finite union of separated sequences, the sequence $\Lambda \cap \{z : \Im z > -\varepsilon\}$ is separated for some $\varepsilon > 0$, and $\{\eta_k\}$ is a discrete (A_2) weight.* The analogue of Theorem 1 for PW^2_ψ is used to prove the last part of this statement.

REFERENCES

1. B. Berndtsson and J. Ortega-Cerdà, *On interpolation and sampling in Hilbert spaces of analytic functions*, J. Reine Angew. Math. **464** (1995), 109–120.
2. A. Beurling, *The Collected Works of Arne Beurling*, vol. 2, Ed. L. Carleson et al., Birkhäuser, Boston, 1989, pp. 341–365.
3. A. Beurling and P. Malliavin, *On Fourier transforms of measures with compact support*, Acta Math. **107** (1962), 291–309.
4. L. de Branges, *Hilbert Spaces of Entire Functions*, Prentice-Hall, Englewood Cliffs, N.J., 1968.
5. R. J. Duffin and A. C. Schaeffer, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Soc. **72** (1952), 341–366.
6. H. Hedenmalm, S. Richter, and K. Seip, *Interpolating sequences and invariant subspaces of given index in the Bergman spaces*, J. Reine Angew. Math. **477** (1996), 13–30.
7. R. Hunt, B. Muckenhoupt, and R. Wheeden, *Weighted norm inequalities for the conjugate function and Hilbert transform*, Trans. Amer. Math. Soc. **176** (1973), 227–251.
8. A. E. Ingham, *Some trigonometrical inequalities with applications to the theory of series*, Math. Z. **41** (1936), 367–379.
9. S. V. Khrushchev, N. K. Nikol'skii, and B. S. Pavlov, *Unconditional bases of exponentials and reproducing kernels*, in "Complex Analysis and Spectral Theory", Lecture Notes in Math. Vol. 864, Springer-Verlag, Berlin/Heidelberg, 1981, pp. 214–335.
10. P. Koosis, *Leçons sur le Théorème de Beurling et Malliavin*, Les Publications CRM, Montréal, 1996.
11. B. Ya. Levin, *Lectures on Entire Functions*, Transl. Math. Monographs, Vol. 150, American Mathematical Society, Providence, 1996.
12. Yu. I. Lyubarskii and K. Seip, *Complete interpolating sequences for Paley-Wiener spaces and Muckenhoupt's A_p condition*, Rev. Mat. Iberoamericana **13** (1997), 361–376.
13. Yu. I. Lyubarskii and K. Seip, *Weighted Paley-Wiener spaces* (to appear).
14. Yu. I. Lyubarskii and M. L. Sodin, *Analogues of sine type functions for convex domains*, Preprint no. 17, Inst. Low Temperature Phys. Engrg., Ukrainian Acad. Sci., Kharkov (1986), (Russian).
15. A. M. Minkin, *Reflection of exponents, and unconditional bases of exponentials*, St. Petersburg Math. J. **3** (1992), 1043–1068.
16. J. Ortega-Cerdà and K. Seip, *Beurling-type density theorems for weighted L^p spaces of entire functions*, J. Analyse Math. (1998) (to appear).
17. J. Ortega-Cerdà and K. Seip, *Multipliers for entire functions and an interpolation problem of Beurling* (to appear).
18. R. E. A. C. Paley and N. Wiener, *Fourier Transforms in the Complex Domain*, American Mathematical Society, New York, 1934.
19. B. S. Pavlov, *Basicity of an exponential system and Muckenhoupt's condition*, Dokl. Akad. Nauk SSSR **247** (1979), 37–40; English transl. in Sov. Math. Dokl. **20** (1979).
20. K. Seip, *Beurling type density theorems in the unit disk*, Invent. Math. **113** (1993), 21–39.
21. K. Seip, *On the connection between exponential bases and certain related sequences in $L^2(-\pi,\pi)$*, J. Funct. Anal. **130** (1995), 131–160.
22. K. Seip, *On Korenblum's density condition for the zero sequences of $A^{-\alpha}$*, J. Analyse Math. **67** (1995), 307–322.

Kristian Seip
Dept. of Mathematical Sciences
Norwegian University of
Science and Technology
N-7034 Trondheim
Norway
seip@math.ntnu.no

Wave Equations with Low Regularity Coefficients

Hart F. Smith

Abstract. We illustrate how harmonic analysis techniques that were developed to understand the L^p mapping properties of oscillatory integral and Fourier integral operators lead to an understanding of solutions to the wave equation on Riemannian manifolds with metrics of limited differentiability.

1991 Mathematics Subject Classification: 35L05, 35R05, 35S30, 35S50
Keywords and Phrases: Wave equation, Strichartz estimates

1. L^p mapping properties of Fourier integral operators

For the purposes of this section, a standard Fourier integral operator of order m is a finite sum of operators of the form

$$Tf(x) = \int e^{i\varphi(x,\xi)}\, a(x,\xi)\, \widehat{f}(\xi)\, d\xi\,. \tag{1.1}$$

The phase function $\varphi(x,\xi)$ is real, homogeneous of degree 1 in ξ, smooth for $\xi \neq 0$, and satisfies the nondegeneracy condition

$$\det\left[\frac{\partial^2\varphi}{\partial x_i \partial \xi_j}\right] \neq 0\,.$$

The amplitude $a(x,\xi)$ is a standard amplitude of order m, which for convenience is also taken compactly supported in x:

$$\left|\partial_x^\beta \partial_\xi^\alpha a(x,\xi)\right| \leq C_{\alpha,\beta}\left(1+|\xi|\right)^{m-|\alpha|}.$$

The most important examples are the two terms of the wave group:

$$\begin{aligned}
\mathbf{C}_t f(x) &= \int e^{i\langle x,\xi\rangle}\, \cos\bigl(t\,|\xi|\,\bigr)\, \widehat{f}(\xi)\, d\xi\,,\\
\mathbf{S}_t f(x) &= \int e^{i\langle x,\xi\rangle}\, \frac{\sin\bigl(t\,|\xi|\,\bigr)}{|\xi|}\, \widehat{f}(\xi)\, d\xi\,.
\end{aligned}$$

For each fixed t, these are standard Fourier integral operators, respectively of order 0 and -1, with two phases $\varphi^{\pm}(x,\xi) = \langle x,\xi\rangle \pm t|\xi|$. The importance of these operators is that the solution to the Cauchy problem for the wave equation

$$\begin{aligned}
\partial_t^2 u(t,x) &= \sum_{j=1}^n \partial_{x_j}^2 u(t,x)\,,\\
u(0,x) &= f(x)\,,\\
\partial_t u(0,x) &= g(x)\,,
\end{aligned}$$

is given by $u(t,x) = \mathbf{C}_t f(x) + \mathbf{S}_t g(x)$.

By a theorem of Hörmander [H] and Eskin [E], Fourier integral operators of order 0 are bounded on the space $L^2(\mathbb{R}^n)$. For the spaces $L^p(\mathbb{R}^n)$, $p \neq 2$, this is not the case, and examples of Littman [Li] show that the following result of Seeger-Sogge-Stein [SSS] is of the best possible nature.

THEOREM. *Let T be a Fourier integral operator of order $m = -(n-1)\,|1/p - 1/2|$, and $1 < p < \infty$. Then T is a bounded operator on $L^p(\mathbb{R}^n)$. If $m = -(n-1)/2$, then T is a bounded operator on the local Hardy space $h^1(\mathbb{R}^n)$.*

Lipschitz and L^p estimates for the wave equation on compact manifolds were obtained by Colin de Vèrdiere and Frisch [CFr]. For operators related to the wave equation the above was demonstrated by Peral [Per], M. Beals [Be], and Miyachi [M].

The key to establishing the above theorem is to break up the operator (1.1) into simple pieces using a partition of unity in the ξ variable. The first step is to make a Littlewood-Paley decomposition by splitting the ξ space into dyadic annuli $2^{k-1} \leq |\xi| \leq 2^{k+1}$. A finer decomposition involving the angular variable is then made in a parabolic manner: the shell $|\xi| \approx \lambda$ is divided into conic sets of opening angle $\lambda^{-1/2}$.

The motivation behind this decomposition is that on each resulting region in ξ, the homogeneous phase function $\varphi(x,\xi)$ is well approximated by a phase which is *linear* in ξ, in the sense that the error is uniformly controlled. Each piece of the operator is then essentially a localisation of $\widehat{f}(\xi)$ to a cube, followed by a change of coordinates, and has uniformly bounded norm on $L^1(\mathbb{R}^n)$. The loss of $(n-1)/2$ derivatives on the local Hardy space results from the fact that, at frequencies comparable to λ, the operator is a sum of $\lambda^{(n-1)/2}$ pieces, each of which acts independently on $h^1(\mathbb{R}^n)$. For details, see [SSS] or chapter IX of [St].

This dyadic-parabolic decomposition is implicit in the work of C. Fefferman [F], where it was exploited to understand spherical summation multipliers. For the wave operators, it is related to approximate plane-wave decompositions of solutions: if the function $f(x)$ has Fourier transform localised to a dyadic shell $|\xi| \approx \lambda$, and within angle $\lambda^{-1/2}$ about some direction ω, then for $|t| \lesssim 1$,

$$\mathbf{C}_t f(x) \approx \frac{1}{2}\big(\, f(x + t\omega) - f(x - t\omega)\,\big)\,,$$

with errors that can be uniformly controlled as $\lambda \to \infty$.

The above theorem can be sharpened by the following result [Sm1], which is natural in view of the fact that order 0 Fourier integral operators form an algebra.

THEOREM. *There exists a function space $\mathcal{H}^1_{\text{FIO}}(\mathbb{R}^n)$, with continuous mappings*

$$D^{-(n-1)/2} : h^1(\mathbb{R}^n) \longrightarrow \mathcal{H}^1_{\text{FIO}}(\mathbb{R}^n) \longrightarrow h^1(\mathbb{R}^n)\,,$$

on which the order 0 Fourier integral operators are bounded mappings.

The norm of a function f in $\mathcal{H}^1_{\text{FIO}}$ is defined by the integral of a quadratic expression in f, analogous to the Lusin area characterisation of functions in the real Hardy space of Fefferman-Stein [FS]. The appropriate area function for $\mathcal{H}^1_{\text{FIO}}$ is evaluated at a spatial point and a direction; in essence, each dyadic-parabolic piece of f is treated independently.

2. Strichartz Estimates

Of greater interest for nonlinear wave equations than the preceeding fixed time estimates for solutions of the wave equation are the family of *Strichartz estimates*, which control mixed L^p norms of a solution over space and time, in terms of Sobolev norms of the initial data. For simplicity, we restrict attention here to space dimension $n = 3$.

Theorem. *Let $u(t,x) = \mathbf{C}_t f(x) + \mathbf{S}_t g(x)$ be the solution to the Cauchy problem for the wave equation. Then for $2 \le q < \infty$, if $1/p + 1/q = 1/2$, the following hold:*

$$\|u\|_{L_t^p L_x^q(\mathbb{R}^{1+3})} \le C_q \left(\|f\|_{\dot{H}^{1-2/q}(\mathbb{R}^3)} + \|g\|_{\dot{H}^{-2/q}(\mathbb{R}^3)} \right). \tag{2.1}$$

In the form stated here, (2.1) is due to Pecher [Pec]. The original Strichartz estimate [Str1,2] is the case $p = q = 4$. More general estimates of this type, in general dimensions, have been developed by several authors, including Brenner [Br], Ginibre and Velo [GV1,2], Kapitanski [K], Keel and Tao [KT], and Lindblad and Sogge [LS].

In case $q = 2, p = \infty$, estimate (2.1) is an energy inequality. The other endpoint estimate at $q = \infty$ does not hold; this would state that

$$\|u\|_{L_t^2 L_x^\infty(\mathbb{R}^{1+3})} \le C \left(\|f\|_{\dot{H}^1(\mathbb{R}^3)} + \|g\|_{L^2(\mathbb{R}^3)} \right).$$

(The failure of this estimate motivates the study of the important *null form* estimates of Klainerman-Machedon [KM].) However, a substitute estimate does hold, which is sufficient to obtain (2.1) for $q < \infty$ by interpolation. To state this estimate, let

$$\mathbf{C}_t^\lambda f(x) = \int e^{i\langle x,\xi\rangle} \cos\big(t\,|\xi|\,\big)\, \phi\big(\lambda^{-1}\,|\xi|\,\big)\, \widehat{f}(\xi)\, d\xi\,,$$

where $\phi(s)$ is supported in $1/2 \le s \le 2$. Then

$$\|\mathbf{C}_t^\lambda f\|_{L^\infty(\mathbb{R}^3)} \le C\,\lambda^2\, t^{-1}\, \|f\|_{L^1(\mathbb{R}^3)}\,. \tag{2.2}$$

This says that the convolution kernel associated to $\mathbf{C}_t^\lambda$ is pointwise bounded by $\lambda^2\, t^{-1}$, which can be demonstrated by stationary phase arguments.

For $n = 3$, a proof of (2.2) (for $|t| \lesssim 1$ and $\lambda \ge 1$) can be obtained using the dyadic-parabolic decomposition mentioned in the first section of this paper; this is important since it allows for a broader class of amplitudes in the Fourier integral operator, which is crucial for low regularity wave equations.

To begin, write

$$\phi\big(\lambda^{-1}\,|\xi|\,\big) = \sum_\omega \widehat{\psi}_\lambda^\omega(\xi)\,, \tag{2.3}$$

where $\widehat{\psi}_\lambda^\omega(\xi)$ is supported in a cone of angle $\lambda^{-1/2}$ about the direction ω, and ω varies over λ indices evenly distributed over the unit sphere. The function $\psi_\lambda^\omega(x)$

is of L^∞ norm comparable to λ^2, and is concentrated in a box with two sides of length $\lambda^{-1/2}$, and one side of length λ^{-1}, the last along the direction ω. The function

$$\int e^{i\langle x,\xi\rangle - it|\xi|}\, \widehat{\psi}^\omega_\lambda(\xi)\, d\xi \tag{2.4}$$

is a "coherent wave packet of frequency λ", in the sense that for $|t| \lesssim 1$ it travels along a ray without significantly changing its shape. We remark that this function is also critical for the Strichartz estimates, in that the two sides of (2.1) are comparable as $\lambda \to \infty$.

The convolution kernel of $\mathbf{C}^\lambda_t$ splits into a sum

$$\sum_\omega \int e^{i\langle x,\xi\rangle}\, \cos\big(t\,|\xi|\,\big)\, \widehat{\psi}^\omega_\lambda(\xi)\, d\xi \approx \frac{1}{2}\sum_\omega \psi^\omega_\lambda(x+t\omega) + \psi^\omega_\lambda(x-t\omega)\,.$$

Then (2.2) follows by showing that the overlap of "supports" of the $\psi^\omega_\lambda(x+t\omega)$ is bounded by t^{-1}, which is a simple exercise in geometry. (We remark that this simple proof fails in space dimension $n \geq 4$, where the overlap count is too high.)

3. The wave equation on Riemannian manifolds

Let

$$\Delta_{\mathbf{g}} f(x) = \frac{1}{\sqrt{\mathbf{g}(x)}} \sum_{i,j=1}^n \partial_{x_i}\Big(\sqrt{\mathbf{g}(x)}\; \mathbf{g}^{ij}(x)\, \partial_{x_j} f(x)\Big)$$

be the Laplace-Beltrami operator for a smooth Riemannian metric $\mathbf{g}$ in a coordinate patch. The Cauchy problem for the wave equation

$$\begin{aligned} \partial_t^2 u(t,x) &= \Delta_{\mathbf{g}} u(t,x)\,, \\ u(0,x) &= f(x)\,, \\ \partial_t u(0,x) &= g(x)\,, \end{aligned} \tag{3.1}$$

has finite propagation speed, so for small time intervals it suffices to work in a coordinate neighborhood.

To solve the Cauchy problem, one seeks the analogue of the plane wave solutions $\exp(i\langle x,\xi\rangle \pm it|\xi|)$. Lax [Lax] provided an asymptotic construction of solutions for small t of the form

$$e^{i\varphi^\pm(t,x,\xi)}\, a_\pm(t,x,\xi)\,,$$

where $a_\pm(t,x,\xi)$ is a standard amplitude of order 0 (which equals 1 at $t=0$), and the real phase $\varphi^\pm(t,x,\xi)$ satisfies the eikonal equation

$$\begin{aligned} \partial_t \varphi^\pm(t,x,\xi) &= \pm\big\|d_x\varphi^\pm(t,x,\xi)\big\|_{\mathbf{g}}\,, \\ \varphi^\pm(0,x,\xi) &= \langle x,\xi\rangle\,. \end{aligned}$$

The solution to the Cauchy problem (for initial condition $g = 0$) can be written (up to an error which is a smooth integral kernel aplied to f) in the form

$$u(t,x) = \frac{1}{2} \sum_{\pm} \int e^{i\varphi^{\pm}(t,x,\xi)}\, a_{\pm}(t,x,\xi)\, \widehat{f}(\xi)\, d\xi\,.$$

Using stationary phase techniques, the estimate (2.2) can be shown to hold for small time intervals, and together with L^2 bounds on Fourier integral operators this implies the Strichartz estimates (2.1) locally, as shown by Kapitanski [K], and Mockenhaupt-Seeger-Sogge [MSS].

4. Low regularity metrics

Consider the following question: what is the minimal regularity condition on the metric coefficients $\mathbf{g}^{ij}(x)$ which insures that the Strichartz estimates hold for solutions $u(t,x)$ to the Cauchy problem (3.1)?

A natural condition for geometric optics is that the metric coefficients possess two bounded derivatives; that is, $\mathbf{g}^{ij}(x) \in C^{1,1}(\mathbb{R}^n)$. This is the minimal regularity condition in the Hölder classes which yields a unique, bilipschitz geodesic flow. That this condition is also optimal among the Hölder classes for Strichartz estimates is shown by the following counterexamples of the author and Sogge [SS].

Theorem. *For $n \geq 3$, and any $\alpha < 1$, there exists $h_\alpha(x) \in C^{1,\alpha}(\mathbb{R}^n)$, and a solution $u(t,x)$ to the Cauchy problem for*

$$\partial_t^2 u(t,x) = h_\alpha(x)\, \Delta u(t,x)\,,$$

for which the Strichartz estimates do not hold.

The function $h_\alpha(x)$ is constructed so that the geodesic flow is singularly focused along some ray. This permits the construction of coherent wave packets travelling along the ray which, due to the singular focusing, are contained in smaller sets than the coherent wave packets (2.4) that are critical for the Strichartz estimates.

On the other hand, the arguments at the end of section 2 show that a positive proof of the Strichartz estimates (in space dimensions 2 and 3) for a metric $\mathbf{g}$ can be reduced to studying wave packets. Roughly, one needs to show that the solution to the Cauchy problem with initial condition $\psi_\lambda^\omega(x)$ is a coherent wave packet that travels along the geodesic x in direction ω. Together with a bilipschitz geodesic flow, this implies the analogue of estimate (2.2).

In [Sm2], this idea was coupled with a decomposition of functions into wave packets to construct the wave group for metrics $\mathbf{g}(t,x) \in C^{1,1}\big(\mathbb{R}^{1+n}\big)$. Modifying techniques of Frazier and Jawerth [FJW] permits the construction of a spanning set of functions for $L^2(\mathbb{R}^n)$ consisting of translates of the $\psi_\lambda^\omega(x)$. The ansatz that the function $\psi_\lambda^\omega(x)$ is rigidly transported along the geodesic flow leads to an inverse for the wave equation, modulo an error that can be eliminated by iteration.

To obtain a manageable class of operators, however, a modification is needed: the function ψ_λ^ω is transported not along the geodesic flow of the metric $\mathbf{g}$, but rather along the flow of a smooth approximation $\mathbf{g}_\lambda$ to $\mathbf{g}$, where the approximation

is chosen depending on the frequency λ, analogous to the paraproduct/multilinear Fourier analysis techniques of Bony [Bo], Coifman and Meyer [CM]. We outline this approximation in the next section in the context of a modified parametrix construction for metrics of bounded curvature.

5. Metrics of bounded sectional curvature

In this section, we assume $\mathbf{g}(x)$ to be a Riemannian metric such that all sectional curvatures are pointwise bounded by some constant C; this is the notion of L^∞ pinched curvature. Some a priori regularity is necessary to make sense of the Riemann curvature tensor, which is a nonlinear expression in the derivatives of $\mathbf{g}$; the condition $\nabla_x \mathbf{g}^{ij}(x) \in L^q(\mathbb{R}^n)$ for some $q > n$ is sufficient. This is also sufficient to construct local harmonic coordinates for $\mathbf{g}$. Lanczos [Lan] observed that in these coordinates the Ricci curvature is an elliptic expression in terms of $\mathbf{g}$ (see DeTurk and Kazdan [DK]); consequently in such coordinates the metric has all second partial derivatives belonging to $\mathrm{BMO}(\mathbb{R}^n)$, which we henceforth assume.

Take a sequence of smooth approximating metrics $\mathbf{g}_k(x)$ to $\mathbf{g}(x)$ by the rule

$$\mathbf{g}_k^{ij}(x) = \big(\, \phi_k * \mathbf{g}^{ij} \,\big)(x)\,,$$

where $\phi_k(x) = 2^{nk/2}\phi(2^{k/2}x)$, with $\phi(x)$ a smooth bump function of integral 1.

It follows from the condition $\nabla_x^2 \mathbf{g}^{ij}(x) \in \mathrm{BMO}(\mathbb{R}^n)$ that

$$\big\|\, \mathbf{g}_k^{ij} - \mathbf{g}^{ij} \,\big\|_{L^\infty(\mathbb{R}^n)} \lesssim 2^{-k}\,. \tag{5.1}$$

Let $\varphi_k^\pm(t,x,\xi)$ be the solutions to the eikonal equations for $\mathbf{g}_k$:

$$\begin{aligned} \partial_t \varphi_k^\pm(t,x,\xi) &= \pm \big\| d_x \varphi_k^\pm(t,x,\xi) \big\|_{\mathbf{g}_k}\,, \\ \varphi_k^\pm(0,x,\xi) &= \langle x,\xi\rangle\,. \end{aligned} \tag{5.2}$$

It follows from the bounded sectional curvature condition that the geodesic flow of $\mathbf{g}_k$ is bilipschitz, uniformly in k, hence that the $\varphi_k^\pm(t,x,\xi)$ form a bounded sequence in $C^2(\mathbb{R}^7)$. Let

$$\mathbf{S}_t g(x) = \frac{1}{2i} \sum_{k=0}^\infty \int \Big(e^{i\varphi_k^+(t,x,\xi)} - e^{i\varphi_k^-(t,x,\xi)} \Big)\, \|\xi\|_{\mathbf{g}_k(x)}^{-1}\, \widehat{g}_k(\xi)\, d\xi\,, \tag{5.3}$$

where $g = \sum_k g_k$ is a Littlewood-Paley decomposition of g. It then follows from (5.1) and (5.2) that

$$\big(\partial_t^2 - \Delta_{\mathbf{g}}\big)\mathbf{S}_t = \mathbf{R}_t$$

is a bounded operator on the Sobolev spaces $H^\gamma(\mathbb{R}^n)$, for $-1 \le \gamma \le 2$, with norms uniformly bounded in t.

One then seeks a solution to the inhomogeneous Cauchy problem

$$\begin{aligned} \partial_t^2 u(t,x) &= \Delta_{\mathbf{g}} u(t,x) + F(t,x)\,, \\ u(0,x) &= 0\,, \\ \partial_t u(0,x) &= 0\,, \end{aligned}$$

in the form

$$u(t,x) = \int_0^t \mathbf{S}_{t-s}\, G(s,x)\, ds\,.$$

This leads to a Volterra equation,

$$F(t,x) = G(t,x) + \int_0^t \mathbf{R}_{t-s} G(s,x)\, ds$$

which may be solved by iteration.

Estimates of the form (2.1) are thus reduced to $L^2 \to L^p$ mapping properties of operators of the form (5.3). The symbols and phases of these operators satisfy exactly the estimates needed to use the decomposition of Seeger-Sogge-Stein. In particular, functions of the form ψ_λ^ω (see (2.3)) are mapped to coherent wave packets. Combined with the ideas at the end of section 2, this yields the following (for details see [Sm3]).

Theorem. *Let* $\mathbf{g}(x)$ *be a Riemannian metric on an open ball in* $\mathbb{R}^3$ *such that, for* $1 \le i,j \le 3$, $\nabla_x^2 \mathbf{g}^{ij}(x) \in \mathrm{BMO}(\mathbb{R}^3)$. *Suppose also that the components of the Riemannian curvature tensor satisfy, for all indices,* $R^{ijkl}(x) \in L^\infty(\mathbb{R}^3)$. *Then for* t *in some interval about* 0, *solutions to the Cauchy problem* (3.1) *satisfy the Strichartz estimates* (2.1).

Acknowledgments

The author gratefully acknowledges the support of the National Science Foundation, and of an Alfred P. Sloan Research Fellowship, for the research presented in this paper.

References

[Be] M. Beals, *L^p Boundedness of Fourier integral operators*, Memoirs of the Amer. Math. Soc. **264** (1982).

[Bo] J.N. Bony, *Calcul symbolique et propagation des singularités pour les équations aux dérivées partielles non linéaires*, Ann. Scient. E.N.S. **14** (1981), 209–246.

[Br] P. Brenner, *On $L^p - L^{p'}$ estimates for the wave equation*, Math Z. **145** (1975), 251–254.

[CM] R.R. Coifman and Y. Meyer, *Au delà des opérateurs pseudo-differentials,* Astérisque **57**, Soc. Math. France, 1978.

[CFr] Y. Colin de Verdière and M. Frisch, *Régularité Lipschitzienne et solutions de l'équation des ondes sur une variété Riemannienne compacte*, Ann. Scient. Ecole Norm. Sup. **9** (1976), 539–565.

[CFe] A. Cordoba and C. Fefferman, *Wave packets and Fourier integral operators*, Comm. Partial Differential Equations **3(11)** (1978), 979–1005.

[DK] D. DeTurk and J. Kazdan, *Some regularity theorems in Riemannian geometry*, Ann. Sci. Ecole. Norm. Sup **14** (1981), 249–260.

[E] G. I. Eskin, *Degenerate elliptic pseudodifferential operators of principal type*, Mat. Sbornik **82** (1970), 585-628 (Russian); English transl. in Math. USSR Sbornik **11** (1970), 539–582.

[F] C. Fefferman, *A note on spherical summation multipliers*, Israel J. Math. **15** (1973), 44–52.

[FS] C. Fefferman and E. M. Stein, *H^p spaces of several variables*, Acta. Math. **129** (1972), 137–193.

[FJW] M. Frazier, B. Jawerth, and G. Weiss, *Littlewood-Paley Theory and the Study of Function Spaces*, CBMS Regional Conference Ser. # 79, American Math. Society, 1991.

[GV1] J. Ginibre and G. Velo, *Conformal invariance and time decay for nonlinear wave equations II*, Ann. Inst. Poincaré (Physique Théorique) **47** (1987), 263–276.

[GV2] ———, *Scattering theory in the energy space for a class of nonlinear wave equations*, Comm. Math. Phys. **123** (1989), 535–573.

[H] L. Hörmander, *Fourier integral operators I*, Acta Math. **127** (1971), 79–183.

[K] L. Kapitanski, *Some generalizations of the Strichartz-Brenner inequality*, Leningrad Math. J. **1** (1990), 693–726.

[KT] M. Keel and T. Tao, *Endpoint Strichartz estimates*, Amer. J. Math (to appear).

[KM] S. Klainerman and M. Machedon, *Space-time estimates for null forms and the local existence theorem*, Comm. Pure. Appl. Math. **46** (1993), 1221–1268.

[Lan] C. Lanczos, *Ein vereinfachendes koordinatesystem fur die Einsteinschen gravitationsgleichungen*, Phys. Z **23** (1922), 537–539.

[Lax] P. Lax, *Asymptotic solutions of oscillatory initial value problems*, Duke Math. J. **24** (1957), 627–646.

[LS] H. Lindblad and C. Sogge, *On existence and scattering with minimal regularity for semilinear wave equations*, J. Funct. Anal. **130** (1995), 357–426.

[Li] W. Littman, *$L^p - L^q$ estimates for singular integral operators*, Proc. Symp. Pure and Appl. Math. AMS **23** (1973), 479–481.

[M] A. Miyachi, *On some estimates for the wave equation in L^p and H^p*, J. Fac. Sci. Tokyo **27** (1980), 331–354.

[MSS] G. Mockenhaupt, A. Seeger and C. D. Sogge, *Local smoothing of Fourier integrals and Carleson-Sjölin estimates*, J. Amer. Math. Soc. **6** (1993), 65–130.

[Pec] H. Pecher, *Nonlinear small data scattering for the wave and Klein-Gordan equations*, Math. Z. **185** (1984), 261–270.

[Per] J. Peral, *L^p Estimates for the wave equation*, J. Funct. Anal. **36** (1980), 114–145.

[SSS] A. Seeger, C.D. Sogge, and E.M. Stein, *Regularity properties of Fourier integral operators*, Annals Math. **133** (1991), 231–251.

[Sm1] H. Smith, *A Hardy space for Fourier integral operators*, Jour. Geom. Anal. **7** (1997).

[Sm2] ———, *A parametrix construction for wave equations with $C^{1,1}$ coefficients*, Annales de l'Institut Fourier **48** (1998).

[Sm3] ———, *Strichartz and nullform estimates for metrics of bounded curvature*, Preprint.

[SS] H. Smith and C. Sogge, *On Strichartz and eigenfunction estimates for low regularity metrics*, Math. Res. Lett. **1** (1994), 729–737.

[St] E. M. Stein, *Harmonic Analysis: Real Variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton University Press, Princeton, 1993.

[Str1] R. Strichartz, *A priori estimates for the wave equation and some applications*, J. Funct. Anal. **5** (1970), 218–235.

[Str2] ———, *Restrictions of Fourier transforms to quadratic surfaces and decay of solutions to wave equations*, Duke Math. J. **44** (1977), 705–714.

Hart F. Smith
University of Washington
Department of Mathematics
Box 354350
Seattle, Washington 98195-4350

From Finite- to Infinite-Dimensional Phenomena in Geometric Functional Analysis on Local and Asymptotic Levels

Nicole Tomczak-Jaegermann[1]

1991 Mathematics Subject Classification: 46B20, 46B07, 46B09

The geometry and linear-metric structure of high dimensional convex bodies make an essential contribution to the understanding of the geometry, structure and some purely infinite-dimensional properties of Banach spaces. An asymptotic approach that studies finite-dimensional geometric properties "stabilized at infinity" makes it possible to identify regularities behind an apparent lack of structure. Recently, a deeper understanding of the infinite nature of Banach spaces has opened possibilities to study some previously intractable linear-topological problems by refined essentially finite-dimensional methods. By putting together certain sophisticated finite-dimensional random constructions we can create new phenomena of infinite flavour in arbitrary Banach spaces.

1. Finite-dimensional phenomena We start the discussion of "random quotients" of finite-dimensional normed spaces. Properties of such spaces reveal a striking interplay between high dimensional geometry and the linear structure of normed spaces. We shall also briefly mention some related properties of Gaussian matrices. Consider the following theorem (the terminology is explained below).

Theorem 1 *For $0 < \alpha < 1$ and $K \geq 1$ there exists $f(\alpha, K) > 0$ such that for every $n \geq 1$, whenever X is an n-dimensional normed space all of whose $[\alpha n]$-dimensional subspaces are K-isomorphic, then X is $f(\alpha, K)$-isomorphic to ℓ_2^n.*

This result is an isomorphic finite-dimensional version of two questions from Banach's book ([Ba32]): regarding an n-dimensional symmetric convex body all of whose k-dimensional sections are affinely equivalent, and the homogeneous Banach space problem. For the former question see Gromov's work [Gr67]; the solution to the latter was obtained by Gowers [G94a], in conjunction with [KT95], and will be discussed later.

Theorem 1 was proved by Bourgain in [B87] for sufficiently small α, and in [MT88] for all α, yielding the function $f(\alpha, K)$ less than $cK^{3/2}$ for $0 < \alpha < 2/3$, and cK^2, for $2/3 \leq \alpha < 1$, where $c = c(\alpha)$. The general line of an argument (the same in both papers) depends on two separate parts of the theory. The first part studies Euclidean sections of convex bodies, yielding upper estimates for the distance of such sections to ellipsoids. It was initiated in the late 60s and has

[1]The author held Canada Council Killam Research Fellowship in 1997/99.

been well developed throughout the intervening period, resulting in the discovery of many deep results relating a variety of geometric characteristics (see Milman's surveys [Mi86b] and [Mi96]). The other part investigates lower estimates; it was initiated by the Gluskin result (below); its development in a general form started with the present theorem. For lower estimates it is natural to work with quotient spaces (corresponding to projections of convex bodies); then the results for subspaces (corresponding to sections of convex bodies) follow by duality. Volume-type characteristics that appear in both parts are different, though, and it requires additional ingenuous arguments to put them together.

We need some notation. For convenience, we describe the real case only, the complex case follows by standard modifications. On $\mathbb{R}^n$ we consider the natural Euclidean norm $\|\cdot\|_2$, and by B_2^n we denote the closed Euclidean unit ball. By ℓ_1^n we denote $\mathbb{R}^n$ with the unit ball $B_1^n = \{x = (a_i) \mid \sum |a_i| \le 1\}$. Any convex compact centrally symmetric body $B \subset \mathbb{R}^n$ determines the normed space E for which B is the unit ball, and any n-dimensional normed space has (many) representations of such a form. The polar body B° is the unit ball in the dual space E^*. By $\{e_i\}$ we denote the unit vector basis of $\mathbb{R}^n$. By $\mathrm{vol}_n(\cdot)$ we denote the n-dimensional Lebesgue measure. If X_1, X_2 are isomorphic Banach spaces, the Banach–Mazur distance is defined by $d(X_1, X_2) = \inf \|T\| \, \|T^{-1}\|$, with the infimum running over all isomorphisms T from X_1 onto X_2; if $d(X_1, X_2) \le d$, we say that the spaces are d-isomorphic. For $B \subset \mathbb{R}^n$ we let

$$v_k(B) = \sup_{F,\, \dim F = k} \left(\mathrm{vol}_k P_F B \big/ \mathrm{vol}_k B_2^k\right)^{1/k}, \qquad \text{for } 1 \le k \le n,$$

where P_F is the orthogonal projection on a subspace $F \subset \mathbb{R}^n$.

Let $E = (\mathbb{R}^n, B_E)$; by properly identifying E with $\mathbb{R}^n$ we may further assume that B_2^n is the ellipsoid of minimal volume containing B_E. The simplest form of a lower estimate used in the proof in Theorem 1 says ([MT88]): *Let $0 < \alpha < 1$. There exist $m = [\alpha n]$-dimensional quotients F_1, F_2 of E such that*

$$d(F_1, F_2) \ge c(\alpha) v_{[m/4]}(B_E)^{-2}. \tag{1}$$

(In fact, these quotients are "random", in sense to be explained shortly.)

Estimates as in (1) are a conceptualization of the discovery of Gluskin in [Gl81a] (see also [Gl86]), who determined an asymptotic growth of the diameter of the Minkowski compactum of all m-dimensional normed spaces, by showing that: *There exists $c > 0$ such that for every $m \ge 1$ there exist ("random") m-dimensional quotients of ℓ_1^{3m}, F_1 and F_2 such that $d(F_1, F_2) \ge cm$.* (By the classical John theorem, $d(F_1, F_2) \le m$, for all m-dimensional normed spaces.)

Gluskin's new point of view triggered extensive investigations of a natural class of "random"quotients of ℓ_1^n, by a number of researchers (including Mankiewicz and Szarek among others). Further studies showed that the resulting bodies are "rigid", in a sense that the underlying normed spaces admit few well bounded linear operators. This is nicely expressed using the notion of mixing operators ([S86], [M88]). An operator $T \in L(\mathbb{R}^m)$ is called k-mixing, where $1 \le k \le m/2$, if there exists a subspace $H \subset \mathbb{R}^m$ with $\dim H \ge k$ such that $\|P_{H^\perp} T x\|_2 \ge \|x\|_2$, for every $x \in H$ (where $P_{H^\perp}$ is the orthogonal projection from $\mathbb{R}^m$ onto $H^\perp$). Note that for every projection P with $k = \operatorname{rank} P \le m/2$, $2P$ is k-mixing. We shall concentrate on quotients of proportional dimension.

Theorem 2 ([S83], [S86]) *There is $c > 0$ such that for every integer $m \geq 1$ there is an m-dimensional quotient of ℓ_1^{2m}, F, such that every projection P on F with $m/4 \leq \operatorname{rank} P \leq 3m/4$, satisfies $\|P : F \to F\| \geq c\sqrt{m}$. More generally, for every $[m/4]$-mixing operator T on $\mathbb{R}^m$, $\|T : F \to F\| \geq c\sqrt{m}$.*

The first statement ([S83]) settled the so-called finite-dimensional basis problem (solved independently in [Gl81b]), showing an example of a sequence of finite-dimensional spaces F_n with $\mathrm{bc}(F_n) \to \infty$. Let us recall the fundamental classical definition. A sequence $\{x_i\}$ in a Banach space X is a Schauder basis, if every $x \in X$ admits the unique representation as a convergent series $x = \sum_i a_i x_i$. In such a situation, for $k = 1, 2, \ldots$, define the projections $P_k : X \to X$ by $P_k(x) = \sum_{i=1}^k a_i x_i$, for $x \in X$. Then $\mathrm{bc}(\{x_i\}) = \sup_k \|P_k\| < \infty$. If a Banach space X has a Schauder basis, the basis constant of X is defined as $\mathrm{bc}(X) = \inf \mathrm{bc}(\{x_i\})$, where the infimum is taken over all bases $\{x_i\}$ in X. So clearly, $\mathrm{bc}(F) \geq c\sqrt{m}$, for F as in the theorem.

An important aspect of these constructions is their random character. For problems requiring technically involved geometric phenomena the Gaussian setting is the easiest to use. Let $\gamma_1, \gamma_2, \ldots$ be independent real valued Gaussian variables with distribution $N(0,1)$. Let $m \geq 1$. Set $g = m^{-1/2} \sum_{i=1}^m \gamma_i e_i \in \mathbb{R}^m$. Let $n > m$ and $k = n - m$. Let $g_1, \ldots, g_k$ be independent $\mathbb{R}^m$-valued variables with the same distribution as g. Consider a *Gaussian projection* $Q_\omega : \mathbb{R}^n \to \mathbb{R}^m$ defined on the unit vector basis in $\mathbb{R}^n$ by $Q_\omega(e_i) = e_i$ for $1 \leq i \leq m$ and $Q_\omega(e_i) = g_{i-m}(\omega)$ for $m < i \leq n$. Given a normed space $E = (\mathbb{R}^n, B_E)$, by an m-dimensional *Gaussian quotient* of E we understand the space $F_\omega = (\mathbb{R}^m, B_{F_\omega})$ where $B_{F_\omega} = Q_\omega(B_E)$.

Of course this approach is related to Gaussian matrices. Convex geometric analysis discovered, often for its own needs, some deep results about such matrices. Since they may be of importance for many other areas of mathematics, we shall briefly digress to comment upon them.

Let $G = G_n(\omega)$ be an $n \times n$ matrix with independent Gaussian $N(0, 1/n)$ entries. Let $\{s_k(G)\}_{k \geq 1}$ be the sequence of singular (s-)numbers of G (i.e., the eigenvalues of $(G^*G)^{1/2}$ arranged in the non-increasing order, counting multiplicities). Their distribution is described by the classical Wigner Semi-circle Law [W55], which however has a qualitative character only. A quantitative distributional inequality was proved by Szarek in [S90]: *For $d \leq n/2$, $\mathcal{P}\{c_1 d/n \leq s_{n-d}(G) \leq c_2 d/n\} \geq 1 - C\exp(-cd^2)$, where $c_1, c_2, c, C > 0$ are absolute constants.* For further refinements and references see [S91]. In the other direction, Gordon studied (cf. e.g., [Go88], [Go92]) the majorization of Gaussian processes, in particular the maximum and the minimum of $\|G(x)\|_2$ over all $x \in B$, for an arbitrary symmetric convex body $B \subset \mathbb{R}^n$. Here G is possibly a rectangular $m \times n$ Gaussian matrix. For example, if $m = \alpha n < n$, this easily implies sharp estimates for the norms of G and of G^{-1} depending on α (established earlier e.g., via complicated combinatorial arguments [Ge80]), and many other geometric applications.

The above quotients of ℓ_1^n can be taken as Gaussian quotients; and the sets of (pairs of) ω's for which the lower estimates do not hold, have the measure exponentially small in n. In the last decade many sophisticated properties of random Gaussian quotients F of ℓ_1^n have been established, connected with factorizations of operators and the distance to the cube ([S90]), actions of compact groups of

operators ([M88], [M98]) and others.

When these constructions are considered in the framework of *arbitrary* normed spaces, as for example in Theorem 1, their random character becomes even more crucial. Randomness is the main reason for the connection of linear structure and volumes. Also, families of random projections (or sections) of high-dimensional convex bodies display a curious dichotomous behaviour: they are either nearly Euclidean or else, they have an unusually rigid structure as discussed above. Thus the rigidity becomes a "random alternative" to being Euclidean.

THEOREM 3 ([MT88], [MT94]) *Let $n \geq 1$ and let E' be a $2n$-dimensional normed space. There exists a quotient space E of E' with $\dim E = n$, and a Euclidean norm on E such that identifying E with $\mathbb{R}^n$ (and the Euclidean norm with $\|\cdot\|_2$), condition (1) is satisfied for a random pair of m-dimensional Gaussian quotients of E ($m = [\alpha n]$). Furthermore, letting $m = [99n/100]$, for a random m-dimensional Gaussian quotient F of E the estimate $\|T : F \to F\| \geq cv_{[m/100]}(B_E)^{-1}$ is valid for all $[m/10]$-mixing operators $T \in L(\mathbb{R}^m)$, with an absolute constant $c > 0$. Hence $\mathrm{bc}(F) \geq cv_{[m/100]}(B_E)^{-1}$ as well.*

For technical reasons, m has to be sufficiently close to n, but its specific value is of no importance. The estimates obtained are sharp: for $E = \ell_1^n$ we recover Gluskin's diameter result and Theorem 2. The preliminary step of passing from E' to E is designed to get the unit ball B_E in a "special position", i.e., having certain additional geometric properties with respect to the Euclidean structure in $\mathbb{R}^n$. This was achieved by using deep results from the convex geometric analysis: the inverse Brunn-Minkowski inequality ([Mi86a]) and the proportional Dvoretzky-Rogers Lemma ([BS88]).

It is also worthwhile to consider a more geometric approach to random families (of subspaces or quotients), through the orthogonal group. Let $\mathcal{G}_{n,m}$ denote the Grassmann manifold of all m-dimensional subspaces of $\mathbb{R}^n$, with the Haar measure. For $F \in \mathcal{G}_{n,m}$ let P_F be the orthogonal projection onto F. If $E = (\mathbb{R}^n, B_E)$ then F endowed with the unit ball $B_F = P_F(B_E)$ is a quotient space of E. We should mention, however, that the orthogonal approach is not equivalent to the Gaussian one; for example, it may be less sensitive to some involved structural properties of normed spaces.

Using more involved arguments it is possible to study invariants like these in Theorem 3 or others, for random quotients of the original space E, without passing to a special position. This reveals a striking threshold phenomenon, which, however, for some invariants can be quite indirect. For example, we have ([MT98b]): *Let E be an n-dimensional space identified with $\mathbb{R}^n$ in such a way that B_2^n is the ellipsoid of minimal volume containing B_E. There exists $1 \leq \varphi = \varphi_E$ such that:*

(i) "random" $(F_1, F_2) \in \mathcal{G}_{n,[n/2]} \times \mathcal{G}_{n,[n/2]}$ satisfies $d\big(P_{F_1}(B_E), P_{F_2}(B_E)\big) \geq \varphi$;

(ii) "random" $F \in \mathcal{G}_{n,[n/8]}$ satisfies $(c/\sqrt{\varphi})P_F(B_2^n) \subset P_F(B_E) \subset P_F(B_2^n)$, where $c > 0$ is an absolute constant.

(Here "random" means "on a set of positive measure".) Intuitively, for any normed space E, identified with $\mathbb{R}^n$ as above, for any fixed K, the only way in which a random pair of $[n/2]$-dimensional quotients of E may be closer together than K is

that random $[n/8]$-dimensional quotients of E are $C\sqrt{K}$-Euclidean. A kind of converse statement is trivially true: the distance between random $[n/8]$-dimensional quotients admits an upper bound by comparison with Euclidean space.

A detailed presentation of random quotients of finite-dimensional spaces, related infinite-dimensional constructions and an extensive bibliography, can be found in [MT98a].

2. Infinite-dimensional constructions A strong case for the emerging integration of finite-dimensional properties and the linear-topological structure of Banach spaces is made by the use of random quotient phenomena for constructions "inside" arbitrary Banach spaces. The first example combining finite-dimensional random quotients of ℓ_1^n into an infinite-dimensional space was given by Bourgain ([B86]), who constructed a real Banach space that admits two non-isomorphic complex structures. Then Szarek ([S87]) constructed a space without a sequence of uniformly bounded projections $\{P_n\}$ with $\sup_n \operatorname{rank}(P_n - P_{n-1}) < \infty$, hence without a Schauder basis.

At the root of these constructions lies a property still stronger than those discussed before: even adding to a quotient F the most regular space of all, does not remove an essential lack of well bounded operators. For example ([S86]): *A space F from Theorem 2 satisfies* $\operatorname{bc}(F \oplus_2 \ell_2) \geq cm^{1/4}$. This property can be formally deduced from a lower estimate for norms of all k-mixing operators on $\mathbb{R}^m$ ([MT94]), so an analogous fact holds in general too, by Theorem 3.

Before stating the next theorem recall that if X_n are Banach spaces, the ℓ_2-sum, $(\bigoplus X_n)_{\ell_2}$, is the Banach space of all sequences of vectors $z = (z_n)$, with $z_n \in X_n$ for all n, such that $\|z\|_{\oplus X_n} = (\sum \|z_i\|_{X_n}^2)^{1/2} < \infty$. If $X_n = X$ for all n, we write $\ell_2(X)$ instead of $(\bigoplus X)_{\ell_2}$.

The first construction of "gluing" together random quotients of finite-dimensional subspaces of an arbitrary Banach space X was done in [MT94] and it led to some interesting structural characterizations of Hilbert space. We give just one example.

Theorem 4 [MT94] *Let X be a Banach space such that every subspace of every quotient of $\ell_2(X)$ has a Schauder basis. Then X is isomorphic to Hilbert space.*

Thus the theory has made a full circle, that started from Enflo's example of a Banach space without the approximation property ([E73]). Spaces Z without a Schauder basis can now be constructed in just three canonical operations, of the ℓ_2-sum and taking subspaces and quotients, starting from an *arbitrary* Banach space X not isomorphic to ℓ_2. Moreover, such spaces are of the form $Z = (\bigoplus Z_n)_{\ell_2}$, where Z_n are *finite-dimensional* quotients of subspaces of $\ell_2(X)$. It should be noted that the presence of the ℓ_2-sum is necessary for a characterization of Hilbert space. Johnson ([J79]) constructed a Banach space X not isomorphic to ℓ_2, all of whose quotients of subspaces have a basis.

It should be emphasised that the hypothesis of the theorem gives no *a priori* information on uniform boundedness of the basis constants involved. This produces a strong infinite-dimensional flavor, which could not exist in specific examples. It is surprising that this effect has been obtained by a fundamentally local (finite-dimensional) approach.

Another direction of the interplay between finite- and infinite-dimensional techniques is illustrated by the homogeneous Banach space problem ([Ba32]): *If an infinite-dimensional Banach space X is isomorphic to all of its infinite-dimensional closed subspaces, is X isomorphic to ℓ_2?* As already mentioned, the problem was solved in the positive by combining Gowers' dichotomy theorem [G94a] (Theorem 6 below) and a result from [KT95]. Its history has been explained in detail in [G94b] so we wish to limit ourselves to just a few comments on the local approach involved.

Before going on, we recall the classical definition that non-zero vectors $\{z_i\}$ in a Banach space are *unconditional* if there is C such that for any scalars $\{a_i\}$ and a sequence $\{\varepsilon_i\}$ of signs, one has $\|\sum \varepsilon_i a_i z_i\| \leq C\|\sum a_i z_i\|$.

The first obvious difficulty in attacking the homogeneous space problem is the lack of information on uniform boundedness of norms of the isomorphisms. (Even up to this day no direct proof is known that if X is homogeneous then X is *uniformly* isomorphic to all of its infinite-dimensional subspaces.) Luckily, Gowers' dichotomy theorem combined with properties of H.I. spaces (discussed in the next section), enables us to quickly overcome this difficulty and to conclude that a homogeneous space X must have an unconditional basis. Then the theorem is concluded by a result from [KT95]:

THEOREM 5 *Let X be a Banach space with an unconditional basis. Then X contains either ℓ_2 or a subspace without an unconditional basis.*

There are two points worth making. Firstly, there exists a property of a space X, slightly weaker than having an unconditional basis, which is "local", that is, is determined by the behaviour of a certain numerical invariant on all finite-dimensional subspaces of X. Thus, once we find a sequence of finite-dimensional subspaces of X with this invariant tending to infinity, the closed span of these subspaces is a subspace without an unconditional basis. The second point is that under very mild geometric assumptions on X, the construction in Theorem 5 is combinatorial, and the resulting subspace preserves a lot of the structure of X. For example, it admits an unconditional decomposition into 2-dimensional subspaces (we omit a precise definition), and this implies that it has a Schauder basis.

3. ASYMPTOTIC INFINITE-DIMENSIONAL GEOMETRY The asymptotic approach to geometric infinite-dimensional properties can be exemplified by the notion of an asymptotic structure, which depends on possibility of "stabilizing" finite-dimensional subspaces "at infinity" ([MiT93], [MMT95]; a forerunner of this notion was studied in [MiSh79]). We shall give just few examples to indicate the possibilities and directions of such results. The main point is that this is the most general asymptotic geometric notion that can be defined for an *arbitrary* Banach space. It clearly yields a richer theory than the classical notion of spreading models, based on Ramsey's combinatorial theorem (see e.g. in [BL84] for the definition).

For simplicity, we consider a Banach space X with a basis $\{x_i\}$. We need some notation. The set of all positive integers is denoted by $\mathbb{N}$. For $F, G \subset \mathbb{N}$ we write $F < G$ whenever $\max F < \min G$, or either F or G is empty. For a vector $z = \sum a_i x_i \in X$, the support of z is $\operatorname{supp}(z) = \{i \mid a_i \neq 0\}$. A block is a vector with a finite support; blocks are successive, $z_1 < z_2$, whenever $\operatorname{supp}(z_1) <$

supp (z_2). Sequences of vectors $\{e_i\}$ and $\{z_i\}$ are C-equivalent ($C \geq 1$) if for all sequences of scalars $\{a_i\}$ we have $(1/\sqrt{C})\|\sum a_i z_i\| \leq \|\sum a_i e_i\| \leq \sqrt{C}\|\sum a_i z_i\|$.

An n-dimensional normed space E with a basis $\{e_i\}$ is an asymptotic space of X (we write $E \in \{X\}_n$), if there exist successive blocks $z_1, \ldots, z_n$, as close to $\{e_i\}$ as we wish, and arbitrarily far and arbitrarily spread out with respect to the basis. Precisely, given $\varepsilon > 0$, for an arbitrarily large m_1 there is a block z_1 with $\{m_1\} <$ supp (z_1) such that for an arbitrarily large m_2 there is a block z_2 with $\{m_2\} <$ supp (z_2), etc., such that the blocks $\{z_1, \ldots, z_n\}$ obtained after n steps are successive and $(1+\varepsilon)$-equivalent to $\{e_i\}$. The asymptotic structure of X consists of all asymptotic spaces of X.

The concept of asymptotic structure in a natural way describes classes of Banach spaces rather than individual spaces. For example, a space X is called an *Asymptotic-ℓ_p* space, $1 \leq p \leq \infty$ (note the capital A) if there exists C such that for all n and $E \in \{X\}_n$, the basis in E is C-equivalent to the unit vector basis in ℓ_p^n. Thus an Asymptotic-ℓ_p space has the simplest possible asymptotic structure—recall that by Krivine's theorem ([K76]) for every X there is $1 \leq p \leq \infty$ such that $\ell_p^n \in \{X\}_n$ for every n. In fact, a block structure is not so very important in this definition: if the equivalence condition is relaxed to the condition that for all n, all $E \in \{X\}_n$ are C-isomorphic to ℓ_p^n, we still get the same class of Asymptotic-ℓ_p spaces, for $1 < p < \infty$ ([MMT95]).

It can be shown ([MMT95]) that: *If X is an Asymptotic-ℓ_p space ($1 < p < \infty$), there exists C satisfying the condition that for all n, representations of all $E \in \{X\}_n$ which are C-complemented by block projections can be found arbitrarily far and arbitrarily spread out. Conversely, the complementation condition implies that X is an Asymptotic-ℓ_p space for some $1 \leq p \leq \infty$.* (A block projection is a projection of a form $Px = \sum z_i^*(x) z_i$, where the sets supp $(z_i) \cup$ supp (z_i^*), for $i = 1, 2, \ldots$, are successive.) For classical spaces ℓ_p and c_0 the first statement is trivial; but in the asymptotic setting it requires a non-obvious stabilization step. The converse statement seems to have a truly asymptotic nature: the validity of its classical analogue requires strong additional assumptions ([LT71]).

A general stabilization argument shows ([KOS98], [MiT95]) that in every Asymptotic-ℓ_p space X even a higher level of structure can be automatically reached: X contains a subspace Y with a basis such that there exists C that for every n, any n normalized blocks of the basis with supports after n, are C-equivalent to the unit vector basis in ℓ_p^n. Such spaces are called *asymptotic-ℓ_p*, $1 \leq p \leq \infty$.

The first truly non-classical Banach space was discovered by Tsirelson [Ts74]. The implicit definition of its unit ball effectively *saturates* the space with a certain geometric property (i.e., each infinite-dimensional subspace has this property) which prevents the space and its dual from containing ℓ_p, for $1 \leq p < \infty$, or c_0. Saturation of spaces with desired (often complicated) properties is the fundamental ingredient of some spectacular developments of recent years. In the dual setting, put forward in [FJ74], the norm on space T is defined implicitly as the solution of an equation. T and T^* are an asymptotic-ℓ_1 and an asymptotic-ℓ_∞, respectively. A detailed study of these spaces and some of their variants appears in [CS89].

More generally, an investigation of the successive block structure of

asymptotic-ℓ_1 spaces was done in [OTW97]. Among other results, natural geometric invariants have been introduced, localized to the Schreier families mentioned below, and certain regular behaviour was established. However, a non-block geometric structure of asymptotic-ℓ_1 spaces may be very diverse: for example a space may contain uniform copies of ℓ_∞^n for all n ([ADKM98]).

On the other hand, truly infinite-dimensional phenomena in general may not stabilize. This was first discovered as a conjunction of two results: a theorem by Milman [Mi69] and the above example by Tsirelson. However, this direction was not pursued for about 15 years. Only in the early 90s it became a central leitmotif in a series of breakthrough results by Gowers and Maurey [GM93], Odell and Schlumprecht [OS93] and Gowers [G94a] (see also the surveys [G94b] and [OS94]).

A passage between finite- and infinite-dimensional geometry may then be achieved by alternating localization and stabilization (as long as possible) of suitable invariants along hierarchies of families (with increasing complexity) of finite subsets of $\mathbb{N}$. This would result in saturating a space with combinatorial structures having required properties: each infinite-dimensional subspace would contain such a structure. An important, and in a sense universal, example of such a hierarchy, which unfortunately we have no place to describe, is given by Schreier families $\{\mathcal{S}_\alpha\}_{\alpha<\omega_1}$, introduced in [AA92]. (The concept of the asymptotic structure discussed above corresponds to family $\mathcal{S}_1$.)

Before we proceed, we need to briefly recall some of the phenomena involved.

In connection with the construction of a Banach space no subspace of which has an unconditional basis, a stronger property was identified in [GM93]: a space X is called *hereditarily indecomposable* (in short, H.I.) if no closed subspace Y of X can be written as a topological direct sum $W \oplus Z$, where W and Z are closed infinite-dimensional subspaces. The space constructed by Gowers and Maurey is H.I. The structure of the algebra $L(X)$ of bounded operators on an H.I. space X is particularly simple ([GM93]): *If X is H.I. and $T \in L(X)$ then $T = \lambda I + S$, where S is a strictly singular operator and λ is a scalar.* It is still an open question whether there exists a Banach space on which every bounded operator is a *compact* perturbation of a scalar, hence admits a non-trivial invariant subspace.

Another inspiring example was constructed by Argyros and Deliyanni [AD97]; their space is H.I. and asymptotic-ℓ_1: any n normalized blocks of the basis with supports after n are 2-equivalent to ℓ_1^n, the lack of stabilization, required in order that a space be H.I., depends on the Schreier families $\mathcal{S}_k$, when $k \to \infty$.

Recall the Gowers dichotomy theorem:

THEOREM 6 ([G94a]) *Every Banach space contains a subspace that either has an unconditional basis or is hereditarily indecomposable.*

A Banach space X is called λ-*distortable* if there exists an equivalent norm $|\cdot|$ on X such that $\inf_{Y\subset X} \sup\{|x|/|y| \mid x, y \in Y, \|x\| = \|y\| = 1\} > \lambda$; and is *arbitrarily distortable* if it is λ-distortable for every $\lambda > 1$. For a detailed report on this notion, and in particular, on Schlumprecht's example [Sch91], we refer to [OS93] and [O98]. Here let us only recall the solution of the distortion problem:

THEOREM 7 ([OS93]) *ℓ_p for $1 < p < \infty$ is arbitrarily distortable. Every Banach space contains ℓ_1 or c_0 or a λ-distortable subspace, for some $\lambda > 1$.*

A complete characterization of Banach spaces containing arbitrarily distortable subspaces is still unclear. Every Banach space either contains an arbitrarily distortable subspace or it contains a subspace of bounded distortion. This latter property means that there is $C < \infty$ such that any equivalent norm can be stabilized up to C, on a certain infinite-dimensional subspace of any given subspace Y. It was shown in [MiT93] that a space of bounded distortion contains an asymptotic-ℓ_p subspace, for some $1 \le p \le \infty$; and it was proved by Maurey ([Ma95]) that an asymptotic-ℓ_p space of type r for some $r > 1$, in which the basis is unconditional, is arbitrarily distortable. (A space has type r for some $r > 1$ if it does not contain copies of ℓ_1^n uniformly for all n.) Having the problem settled for a large class of spaces with an unconditional basis, Theorem 6 suggests that the next important case is that of hereditarily indecomposable spaces. It was widely expected that H.I. spaces should be arbitrarily distortable, and it is indeed so.

Theorem 8 ([T96]) *A Banach space X of bounded distortion contains a subspace with an unconditional basis. Consequently, any H.I. space is arbitrarily distortable.*

The main part of the argument uses the condition of bounded distortion to construct, for some fixed C, trees in X whose finite branches are built from C-unconditional sequences of successive blocks, and which have arbitrarily large countable ordinal index. An easy application of Kunen–Martin boundedness principle (see e.g., [D77]) shows the existence of a C-unconditional tree with an infinite branch, whose linear span will be the subspace with a C-unconditional basis.

As an immediate corollary we get that: *Every Banach space of type r for some $r > 1$ contains an arbitrarily distortable subspace.* This substantially limits the hypothetical possibility of the existence of a distortable space of bounded distortion. It would be very interesting if such a space existed, as it would demonstrate new geometric and combinatorial phenomena. The most prominent candidate is Tsirelson's space T (cf. e.g., [OTW97], [OT98]).

Returning to H.I. spaces, although their structure theory appears to have no bearing on spaces with an unconditional basis, a recent surprising and beautiful result of Argyros and Felouzis [AF98] shows that there is a direct connection between these two classes.

Theorem 9 ([AF98]) *Every Banach space either contains a subspace isomorphic to ℓ_1 or a subspace which is a quotient of an H.I. space. Furthermore, the class of Banach spaces which are quotients of H.I. spaces contains among others: spaces of type r for some $r > 1$ with an unconditional basis (in particular ℓ_p and L_p for $1 < p < \infty$), c_0, Tsirelson's space T and its dual.*

The proof of this result consists of two new essential ingredients. The first is an abstract interpolation scheme (originating in [DFJP74]) that yields a factorization of certain operators through H.I. spaces. The second is a geometric concept of thin sets, combined with an ingenious combinatorial construction of thin norming sets. The proof of the latter statement is geometric, while the former statement uses a rather complicated saturation argument.

References

[AA92] D. Alspach and S. Argyros: *Complexity of weakly null sequences*, Diss. Math., 321.

[AD97] S. Argyros and I. Deliyanni: *Examples of asymptotically ℓ^1 Banach spaces*, Trans. A.M.S., 349, 973–995.

[ADKM98] S. Argyros, I. Deliyanni, D. Kutzarova, A. Manoussakis: *Modified mixed Tsirelson spaces*, preprint.

[AF98] S. Argyros and Ferouzis: *Interpolating hereditarily indecomposable Banach spaces*, preprint.

[Ba32] S. Banach: *Theorie des operations lineaires*, Warszawa.

[BL84] B. Beauzamy and J.-T. Lapresté: *Modèles étalés des espace de Banach* Travaux en Cours, Herman, Paris.

[B86] J. Bourgain: *Real isomorphic complex Banach spaces need not to be complex isomorphic*, Proc. A.M.S., 98, 221–226.

[B87] J. Bourgain: *On finite-dimensional homogeneous Banach spaces*, Geom. Aspects of Funct. Anal., Israel Seminar 1986-87, Springer-Verlag LNM 1317, 232–239.

[BS88] J. Bourgain and S. J. Szarek: *The Banach-Mazur distance to the cube and the Dvoretzky-Rogers factorization*, Israel J. Math., 62, 169–180.

[CS89] P. G. Casazza and T. J. Shura: *Tsirelson's Space*, Springer-Verlag LNM 1363,

[DFJP74] W. J. Davis, T. Figiel, W. B. Johnson and A. Pełczyński: *Factoring weakly compact operators*, J. Funct. Anal. 17, 311–327.

[D77] C. Dellacherie: *Les derivations en theorie descriptive des ensembles et le theoreme de la borne*, Seminare de Prob. XI, Springer-Verlag LNM 581, 34–46.

[E73] P. Enflo: *A counterexample to the approximation problem in Banach spaces*, Acta Math., 130 (1973), 309–317; *Recent results on general Banach spaces*, Proc. I.C.M. Vancouver, 1974, Canad. Math. Congress, Montreal, 1975, pp. 53–55.

[FJ74] T. Figiel and W. B. Johnson: *A uniformly convex Banach space which contains no ℓ_p*, Comp. Math., 29, 179–190.

[Ge80] S. Geman: *A limit theorem for the norm of random matrices*, Ann. Probab., 8, 252–261.

[Gl81a] E. D. Gluskin: *The diameter of the Minkowski compactum is roughly equal to n*, Funct. Anal. Appl., 15, 72–73.

[Gl81b] E. D. Gluskin: *Finite-dimensional analogues of spaces without basis*, Doklady Acad. Nauk SSSR, 216, 5, 146–150.

[Gl86] E. D. Gluskin: *Probability in geometry of Banach spaces*, Proc. I.C.M. Berkeley, CA 1986, AMS Providence, RI, 1987, 924–938.

[Go88] Y. Gordon: *Gaussian processes and almost spherical sections of convex bodies*, Ann. Probab. 16, 180–188.

[Go92] Y. Gordon: *Majorization of Gaussian processes and geometric applications*, Probab. Theory Related Fields 91, 251–267.

[G94a] W. T. Gowers: *Analytic sets and games in Banach spaces*, preprint IHES M/94/42; *A new dichotomy for Banach spaces*, GAFA 6 (1996), 1083–1093.

[G94b] W. T. Gowers: *Recent Results in the Theory of Infinite-Dimensional Banach Spaces*, Proc. I.C.M. Zürich 1994, Birkhäuser Verlag, Basel 1995, 933–942.

[GM93] W. T. Gowers and B. Maurey: *The unconditional basic sequence problem*, Journal. A.M.S. 6, 851–874.

[Gr67] M. Gromov: *On a geometric conjecture of Banach*, Izv. Akad. Nauk SSSR, Ser. Mat. 31, 1105–1114.

[J79] W. B. Johnson: *Banach spaces all of whose subspaces have the Approximation Property*, in Special Topics in Appl. Math., Proc. Bonn 1979, North Holland, 1980, 15–26.

[KOS98] H. Knaust, E. Odell and Th. Schlumprecht: *On asymptotic structure, the Szlenk index and UKK properties in Banach space*, preprint.

[KT95] R. Komorowski and N. Tomczak-Jaegermann: *Banach spaces without local unconditional structure*, Israel J. Math., 89 (1995), 205–226; *Erratum to "Banach spaces without local unconditional structure*, Israel J. Math. (to appear).

[K76] J. L. Krivine: *Sous espaces de dimension finie des espaces de Banach réticulés*, Ann. of Math. (2) 104, 1–29.

[LT71] J. Lindenstrauss and L. Tzafriri: *On complemented subspaces problem*, Israel J. Math. 9, 263–269.

[M88] P. Mankiewicz: *Subspace mixing properties of operators in $\mathbb{R}^n$ with applications to Gluskin spaces*, Studia Math., 88, 51–67.

[M98] P. Mankiewicz: *Compact groups of operators on proportional quotients of ℓ_1^n*, Israel J. Math., to appear.

[MT88] P. Mankiewicz and N. Tomczak-Jaegermann: *Random subspaces and quotients of finite-dimensional Banach spaces*, Odense Univ./1988/9; *The solution of finite-dimensional homogeneous Banach space problem*, Israel J. Math., 75 (1991), 129–159.

[MT94] P. Mankiewicz and N. Tomczak-Jaegermann: *Schauder bases in quotients of subspaces of $\ell_2(X)$*, Amer. J. Math., 116, 1341–1363.

[MT98a] P. Mankiewicz and N. Tomczak-Jaegermann: *Random Banach spaces*, in "Handbook on Banach Spaces", (W.B. Johnson and J. Lindenstrauss, eds.), Elsevier, in preparation.

[MT98b] P. Mankiewicz and N. Tomczak-Jaegermann: in preparation.

[Ma95] B. Maurey: *A remark about distortion*, Oper. Theory: Adv. Appl., 77, 131–142.

[MMT95] B. Maurey, V. D. Milman and N. Tomczak-Jaegermann: *Asymptotic infinite-dimensional theory of Banach spaces*, Oper. Theory: Adv. Appl. 77, 149–175.

[Mi69] V. D. Milman: *The spectrum of bounded continuous functions on the unit sphere of a Banach space*, Funct. Anal. Appl., 3 (1969), 67–79; *Geometric theory of Banach spaces II, geometry of the unit sphere*, Russian Math. Survey 26 (1971), 79–163.

[Mi86a] V. D. Milman: *Inégalité de Brunn-Minkowski inverse et applications à la théorie local des espaces normés*, C.R.A.S. 302, 25–28.

[Mi86b] V. D. Milman: *The concentration phenomenon and linear structure of finite-dimensional normed spaces*, Proc. I.C.M. Berkeley, CA 1986, AMS Providence, RI, 1987, 961–975.

[Mi96] V. D. Milman: *Surprising geometric phenomena in high dimensional convexity theory*, Proc. E.M.C.2, Budapest, 1996, Birkhäuser Verlag, 1998.

[MiSh79] V.D. Milman and M. Sharir: *Shrinking minimal systems and complementation of ℓ_p^n-spaces in reflexive Banach spaces*, Proc. L.M.S., 39, 1–29.

[MiT93] V. D. Milman and N. Tomczak-Jaegermann: *Asymptotic ℓ_p spaces and bounded distortions*, (Bor-Luh Lin and W.B. Johnson, eds.), Contemp. Math., 144, 173–195.

[MiT95] V. D. Milman and N. Tomczak-Jaegermann: unpublished.

[O98] E. Odell: *On subspaces, asymptotic structure and distortion of Banach spaces; connections with logic*, preprint.

[OS93] E. Odell and Th. Schlumprecht: *The distortion problem*, GAFA, 3 (1993), 201–207; *The distortion problem*, Acta Math., 173 (1994), 259–281.

[OS94] E. Odell and Th. Schlumprecht: *Distortion and Stabilized Structure in Banach Spaces; New Geometric Phenomena for Banach and Hilbert Spaces*, Proc. I.C.M. Zürich 1994, Birkhäuser Verlag, Basel 1995, 955–965.

[OT98] E. Odell and N. Tomczak-Jaegermann: *On certain equivalent norms on Tsirelson's space*, preprint.

[OTW97] E. Odell, N. Tomczak-Jaegermann and R. Wagner: *Proximity to ℓ_1 and Distortion in Asymptotic ℓ_1 Spaces*, Jour. of Funct. Anal., 150, 101–145.

[Sch91] Th. Schlumprecht: *An arbitrarily distortable Banach space*, Israel J. Math., 76, 81–95.

[S83] S. J. Szarek: *The finite-dimensional basis problem with an appendix on nets of Grassman manifold*, Acta. Math., 151, 153–179.

[S86] S. J. Szarek: *On the existence and uniqueness of complex structure and spaces with "few" operators*, Trans. A.M.S., 293, 339–353.

[S87] S. J. Szarek: *A Banach space without a basis which has the bounded approximation property*, Acta Math.m 159, 81–98.

[S90] S. J. Szarek: *Spaces with large distance to ℓ_∞^n and random matrices*, Amer. J. of Math., 112, 899–942.

[S91] S. J. Szarek: *Condition numbers of random matrices*, J. of Complex., 7, 131–149.

[T96] N. Tomczak-Jaegermann: *Banach spaces of type $p > 1$ have arbitrarily distortable subspaces*, GAFA 6, 1074–1082.

[Ts74] B. S. Tsirelson: *Not every Banach space contains ℓ_p or c_0*, Funct. Anal. Appl., 8, 138–141.

[W55] E. Wigner: *Characteristic vectors of bordered matrices with infinite dimensions*, Ann. Math., 62 (1955), 548–564; *On the distribution of the roots of certain symmetric matrices*, Ann. Math., 67 (1958), 325–327.

Dept. of Math. Sci., Univ. of Alberta
Edmonton, Alberta, Canada T6G 2G1

Discrete Analogues of Singular and Maximal Radon Transforms

Stephen Wainger*

Abstract. We describe recent results concerning ℓ^p estimates for certain discrete operators and the application of methods of analytic number theory in the treatment of these operators.

1991 Mathematics Subject Classification: 42B20, 42B25, 11L15

We would like to discuss recent joint work with E. M. Stein concerning estimates for certain "discrete" operators of harmonic analysis, the difference between these operators and analogous older "continuous" operators, and the role ideas of analytic number theory play in resolving the extra difficulties arising in studying these discrete operators.

We begin by recalling the continuous operators we have in mind. For each x in R^ℓ, we let $\gamma(x,t)$ be a smooth k-dimensional surface passing through x. That is $\gamma(x,t)$ is a smooth mapping of $R^\ell \times R^k \to R^\ell$ with $\gamma(x,0) = x$. We also let $K(t)$ be a smooth Calderon-Zygmund kernel on R^k. That is $K(t)$ is smooth away from the origin, for $0 < a < b$, $\int_{a\le|t|\le b} K(t)dt = 0$, and for positive λ, $K(\lambda t) = \lambda^{-k}K(t)$. We set

$$Sf(x) = \int f(\gamma(x,t))K(t)dt,$$

and

$$Mf(x) = \sup_R \frac{1}{|B(R)|}\int_{B(R)} f(\gamma(x,t))dt.$$

The following is a rough version of the type of result we have in mind.

Theorem 1: [Christ, Nagel, Stein, Wainger]. See [CNSW].

If $\gamma(x,t)$ satisfies an appropriate curvature condition, S is locally bounded in $L^p(R^\ell)$, $1 < p < \infty$ and Mf is locally bounded in $L^p, 1 < p$.

Here $B(R)$ is the ball in R^k of radius R centered at the origin, and $|B(R)|$ denotes its measure. S and M are called the singular and maximal Radon transforms respectively.

To make the rough statement correct one has to modify the definitions of S and M by introducing appropriate cut off functions. For our purposes it will not be necessary to know the precise formulation of the curvature condition,but for the sake of completeness we give two of several equivalent formulations. One way of expressing the curvature condition is in terms of vector fields. It can be shown

*Supported in part by an NSF grant at the University of Wisconsin-Madison.

that for any smooth $\gamma(x,t)$ there is a unique family of vector fields X_α on R^ℓ so that we have an asymptotic formula

$$\gamma(x,t) \sim \exp[\Sigma t^\alpha X_\alpha](x)$$

$\alpha = (\alpha_1, \cdots, \alpha_k)$ with $\alpha_1, \ldots, \alpha_k$ integers. Here exp is the ordinary exponential map and the meaning of $\sim$ is that if we only include terms with $\alpha_1 + \cdots + \alpha_k \leq N$, the error is $\mathcal{O}((t)^{N+1})$. Then the curvature condition is satisfied if the X_α and their commutators span R^ℓ at every x. If $\gamma(x,t)$ is real analytic, the curvature condition can be expressed in terms of invariant manifolds of the flow $t \to \gamma(x,t)$. If $\gamma(x,t)$ is real analytic the curvature condition is satisfied if for no x there is a small piece of submanifold passing through x of positive codimension invariant under the flow $t \to \gamma(x,t)$. If $\gamma(x,t)$ is smooth the curvature condition may be expressed by saying there is no submanifold of positive codimension invariant to infinite order in an appropriate sense.

We have the following corollary of Theorem 1.

COROLLARY: *If $\gamma(x,t)$ is real analytic and f is in $L^p, 1 < p$*

$$\lim_{\epsilon \to 0} \frac{1}{|B(\epsilon)|} \int_{B(\epsilon)} f(\gamma(x,t))dt = f(x) \quad a.e.$$

Theorem 1 has a history of over 30 years, and others have contributed steps leading to the proof. Among these people are Fabes, Geller, Greanleaf and Uhlman, D. Mueller, Phong, Ricci, and Riviere. See references cited in [CNSW].

The conclusion of Theorem 1 does not hold for an arbitrary smooth $\gamma(x,t)$. See [NW] and [SW1]. The conclusion may however hold in some cases where the curvature condition fails. For example, the conclusions hold if $\gamma(x,t) = x + \Gamma(t)$ and $\Gamma(t)$ is a straight line through the origin. Some of the people who considered the problem of obtaining L^p estimates for S and M when the curvature condition fails are Carbery, H. Carlsson, Christ, Cordoba, Duoandikoetxea, Nagel, Rubio de Francia, Seeger, Vance, Wainger, Weinberg, and Ziesler. See references cited in [CWW] and [WWZ].

The effect of curvature is more dramatic in a related question – that of spherical averages, and we digress to discuss this problem. Denote by

$$Af(x) = \sup_{r>0} \int_\Sigma |f(x - ry'))d\sigma(y')$$

where $\sum$ is the unit sphere in R^ℓ, $\ell \geq 2$, and $d\sigma(y')$ is normalized rotationally invariant measure on Σ.

THEOREM 2: [STEIN $\ell \geq 3$, BOURGAIN $\ell = 2$]

$$\|Af\|_{L^p} \leq C(\ell, p)\|f\|_{L^p}$$

if $p > \frac{\ell}{\ell - 1}$ and $\ell \geq 2$.

See [S], [B1] and also [MSS]. As a corollary one finds that as $r \to 0$

$$\int_\Sigma f(x - ry')d\sigma(y') \to f(x) \qquad \text{a.e.}$$

if f is in $L^p(R^\ell), p > \frac{\ell}{\ell-1}$.

To see the effect of curvature consider

$$Bf(x) = \sup_{r>0} \int_{Q_r} f(x - y')dq_r(y')$$

where Q_r is the boundary of a cube of diameter r and faces parallel to the coordinate hyperplanes, and dq_r is $\ell-1$ dimensional Lebesgue measure on Q_r normalized so that Q_r has measure 1. Let U be the set of all points which are on those hyperplanes which are parallel to a fixed coordinate hyperplane and at a rational distance from it. Take f to be the characteristic function of U. Then $f = 0$ a.e. and $Bf = 1/2^\ell$ a.e. so there can be no analogue of Theorem 2 in this setting.

We now describe the discrete analogues of S and M. Let $P(x,t)$ be a polynomial mapping from $R^\ell \times R^k$ with integer coefficients. Denote by Z^ℓ the lattice points in R^ℓ, that is points with integral coordinates. Let f be a function defined on Z^ℓ. For m in Z^ℓ, set

$$\mathcal{S}f(m) = \sum_{\substack{n \in Z^k \\ n \neq 0}} K(n)f(P(m,n)),$$

and

$$\mathcal{M}f(m) = \sup_R \sum_{\substack{n \in Z^k \\ |n| \leq R}} |f(P(M,n))|.$$

We are then interested in obtaining estimates for $\mathcal{S}$ and $\mathcal{M}$ in $\ell^p(Z^\ell)$. The first results were in the translation invariant case, namely the case that $P(m,n) = m - Q(n)$ where Q is a polynomial mapping from R^k to R^ℓ with integer coefficients. (In the continuous situation results in the translation invariant case were also obtained many years before Theorem 1 was proved in full generality). The known results in this translation invariant case are the following:

Theorem 3: Arkhipov and Oskolkov [1987] for $k = 1$, Stein and Wainger [1990] for general k. See [A0] and [SW2].

$$\|\mathcal{S}f\|_{\ell^2} \leq A\|f\|_{\ell^2}.$$

Theorem 4: Bourgain [1988-1989]. See [B2].

$$\|\mathcal{M}f\|_{\ell^p} \leq A_p\|f\|_{\ell^p}, 1 < p.$$

Theorem 5: Stein and Wainger [1990]. See [SW2].

$$\|\mathcal{S}f\|_{\ell^p} \leq A_p\|f\|_{\ell^r}, \qquad \frac{3}{2} < p < \frac{5}{2}.$$

There is a recent result in which the operator is not translation invariant. For example if $\ell = 2$ and $k = 1$ we might take

$$P(m_1, m_2, n) = (m_1 - n, m_2 - nm_1^2).$$

More generally we assume $m = (m_1, m_2)$ with m_1 in Z^k and m_2 in $Z^{\ell-k}$ and we consider operators commuting with translations in the m_2 directions. That is $P(m, n) = (m_1 - n, m_2 - Q(m_1, n))$ where Q is a polynomial mapping from $R^k \times R^k \to R^{\ell-k}$ with integer coefficients. In this situation we have the following result.

THEOREM 6: STEIN AND WAINGER [1997]. See [SW3].

$$\|\mathcal{S}f\|_{\ell^2(Z^\ell)} \leq A\|f\|_{\ell^2(Z^\ell)}.$$

We would also like to mention two related results. Suppose p_n denotes the nth prime. For m an integer, set

$$\mathcal{M}f(m) = \sup_R \frac{1}{R} \sum_{n=1}^{R} |f(m - p_n)|.$$

Then we have the following result.

THEOREM 7: WIERDL [1988]. See [W].

$$\|\mathcal{M}f\|_{\ell^p(Z)} \leq A_p\|f\|_{\ell^p(Z)}, p > 1.$$

Finally there is a partial analogue of Theorem 2. We let $N(\rho)$ denote the number of lattice points on the sphere of radius ρ in R^ℓ. ($N(\rho) = 0$ unless ρ is the square root of an integer). For m in Z^ℓ, let

$$\mathcal{A}_s f(m) = \sup_{\substack{\rho \\ s \leq \rho \leq 2s}} \frac{1}{N(\rho)} \sum_{|n|=\rho} |f(m - n)|.$$

We then have the following result.

THEOREM 8: MAGYAR [1996]. See [M].
For $\ell \geq 5$ and $p > n/(n-2)$

$$\|\mathcal{A}_s f\|_{\ell^p(Z^\ell)} \leq C_p\|f\|_{\ell^p(Z^\ell)}.$$

Theorem 3) and Theorem 7) have applications to ergodic theory. Let T be a measure preserving invertible transformation on a probability space Ω, and set $\tau f(x) = f(Tx)$. Then Theorem 7 is an important ingredient in Bourguin's ergodic theorem. An important special case of Bourgain's theorem is the following:

THEOREM 9: BOURGUIN 1988-1989. See [B2].

For any integer r and f in $L^p(\Omega), p > 1$,

$$\frac{1}{N}\sum_{n=1}^{N} \tau^{n^r} f(x)$$

converges almost everywhere in Ω.

A similar result holds if the sequence n^r is replaced by the sequence of primes.

We want to deal with the following question:

QUESTION: What is the difference between the continuous and discrete problems?

SHORT ANSWER: The difference between sums and integrals.

One of the most dramatic differences between sums and integrals can be seen by considering two functions $A(s) = \int_1^\infty \frac{dt}{t^s}$ and $B(s) = \sum_{n=1}^\infty \frac{1}{n^s}$. Both $A(s)$ and $B(s)$ are defined for Re $s > 1$ and have meromorphic continuations to the entire complex plane. But there the similarity stops. $A(s) = \frac{1}{s-1}$, and $B(s)$ is not bounded for s away from 1. And in fact the correct growth of $B(s)$ is one of the hardest problems in mathematics.

More to the point, certain changes of variables in integrals have no analogues for sums, and in fact estimates for integrals provide a wrong guess for analogous sums. Let λ be large and set $A(\lambda) = \int_a^b e^{2\pi i \lambda x^2} dx$ and $B(\lambda) = \sum_{a \le n \le b} e^{2\pi i \lambda n^2}$. To study $A(\lambda)$ we make a change of variables $u = \sqrt{\lambda} x$, and see $A(\lambda) = \frac{1}{\sqrt{\lambda}} \int_{a'}^{b'} e^{2\pi i u^2} du$. In effect we have normalized the situation to the case that the coefficient of u^2 is 1. Normalization procedures amounting to changes of variables in integrals, though of a more complicated nature, play an important part in the proof of Theorem 1, and these changes of variables are not available in the discrete problems. It of course follows that $|A(\lambda)| \le \frac{C}{\sqrt{\lambda}}$. On the other hand if we take λ to be an integer, $B(\lambda) \sim b - a$.

We now wish to compare continuous and discrete operators. Perhaps the easiest operators to consider are those of fractional integration of imaginary order. For j a positive integer let

$$C_j f(x) = \int_1^\infty f(x - y^j) \frac{dy}{y^{1+i\gamma}} \quad \text{and} D_j f(m) = \sum_{n=1}^\infty f(m - n^j) \frac{1}{n^{1+i\gamma}},$$

with γ real. The proof of the $L^p(R)$ boundedness of C_1 and the $\ell^p(Z)$ boundedness of D_1 are similar. The change of variables $u = y^j$ reduces the study of C_j to C_1. No such change of variables is possible for D_j and in fact D_j for $j \ge 2$ is much different from C_j or D_1 (which as we have said are similar).

We would like to consider the difference in proving the ℓ^2 boundedness D_1 and D_2. Let $\mu_j(\theta) = \sum_{n=1}^\infty \frac{1}{n^{1+i\gamma}} \exp(2\pi i n^j \theta)$ for $j = 1$ and $j = 2$. To show D_j is bounded in ℓ^2 it is sufficient (and necessary) to show $\mu_j(\theta)$ is a bounded function. Let

$$S_N^j(\theta) = \sum_{1 \le n \le N} \exp 2\pi i n^j \theta,$$

QUESTION: For what θ is

$$|S_N^j(\theta)| \le AN^{1-\delta} \text{ ?} \tag{1}$$

If 1) holds for an interval of θ, we may sum by parts in the expression for $\mu_j(\theta)$ and conclude that $\mu_j(\theta)$ is bounded in that range of θ. Let us compare $S_n^1(\theta)$ and $S_N^2(\theta)$ at a rational point $\theta = \frac{p}{q}$ with $(p,q) = 1$. (For simplicity take N to be a multiple of q). We are then considering $S_N^j = \sum_{n=1}^N \exp 2\pi i n^j \frac{p}{q}$. We want to write $n = mq + \ell$ where m runs from 1 to $\frac{N}{q}$ and ℓ goes from 1 to q. Then

$$S_N^j\left(\frac{p}{q}\right) = \sum_{m=1}^{\frac{N}{q}} \sum_{\ell=1}^{q} \exp 2\pi i (mq+\ell)^j \frac{p}{q}.$$

$$\exp 2\pi i (mq+\ell)^j \frac{p}{q} = \exp 2\pi i (u + \ell^j)\frac{p}{q},$$

where u is an integer divisible by q. So

$$S_N^j\left(\frac{p}{q}\right) = \sum_{m=1}^{\frac{N}{q}} \sum_{\ell=1}^{q} \exp 2\pi i \ell^j \frac{p}{q},$$

and the sum on ℓ is independent of m. Thus

$$\begin{aligned} S_N^j\left(\tfrac{p}{q}\right) &= \tfrac{N}{q} \cdot \textstyle\sum_{\ell=1}^{q} \exp 2\pi i \ell^j \frac{p}{q} \\ &= \tfrac{N}{q} G_j(p,q) \end{aligned}$$

where $G_j(p,q) = \sum_{\ell=1}^{q} \exp 2\pi i \ell^j \frac{p}{q}$. If $j = 1$ $G_j(p,q) = 0$. If $j = 2$, $G_j(p,q)$ is not necessarily 0. In fact

$$|G_2(p,q)| = \begin{cases} \sqrt{q} & \text{if } q \text{ is odd} \\ \sqrt{2q} & \text{if } q \equiv 0 \pmod 4 \\ \sqrt{q} & \text{if } q \equiv 2 \pmod 4. \end{cases}$$

So

$$S_N^2\left(\frac{p}{q}\right) = \frac{N}{q} G_2(p,q) \neq \mathcal{O}(N^{1-\delta}) \tag{2}$$

in general. The upshot is that to prove the boundedness of $\mu_1(\theta)$ we need to use the cancellation in $\Sigma \frac{1}{n^{1+i\gamma}}$ only when θ is near an integer, while to prove the boundedness of $\mu_2(\theta)$ we need to use the cancellation of $\Sigma \frac{1}{n^{1+i\gamma}}$ "near" each rational $\frac{p}{q}$.

The motivation for the proof of the boundedness of $\mu_j(\theta)$ for $j \ge 2$ comes from ideas of Hardy, Littlewood, Ramanujan and Vinogradov in analytic number theory. A typical problem concerns the number of solutions in positive integers of

the equation $k = n_1^r + \cdots + n_\ell^r$ for fixed integers r and ℓ. Let us denote by $H(k)$ the number of solutions. Let $S_N(\theta) = \sum_{n=1}^{N} \exp 2\pi i n^r \theta$. Then

$$H(k) = \int_0^1 e^{-2\pi i k\theta} [S_N(\theta)]^\ell d\theta, \tag{3}$$

for

$$[S_N(\theta)]^\ell = \sum_{n_1, n_2, \ldots, n_\ell} \exp 2\pi i (n_1^r + \cdots + n_\ell^r)\theta.$$

We then get a contribution to the integral in 3) exactly when $k = n_1^r + \cdots + n_\ell^r$ for some choice of integers $n_1, n_2, \ldots, n_\ell$. The idea of Hardy and Littlewood is that the main contribution to the integral in 3) comes from small intervals around rationals with denominators that are small compared to N and that if θ is in such an interval a convenient approximation to $S_N(\theta)$ can be found. See for example [HL]. Notice that 2) suggests that for θ "near" a rational with large denominator q, that is $q > N^\epsilon$. there is a non-trivial estimate for $S_N^j(\theta)$. However in the derivation of 2) we also assumed that $q << N$. One can observe that if $q >> N^j$, all the exponentials, $\exp 2\pi i n^j \frac{p}{q}$, would point in the same direction so that no cancellation could occur in the sum for $S_N^j(\theta)$. So to obtain cancellation in the sum for $S_N^j(\theta)$, we require θ to be near $\frac{a}{q}$ with $N^\epsilon \leq q \leq N^{j-\theta}$. In fact one can prove the following lemma which will be important in the sequel.

LEMMA 10: *For every $\epsilon > 0$, there are constants A and δ (depending on j) such that if*

$$|\theta - \frac{p}{q}| \leq \frac{1}{q^2},\ (p,q) = 1, \quad \textit{and} \quad N^\epsilon \leq q \leq N^{j-\epsilon},$$

then

$$|S_N^j(\theta)| \leq A N^{1-\delta}.$$

See [V], where the estimate is stated in a more precise form. In our discussion of $\mu_2(\theta)$ below we shall see why good convenient approximations can be made to sums like $S_N(\theta)$ if θ is near a rational with small denominator.

The idea of using number theoretic methods to study these discrete problems was due independently to Arkhipov and Oskolkov [AO] and Bourgain [B2]. Let us show how to prove $\mu_2(\theta)$ is bounded. To simplify the notation, we shall take $\gamma = \frac{2\pi}{\ln 2}$ so that $\int_1^2 \frac{dt}{t^{1+i\gamma}} = 0$. We want to write

$$\mu_2(\theta) = \sum_{\substack{p,q \\ (p,q)=1}} M^{p,q}(\theta) + \quad \text{Error} \tag{4}$$

where $M^{p,q}$ is the contribution from rationals $\frac{p}{q}$ near θ, where in some sense q should have small denominator. To this end we fix θ and write

$$\mu_2(\theta) = \sum_{j \geq 1} H_j(\theta)$$

where

$$H_j(\theta) = \sum_{2^j \le n < 2^{j+1}} \frac{1}{n^{1+i\gamma}} \exp 2\pi i n^2 \theta.$$

We then set

$$M^{p,q}(\theta) = \sum_{j \atop |\theta - \frac{p}{q}| < 2^{-j(2-\epsilon)}, q < 2^{\epsilon j}} H_j(\theta).$$

Lemma 10 (together with Dirichilet's principle) implies that

$$\mu_2(\theta) = \sum_{p,q} M^{(p,q)}(\theta) + \quad \text{bounded error.}$$

Next we want to show that

$$M^{(p,q)}(\theta) = (\text{small in } q) \cdot \text{Integral} + \text{Error},$$

We will then be able to make appropriate changes of variables in the integral. In fact we will see that

$$5) \qquad M^{(p,q)}(\theta) = \frac{1}{q} G(p,q) I(2^{2j}(\theta - \frac{a}{q})) \quad + \quad \text{bounded error},$$

where $I(\phi) = \int_1^2 e^{2\pi i t^2 \phi} \frac{dt}{t^{1+i\gamma}}$ and $G(p,q) = \sum_{j=1}^{q} e^{2\pi i j^2 \frac{p}{q}}$.

Let us assume 5) is true for the moment. Then

$$6) \qquad |G(p,q)| \le A q^{1-\delta}$$

by Lemma 10. Also since $\int_1^2 \frac{dt}{t^{1+i\gamma}} = 0$

$$7) \qquad |I(\phi)| \le A|\phi|.$$

Now a change of variables shows that for $1 \le s \le 2$

$$\left| \int_1^s e^{2\pi i t^2 \phi} dt \right| \le \frac{A}{\sqrt{|\phi|}}.$$

So integrating by parts we see

$$8) \qquad |I(\phi)| \le \frac{A}{|\phi|^{1/2}}.$$

Finally it is possible to show that for a fixed θ

9) the number of (p,q) that occur with $2^s \le q < 2^{s+1}$ is uniformly bounded.

If the estimates 6,7,8, and 9 are substituted into 5), it is easy to see that $\mu_2(\theta)$ is bounded. So we are faced with trying to write $M^{(p,q)}$ as a product $\frac{1}{q} G(p,q) \cdot$ Integral.

$$M^{p,q}(\theta) = \sum_j{}'' H_j(\theta).$$

Recall that if we write $\theta = \frac{9}{q} + \beta$, then for j to occur in Σ'', $|\beta| < 2^{-(2-\epsilon)j}$ and $q < 2^{\epsilon j}$. Again

$$H_j = \sum_{2^j < n < 2^{j+1}} \frac{1}{n^{1+i\gamma}} exp(2\pi i n^2 \theta).$$

We write $n = mq + \ell$ with $0 \leq \ell \leq q-1$. Then since q is small

$$10) \qquad \frac{1}{n^{1+i\gamma}} = \frac{1}{(mq+\ell)^{1+i\gamma}} \sim \frac{1}{(mq)^{1+i\gamma}}$$

since $\ell < q$ is small. A more subtle point is that

$$11) \qquad \exp 2\pi i n^2 \theta = \exp 2\pi i m^2 q^2 \beta \cdot \exp 2\pi i \ell^2 \frac{p}{q} + \mathcal{O}(2^{-j/2})$$

(if ϵ is sufficiently small). To see 11) note that $n^2\theta = (mq+\ell)^2(\frac{a}{q}+\beta) = m^2q^2\beta + \ell^2\frac{a}{q} + 2m\ell q\beta +$ integer, and now since $m < 2^j$, $q < 2^{\epsilon j}$ and $|\beta| < 2^{-(2-\epsilon)j}$ the term $2m\ell q\beta$ may be dropped by making an error $\mathcal{O}(2^{-j/2})$, which gives 11. (When we come to the non-translation invariant problems we will arrive at an analogous point, however we will not have control on the size of m i.e. $m < 2^j$ which will cause a major difficulty.)

Thus 11) is established, and in $\sum''$ we may replace H_j by

$$12) \qquad \overline{H}_j = \frac{1}{q^{1+i\gamma}} G(p,q) \sum_{\frac{2^j}{q} \leq m < \frac{2^{j+1}}{q}} \frac{1}{m^{1+i\gamma}} \exp 2\pi i m^2 q^2 \beta,$$

Using once again the facts that $m < 2^j, q < 2^{\epsilon j}$ and $|\beta| < 2^{-(2-\epsilon)j}$, we see that we may replace the sum in (2) by an integral making an error which is $\mathcal{O}(2^{-j/2})$. Then a change of variables in the integral gives us 5). The proof that $\mu_2(\theta)$ is bounded is now complete.

Estimates for the maximal function as well as ℓ^p estimates are much more difficult because it does not suffice to deal with one fixed θ.

Let us try to see what is involved in proving the ℓ^2 boundedness in a non-translation invariant case. For (m,ℓ) in Z^2, and f defined on Z^2, we set

$$Sf(m,\ell) = \sum_{\substack{n \\ m-n \neq 0}} f(n, \ell - m^2 n) \cdot \frac{1}{m-n}.$$

We wish to show

$$13) \qquad \|Sf\|_{\ell^2(Z^2)} \leq A \|f\|_{\ell^2(Z^2)}.$$

For f defined on Z^1, we set

$$14) \qquad S_\theta f(m) = \sum_{\substack{n \\ m-n \neq 0}} \exp(2\pi i m^2 n \theta) \frac{f(n)}{m-n}.$$

By using a well known technique of taking the Fourier transform in the ℓ variable and using Plancheral's theorem, we see that to prove 13 it suffices to show

$$\|S_\theta f\|_{\ell^2(Z)} \leq A\|f\|_{\ell^2(Z)}, \tag{15}$$

uniformly in θ. We shall try to follow the lines of the proof of the boundedness of $\mu_2(\theta)$. The main idea is to replace the formula 5) by writing S_θ as a tensor product of an operator variant of the expression $\frac{1}{q}G(a,q)$ and an integral operator. We proceed with an operator valued version of the treatment of $\mu_2(\theta)$ above. We define operator valued analoges of the H_j, namely

$$H_j(\theta)f(m) = \sum_{\substack{n \\ 2^j \leq |m-n| < 2^{j+1}}} \exp(2\pi i m^2 n\theta) \quad \frac{f(m)}{m-n} \tag{16}$$

and set

$$M^{(p,q)}(\theta)f(m) = \sum_{\substack{j \\ |\theta - \frac{a}{q}| \leq 2^{-(3-\epsilon)j}, q < 2^{\epsilon j}}} H_j(\theta)f(m).$$

Then it is possible to prove an operator valued version of Lemma 10 so that

$$S_\theta = \sum_{p,q} M^{(p,q)}(\theta) \ + \ \text{bounded operator}.$$

We now want to write in analogy with 5)

$$M^{p,q}(\theta) \sim \frac{1}{q}G(p,q) \otimes I_\theta$$

where $G(p,q)$ is an operator valued analogue of G and I_θ is an integral operator. In analogy with the argument proving 5) in the expression 16) for H_j, we set

$$m = m_1 q + \mu \quad \text{and} \quad n = n_1 q + \nu.$$

Following the lines of the argument in the translation invariant case we would like to write $\theta = \frac{a}{q} + \beta$, and would like to say that

$$\begin{aligned} \exp 2\pi i \ & (m_1 q + \mu)^2 (n_1 q + \nu)(\tfrac{a}{q} + \beta) \\ & = \exp 2\pi i m_1^2 n_1 q^3 \cdot \exp 2\pi i \mu^2 \nu \tfrac{a}{q} + \text{ small error}. \end{aligned} \tag{17}$$

Unfortunately we can not do this because while we have control on the size of $m_1 - n_1$, we have no control of the size of m_1 or n_1. And even if we could prove 17) we could not replace a sum on n_1 by an integral because we have no estimate on the size of m_1 and n_1. The main idea in getting around this difficulty is to note that in dealing with $M^{p,q}(\theta)$, $2^j < \left(\frac{1}{\beta}\right)^{\frac{1}{3-\epsilon}}$. Thus to obtain an estimate for $M^{p,q}(\theta)$, it suffices to obtain estimates of translates of the operators $M^{p,q}$ where however m and n can be assumed to be at most $\left(\frac{1}{\beta}\right)^{\frac{1}{3-\epsilon}}$. We refer to [SW3] where the complicated details are carried out.

References

[AO] Arkhipov, G. I. and Oskolkov, K. I. (1987). On a special trigonometric series and its applications. *Mat. Sb.*, 134, 147–158.

[B1] Bourgain, J. (1986). Averages in the plane over convex curves and maximal operators. *J. Analyse Math.*, 47, 69–85.

[B2] Bourgain, J. (1989). Pointwise ergodic theorems for arithmetic sets. *Inst. Hautes Etudes Sci. Publ. Math.*, 69, 5–45.

[CWW] Carbery, A., Wainger, S., and Wright, J. (1995). Hilbert transforms and maximal functions associated to flat curves on the Heisenberg group. *J. of Amer. Math. Soc.*, 8, 141–179.

[CNSW] Christ, M., Nagel, A., Stein, E. M., Wainger, S., Singular and maximal radon transforms: analysis and geometry. to appear.

[HL] Hardy, G. H. and Littlewood, J. E. (1920). A new solution of Waring's Problem. *Quart. J. of Math.*, 48, 272–293.

[M] Magyar, A. (1997). L^p-bounds for spherical maximal operators on Z^n. *Revista Matematica Iberoamericana*, 13, 1-11.

[MSS] Mockenhaupt, G. Seeger, A., and Sogge, C. D. (1992). Wave front sets, local smoothing, and Bourgain's circular maximal theorem. *Ann. of Math.*, 136, 207-218.

[NW] Nagel, A. and Wainger, S. (1976). Hilbert transforms associated with plane curves. *Trans. Amer. Math. Soc.*, 223, 235–252.

[S] Stein, E. M. (1976). Maximal functions: Spherical means. *Proc. Nat. Acad. Sci. U.S.A.*, 73, 2174–2175.

[SW1] Stein, E. M. and Wainger, S. (1978). Problems in harmonic analysis related to curvature. *Bull. Amer. Math. Soc.*, 84, 1239–1295.

[SW2] Stein, E. M. and Wainger, S. (1990). Discrete analogues of singular Radon transforms. *Bull. Amer. Math. Soc.*, 23, 537–544.

[SW3] Stein, E. M. and Wainger, S. Discrete analogues of singular Radon transforms: a non translation invariant case. to appear.

[V] Vinogradov, I. (1954). *The method of trigonometrical sums in the theory of numbers.* Interscience, New York.

[WWZ] Wainger, S., Wright, J. and Ziesler, S. Singular integrals associated to hypersurfaces: L^2 theory. MSRI preprint #1997–106.

[W] Wierdl, M. (1988). Pointwise ergodic theorem along the prime numbers. *Israel J. Math.*, 64, 315–336.

Stephen Wainger
Department of Mathematics
University of Wisconsin
Madison, WI 53706
wainger@math.wisc.edu

Maximal Averages and Packing of One Dimensional Sets

Thomas Wolff

Abstract. We discuss recent work of several authors on the Kakeya needle problem and other related problems involving nonexistence of small sets containing large families of one dimensional objects.

1991 Mathematics Subject Classification: 42B99
Keywords and Phrases: Kakeya set, maximal function

My purpose here is to summarize some recent work in real analysis related to Kakeya type maximal functions. I will take a fairly narrow point of view; specifically, will only consider the classical situations of lines and circles, and will not discuss the recent work on related problems involving oscillatory integrals, to be found for example in [2] and [18]. For a more detailed survey see [22].

The basic open problem in this area, known as the Kakeya problem, has several (morally but not formally equivalent) formulations. We state them below in increasing order of "strength." One defines a Kakeya set to be a compact set $E \subset \mathbb{R}^n$ which contains a unit line segment in each direction,

$$\forall e \in \mathbb{P}^{n-1} \exists x \in \mathbb{R}^n : x + te \in E \ \forall t \in [-\frac{1}{2}, \frac{1}{2}]$$

where we regard $\mathbb{P}^{n-1}$ as being the unit sphere with antipodal points identified. If δ is a small positive number and $f : \mathbb{R}^n \to \mathbb{R}$ then one defines the Kakeya maximal function of f, $f^*_\delta : \mathbb{P}^{n-1} \to \mathbb{R}$ via

$$f^*_\delta(e) = \sup_a \frac{1}{|T^\delta_e(a)|} \int_{T^\delta_e(a)} |f(x)| dx$$

where $T^\delta_e(a)$ is the cylinder centered at a with length 1, cross section radius δ and axis in the e direction. Also define the δ-entropy $\mathcal{N}_\delta(E)$ to be the maximum possible cardinality for a δ-separated subset. Then the following are all open questions if $n \geq 3$.

1. Is it true that if E is a Kakeya set in $\mathbb{R}^n$ then $\limsup_{\delta \to 0} \frac{\log \mathcal{N}_\delta(E)}{\log \frac{1}{\delta}} = n$?
2. Is it true that a Kakeya set in $\mathbb{R}^n$ must have Hausdorff dimension n?

3. Is the following estimate true?

$$\forall \epsilon > 0 \exists C_\epsilon : \|f_\delta^*\|_{L^n(\mathbb{P}^{n-1})} \leq C_\epsilon \delta^{-\epsilon} \|f\|_{L^n(\mathbb{P}^{n-1})} \tag{1}$$

We discuss below the partial results that have been proved on this problem and some work on a class of related problems involving circles in the plane.

BACKGROUND

Let me mention some results that were proved by 1990.

1. Two dimensional Kakeya maximal theorem, cf. Davies [7], Cordoba [6], Bourgain [2]. In $n = 2$ dimensions the above three statements are known to be true. The first two were proved in [7] while the last was proved in [6] in a slightly different formulation and at the beginning of [2] as stated; the latter paper also introduced the particular definition of Kakeya maximal function adapted above.

Statement 3. in $\mathbb{R}^2$ can be proved by an elementary geometric-combinatorial argument exploiting the fact that two lines intersect in at most one point, and the size of the intersection of the corresponding tubes is determined by the angle of intersection: if e_1 and e_2 determine an angle θ, then $T^\delta_{e_1}(a_1) \cap T^\delta_{e_2}(a_2)$ is contained in a tube of length $\frac{\delta}{\delta+\theta}$. This was the approach in [6]. Alternately, it can be proved using the Plancherel theorem (e.g. [2]).

2. L^p estimates for the X-ray transform. In higher dimensions the strongest result connected with 1,2,3 which was proved before 1990 was the "space-time" estimate of Drury [8] and Christ [4], which was motivated by a similar result of Oberlin-Stein for the Radon transform. We explain this briefly. "Space-time" is ad hoc terminology but it is convenient and is intended to convey the analogy with estimates for the wave equation in space-time. (Indeed, it is possible to view the X-ray transform as a Fourier integral operator, although we do not take this point of view here)

There is a hierarchy of possible partial results on (1), namely the conjectural bounds ($1 \leq p \leq n$)

$$\forall \epsilon \exists C_\epsilon : \|f_\delta^*\|_{L^q(\mathbb{P}^{n-1})} \leq C_\epsilon \delta^{-(\frac{n}{p}-1+\epsilon)} \|f\|_p, \; q = (n-1)\frac{p}{p-1} \tag{2}$$

which would follow from (1) by interpolating with the trivial $\|f_\delta^*\|_\infty \lesssim \delta^{-(n-1)}\|f\|_1$. Notice that the partial result becomes stronger as p increases. Now let G be the space of lines in $\mathbb{R}^n$. Then G can be identified with the tangent bundle to $\mathbb{P}^{n-1}$ by mapping a line ℓ to its direction e and its closest point to the origin x, which is orthogonal to e, and one gives G the resulting volume form etc. The X-ray transform of a function f is the function $Xf : G \to \mathbb{R}$ defined by $Xf(\ell) = \int_\ell f$. There is a natural splitting of directions, so it is natural to consider estimates for the operator X from L^p (or L^p Sobolev spaces $W^{p,\alpha}$) to mixed norm spaces $L^q_e(L^r_x)$, where $e \in \mathbb{P}^{n-1}$ and $x \perp e$. The Kakeya conjecture in form 3. is equivalent to the assertion that X maps $W^{n,\epsilon}_{loc}$ to $L^n_e(L^\infty_x)$ for each $\epsilon > 0$. On the

other hand, it is shown in [4] that the pure L^p estimate $\|Xf\|_{L^{n+1}(G)} \leq C\|f\|_{\frac{n+1}{2}}$ is valid. From this, one can easily conclude (2) with $p = \frac{n+1}{2}$.

In [2], Bourgain gave a different, combinatorial approach not going through the space-time estimate, and used it to obtain (2) for $p = \frac{n+1}{2} + \epsilon_n$ (actually, he assumed $q = p$ in (2)) where $\epsilon_3 = \frac{1}{3}$ and ϵ_n is given by an inductive formula. This bound has since been improved in [19] and [3] as we will explain below.

3. Spherical maximal theorem of Stein-Bourgain. Let σ be surface measure on the unit sphere $S^{n-1} \subset \mathbb{R}^n$ and let

$$\mathcal{M}f(x) = \sup_r \int |f(x + r\omega)| d\sigma(\omega)$$

Then

$$\|\mathcal{M}f\|_{L^p(\mathbb{R}^n)} \lesssim \|f\|_{L^p(\mathbb{R}^n)},\ p > \frac{n}{n-1} \tag{3}$$

Stein [17] proved this in three or more dimensions, Bourgain [1] in two dimensions, and independently of Bourgain, Marstrand [10] proved the following geometric consequence or special case analogous to formulations 1. and 2. of the Kakeya problem: a set in $\mathbb{R}^2$ containing a circle with each center has positive measure.

The techniques involved in proving (3) are two fold:

Fourier analysis: the Plancherel theorem and stationary phase asymptotics for $\hat{\sigma}$, e.g. the fact that $|\hat{\sigma}(x)| \lesssim |x|^{-\frac{n-1}{2}}$.

Geometry: we restrict the discussion here to the two dimensional case. The issue, which is not as trivial as it may sound, is to understand how thin annuli intersect. In contrast to the situation for the two dimensional Kakeya problem, the shape of the intersection of two annuli is not determined by the arguments of the maximal function, i.e. centers of the circles, but depends also on how the circles are drawn, and the area will be largest when they are tangent. Let $C(x, r)$ be the circle with center x and radius r and $C_\delta(x, r)$ its δ-neighborhood. We will always assume for simplicity that $\frac{1}{2} \leq r \leq 2$ and $|x| < \frac{1}{4}$. This assumption precludes "external" tangencies so two circles $C(x_1, r_1)$ and $C(x_2, r_2)$ are tangent precisely when the quantity $\Delta((x_1, r_1), (x_2, r_2)) = ||x_1 - x_2| - |r_1 - r_2||$ is equal to zero. we will say they are δ-tangent at a if the parameter Δ is $\leq \delta$ and $a \in C_\delta(x_1, r_1) \cap C_\delta(x_2, r_2)$. If we assume that $|x_1 - x_2| + |r_1 - r_2|$ is bounded from below then the intersection will have area $\approx \frac{\delta^2}{\sqrt{\Delta + \delta}}$. Compensating for this is a significant fact discovered by Marstrand (the "three circle lemma"; [10], Lemma 5.2) which is a quantitative version of the circles of Appolonius. We state only a special case. Fix three circles $C_i = C(x_i, r_i)$. Consider a set of (x, r) with $|x - x_i| + |r - r_i|$ bounded from below and such that $C(x, r)$ is δ-tangent to each C_i at three points whose mutual distances are bounded from below. This set of (x, r) is then contained in the union of two balls of radius approximately δ. This is proved in the following way: (i) a version of the circles of Appolonius covers the limiting case $\delta = 0$ - there are at most two circles tangent (in the above sense) to three given circles at distinct

points, and (ii) the $\delta-$tangent circles must be close to one of these circles, as may be seen by applying the inverse function theorem in an appropriate manner.

Roughly, although various different arguments are possible, the Fourier analysis arguments work best in higher dimensions, while in two dimensions either a purely geometric approach or a combination of the two is used. The first was Marstrand's approach based on the three circle lemma, and was recently extended to a proof of (3) by Schlag [14]. The second was Bourgain's approach.

Recent work related to Bourgain-Marstrand

It turns out that quite a bit of more detailed information can be obtained by combining geometric facts like the 3-circle lemma with some combinatorial techniques. This work was largely motivated by the following question which arises naturally in connection with the three dimensional Kakeya problem. Indeed the special case of the inequality (1) for functions in $\mathbb{R}^3$ invariant by rotations around the x_3 axis is a two dimensional problem which turns out to be a variant on (4) below.

Suppose a set in $\mathbb{R}^2$ contains a circle of every radius. Then must it have Hausdoff dimension two?

It is known that such a set can have measure zero so one expects to be working with an "almost maximal inequality," i.e. a bound for averages over δ-neighborhoods of circles with less than power dependence on δ, analogous to (1). The relevant maximal function is the following one: $\mathrm{M}_\delta f(r) = \sup_x \frac{1}{|C_\delta(x,r)|}\int_{C_\delta(x,r)}|f|$ which we regard as having domain $[\frac{1}{2},2]$. The following result was proved in [20]:

$$\forall\epsilon\exists C_\epsilon : \|\mathrm{M}_\delta f\|_3 \le C_\epsilon\delta^{-\epsilon}\|f\|_3 \tag{4}$$

It follows from this that the answer to the above question is affirmative.

Prior to [20] several other related results were proved. The basic technique used is from a paper of Kolasa and the author [9] which was written in 1994, and can be described in the following way. The difficulty is to control intersections between δ-annuli, and the main difficulty in doing this occurs when the annuli are δ-tangent. Accordingly one needs to control the number of δ-tangencies among annuli. Marstrand's lemma makes it possible to view this as a continuum analogue of the following discrete problem: given N circles, bound the number of pairs of tangent circles, assuming a nondegeneracy condition such as that no three circles are tangent at a point. The circles of Appolonius and the "Zarankiewicz problem" in elementary graph theory give a bound $\mathcal{O}(N^{5/3})$ which was used in [9] to prove the partial result on (4) obtained by interpolating with an L^1 to L^∞ estimate as in (2) and then setting $p = \frac{8}{3}$. Later on the author found the paper [5] whose techniques imply a bound $\mathcal{O}(N^{\frac{3}{2}+\epsilon})$ in the discrete problem, and with some effort [20] one can obtain from this a proof of (4). In the intervening time, Schlag [13] was able to prove a sharp L^p to L^q almost maximal estimate in the setting of Bourgain's theorem using a combination of this technique and the Plancherel

theorem. His result is

$$\forall \epsilon \exists C_\epsilon : \|M_\delta f\|_5 \leq C_\epsilon \delta^{-\epsilon} \|f\|_{\frac{5}{2}}$$

where $M_\delta f(x) = \sup_{\frac{1}{2} \leq r \leq 2} \frac{1}{|C_\delta(x,r)|} \int_{C_\delta(x,r)} |f|$. This was then extended using different techniques to a space-time estimate by Schlag-Sogge [16]. With hindsight these results are also corollaries of (4), see [20], p. 987. A further result (see [20]) is that a set in $\mathbb{R}^2$ containing circles whose centers contain a set of dimension $\alpha \leq 1$ will have dimension at least $\alpha + 1$. This has recently been improved (in a certain sense) by T. Mitsis [11]: a set containing circles whose centers have dimension $> \frac{3}{2}$ will have positive measure.

Approaches to Kakeya

In the rest of the article we will explain what is known about the Kakeya problem. At present the following results are known:

1. Estimate (2) holds when $p = \frac{n+2}{2}$, in particular Kakeya sets have dimension at least $\frac{n+2}{2}$. This result is from [19].

2. In the three dimensional case, an improvement of the latter result to a mixed norm space-time type estimate [21]. This can be described as follows: interpolate in an appropriate manner between the Drury-Christ estimate $\|Xf\|_{n+1} \lesssim \|f\|_{\frac{n+1}{2}}$ and the conjecture (1). This results in a collection of conjectural bounds for the X ray transform from $W^{p,\epsilon}(\mathbb{R}^n)$ for any $\epsilon > 0$ to the mixed $L^q_e(L^r_x)$ spaces on the space of lines G. For given p, the estimate on L^p improves over the corresponding estimate (2), in the same sense as the result of [8] improves over the $p = \frac{n+1}{2}$ case of (2). It turns out [21] that one can prove the mixed norm estimate when $n = 3$, $p = \frac{5}{2}$ (hence $q = \frac{10}{3}$, $r = 10$).

3. The Hausdorff dimension of a Kakeya set in $\mathbb{R}^n$ is at least $\alpha(n-1)+1$ for suitable explicit $\alpha > \frac{1}{2}$. This result and a related result for the Kakeya maximal function are from very recent work of Bourgain [3]. It is clearly a substantial improvement in high dimensions although, as of this writing, the argument does not give anything new in three dimensions.

We briefly describe the idea of [19] (which can be considered a variant on an idea in [2]), as it applies to the entropy formulation 1. of the Kakeya problem. Namely, if E is a Kakeya set then $\mathcal{N}_\delta(E) \geq C_\epsilon \delta^{-\frac{n+2}{2}+\epsilon}$ for any $\epsilon > 0$. To prove this consider a maximal δ-separated subset $\{e_j\}$ of $\mathbb{P}^{n-1}$. For each j there is a segment in the e_j direction contained in E and we let T_j be the cylinder obtained by "thickening" it by δ. For an appropriately chosen N, if half the points in each T_j belong to $< N$ other T_i's, then one immediately gets a lower bound on the volume of the union (hence on $\mathcal{N}_\delta(E)$) since $\sum_j |T_j| \approx 1$. On the other hand, if half the points of some T_j belong to $\geq N$ other T_i's, then one obtains a large family of tubes intersecting a line segment. Each of these belongs to a δ-neighborhood of an essentially unique 2-plane through the line segment and then one can obtain a lower bound for the volume of the union by applying the two dimensional results. The proof in [21] is also based on a (quite complicated) elaboration of this idea.

The above argument is rather unsophisticated. It is tempting to think that one should be able to incorporate techniques related to [5], but this appears difficult to do. We refer though to [15] which contains an analogue of the three circle lemma and to [22] for some further discussion and references.

Bourgain [3] uses a different type of combinatorics. We finish by stating one of his lemmas and explaining how it implies an improved partial result in formulation 1. of the Kakeya problem; corresponding improvements in the other formulations are also in [3] but require some further ideas. It is not really used that the set contains an entire line segment in each direction, just that it contains three well separated points in arithmetic progression on such a line segment. The lemma in question is

Lemma Let A and B be subsets of $\mathbb{Z}^n$ for some n, Γ a subset of $A \times B$ and define $S = \{a + b : (a, b) \in \Gamma\}$, $D = \{a - b : (a, b) \in \Gamma\}$. Assume that A, B and S have cardinality less than N. Then D has cardinality less than $CN^{2-\epsilon}$. Here $\epsilon > 0$ is an explicit numerical constant, and in particular is independent of n.

The value of ϵ is given in [3]. We note that the question of the relative size of sumsets and difference sets is a deep question in combinatorial number theory and refer the reader to Ruzsa's work, for example the survey article [12].

Given the lemma, one can see that a Kakeya set E satisfies $\mathcal{N}_\delta(E) \geq \delta^{-\alpha(n-1)}$ with $\alpha > \frac{1}{2}$ in the following way [3]. Let G be the lattice $\delta\mathbb{Z}^n \subset \mathbb{R}^n$, and for each of the segments $\{x + te : |t| \leq \frac{1}{2}\}$ in the definition of Kakeya set, let x^+ and x^- be the elements of G closest to $x + \frac{1}{2}e$ and $x - \frac{1}{2}e$ respectively. Let A be the set whose elements are the various x^+ and x^- and define $\Gamma \subset A \times A$ to be the set of pairs (x^+, x^-); then let S be the set of sums $x^+ + x^-$. Evidently, $|A| \lesssim \mathcal{N}_\delta(E)$, and in addition, $|S| \lesssim \mathcal{N}_\delta(E)$, since the midpoint $\frac{1}{2}(x^+ + x^-)$ is within $C\delta$ of $x \in E$. But it is equally clear that each point of $\mathbb{P}^{n-1}$ is within $C\delta$ of some difference $x^+ - x^-$. Thus $\delta^{-(n-1)} \lesssim \mathcal{N}_\delta(E)^{2-\epsilon}$, as claimed.

References

[1] J. Bourgain, *Averages in the plane over convex curves and maximal operators*, J. Analyse Math. 47(1986), 69-85.

[2] J. Bourgain, *Besicovitch type maximal operators and applications to Fourier analysis*, Geometric and Functional Analysis 1 (1991), 147-187.

[3] J. Bourgain, in preparation.

[4] M. Christ, *Estimates for the k-plane transform*, Indiana Univ. Math. J. 33(1984), 891-910.

[5] K. L. Clarkson, H. Edelsbrunner, L. J. Guibas, M. Sharir, E. Welzl, *Combinatorial complexity bounds for arrangements of curves or spheres*, Discrete Comput. Geom. 5(1990), 99-160.

[6] A. Cordoba, *The Kakeya maximal function and spherical summation multipliers*, Amer. J. Math. 99 (1977), 1-22.

[7] R. O. Davies, *Some remarks on the Kakeya problem*, Proc. Cambridge Phil. Soc. 69(1971), 417-421.

[8] S. Drury, *L^p estimates for the x-ray transform*, Ill. J. Math. 27(1983), 125-129.

[9] L. Kolasa, T. Wolff, *On some variants of the Kakeya problem*, preprint.

[10] J. M. Marstrand, *Packing circles in the plane*, Proc. London Math. Soc. 55 (1987), 37-58.

[11] T. Mitsis, *A problem related to sphere and circle packing*, preprint.

[12] I. Z. Ruzsa, *Sums of finite sets*, Number theory (New York, 1991–1995), 281–293, Springer, New York, 1996.

[13] W. Schlag, *A generalization of Bourgain's circular maximal theorem*, J. Amer. Math. Soc. 10(1997), 103-122.

[14] W. Schlag, *A geometric proof of the circular maximal theorem*, preprint.

[15] W. Schlag, *A geometric inequality with applications to the Kakeya problem in three dimensions*, preprint.

[16] W. Schlag, C. Sogge, *Local smoothing estimates related to the circular maximal theorem*, Math. Research Letters 4(1997), 1-15.

[17] E. M. Stein, *Maximal functions: spherical means*, Proc. Nat. Acad. Sci. USA 73(1976), 2174-2175.

[18] T. Tao, A. Vargas, L. Vega, *A bilinear approach to the restriction and Kakeya conjectures*, preprint; and a second paper in preparation.

[19] T. Wolff, *An improved bound for Kakeya type maximal functions*, Revista Math. Iberoamericana 11 (1995), 651-674.

[20] T. Wolff, *A Kakeya type problem for circles*, Amer. J. Math. 119(1997), 985-1026.

[21] T. Wolff, *A mixed norm estimate for the x-ray transform*, preprint.

[22] T. Wolff, *Recent work connected with the Kakeya problem*, Princeton University 250th Anniversary Proceedings, in press.

Thomas Wolff
Department of Mathematics
253-37 Caltech
Pasadena, Ca 91125, USA
wolff@cco.caltech.edu

Section 9

Ordinary Differential Equations and Dynamical Systems

In case of several authors, Invited Speakers are marked with a *.

Rigidity and Renormalization in One Dimensional Dynamical Systems

W. de Melo[1]

Abstract. If two smooth unimodal maps or real analytic critical circle maps have the same bounded combinatorial type then there exists a $C^{1+\alpha}$ diffeomorphism conjugating the two maps along the corresponding critical orbits for some $\alpha > 0$. The proof is based on a detailed understanding of the orbit structure of an infinite dimensional dynamical system: the renormalization operator.

1. Introduction

A smooth discrete dynamical system is generated by a smooth transformation $f: M \to M$ of a compact manifold M called the phase space. Its dynamics involves an infinite number of maps, the iterates of f, defined inductively by $f^1 = f$, $f^n = f \circ f^{n-1}$. Accordingly, for points x in the phase space we have the notions of positive orbit, $\{x \in M\colon\ f^n(x); n \geq 0\}$, negative orbit, $\{y; f^m(y) = x, m \geq 0\}$ and the grand orbit, $\{y; f^m(y) = f^n(x), m, n \geq 0\}$. In the qualitative theory of dynamical systems, the natural equivalence relation to express the notion of "same dynamics" is conjugacy: f and g are conjugate if there exists a homeomorphism $h: M \to M$ such that $h \circ f = g \circ h$. Such a homeomorphism, called a conjugacy between f and g, maps orbits of f into orbits of g. If $0 \leq r \leq \infty$ then the space of C^r dynamical systems with the C^r topology is a Baire space and if $r < \infty$ it is even a Banach manifold. Similarly, if the parameter space, say P, is also a compact manifold then the space of C^r families of mappings $F: P \times M \to M$ is also a Baire space. Hence, we can talk about typical dynamical systems or typical parametrized families when they belong to a residual subset of the full space (in particular dense). In the case of a given specific parametrized family of dynamical systems, we have a different notion of typical: a property is Lebesgue typical if it is satisfied for maps corresponding to a full Lebesgue measure in the parameter space.

In real one-dimensional dynamical systems, the phase space is either a compact interval or the unit circle. In both cases we have an order structure and we say that two orbits have the same combinatorial type if the mapping that sends the $i-th$ element of one orbit into the $i-th$ element of the other orbit is order preserving. If this correspondence is smooth we say that the orbits have the same geometric type: indeed, a smooth map, being infinitesimally affine, preserves the small scale geometric properties of the orbits (for instance, the Hausdorff dimensions of the closures of the two orbits are the same).

Let us consider the following parametrized families of maps.

$$q_a: [-1,1] \to [-1,1], \quad q_a(x) = -ax^2 + 1, \quad 0 < a \leq 2 \tag{1}$$

[1]This work has been partially supported by the Pronex Project on Dynamical Systems

$$A_{a,b}\colon \mathbf{S}^1 \to \mathbf{S}^1; \qquad A_{a,b}(x) = x + a + \frac{b}{2\pi}\sin(2\pi x) \pmod 1, \tag{2}$$
$$\text{where } 0 \le a \le 2 \quad \text{and } 0 \le b$$

All the maps in the family (1) have a unique critical point. If $b < 1$, the mapping $A_{a,b}$ is a circle diffeomorphism; if $b = 1$, it is a critical circle mapping: a smooth homeomorphism with a unique critical point of cubic type and, lastly, for $b > 1$ it is not invertible and the dynamics becomes "chaotic".

These maps have played a crucial role in the development of one-dimensional dynamics for two reasons. First, they exhibit all possible combinatorial behavior for a large class of maps, namely, the maps of the form $\phi \circ q_1 \circ \psi$, where ϕ and ψ are interval diffeomorphisms, which we call *fold maps* (or unimodal maps) and the circle maps of the form $\Phi \circ A_{0,1} \circ \Psi$, where Φ and Ψ are circle diffeomorphisms, and these are called *critical circle maps*. The second reason is that maps in these families exhibit a very complex dynamical behavior, varying wildly with the parameter and thus producing a rich bifurcation set. All these facts were already well established at the end of the 70's when new unexpected quantitative discoveries greatly enhanced the interest of mathematicians and physicists in the subject.

These discoveries came from two different sources. From pure mathematics through the fundamental work of M. Herman on the smoothness of conjugacies between circle diffeomorphisms in [H], [Yoa]. The other input came from physics. Inspired by the scaling laws observed in phase transition and the renormalization group ideas developed in statistical mechanics to explain these phenomena, [W], Feigenbaum [F] and independently Coullet-Tresser [CT], performing numerical experiments with parametrized families of interval maps similar to (1) above, detected similar scaling laws, both in the phase space and parameter space, that were universal in the sense that they were independent of the particular family under consideration. Furthermore, they conjectured that these quantitative properties might be explained using renormalization group ideas adapted to this setting. It would correspond to a dynamical system (the renormalization operator) acting in the space of one-dimensional dynamical systems and the scaling laws observed would be a consequence of the hyperbolicity of a fixed point of this operator. The renormalization operator, which will be defined in section 3, is just the first return map of the original dynamical system to a smaller interval around the critical point, rescaled to the original size. Hence, the iterates of the renormalization operator reveals the small scale geometric properties of the critical orbit. Similar experiments were performed for critical circle mappings in [FKS] and analogous conjectures were formulated [La2], [Ra].

After a computer assisted proof of these conjectures [La1], a great effort was made to provide a conceptual proof of these and some extended conjectures. The main contribution came from Sullivan [Su], who introduced in the theory several new ideas and tools from real and complex analysis. He was able to prove the existence of a Cantor set in the space of fold maps that is invariant under iteration of the renormalization operator and such that the restriction of the operator to this set is a homeomorphism topologically conjugate to a full shift of a finite number of symbols. The Feigenbaum-Coullet-Tresser's fixed point corresponds to one of the fixed point of the shift map. Furthermore, for each fold map f with bounded

combinatorial type there exists a map g in the Cantor set, such that the iterates of the renormalization operator at f and g converges to each other. Also, the maps in the Cantor attractor are real analytic with very nice holomorphic extensions to the complex plane: they belong to a compact set in the space of quadratic-like maps in the sense of Douady and Hubbard, [DH]. After this, the powerful arsenal from conformal dynamics could be used and McMullen was able to prove the exponential convergence of iterates of the renormalization operator at quadratic like maps with the same bounded combinatorial type, thus establishing a strong rigidity result for such maps: their critical orbits have the same geometric type (see his paper in this volume and Theorem 3.5 in section 3). Finally, Lyubich in [Ly] proved the full hyperbolicity of the invariant Cantor set in the context of germs of quadratic-like maps. Once again Smale's horse-shoe shows up as a basic ingredient of dynamics!

To extend this result to the setting of smooth dynamical systems one has to overcome many technical difficulties most of them arising from the fact that although the space of dynamical systems is a nice Banach manifold the renormalization operator is not differentiable. The rigidity of the critical orbit of infinite renormalizable smooth maps of bounded type is related to the exponential convergence of iterates of the renormalization operator (Theorem 3.4 below) and an extension of the rigidity result for smooth mappings is discussed in [dMP] A partial result on the hyperbolicity in the setting of smooth maps was obtained in [D] and [FMP] to be discussed in the next section.

The results of Herman for circle diffeomorphisms can also be treated using renormalization ideas, as done in [SK], using only arguments from real analysis as in Herman's original proof.

We also point out that Martens in [M] proved the existence of the periodic points of the renormalization operator using only real analysis, extending the result to a broader class of maps having non-integer power law critical points.

2. Rigidity in phase space and parameter space.

A *fold map* (or unimodal map) is a smooth map f of a compact interval I that has a unique quadratic critical point $c_f \in I$, namely, $f = \phi \circ q \circ \psi$ where ϕ, ψ are $C^r, r \geq 2$, diffeomorphisms of compact intervals and $q(x) = x^2$.

The combinatorial type of any orbit of a fold map is determined by the combinatorial type of the critical orbit (see [MS], pp. 92). Therefore, we say that two fold maps f and g have the same *combinatorial type* if the mapping $f^i(c_f) \mapsto g^i(c_g)$, for $i \in \mathbf{N}$ is order preserving, where $f^1 = f$ and $f^i = f \circ f^{i-1}$ is the i-th iterate of f.

A fold map f is renormalizable if there exists a periodic interval J around the critical point of period $p \geq 2$, i.e., $f^p(J) \subset J$ and the interior of the intervals $J, f(J), \ldots, f^{p-1}(J)$ are pairwise disjoint. Hence, the restriction of f^p to J is again a fold map. To a renormalizable map f we can associate the set of positive integers $\mathcal{P}_f = \{2 \leq q_1 < q_2, \ldots\}$ of periods of renormalization. We say that f is infinitely renormalizable with bounded combinatorial type if the cardinality of $\mathcal{P}_f$ is infinite and the quotient of any two consecutive elements of $\mathcal{P}_f$ is bounded by some integer N. For ∞-renormalizable C^2 maps without periodic attractors, the critical orbit has an even stronger role since its closure is the global attractor: the

ω-limit set of Lebesgue almost all points in the dynamical interval is the closure of the critical orbit.

By combining the results of Sullivan [Su], McMullen, [Mc], [Mcb] and Lyubich [Ly], we prove in [dMP] the following rigidity result:

Theorem 2.1 (Rigidity in phase space). *If f and g are C^2 ∞-renormalizable fold maps with the same bounded combinatorial type, then there exists a $C^{1+\alpha}$ diffeomorphism $h:\mathbf{R} \to \mathbf{R}$ such that $h(f^i(c_f)) = g^i(c_g)$ for all $i \in \mathbf{N}$, where the Hölder exponent $\alpha > 0$ depends only on the bound on the combinatorial type.*

Remark 1.. An example in [FM1] can be adapted to show that the above result is false if the combinatorial type is not bounded.

Remark 2.. Even if the maps are very smooth we cannot expect the mapping h to be much smoother. This is in contrast with the case of circle diffeomorphisms treated by Herman, where, if the combinatorics is correct, the conjugacy is C^∞ if the diffeomorphisms are C^∞.

Let us give a geometric interpretation of the above theorem. Consider the complement of the closure of the critical orbit in the complex plane, $S_f = \mathbf{C} \setminus \text{Closure } (\{f^n(c_f), n \geq 0\})$, endowed with the hyperbolic metric (complete Riemannian metric of constant curvature -1). If f has bounded combinatorics, then there exists a family of closed geodesics that partition S_f in a countable number of pairs of pants and the lengths of the geodesics are uniformly bounded from above and from below (Corollary 3.1 of section 3). S_f is a tree of pairs of pants connected by the closed geodesics in the boundary, so each pair of pants has a "height" in this tree. If two map f and g have the same combinatorial type, then the partition of S_f and S_g into pairs of pants are isormorphic: there exists a homeomorphisms between the two surfaces respecting the partition and the height. Now, Theorem 2.1 implies that the differences between the lengths of the corresponding geodesics converge to zero exponentially fast so that the corresponding pair of pants becomes closer and closer to being isometric.

Let us formulate a rigidity conjecture in the parameter space.

Conjecture (Rigidity in the parameter space). *Let q_a , $1 \leq a \leq 2$ be the quadratic family $q_a(x) = -ax^2 + 1$. Given $N \geq 2$ and a typical family f_t of C^r fold maps, there exists a $C^{1+\epsilon}$ map $k_N:\mathbf{R} \to \mathbf{R}$ such that*

a) f_t is ∞-renormalizable with combinatorial type bounded by N if and only if $q_{k_N(t)}$ has the same combinatorial type as f_t;

b) k_N is piecewise monotone with a finite number of turning points corresponding to maps that are hyperbolic (the critical point of the mapping belongs to the basin of attraction of a periodic point).

Another way to formulate this conjecture is to say that the space of maps of combinatorial type bounded by N is laminated by smooth codimension-one submanifolds consisting of maps with the same combinatorial type and the holonomy of this lamination is $C^{1+\epsilon}$. The quadratic family intersects transversally each leaf of the lamination in a unique point and intersects the lamination in a Cantor set of Hausdorff dimension bigger than 0 and smaller than 1. A typical family is also transversal to the leaves and intersects each leaf in at most a bounded number of

points. This conjecture was verified for analytic families of quadratic like maps in [Ly].

We are still far from proving this conjecture in the setting of smooth maps. In this direction, by combining the results of A. M. Davie [D] and of Lyubich [Ly], we get as a consequence of [FMP]:

THEOREM 2.2. *If r is big enough then, in the space of C^r fold maps, the set of ∞-renormalizable fold maps with the same combinatorial type bounded by N is a C^1 codimension-one Banach submanifold.*

The combinatorial behavior of a critical circle mapping without periodic points is characterized by a unique real number since any such a map f is combinatorially equivalent to a rigid rotation $R_\alpha: x \mapsto x + \alpha \mod 1$, where the irrational number α is called the rotation number of f. Even more related to the dynamics is the set of positive integers that give the continued fraction decomposition of $\alpha = [a_0, a_1, \dots, a_n, \dots]$,

$$[a_0, a_1, \dots, a_n, \dots] = \cfrac{1}{a_0 + \cfrac{1}{a_1 + \cfrac{1}{\cdots \cfrac{}{a_n + \cfrac{1}{\cdots}}}}},$$

We say that the combinatorial type of f is bounded by N if $a_i \leq N$ for all i. The main result of [FM1] and [FM2] is

THEOREM 2.3. *If f and g are real analytic critical circle mappings with the same bounded combinatorial type, then there exists a $C^{1+\alpha}$ conjugacy between f and g where $\alpha > 0$ depends only on the bound on the combinatorial type.*

As in the case of fold maps we cannot expect to have a much better regularity of the conjugacy. We expect the result to hold also for smooth critical circle maps, except when the combinatorics is unbounded in which case we have a counter example in [FM1]. However, Yoccoz proved in [Yob] that two critical circle mappings with the same irrational rotation number are topologically conjugate and in fact, as he proved later in an unpublished paper (see [FM1]), the conjugacy is always quasi-symmetric. This is again in contrast to the situation of circle diffeomorphisms where the conjugacy is not in general quasi-symmetric if the rotation number is Liouville.

The same type of rigidity in parameter space is expected for critical circle maps: for typical one parameter families, the rotation number is a piecewise monotone function of the parameter and the correspondence between parameters corresponding to maps having bounded combinatorial type should be $C^{1+\epsilon}$.

In [Mcc], McMullen proved a result similar to Theorem 2.3 in the context of Siegel discs of quadratic-like maps, see also his article in this volume.

3. The renormalization operator

Any C^r fold map is smoothly conjugate to a C^r fold map $f\colon[-1,1]\to[-1,1]$ of the form $f=\phi\circ q$ where $q(x)=x^2$ and $\phi\colon[0,1]\to[-1,1]$ is a C^r embedding with $\phi(0)=1$. Hence we can restrict our attention to the space $\mathcal{F}^r$ of C^r fold maps so normalized. Let $\mathcal{D}$ be the set of maps in $\mathcal{F}$ that are renormalizable. If $f\in\mathcal{D}$ and its minimum period of renormalization is p we set

$$\mathcal{R}(f)(x)=\frac{1}{\lambda}f^p(\lambda x)\quad\text{where}\quad\lambda=f^p(0) \tag{3}$$

The mapping $\mathcal{R}\colon\mathcal{D}\to\mathcal{F}$ is the *renormalization operator*. To each $f\in\mathcal{D}$ with minimum renormalization period p we can associate a permutation γ of $\{1,2,\dots,p\}$ as follows. Let $I_1,\dots,I_p$ be a labeling of the intervals J, with endpoints $f^p(0),-f^p(0)$, $f(J),\dots,f^{p-1}(J)$ compatible with their ordering in the real line. Then γ is defined by $f(I_j)\subset I_{\gamma(j)}$. We call γ a unimodal permutation because it has the following properties: i) if we plot the graph of γ and connect the consecutive points by a line segment we get a unimodal map; ii) the iterates of any point by γ is the whole set; iii) there is no partition of the domain in disjoint subsets that are permuted by γ. Conversely, if γ is a unimodal permutation there exists a function in the quadratic family that is renormalizable and has γ as the associated permutation. Therefore, we can write the domain $\mathcal{D}$ of the renormalization operator as a disjoint union of $\mathcal{D}_{\gamma_i}$, where $\{\gamma_i,\quad i=1,2\dots\}$ is the set of all unimodal permutations. If $r\le\infty$, then $\mathcal{D}_{\gamma_i}$ are clearly the connected components of $\mathcal{D}$. The intersection of each $\mathcal{D}_{\gamma_i}$ with the quadratic family is an interval (see [MS], pp. 194). From a profound theorem of Yoccoz, [Hu], it follows also that the intersection of $\mathcal{D}$ with the quadratic family is dense in the interval $(0,1]$.

Two $\infty-$renormalizable fold maps f and g have the same combinatorial type if and only if $\mathcal{R}^n(f)$ and $\mathcal{R}^n(g)$ belong to the same $\mathcal{D}_{\gamma(i)}$ for every $n\ge0$. Hence, each combinatorial type is given by a sequence $\gamma(i),\quad i\ge0$, of unimodal permutations and vice-versa.

The non-linearity of a C^2 interval map ϕ at a point $x\in$ is defined as

$$N\phi(x)=\frac{D^2f(x)}{Df(x)}.$$

Theorem 3.1 (Sullivan [Su]). *There exists a universal constant $d>0$ such that if $f\in\mathcal{F}^r,\quad r\ge2$, is infinitely renormalizable with bounded combinatorial type then $\mathcal{R}^n(f)=\phi_n\circ q$, where the non-linearity of ϕ_n is bounded by d for all $n\ge n_0=n_0(f)$.*

Notice that this is a compactness kind of result since, by Ascoli's theorem, it implies the existence of a subset $\mathcal{K}\subset\mathcal{F}$, which is compact in the C^k topology for $k<r$, such that the iterates of any map by the renormalization operator eventually belong to $\mathcal{K}$. If f is at least C^3, we can also prove theorem 3.1 for maps with unbounded combinatorics.

There is an important consequence on the geometry of the critical orbit. We say that a Cantor set in the complex plane has bounded geometry if there exists a family of closed geodesics of the hyperbolic metric of the complement of the Cantor set that are uniformly bounded from above and from below and that separates the space into pair of pants.

COROLLARY 3.1. *If f is infinitely renormalizable of bounded combinatorial type, then the closure of the critical orbit is a Cantor set of bounded geometry. In particular, its Hausdorff dimension is bigger than 0 and smaller than one.*

Another important consequence of Theorem 3.1 is that the limit set of the renormalization operator restricted to the subset $\mathcal{D}^{(N)} \subset \mathcal{D}$ of fold maps with renormalizable minimum period $\leq N$, is a compact set in a much stronger topology. To state this result we need the basic definition below.

DEFINITION 3.1. A fold map $f = \phi \circ q$ belongs to the Epstein class $\mathcal{E}$, if ϕ has a holomorphic extension Φ to a topological disc such that: 1) Φ is one-to-one; 2) the image of ϕ is equal to the topological disc $\mathbf{C}(L) = (\mathbf{C} \setminus \mathbf{R}) \cup L$, where L is an interval containing the image of ϕ.

For holomorphic maps defined on topological discs we may consider the Carathéodory topology [Mc]. This is a very strong topology. Indeed, if a holomorphic extension F_n of a fold map f_n converges to a holomorphic extension F of f, then there exist a neighborhood U of the dynamical interval which is contained in the domains of all the maps and such that the restriction of the sequence to U converges uniformly to the restriction of f. In particular f_n converges to f in the C^r topology for any r, namely $|f_n - f|_r \to 0$, where $|f_n - f|_r = \sup\{|f_n(x) - f(x)|, \ldots, |D^r f_n(x) - D^r f(x)|\}$.

COROLLARY 3.2. *There exists a compact subset $\mathcal{C}_N \subset \mathcal{E}$ in the Carathéodory topology such that if f is an ∞-renormalizable C^2 fold map of combinatorial type bounded by N , there are constants $C > 0$ and $0 < \lambda < 1$ and a sequence f_n having a holomorphic extension in $\mathcal{C}_N$ such that $|\mathcal{R}^n(f) - f_n|_0 \leq C\lambda^n$. In particular all the limit set of iterates of the renormalization operator at maps of combinatorial type bounded by N is contained in $\mathcal{C}_N$.*

The proof of Theorem 3.1 and its corollaries involve only real analysis and the main ingredient is the Koebe's distortion theorem and the control of the distortion of the cross ratio under iteration [MS].

The maps of the Epstein class have the important property that each branch of its inverse contracts the hyperbolic map of the upper half space by the Schwarz Lemma. This is the basic tool in the proof of the next theorem which is a bridge between real and complex dynamics.

DEFINITION 3.2. A quadratic-like map is a holomorphic map $F\colon U \to V$ between topological discs U and V such that F is proper, two-to-one, and the open disc V contains the closure of U. The modulus of F is defined as the conformal modulus of the annulus $V \setminus \text{Closure}(U)$.

THEOREM 3.2 (SULLIVAN [Su]). *For each N, there exists a subset $\mathcal{S}$ of quadratic-like maps which is compact in the Carathéodory topology and $0 < \lambda < 1$ such that if f is an infinitely renormalizable C^2 fold map of combinatorial type bounded by N, then there exists a sequence of fold maps f_n having holomorphic extensions in $\mathcal{S}$ and a positive constant $C > 0$ so that $|\mathcal{R}^n(f) - f_n|_0 \leq C\lambda^n$. In particular all maps in the limit set of the renormalization operator restricted to the set of maps of combinatorial type bounded by N have holomorphic extensions in $\mathcal{S}$.*

If $F: U \to V$ is a quadratic-like map then the set $K_F = \{z \in U; F^n(z) \in U \ \forall n \geq 0\}$ is called the filled-in Julia set of F and its boundary is the Julia set of F. If F is the extension of a fold map, then K_F contains the dynamical interval $[-1, 1]$ and all its pre-images. If f is infinitely renormalizable the filled-in Julia set has empty interior and is equal to the Julia set. Theorem 3.1 combined with Sullivan's pull-back argument gives the following:

Corollary 3.3. *If F and G are quadratic-like holomorphic extensions of ∞-renormalizable fold maps of the same bounded combinatorial type then there exists a quasi-conformal conjugacy between F and G in a neighborhood of the Julia set.*

Using Corollary 3.3 and the measurable Riemann mapping theorem, one can perform quasi-conformal deformations of germs of quadratic-like maps that are extensions of fold maps. Using this and some extensions of the Teichmüler theory to Riemann surfaces laminations connected to such germs, Sullivan arrived at the theorem below that describes completely the dynamics of the renormalization operator in the space of maps of bounded combinatorial type. Let P_N be the finite set of unimodal permutations of length at most N. Let Σ_N be the set of biinfinite sequences $\theta: \mathbf{Z} \to P_N$ endowed with the product topology and let $\sigma: \Sigma_N \to \Sigma_N$ be the shift homeomorphism $\sigma(\theta)(i) = \theta(i+1)$.

Theorem 3.3 (Sullivan [Su]). *There exists a one to one continuous mapping $\theta \in \Sigma_N \mapsto f_\theta \in \mathcal{F}$ with the following properties:*

1. *f_θ is ∞-renormalizable of combinatorial type $(\theta(0), \theta(1), \dots)$ and has a quadratic like extension that belongs to the compact set $\mathcal{S}$.*
2. *$\mathcal{R}(f_\theta) = f_{\sigma(\theta)}$*
3. *If f is a C^2 infinitely renormalizable map of the same combinatorial as f_θ then $|\mathcal{R}^n f - \mathcal{R}^n f_\theta|_0$ converges to zero as $n \to \infty$.*

The relevance of the convergence of the renormalization operator to the rigidity problem in phase space is given by the following result, which is proved using again Theorem 3.1 (see [MS], pp. 546):

Theorem 3.4. *Let f and g be two C^2, infinitely renormalizable, fold maps with the same bounded combinatorial type. If $|\mathcal{R}^n(f) - \mathcal{R}^n(g)|_0$ converges to zero exponentially fast, then there exists a $C^{1+\alpha}$ diffeomorphism of the real line that conjugates the two maps along the critical orbits.*

The compact set $\Lambda_N = \{f_\theta; \theta \in \Sigma_N\}$ is called the renormalization limit set since, by Theorem 3.3, the union of the basin of attraction of the functions in Λ_N is equal to the set of all infinitely renormalizable maps of combinatorial type bounded by N. However Theorem 3.3 does not give yet a rate of convergence. The first main step in getting the exponential convergence needed in Theorem 3.4 comes from the work of McMullen. Using Sullivan's compactness theorem and a rigidity result a la Mostow in the geometric limit of renormalization, he was able to prove that the quasi-conformal conjugacy in Corollary 3.3 is in fact $C^{1+\alpha}$ at the critical point (see [Mc], [Mcb] and his article in this volume) proving the following:

Theorem 3.5 (McMullen). *Let f and g be ∞-renormalizable fold maps with the same bounded combinatorial type. If f and g have quadratic-like extensions then $|\mathcal{R}^n(f) - \mathcal{R}^n(g)|_0$ converges to zero exponentially fast.*

The final basic step is Lyubich's hyperbolicity of the renormalization limit set Λ_N in the space of germs of quadratic like maps [Ly]. His result implies that the iterates of the renormalization operator expand exponentially the distance between two maps with different combinatorics. Even the precise formulation of this statement is subtle because there is no natural domain of definition for the holomorphic extensions of the maps in the Cantor set Λ_N and, therefore, we do not have a Banach space of maps where the operator acts smoothly. One of the major achievements is Lyubych's paper is to provide a natural complex analytic structure to the set of germs of quadratic-like maps (modeled in a direct set of Banach spaces) and a complex holomorphic operator with respect to this structure that restricts to the renormalization operator. It is with respect to this structure that he formulates and proves the hyperbolicity of the Λ_N.

In [dMP] we use the expanding direction of Λ_N, in the space of germs of quadratic-like maps, to improve Theorem 3.2: the iterates of the renormalization operator at an infinitely renormalizable map f of bounded combinatorial type can be exponentially approximated by a map f_n in $\mathcal{S}$ with the same combinatorics as $\mathcal{R}^n(f)$. Combining this with McMullen's exponential contraction we arrive at Theorem 2.1.

In [FMP], we translate Lyubich's hyperbolicity statement in terms of an operator acting on an open set of a finite union of Banach spaces of holomorphic functions containing the renormalization limit set Λ_N as a hyperbolic set in the usual sense and such that the new operator restricts to an iterate of the renormalization operator. Extending to this setting the analytic estimates of [D], we show that the hyperbolicity feature persists in the space of C^r fold maps for r big enough with C^1 local stable manifolds (and real analytic local unstable manifolds given by Lyubich).

In a remarkable paper [Lyb], Lyubich proved the hyperbolicity of the renormalization operator in the space of all renormalizable maps including those of unbounded type. In particular he was able to extend Theorem 3.5 to maps with any combinatorial type. However, this is not sufficient to establish a rigidity result which, as we pointed out before, is false at least for smooth mappings.

4. Renormalization of critical circle mappings

Let us consider a critical circle mapping f without periodic points whose rotation number has the following continued fraction expansion: $\rho(f) = [a_0, \dots a_n, \dots]$. When the partial quotients a_n are bounded, we say that $\rho(f)$ is a number of *bounded type*

The denominators of the convergents of $\rho(f)$, defined recursively by $q_0 = 1$, $q_1 = a_0$ and $q_{n+1} = a_n q_n + q_{n-1}$ for all $n \geq 1$, are the *closest return times* of the orbit of any point to itself. We denote by Δ_n the closed interval containing c whose endpoints are $f^{q_n}(c)$ and $f^{q_{n+1}}(c)$. We also let $I_n \subseteq \Delta_n$ be the closed interval whose endpoints are c and $f^{q_n}(c)$. Observe that $\Delta_n = I_n \cup I_{n+1}$. The

most important combinatorial fact in the study of the geometry of a circle map is that for each n the collection of intervals

$$\mathcal{P}_n = \left\{I_n,\, f(I_n), \dots,\, f^{q_{n+1}-1}(I_n)\right\} \cup \left\{I_{n+1},\, f(I_{n+1}), \dots,\, f^{q_n-1}(I_{n+1})\right\}$$

constitutes a partition of the circle (modulo endpoints), called *dynamical partition of level* n of the map f. Note that, for all n, $\mathcal{P}_{n+1}$ is a refinement of $\mathcal{P}_n$.

Of course, these definitions make sense for an arbitrary homeomorphism of the circle. For a rigid rotation, we have $|I_n| = a_{n+1}|I_{n+1}| + |I_{n+2}|$. Therefore, if a_{n+1} is very large then I_n is much longer than I_{n+1}. It is a remarkable fact, first proved by Świątek and Herman, that this cannot happen for a critical circle map! Indeed, the dynamical partitions $\mathcal{P}_n$ have *bounded geometry*, in the sense that adjacent atoms have comparable lengths.

Commuting pairs and renormalization. Let f be a critical circle map as before, and let $n \geq 1$. The first return map $f_n : \Delta_n \to \Delta_n$ to $\Delta_n = I_n \cup I_{n+1}$, called the *n-th renormalization of f without rescaling*, is determined by a pair of maps, namely $\xi = f^{q_n} : I_{n+1} \to \Delta_n$ and $\eta = f^{q_{n+1}} : I_n \to \Delta_n$. This pair (ξ, η) is what we call a *critical commuting pair*. Each f_{n+1} is by definition the *renormalization without rescaling* of f_n. Conjugating f_n by the affine map that takes the critical point c to 0 and I_n to $[0, 1]$ we obtain $\mathcal{R}^n(f)$, the n-th renormalization of f.

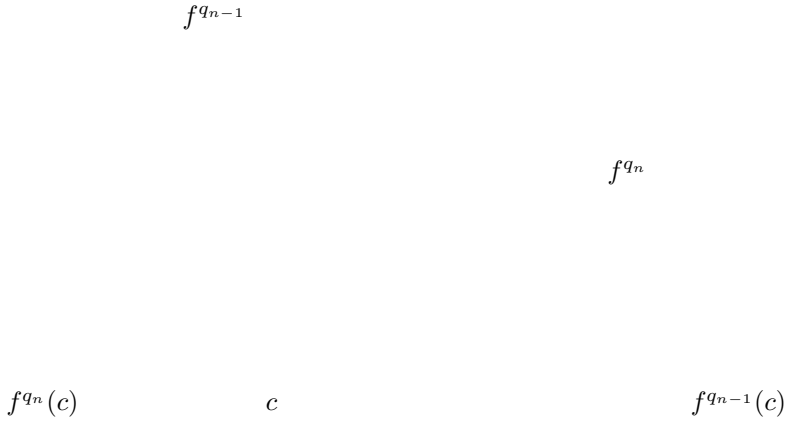


Figure 1. Two consecutive renormalizations of f.

HOLOMORPHIC PAIRS. The concept of *holomorphic commuting pair* was introduced in E. de Faria's thesis,[dF], and plays a crucial role in the proof of theorem 2.2. We recall the definition and some of the relevant properties of these objects, henceforth called simply *holomorphic pair*. Assume we are given a configuration of four simply-connected domains $\mathcal{O}_\xi, \mathcal{O}_\eta, \mathcal{O}_\nu, \mathcal{V}$ in the complex plane, called a *bowtie*, such that

(a) Each $\mathcal{O}_\gamma$ is a Jordan domain whose closure is contained in $\mathcal{V}$;

(b) We have $\overline{\mathcal{O}}_\xi \cap \overline{\mathcal{O}}_\eta = \{0\} \subseteq \mathcal{O}_\nu$;

(c) The sets $\mathcal{O}_\xi \setminus \mathcal{O}_\nu$, $\mathcal{O}_\eta \setminus \mathcal{O}_\nu$, $\mathcal{O}_\nu \setminus \mathcal{O}_\xi$ and $\mathcal{O}_\nu \setminus \mathcal{O}_\eta$ are non-empty and connected.

A holomorphic pair with domain $\mathcal{U} = \mathcal{O}_\xi \cup \mathcal{O}_\eta \cup \mathcal{O}_\nu$ is the dynamical system generated by three holomorphic maps $\xi : \mathcal{O}_\xi \to \mathbb{C}$, $\eta : \mathcal{O}_\eta \to \mathbb{C}$ and $\nu : \mathcal{O}_\nu \to \mathbb{C}$ satisfying the following conditions.

[H_1] Both ξ and η are univalent onto $\mathcal{V} \cap \mathbb{C}(\xi(J_\xi))$ and $\mathcal{V} \cap \mathbb{C}(\eta(J_\eta))$ respectively, where $J_\xi = \mathcal{O}_\xi \cap \mathbb{R}$ and $J_\eta = \mathcal{O}_\eta \cap \mathbb{R}$. (Notation: $\mathbb{C}(I) = (\mathbb{C} \setminus \mathbb{R}) \cup I$.)

[H_2] The map ν is a 3-fold branched cover onto $\mathcal{V} \cap \mathbb{C}(\nu(J_\nu))$, where $J_\nu = \mathcal{O}_\nu \cap \mathbb{R}$, with a unique critical point at 0.

[H_3] We have $\mathcal{O}_\xi \ni \eta(0) < 0 < \xi(0) \in \mathcal{O}_\eta$, and the restrictions $\xi|[\eta(0), 0]$ and $\eta|[0, \xi(0)]$ constitute a critical commuting pair.

[H_4] Both ξ and η extend holomorphically to a neighborhood of zero, and we have $\xi \circ \eta(z) = \eta \circ \xi(z) = \nu(z)$ for all z in that neighborhood.

[H_5] There exists an integer $m \geq 1$, called the *height* of Γ, such that $\xi^m(a) = \eta(0)$, where a is the left endpoint of J_ξ; moreover, $\eta(b) = \xi(0)$, where b is the right endpoint of J_η. The interval $J = [a, b]$ is called the *long dynamical interval* of Γ, whereas $\Delta = [\eta(0), \xi(0)]$ is the *short dynamical interval* of Γ. They are both forward invariant under the dynamics. The *rotation number* of Γ is by definition the rotation number of the critical commuting pair of Γ (condition H_3).

In the figure below the solid lines are mapped into the real axis and the heavier solid lines are mapped into the interval J. Notice that ν is not a polynomial-like map of degree three because the interval to the left of $\eta(0)$ in the domain of ν is in the image of the boundary of $\mathcal{O}_\nu$.

Figure 2. Holomorphic commuting pair.

A fundamental step in the proof of Theorem 2.3 is the statement that a sufficiently high renormalization of a real analytic critical circle map has holomorphic extension belonging to a compact set of holomorphic pairs. From that we prove that the limit set of this pair is "chaotic" and that the critical point is a deep point in the limit set in the sense of [Mcb]. From this point on we use McMullen's machinery, developed in chapter 9 of [Mcb], to prove Theorem 2.3.

References

[CT] P. Coullet and C. Tresser, *Itération d'endomorphismes et groupe de renormalization.*, C. R. Acad. Sci. Paris **287A** (1978), 577–580.

[D] A.M. Davie, *Periodic doubling for $C^{2+\epsilon}$ mappings*, Commun. Math. Phys. **176** (1996), 262–272.

[DH] A. Douady and J.H. Hubbard, *On the dynamics of polynomial-like maps*, Ann.Sc.Éc.Norm.Sup. **18** (1985), 287–346.

[dF] E. de Faria, *Asymptotic rigidity of scaling ratios for critical circle mappings*, IMS Stony Brook preprint 96/13, to appear in Erg.Th.Dynam.Sys..

[FM1] E. de Faria and W. de Melo, *Rigidity of critical circle mappings I*, IMS Stony Brook preprint 97/16.

[FM2] E. de Faria and W. de Melo, *Rigidity of critical circle mappings II* (1997), IMS Stony Brook preprint 97/17.

[FMP] E. de Faria W. de Melo and A. A. Pinto, *Global hyperbolicity of renormalization for C^r unimodal mappings*, In preparation..

[F] M. J. Feigenbaum, *Quantitative universality for a class of non-linear transformations*, J. Stat. Phys. **19** (1978), 25–52.

[FKS] M. J. Feigenbaum L. P. Kadanoff and S. J. Shenker, *Quasiperiodicity in dissipative systems: a renormalization group analysis.*, Phys. D **5** (1982), 370–386.

[H] M. Herman, *Sur la conjugaison differentiable des difféomorphismes du cercle a des rotations*, Publ. Math. IHES **49** (1979), 5–234.

[Hu] J. H. Hubbard, *Local connectivity of Julia sets and bifurcation loci: three theorems of J.-C. Yoccoz*, Topological Methods in Modern Mathematics, A Symposium in Honor of John Milnor, Publish or Perish, Inc, 1993, pp. 467–511.

[La1] O. Lanford III, *A computer assisted proof of the Feigenbaum conjectures.*, Bull. Amer. Math. Soc. **6** (1982), 427–434.

[La2] O. Lanford III, *Renormalization group methods for critical circle mappings with general rotation number*, VIIIth International Congress on Mathematical Physics (Marseille, 1986), World Sci. Publishing, Singapore, 1987, pp. 532–536.

[Ly] M. Lyubich, *Feigenbaum-Coullet-Tresser Universality and Milnor's Hairiness Conjecture*, Annals of Mathematics, to appear (1998).

[Lyb] M. Lyubich, *Almost every real quadratic map is either regular or stochastic*, IMS Stony Brook preprint 97/8 (1998).

[M] M.. Martens, *The periodic points of renormalization.*, IMS Stony Brook preprint 96/3, to appear in Annals of Math. (1996).

[Mc] C. McMullen, *Complex dynamics and renormalization*, Annals of Math Studies **135** (1994), Princeton University Press, Princeton.

[Mcb] ———, *Renormalization and 3-manifolds which fiber over the circle*, Annals of Math Studies **142** (1996), Princeton University Press, Princeton.

[Mcc] ———, *Self-similarity of Siegel disks and the Hausdorff dimension of Julia sets.*, Acta Mathematica, to appear.

[dMP] W. de Melo and A. A. Pinto, *Rigidity of C^2 infinitely renormalizable quadratic maps*, In preparation..

[MS] W. de Melo and S. van Strien, *One dimensional dynamics*, Springer-Verlag, Berlin and New York, 1993.

[Ra] D. Rand, *Global phase-space universality, smooth conjugacies and renormalization: the $C^{1+\alpha}$ case*, Nonlinearity **1** (1988), 181–202.

[SK] Ya. G. Sinai, K.M Khanin, *Smoothness of conjugacies of diffeomorphisms of the circle with rotations.*, Russian Math. Surveys **45** (1989), 69–99.

[Su] D. Sullivan, *Bounds, quadratic differentials and renormalization conjectures*, Mathematics into the Twenty-First Century, Amer. Math. Soc. Centennial Publication, vol.2, Amer. Math. Soc., Providence, RI, 1991.

[W] K. G. Wilson, *The renormalization group and critical phenomena*, Reviews of Modern Physics **55** (1983), 583–600.

[Yoa] J. C. Yoccoz, *Conjugaison differentiable des diffeomorphismes du cercle dont le nombre de rotation verifie une condition Diophantienne*, Ann. Sci. de l'Ec. Norm. Sup. **17** (1984), 333–361.

[Yob] J. C. Yoccoz, *Il n'y a pas de contre-example de Denjoy analytique,*, C. R. Acad. Sci. Paris **298** (1984), 141–144.

W. de Melo
IMPA,
Estrada Dona Castorina 110,
Jardim Botânico, CEP22460-320
Rio de Janeiro RJ - Brasil
email: demelo@impa.br

Reducibility and Point Spectrum for Linear Quasi-Periodic Skew-Products

L. H. Eliasson

Abstract. We consider linear quasi-periodic skew-product systems on $\mathbb{T}^d \times G$ where G is some matrix group. When the quasi-periodic frequencies are Diophantine such systems can be studied by perturbation theory of KAM-type and it has been known since the mid 60's that most systems sufficiently close to constant coefficients are reducible, i.e. their dynamics is basically the same as for systems with constant coefficients. In the late 80's a perturbation theory was developed for the other extreme. Fröhlich-Spencer-Wittver and Sinai, independently, were able to prove that certain discrete Schrödinger equations sufficiently far from constant coefficients have pure point spectrum, which implies a dynamics completely different from systems with constant coefficients. In recent years these methods have been improved and in particular $SL(2,\mathbb{R})$ — related to the the Schrödinger equation — and $SO(3,\mathbb{R})$ have been well studied.

1991 Mathematics Subject Classification: 34, 58, 81
Keywords and Phrases: skew-product, quasi-periodicity, reducibility, point spectrum

1. Introduction.

A *linear quasi-periodic skew-product system* on $\mathbb{T}^d \times G$ is a mapping

$$(\theta, X) \longmapsto (\theta + \omega, A(\theta)X) \tag{1}$$

where θ belongs to the d-dimensional torus $\mathbb{T}^d$, $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$, ω is a vector in $\mathbb{R}^d$ and A is a continuous function on $\mathbb{T}^d$ with values in some matrix subgroup G of $GL(D,\mathbb{R})$, for example $SL(2,\mathbb{R})$ or $SO(3,\mathbb{R})$. This system is often given as a time-one map of a system of linear differential equation

$$\frac{d}{dt}X(t) = A(\theta + t\omega)X(t), \tag{2}$$

in which case we talk about a *time-continuous* system, and it often naturally contains parameters.

What interests us is the time-evolution of the system (1). At time n it is described by a matrix product

$$A_n(\theta) = A(\theta + (n-1)\omega)\dots A(\theta + \omega)A(\theta), \tag{3}$$

whose behavior we want to study when $n \to \infty$. We are for example interested in if the product becomes unbounded or remains bounded and in the behavior of the eigenvalues.

As example we can consider *the time-discrete Schrödinger equation*

$$-(u_{n+1} + u_{n-1}) + V(\theta + n\omega)u_n \;=\; Eu_n \tag{4}$$

with spectral parameter E. This equation can be written as (1) with

$$A(\theta) \;=\; \begin{pmatrix} 0 & 1 \\ -1 & V(\theta) - E \end{pmatrix} \;\in\; SL(2, \mathbb{R}),$$

where E occurs as a free parameter. Another example is *the time-continuous Schrödinger equation*

$$-\frac{d^2}{dt^2}y(t) + V(\theta + t\omega)y(t) \;=\; Ey(t) \tag{5}$$

which can be written as a first order system of the type in (2) with

$$A(\theta) \;=\; \begin{pmatrix} 0 & 1 \\ V(\theta) - E & 0 \end{pmatrix}.$$

The fundamental solution (or monodromy matrix, or time-evolution operator, or propagator, or...) $\Phi_t(\theta, E)$ of this system is a matrix in $SL(2, \mathbb{R})$ and its time-evolution is determined by (1) if we let $A(\theta, E) = \Phi_1(\theta, E)$.

If $\omega/2\pi = (p_1/q_1, \ldots, p_d/q_d) \in \mathbb{Q}^d$ the system is *periodic* and otherwise it is *quasi-periodic*. If it is quasi-periodic one can without restriction assume that $\tilde{\omega} = (\omega, 2\pi)$ is rationally independent, i.e.

$$< k, \tilde{\omega} > \neq 0 \quad \text{whenever} \quad k \in \mathbb{Z}^{d+1} \setminus 0.$$

Here we distinguish two cases. We say that the frequencies are *Diophantine* if the vector $\tilde{\omega}$ is Diophantine, i.e.

$$|< k, \tilde{\omega} >| \geq \frac{\kappa}{|\, k \,|^\tau}, \quad \forall k \in \mathbb{Z}^{d+1} \setminus 0 \tag{6}$$

for some $k, \tau > 0$. If they are not Diophantine we say that they are *Liouville*. The set of Diophantine vectors is of full measure and the set of Liouville vectors is topologically generic, i.e. it is a dense $\mathcal{G}_\delta$.

2. Basic Concepts.

We consider first the periodic case $\omega = 2\pi(p_1/q_1, \ldots, p_d/q_d)$. The time-evolution of (1) for a given θ is determined by the spectral properties, in particular the eigenvalues, of the matrix $A_q(\theta)$, where q is a common multiple of all the q_i's. The best way to describe this evolution is to transform the system to a constant coefficient system. That this is possible for a time-continuous periodic system

(2) was shown by Floquet by an easy argument. For a discrete system (1) the argument is even easier and it gives that there exists a change of variables on $(2\mathbb{T})^d \times G$, as smooth as A but only piecewise,

$$(\theta, X) \longmapsto (\theta, C(\theta)X)$$

which conjugates (1) to another skew-system on $(2\mathbb{T})^d \times G$

$$(\theta, X) \longmapsto (\theta + \omega, B(\theta)X)$$

where B is constant along the orbits $\{\theta + k\omega\}_{k\in\mathbb{Z}}$, i.e.

$$B(\theta) = B(\theta + k\omega), \quad \forall\theta \in \mathbb{T}^d, \ \forall k \in \mathbb{Z}.$$

An equivalent formulation is that there exists a matrix $C(\theta)$ such that

$$A(\theta) = C(\theta + \omega)B(\theta)C^{-1}(\theta). \tag{7}$$

(The "period-doubling" which reflects that C is defined on $(2\mathbb{T})^d$ and not on $\mathbb{T}^d$ is necessary if one doesn't want to complexify the system.)

This illustrates the concept of *reducibility*, which was first considered by Lyapunov [1]. It is not obvious what conditions one should require of the transformation C but for periodic and quasi-periodic systems we shall demand that C is defined on some finite covering of the torus and is piecewise continuous. With such a choice periodic systems are always reducible while quasi-periodic systems, as we shall discuss below, turns out not to be. If a quasi-periodic system is reducible however, then the matrix B will be independent of θ. If the transformation C is, say, analytic then we talk about *analytic reducibility*. One could also consider weaker conditions on C: a transformation that is only measurable would a priori be interesting but no results are known in this direction.

A reducible system has Floquet exponents which are nothing but the eigenvalues of the matrix B. The imaginary parts of the Floquet exponents are only defined modulo

$$\{\frac{i}{2} < k, \tilde{\omega} > : k \in \mathbb{Z}^{d+1}\}.$$

In general there is no unique and independent way to specify these imaginary parts except in the case $SL(2, \mathbb{R})$ where they are identified as $\pm$*the rotation number* [2].

The real parts of the Floquet exponents have an independent characterization as *the Lyapunov exponents* which exist for all quasi-periodic skew-products. In fact by a theorem of Oseledet's for a.e. θ there is a measurable decomposition of $\mathbb{R}^D$ into a sum of invariant subspaces

$$\mathbb{R}^D = \bigoplus_i W_i(\theta), \quad \dim W_i(\theta) = m_i \tag{8}$$

such that

$$\lim_{n\to\infty} \frac{1}{n} \log \mid A_n(\theta)\bar{u} \mid = \lambda_i, \qquad \forall \bar{u} \in W_i(\theta). \tag{9}$$

We call the λ_i's and their multiplicities m_i *the Lyapunov spectrum* of the system. If the system is reducible then the Lyapunov spectrum coincides with the real part of the spectrum of B and it is *uniform* — the decomposition (8) is continuous and the limits (9) exist for all θ. There is a somewhat weak converse of this result when all exponents have multiplicity one: if the system has uniform and simple Lyapunov spectrum and if ω is Diophantine then it is reducible [3]. The assumptions of this theorem are however hard to verify in general.

We now turn to the quasi-periodic case. In distinction to the periodic case, quasi-periodic systems are not always reducible. For example, a reducible system must be *regular* in the sense of Lyapunov, i.e.

$$\sum(\text{Lyapunov exponents}) \;=\; \lim_{n\to\infty}\frac{1}{n}\sum_{l=0}^{n-1}\mathrm{Re}\,(Tr(A(\theta+l\omega))$$

where the Lyapunov exponents are counted with multiplicities [1]. Examples are known at least since the 60's of irregular time-continuous quasi-periodic systems [4,5]. This notion however provides little insight into the dynamics of the system.

A reducible system cannot have a *point eigenvalue*, i.e. there cannot exist a sequence of vectors $\{v_n : n \in \mathbb{Z}\}$ in $l^2(\mathbb{Z}) \otimes \mathbb{R}^D$ such that

$$v_{n+1} - A(\theta + n\omega)v_n \;=\; Ev_n, \quad \forall n \in \mathbb{Z},$$

for some constant E. Examples of time-continuous quasi-periodic systems with point eigenvalues were given in [6,7]. These examples are not smooth on the torus $\mathbb{T}^d$ however. The first smooth example came as a consequence of a famous theorem on reducibility [8]. The almost Mathieu equation

$$-(u_{n+1} + u_{n-1}) + K\cos(\theta + n\omega)u_n \;=\; Eu_n$$

with Diophantine frequencies was proven to be reducible for small enough K and for certain values of E. As a consequence of the "self-dual" character of this equation under the Fourier transform it must therefore have point eigenvalues for K large enough [9].

A reducible system must be integrable in the sense that it has an invariant foliation of the space $\mathbb{T}^d \times G$ into submanifolds whose dimension equals d plus the dimension of the center of G. In particular such a system cannot be *transitive*, much less be *ergodic*. Examples of ergodic quasi-periodic skew-products were constructed in [10]. These examples are smooth on $\mathbb{T}^d$ but the frequencies are Liouville.

Nor can reducible systems be *non-uniformly hyperbolic* because, as we mentioned above, the Lyapunov spectrum of a reducible system must be uniform. Examples of non-uniformly hyperbolic systems of Schrödinger type are given in [2,11] and of other types in [12].

Because of the existence of both reducible and non-reducible quasi-periodic systems two questions occur naturally. What is the structure of the set of reducible and non-reducible systems respectively? And what are the typical dynamical properties of the non-reducible systems. We shall provide some answers to these questions in the case when ω is Diophantine (6) and the system is analytic and either close to or far from constant coefficients.

3. Close to Constant Coefficients.

Reducibility of quasi-periodic skew-products close to constant coefficients was obtained by KAM-arguments already in the 60's. The first results were proven under the assumption of sufficiently many parameters [13]. That in general only one parameter is needed became obvious in [14] — where in particular the case $Sp(n,\mathbb{R})$ is treated — and it was proven in general in [15]. These results give reducibility for all parameter values except a small but positive measure set, but the following stronger statement should be true.

Conjecture. *Any generic analytic one-parameter family of skew-systems* (1) *sufficiently close to constant coefficients is reducible for a.e. parameter value.*

The first verification, and the motivation, of this conjecture was done in [16], where previous results [8,17] on the quasi-periodic Schrödinger equation were extended. The main result in [16] is the following

Theorem 1. *Assume that ω satisfies* (6) *and that V is analytic in the complex strip $\mid Im\,\theta \mid < r$. Then there exists a constant $\varepsilon_0 = \varepsilon_0(r,\kappa,\tau)$ such that if*

$$\sup_{|\mathrm{Im}\,\theta|<r} \mid V(\theta) \mid \; < \; \varepsilon_0$$

then (4) *is reducible for a.e. E and all θ, i.e. the fundamental solution can be written*

$$A_n(\theta,E) \;=\; C(\theta+n\omega,E)e^{nB(E)}C^{-1}(\theta,E),$$

with $C(\cdot,E) : (2\mathbb{T})^d \to SL(2,\mathbb{R})$ analytic and $B(E) \in sl(2,\mathbb{R})$. The set of admissible E's depends on the potential V.

The theorem was proven in the time-continuous case but the proof carries over easily to the discrete case. There is probably a corresponding result for Gevrey classes but if it holds also in $\mathcal{C}^\infty$-category or in finite differentiability is unclear.

There is also a result in [16] stating that the (possible) non-reducible systems in this one-parameter family must have Lyapunov exponents = 0, and that for generic potentials not *all* systems in the family are reducible. This non-reducibility is shown by constructing solutions that are unbounded but increases more slowly than linearly. So even near to constant coefficients there is some delicate mixture of reducible and non-reducible systems. Best known is this mixture in the compact case $SO(3,\mathbb{R})$.

4. Compact Case.

Let's first observe that a matrix in $SO(3,\mathbb{R})$ has three eigenvalues $e^{\pm i\alpha}, 0$ for some real number α. Hence, if A is constant then $\mathbb{T}^d \times SO(3,\mathbb{R})$ is foliated into invariant tori of dimension $d+1$ and the the orbits of (1) are dense on these tori if and only if α is irrational.

Assume that $A_0 \in SO(3,\mathbb{R})$ and that $A : \mathbb{T}^d \to SO(3,\mathbb{R})$ is analytic in $\mid \mathrm{Im}\,\theta \mid < r$. The following result is due to R. Krikorian [18].

THEOREM 2. *If ω satisfies* (6) *and if $A_0 \neq 0$, then there exists an $\varepsilon_1 = \varepsilon_1(\kappa, \tau, r, A_0)$ such that if*

$$\sup_{|\operatorname{Im}\theta|<r} |\, \hat{A}(\theta)\, | < \varepsilon_0,$$

then the skew-product (2) *with $A(\theta) = \lambda A_0 + \hat{A}(\theta)$ is analytically reducible for a.e. λ.*

Though this is not exactly the statement of the conjecture since it only refers to a particular one-parameter family it is pretty close. It justifies therefore that we think of the reducible systems as being of "full measure" close to constant coefficients. Hence, in a measure sense the typical system has an invariant foliation into $d+1$-dimensional tori. A natural question is: what can one say about the complementary set?

On the compact manifold $\mathbb{T}^d \times SO(3,\mathbb{R})$ the dynamical system (2) preserves the product Haar measure $\mu \times \nu$. In the reducible case this measure is certainly not ergodic and there are invariant measures supported on each invariant torus. However, for the topologically generic system there is no trace of any invariant set whatsoever, not even of measurable invariant sets. This is the content of the next theorem [19]

THEOREM 3. *There exists an $\varepsilon_0 = \varepsilon_0(\kappa, \tau, r, A_0)$ such that for the generic $\hat{A}(\theta)$ in*

$$\sup_{|\operatorname{Im}\theta|<r} |\, \hat{A}(\theta)\, | < \varepsilon_0$$

the skew-product (2) *with $A(\theta) = A_0 + \hat{A}(\theta)$ is uniquely ergodic.*

Together these two theorems give a very nice picture of the behavior of analytic systems close to constant coefficients: the reducible and uniquely ergodic systems are mixed like the Diophantine and the Liouville numbers. Is it possible that this is also the global situation? A *strong version* of this question is if it is possible to conjugate any system to a close-to-constant-coefficient system in an analytical (or possibly a weaker) topology. A *weak version* would be if any system can be approximated by a reducible system in an analytical (or possibly a weaker) topology.

The weak version is completely open, and the strong version is doubtful because of an example by M. Rychlik. The example is a skew-product on $\mathbb{T} \times SU(2,\mathbb{C})$ [20], which is even a time-one map of a $\mathcal{C}^\infty$-system, given by the matrix

$$A(\theta) \;=\; \begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}.$$

With this matrix the system (1) is not reducible, and it seems unlikely that it can be conjugated close to constant coefficient. Notice however that the system, though not reducible, has an invariant foliation into 2-dimensional tori on which the orbits are dense.

Theorem 2 also holds for general compact matrix groups [21]. A weaker result than Theorem 3 was proven for $SU(2,\mathbb{C})$ in [22].

5. Far from Constant Coefficients.

If little is known for systems in $\mathbb{T}^d \times SO(3,\mathbb{R})$, or in other compact groups, unless they are close to constant coefficients, we have additional information in the non-compact case $\mathbb{T} \times SL(2,\mathbb{R})$. The subharmonic argument in [2,11] gives a class of discrete Schrödinger equations (4) with large potential V which has non-zero Lyapunov exponents for all E, and therefore must be non-uniformly hyperbolic at least for some E. But more is true. The discrete Schrödinger equation on $\mathbb{T}$ with Diophantine frequencies often has a complete set of eigenvectors in $l^2(\mathbb{Z})$ if the potential is large enough. We shall give a precise formulation of this result.

Assume that V satisfies the transversality conditions

$$(10) \quad \begin{cases} \max\limits_{0\le\nu\le s} \mid \partial_x^\nu (V(\theta+x)-V(\theta)) \mid \ge \ \xi \ > \ 0, & \forall\theta \\ \max\limits_{0\le\nu\le s} \mid \partial_\theta^\nu (V(\theta+x)-V(\theta)) \mid \ge \ \xi \inf_{m\in\mathbb{Z}} \mid x-2\pi m \mid, & \forall\theta\ \forall x. \end{cases}$$

These two conditions can be understood as requiring that the potential is always "different from a constant" and always "different from itself under translations". They are fulfilled, for appropriate values of s and ξ, for any analytic function defined in θ with no shorter period than 2π.

Theorem 4. *Assume that ω satisfies* (6) *and that V is analytic and bounded by a constant γ in the complex strip $\mid Im\,\theta \mid < r$ and satisfies condition* (10). *Then there exists a constant $\varepsilon_0 = \varepsilon_0(\gamma, r, \kappa, \tau, s, \xi)$ such that if $\mid \varepsilon \mid < \varepsilon_0$*

$$(11) \qquad -\varepsilon(u_{n+1}+u_{n-1}) + V(\theta+n\omega)u_n \ = \ Eu_n$$

has a complete set of eigenvectors in $l^2(\mathbb{Z})$ for almost all θ.

The closure of the set of eigenvalues — the spectrum — is a set whose complement in $[\inf V, \sup V]$ has measure that goes to 0 with ε.

It follows then from a theorem of Kotani [23, Proposition VII.3.3] that (11) must have non-zero Lyapunov exponents for almost all E and hence must be non-uniformly hyperbolic for almost all parameter values E in the spectrum.

These systems can not be conjugated close to constant coefficients in an analytic topology because close to constants there are no non-uniformly hyperbolic systems by Theorem 1. The question if they can be approximated by reducible systems in an analytic (or possibly weaker) topology is however still open.

Theorem 3 was first proven for a "cosine"-like potential in [24,25], and in [26] for a more general difference operator. The general version given here is from [27]. It holds not only for analytic potentials but also for smooth ones belonging to some Gevrey class. It also holds if one replaces the nearest neighbor operator $(u_{n+1}+u_{n-1})$ by a symmetric operator

$$\sum_{-\infty<k<\infty} a_k(\theta+n\omega)u_{n+k}$$

where the a_k's are assumed to be analytic in $\mid \mathrm{Im}\,\theta \mid < r$ (or more generally of some Gevrey class) and decay exponentially in k.

The situation with a higher-dimensional torus is much more complex. The case $\mathbb{T}^2$ has been analyzed in [28] but complete proofs has not been published.

It is not known what other kinds of dynamics can occur for a skew-system on $\mathbb{T} \times SL(2, \mathbb{R})$ than the types described here.

The proofs of Theorem 1-3 have been obtained by ODE-methods applied to a particular skew-system (1) or (2) in order to conjugate, or try to conjugate, it to constant coefficients. The proof of Theorem 4 is different. Here one considers the Schrödinger equation (3) not as a dynamical system but as an operator on $l^2(\mathbb{Z})$ which one tries to conjugate to diagonal form. Both types of results are therefore conjugation results involving small divisor problems, but one is in finite-dimensional dynamical space and the other is in an infinite-dimensional space.

References

1. V.V. Nemytskii, V.V. Stepanov, *Qualitative theory of differential equations*, Princeton University Press, Princeton, New Jersey, 1960.
2. M. Herman, *Une méthode pour minorer les exposants de Lyapunov*, Comment. Math. Helvetici **58** (1983), 453-502.
3. R. A. Johnsson, G. R. Sell, *Smoothness of spectral subbundles and reducibility of quasi-periodic linear differential systems*, J. Diff. Eq. **41** (1981), 262-288.
4. V.M. Millionščikov, *Proof of the existence of irregular systems of linear differential with almost periodic coefficients*, Diff. Equations **4** (1968), 203-205.
5. R. A. Johnson, *The recurrent Hill's equation*, J. Diff. Equations **46** (1982), 165-193.
6. A. Gordon, *On the point spectrum of the one-dimensional Schrödinger operator*, Russ. Math. Surveys **31** (1976), 457-258.
7. R. A. Johnson, J. Moser, *The rotation number for almost periodic potentials*, Commun. Math. Phys. **84** (1982), 403-438.
8. E. I. Dinaburg, Ya. G. Sinai, *The one-dimensional Schrödinger equation with quasi-periodic potential*, Funkt. Anal. i. Priloz. **9** (1975), 8-21.
9. B. Simon, *Almost periodic Schrödinger operators: a review*, Adv. Appl. Math. **3** (1982), 463-490.
10. M. Nerurkar, *On the construction of smooth ergodic skew-products*, Ergod. Th. & Dynam. Sys. **8** (1988), 311-326.
11. E. Sorets, T. Spencer, *Positive Lyapunov exponents for Schrödinger operators with quasi-periodic potentials*, Commun. Math. Phys. **142** (1991), 543-566.
12. L. S. Young, *Lyapunov exponents for some quasi-periodic cocycles*, Ergod. Th. & Dynam. Sys. **17** (1997), 483-504.
13. J. Moser, *Convergent series expansions for quasiperiodic motions*, Math. Ann. **169** (1967), 136-176.
14. L. H. Eliasson, *Perturbations of stable invariant tori for Hamiltonian systems*, Ann. Sc. Norm. Sup. Pisa Cl. Sci. **15** (1988), 115-147.
15. A. Jorba, C Simo, *On the reducibility of linear differential equations with quasi-periodic coefficients*, J. Diff. Eq. **98** (1992), 111-124.
16. L. H. Eliasson, *Floquet solutions for the one-dimensional quasi-periodic Schrödinger equation*, Commun. Math. Phys. **146** (1992), 447-482.

17. J.Moser, J. Pöschel, *An extension of a result by Dinaburg and Sinai on quasi-periodic potentials*, Comment. Math. Helvetici **59** (1984), 39-85.
18. R. Krikorian, *Réductibilité presque partout des systèmes quasi-périodiques dans le cas* $SO(3,R)$, C. R. Acad. de Paris, Série I **321** (1995), 1039-1044.
19. L. H. Eliasson, *Ergodic skew systems on* $SO(3,\mathbb{R})$, preprint ETH-Zürich (1991).
20. M. Rychlik, *Renormalization of cocycles and linear ODE with almost periodic coefficients*, Invent. Math. **110** (1992), 173-206.
21. R. Krikorian, *Reductibilité presque partout des flots fibrés quasi-périodiques á valeurs dans des groupes compacts*, Ann. Sci. École Norm. Sup. (to appear).
22. M. R. Herman, *Non topological conjugacy of skew products on* $SU(2)$, manuscript (1989).
23. R. Carmona, J. Lacroix, *Spectral Theory of Random Schrödinger Operators*, Birkhäuser, Boston, 1990.
24. Ya. G. Sinai, *Anderson localization for the one-dimensional difference Schrödinger operator with a quasi-periodic potential*, J. Stat. Phys. **46** (1987), 861-909.
25. J. Fröhlich, T. Spencer, P. Wittver, *Loclization for a class of one-dimensional quasi-periodic Schrödinger operators*, Commun. Math. Phys. **132** (1990), 5-25.
26. V. A. Chulaevsky, E. I. Dinaburg, *Methods of KAM-Theory for Long-Range Quasi-Periodic Operators on* $\mathbb{Z}^\nu$. *Pure Point Spectrum*, Commun. Math. Phys. **153** (1993), 559-577.
27. L. H. Eliasson, *Discrete one-dimensional quasi-periodic Schrödinger operators with pure point spectrum*, Acta Math. **179** (1997), 153-196.
28. V. A. Chualevsky, Ya. G. Sinai, *Anderson localization and KAM-theory*, P. Rabinowitz, E. Zehnder (eds.): Analysis etcetera, Academic Press, New York, 1989.

L. H. Eliasson
Royal Institute of Technology
S-10044 Stockholm
Sweden
hakane@math.kth.se

Hyperbolicity, Stability, and the Creation of Homoclinic Points

Shuhei Hayashi

1 The connecting lemma

The importance of connecting invariant manifolds by small perturbations of dynamical systems has been realized through the solution of the C^1 Stability and Ω-Stability Conjectures for diffeomorphisms, respectively by Mañé ([M3]) and Palis ([P2]). Moreover, the extension of their results was done through the creation of homoclinic points by C^1 small perturbation ([H1]).

Let M be a compact manifold without boundary and let $\mathrm{Diff}^1(M)$, resp. $\mathcal{X}^1(M)$, be the set of C^1 diffeomorphisms of M, resp. vector fields on M, with the C^1 topology. Denote by X_t, $t \in \mathrm{R}$, the C^1 flow on M generated by $X \in \mathcal{X}^1(M)$. A set $\Lambda \subset M$ is *hyperbolic* for $f \in \mathrm{Diff}^1(M)$, resp. $X \in \mathcal{X}^1(M)$, if it is compact, invariant, i.e., $f(\Lambda) = \Lambda$, resp. $X_t(\Lambda) = \Lambda$ for all $t \in \mathrm{R}$, there exists a continuous splitting $TM|\Lambda = E^s \oplus E^u$, resp. $TM|\Lambda = E^0 \oplus E^s \oplus E^u$ that is invariant under Df, resp. DX_t for all $t \in \mathrm{R}$, and there exist constants $K > 0$, $0 < \lambda < 1$ such that

$$||(Df^n)|E^s(x)|| \leq K\lambda^n, \quad n \geq 0$$

and

$$||(Df^{-n})|E^u(x)|| \leq K\lambda^n, \quad n \geq 0,$$

resp.

$$E^0(x) = \mathrm{R} \cdot X(x)$$
$$||(DX_t)|E^s(x)|| \leq K\lambda^t, \quad t \geq 0$$

and

$$||(DX_{-t})|E^u(x)|| \leq K\lambda^t, \quad t \geq 0$$

for all $x \in \Lambda$. In particular, Λ is called *isolated* if there exists a compact neighborhood U of Λ such that

$$\bigcap_{n \in \mathbf{Z}} f^n(U) = \Lambda,$$

resp.

$$\bigcap_{t \in \mathbf{R}} X_t(U) = \Lambda.$$

We say that p is a *homoclinic point* associated to Λ if

$$p \in W^s(\Lambda) \cap W^u(\Lambda) - \Lambda.$$

THEOREM 1.1 [H2]. If a C^1 dynamical system has an almost homoclinic sequence associated to an isolated hyperbolic set Λ, then there exists a dynamical system C^1 arbitrarily close to the original one, coinciding with it in a neighborhood of Λ, and having a homoclinic point associated to Λ.

For the definition of almost homoclinic sequences, see [H2]. For instance, a sequence of periodic orbits outside Λ accumulating on Λ gives an almost homoclinic sequence associated to Λ. In the proof of Theorem 1.1, we only use Pugh's perturbation technique in his Closing Lemma and don't need the hyperbolicity of Λ. So, we can apply the perturbation in Theorem 1.1 to more general situation. In particular, we get Theorem 1.2 below.
For $X \in \mathcal{X}^1(M)$ and a point $x \in M$, the *ω-limit set* of x, $\omega(x)$ is defined by $\omega(x) = \{y \in M | \exists t_i \to +\infty \text{ such that } X_{t_i}(x) \to y\}$; the *$\alpha$-limit set* of x, $\alpha(x)$ is defined similarly with $t_i \to -\infty$ instead of $t_i \to +\infty$.

THEOREM 1.2 [H3]. Let $\mathcal{U}$ be a neighborhood of $X \in \mathcal{X}^1(M)$ and $p, q \in M$ with $q \in \omega(p) - \omega(q)$ be given. Then, there exists $Y \in \mathcal{U}$ coinciding with X outside an arbitrarily small neighborhood of $\{X_t(q) | -T \le t \le 0\}$ for some $T(\mathcal{U}, q, X) > 0$ and having an orbit including p and $\{X_t(q) | t \ge 0\}$.

There still remains the other type of connecting problem even for the C^1 case.

PROBLEM. For p and q respectively belonging to the unstable and stable manifolds of a hyperbolic singularity (or periodic orbit), if $\omega(p)$ meets $\alpha(q)$, then is it possible to have a homoclinic point associated to it by a C^1 small perturbation?

This problem is mentioned in [PM] and [Pu]. Pugh gave an example in [Pu] showing that it is not always possible even for a C^1 vector field when the ambient manifold is not compact. Theorem 1.3 below is not the complete solution of the problem when the manifold is compact, but, using it together with Theorem 1.2, we get a C^1 Make or Break Lemma (Theorem 1.4), which gives an affirmative answer to a question suggested by Mañé. Denote by $\mathrm{Sing}(X)$ and $\mathrm{Per}(X)$ respectively the set of singularities of X and that of periodic points of X.

THEOREM 1.3 [H3]. Let $\mathcal{U}$ be a neighborhood of $X \in \mathcal{X}^1(M)$ and $p, q \in M$ with $\omega(p) \cap \alpha(q) - (\mathrm{Sing}(X) \cup \mathrm{Per}(X)) \neq \emptyset$ be given. Then, for each $\tilde{p} \in \omega(p) \cap \alpha(q) - (\mathrm{Sing}(X) \cup \mathrm{Per}(X))$, there exists $Y \in \mathcal{U}$ coinciding with X outside an arbitrarily small neighborhood of $\{X_t(\tilde{p}) | 0 \le t \le T\}$ for some $T(\mathcal{U}, \tilde{p}, X) > 0$ and having an orbit including p and q.

THEOREM 1.4 [H3]. Given $p, q \in M$ with $\omega(p) \cap \alpha(q) \neq \emptyset$, there exists a vector field Y C^1 close to $X \in \mathcal{X}^1(M)$ such that either (a) Y has an orbit including p and q, or (b) $\omega(p) \cap \alpha(q) = \emptyset$.

2 The Stability Conjecture The concept of (structural) stability goes back to the work of Andronov and Pontryagin [AP] in 1937. They considered the necessary and sufficient conditions for vector fields on a two-dimensional disk to be structurally stable. Here the *structural stability* deals with the topological persistence under small perturbations of the orbit structure of a dynamical system, which is expressed by a homeomorphism of the ambient manifold sending orbits of the initial one onto orbits of the perturbed system preserving their orientations.
In the late fifties, Peixoto extended this characterization to closed orientable surfaces and subsequently proved the density of stable two-dimensional flows. At this point, Smale thought that perhaps such kind of result could be true in any dimension for both diffeomorphisms and flows. To that end, he formulated what is now called a Morse-Smale system: the limit set consists of finitely many fixed or periodic hyperbolic orbits with their stable and unstable manifolds being transverse. And he conjectured that (a) they are (structurally) stable and (b) they are dense among all C^r dynamical systems, $r \geq 1$. Part (b) of the conjecture was soon shown not to be true, due to the existence (and persistence) of transversal homoclinic orbits. Remarkably, Smale responded by discovering that a transversal homoclinic orbit implies the existence of a new prototype of dynamics: the horseshoe transformation, whose limit set consists of a Cantor set in which the (infinitely many) periodic orbits are dense. In the mid sixties, motivated by Smale's questions, Anosov proved that globally hyperbolic systems are stable. Soon afterwards, Palis and Smale proved that the Morse-Smale systems are stable, so (at least) part (a) of Smale's initial conjecture is correct. Their methods were quite distinct from those of Anosov, since for the Morse-Smale systems there are several hyperbolic "pieces" (fixed or periodic orbits), with stable and unstable manifolds of different dimensions. At this point, putting together their result and that of Anosov, Palis and Smale formulated in 1967 the celebrated Stability Conjecture: a system is stable if its limit set is hyperbolic and all the stable and unstable manifolds are transversal. Instead of hyperbolicity of the limit set, we can require the nonwandering set to be hyperbolic and the periodic orbits to be dense on it (*Axiom A* or *hyperbolic* systems).
In the beginning of 1970's, Robbin, de Melo, and Robinson proved that the properties of Axiom A together with the transversality condition between stable and unstable manifolds (the strong transversality condition) is sufficient for the structural stability.

Theorem 2.1 (Robbin [R], de Melo [dM], Robinson [Ro]). C^1 dynamical systems satisfying Axiom A and the strong transversality condition are C^1 structurally stable.

In fact, a celebrated conjecture, the so-called Stability Conjecture had been raised by Palis and Smale [PSm] in the late 1960's, saying that the condition is the necessary and sufficient conditions for structural stability. The sufficient condition was proved relatively early, but it took much longer time to solve the converse. After contributions by many mathematicians, it was finally solved for

the C^1 case by Mañé for diffeomorphisms and by Hayashi for flows. (See also Wen [W].)

THEOREM 2.2 (Mañé [M3], Hayashi [H2]). C^1 structurally stable dynamical systems satisfy Axiom A and the strong transversality condition.

The Ω-stability is a stability restricted to the nonwandering set (so it is weaker than the structural stability), and there is a similar conjecture, the so-called Ω-Stability Conjecture. See Palis [P2] and Hayashi [H2] for the proof. Thus, Palis-Smale's conjecture characterizing stable dynamical systems has completed (for the C^1 case). As a consequence, it turns out that the two concepts, hyperbolicity and stability are essentially equivalent to each other for C^1 dynamical systems of a compact manifold.

As to Theorem 2.2 for flows, the biggest difficulty is the existence of singularities (which are all hyperbolic by stability). In fact, if a sequence of periodic points is accumulating on a singularity, similar arguments to the diffeomorphism case cannot be applied, and if singularities are separated from periodic points, taking the time-one map, parallel arguments to those of diffeomorphism case are available in principle. The separation, the crucial step in the proof of Theorem 2.2 for flows, is obtained by Theorem 1.1. In fact, a sequence of periodic points accumulating on a (hyperbolic) singularity gives an almost homoclinic sequence, so Theorem 1.1 implies that a homoclinic point associated to the singularity can be created by a C^1 small perturbation, which belongs to an unstable saddle connection and contradicts the stability of the vector field. Thus we get the following: for C^1 stable vector fields, singularities are not in the closure of the set of periodic points. In other words, singularities are isolated in the nonwandering set.
However, there is still an essential difference between diffeomorphisms and flows; that is, even though the set of periodic points with the same index (the dimension of the stable subspace) is hyperbolic (then it can be decomposed into disjoint finite union of isolated hyperbolic sets each of which has a dense orbit), a periodic point with the same index might appear far from the original hyperbolic set by arbitrarily small perturbation, which never occurs in stable diffeomorphisms. This phenomenon cause a difficulty in proving the hyperbolicity, but we can take a dense subset in the set of stable vector fields in which the phenomenon never occurs. So, by a similar method to the diffeomorphism case, hyperbolicity of each vector field in the dense subset is obtained. After that, every stable vector field is finally proved to be hyperbolic.

3 BEYOND HYPERBOLICITY

As mentioned in Section 2, Smale expected that the stable systems would be dense in the set of all dynamical systems. This "dream" has collapsed through many examples: there are open sets of unstable or even Ω-unstable systems (see [PT]). Still, one can ask:

Conjecture 3.1 (Palis). The set of Morse-Smale dynamical systems together with the systems having a transversal homoclinic point forms a dense subset in the space of dynamical systems.

Since Axiom A systems which are not Morse-Smale ones have transversal homoclinic points, this conjecture can be also considered as a step toward another conjecture by Palis [PT]:

Conjecture 3.2 (Palis). Every diffeomorphism on a compact manifold can be approximated by a diffeomorphism satisfying Axiom A or else by one exhibiting a bifurcation involving the creation of homoclinic points (homoclinic bifurcation).

This is in the direction of Palis' program aiming at the global understanding of dynamical systems in the complement of the closure of the set of hyperbolic (or stable) ones. One of his main conjectures is to ask if densely one has finitely many attractors with Sinai-Ruelle-Bowen invariant measures and whose basins cover Lebesgue almost all points in the ambient manifold. Moreover, the attractors should be stochastically stable (see [P3]). In a probabilistic way that would rescue the lost dream of the sixties mentioned in the beginning of this section. Another way is to find a dense subset in the complement having a dynamical feature and find out some mechanism to investigate the bifurcations around each element of the dense subset. It is known that homoclinic tangencies (nontransversal intersection of a stable and an unstable manifold of the same hyperbolic fixed point or periodic point) yields rich phenomena of nonhyperbolic dynamics, such as infinitely many sinks and strange attractors. See [P3] and [PT] for more on this program.

For the two-dimensional case, Conjecture 3.2 was solved recently by Pujals and Sambarino [PS]; that is, every C^1 surface diffeomorphism can be approximated by Axiom A diffeomorphism or else by one exhibiting a homoclinic tangency . In higher dimensions, there are examples of open sets of nonhyperbolic diffeomorphisms where elements exhibit no homoclinic tangencies. In fact, Diaz [D] constructed examples which is obtained after unfolding a three-dimensional cycles with two hyperbolic fixed points p and q of saddle type having different indices containing a transversal intersection of $W^u(p)$ and $W^s(q)$ and a nontransversal orbit of intersection between $W^u(q)$ and $W^s(p)$.

As a first step to have the hyperbolicity in the complement of the closure of the set of diffeomorphisms exhibiting a homoclinic tangency, it is natural to consider showing the existence of dominated splitting on the supports of ergodic probability measures. Here a *dominated splitting* on a compact invariant set Λ is a continuous, f-invariant (i.e., invariant under the derivative of f) splitting

$$TM|\Lambda = E \oplus F$$

such that there exist constants $m \in \mathrm{Z}^+$, $0 < \lambda < 1$ satisfying

$$||(Df^m)|E(x)|| \cdot ||(Df^{-m})|F(f^m(x))|| < \lambda.$$

We know that there exist the Lyapunov splittings in a dense subsets of the supports of ergodic measures by Oseledec's theorem. Let us recall the Oseledec's theorem. Denote by $\Lambda(f)$ for $f \in \mathrm{Diff}^1(M)$ the set of points satisfying the following properties: there exists a splitting $T_xM = \bigoplus_{i=1}^m E_i(x)$ (the Lyapunov splitting at x) and numbers $\lambda_1(x) > \cdots > \lambda_m(x)$ (the Lyapunov exponents at x) such that $\lim_{n\to\pm\infty} \frac{1}{n} \log ||(D_x f^n)v|| = \lambda_i(x)$ for every $1 \le i \le m$ and $0 \ne v \in E_i(x)$. Oseledec proved that $\mu(\Lambda(f)) = 1$ for every f-invariant probability measure μ on the Borel σ-algebra of M. Here $E_i(x)$ $(1 \le i \le m)$ are just measurable functions of x. In this direction, there is a theorem by Mañé [M2], saying that for C^1 generic f (elements in a residual subset of $\mathrm{Diff}^1(M)$), there is a residual subset $\mathcal{R}$ in the space of ergodic measures $\mathcal{M}_e(f)$ of f such that each $\mu \in \mathcal{R}$ has a dominated splitting on its support $s(\mu)$ coinciding with the Lyapunov splitting at μ-a.e. point of $s(\mu)$. As Mañé mentioned in [M2], generic elements of $\mathcal{M}_e(f)$ fail to reflect the dynamical complexity of f in the sense that C^1 generically, the entropy $h_\mu(f)$ is zero for generic μ. For instance, $h_\mu(f) = 0$ when μ is supported on a single periodic orbit. So he suggested to work in the space $\mathcal{M}_e^c(f) = \{\mu \in \mathcal{M}_e(f) | h_\mu(f) > c\}$ and prove that generic measures in $\mathcal{M}_e^c(f)$ satisfy a strong form of Oseledec's theorem.
The following result is in the direction of the combination of proposals of Mañé and Palis. Let $\mathcal{H}^1(M)$ be the set of C^1 diffeomorphisms exhibiting a homoclinic tangency.

THEOREM 3.3 [H5]. There is a dense subset $\mathcal{D}$ in the complement of $\overline{\mathcal{H}^1(M)}$ such that if $f \in \mathcal{D}$, for every $\mu \in \mathcal{M}_e(f)$ which is not supported on a single periodic orbit, either $\lim_{n\to\pm\infty} \frac{1}{n} \log ||(D_x f^n)v|| = 0$ for μ-a.e. x and every $0 \ne v \in T_xM$ or there exist dominated splittings

$$TM|s(\mu) = E^- \oplus F,$$

$$TM|s(\mu) = E \oplus F^+$$

such that

$$E^-(x) = \bigoplus_{\lambda_j(x)<0} E_j(x), \quad F = \bigoplus_{\lambda_j(x)\ge 0} F_j(x),$$

$$E(x) = \bigoplus_{\lambda_j(x)\le 0} E_j(x), \quad F^+ = \bigoplus_{\lambda_j(x)>0} F_j(x)$$

at μ-a.e. x of $s(\mu)$.

For the proof, we need an improved version of Mañé's ergodic closing lemma ([M1]) and a theorem in Pesin theory. Using this theorem and Theorem 1.1, we get the following result toward the solution of Conjecture 3.1.

Theorem 3.4 [H5]. In the complement of the closure of the set of diffeomorphisms exhibiting a transversal homoclinic point together with the Morse-Smale ones, every diffeomorphism can be C^1 approximated by one having an ergodic measure whose support has dominated splittings as in Theorem 3.3.

Acknowledgment: The author would like to thank Professor J. Palis for his many suggestions on this paper.

References

[A] D. Anosov, *Geodesic flows on closed Riemannian manifolds of negative curvature*, Proc. Steklov Inst. 90 (1967) (Amer. Math. Soc. Transl. 1969).

[AP] A.A. Andronov and L. Pontryagin, *Systems Grossiers*, Dokl. Akad. Nauk. USSR 14 (1937), 247–251.

[dM] W. de Melo, *Structural stability of diffeomorphisms on two-manifolds*, Inventiones Math. 21 (1973), 233–246.

[D] L. Diaz, *Robust nonhyperbolic dynamics and heterodimensional cycles*, Ergod. Th. and Dynam. Sys. 15 (1995), 291–315.

[H1] S. Hayashi, *Diffeomorphisms in $\mathcal{F}^1(M)$ satisfy Axiom A*, Ergod. Th. and Dynam. Sys. 12 (1992), 233–253.

[H2] ——, *Connecting invariant manifolds and the solution of the C^1 stability and Ω-stability conjectures for flows*, Ann. of Math. 145 (1997), 81-137.

[H3] ——, *A C^1 Make or Break Lemma*, preprint.

[H4] ——, *On non-singular vector fields in $\mathcal{G}^1(M)$*, in preparation.

[H5] ——, in preparation.

[M1] R. Mañé, *An ergodic closing lemma*, Ann. of Math. 116 (1982), 503–540.

[M2] ——, *Oseledec's Theorem from the Generic Viewpoint*, in the Proceeding of the International Congress of Mathematicians, Warszawa 1983, 1269–1276.

[M3] ——, *A proof of the C^1 Stability Conjecture*, Publ. Math. IHES 66 (1988), 161–210.

[P1] J. Palis, *On the Contribution of Smale to Dynamical Systems*, From Topology to Computation: Proceedings of the Smalefest (M. Hirsch, J. Marsden and M. Shub, eds.), Springer-Verlag, 1993, pp. 165–178.

[P2] ——, *On the C^1 Ω-Stability Conjecture*, Publ. Math. IHES 66 (1988), 211–215.

[P3] ——, *A global view of dynamics and a conjecture on the denseness of finitude of attractors*, to appear in Astérisque.

[PM] J. Palis and W. de Melo, *Geometric theory of dynamical systems*, Springer-Verlag, 1982.

[PSm] J. Palis and S. Smale, *Structural stability theorems.* Global Analysis, Proc. Sympos. Pure Math. vol. 14, Amer. Math. Soc., 1970, 223–231.

[PT] J. Palis and F. Takens, *Hyperbolicity and Sensitive Chaotic Dynamics at Homoclinic Bifurcations*, Cambridge University Press, Cambridge, New York. Springer-Verlag, 1987.

[Pe] M. Peixoto, *Structural stability on two dimensional manifolds*, Topology 1 (1962), 101–120.

[Pu] C. Pugh, *The C^1 Connecting Lemma*, J. Dyn. and Diff. Eq. 4 (1992), 545–553.

[PR] C. Pugh and C. Robinson, *The C^1 Closing Lemma, including Hamiltonians*, Ergod. Th. and Dynam. Sys. 3 (1983), 261–313.

[PS] E. Pujals and M. Sambarino, *Homoclinic tangencies and hyperbolicity for surfaces diffeomorphisms: a conjecture of Palis*, preprint.

[R] J. Robbin, *A structural stability theorem*, Ann. of Math. 94 (1971), 447–493.

[Ro] C. Robinson, *Structural stability of vector fields*, Ann. of Math. 99 (1974), 154–175.

[W] L. Wen, *On the C^1 stability conjecture for flows*, J. Diff. Eq. 129 (1996), 334–357.

Shuhei Hayashi
School of Commerce,
Waseda University
Shinjuku, Tokyo 169-50,
Japan

Some Open Problems in Dynamical Systems

Michael Herman[1]

"Ce que nous savons est peu de choses ; ce que nous ignorons est immense".
P. S. Laplace

1 Siegel singular disks.

Let $P_\alpha(z) = e^{2i\pi\alpha}z + z^2$, $\alpha \in \mathbb{R} - \mathbb{Q}$, α is a Bruno number and we denote by S_α the maximal connected open set containing 0 on which P_α is linearizable (see [Y$_2$]) (the Siegel singular disk of P_α ; we use the word singular to avoid any confusion with Siegel domains).

Questions.

1. *Does there exist α_1 a Bruno number such that ∂S_{α_1} (the boundary of S_{α_1}) is a C^∞ submanifold of $\mathbb{C}$ (a circle)* ? (see R. Pérez-Marco) [P]).

2. *Does there exist α_2 a Bruno number such that ∂S_{α_2} (the boundary of S_{α_2}) is not a Jordan curve* ? (see A. Douady) [D]).

I conjecture that the answers are positive, but for question 2 one might need to consider general polynomials.

2 Invariant tori.

We suppose $\alpha \in \mathbb{T}^n$ satisfies a diophantine condition (and write $\alpha \in DC$).

$$\exists\, \gamma > 0\,,\ \exists \beta \geq 0\,,\ \forall\, k \in \mathbb{Z}^n - \{0\}\ ,\ |||\langle k, \alpha\rangle||| \geq \frac{\gamma}{\|k\|^{n+\beta}}\ ,$$

where $\langle k, \alpha\rangle = \sum_j k_j\alpha_j$, $|||\ |||$ denotes the distance to nearest integer and $\|k\| = \sum_j |k_j|$.

[1] C.N.R.S.

Let $C_0^\infty(\mathbb{T}^n)$ the space of C^∞ functions on $\mathbb{T}^n$ of Haar integral 0. To each $\varphi \in C_0^\infty(\mathbb{T}^n)$ we associate the C^∞ exact symplectic diffeomorphisms on $T^*\mathbb{T}^n \cong \mathbb{T}^n \times \mathbb{R}^n$ with coordinates $(\theta, r) \in \mathbb{T}^n \times \mathbb{R}^n$,

$$F_\varphi(\theta, r) = \left(\theta + r, r + \frac{\partial \varphi}{\partial \theta}(\theta + r)\right), \frac{\partial}{\partial \theta} = \left(\frac{\partial}{\partial \theta_1}, \ldots, \frac{\partial}{\partial \theta_n}\right)$$

(if $\varphi \in C^k$, F is C^{k-1}).

For $\alpha \in DC$ let $U_\alpha = \{\varphi \in C_0^\infty(\mathbb{T}^n),\ F_\varphi$ leaves invariant a torus $T_{\alpha,\varphi}$ graph of $\psi \in C^\infty(\mathbb{T}^n, \mathbb{R}^n)$ on which F_φ is C^∞-conjugated to $R_\alpha : \theta \to \theta + \alpha\}$. Using KAM, $0 \in U_\alpha$ and U_α is C^∞ open (cf. [H_8]). The open set is not all $C_0^\infty(\mathbb{T}^n)$ since (cf. [H_3] and [H_4]) if $\varphi \in U_\alpha$ the symmetric matrix $I + 1/2D^2\varphi(\theta)$ must be strictly positive definite for all $\theta \in \mathbb{T}^n$ (I denotes the unit $n \times n$ matrix) : if one remplaces φ by $\lambda\varphi$, supposing that $\varphi \not\equiv 0$ (is not identically zero), the above condition will be violated when λ is large.

2.1 - QUESTION.

Is the boundary ∂U_α of U_α in $C^\infty(\mathbb{T}^n)$ a locally flat C^k-submanifold (for some $k \geq 0$) ?
Nothing is known (e.g. is every point $x \in \partial U_\alpha$ accessible from U_α and the exterior of U_α ?).

2.2 - *Let $\bigcup_{\alpha \in DC} U_\alpha = U$; what is the minimal number $k \in \mathbb{R}_+^*$ for which U is open near 0, in the C^k-topology* ?
(One shows, [H_3], that $k \geq n+2$). For many reasons it is not unnatural to work in Sobolev $W^{k,p}$ topology, $k \in \mathbb{R}_+^*$, $p \geq 1$: the topology defined by asking that φ has [k] derivatives in the distributional sense in $L^p(\mathbb{T}^n, d\theta)$ and $t \in \mathbb{T}^n \mapsto D^{[k]}\varphi \circ R_t \in L^p$ satisfies a Hölder condition of exponent $k - [k]$.
One easily shows, [H_3], that U is not open (near 0) in the $W^{2n+2-\varepsilon,1}$ topology ($\forall\, \varepsilon > 0$), and the author showed, [H_2], that,
when $n = 1$ and α is of constant type, $\left(\exists\, \gamma > 0,\ \forall\, p/q,\ |\alpha - p/q| \geq \frac{\gamma}{q^2}\right)$, U_α is open near 0, in the $W^{4,p}$, $p > 1$, topology.

The author has been unable to decide if this is also true when $p = 1$!

One proves, using [S], that U_α is open, near 0, in the C^k-topology when $k \geq 2(n + \beta) + 2 + \varepsilon$, $\forall\, \varepsilon > 0$, and is not open in the C^k-topology, near 0, when $k \leq 2(n + \beta) + 2 - \varepsilon$, $\forall\, \varepsilon > 0$ (when $n = 1$, this is proved in [H_1] and if $n \geq 2$, it is really not difficult to adapt the examples, with $n = 1$, of [H_1] to this case).

REMARK. There is a difference between $n = 1$ and $n \geq 2$. When $n = 1$, the global conjugacies theorems of C^∞-diffeomorphisms of the circle to diophantine rotations (cf. [Y_1]) imply that ψ has to loose its smoothness when $\varphi \in \partial U_\alpha$, see [$H_2$, p.46-49]. If $n \geq 2$, there are no global conjugacy theorems to diophantine translations of $\mathbb{T}^n$ (cf. [H_5]) which makes question 2.1 harder since one has to solve 2 problems.

3 Questions on measure preserving diffeomorphims of 2 manifolds.

The following question would be to give a counterexample to a conjecture (or question) of Birkhoff [B_1].

3.1 - Question.

Does there exist a $C^\omega(=\mathbb{R}$-analytic)-diffeomorphism, homotopic to the identity, Lebesgue measure-preserving, of $\mathbb{T}^1\times[-1,1]$ or S^2, with a finite number of periodic orbits and a dense orbit ?

The answer is positive for C^∞-diffeomorphisms (Anosov, Katok [AK]).

The problem would be to create methods to prove similar results as in [AK] for C^ω-diffeomorphisms ; as the group of C^ω-diffeomorphisms of a compact C^ω-manifold with the natural C^ω-topology is not a Baire space, one has to work in the complex ; which requires new methods (cf. [Y][PY]).

3.2 - Question. *Let f be a C^∞-diffeomorphism preserving the Lebesgue measure of $\mathbb{T}^1\times[-1,1]$, homotopic to the identity, that has a finite number of periodic points* (in fact, by a result of Franks [ICM 94] also obtained independly by P. Le Calvez, no periodic points) *and is such that the rotation number $\rho(f_{|\mathbb{T}^1\times\{-1\}})=\alpha$ satisfies a diophantine condition. Is f C^∞-conjugated to $R_\alpha(\theta,r)=(\theta+\alpha,r)$?*

I would expect a counter-example, but there is some evidence in the opposite direction.

We will show elsewhere this is the case if f is C^∞-close to R_α and f is always C^∞-conjugated to R_α near $\mathbb{T}^1\times\{\pm1\}$.

When α is a Liouville number a negative answer to the question can happen in the exotic examples constructed by M. Handel [HA] (see also [H_9]).

We can ask the same question replacing the condition that f is Lebesgue measure preserving by the intersection property (this could make a counter-example easier).

We also ask the similar question for $\mathbb{D}^2=\{z\in\mathbb{C},|z|\leq 1\}$. When f is Lebesgue measure preserving and f has a finite number of periodic points is f C^∞-conjugated to r_α ?
This is always the case near 0 and $\partial\mathbb{D}^2$ and globally when f is C^∞ near $r_\alpha, r_\alpha z=e^{2i\pi\alpha}z$.

3.3 - Let $f: z\in\mathbb{R}^{2n}\to Az+O(z^2)\in\mathbb{R}^{2n}$ be a germ of symplectic diffeomorphisms such that $A\in Sp(2n,\mathbb{R})$ is conjugated in $Sp(2n,\mathbb{R})$ to $r_{\alpha_1}\times\cdots\times r_{\alpha_n}$, $\alpha=(\alpha_1,\ldots,\alpha_n)\in DC$.

3.4 - Conjecture.
If f is real analytic, then f leaves invariant, in any small neighbourhood of 0, a set of positive Lebesgue measure of Lagrangian tori.

3.5 - Remarks.

By Moser's theorem and a theorem of Rüssmann [R], the conjecture is true when $n=1$. We insist that we ask for a set of positive measure of invariant tori in the conjecture. If we ask the conjecture when f is only C^∞, the conjecture, as we will show elsewhere is correct if $n=1$, unknown if $n=2$, and false if $n\geq 3$.
There are many cases when the conjecture is true.

After Birkhoff normal form in polar symplectic coordinates $x_j = \sqrt{r_j}\cos 2\pi\theta_j$ $y_j = -\sqrt{r_j}\sin 2\pi\theta_j$, $\alpha \in DC$, $f(\theta, r) = (\theta+\alpha+\ell(r), r)+(\ell(0) = 0)$, $\ell \in (\mathbb{R}[[r_1, \ldots, r_n]])^n$ (i.e. formal power series).

The conjecture is true when the components of ℓ are independant over $\mathbb{R}$ as formal power series (this follows from a theorem of Rüssmann (cf. [BHS]) and with this hypothesis, the conjecture is true even when f is C^∞) or $\ell = 0$ [R].

4 Entropy and exponents.

4.1 - Let M^n be a compact connected C^∞-manifold, $f : M^n \to M^n$ a C^∞-diffeomorphism that leaves invariant a probability measure μ on M^n. We denote by $Df(x)$ the tangent of f at $x \in M$.
Let

$$\lambda_+(f,\mu) = \lim_{k\to+\infty} \frac{1}{k}\int_M \log \|Df^k(x)\|_x \, d\mu(x) = \inf_{k\geq 1} \frac{1}{k}\int_M \log \|Df^k(x)\|_x \, d\mu(x)$$

where Df^k is tangent map for f^k, $\| \ \|_x$ is the operator norm induced by a Riemannian metric on TM of $Df^k(x) : T_xM \to T_{f^k(x)}M$.
When μ is a C^∞ probability measure (in every chart $\mu = \varphi dx_1 \otimes \cdots \otimes dx_n$, $\varphi \in C^\infty$, $\varphi > 0$) by Pesin's formula the entropy of f for the measure μ satisfies $n\lambda_+(f,\mu) \geq h_\mu(f) \geq \lambda_+(f,\mu)$ so that $\lambda_+(f,\mu) = 0 \Rightarrow h_\mu(f) = 0$ (also is true if f is only C^1 by Margulis-Ruelle).
$G^\infty_\mu(\mathbb{D}^2) = \{f \in \mathrm{Diff}^\infty(\mathbb{D}^2)$, support$(f) \subset \mathbb{D}^2$, $f_*\mu = \mu =$ Lebesgue measure normalised$\}$.

4.2 - Conjecture.

For every C^∞-neighbourhood W of the identity in G^∞_μ there exists $g \in W$ such that $h_\mu(g) > 0$.

See [K_1] for an example of $f \in G^\infty_\mu(\mathbb{D}^2)$ with $h_\mu(f) > 0$.

Let us remark that since $f \to \lambda_+(f,\mu) \geq 0$ is upper semicontinuous, the set $G = \{f, \lambda_+(f,\mu) = 0\}$ is a G_δ in $G^\infty_\mu(\mathbb{D}^2)$ for the C^∞-topology.

4.3 - *One of the 2 exclusive possibilities is true by Baire category :*

a) The set G *is a dense* G_δ.

b) There exists $U \subset G^\infty_\mu$, a C^∞ open set, $U \neq \emptyset$, for every $g \in U$ one has $h_\mu(g) \geq \delta$, for some $\delta > 0$.

Question. *Is a) or b) true ?*

Mañé claims [ICM 82] that, for $G^\infty_\mu(M^2)$ where (M^2, μ) is a compact 2-manifold without boundary and μ is C^∞ probability measure, then a) is true for G a dense G_δ in the C^1 topology on $G^\infty_\mu(M^2) - \{f, f$ is an Anosov diffeomorphism$\}$.
For an outline of a possible proof, see [M].

A similar question can be asked for smooth measure preserving diffeomorphisms on (M^n, μ), any compact manifold of dimension $n \geq 2$, or for symplectic diffeomorphisms on symplectic manifolds (M^{2n}, w), $n \geq 1$, without boundaries.

4.4 - There are cases when b) has a positive answer : when f is partially hyperbolic (a C^1 open property) : the tangent map of f, Df, leaves invariant

$TM = E^s \oplus E^c \oplus E^u$, 3 continuous bundles and there exists $C_j > 0$, $1 \le j \le 4$, such that 0, for every $k \in \mathbb{N}$,

$\|Df^k v\|_{f^k(x)} \ge C_1\, \lambda_1^k \|v\|_x,\ \lambda_1 > 1,\ v \in E^u_x$,
$\|Df^k v\|_{f^k(x)} \le C_2\, \lambda_2^k \|v\|_x,\ \lambda_2 < 1,\ v \in E^s_x$,
$C_4\, \lambda_4^k \|v\|_x \le \|Df^k v\|_{f^k(x)} \le C_3\, \lambda_3^k \|v\|_x,\ v \in E^c_x$,
$0 < \lambda_2 < \lambda_4 \le 1 \le \lambda_3 < \lambda_1$.

4.5 - More generally and in particular in the case volume preserving, there is the notion of *dominated splitting* : $TM = E^s \oplus E^u$, continuous invariant bundles, invariant by Df, such that : $\|Df^k_{|E^s_x}\|_x\, \|Df^{-k}_{|E^u_{f^k(x)}}\|_{f^k(x)}\| \le C\, \lambda^k$, $0 < \lambda < 1$, $C > 0$.

4.6 - *The notion of stably exponents* $= SE^\infty_*$.

Let $f \in G^\infty_*(M)$ where $* = \mu$ for smooth measure preserving diffeomorphisms, or $* = w$ for symplectic form.

4.7 - Definition.

Let $f \in G^\infty_*(M)$. We say that $f \in SE^\infty_*$ if there exists $\delta > 0$ and there exists V a C^1-neighbourhood of f such that every $g \in V$, and for every periodic orbit of $g, 0_{p/q} = \{g^j(x_0),\ 0 \le j \le q-1, g^q(x_0) = x_0\}$,

$$\lambda_+(g, \mu_{p/q}) \ge \delta > 0 \ \text{ where } \mu_{p/q} = \frac{1}{q} \sum_0^{q-1} \delta_{g^j(x_0)},\ x_0 \in 0_{p/q}$$

and δ_y denotes the Dirac mass at $y \in M^n$.

We take the C^1-topology only in order to use closing lemma arguments.

Using the ergodic closing lemma of Mañé ([A_2]), $f \in SE^\infty_*$ implies for every υ probability measure invariant by f satisfies $\lambda_+(f, \upsilon) \ge \delta$.

Remark. When f is a general diffeomorphism we ask in the definition of stable exponents that $\inf(\lambda_+(g, \mu_{p/q})\, ,\ \lambda_+(g^{-1}, \mu_{p/q})) \ge \delta$.

4.8 - Remarks.

If $f \notin SE^\infty_*$, there exists a C^1-perturbation g of f, $g \in G^\infty_*(M)$, such that g has a totally elliptic periodic orbit : there exists $x_0, g^q(x_0) = x_0$ and all the eigenvalues of $Dg^q(x_0)$ are on the unit circle.
Then by a C^1-perturbation we can suppose that g is periodic on a small neighbourhood of x_0. After a C^∞-perturbation one falls in the KAM world (symplectic, or for volume preserving the existence of codimension 1 periodic invariant diophantine tori : Cheng Sun [C], Yoccoz [Y], Xia [X], and by the similar proof as in Yoccoz [Y], see also [BHS]). The above results do not violate that C^∞-generically in $G^\infty_\mu(M^n)$, when $n \ge 3$, all the periodic points are hyperbolic.

When $f \in SE^\infty$ is not measure preserving, using the definition in the remark of 4.7, then by a C^1-perturbation, one can create a sink.

Of course, by definition SE_* is C^1-open in $G^\infty_*(M)$.

4.9 - Questions.

a) *Is the set $\{f \in G^\infty_*(M), f$ has a dominated splitting on $M\}$ C^1-dense in SE^∞_* ?* (it is C^1-open).

For many reasons it seems reasonable to conjecture that the answer is positive (see ([DPU]) on 3-manifolds for general diffeomorphisms "stably transitive").

It follows from Newhouse [N] that if $f \in G_w^\infty(M^{2n})$ is such that all periodic points are C^1-stably hyperbolic, then f is an Anosov diffeomorphism.

b) *Does there exist a connected and simply connected manifold M^n with a partially hyperbolic diffeomorphism* ?

The answer is not even known if the dimension of M^n is $n = 3$. If $E^c \oplus E^u$ or $E^c \oplus E^s$ defines a C^0- foliation, it would follow that it is negative by the C^0-Novikov theorem, but it is an open problem to prove, when $n = 3$, that $E^c \oplus E^u$ defines a C^0-foliation !

c) *On $G_*^\infty(M) - SE_*^\infty$, which of the two possibilities of* 4.3 *is true* ?

4.10 - Let M^n be compact connected manifold and f a strictly ergodic C^∞-diffeomorphism (minimal (i.e. every orbit is dense) and uniquely ergodic).

CONJECTURE. *The topological entropy of f equals* 0.

The conjecture is true when $n = 2$, Katok [K] ([ICM 82]) ; A.B. Katok remarked the minimal diffeomorphisms with positive topological entropy constructed in [H_6] are not uniquely ergodic.

5 EXISTENCE OF PERIODIC ORBITS.

There is some evidence, [G][H_7], that the closing lemma of Pugh will not generalize to the C^∞-topology. So we will ask some precise questions, hoping they might lead to some better understanding of the problem.

5.1 - Let $\mathrm{Diff}_o^\infty(\mathbb{T}^n)$, the group of diffeomorphisms of $\mathbb{T}^n$, C^∞-isotopic to the identity endowed, with the C^∞-topology.

Let $G = \{f \in \mathrm{Diff}_0^\infty(\mathbb{T}^n),\ f$ has no periodic points$\}$ (that is a G_δ) and $\overline{G}$ the C^∞-closure of G $\mathrm{Diff}_0^\infty(\mathbb{T}^n)$.

5.2 - QUESTION. *Does the closed set $\overline{G}$ have no interior ($n \geq 2$) in* $\mathrm{Diff}_0^\infty(\mathbb{T}^n)$?

One can ask the above for volume-preserving diffeomorphisms when $n \geq 3$. In counterpart, it follows from Conley-Zehnder [HZ] (see also Zehnder [ICM 86]), for $\mathrm{Diff}_{0,w}^\infty(\mathbb{T}^{2n})$ when w is a constant symplectic form and $0, w$ denotes that the isotopy is symplectic, that $\overline{G_w}$ has no interior in $\mathrm{Diff}_{0,w}^\infty(\mathbb{T}^{2n})$, where $G_w = G \cap \mathrm{Diff}_{0,w}^\infty(\mathbb{T}^{2n})$.

5.3 - Along the lines of Poincaré [P_1] (see [H_7]), we ask :

Let $H_0(r)$ be a function defined *on* $T^*\mathbb{T}^n = \mathbb{T}^n \times \mathbb{R}^n$, $n \geq 2$. *We suppose H_0 is C^ω-convex with superlinear growth* when $\|r\| \to +\infty$, $H_0(r)/\|r\| \to +\infty$), and $\frac{\partial H_0}{\partial r}(0) = 0$.

QUESTION. *Can one find a C^ω-family $H_\varepsilon(\theta, r) = H_0(r) + \varepsilon H_0(\theta, r, \varepsilon)$, every thing being $\mathbb{R}$-analytic, such that for some ε_0 small on $H_{\varepsilon_0}^{-1}(1)$, the periodic orbits of the Hamiltonian flow of H_{ε_0} are not dense* ?

The only real restrictions we ask for an example is that for every $\theta, r \to H_{\varepsilon_o}(\theta, r)$ is strictly convex and $H_{\varepsilon_o}^{-1}(]-\infty, 1[)$ countains the zero section.
One easily constructs C^∞-counter examples, but this is not the question asked !

5.4 - Recently V. Ginzburg, [G$_1$], and the author constructed $M^{2n-1} \subset \mathbb{R}^{2n}$, $n \geq 4$, compact connected, C^∞-hypersurface such that characteristic flow on M^{2n-1} has no periodic orbits (for C^∞-compact connected examples when $n = 3$, see [G$_2$]).

Let us write $M^{2n-1} = H^{-1}(1)$, where H is a C^∞-function ans 1 is a regular value. A theorem of Hofer and Zehnder [HZ] says that for $\varepsilon_n \to 1$, the Hamiltonian flow of H on $H^{-1}(\varepsilon_n)$ has at least one periodic orbit P_{ε_n} and one wonders how bad P_{ε_n} behaves, when $\varepsilon_n \to 1$, in the space of compact sets on $H^{-1}_{\varepsilon_n}$.

5.5 - QUESTIONS.

1. *When $n \geq 2$ can one find $M^{2n-1} \subset \mathbb{R}^{2n}$, a C^∞-compact connected hypersurface, such that on M^{2n-1} the characteristic flow is minimal (every orbit is dense)* ?

2. *Can the characteristic flow on M^{2n-1} be an Anosov flow* ?

5.6 - Let M^n be a compact connected oriented manifold of dimension $n \geq 3$ and with Euler-Poincaré characteristic $\chi(M^n) = 0$. We consider C^∞, Ω-divergence free vectors-fields where Ω is a C^∞-volume form on M^n. We suppose $X(x) \neq 0, \forall\, x \in M^n$. We denote by f_t^X the flow of the vector-field X.

5.7 - QUESTIONS.

a) *Does there exist vector fields as above (in every homotopy class of non-zero vector fields) such that the flow of f_t^X has no periodic orbit* ?

b) *Does there exist vector-fields as above such that the flow f_t^X is Ω-ergodic* ?

(Anosov [A] constructed an ergodic flows f_t^X on every (M^n, Ω) orientable manifold ($n \geq 3$) but the vector field in Anosov's construction has zeros (which of course will be the case when $\chi(M^n) \neq 0$). For $n = 2$, Anosov's result is also true except, by Poincaré-Bendixon's theorem, on the 2-sphere (we suppose that M^n is orientable otherwise, one has to replace Ω by a smooth measure and Anosov's theorem is still true except for the projective plane or the Klein bottle). I certainly believe the answers are positive when n is large. The case $n = 3$ is much more delicate (in the C^1 case, for question a), see [K$_2$]).

5.8 - Let (M^{2n}, w) be a C^∞-compact symplectic manifold. Given a closed 1-form, v we define a symplectic vector-field X by $i_X w = v$. The following statements are equivalent :

- The vector-field X has no zeros ;
- v has no zeros ;
- M^{2n} fibers of S^1 (Tischler [T]).

Claim : There exists a symplectic manifold M^{2n} with $\chi(M^{2n}) = 0$, but M^{2n} does not fiber over S^1.

To see this, we consider $M^4 = (\mathbb{P}_1(\mathbb{C}) \times M_g) \sharp_k \overline{\mathbb{P}}_2(\mathbb{C})$ where M_g is a compact surface of genus $g \geq 2$ and we blow up k points of $\mathbb{P}_1(\mathbb{C}) \times M_g$. We have $\chi(M^4) = 2(2 - 2g) + k = 0$ when $k = 2(-2 + 2g)$. The manifold M^4 does not fiber over S^1 since the fiber, if it existed, would have a non finitely generated fundamental group.

CONJECTURE.

Every symplectic diffeomorphism homotopic to the identity of (M^{2n}, w), where $\chi(M^{2n}) = 0$ and M^{2n} does not fiber over S^1, has a fixed point (in fact a number of fixed points $\geq$ minimal number of zeros of any symplectic vector-field).

When f is C^1-near the identity, the conjecture is easy and true.

6 INSTABILITIES OF HAMILTONIAN FLOWS ON $T^*\mathbb{T}^n$, $n \geq 3$, AND THE PROBLEM OF TOPOLOGICAL STABILITY.

6.1 - A part from the results of Arnold [A_3], R. Douady and P. Le Calvez [DL_2][D_2], and the beautiful work of John Mather for twist maps [M_1] (in [LC_2] P. Le Calvez proves some of Mather's results by topological methods following in part Birkhoff) and partial generalizations to higher dimensions by Mather [MA_2] [MA_3] and a survey [MA_4], the subject lacks any non-trivial result.

6.2 - QUESTIONS.

1. *Can one find an example of a C^∞-Hamiltonian H in a small C^k-neighbourhood $k \geq 2$ of $H_0(r) = 1/2\|r\|^2$ such that, on $H^{-1}(1)$, the Hamiltonian flow of H has one dense orbit ?*

See the example of Donnay Liverani [DL] of a C^∞-Riemannian metric on $\mathbb{T}^2$ with an ergodic geodesic flow.

Many people believe that examples as above do exist and are C^∞-generic (cf. P. and T. Ehrenfest [E], G.D Birkhoff [B_2], V.I. Arnold [ICM 66]) and these questions have been called by the author the quasi-ergodic hypothesis, following [E] (but in this reference, no clear distinction is made between "every orbit is dense" and "one orbit is dense", and we choose "one orbit is dense"). The negation of the quasi-ergodic hypothesis is topological stability.

In the following questions I suppose that H is in a C^2-neighbourhood of $H_0(r)$, that is convex, of super linear growth, and such that $\frac{\partial H_0}{\partial r}(0) = 0$.

(What we really want is that $H^{-1}(1)$ be connected and that the set $H^{-1}(]-\infty, 1[)$ contains a Lagrangian manifold, that is the graph over $\mathbb{T}^n$).

2. *For the C^∞-generic H and the generic ergodic minimal measure μ of Mather* [MA_2], *is the flow $f^H_{t|\mathrm{supp}(\mu)}$ on $H^{-1}(1)$ a hyperbolic flow ?*

(For the twist map case, see Le Calvez [LC_1]).

3. *For the C^∞-generic H on $H^{-1}(1)$ are the probability measures defined by periodic orbits, dense (in the weak topology on probability measures) in the ergodic minimal measures ?*

(For monotone twist maps the result is known by Aubry Mather's theory [MA_4]. The higher dimensional case is much more delicate (cf. [A_1], see also [H_4, §3.3, p. 53]).

4. *Can one find C^∞-generically a closed connected set F, invariant by the Hamiltonian flow f^H_t of H, $F \subset H^{-1}(1)$, such that F contains all the supports of the ergodic minimal measures, in F the orbits asymptotic to the locally minimal orbits* (cf. [LC_2]) *are dense in F, and on F, $f^H_{t|F}$ has a dense orbit ?*

The above questions are precise, compatible with what J. Mather proved for the dynamics of monotone twist maps in Birkhoff zones of instabilities (the existence

of a dense orbit is an open question). The author conjectures the answer of the question 4 is positive and the methods of Mather, specially variational methods, make that the question is not desperate and furtheremore no hypothesis of small C^∞-perturbations of completely integrable systems is needed. In counterpart, as we asked in the question that the orbits asymptotic to the locally minimal orbits (minimizing with constraints) are dense in F (condition that J. Mather uses in most proofs (cf. [MA_4],p. 162-169)), C^∞-generically the locally minimal orbit (that is a closed set) are *nowhere dense in* $H^{-1}(1)$ (so the question is not really related to the quasi-ergodic hypothesis), but would show how to move around frequencies (rotation vectors of the ergodic minimizing measures). For some examples, when $n = 2$ and large perturbations, see V. Bangert [BA,9.12].

7 The oldest open question in dynamical systems.

7.1 - I. Newton, [N_1], certainly believed that the n-body problem, $n \geq 3$, (n particles moving under universal gravitation) is topological instable and, to paraphrase Laplace, makes the hypothesis that God solves the problem and controls the instabilities (hypothesis criticized by Leibniz and all the enlighted XVIII$^{\text{th}}$ century). The question we will ask for the n-body problem is a special case when the energy surface is not compact and the volume form on the energy surface is not finite (hypotheses at infinity are of course necessary), but we will not formulate any general question.

7.2 - For the n-body problem in space, we will suppose $n \geq 3$.

- The center of mass is fixed at 0.
- On the energy surface we C^∞-reparametrize the flow by a C^∞ function φ_e (after reduction of the center of mass) such that the flow is complete : we replace H by $\varphi_e(H-e) = H_e$ so that the new flow takes an infinite time to go to collisions ($\varphi_e > 0$ is a C^ω function outside collisions).

7.3 - Following G.D. Birkhoff [B_3] (who only considers the case $n = 3$ and the angular momentum $\neq 0$) (see also A.N. Kolmogorov [ICM 54]), we ask :

Question. *Is for every e the non wandering set of the Hamiltonian flow of* H_e *on* $H_e^{-1}(0)$ *nowhere dense in* $H_e^{-1}(0)$?

In particular, this would imply that the bounded orbits are nowhere dense and no topological stability occurs.

It follows from the identity of Jacobi-Lagrange that, when $e \geq 0$, every point such that its orbit is defined for all times, is wandering.

The only thing known is that, even when $e < 0$, wandering sets do exist (Birkhoff and Chazy, see V.M. Alexeyev [ICM 70]).

The fact that the bounded orbits have positive Lebesgue-measure when the masses belong to a non empty open set, is a remarkable result announced by V.I. Arnold [A_4] (Arnold only gives a proof for planar 3-body problem and if the author is not mistaken, Arnold's claim is correct).

In some respect Arnold's claim proves that Lagrange and Laplace, against Newton, are correct in the sense of measure theory and that in the sense of topolgy, the above question, in some respect, could show Newton is correct. For some soft

(almost explicit) examples of dissipative nature of Hamiltonian on non-compact energy surfaces with infinite volume we refer the reader to [H_8].
What seems not an unreasonable question to ask (and possibly prove in a finite time with a lot of technical details) is that :

In one of the masses $m_0 = 1$ and the other masses $m_j = \varepsilon M_j$, $1 \leq j \leq n-1$, $M_j > 0$, $\varepsilon > 0$, then in any neighbourhodd of fixed different circulat orbits arounds m_0 moving in the same direction in a plane, when ε is small, there are wandering domains.

But in many respects this is not the question asked that is more global.

In Herman [ICM 78] (that refers to the article of M.R. Herman in the Proceedings of the ICM held in 1978) the reader will find other open problems some of which are still insolved. For other problems on Siegel's linearization theorem (most are still open) we refer to [H_{10}].

References

[A] D.V. Anosov, *Existence of smooth ergodic flows on smooth manifolds*, Math. U.R.S.S. Izvestija **8** (1974), p. 525-552.

[AK] D.V. Anosov, A.B. Katok, *New examples in smooth ergodic theory*, Trans. Moscow Math. Soc. **23** (1970), p. 1-35.

[A_1] Marie-Claude Arnaud, *Type des points fixes des difféomorphismes symplectiques de $\mathbb{T}^n \times \mathbb{R}^n$*, Mémoire (nouvelle série) n° 48, Suppl. Bull. Soc. Math. Fr. t.**120**, Fasc. 1, (1992).

[A_2] Marie-Claude Arnaud, *Le "closing Lemma" en topologie C^1*, to appear in Mémoires de Soc. Math. Fr.

[A_3] V.I. Arnold, *Instability of dynamical systems with several degrees of freedom*, Sov. Math. Dok **5** (1964), p. 581-585.

[A_4] V.I. Arnold, *Small denominators and problems of stability of motion in classical and celestial mechanics*, Russ. Math. Surveys **18**,6 (1963), p. 85-191.

[BA] V. Bangert, *Mather sets for twist maps and geodesics on tori, Dynamics reported*, Vol.1, Wiley & Sons (1988), p. 1-56.

[BHS] H.S. Broer, G.B. Huitema, M.B. Sevryuk, *Quasi-periodic motions in families of dynamical systems*, Springer Lect. Notes in Math. n° **1645** (1996).

[B_1] G.D. Birkhoff, Collected Math. Papers, Vol.2, Dover, p. 712.

[B_2] G.D. Birkhoff, Collected Math. Papers, Vol.2, Dover, p. 462-465.

[B_3] G.D. Birkhoff, Dynamical Systems, Amer. Math. Soc. Colloquium Pub. Vol. IX (1966), p. 290.

[C] Ch.-Q. Cheng, Y.S. Sun, *Existence of invariant tori in three-dimensional measure-preserving mappings*, Celest. Mech. Dyn. Astr. **47** (1990), p. 275-292.

[DPU] L.J. Diaz, E.R. Pujals, R, Ures, *Partial hyperbolicity and robust transitivity*, Preprint (1997).

[DL_1] V.J. Donnay, C. Liverani, *Potentials on the two-torus for which the Hamiltonian is ergodic*, Commun. Math. Phys. **135** (1991), p. 267-302.

[D_1] Adrien Douady, *Disques de Siegel et anneaux de Herman*, Séminaire Bourbaki n° 677, Astérisque **152-153**, Soc. Math. Fr. (1987), p. 151-172.

[D_2] Raphaël Douady, *Stabilité ou instabilité des points fixes elliptiques*, Ann. Sci. Ec. Norm. Sup., t.**21** (1988), p. 1-46.

[DL$_2$] Raphaël Douady, P. Le Calvez, *Exemple de point fixe elliptique non topologiquement stable en dimension 4*, C.R. Acad. Sci. Paris, t. 296 (1983), p. 895-898.

[E] P. and T. Ehrenfest, The conceptual foundations of the statistical approach in mechanics, Dover.

[G$_1$] V.L. Ginzburg, *An embedding $S^{2n-1} \to \mathbb{R}^{2n}$, $2n-1 \geq 7$, whose Hamiltonian flow has no periodic trajectories*, I.M.R.N. 2 (1995), p. 83-98.

[G$_2$] V.L. Ginzburg, A smooth counterexample to the Hamiltonian Seifert conjecture in $\mathbb{R}^6$, Preprint, Santa-Cruz (1997).

[G] C. Gutirrez , *A counterexample to a C^2-closing lemma*, Erg. Th. and Dyn. Systems 7 (1987), p. 509-530.

[HA] M. Handel , *A pathological area preserving C^∞-diffeomorphisms of the plane*, Proc. Amer. Math. Soc. 86 (1982), p. 163-169.

[H$_1$] M.R. Herman, *Sur les courbes invariantes par les difféomorphismes de l'anneau*, Vol.1, Astérisque 103-104, Soc. Math. Fr. (1983).

[H$_2$] M.R. Herman, *Sur les courbes invariantes par les difféomorphismes de l'anneau*, Vol.2, Astérisque 144, Soc. Math. Fr. (1986).

[H$_3$] M.R. Herman, *Existence et non existence de tores invariants par des difféomorphismes symplectiques*, Sém. E.D.P. 1987-1988, Exposé XIV, Centre de Mathématiques de l'Ecole Polytechnique, p. 1-24.

[H$_4$] M.R. Herman, *Inégalités a priori des tores lagrangiens par des difféomorphismes symplectiques*, Publ. Math. I.H.E.S. n° 70 (1989), p. 47-100.

[H$_5$] M.R. Herman, *Une méthode pour minorer les exposants de Lyapunov et quelques exemples montrant le caractère local d'un théorème d'Arnold et de Moser sur le tore de dimension 2*, Commentarii Math. Helv. 58 (1983), p. 457-502.

[H$_6$] M.R. Herman, *Construction d'un difféomorphisme minimal d'entropie topologique non nulle*, Erg. Th. and Dyn. Systems 1 (1981), p. 61-76.

[H$_7$] M.R. Herman, *Exemples de flots hamiltoniens dont aucune perturbation en topologie C^∞ n'a d'orbites périodiques sur un ouvert de surfaces d'énergie*, C.R. Acad. Sci. Paris, t. 312 (1991), p. 989-994.

[H'$_7$] M.R. Herman, *Différentiabilité optimale et contre-exemples à la fermeture en topologie C^∞ des orbites récurrentes de flots hamiltoniens*, C.R. Acad. Sci. Paris, t. 313 (1992) p. 153-182.

[H$_8$] M.R. Herman, *Dynamics connected with indefinite normal torsion*, twist mappings and their applications, Ed. R. Mc Gehee and K. Meyer, IMA Vol.44, Springer (1992), p. 153-182, see § 4.20-4.22.

[H$_9$] M.R. Herman, *Construction of some curious diffeomorphisms of the Riemann sphere*, J. London Math. Soc. 34 (1986), p. 375-384.

[H$_{10}$] M.R. Herman, *Recent results and some open questions on Siegel's linearization theorem of germs of complex analytic diffeomorphism of $\mathbb{C}^n$ near a fixed point*, Proc. VIII Int. Conf. Math. Phys. (Marseille 1989), World Scientific (1987), p. 138-198.

[HZ] H. Hoffer, E. Zehnder, *Symplectic invariants and Hamiltonian dynamics*, Birkhaušer (1994).

[K] A.B. Katok, *Lyapunov exponents, entropy and periodic orbits for diffeomorphisms*, Publ. Math. I.H.E.S. 51 (1980), p. 137-173.

[K$_1$] A.B. Katok, *Bernoulli diffeomorphisms on surfaces*, Ann. Math. 110 (1979), p. 529-547.

[K$_2$] G. Kuperberg, *A volume-preserving counterexample to the Seifert conjecture*, Commentarii Math. Helv. 71 (1996), p. 70-97.

[LC$_1$] P. Le Calvez, *Les ensembles d'Aubry-Mather d'un difféomorphisme de l'anneau déviant la verticale sont en général hyperboliques*, C.R. Acad. Sci. Paris, t. 306 (1998), p. 51-54.

[LC$_2$] P. Le Calvez, *Propriétés dynamiques des régions d'instabilité*, Ann. Sci. Ec. Norm. Sup. 20 (1987), p. 443-464.

[M] R. Mañé, *The Lyapunov exponents of generic area preserving diffeomorphisms*, Int. Conf. Dyn. Syst. (Montevideo), Pitman Res. Notes Math. Ser. 362 (1996), p. 110-119.

[MA$_1$] J. Mather, *Variational construction of orbits for twist diffeomorphisms*, J. Amer. Math. Soc. 4 (1991), p. 203-267.

[MA$_2$] J. Mather, *Action minimizing invariant measures for positive definite Lagrangian systems*, Math. Z. 207 (1991), p. 169-207.

[MA$_3$] J. Mather, *Variational construction of connecting orbits*, Ann. Inst. Fourier 43 (1993), p. 1349-1386.

[MA$_4$] G. Forni, J.N. Mather, *Action minimizing orbits in Hamiltonian systems*, Springer Lect. Notes in Math. n° 1589, p. 91-186.

[N] S.E. Newhouse, *Quasi-elliptic periodic points in conservative dynamical systems*, Amer. J. Math. 99 (1977), p. 1061-1087.

[N$_1$] I. Newton,

• *Principa, General Scholium*, End of Book III, Translation Mottes, Vol. II, Univ. California Press, p. 544.

• *Optics*, Dover, Query 31, p.402, see also Query 27, p. 369.

[P] R. Perez-Marco, *Siegel disks with quasi-analytic boundary*, Preprint Univ. Paris-Sud (1997).

[PY] R. Perez-Marco, J.-C. Yoccoz, *Germes de feuilletages holomorphes*, Astérisque 222, Soc. Math. Fr. (1994), p. 345-371.

[P$_1$] H. Poincaré *Les méthodes nouvelles de la mécanique céleste*, t.I, Gauthier-Villars (1892) §32, p. 82 that refers to §13.

[R] H. Rüssmann, *On the convergence of power series transformations of analytic mappings near a fixed point into a normal form*, Preprint, I.H.E.S. (1977).

[S] D. Salamon, *The Kolmogorov-Arnold-Moser theorem*, Preprint ETH Zürich (1986).

[T] T. Tischler, *On fibering certain foliated manifolds*, Topology 9 (1970), p. 153-154.

[X] Zh. Xia, *Existence of invariant tori in volume-preserving diffeomorphisms*, Erg. Th. and Dyn. Systems 12 (1992), p. 621-631.

[Y] J.-C. Yoccoz, *Travaux de Herman sur les tores invariants*, Séminaire Bourbaki n° 754, Astérisque 206, Soc. Math. Fr. (1992), p. 311-344.

[Y$_1$] J.-C. Yoccoz, *Conjugaison différentiable des difféomorphismes du cercle dont le nombre de rotation vérifie une condition diophantienne*, Ann. Sc. E.N.S. t.17, (1984), p. 333-359.

[Y$_2$] J.-C. Yoccoz, *Petits diviseurs en dimension 1*, Astérisque n° 231, Soc. Math. Fr. (1995).

Michael Herman
Univ. Denis Diderot/Paris 7
Inst.de Math.- Géométrie & Dynamique
Tour 45/55 - 5e étage -
2, Place Jussieu
75251 Paris Cedex 05

Random Dynamics and its Applications

Yuri Kifer

Abstract. Random transformations emerge in a natural way as a model for description of a physical system whose evolution mechanism depends on time in a stationary way. This leads to the study of actions of compositions of different maps chosen from a typical sequence of transformations. The question whether such actions are chaotic can be dealt with employing the random thermodynamic formalism developped in recent years. This theory has nice applications to random networks, fractal dimensions of random sets and other models.

1991 Mathematics Subject Classification: Primary 58F11; Secondary 60J99, 60F15
Keywords and Phrases: random transformations, thermodynamic formalism, limit theorems

1. Introduction

Evolution of many physical systems can be better described by compositions of different maps, i.e. time dependent transformations, rather than by repeated application of exactly the same transformation. It is natural to study such problems for typical in some sense sequences of maps which leads to the framework of random transformations.

This set up was discussed already in Ulam and von Neumann [UN] and in Kakutani [Ka] in connection with random ergodic theorems. Later this topic was studied in the framework of the relativized ergodic theory (Thouvenot [Th], Ledrappier and Walters [LW]) but the real push this subject received in 80-ies when stochastic flows appearing as solutions of stochastic differential equations provided a rich source of random diffeomorphisms (see references and the historical review in Arnold [Ar]). This prompted, in particular, the book [Ki1] which, in turn, played a role in motivating other work such as the general relativized variational principle (Bogenschütz [Bo]) and some results in smooth random dynamics. Further developments of the latter included random invariant manifolds, Lyapunov exponents, and a random bifurcation theory (see Arnold [Ar] and references there), random versions of the Margulis-Ruelle entropy inequality (see Kifer [Ki1], Liu and Qian [LQ], Bahnmüller and Bogenschütz [BB]) and of the Pesin entropy formula and the corresponding characterization of the random Sinai-Ruelle-Bowen measures (see Ledrappier and Young [LY], Liu and Qian [LQ], Bahnmüller and Liu [BL]), and the random thermodynamic formalism (see Kifer [Ki2], Bogenschütz and Gundlach [BG], Khanin and Kifer [KK]).

The formal set up consists of a probability space $(\Omega, \mathcal{A}, P)$ together with a $P-$preserving ergodic invertible map $\theta : \Omega \to \Omega$, of another measurable space $(\mathbf{X}, \mathcal{B})$, and of a measurable subset X of $\mathbf{X} \times \Omega$ with fibers $X^\omega = \{x \in \mathbf{X} : (x, \omega) \in X\} \in \mathcal{B}$. The dynamics is given by a mesurable map $\tau : X \to X$ which is a skew product transformation $\tau(x, \omega) = (F_\omega x, \theta\omega)$ where the fiber maps $F_\omega : X^\omega \to X^{\theta\omega}$ with the composition rule $F^n_\omega = F_{\theta^{n-1}\omega} \circ \cdots \circ F_{\theta\omega} \circ F_\omega : X^\omega \to X^{\theta^n\omega}$ are called random transformations.

Theory of random transformations concerns mainly with actions of F^n_ω for typical $\omega \in \Omega$, i.e. except of ω's forming a set of zero $P-$measure. I shall discuss here only certain aspects of ergodic theory of random transformations related mainly to the random thermodynamic formalism which is crucial in describing chaotic (stochastic) spatial behaviour of compositions F^n_ω for typical ω. Familiar signs of stochastic behavior are the central limit theorem (CLT), the law of iterated logarithm (LIL), large deviations (LD) etc. which hold true for some classes of random transformations such as random expanding in average maps, random subshifts of finite type and certain random hyperbolic diffeomorphisms. The theory has applications to random networks, computations of fractal dimensions of random sets, and to random walks with stationarily changing distributions which also will be discussed in this paper. Recently random diffeomorphisms were employed in some models of statistical physics (see Ruelle [Rue]).

2. Random Thermodynamic Formalism

Let $\mathcal{P}_P(X)$ be the space of probability measures on X whose marginal on Ω coincides with P. I assume that all spaces under consideration are Borel subsets of Polish spaces, and so any $\mu \in \mathcal{P}_P(X)$ has an essentially unique desintegration $\mu(dx, d\omega) = \mu^\omega(dx)P(d\omega)$ with $\mu^\omega, \omega \in \Omega$ being a measurable family of probability measures on X^ω. It is easy to see that μ is $\tau-$invariant if and only if $F_\omega\mu^\omega = \mu^{\theta\omega}$ $P-$almost surely (a.s.). Accordingly, a measurable set $A \subset X$ is $\tau-$invariant if and only if its fibers $A^\omega = \{x : (x, \omega) \in A\}$ satisfy $F_\omega A^\omega = A^{\theta\omega}$ $P-$a.s.

Given $\mu \in \mathcal{P}_P(X)$ the relativized or fiber entropy $h^{(r)}_\mu(\tau)$ of τ can be defined as the conditional entropy of τ with respect to the $\sigma-$algebra $\pi^{-1}\mathcal{A}$ where $\pi : X \to \Omega$ is the natural projection to the second factor (see [LW], [Ki1], [Bo]). Another way to obtain $h^{(r)}_\mu(\tau)$, somewhat similar to the deterministic case, is via finite partitions $\mathcal{R} = \{R_1, ..., R_n\}$ of X into measurable sets. Set $R^\omega_i = \{x : (x, \omega) \in R_i\}$ and $\mathcal{R}^\omega = \{R^\omega_1, ..., R^\omega_n\}$ then $P-$a.s.,

$$h^{(r)}_\mu(\tau) = \sup_{\mathcal{R}} \lim_{n\to\infty} \frac{1}{n} H_{\mu_\omega}\left(\bigvee_{i=0}^{n-1} (F^i_\omega)^{-1}\mathcal{R}^{\theta^i\omega}\right). \tag{2.1}$$

Existence of this limit follows from Kingman's subadditive ergodic theorem (see [Ki1]).

Assume now that $\mathbf{X}$ is compact and the fibers X^ω are Borel subsets of $\mathbf{X}$. For continuous random transformations F_ω and any measurable function g on X such that $g_\omega(x)$ is continuous in x and $\sup_x |g_\omega(x)| \in L^1(\Omega, P)$ introduce another useful

quantity, called the relativized (fiber) topological pressure $Q_\tau(g)$, by

$$Q_\tau(g) = \sup_{\mu\in\mathcal{P}_P(X)} (\int g d\mu + h_\mu^{(r)}(\tau)). \tag{2.2}$$

The number $Q_\tau(0)$ denoted by $h_{\text{top}}^{(r)}(\tau)$ is called the relativized topological entropy. Actually, similarly to the deterministic case the proper definition of $Q_\tau(g)$ is via $(\omega, n, \varepsilon)-$separated sets (see [Ki1] and [Bo]) and then (2.2), called the relativized variational principle, is derived as a theorem. Under rather general conditions, called (random) expansivity, one can show that $h_\mu(\tau)$ is upper semicontinuous in μ, and so the supremum in (2.2) is attained at some $\mu \in \mathcal{P}_P(X)$ which is called an equilibrium state. If a maximizing measure is unique it has usually nice properties. Equilibrium states are related to Gibbs measures and both have their roots in statistical mechanics where g plays the role of a potential.

Next, I shall describe specific models of random transformations which will appear in the following exposition. I shall start with random subshifts of finite type (see [BG] and [KK]) where $X^\omega = X_A^\omega = \{x = (x_0, x_1, ...) : x_i \in \{1, ..., \ell(\theta^i\omega)\}$ and $a_{x_i x_{i+1}}(\theta^i\omega) = 1\, \forall i = 0, 1, ...\}$, $\ell : \Omega \to \mathbb{Z}_+ = \{1, 2, ...\}$ satisfies $0 < \int \log \ell dP < \infty$, and $A(\omega) = ((a_{ij}(\omega)))$, $\omega \in \Omega$ is a measurable family of $\ell(\omega)\times\ell(\theta\omega)-$ matrices with 0 and 1 entries such that $P-$a.s. $A(\omega)$ has no zero row. Random transformations F_ω act on X^ω as left shifts $(F_\omega x)_i = x_{i+1}$. A random subshift of finite type is called topologically mixing if there exists a $\mathbb{Z}_+-$ valued random variable $0 < N = N_\omega < \infty$ so that $P-$a.s. $A(\theta^{-N}\omega)\cdots A(\theta^{-2}\omega)A(\theta^{-1}\omega)$ is a matrix with positive entries. The random Ruelle-Perron-Frobenius (RPF) operator $\mathcal{L}_g^\omega$ corresponding to a function $g = g_\omega(x)$ on X maps functions on X^ω to functions on $X^{\theta\omega}$ by the formula

$$\mathcal{L}_g^\omega q(x) = \sum_{y\in F_\omega^{-1}x} e^{g_\omega(y)} q(y). \tag{2.3}$$

Suppose that $E \sup_x |g_\omega(x)| < \infty$ and

$$|g_\omega(x) - g_\omega(y)| \le K_g(\omega)(\text{dist}(x, y))^\kappa \tag{2.4}$$

for some $\kappa > 0$ and a random variable $K_g(\omega) > 0$ with $E|\log K_g| < \infty$, where $\text{dist}(x, y) = e^{-\min\{i\ge 0: x_i\ne y_i\}}$ and E denotes the expectation on $(\Omega, \mathcal{A}, P)$. Then the random RPF theorem ([KK], [BG]) yields that there exists a unique positive random variable $\lambda = \lambda_\omega$ with $E|\log\lambda| < \infty$, a positive function $h = h_\omega(x)$, and $\nu \in \mathcal{P}_P(X)$ having desintegrations ν^ω such that

$$\mathcal{L}_g^\omega h_\omega = \lambda_\omega h_{\theta\omega},\ (\mathcal{L}_g^\omega)^* \nu^{\theta\omega} = \lambda_\omega \nu^\omega, \text{ and } \int_{X^\omega} h_\omega d\nu^\omega = 1. \tag{2.5}$$

Then the relativised topological pressure of g has the form $Q_\tau(g) = \int \log\lambda_\omega dP(\omega)$ and $\mu \in \mathcal{P}_P(X)$ having desintegrations μ^ω satisfying $d\mu^\omega = h_\omega d\nu^\omega$ is $\tau-$invariant and it is the unique equilibrium state for g.

This set up is quite appropriate to study randomly evolving graphs or random networks $\mathcal{N}$ where $V(\omega) = \{1, 2, ..., \ell(\omega)\}$ is the set of vertices for an (evironment) ω and I connect by an arrow $i \in V(\omega)$ to $j \in V(\theta\omega)$ iff $a_{ij}(\omega) = 1$. A sequence $(i_0, i_1, ..., i_n)$ is a path in $\mathcal{N}$ iff $a_{i_k i_{k+1}}(\theta^k\omega) = 1 \, \forall k = 0, 1, ..., n-1$. The topologically mixing condition formulated above yields that for any $i \in V(\omega)$ and $j \in V(\theta^n\omega)$ with n sufficiently large there exists a path of length n in $\mathcal{N}$ starting at i and ending at j. In the next section I shall formulate general limit theorems which can be applied, in particular, to describe statistical properties of paths in such random networks. More general models of multidimensional random subshifts of finite type and of random sofic shifts, which also have combinatorial applications, were studied in [Ki3] and [GK2], respectively.

By constructing random Markov partitions and employing random subshifts of finite type one can study also random (spatially uniform) hyperbolic diffeomorphisms which have random expanding and contracting (in average) invariant subbundles (see [Li], [GK1]). As an example of this situation take, for instance, $F_\omega = f_\omega^{n(\omega)}$ where f_ω is a random diffeomorphism whose all realizations belong to a small C^2 neighborhood of one Anosov diffeomorphism (or a diffeomorphism having a basic hyperbolic set) and $n = n(\omega)$ is a random variable taking values $0, 1, 2, ...$ with $0 < \int \log(1+n) dP < \infty$. Another interesting example of a random Anosov diffeomorphism is due, essentially, to Arnoux and Fisher. Let $\sigma : \Omega \to \Omega$ be a $P-$preserving ergodic invertible map and assume that $\theta = \sigma^2$ is also ergodic. Random transformations here are automorphisms of the torus $\mathbb{T}^2$ given by $F_\omega = \begin{pmatrix} 1 + n(\omega)n(\sigma\omega) & n(\sigma\omega) \\ n(\omega) & 1 \end{pmatrix}$ where $n = n(\omega)$ is a $\mathbb{Z}_+-$valued random variable with $\log n \in L^1(\omega, P)$. Denote by $[k_1, k_2, ...]$ the continued fraction $\cfrac{1}{k_1 + \cfrac{1}{k_2 + \cdots}}$ and set $a(\omega) = [n(\omega), n(\sigma\omega), n(\sigma^2\omega), ...]$, $b(\omega) = [n(\sigma^{-1}\omega), n(\sigma^{-2}\omega), ...]$. Define $\xi(\omega) = \begin{pmatrix} a(\omega) \\ -1 \end{pmatrix}$, $\eta(\omega) = \begin{pmatrix} 1 \\ b(\omega) \end{pmatrix}$, $\lambda(\omega) = a(\omega)a(\sigma\omega)$ and $\gamma(\omega) = (b(\sigma\omega)b(\sigma^2\omega))^{-1}$. Then $F_\omega\xi(\omega) = \lambda(\omega)\xi(\theta\omega)$, $F_\omega\eta(\omega) = \gamma(\omega)\eta(\theta\omega)$, $\lambda(\omega) < 1$, $\gamma(\omega) > 1$, and so, ξ and η span the contracting and expanding (in average) directions, respectively. Allowing also zero values of $n(\omega)$ one can achieve even that the angles between these directions may approach zero arbitrarily close. All these constructions fall into a more general class of random diffeomorphisms having in the tangent bundle random invariant (expanding in average) cone families (see [GK1]).

In the continuous time case the situation is more complicated and, essentially, no ergodic theory of random (spatially uniform) hyperbolic flows exists, as yet, which could provide constructions of equilibrium states via a thermodynamic formalism approach (cf. [GK1]). A successful theory should include natural perturbation models such as a random flow generated by a random vector field whose all realizations are close to a deterministic vector field generating an Anosov flow. Meanwhile, only simple examples can be dealt with. Consider, for instance, a random flow F_ω^t given by the equation $\frac{dF_\omega^t x}{dt} = q_{\theta^t\omega}(F_\omega^t x)B(F_\omega^t x)$ where θ^t is an

ergodic $P-$preserving flow on Ω, q_ω is a measurable family of smooth positive functions on a compact Riemannian manifold M, and B is a vector field generating a transitive Anosov flow f^t on M. Then F^t_ω is obtained from f^t by the random time change and both flows have the same orbits. Using a Markov partition for f^t one can represent F^t_ω as a suspension over a random subshift of finite type with a random ceiling function bounded away from zero and infinity (see [Ki6]).

Another model I have in mind is the case of expanding in average smooth random maps considered in [KK] which can be studied directly without a symbolic representation. Assume for simplicity that all X^ω's coincide with one compact connected $d-$dimensional C^2 Riemannian manifold M and all $F_\omega : M \to M$ are C^2 endomorphism of M such that $\log \|DF^{-1}_\omega\|, \log \|DF_\omega\| \in L^1(\Omega, P)$ and $\int \log \|DF^{-1}_\omega\| dP(\omega) < 0$ where DF is the differential of F and $\|\cdot\|$ is the supremum norm. The random RPF operator $\mathcal{L}^\omega_g$ is defined again by (2.3) and if $g'_\omega s$ are Hölder continuous, i.e. (2.4) is satisfied with an integrable $\log K_g$, then the random RPF theorem holds true yielding a random variable $\lambda_\omega > 0$, a function $h = h_\omega(x) > 0$ on $M \times \Omega$, and probability measures ν^ω on M satisfying (2.5) so that $\mu \in \mathcal{P}_P(X)$ with desintegrations μ^ω satisfying $d\mu^\omega = h_\omega d\nu^\omega$ is the unique equilibrium state for g. Both in this model and in the case of random Anosov (hyperbolic) diffeomorphisms there are relativized Sinai-Ruelle-Bowen measures μ_{SRB} having special properties which are equilibrium states for the functions $\varphi^u_\omega(x)$ equal minus logarithm of the Jacobian of either $D_x F_\omega$ (in the expanding case) or of the restriction of $D_x F_\omega$ to the random expanding subbundle (in the hyperbolic case).

3. Limit theorems for random transformations

In this section I shall formulate the LD, CLT, and LIL results for random transformations F_ω belonging to one of the specific classes considered in the previous section. Set $I_g(\nu) = Q_\tau(g) - \int g d\nu - h^{(\tau)}_\nu(\tau)$ if $\nu \in \mathcal{P}_P(X)$ and ν is $\tau-$invariant, while $I_g(\nu) = \infty$, otherwise, and put $J_{g,q}(r) = \inf\{I_g(\nu) : \int q d\nu = r\}$ if a $\nu \in \mathcal{P}_P(X)$ satisfying conditions in brackets exists and $J_{g,q}(r) = \infty$, otherwise.

3.1. Theorem. *(cf.[Ki2]) Suppose that Ω is a locally compact space. Let $\mu \in \mathcal{P}_P(X)$ with desintegrations $d\mu(x,\omega) = d\mu^\omega(x) dP(\omega)$ be the unique equilibrium state for a function g satisfying conditions of the corresponding RPF-theorem (i.e. (2.4) holds true). Set $\zeta^n_{x,\omega} = \frac{1}{n}\sum_{k=0}^{n-1} \delta_{\tau^k(x,\omega)}$, where δ_z is the Dirac measure at z, and $S_n q(x,\omega) = n \int q d\zeta^n_{x,\omega} = \sum_{k=0}^{n-1} q \circ \tau^k(x,\omega)$. Then for each bounded continuous function q and any numbers $r_1 < r_2$,*

$$-\inf_{r\in[r_1,r_2]} J_{g,q}(r) \geq \limsup_{n\to\infty} \frac{1}{n} \log \mu^\omega\{x \in X^\omega : \frac{1}{n} S_n q(x,\omega) \in [r_1, r_2]\} \tag{3.1}$$

$$\geq \liminf_{n\to\infty} \frac{1}{n} \log \mu^\omega\{x \in X^\omega : \frac{1}{n} S_n q(x,\omega) \in (r_1, r_2)\} \geq -\inf_{r\in(r_1,r_2)} J_{g,q}(r)$$

$P-$a.s. The large deviations estimates for occupational measures $\zeta^n_{x,\omega}$, i.e. the upper and lower bounds for the limits of $n^{-1} \log \mu^\omega\{x \in X^\omega : \zeta^n_{x,\omega} \in G\}$, (with G being a closed or open set of probability measures on $\mathbf{X} \times \Omega$) hold true, as well, with

the rate functional $I_g(\nu)$. In the case of random expanding in average transformations of a compact Riemannian manifold M or random Anosov diffeomorphisms the results remain true if μ^ω is replaced by the normalized Riemannian volume m on M and one takes $g_\omega(x) = \varphi^u_\omega(x)$.

Observe, that $I_g(\nu) = 0$ and $J_{g,q}(r) = 0$ if and only if $\nu = \mu$ and $r = \int q d\mu$. Therefore, (3.1) estimates large deviations from the ergodic theorem, i.e. it describes the decay of $\mu-$measure of points having irregular with respect to μ behavior.

In the case of random subshifts of finite type Theorem 3.1 can be modified to become a combinatorial statement on random networks (see [Ki5]). For $\alpha = (\alpha_0, ..., \alpha_n)$ with $a_{\alpha_i,\alpha_{i+1}}(\theta^i\omega) = 1\,\forall i = 0, ..., n-1$ set $C^\omega_\alpha = \{x \in X^\omega_A : x_i = \alpha_i\,\forall i = 0, 1, ...n\}$ which is called an $n-$cylinder set. Denote by $\Pi^\omega_n(a, b)$ the set of all $n-$cylinders $C^\omega_{\alpha_0,...,\alpha_n}$ with $\alpha_0 = a \in V(\omega)$ and $\alpha_n = b \in V(\theta^n\omega)$ and by $|R|$ the cardinality of a set R. Let $I(\nu) = I_0(\nu) = h^{(r)}_{\rm top}(\tau) - h^{(r)}_\nu(\tau)$ if $\nu \in \mathcal{P}_P(X)$ and ν is $\tau-$invariant, while $I(\nu) = \infty$, otherwise, and put $J_q(r) = J_{0,q}(r)$. Then for any bounded continuous function q, $a \in V(\omega)$, $b_n \in V(\theta^n\omega)$, $x_\alpha \in C^\omega_\alpha$, and numbers $r_1 < r_2$ with probability one as $n \to \infty$, $|\Pi^\omega_n(a, b_n)|^{-1}|\{C^\omega_\alpha \in \Pi^\omega_n(a, b_n) : n^{-1}(S_nq)(x_\alpha, \omega) \in (r_1, r_2)\}| \asymp \exp(-n \inf_{r\in(r_1,r_2)} J_q(r))$. Here $\asymp$ means that both sides of the formula have the same logarithmic asymptotical behavior in the sense of inequalities in (3.1). In particular, if I assign to each edge e of the network $\mathcal{N}(\omega)$ its length $l_\omega(e)$ and set $q_\omega(x) = l_\omega(x_0, x_1)$, which gives a continuous function, then this yields large deviations for the average length of paths with n vertices. The corresponding second level of large deviations estimates $|\Pi^\omega_n(a, b_n)|^{-1}|\{C^\omega_\alpha \in \Pi^\omega_n(a, b_n) : \zeta^n_{x_\alpha,\omega} \in G\}| \asymp \exp(-n \inf_{\nu\in G} I(\nu))$ for occupational measures holds true, as well.

Next, I formulate the CLT and the LIL from [Ki6]. Let μ be as in Theorem 3.1 and $\varphi = \varphi_\omega(x)$ satisfying $\int \varphi_\omega d\mu^\omega = 0$ be a Hölder continuous in x random function with an exponent $\kappa > 0$ and a random variable K_φ, i.e. φ satisfies (2.4). For a random variable $L = L(\omega)$ and a constant C set $Q_{L,C} = \{\omega : L(\omega) \le C\}$ and $k_{L,C}(\omega) = \min\{n : \theta^n\omega \in Q_{L,C}\}$. I say that the L, C integrability condition for φ holds true if $\int (\sum_{i=0}^{k_{L,C}-1}(\|\varphi\| + K_\varphi) \circ \theta^i)^2 dP < \infty$ where $\|\varphi\|_\omega = \sup_x |\varphi_\omega(x)|$.

3.2. THEOREM. *There exist a random variable $L = L(\omega)$ and a constant C (which can be written explicitly for specific models above) such that if the L, C integrability condition holds true then $P-$a.s. the limit $\sigma^2 = \lim_{n\to\infty} \frac{1}{n} \int (\sum_{j=0}^{n-1} \varphi_{\theta^j\omega} \circ F^j_\omega)^2 d\mu^\omega$ exists and for $P-$a.a. ω and any number a,*

$$\lim_{n\to\infty} \mu^\omega \left\{ x \in X^\omega : n^{-1/2} \sum_{i=0}^{n-1} (\varphi \circ \tau^i)(x, \omega) \le a \right\} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{u^2}{2\sigma^2}} du \tag{3.2}$$

where in the case $\sigma = 0$, the normal distribution in the right hand side of (3.2) should be understood as the Dirac measure at 0.

Assuming that $\sigma > 0$ the invariance principle for the LIL holds true. Namely, if $\zeta(t) = (2t \log\log t)^{1/2}$ and $\eta_n(t) = (\zeta(\sigma^2 n))^{-1}(\sum_{j=0}^{k-1} \varphi \circ \tau^j + (nt - k)\varphi \circ \tau^k)$ for $t \in [\frac{k}{n}, \frac{k+1}{n})$, $k = 0, 1, ..., n-1$ then $\mu-$a.s. the set of limit points in $C[0, 1]$ of

functions $\eta_n(t)$ *as* $n \to \infty$ *coincides with the set of absolutely continuous* $\eta \in C[0,1]$ *with* $\int_0^1 (\dot{\eta}(t))^2 dt \leq 1$.

The role of $L = L(\omega)$ emerging in Theorem 3.2 is to offset the nonuniformity in ω of the models above so that, for instance, F_ω^n will be uniformly expanding for $n \geq L(\omega)$ or, in the case of random subshifts of finite type, all matrices $A(\omega)A(\theta\omega)\cdots A(\theta^n\omega)$ will be positive for any $n \geq L(\omega)$. In addition, $L(\omega)$ bounds some parameters related to the functions g_ω and φ_ω appearing in Theorem 3.2.

Observe that Theorem 3.2 yields fiber-wise CLT and LIL for some deterministic skew product transformations. For instance, consider an expanding map of the 3-dimensional torus $\mathbb{T}^3 = \mathbb{T}^1 \times \mathbb{T}^2$ given by the formula $\tau(x,y) = (F_y x, \theta y)$ where θ is an ergodic automorphism of $\mathbb{T}^2$ and $F_y x = \gamma(y) + n(y)x \,(\mathrm{mod}\, 1)$ where $\gamma(y) \in \mathbb{R}$, $n(y) \in \mathbb{Z}_+$ are measurable functions with $0 < \int_{\mathbb{T}^2} \log n(y) dy < \infty$. Since both θ and F_y's preserve the Lebesgue measures (denoted Leb below) on $\mathbb{T}^2$ and on $\mathbb{T}^1$, respectively, I can view F_y's as "random" expanding maps of $\mathbb{T}^1$ with $\Omega = \mathbb{T}^2$, P =Leb, $M = \mathbb{T}^1$, and μ^y =Leb (which is a "random" Gibbs measure corresponding to the function $g_y = \log n(y)$). Theorem 3.2 yields now that for Leb-a.a.y, $\mathrm{Leb}\{x : \frac{1}{\sqrt{n}} \sum_{l=0}^{n-1} \varphi \circ \tau^l(x,y) \leq a\}$ converges as $n \to \infty$ to the right hand side of (3.2) and the corresponding LIL follows, as well.

4. Fractal dimensions of random sets

Any $x \in [0,1)$ can be represented in the form of a "random base expansion"

$$x = \sum_{i=0}^{\infty} \frac{x_i(\omega)}{\ell(\omega)\ell(\theta\omega)\cdots\ell(\theta^i\omega)},\ x_i(\omega) \in \{0,1,...,\ell(\theta^i\omega)-1\} \tag{4.1}$$

where, again, ℓ is a $\mathbb{Z}_+-$valued random variable satisfying $0 < \int \log \ell dP < \infty$. To make this representation unique one can forbid the tails of the form $x_i(\omega) = \ell(\theta^i\omega) - 1\, \forall i \geq n$. Identify 0 and 1 then $F_\omega x = \ell(\omega)x \,(\mathrm{mod}\, 1)$ can be considered as a random expanding transformation of the unit circle $\mathbb{T}^1$. If $\tau(x,\omega) = (F_\omega x, \theta\omega)$ is the skew product transformation and $\phi(x,\omega) = x_0(\omega)$ then

$$x_i(\omega) = (\phi \circ \tau^i)(x,\omega). \tag{4.2}$$

Observe that all F_ω preserve the Lebesgue measure m on $[0,1)$ and the corresponding measure $m \times P$ is $\tau-$invariant, has the (maximal) relativized entropy $\int \log \ell dP$, and it is the unique equilibrium state for the function $-\log \ell$. Moreover, it is ergodic, and so $m \times P-$a.s., $\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} x_i(\omega) = \frac{1}{2} \int (\ell - 1) dP$ assuming that the right hand side exists. In view of (4.2) and Theorem 3.1 one has the large deviations estimates for $m\{x : \frac{1}{n} \sum_{i=0}^{n-1} x_i(\omega) \in [r_1, r_2]\}$. Furthermore, by Theorem 3.2 $P-$a.s. for any number a, $\lim_{n\to\infty} m\{x : n^{-1/2} \sum_{i=0}^{n-1} (x_i(\omega) - \frac{1}{2}(\ell(\theta^i\omega) - 1)) \leq a\} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{u^2}{2\sigma^2}} du$. Moreover, σ can be computed here precisely since, when ω is fixed, $x_0(\omega), x_1(\omega), ...$ are independent random variables (with stationarily changing distributions) on the space $([0,1], m)$, which gives $\sigma^2 = \frac{1}{12} \int (\ell^2 - 1) dP$ provided the right hand side exists.

Consider the sequence space $X^\omega = \{x = (x_0, x_1, ...) : x_i \in \{0, 1, ..., \ell(\theta^i\omega) - 1\}\}$ and the map $\pi^\omega : [0,1) \to X^\omega$, $\pi^\omega(x) = (x_0(\omega), x_1(\omega), ...)$ then $F_\omega \pi^\omega = \pi^\omega f_\omega$ where f_ω is the left shift on X^ω. Thus π^ω is a semi-conjugacy and, in fact, this symbolic representation comes from the random Markov partition of $[0,1)$ into $\ell(\omega)$ equal subintervals.

A measurable set $G \subset \mathbb{T}^1 \times \Omega$ is $\tau-$invariant iff $F_\omega G^\omega = G^{\theta\omega}$ where $G^\omega = \{x : (x, \omega) \in G\}$. Such sets can be obtained, for instance, considering x whose expansion (4.1) does not contain certain prescribed digits which may be called random Cantor sets. Assuming that all G^ω are compact one has the following formula for their Hausdorff dimension (see [Ki4]),

$$HD(G^\omega) = \frac{h^{(r)}_{\rm top}(\tau, G)}{\int \log \ell dP} \ \ P - \text{a.s.} \tag{4.3}$$

where $h^{(r)}_{\rm top}(\tau, G)$ is the relativised tological entropy of τ restricted to G. Next, I consider another class of random invariant sets which are dense in $[0,1)$. Set $N^\omega_{kl}(x,n) = |\{j \ge 0, j < n : \ell(\theta^j\omega) = k, x_j(\omega) = l - 1\}|$ and $N^\omega_l(x,n) = \sum_{k \in \mathbb{Z}_+} N^\omega_{kl}(x,n)$ where $|\{\cdot\}|$ denotes the cardinality of a set $\{\cdot\}$. Let $r = (r_k, k \in \mathbb{Z}_+)$ be an infinite probability vector and $A = (a_{kl}, k, l \in \mathbb{Z}_+)$ be an infinite probability matrix such that $a_{kl} = 0$ unless $l \le k$. Define the sets $U^\omega_r = \left\{x \in [0,1) : \lim_{n\to\infty} \frac{1}{n} N^\omega_l(x,n) = r_l \ \forall l \in \mathbb{Z}_+\right\}$ (i.e. prescribing frequencies of digits) and $V^\omega_A = \left\{x \in [0,1) : \lim_{n\to\infty} \frac{1}{n} N^\omega_{kl}(x,n) = q_k a_{kl} \ \forall k, l \in \mathbb{Z}_+\right\}$ where $q_k = P\{\ell = k\}$.

4.1. THEOREM. *([Ki4]) With probability one,*

$$HD(V^\omega_A) = \frac{-\sum_{k\in\mathbb{Z}_+} q_k \sum_{l\le k} a_{kl} \log a_{kl}}{\int \log \ell dP} \stackrel{\rm def}{=} H_A, \tag{4.4}$$

and so $HD(V^\omega_A) = 1$ iff $a_{kl} = k^{-1}$ for all $l \le k$ and any $k \in \mathbb{Z}_+$ such that $q_k \ne 0$. In the last case with probability one V^ω_A has also the Lebesgue measure one. The sets U^ω_r have the Lebesgue measure one for P-a.a.ω iff $r_l = \sum_{k\ge l} q_k k^{-1}$ for all $l \in Z_+$ (which is a random version of Borel's normal number theorem). Furthermore, for P-a.a.ω, $HD(U^\omega_r) = \sup_{A\in\mathcal{A}_{qr}} H_A \stackrel{\rm def}{=} H$, where the supremum is taken over the set $\mathcal{A}_{qr}$ of all infinite probability matrices $A = (a_{kl})$ such that $a_{kl} = 0$ unless $l \le k$ and $qA = r$ with q and r considered as the row vectors. The set $\mathcal{A}_{qr}$ is nonempty iff $\sum_{l\in F} q_l \ge \sum_{l\in F} r_l$ for any filter $F \in \mathcal{F}$ in $\mathbb{Z}_+$, (i.e. if $l \in F$ and $l \le k$ then $k \in F$). If $\mathcal{A}_{qr}$ is empty then with probability one U^ω_r is empty too.

The expression in the numerator of the right hand side of (4.4) is the fiber entropy of certain random Bernoulli measure which emerges naturally in the proof. Computations of dimensions of different other invariant sets of random tranformations, as well as multidimensional generalizations, can be found in [Ki4].

5. "RANDOM" RANDOM WALKS ON GROUPS

Markov chains with random transition probabilities emerge directly from random subshifts of finite type taken with random Markov measures but also they are

closely related ideologically to random transformations (cf. [Ki6]). In this section I consider random walks with stationarily changing distributions on discrete groups and demonstrate how a relativized entropy like characteristic describes their asymptotic behavior.

Let G be a discrete group and $\mu^\omega, \omega \in \Omega$ be a measurable family of probability measures on G. Next, I consider the Markov chain X_n^ω with random transitions on G, which I call "random" random walk, with $n-$step transition probabilities

$$P^\omega(n, g_1, g_2) = \mu^\omega * \mu^{\theta\omega} * \cdots * \mu^{\theta^{n-1}\omega}(g_2 g_1^{-1}) \tag{5.1}$$

where $*$ denotes the usual convolution of measures on groups. Following [Rub] call a measurable in ω and x function h random harmonic if

$$\sum_{r\in G} P^\omega(1, g, r) h_{\theta\omega}(r) = \sum_{r\in G} h_{\theta\omega}(rg)\mu^\omega(r) = h_\omega(g). \tag{5.2}$$

The next natural goal is to describe spaces of random harmonic functions which is related to the asymptotic behavior of X_n^ω.

Suppose that h is random harmonic and $c(\omega) = \sup_x |h_\omega(x)| < \infty$. Since I assume that θ is ergodic it follows from (5.2) that c is constant $P-$a.s., and so h is bounded. Let e be the identity of G and P^ω be the path distribution of the Markov chain X_n^ω, $n \geq 0$, $X_0^\omega = e$. Is is easy to see that $h_{\theta^n\omega}(gX_n^\omega)$ is a martingale under P^ω, and so for all $g \in G$ and $P^\omega-$a.a. paths $\xi \in G^{\mathbb{Z}_+}$ the limit $\lim_{n\to\infty} h_{\theta^n\omega}(gX_n^\omega) = \varphi_\omega(g\xi)$ exists where $g\xi = (g\xi_0, g\xi_1, ...)$ for $\xi = (\xi_0, \xi_1, ...)$ determines the action of G on paths $\xi \in G^{\mathbb{Z}_+}$. Moreover, for any $g \in G$, $P-$a.a.ω, and $P^\omega-$a.a.ξ one has $\varphi_\omega(g\xi) = \varphi_{\theta\omega}(g\sigma\xi)$ where σ is the left shift. Let $\tau(\xi, \omega) = (\sigma\xi, \theta\omega)$ and $\mathcal{F}$ be the $\sigma-$algebra of $\tau-$invariant measurable sets from $G^{\mathbb{Z}_+} \times \Omega$. Set $\mathcal{F}^\omega = \{A^\omega = \{\xi : (\xi, \omega) \in A\} : A \in \mathcal{F}\}$ and let π_ω be the factorizing map of $(G^{\mathbb{Z}_+}, P^\omega)$ to the quotient space corresponding to the measurable partition attached to $\mathcal{F}^\omega$. Then one has a Poisson type representation $h_\omega(g) = \int \varphi_\omega \circ \pi_\omega dg\nu^\omega$ where $\nu^\omega = \pi_\omega P^\omega$ satisfying $\mu^\omega * \nu^{\theta\omega} = \nu^\omega$ is naturally to call a random harmonic measure.

For any probability measure η on G set $H(\eta) = -\sum_{g\in \mathrm{supp}\,\eta} \eta(g) \log \eta(g)$ and assume that $\int H(\mu^\omega) dP(\omega) < \infty$. Let $\mu_n^\omega = \mu^\omega * \mu^{\theta\omega} * \cdots \mu^{\theta^{n-1}\omega}$ and $\mathbf{h}_n^\omega = H(\mu_n^\omega)$. Then $\mathbf{h}_{n+m}^\omega \leq \mathbf{h}_n^\omega + \mathbf{h}_m^{\theta^n\omega}$ and by the subadditive ergodic theorem $P-$a.s. the limit, called the fiber (or relativized) Avez entropy, $\mathbf{h}(G, \mu) = \lim_{n\to\infty} \frac{1}{n}\mathbf{h}_n^\omega$ exists and it is not random. Let G^ω be the support of the measure $\sum_{n=1}^\infty 2^{-n}\mu_n^\omega$ and assume that $G^\omega = G$ $P-$a.s.

5.1. THEOREM. *(i) For $P-$a.a.ω $P^\omega-$a.s., $\lim_{n\to\infty} \frac{1}{n} \log \mu_n^\omega(X_n^\omega) = -\mathbf{h}(G, \mu)$; (ii) $\mathbf{h}(G, \mu) = 0$ iff there are no random bounded harmonic functions except $\mu-$a.s. constants (where $d\mu(\xi, \omega) = d\mu^\omega(\xi)dP(\omega)$).*

In some cases, for instance, when G is a free group, one can also obtain Hausdorff dimensions of random harmonic measures via $\mathbf{h}(G, \mu)$ and the speed of convergence of X_n^ω to infinity. Other results concerning this set up will appear in a forthcoming paper joint with Kaimanovich and Rubshtein. Results on "random" random walks on continuous groups, in particular, products of independent random matrices with stationarily changing distributions will appear elsewhere.

REFERENCES

[Ar] L. Arnold., *Random Dynamical Systems*, Springer Verlag, Berlin, 1998.

[Bo] T. Bogenschütz, *Entropy, pressure, and a variational principle for random dynamical systems*, Random&Comp.Dyn. **1** (1992), 99-116.

[BB] J. Bahnmüller and T. Bogenschütz, *A Margulis–Ruelle inequality for random dynamical systems*, Arch. Math. **64** (1995), 246-253.

[BG] T. Bogenschutz and V. M. Gundlach, *Ruelle's transfer operator for random subshifts of finite type*, Ergod. Th.& Dynam. Sys. **15** (1995), 413-447.

[BL] J.Bahnmüller and P.-D. Liu, *Characterization of measures satisfying Pesin's entropy formula for random dynamical systems*, J. Dynam. and Diff. Equat. **10** (1998).

[GK1] V. M. Gundlach and Y. Kifer, *Random hyperbolic systems*, Preprint, 1998.

[GK2] V. M. Gundlach and Y. Kifer, *Random sofic shifts*, Preprint, 1998.

[Ka] S. Kakutani, *Random ergodic theorems and Markoff processes with a stable distribution*, Proc. 2nd Berkeley Symp. on Math. Stat. and Probab., 1951, pp. 247-261.

[Ki1] Y. Kifer, *Ergodic Theory of Random Transformations*, Birkhäuser, Boston, 1986.

[Ki2] Y. Kifer, *Equilibrium states for random expanding transformations*, Random&Comp. Dynam. **1** (1992), 1-31.

[Ki3] Y.Kifer, *Multidimensional random subshifts of finite type and their large deviations*, Probab. Th. Rel. Fields **102** (1995), 223-248.

[Ki4] Y. Kifer, *Fractal dimensions and random transformations*, Trans. Amer. Math. Soc. **348** (1996), 2003-2038.

[Ki5] Y. Kifer, *Large deviations for paths and configurations counting*, Ergodic Theory of $\mathbb{Z}^d$-Actions (M, Pollicott and K. Schmidt, eds.), London Math. Soc. Lecture Note Series, vol. 228, Cambridge University Press, Cambridge, 1996, pp. 415-432.

[Ki6] Y. Kifer, *Limit theorems for random transformations and processes in random environments*, Trans.Amer.Math.Soc. **350** (1998), 1481-1518.

[KK] K. Khanin and Y. Kifer, *Thermodynamic formalism for random transformations and statistical mechanics*, in: Sinai's Moscow Seminar on Dynamical Systems (L.A.Bunimovich, B.M.Gurevich, Ya.B.Pesin, eds.), vol. 171, American Mathematical Society Translations–Series 2, 1996, pp. 107-140.

[Li] P.-D. Liu, *Random perturbations of Axiom A basic sets*, J. Stat. Phys. **90** (1998), 467-490.

[LQ] P.-D. Liu and M. Qian., *Smooth Ergodic Theory of Random Dynamical Systems*, Lecture Notes in Math., vol. 1606, Springer Verlag, Berlin, 1995.

[LW] F. Ledrappier and P. Walters, *A relativized variational principle for continuous transformations*, J. London Math. Soc. **(2) 16** (1977), 568-576.

[LY] F. Ledrappier and L.-S. Young, *Entropy formula for random transformations*, Probab. Th. Rel. Fields **80** (1988), 217-240.

[Rub] B.-Z. Rubshtein, *Convolutions of random measures on compact groups*, J. of Theoret. Probab. **8** (1995), 523-538.

[Rue] D. Ruelle, *Positivity of entropy production in the presence of a random thermostat*, J. Stat. Phys. **86** (1997), 935-951.

[Th] J.-P. Thouvenot, *Quelques proprietes des systemes dynamiques qui se decomposent en un produit de deux systemes dont l'un est un schema de Bernoulli*, Isr. J. Math. **21** (1975), 177-207.

[UN] S.M. Ulam and J. von Neumann, *Random ergodic theorems*, Bull. Amer. Math. Soc. **51** (1945), 660.

Yuri Kifer
Institute of Mathematics,
Hebrew University of Jerusalem,
Givat Ram, Jerusalem 91904,
ISRAEL

Elements of a Qualitative Theory of Hamiltonian PDEs

Sergei B. Kuksin

We discuss nonlinear Hamiltonian partial differential equations (PDEs) and consider the finite-volume case only. That is, we are concerned with equations for functions (or vector-functions) $u(t,x)$, where the space-variable x belongs to a bounded domain and the equations are supplemented by appropriate boundary conditions. We treat them as ordinary differential equations in infinite-dimensional function spaces formed by functions of x and assume that they can be written in the Hamiltonian form:

$$\dot u(t) = J\nabla H(u(t)). \tag{1}$$

Here J is an anti self-adjoint operator in the space of square-integrable functions, H is a hamiltonian of the equation and ∇H is its L_2-gradient (if H is a functional of the calculus of variations, then ∇H equals to its variational derivative). The equation (1) is Hamiltonian with respect to a symplectic structure, defined in the function space by the form α_2,

$$\alpha_2(\xi(x),\eta(x)) = \langle(-J)^{-1}\xi(x),\eta(x)\rangle_{L_2}.$$

Hamiltonian PDEs are of extreme physical importance since they describe processes without dissipation of energy: (usually) the system's energy equals the hamiltonian H and preserves due to the same trivial arguments as in the finite-dimensional case.

Below we discuss three groups of results concerning qualitative behaviour of Hamiltonian PDEs. We have selected them according to our own taste, the references are by no means complete.

1. Nearly integrable PDEs

Some of nonlinear Hamiltonian PDEs with one-dimensional space variable x in a finite segment, supplemented by appropriate boundary conditions, are known to be integrable. For example, the Korteweg - de Vries equation (KdV) under zero-meanvalue periodic boundary conditions:

$$\text{(KdV)}\qquad \dot u = \frac{\partial}{\partial x}(-u_{xx}+3u^2),\qquad x\in S^1=\mathbb{R}/2\pi\mathbb{Z},\quad \int_0^{2\pi} u(t,x)\,dx = 0,$$

and the Sine-Gordon equation (SG) under Dirichlet boundary conditions:

$$\text{(SG)} \qquad \ddot{u} = u_{xx} - A\sin Bu, \qquad u(t,0) = u(t,\pi) = 0,$$

where $A, B > 0$. We view the equations as dynamical systems in appropriate Sobolev spaces Z^s, formed by functions $u(x)$ which respect the boundary conditions. (For the KdV, Z^s is the Sobolev space $H_0^s(S^1)$, formed by zero-meanvalue functions on the circle S^1. For the SG equation Z^s is the Sobolev space formed by odd 2π-periodic functions – these functions vanish for $x = 0$ and $x = \pi$).

Integrability of the KdV equation manifests itself in the following properties of a dynamical system which the equation defines in the spaces Z^s, discovered twenty years ago by P.Lax and S.P.Novikov (see [DMN]): For $n = 1, 2, \dots$ the space $\cap Z^s$ contains a smooth $2n$-dimensional manifold $\mathcal{T}^{2n}$, invariant for the KdV-flow, such that:

a) restriction of the equation to $\mathcal{T}^{2n}$ defines a Liouville-Arnold integrable Hamiltonian system,

b) $\mathcal{T}^{2n} \subset \mathcal{T}^{2m}$ if $m > n$,

c) union of all manifolds $\mathcal{T}^{2n}$ is dense in each space Z^s.

For the SG equation everything is much the same but the manifolds $\mathcal{T}^{2n}$ have algebraic singularities and their union is only proven to be dense in the vicinity of the origin.

The invariant manifolds $\mathcal{T}^{2n}$ are filled with time-quasiperiodic solutions $u_n(t,x)$ (so-called *n-gap solutions*). An n-gap solution u_n depends on an n-dimensional action $p \in \mathbb{R}^n_+$ and on n-dimensional angle $q \in \mathbb{T}^n$: $u_n(t,x;p,q) = \Phi_n(W_p t + q, x, p)$. The function $\Phi_n(q,x,p)$ is analytic and can be explicitly written in terms of theta-functions (the Its-Matveev formula, see [DMN]); this is another manifestation of integrability of KdV and SG equations. The n-vector W_p is called the *frequency vector*. The union in q and t of the curves $u_n(t,\cdot;p,q)$ is a smooth invariant n-torus in the space $\cap Z^s$, called the *n-gap torus*.

1.1. THE PROBLEM OF PERSISTENCE. Since both KdV and SG equations do not arise in mathematical physics in their exact form (as, for example, Navier - Stokes equations do), but only present simplified forms of some real physical equations, then it is important to understand if the finite-gap solutions u_n have something to do with "real" equations. Assuming that a "real" equation is Hamiltonian, that (say) the KdV equation comprises its highest derivatives and that the equation is local[1] (i.e. it does not contain integral terms), we write it as

$$\dot{u} = \frac{\partial}{\partial x}(-u_{xx} + 3u^2 + \varepsilon \frac{\partial}{\partial u} h(u,x)), \tag{2}$$

where the function h is assumed to be analytic in u.

1.2. KAM FOR PDES. The question we posed in the previous section can be understood in the following way: Does a finite-gap solution $u_n(t,x)$ of the KdV equation persist as a time-quasiperiodic solution for equation (2) (i.e., does (2)

[1] this assumption is imposed only for simplicity

have a time-quasiperiodic solution u_n^ε, close to u_n)? The affirmative answer is given by the following KAM for PDEs theorem:

For most (in the sense of measure) values of the action p, the n-gap solution $u_n(t,x;p,q)$ for the KdV-equation persists as a time-quasiperiodic solution $u_n^\varepsilon(t,x;p,q)$ for equation (2). The solution u_n^ε is linearly stable. Its closure in any space Z^s forms an invariant smooth n-torus.

The persisted solutions u_n^ε have the form $u_n^\varepsilon(t,x;p,q) = \Phi_n^\varepsilon(W_p^\varepsilon t+q,\, x,\, p)$. The new frequency vector W_p^ε is $O(\varepsilon)$-close to W_p and the function Φ_n^ε is $O(\varepsilon^\rho)$-close to Φ_n for any $\rho < 1$. In particular, for most p the theta-formula for an n-gap solution with the corrected frequency vector gives the function $\Phi_n(W_p^\varepsilon t + q,\, x, p)$, which is *forever* $O(\varepsilon^\rho)$-close to an exact solution of (2). The corrected frequency vector is $W_p^\varepsilon = W_p + \varepsilon W_p^1 + o(\varepsilon)$, where W_p^1 equals to averaging along the corresponding n-gap torus of a vector-function, constructed in terms of a hamiltonian of the perturbation.

A union (in n, p and q) of all persisted solutions, treated as curves in a space Z^s, becomes dense in Z^s as $\varepsilon \to 0$.

Similar results hold for the perturbed SG equation:

$$u_{tt} = u_{xx} - A\sin Bu + \varepsilon g(u,x) = 0. \tag{3}$$

The differences are that, first, large-amplitude solutions both for SG equation and for (3) are not linearly stable and, second, we do not know if the persisted solutions jointly are asymptotically dense as $\varepsilon \to 0$.

The KAM-theorem for PDEs is an infinite-dimensional version of the classical finite-dimensional theorem due to Kolmogorov-Arnold-Moser. Essential difference is that in the finite-dimensional case persisted time-quasiperiodic solutions fill the phase-space up to a set of small measure, while in the PDE-case the solutions which persist *due to the theorem* jointly have zero measure (for any reasonable measure in the corresponding function space).

The theorem we discussed in this section applies to quasilinear perturbations of all "classical" integrable PDEs with one-dimensional space variable, including all equations from the KdV hierarchy, etc. It is based on an abstract infinite-dimensional KAM-theorem. For exact statements and proofs see [K1, K3, K4, P1].

It is unknown what happens to infinite-gap solutions (and the corresponding infinite-gap tori) under Hamiltonian perturbations of the integrable equations.

1.3. SMALL OSCILLATIONS. The persistence problem posed in section 1.1 admits another understanding: does a small-amplitude finite-gap solution for the KdV or for SG equation persist after we have perturbed the equation by a higher-order at zero term? The affirmative answer follows from the same abstract KAM-theorem which implies the results of the previous section, see [BoK]. In particular, since $\sin Bu = Bu - B^3u^3/6 + O(u^5)$, then most of small amplitude finite-gap solutions of the SG equation (with appropriate A and B) persist in the φ^4-equation:

$$u_{tt} = u_{xx} - mu + \gamma u^3, \quad m,\gamma > 0. \tag{φ^4}$$

Small solutions for this (and similar) equations can be also constructed treating (φ^4) as a perturbation of another integrable infinite-dimensional system, namely its Birkhoff normal form at zero, see [KP, P2] (also see [W] and the Introduction to [K2] for related results).

Nothing is known about small time-quasiperiodic solutions for the φ^4-equation with $m = 0$.

1.4. Closely related persistence problem arises when we examine an *equation (1) with a small nonlinearity* and with a linear part with pure imaginary spectrum in order to prove that time-quasiperiodic solutions of the linear equation persist in the nonlinear equation (1). If the linear equation depends on an additional finite-dimensional parameter in a non-degenerate way, then any its time-quasiperiodic solution persists in the nonlinear equation for most values of the parameter, provided that: 1) the space-variable x belongs to a finite segment and 2) the perturbed equation is quasi-linear (i.e., the nonlinear term of (1) contains less derivatives than its linear part). – This follows from the same abstract KAM-theorem as above, see [K2, P1].

Some years ago J.Bourgain [B1] developed another KAM-approach, originally proposed by Craig - Wayne in [CW], and successfully used it to study the persistence problem which we discuss in this subsection. The main advantage of this approach is that it applies to two-dimensional (in space) Schrödinger equation. A disadvantage is that it applies only to semilinear equations (i.e. to equations where the nonlinear term contains no derivatives).

We do not know what happens to invariant tori of an n-dimensional (in space) linear Schrödinger equation with $n \geq 3$ and of a linear wave equation with $n \geq 2$ under Hamiltonian perturbations.

1.5. AVERAGING THEOREMS. Due to the KAM-results presented in section 1.2, the perturbed KdV equation (2) contains invariant finite-dimensional tori, filled with linearly stable time-quasiperiodic solutions, and union of these tori is asymptotically dense in any space Z^s as $\varepsilon \to 0$. Hence, any solution of equation (2) with sufficiently small ε for long time stays close to some n-gap torus. This result does not specify the persistence time. For a *finite-dimensional* nearly-integrable system this time is known to grow at least as $\exp \varepsilon^{-a}$, $a > 0$ (Nekhoroshev's theorem). To obtain an analogy of this result for equation (2) is an intriguing open problem. What is known is a local theorem which applies to a class of nearly integrable PDEs and states that for solutions of these equations with small analytical initial data the persistence time is bigger than $C_M \varepsilon^{-M}$ for each M, see [Bam] (also see there references for related results concerning some parameter-depending equations with small nonlinearities).

2. SYMPLECTIC INVARIANTS AND GROMOV'S NON-SQUEEZING PROPERTY.

2.1. GIBBS MEASURE. Flow-maps $\{S_t\}$ of any Hamiltonian PDE (1) preserve the symplectic form α_2 (see the introduction), provided that they are C^1-smooth. For a finite-dimensional Hamiltonian system in the space $(\mathbb{R}^{2N}_{p,q}, dp \wedge dq)$ symplecticity of the flow-maps of a Hamiltonian vector-field yields that they preserve the

Lebesgue measure $dpdq$ as well as the Gibbs measure $\exp(-\mathcal{H}(p,q))dpdq$, where $\mathcal{H}$ is the hamiltonian. In an infinite-dimensional function space $\{u(x)\}$ a Lebesgue measure $du(\cdot)$ does not exist, but the Gibbs measure $\mu_H = \exp(-H(u(\cdot)))\, du(\cdot)$, where H is a hamiltonian of the PDE, often is well defined if $\dim x = 1$ or 2. Its construction is well known from the quantum field theory (see [GJ]). A difficulty is that the measure μ_H is supported by a space of functions of low smoothness. To prove invariance of μ_H, a flow of the equation (1) has to be proven to exist in the corresponding low-smoothness space and to possess some regular properties. This can be done for many one-dimensional and for some two-dimensional equations, see [B2, B5, MV] and references therein.

It is an open problem whether a non-integrable Hamiltonian PDE has an invariant measure, supported by smooth functions (this measure should not be supported by a trivial invariant set like a periodic trajectory of the equation). This problem is closely related to the following question: is it true that high Sobolev norms of typical solutions for a non-integrable Hamiltonian PDE grow with time unboundedly, see [B3, B4].

2.2. SYMPLECTIC CAPACITY. The Gibbs measure μ_H corresponds to a subset of the function phase-space of a Hamiltonian PDE a flow-invariant quantity, namely its measure. This is not a unique invariant characteristic of subsets. Existence of another symplectic invariant for finite-dimensional Hamiltonian systems, called *symplectic capacity*, follows from Gromov's non-squeezing theorem (or can be constructed independently to prove the theorem), see in [HZ]. To discuss a version of this invariant applicable to (1), we need a notion of a *Darboux phase-space* Z_D for this equation[2] : Z_D is a Hilbert space which admits an orthonormal Hilbert basis $\{\varphi_j \mid j \in \mathbb{Z}_0\}$ ($\mathbb{Z}_0$ is the set of non-zero integers), which is a Darboux basis for the equation's symplectic structure, i.e., $\alpha_2[\varphi_j, \varphi_{-k}] = \delta_{j,k}$ for any $j \in \mathbb{N}$ and for all k.

EXAMPLES. 1) For the KdV equation (and its perturbation (2)), Z_D is the Sobolev space $H_0^{-1/2}(S^1)$. 2) A nonlinear wave equation

$$\ddot{u} - \delta\triangle u + mu + f(u,x) = 0, \quad u = u(t,x),\ x \in \mathbb{T}^n,$$

where $m > 0$ and f is a smooth function, can be written in the following Hamiltonian form:

$$\dot{u} = -Lw, \ \ \dot{w} = Lu + L^{-1}f(u,x), \tag{4}$$

where $L = (-\delta\triangle + m)^{1/2}$. For this equation $Z_D = Z^{1/2} = H^{1/2}(\mathbb{T}^n) \times H^{1/2}(\mathbb{T}^n)$ (see [K5, K6]). 3) If $f = 0$ (so the equation (4) is linear), then any space Z^s is a Darboux space. (On the contrary, for a typical *nonlinear* equation (4) a space Z^s with $s \geq 5$ is not Darboux. It is plausible that $Z^{1/2}$ is the unique Darboux space).

Let Z_D be a Darboux space for a symplectic form α_2 and $\{\varphi_j \mid j \in \mathbb{Z}_0\}$ be its basis as above. A map c which corresponds to an open subset $O \subset Z_D$ a number $c(O) \in [0, \infty]$ is called a (symplectic) capacity if

[2] in fact, for its symplectic structure.

α) c is translationary invariant, i.e., $c(O) = c(O + \xi)$ for $\xi \in Z_D$; β) c is monotonic, i.e. a bigger set has a bigger capacity; γ) c is 2-homogeneous, i.e. $c(\tau O) = \tau^2 c(O)$; δ) $c(B_r) = c(\Pi_r^{(k)}) = \pi r^2$, where B_r is the r-ball in Z_D, centered at the origin, and $\Pi_r^{(k)}$ is the cylinder formed by all vectors $\sum z_l \varphi_l$ such that $z_k^2 + z_{-k}^2 \le r^2$.

A finite-dimensional symplectic space $(\mathbb{R}^{2n}_{p,q}, dp \wedge dq)$ admits a symplectic capacity, invariant for symplectomorphisms [HZ]. A Darboux space Z_D also admits one. This capacity is invariant for flow-maps $\{S_t\}$ of a Hamiltonian equation (1), provided that

$$S_t = \text{linear operator} + \text{compact smooth operator}, \tag{5}$$

where the linear operator is a direct sum of rotations in the planes, spanned by the vectors φ_j and φ_{-j}, $j = 1, 2, \dots$ (see [K5]).

The assumption (5) is met by the nonlinear wave equation (4) if $n = 1$ and $f(u, x)$ has a polynomial growth in u, or $n = 2, 3$ and f as a function of u is a polynomial of a sufficiently low degree, see [K5,K6] and [B4].

The symplectic capacity is an invariant of the flow of a Hamiltonian PDE in a function space of low smoothness, as well as the Gibbs measure. An essential difference between these two invariants is that the former is constructed in terms of the equation's symplectic structure, while the latter – in terms of its hamiltonian (the same is true for the corresponding function spaces, so usually they are different).

An immediate consequence of existence of a symplectic capacity is that the flow-maps $\{S_t\}$, satisfying (5), can not squeeze a ball in a Darboux space Z_D to a cylinder of a smaller radius[3]; cf. the properties α), β) and δ). This is Gromov's non-squeezing property.

On the contrary, the squeezing (and a closely related pulling-through phenomenon, see below) both are possible (and are typical under some circumstances) if we consider the equation in a function space of high smoothness, i.e. study its classical solutions rather than generalised ones. In particular, the flow $\{S_t\}$ of equation (4) in a Sobolev space Z^s, $s \ge 5$, squeezes a typical ball of a radius of order one to a cylinder $\Pi_\rho^{(k)}$ with $\rho \sim (\ln \delta^{-1})^{-1}$, provided that the nonlinearity f is also typical,[4] see [K6].

3. Small-dispersion/dissipation equations

Let us consider the following class of PDEs:

$$\begin{aligned}\langle\text{non-linear homogeneous Hamiltonian equation}\rangle + \langle\delta_1\text{-small linear damping}\rangle \\ + \langle\delta_2\text{-small linear dispertion}\rangle = \zeta(t, x),\end{aligned} \tag{6}$$

where $\delta_1 \ge 0$, $\delta_2 \ge 0$ and $\delta := \sqrt{\delta_1^2 + \delta_2^2} > 0$. If $\delta_1 = 0$, then this equation is Hamiltonian. Still, the most important are equations with $\delta_1 > 0$ since they describe turbulence in different physical media.

[3] It is unknown if the assumption (5) is superfluous and can be dropped.

[4] Clearly, $\rho \ll 1$ if $\delta \ll 1$. So Gromov's property fails in this space.

The Navier-Stokes (NS) equations have the form (6) with the Euler equations for the homogeneous Hamiltonian equation and with $\delta_1 > 0$, $\delta_2 = 0$. Another good example of equation (6) is given by the damped/driven nonlinear Schrödinger equation:

$$\dot{u} - \delta_1 \triangle u + i\delta_2 \triangle u - i|u|^{2p}u = \zeta(t,x), \quad p \in \mathbb{N},\ \delta_1, \delta_2 \geq 0, \tag{7}$$

which we shall consider for $x \in \mathbb{R}^n$, $n \leq 3$, under the odd periodic boundary conditions:

$$u(t,x) = u(t, x_1, \dots, x_j + 2, \dots) = -u(t, x_1, \dots, -x_j, \dots) \quad \forall j$$

(they imply that $u(t,x)$ vanishes at the boundary of the cube of half-periods $\{0 \leq x_j \leq 1\}$). It is known that (7) has a unique smooth in x solution for any smooth odd periodic initial data $u(0,x) = u_0(x)$ (and for any continuous in t, smooth odd periodic in x function ζ). We shall discuss qualitative behaviour of solutions for equation (7) in the turbulent limit, i.e. when $\delta \ll 1$. We shall state results for equation (7), using some terminology which comes from the hydrodynamical turbulence, i.e. from the NS equations.

3.1. ESSENTIAL PART OF A PHASE-SPACE. Let us first consider equation (7) with $\zeta = 0$, supplemented by an order-one initial condition

$$u|_{t=0} = u_0(x) \in C^\infty, \quad |u_0|_{L_\infty} = U, \tag{8}$$

$U \sim 1$. Due to a trivial a priori estimate, L_2-norm in x of a solution u decays with t at least as $\exp - \delta_1 t$. Hence, the solution practically vanishes by a time $\gg \delta_1^{-1}$. We are interested in its behaviour for $0 \leq t \leq \delta^{-a}$ with $0 < a \leq 1$.

Denoting by $|u|_m$ the C^m-norm of a function $u(x)$, we define the essential part of the smooth phase-space of equation $(7)|_{\zeta=0}$ (with respect to the C^m-norm, $m \geq 2$) as

$$\mathfrak{A}_m = \{u(x) \in C^\infty \mid u \text{ is odd periodic and } |u|_0^{2m\varkappa+1} < K_m \delta^{m\varkappa} |u|_m\}.$$

Here $\varkappa$ is any fixed number $< 1/3$ and $K_m = K_m(\varkappa)$ is some specific constant. This set is formed by fast oscillating functions since $|u|_m \gg |u|_0$ for any $u \in \mathfrak{A}_m$ if $\|u\|_0 \gtrsim 1$ (when $\delta \ll 1$). The set looks like a narrow tube with respect to the C^m-norm since its intersection with a ball $\{|u|_m \leq R\}$ is contained in the narrow cylinder $\Pi_\rho^{(k)}$, formed by complex functions $u = \sum u_s e^{\pi i s \cdot x}$ such that $|u_k| < \rho$, where $\rho = C_m \delta^{1/2 + O(m^{-1})} R^{O(m^{-1})}$.

The set $\mathfrak{A}_m$ is important to understand dynamics of the equation (7) since: *by the time $C_m \delta^{-1}$ the flow of equation $(7)|_{\zeta=0}$ will pull the whole space of smooth odd periodic functions through $\mathfrak{A}_m$.* This pull-through phenomenon can be specified: *a solution u for $(7)_{\zeta=0}$, (8) will visit the set $\mathfrak{A}_m$ by the time $\delta^{-1/3} U^{-4/3}$. By the moment of a first entry to $\mathfrak{A}_m$ the solution will change its supremum-norm no more than twice.*

Hence, by the time $\delta^{-1/3}$ any solution $u(t,x)$ for $(7)_{\zeta=0}, (8)_{U=1}$ will make its C^m-norm as big as $C_m \delta^{-m\varkappa}$.

Equation (7) with $\zeta = 0, \delta_1 = 0$ takes the Hamiltonian form

$$\dot{u} + i\delta\triangle u - i|u|^{2p}u = 0. \tag{7'}$$

It has two integrals of motion: the hamiltonian and the L_2-norm $|u(t,\cdot)|_{L_2}$. Since $|u(t,\cdot)|_{L_\infty} \geq |u(t,\cdot)|_{L_2} = \text{const}$, then any non-zero solution for (7′) will visit $\mathfrak{A}_m$ during any time-interval longer than $\delta^{-1/3}|u|_{L_2}^{-4/3}$. I.e., $\mathfrak{A}_m$ is a recursion subset for this Hamiltonian PDE.

3.2. BOUNDS FOR AVERAGED HIGH NORMS. It turns out that since C^m-norms of a solution for $(7)_{\zeta=0}, (8)_{U=1}$ become big at least once, then they are big at the average; hence, its Sobolev norms are big at the average as well:

$$\delta^a \int_0^{\delta^{-a}} \|u(t,\cdot)\|_m^2 dt \geq C_m \delta^{-2m\varkappa_m}. \tag{9}$$

Here $a \geq 1/3$, $\varkappa_m = \varkappa_m(a) \nearrow 1/3$ and $\|\cdot\|_m$ stands for the norm in the Sobolev space of odd periodic functions. This estimate is essentially nonlinear since it obviously fails if $p = 0$.

The norms of the solution u satisfy usual upper estimates: if $\delta_2 = 0$ and $\zeta = 0$, then

$$\delta_a \int_0^{\delta^{-a}} \|u(t,\cdot)\|_m^2 dt \leq C'_m \delta^{-m}, \tag{10}$$

where the constants C'_m depend on C^m-norms of the initial condition u_0. We stress that the exponents for δ in the r.h.s.'s of (9) and (10) are universal: they do not depend on the nonlinearity $|u|^{2p}u$, the dimension n and the initial condition u_0.

Estimates similar to (9), (10) remain true for solutions of equation (7) with non-zero forcing ζ if we assume that $\zeta = \zeta^\omega(t,x)$ is a random field, smooth odd periodic in x and stationary in t (such equations are believed to present right mathematical description of physical turbulence, see in [EKMS, K8]): If $u^\omega(t,x)$ is a solution for (7) with, say, zero initial condition at $t = 0$ and $\langle\|u\|_m^2\rangle$ is its averaged squared Sobolev norm, $\langle\|u\|_m^2\rangle = \delta^a \int_0^{\delta^{-a}} \mathbf{E}\|u(t,\cdot)\|_m^2 dt$, then

$$C_m^{-1}\delta^{-2m\nu_m} \leq \langle\|u\|_m^2\rangle \leq C_m\delta^{-2m\mu_m} \quad \text{if} \quad a \geq 1, \tag{11}$$

where $\mu_m \nearrow B < \infty$ and $\nu_m \nearrow A > 0$. Moreover, we know that $\frac{3}{17} < A$, $B \leq \frac{3}{2}$ and that (11) remains true if in the definition of $\langle\|u\|_m^2\rangle$ we replace the time-segment $[0, \delta^{-a}]$ by any segment in $[0,\infty)$, longer than δ^{-a}.

A popular mathematical idealisation of the physically correct forcings ζ as above is given by a random field ζ which is white noise in time [EKMS]. For forcings like that the estimates (11) hold with $A = \frac{1}{2}$, $B = 1$.

An important feature of turbulent behaviour of a solution $u_\delta^\omega(t,x)$ is a short size of its space-scale l_x (see e.g. [LL], § 33 and [CDT]). Defining the space-scale as $l_x = \delta^\gamma$, where

$$\gamma = \gamma(u^\omega) = \liminf_{m\to\infty} \liminf_{\delta\to 0} \frac{\ln\langle\|u_\delta\|_m^2\rangle^{1/2m}}{\ln\delta^{-1}}$$

(see [K8]), we get from (11) that $A \leq \gamma \leq B$.

3.3. ASYMPTOTICAL SPECTRAL PROPERTIES OF SOLUTIONS AND THE KOLMOGOROV - OBUKHOV LAW. The estimates for the space-scale l_x of a solution $u^\omega(t,x)$, discussed above, characterise its infinitesimal in x behaviour. Arguments of Tauberian kind transform these estimates to information on asymptotical as $s \to \infty$ behaviour of Fourier coefficients $\hat{u}^\omega_s(t)$ of the solution.[5] To present it we denote by E_s the averaged squared Fourier coefficient $\hat{u}_s$, $E_s = \delta^a \int_0^{\delta^{-a}} \mathbf{E}|\hat{u}^\omega_s(t)|^2\,dt$. (We remark that if u^ω was a space-periodic solution for the NS equations, then E_s would be the energy of the fluid, corresponding to the wave vector s).

The numbers E_s obey the following asymptotic, which hold for any $\varepsilon > 0$ with A, B and γ as in the previous section:

1. $E_s = o(|s|^{-M})$ for $|s| \geq \delta^{-B-\varepsilon}$ with every M, if δ is sufficiently small. If $|s| \geq \delta^{-\gamma-\varepsilon}$, then the same holds true for all δ from an appropriate sequence $\{\delta_j \searrow 0\}$.

2. There exist $c(\varepsilon)$ and $C(\varepsilon)$ such that

$$\delta^c \leq |\mathcal{A}_\varepsilon|^{-1} \sum_{s \in \mathcal{A}_\varepsilon} E_s \leq \delta^C,$$

where $\mathcal{A}_\varepsilon = \{\delta^{-A+\varepsilon} \leq |s| \leq \delta^{-B-\varepsilon}\}$, for all small δ. The same holds true for the smaller set $\mathcal{A}_\varepsilon = \{\delta^{-\gamma+\varepsilon} \leq |s| \leq \delta^{-\gamma-\varepsilon}\}$ with appropriate exponents $c(\varepsilon)$ and $C(\varepsilon)$, for all δ from a sequence $\{\delta_j \searrow 0\}$.

The heuristic Kolmogorov - Obukhov (K-O) law (see [LL], § 33) states that the energy E_s of a wave-vector s is $o(|s|^{-M})$ for every M if $|s| > \delta^{-\gamma^K}$, and

$$\frac{1}{C} \sum_{r \leq |s| \leq r+C} E_s \sim \text{const} \cdot r^\theta \quad \text{for} \quad \delta^{-\gamma^0} < r < \delta^{-\gamma^K}.$$

The inverse threshold δ^{γ^K} is called *Kolmogorov's inner scale* of the turbulent flow. For 3-dimensional NS equations the exponent $\theta = 5/3$ and $\gamma^K = 3/4$, see [LL].

The properties 1 and 2 of a solution $u^\omega(t,x)$ for (7) present a weak form of the K-O law. In particular, if any solution u^ω for (7) satisfies the K-O law, then γ^K must equal the exponent $\gamma(u^\omega)$. Consequently, γ^K must meet the estimate $A \leq \gamma^K \leq B$. It is curious to note that for the forcing $\zeta^\omega(t,x)$ which is white noise in time, the results of section 3.2 imply the bounds $\frac{1}{2} \leq \gamma^K \leq 1$ which remarkably agree with the value $\gamma^K = 3/4$, prescribed by K-O for the 3-dimensional hydrodynamic turbulence.

The property 1 shows that the Fourier modes $\hat{u}^\omega_s e^{\pi i s \cdot x}$ with $|s| > \delta^{-\gamma}$ can be ignored when a solution u is calculated numerically, while the modes with $|s| < \delta^{-\gamma}$ are essential. Hence, a numerical scheme to calculate u has to have dimension of order $\delta^{-2\gamma n}$. This is a very big number since δ corresponding to a turbulent regime is very small (for turbulence in water and in air it is as small as $10^{-7} - 10^{-4}$). – This is why it is so difficult to study turbulence numerically.

Proofs of the results presented in sections 3.1-3.3 see in [K7, K8]. See [EKMS] for the turbulence-limit $\delta \to 0$ in a randomly forced Burgers equation.

[5] We write $u^\omega(t,x)$ as $\sum_{s \in \mathbb{Z}^n} \hat{u}^\omega_s(t) e^{\pi i s \cdot x}$.

REFERENCES

[B1] Bourgain J., *Quasi-periodic solutions of Hamiltonian perturbations for 2D linear Schrödinger equation*, Annals of Mathematics **145** (1998).

[B2] Bourgain J., *Periodic nonlinear Schrödinger equation and invariant measures*, Comm. Math. Phys. **166** (1994), 1-26.

[B3] Bourgain J., *Aspects of long time behaviour of solutions of nonlinear Hamiltonian evolution equations*, Geometric and Functional Analysis **5** (1995), 105-140.

[B4] Bourgain J., *On the growth in time of higher Sobolev norms of smooth solutions of Hamiltonian PDE*, International Mathematics Research Notes (1996), 277-304.

[B5] Bourgain J., *Invariant measures for the 2D-defocusing nonlinear Schrödinger equation*, Comm. Math. Phys. **176** (1996), 421-445.

[Bam] Bambusi D., *Nekhoroshev theorem for small amplitude solutions in NLS equations*, Math. Z. (to appear).

[BoK] Bobenko A.I, Kuksin S.B., *The nonlinear Klein-Gordon equation on an interval as a perturbed Sine-Gordon equation*, Comment. Math. Helv. **70** (1995), 63-112.

[CDT] Constantin P., Doering Ch., Titi E., *Rigorous estimates of small scales in turbulent flow*, J. of Math. Physics **37** (1996), 6142-6156.

[CW] Craig W., Wayne C.E., *Newton's method and periodic solutions of nonlinear wave equation*, Comm. Pure. Appl. Math. **46** (1993), 1409–1501.

[DMN] Dubrovin B.A., Matveev V.F., Novikov S.P., *Nonlinear equations of Korteweg-de Vries type, finite zone linear operators, and Abelian varieties*, Uspekhi Mat. Nauk **31:1** (1976), 55-136; English transl. in Russ. Math. Surv. **31:1** (1976).

[EKMS] E W., Khanin K., Mazel A., Sinai Ya., *Invariant measures for 1-D Burgers equation with stochastic forcing* (to appear).

[GJ] Glimm J., Jaffe A., *Quantum physics: a functional integral point of view, 2nd ed.*, Springer, 1987.

[HZ] Hofer H., Zehnder E., *Symplectic invariants and Hamiltonian dynamics*, Birkhäuser, 1994.

[K1] Kuksin S.B., *Perturbation theory for quasiperiodic solutions of infinite-dimensional Hamiltonian systems, and its applications to the Korteweg-de Vries equation*, Matem. Sbornik **136(178):3** (1988); English transl in Math. USSR Sbornik **64** (1989), 397-413.

[K2] Kuksin S.B., *Nearly integrable infinite-dimensional Hamiltonian systems*, Lecture Notes in Math. 1556 (1993), Springer.

[K3] Kuksin S.B., *A KAM-theorem for equations of the Korteweg - de Vries type*, Rev. Math. & Math. Phys. **10:3** (1998), 1-64.

[K4] Kuksin S.B., *Analysis of Hamiltonian PDEs*, Preprint of a book to be accepted by Oxford University Press.

[K5] Kuksin S.B., *Infinite-dimensional symplectic capacities and a squeezing theorem for Hamiltonian PDEs*, Comm. Math. Phys. **167** (1995), 531-552.

[K6] Kuksin S. B., *On squeezing and flow of energy for nonlinear wave equations*, Geometric and Functional Analysis **5** (1995), 668-701.

[K7] Kuksin S. B., *On turbulence in nonlinear Schrödinger equations*, Geometric and Functional Analysis **7** (1997), 783-822.

[K8] Kuksin S. B., *Spectral properties of solutions for nonlinear PDEs in the turbulent regime*, Geometric and Functional Analysis (to appear).

[KP] Kuksin S.B., Pöschel J., *Invariant Cantor manifolds of quasi-periodic oscillations for a nonlinear Schrödinger equation*, Annals of Mathematics **143** (1996), 149-179.

[MV] McKean H., Vaninsky K., *Statistical mechanics of nonlinear wave equations*, in the "Trends and perspectives in applied mathematics", collection of articles dedicated to Fritz John, 1994.

[P1] Pöschel J., *A KAM-theorem for some nonlinear PDEs*, Ann. Scuola Norm. Sup. Pisa, Cl. Sci., IV Ser. 15 **23** (1996), 119-148.

[P2] Pöschel J., *Quasi-periodic solutions for a nonlinear wave equation*, Comment. Math. Helvetici **71** (1996), 269-296.

[W] Wayne, C.E., *Periodic and quasi-periodic solutions of nonlinear wave equations via KAM theory*, Commun. Math. Phys. **127** (1990), 479-528.

Sergei B. Kuksin
Department of Mathematics
Heriot-Watt University
Riccarton
EH14 4AS Edinburgh. UK
and
Steklov Institute
Moscow
S.B.Kuksin@ma.hw.ac.uk

Counterexamples to the Seifert Conjecture

To my Son Greg

Krystyna Kuperberg[1]

Abstract. Since H. Seifert proved in 1950 the existence of a periodic orbit for a vector field on the 3-dimensional sphere S^3 which forms small angles with the fibers of the Hopf fibration, several examples of aperiodic vector fields on S^3 have been produced as well as results showing that in some situations a compact orbit must exists. This paper surveys presently known types of vector fields without periodic orbits on S^3 and on other manifolds.

1991 Mathematics Subject Classification: Primary 58F25; Secondary 57R25, 35B10, 58F18
Keywords and Phrases: dynamical system, plug, periodic orbit, minimal set, PL foliation

1 Introduction: The Seifert Conjecture.

A *dynamical system* or a *flow* on a metric space X is a *topological group action* of the additive group of reals $\mathbb{R}$ on X, or equivalently a continuous map $\Phi : \mathbb{R} \times X \to X$ such that $\Phi(0, p) = p$ and $\Phi(t + s, p) = \Phi(s, \Phi(t, p))$. If M is a smooth or real analytic manifold and Φ is differentiable, then $\frac{d\Phi}{dt}|_{t=0}$ is the *vector field of* Φ and is in the same smoothness category as Φ. By a standard integration theorem, a C^1 vector field on a closed manifold can be integrated to produce a corresponding dynamical system. A *trajectory* or an *orbit* of a point p is the image of $\Phi(\mathbb{R} \times \{p\})$ in X. A compact trajectory is *periodic*: either consisting of a *fixed point* or homeomorphic to S^1. A dynamical system, or equivalently a vector field, is *aperiodic* if it contains no compact trajectories. A non-compact trajectory is a one-to-one image of $\mathbb{R}$. A compact non-empty set A is *minimal*, if A is the union of trajectories, and no proper subset of A has these properties. A compact orbit is an example of a minimal set. A minimal set is always the closure of a trajectory, but

[1]The author was supported in part by the NSF grants DMS-9401408 and DMS-9704558.

not every set containing a trajectory as a dense subset is minimal. Throughout this paper, it is assumed that all considered vector fields are non-singular, or equivalently, that the dynamical systems possess no fixed points.

Any closed 3-manifold has Euler characteristic zero and hence admits a non-singular vector field. The Hopf fibration of the 3-dimensional sphere S^3 yields a dynamical system on S^3 whose every trajectory is circular. A small perturbation can easily eliminate all but one periodic orbit. In 1950, H. Seifert [34] proved the following:

THEOREM 1 *Suppose that $\mathcal{V}$ is a continuous vector field on S^3 satisfying the uniqueness of solution condition. Then there is an $\epsilon > 0$ such that if the vectors of $\mathcal{V}$ form angles smaller than ϵ with the fibers of the Hopf fibration, then $\mathcal{V}$ has a least one periodic solution.*

Subsequently, Seifert asked whether every dynamical system on S^3 has a periodic trajectory. The conjecture that the answer is "yes," under the natural C^1 differentiability assumption, became known as the Seifert conjecture. Further developments resulted in a stronger statement of the problem, see [39] and [31]:

A MODIFIED SEIFERT CONJECTURE: Every C^1 dynamical system on S^3 possesses a minimal set of covering dimension 1.

The table below illustrates the existing counterexamples to the Seifert conjecture and the modified Seifert conjecture:

flows on S^3	not volume preserving	volume preserving
discrete circular trajectories	C^ω (F. W. Wilson)	C^∞ (G. Kuperberg)
aperiodic, 1-dimensional minimal sets	C^1 (P. A. Schweitzer) $C^{3-\epsilon}$ (J. Harrison)	C^1 (G. Kuperberg)
2-dimensional minimal sets	C^ω (G. Kuperberg, K. Kuperberg)	—

2 DISCRETE CLOSED ORBITS AND THE APPLICATION OF PLUGS.

The presently known examples of aperiodic dynamical systems on S^3 are based on constructions of aperiodic plugs which are used to locally modify a given dynamical system with discrete periodic orbits in order to break these orbits without forming new ones. An example of an n-dimensional C^r plug, $1 \leq r \leq \infty$, can be described as follows. Let $\mathcal{V}$ be a constant vector field on $\mathbb{R}^n$ parallel to a given line L. Suppose that F is an $(n-1)$-dimensional compact connected manifold with boundary allowing an embedding of the Cartesian product of F and the interval I, $F \times I$, in $\mathbb{R}^n$ in such a way that for $p \in F$, $\{p\} \times I$ is a straight line segment parallel to L. A *plug* is a C^r vector field $\mathcal{F}$ on $F \times I$ which coincides with $\mathcal{V}$ in a neighborhood of the boundary $\partial(F \times I)$ and satisfies two additional conditions: 1. there is a trajectory whose positive limit set is inside the set $F \times I$ (*trapped trajectory*); 2. if a trajectory

of $\mathcal{F}$ passes through $F \times I$, then it contains a pair of points $(p, 0)$ and $(p, 1)$ (*matched ends*). The definition of a *twisted plug* is similar, but the requirement on $\mathcal{F}$ on the side boundary is relaxed so that $\mathcal{F}$ is tangent to $(\partial F) \times I$ and there are no minimal sets in the side boundary. A chart in a manifold containing a set homeomorphic to $F \times I$ on which the vector field is conjugate to a constant vector field parallel to the fiber I is replaced by an aperiodic plug matching the end points of its trajectories to the end points of trajectories in the chart. If a segment of a circular orbit is replaced by a trapped orbit, then the periodicity is removed.

Plugs are also defined for real analytic vector fields, piecewise linear foliations, and higher dimensional foliations, see [31], [20] and [21]. As remarked by W. Thurston [35], by the Morrey-Grauert theorem asserting that two analytic manifolds which are diffeomorphic are analytically diffeomorphic, real analytic plugs can be used to alter vector fields and foliations on real analytic manifolds.

One of the two basic properties of a plug, the "trapped trajectory," dates to a classical example of a fixed point free homeomorphism of an acyclic compact subset of $\mathbb{R}^3$ given by K. Borsuk [1] in 1935. In 1966, in a fundamental paper [39] F. W. Wilson introduced a special kind of symmetry of vector fields which implies the other important property of plugs, the "matched ends." He proved the following:

Theorem 2 (Wilson) *Every C^∞ n-manifold without boundary, of Euler characteristic zero or non-compact, admits a C^∞ dynamical system with a discrete collection of minimal sets. Each of the minimal sets is an $(n-2)$-torus $S^1 \times \cdots \times S^1$, and every trajectory originates (resp. limits) on one of these tori.*

Wilson's theorem is actually valid in the C^ω category and it implies that a C^ω analogue to the Seifert conjecture for higher dimensional spheres of odd dimensions does not hold. The minimal sets are of codimension 2 and hence the resulting flows in higher dimensions are aperiodic. In a subsequent paper [28], he and P. B. Percell consider another use of a plug in a flow on a closed manifold: a single plug can capture all trajectories.

The method of "chopping up" trajectories was also used in [22] (see also [23]) to demonstrate the existence of flows with uniformly bounded orbits, specifically:

Theorem 3 *There exists an aperiodic dynamical system on $\mathbb{R}^3$ with each orbit of diameter smaller than 1.*

In dimension 3, Wilson's plug has circular orbits. His theorem asserts the existence of a real analytic vector field with finitely many circular orbits on any closed 3-manifold. A different method is used by G. Kuperberg in [21] to establish a similar fact for volume preserving dynamical systems. He constructs a twisted plug with two circular trajectories on a set homeomorphic to the solid torus $S^1 \times D^2$, copies of which he inserts into the torus $S^1 \times S^1 \times S^1$ furnished with the irrational flow whose every orbit is dense. By the Wallace-Lickorish theorem, any closed orientable 3-manifold can be obtained from any other closed orientable 3-manifold by an integral surgery on a finite link of tori $S^1 \times D^2$. Surgery on non-compact manifolds is handled on a locally finite link. The insertion of each of the Dehn

twisted plugs introduces two circular orbits. The constructions of [21] are volume preserving and yield the following:

THEOREM 4 *Every orientable boundaryless 3-manifold possesses a C^∞ volume preserving dynamical system with a discrete collection of circular trajectories.*

3 COUNTEREXAMPLES TO THE SEIFERT CONJECTURE.

This section lists the known examples of aperiodic flows on S^3 with respect to the degree of differentiability and other properties. In each case a plug is constructed and inserted into a dynamical system on S^3 with one circular orbit. The plug breaks the orbit.

3.1 SCHWEITZER'S VECTOR FIELD.

The first counterexample to the Seifert conjecture came from P. A. Schweitzer in 1972 (published in 1974, see [31]). Schweitzer's construction of an aperiodic C^1 plug is very geometric and astonishing in its simplicity. Unlike Wilson's plug, this vector field is defined on $F \times I$, where F is a non-planar punctured torus. The symmetry guaranteeing the matched ends condition is modeled on two parallel Denjoy minimal sets on which the flow moves in the opposite directions.

THEOREM 5 (Schweitzer) *S^3 admits an aperiodic C^1 vector field.*

3.2 HARRISON'S DIAMOND CIRCLES.

Since the Denjoy vector field on a smooth surface $S^1 \times S^1$ cannot be of class C^2, it seemed impossible to improve the degree of differentiability of Schweitzer's example. J. Harrison [11] embeds the torus $S^1 \times S^1$ in dimension 3 sacrificing the smoothness of the embedding in order to improve the differentiability of the flow on the minimal set. The Denjoy homeomorphism on one of the S^1 factors follows the "diamond circle" pattern.

THEOREM 6 (Harrison) *S^3 admits an aperiodic $C^{3-\epsilon}$ vector field.*

Harrison's construction is limited by the dimension of S^3; thus her method cannot produce a C^3 counterexample to the Seifert conjecture.

3.3 A real analytic counterexample.

The idea behind a smooth aperiodic plug [19] (see also [7]) is to reinsert a Wilson-type plug in itself to cause a recursive breaking of the periodic trajectories. A simple condition prevents the formation of new circular trajectories, even if subjected to the repetitious process of recursion. In [20], G. Kuperberg and K. Kuperberg give specific polynomial formulas for self-insertion performed on a real analytic plug. One of the more interesting features of this construction is that the only minimal set is 2-dimensional, thus the vector field is aperiodic. Hence:

Theorem 7 *There is a C^ω counterexample to the modified Seifert conjecture.*

3.4 Volume preserving aperiodic flows.

H. Hofer [13] proved that a C^1 Reeb vector field on S^3 possesses a closed orbit. This result put a new light on questions related to Hamiltonian flows and volume preserving flows on S^3. In [21], G. Kuperberg adjusts the flow around the Denjoy minimal set in Schweitzer's C^1 plug to make a volume preserving aperiodic C^1 plug, even though the Denjoy dynamical system on $S^1 \times S^1$ is not area preserving. This gives a volume preserving flow without periodic trajectories on S^3, and by Theorem 4, on other 3-manifolds:

Theorem 8 *Every orientable 3-manifold without boundary admits an aperiodic C^1 volume preserving dynamical system.*

At this moment, it is not known whether the differentiability of the volume preserving counterexample to the Seifert conjecture can be improved in a similar fashion as in Harrison's work. However, the intricate formulas of [21] and elaborate computations of [11] emphasize the difficulty in obtaining a C^2 volume preserving aperiodic 3-dimensional plug.

Although the method of self-insertion described [19] and [20] allows quite a lot of flexibility and yields various flows with different degrees of C^r differentiability, the resulting plugs are not volume preserving if $r \geq 1$.

3.5 The structure of minimal sets.

The minimal sets in the counterexamples to the Seifert conjecture, [31], [11] and [21], based on the Denjoy flow, are all homeomorphic to the Denjoy minimal set. The mirror-image symmetry introduced by Wilson is very essential to these flows and always creates two minimal sets. In effect, no example of an aperiodic volume preserving plug with only one minimal set exists.

The plugs described in [19] and [20] contain only one minimal set and every closed 3-manifold admits an analytic flow with only one minimal set. If the construction is at least C^1, then there is a large set of trajectories limiting on the minimal set, preventing the flow from being volume preserving. In contrast to Schweitzer's example, the minimal set is not isolated in the sense of Matsumoto, i.e., every neighborhood contains trajectories that do not belong to the minimal set. It is not known if the minimal set in these constructions (C^1 or better) can be

of dimension 1. A C^0 dynamical system of this type with a 1-dimensional minimal set is given in [20]. In general, if the minimal set is 1-dimensional, then, like the Denjoy sets and solenoids, it is locally homeomorphic to the Cartesian product of the Cantor set and the interval.

In all of the above examples, each of the minimal sets is the inverse limit of polyhedra. Thus a useful tool for classifying these minimal sets is the first cohomology group.

3.6 Flows in higher dimensions.

By Theorem 2, differentiable n-manifolds of Euler characteristic zero or non-compact without boundary, $n \geq 4$, admit smooth aperiodic dynamical systems whose minimal sets are of codimension 2. In particular, this is true for odd dimensional spheres S^n, $n \geq 5$. [20] strengthens Wilson's result:

THEOREM 9 *If M is a closed differentiable or C^ω n-manifold, $n \geq 3$, admitting a dynamical system in the same smoothness category, then there exists an aperiodic dynamical system on M, in the same smoothness category, with only one minimal set whose dimension is $n-1$.*

THEOREM 10 *If M is a differentiable or C^ω manifold without boundary, of dimension at least 3, admitting a dynamical system in the same smoothness category, and $\mathcal{U}$ is an open cover of M, then there exists an aperiodic dynamical system on M, in the same smoothness category, whose orbits are contained in the elements of $\mathcal{U}$, and whose minimal sets have codimension 1.*

The Hamiltonian version of the Seifert conjecture in dimension 3 has not been solved yet, but there are interesting examples in higher dimensions. In 1994, V. Ginzburg [8] and M. Herman [12] independently constructed examples of smooth compact hypersurfaces without closed characteristics in $\mathbb{R}^{2n}$, $n \geq 4$, resolving the case of Hamiltonian flows on spheres of dimension 7 or higher. At the same time, M. Herman [12] found a $C^{3-\epsilon}$ counterexample to the Hamiltonian Seifert conjecture in dimension 5 (i.e., for a compact hypersurface in $\mathbb{R}^6$). In 1997, V. Ginzburg [10] improved the previous results and obtained a smooth proper function $H : \mathbb{R}^{2n} \to \mathbb{R}$, for $2n \geq 6$, with a regular level set on which the Hamiltonian flow has no closed orbits.

3.7 Piecewise linear flows.

PL dynamical systems are thoroughly examined by G. Kuperberg in [21]. A measure on a PL manifold is *simplicial* relative to a triangulation T if on each simplex the measure is given by a linear embedding of the simplex in Euclidean space. The following analogue of Moser's theorem [25], given in [21], demonstrates that simplicial measures are the PL analogue of volume forms:

THEOREM 11 *Two simplicial measures on a connected* PL *manifold M with the same total volume are equivalent by a* PL *homeomorphism. Moreover, any simplicial measure is locally* PL*-Lebesgue.*

The results of [21] related to volume preserving PL flows are:

Theorem 12 *Every orientable 3-manifold without boundary possesses a transversely measured* PL *flow with discrete periodic trajectories.*

Theorem 13 *There is a* PL*, measured, integrally Dehn-twisted plug $\mathcal{D}$ with two closed circular orbits.*

Theorem 14 *Every orientable 3-manifold without boundary possesses a transversely measured* PL *dynamical system with a discrete collection of circular trajectories.*

The main result for PL dynamical systems in [20] is:

Theorem 15 *Let M be a* PL *manifold of dimension $n \geq 3$, $1 \leq k \leq n-1$, and let $\mathcal{U}$ be an open cover of M. A* PL *flow on M can be modified in a* PL *fashion so that the orbits are contained in the elements of $\mathcal{U}$, there are no circular orbits, and all minimal sets are k-dimensional.*

As a corollary, in dimension 3 we have:

Theorem 16 *For $k = 1, 2$, every orientable 3-manifold without boundary admits an aperiodic* PL *flow such that all minimal sets are k-dimensional.*

4 Higher dimensional foliations.

A *k-foliation* on an n-manifold M is an atlas of charts in $\mathbb{R}^n$ that preserve the parallel k-plane foliation of $\mathbb{R}^n$, which is a partition of $\mathbb{R}^n$ into translates of flat $\mathbb{R}^k \subset \mathbb{R}^n$. M is then a *k-foliated manifold*. The foliation structure is in a given category, such as smooth, if the gluing maps are simultaneously in the same category and preserve k-planes.

In [20], G. Kuperberg and K. Kuperberg generalize Theorems 9, 10 and 15 to higher dimensional foliations as follows:

Theorem 17 *If M is a continuous, C^∞, C^ω, or* PL *closed manifold of dimension ≥ 3 admitting a dynamical system in the same smoothness category, then there exists an aperiodic dynamical system on M in the same smoothness category, with exactly one minimal set which is of codimension 1.*

Theorem 18 *If M is a continuous, C^∞, C^ω, or* PL *manifold without boundary of dimension ≥ 3 admitting a dynamical system in the same smoothness category, and $\mathcal{U}$ is an open cover of M, then there exists an aperiodic dynamical system on M in the same smoothness category, whose orbits are contained in the elements of $\mathcal{U}$, and whose minimal sets have codimension 1.*

The above theorems do not carry much information for codimension 1 foliations. There are many results relating to opening closed leaves of such foliations. In particular, S. P. Novikov [27] proved that every C^2 codimension 1 foliation of S^3 has a closed leaf (later extended to continuous foliations), while P. A. Schweitzer [32] showed that it is always possible to modify any codimension 1 foliation in dimension 4 or higher in a C^1 fashion so that it has no compact leaf.

5 Existence of closed orbits

A sequence of fundamental results related to contact forms, Hamiltonian dynamics, and periodic orbits was preceded by a paper of H. Seifert [33] who established the existence of periodic solutions on a fixed energy surface for some Hamiltonians. J. Martinet proved in [24] that every smooth compact 3-manifold possesses a contact form. A tremendous amount of work in this field was done by J. Moser [26], I. Ekeland and J.-M. Lasry [3], A. Weinstein [37], [38], P. Rabinowitz [29], [30], C. Viterbo, Y. Eliashberg, W. Thurston, H. Hofer, E. Zehnder, and others (see [4], [5], [14], [15], [17], [18] for multiple papers and authors). Of particular importance is the 1987 paper by C. Viterbo [36] with a proof of the Weinstein conjecture in $\mathbb{R}^{2n}$: a hypersurface of contact type carries a closed characteristic; and, in relation to both the Seifert and the Weinstein conjectures, the 1993 result of H. Hofer [13] who proved the existence of a closed orbit for a C^1 Reeb vector field on S^3. Subsequently, H. Hofer, K. Wysocki, and E. Zehnder [16] proved that every Reeb vector field on S^3 has an unknotted periodic orbit. K. Cieliebak [2] and V. Ginzburg [9] studied both the existence of periodic orbits and opening closed orbits. J. Etnyre and R. Ghrist [6] proved the Seifert conjecture in hydrodynamics: the C^ω plug [20] cannot be parallel to its curl under any metric.

References

[1] K. Borsuk, *Sur un continu acyclique qui se laisse transformer topologiquement en lui même sans points invariants*, Fund. Math. 24 (1935), 51-58.

[2] K. Cieliebak, *Symplectic boundaries: creating and destroying closed characteristics*, Geom. Funct. Anal. 7 (1997), 269-321.

[3] I. Ekeland and J.-M. Lasry, *On the number of periodic trajectories for a Hamiltonian flow on a convex energy surface*, Ann. of Math. 112 (1980), 283-319.

[4] Y. Eliashberg and H. Hofer, *A Hamiltonian characterization of the three-ball*, Diff. Int. Eqs. 7 (1994), 1303-1324.

[5] Y. Eliashberg and W. P. Thurston, *Confoliations*, University Lecture Series, Amer. Math. Soc., 1997.

[6] J. Etnyre and R. Ghrist, *Contact topology and hydrodynamics*, xxx.lanl.gov e-Print archive, http://front.math.ucdavis.edu/math.DG/9708111.

[7] É. Ghys, *Construction de champs de vecteurs sans orbite périodique (d'après Krystyna Kuperberg)*, Séminaire Bourbaki 78, Juin 1994.

[8] V. L. Ginzburg, *An embedding $S^{2n-1} \to \mathbb{R}^{2n}$, $2n-1 \geq 7$, whose Hamiltonian flow has no periodic trajectories*, Internat. Math. Res. Notices (1995), 83-98.

[9] V. L. Ginzburg, *On the existence and non-existence of closed trajectories for some Hamiltonian flows*, Math. Zeit. 223 (1996), 397-409.

[10] V. L. Ginzburg, *A smooth counterexample to the Hamiltonian Seifert conjecture in* $\mathbb{R}^6$, Internat. Math. Res. Notices (1997), 641-650.

[11] J. Harrison, C^2 *counterexamples to the Seifert conjecture*, Topology 27 (1988), 249-278.

[12] M. Herman, Communication at the conference in honor of A. Douady *Geometrie Complexe et Systemes Dynamiques*, Orsay, France, 1995.

[13] H. Hofer, *Pseudoholomorphic curves in symplectizations with applications to the Weinstein conjecture in dimension three*, Invent. Math. 114 (1993), 515-563.

[14] H. Hofer and C. Viterbo, *The Weinstein conjecture in cotangent bundles and related results*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. 15 (1988), 411-445.

[15] H. Hofer and C. Viterbo, *The Weinstein conjecture in the presence of holomorphic spheres*, Comm. Pure Appl. Math. 45 (1992), 583-622.

[16] H. Hofer, K. Wysocki and E. Zehnder, *Unknotted periodic orbits for Reeb flows on the three-sphere*, Topol. Methods Nonlinear Anal. 7 (1996), 219-244.

[17] H. Hofer and E. Zehnder, *Periodic solutions on hypersurfaces and a result by C. Viterbo*, Invent. Math. 90 (1987), 1-9.

[18] H. Hofer and E. Zehnder, *Symplectic invariants and Hamiltonian dynamics*, Birkhäuser Verlag, Basel, 1994.

[19] K. Kuperberg, *A smooth counterexample to the Seifert conjecture*, Ann. of Math. 140 (1994), 723-732.

[20] G. Kuperberg and K. Kuperberg, *Generalized counterexamples to the Seifert conjecture*, Ann. of Math. 144 (1996), 239-268.

[21] G. Kuperberg, *A volume-preserving counterexample to the Seifert conjecture*, Comment. Math. Helv. 71 (1996), 70-97.

[22] K. M. Kuperberg and C. S. Reed, *A dynamical system on* $\mathbb{R}^3$ *with uniformly bounded trajectories and no compact trajectories*, Proc. Amer. Math. Soc. 106 (1989), 1095-1097.

[23] K. Kuperberg, W. Kuperberg, P. Minc and C. S. Reed, *Examples related to Ulam's fixed point problem*, Topol. Methods Nonlinear Anal. 1 (1993), 173-181.

[24] J. Martinet, *Formes de contact sur les variétés de dimension 3*, Lecture Notes in Math. 209 (1971), 142-163.

[25] J. Moser, *On the volume elements on a manifold*, Trans. Amer. Math. Soc. 120 (1965), 286-294.

[26] J. Moser, *Periodic orbits near an equilibrium and a theorem of A. Weinstein*, Comm. Pure Appl. Math. 29 (1976), 727-747.

[27] S. P. Novikov, *Topology of foliations*, Trans. Math. Moscow Soc. 14 (1967), 268-296.

[28] P. B. Percell, F. W. Wilson, *Plugging flows*, Trans. Amer. Math. Soc. 233 (1977), 93-103.

[29] P. H. Rabinowitz, *Periodic solutions of Hamiltonian systems*, Comm. Pure Appl. Math. 31 (1978), 157-184.

[30] P. H. Rabinowitz, *Periodic solutions of a Hamiltonian system on a prescribed energy surface*, J. Diff. Eqs 33. (1979), 336-352.

[31] P. A. Schweitzer, *Counterexamples to the Seifert conjecture and opening closed leaves of foliations*, Ann. of Math. 100 (1974), 386-400.

[32] P. A. Schweitzer, *Codimension one foliations without compact leaves*, Comment. Math. Helv. 70 (1995), 171-209.

[33] H. Seifert, *Periodische Bewegungen mechanischer Systeme*, Math. Zeit. 51 (1948), 197-216.

[34] H. Seifert, *Closed integral curves in 3-space and isotopic two-dimensional deformations*, Proc. Amer. Math. Soc. 1 (1950), 287-302.

[35] W. P. Thurston, Electronic communication (e-mail).

[36] C. Viterbo, *A proof of Weinstein's conjecture in* $\mathbb{R}^{2n}$, Ann. Inst. H. Poincaré Anal. Non Linéaire 4 (1987), 337-356.

[37] A. Weinstein, *Periodic orbits for convex Hamiltonian systems*, Ann. of Math. 108 (1978), 507-518.

[38] A. Weinstein, *On the hypotheses of Rabinowitz' periodic orbit theorems*, J. Diff. Eqs. 33 (1979), 353-358.

[39] F. W. Wilson, *On the minimal sets of non-singular vector fields*, Ann. of Math. 84 (1966), 529-536.

Krystyna Kuperberg
Department of Mathematics
Auburn University
Auburn, AL 36849-5310
USA

Rigidity and Inflexibility in Conformal Dynamics

Curtis T. McMullen[1]

1 Introduction

This paper presents a connection between the rigidity of hyperbolic 3-manifolds and universal scaling phenomena in dynamics.

We begin by stating an inflexibility theorem for 3-manifolds of infinite volume, generalizing Mostow rigidity (§2). We then connect this inflexibility to dynamics and discuss:

- The geometrization of 3-manifolds which fiber over the circle (§2);
- The renormalization of unimodal maps $f : [0, 1] \to [0, 1]$ (§4),
- Real-analytic circle homeomorphisms with critical points (§5), and
- The self-similarity of Siegel disks (§6).

Chaotic sets for these four examples are shown in Figure 1. The snowflake in the first frame is the limit set Λ of a Kleinian group Γ acting on the Riemann sphere $S^2_\infty = \partial\mathbb{H}^3$. Its center c is a *deep point* of Λ, meaning the limit set is very dense at microscopic scales near c. Because of the inflexibility and combinatorial periodicity of $M = \mathbb{H}^3/\Gamma$, the limit set is also self-similar at c with a universal scaling factor.

The remaining three frames show deep points of the (filled) Julia set for other conformal dynamical systems: the Feigenbaum polynomial, a critical circle map and the golden ratio Siegel disk. Our goal is to explain an inflexibility theory that leads to universal scaling factors and convergence of renormalization for these examples as well.

The qualitative theory of dynamical systems, initiated by Poincaré in his study of celestial mechanics, seeks to model and classify stable regimes, where the topological form of the dynamics is locally constant. In the late 1970s physicists discovered a rich, universal structure in the onset of instability. One-dimensional dynamical systems emerged as elementary models for critical phenomena, phase transitions and renormalization.

In pure mathematics, Mostow and others have developed a rigidity theory for compact manifolds M^n of constant negative curvature, $n \geq 3$, and other quotients of symmetric spaces. This theory shows M is determined up to isometry by $\pi_1(M)$

[1]Research supported in part by the NSF.

1991 Mathematics Subject Classification: 30D05, 30F40, 58F11, 58F23

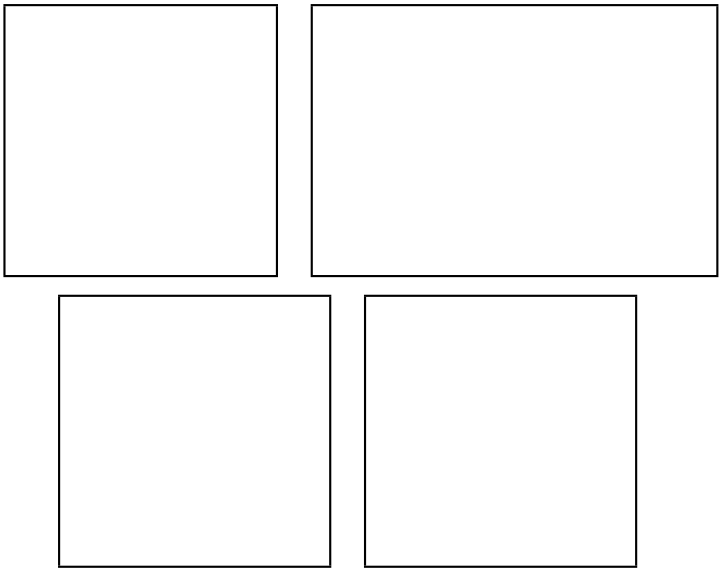


Figure 1. Dynamical systems with deep points: a totally degenerate Kleinian group, the Feigenbaum polynomial, a critical circle map and the golden mean Siegel disk.

as an abstract, finitely-presented group. Remarkably, rigidity of M is established via the ergodic theory of $\pi_1(M)$ acting on the boundary of the universal cover of M.

In our case, $M = \mathbb{H}^3/\Gamma$ is a hyperbolic 3-manifold, the boundary of its universal cover $\mathbb{H}^3$ is isomorphic to S^2, and the action of $\pi_1(M) \subset \mathrm{Isom}^+(\mathbb{H}^3) = PSL_2(\mathbb{C})$ on S^2 is conformal. Similarly, upon complexification, 1-dimensional dynamical systems give rise to holomorphic maps on the Riemann sphere $\widehat{\mathbb{C}} \cong S^2$. Hyperbolic space $\mathbb{H}^3$ enters the dynamical picture as a means to organize *geometric limits* under rescaling (§3). The universality observed by physicists can then be understood, as in the case of 3-manifolds, in terms of rigidity of these geometric limits.

We conclude with progress towards the classification of hyperbolic manifolds (§7), where geometric limits also play a central role.

2 Hyperbolic 3-manifolds and fibrations

A *hyperbolic manifold* is a complete Riemannian manifold with a metric of constant curvature -1. Mostow rigidity states that any two closed, homotopy equivalent hyperbolic 3-manifolds are actually isometric.

In this section we discuss a remnant of rigidity for *open* manifolds. Let $\mathrm{core}(M) \subset M$ denote the *convex core* of M, defined as the closure of the set of geodesic loops in M. The manifold M satisfies $[r, R]$-*injectivity bounds*, $r > 0$, if for any $p \in \mathrm{core}(M)$, the largest embedded ball $B(p, s) \subset M$ has radius $s \in [r, R]$.

Let $f : M \to N$ be a homotopy equivalence between a pair of hyperbolic 3-manifolds. Then f is a *K-quasi-isometry* if, when lifted to the universal covers,

$$\begin{aligned} \operatorname{diam}(\widetilde{f}(B)) &\leq K(\operatorname{diam} B + 1) \quad \forall B \subset \widetilde{M}, \quad \text{and} \\ \operatorname{diam}(\widetilde{f}^{-1}(B)) &\leq K(\operatorname{diam} B + 1) \quad \forall B \subset \widetilde{N}. \end{aligned}$$

A diffeomorphism $f : M \to N$ is an *asymptotic isometry* if f is exponentially close to an isometry deep in the convex core. That is, there is an $A > 1$ such that for any nonzero vector $v \in T_pM$, $p \in \operatorname{core}(M)$, we have

$$\left| \log \frac{|Df(v)|}{|v|} \right| \leq CA^{-d(p,\partial \operatorname{core}(M))}.$$

In [Mc2] we show:

THEOREM 2.1 (GEOMETRIC INFLEXIBILITY) *Let M and N be quasi-isometric hyperbolic 3-manifolds with injectivity bounds. Then M and N are asymptotically isometric.*

Mostow rigidity is a special case: if M and N are closed, then any homotopy equivalence $M \sim N$ is a quasi-isometry, injectivity bounds are automatic, and $\partial \operatorname{core}(M) = \emptyset$, so an asymptotic isometry is an isometry.

To sketch the proof of Theorem 2.1, recall any hyperbolic 3-manifold M determines a conformal dynamical system, namely the action of its fundamental group $\pi_1(M)$ on the sphere at infinity $S^2_\infty = \partial\mathbb{H}^3$ for the universal cover $\widetilde{M} \cong \mathbb{H}^3$. The *limit set* $\Lambda \subset S^2_\infty$ is the chaotic locus for this action; its convex hull covers the core of M. The action is properly discontinuously on the rest of the sphere, and the quotient $\partial M = (S^2_\infty - \Lambda)/\pi_1(M)$ gives a natural Riemann surface at infinity for M.

Figure 2. An observer deep in the convex core sees a kaleidoscopic view of ∂M.

A quasi-isometric deformation of M determines a quasiconformal deformation v of ∂M, which in turn admits a (harmonic) visual extension V to an equivalent deformation of M. The strain $SV(p)$ is the average of the ellipse field $Sv = \overline{\partial} v$ over all visual rays γ from p to ∂M. By our injectivity bounds, γ corkscrews chaotically before exiting the convex core. Thus the ellipses of Sv on ∂M appear in random orientations as seen from p (Figure 2). This randomness provides abundant cancellation in the visual average, and we find the metric distortion

$\|SV(p)\|$ is exponentially small compared to $\|Sv\|_\infty$. Thus V is an infinitesimal asymptotic isometry.

In dimension 3, any two quasi-isometric hyperbolic manifolds are connected by a smooth path in the deformation space, so the global theorem follows from the infinitesimal version.

Inflexibility is also manifest on the sphere at infinity. Let us say a local homeomorphism ϕ on $S^2_\infty \cong \widehat{\mathbb{C}}$ is $C^{1+\alpha}$-*conformal* at z if the complex derivative $\phi'(z)$ exists and

$$\phi(z+t) = \phi(z) + \phi'(z)\cdot t + O(|t|^{1+\alpha}).$$

We say $x \in \Lambda \subset S^2_\infty$ is a *deep point* if Λ is so dense at x that for some $\beta > 0$,

$$B(y,s) \subset B(x,r) - \Lambda \implies s = O(r^{1+\beta}).$$

It is easy to see that a geodesic ray $\gamma \subset \mathbb{H}^3$ terminating at a deep point in the limit set penetrates the convex hull of Λ at a linear rate. ¿From the inflexibility theorem we find:

Corollary 2.2 *Let M and N satisfy injectivity bounds, and let $\phi : S^2_\infty \to S^2_\infty$ be a quasiconformal conjugacy between $\pi_1(M)$ and $\pi_1(N)$. Then ϕ is $C^{1+\alpha}$-conformal at every deep point of the limit set of $\pi_1(M)$.*

The inflexibility theorem is motivated by the following application to 3-manifolds that fiber over the circle. Let S be a closed surface of genus $g \geq 2$ and let $\psi \in \mathrm{Mod}(S)$ be a pseudo-Anosov mapping class. Let

$$T_\psi \;=\; S \times [0,1]/\{(x,0) \sim (\psi(x),1)\}$$

be the 3-manifold fibering over the circle with fiber S and monodromy ψ. By a deep theorem of Thurston, T_ψ is hyperbolic. To find its hyperbolic structure, let $V(S)$ denote the variety of representations $\rho : \pi_1(S) \to \mathrm{Isom}(\mathbb{H}^3)$, and define

$$\mathcal{R} : V(S) \to V(S)$$

by $\mathcal{R}(\rho) = \rho \circ \psi_*^{-1}$. We refer to $\mathcal{R}$ as a *renormalization operator*, because it does not change the group action on $\mathbb{H}^3$, only its marking by $\pi_1(S)$.

Let $QF(S) \overset{Q}{\cong} \mathrm{Teich}(S) \times \mathrm{Teich}(\overline{S}) \subset V(S)$ denote the space of quasifuchsian groups, and define

$$M(X,\psi) = \lim_{n\to\infty} Q(X, \psi^{-n}Y), \quad \text{for any } (X,Y) \in \mathrm{Teich}(S) \times \mathrm{Teich}(\overline{S}).$$

Then $M = M(X,\psi)$ has injectivity bounds, its convex core is homeomorphic to $S \times [0,\infty)$, and the manifolds M and $\mathcal{R}(M)$ are quasi-isometric. By the inflexibility theorem there is an asymptotic isometry $\Psi : M \to M$ in the homotopy class of ψ, so the convex core of M is asymptotically periodic. As n tends to ∞, the marking of $\mathcal{R}^n(M)$ moves into the convex core at a linear rate, and we find:

Theorem 2.3 *The renormalizations $\mathcal{R}^n(M(X,\psi))$ converge exponentially fast to a fixed-point M_ψ of $\mathcal{R}$.*

Since $\mathcal{R}(M_\psi) = M_\psi$, the map ψ is realized by an isometry α on M_ψ, and the quotient $T_\psi = M_\psi/\langle\alpha\rangle$ gives the desired hyperbolic structure on the mapping cylinder of ψ.

This iterative construction of T_ψ hints at a dynamical theory of the action of $\mathrm{Mod}(S)$ on the variety $V(S)$, as does the following result [Kap]:

Theorem 2.4 (Kapovich) *The derivative $D\mathcal{R}_\psi$ is hyperbolic on the tangent space to $V(S)$ at M_ψ for all pseudo-Anosov mapping classes on closed surfaces.*

The snowflake in the first frame of Figure 1 is a concrete example of the limit set Λ for a Kleinian group $\Gamma = \pi_1(M(X,\psi))$ as above. In this example S is a torus, made hyperbolic by introducing a single orbifold point $p \in S$ of order 3; and $\psi = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right) \in SL_2(\mathbb{Z}) \cong \mathrm{Mod}(S)$ is the simplest pseudo-Anosov map. The suspension of $p \in S$ gives a singular geodesic $\gamma \subset T_\psi$ forming the orbifold locus of the mapping torus of ψ.

The picture is centered at a deep point $c \in \Lambda$ fixed by an elliptic element of order 3 in Γ. The limit set Λ is a nowhere dense but very furry tree, with six limbs meeting at c. By general results, Λ is a locally connected dendrite, with Hausdorff dimension two but measure zero [CaTh], [Th1, Ch. 8], [Sul1], [BJ1]; in fact by [BJ2] we have $0 < \mu_h(\Lambda) < \infty$ for the gauge function $h(r) = r^2|\log r \log\log\log r|^{1/2}$.

One can easily construct a quasiconformal automorphism ϕ of Γ, with $\phi(c) = c$ and $\phi \circ \gamma = \psi_*(\gamma) \circ \phi$ for all $\gamma \in \Gamma$. By Corollary 2.2, ϕ is $C^{1+\alpha}$-conformal at c, and we find:

Theorem 2.5 *The limit set Λ is self-similar at each elliptic fixed-point in Λ, with scaling factor $\phi'(c) = e^L$. Here L is the complex length of the singular geodesic γ on T_ψ.*

In particular the self-similarity factor e^L is inherited from the geometry of the rigid manifold T_ψ, and it is universal across all manifolds $M(X,\psi)$ attracted to M_ψ under renormalization.

3 Geometric limits in dynamics

In this section we extend the inflexibility of Kleinian groups and their limit sets to certain other conformal dynamical systems $\mathcal{F}$ and their Julia sets J, where we will find:

> *The conformal structure at the deep points of J is determined by the topological dynamics of $\mathcal{F}$.*

Consider the space $\mathcal{H}$ of all holomorphic maps $f : U(f) \to V(f)$ between domains in $\widehat{\mathbb{C}}$. Introduce a (non-Hausdorff) topology on $\mathcal{H}$ such that $f_n \to f$ if

for any compact $K \subset U(f)$, we have $K \subset U(f_n)$ for all $n \gg 0$ and $f_n|K \to f|K$ uniformly.

A *holomorphic dynamical system* is a subset $\mathcal{F} \subset \mathcal{H}$. Given a sequence of dynamical systems $\mathcal{F}_n \subset \mathcal{H}$, the *geometric limit* $\mathcal{F} = \limsup \mathcal{F}_n$ consists of all maps $f = \lim f_{n_i}$ obtained as limits of subsequences $f_{n_i} \in \mathcal{F}_{n_i}$.

To bring hyperbolic space into the picture, identify $\widehat{\mathbb{C}}$ with the boundary of the Poincaré ball model for $\mathbb{H}^3$, let $\mathrm{F}\mathbb{H}^3$ be its frame bundle, and let $\omega_0 \in \mathrm{F}\mathbb{H}^3$ be a standard frame at the center of the ball. Given any other $\omega \in \mathrm{F}\mathbb{H}^3$, there is a unique Möbius transformation g sending ω_0 to ω, and we define

$$(\mathcal{F}, \omega) = g^*(\mathcal{F}) = \{g^{-1} \circ f \circ g \ : \ f \in \mathcal{F}\}.$$

In other words, $(\mathcal{F}, \omega)$ is $\mathcal{F}$ as 'seen from' ω.

We say $\mathcal{F}$ is *twisting* if it is essentially nonlinear — for example, if there exists an $f \in \mathcal{F}$ with a critical point, or if $\mathcal{F}$ contains a free group of Möbius transformations.

Given a closed set $J \subset \widehat{\mathbb{C}}$, we say $(\mathcal{F}, J)$ is *uniformly twisting* if $\limsup(\mathcal{F}, \omega_n)$ is twisting for any sequence $\omega_n \in \mathrm{F}(\mathrm{hull}(J))$, the frame bundle over the convex hull of J in $\mathbb{H}^3$. Informally, uniform twisting means $\mathcal{F}$ is quite nonlinear at every scale around every point of J.

For a Kleinian group, the pair $(\Gamma, \Lambda(\Gamma))$ is uniformly twisting iff $M = \mathbb{H}^3/\Gamma$ has injectivity bounds. Thus geometric inflexibility, Corollary 2.2, is a special case of [Mc2]:

Theorem 3.1 (Dynamic Inflexibility) *Let $(\mathcal{F}, J)$ be uniformly twisting, and let ϕ be a quasiconformal conjugacy from $\mathcal{F}$ to another holomorphic dynamical system $\mathcal{F}'$. Then ϕ is $C^{1+\alpha}$-conformal at all deep points of J.*

The next three sections illustrate how such inflexibility helps explain universal scaling in dynamics.

4 Renormalization of interval maps

Let $f : I \to I$ be a real-analytic map on an interval. The map f is *quadratic-like* if $f(\partial I) \subset \partial I$ and f has a single quadratic critical point $c_0(f) \in \mathrm{int}(I)$. The basic example is $f(x) = x^2 + c$ on $[-a, a]$ with $f(a) = a$. We implicitly identify maps that are linearly conjugate.

If an iterate $f^p|L$ is also quadratic-like for some interval L, with $c_0(f) \in L \subset I$, then we can take the least such $p > 1$ and define the *renormalization* of f by

$$\mathcal{R}(f) = f^p|L.$$

The order of the intervals $L, f(L), \ldots, f^p(L) = L \subset I$ determines a permutation $\sigma(f)$ on p symbols.

The map f is *infinitely renormalizable* if the sequence $\mathcal{R}^n(f)$ is defined for all $n > 0$. The *combinatorics* of f is then recorded by the sequence of permutations $\tau(f) = \langle \sigma(\mathcal{R}^n(f)) \rangle$. We say f has *bounded combinatorics* if only finitely many permutations occur, and *periodic combinatorics* if $\tau(\mathcal{R}^q f) = \tau(f)$ for some $q \geq 1$.

THEOREM 4.1 *Let $f : I \to I$ be infinitely renormalizable, with combinatorics of period q. Then $\mathcal{R}^{qn}(f) \to F$ exponentially fast as $n \to \infty$, where F is the unique fixed-point of the renormalization operator $\mathcal{R}^q$ with the same combinatorics as f.*

For example, the Feigenbaum polynomial $f(x) = x^2 - 1.4101155\cdots$, arising at the end of the cascade of period doublings in the quadratic family, has $\tau(f) = \langle(12),(12),(12),\ldots\rangle$. Under renormalization, $\mathcal{R}^n(f)$ converges exponentially fast to a solution of the functional equation

$$F \circ F(x) = \alpha^{-1} F(\alpha x).$$

To formulate the speed of convergence more completely, extend $f : I \to I$ to a complex analytic map on a neighborhood of $I \subset \mathbb{C}$, and let $F : W \to \mathbb{C}$ denote the maximal analytic continuation of the renormalization fixed-point. Then we find there is an $A > 1$ such that for any compact $K \subset W$, we have

$$\sup_{z \in K} |\mathcal{R}^n(f)(z) - F(z)| = O(A^{-n}),$$

where $\mathcal{R}^n(f)$ is suitably rescaled.

Now suppose only that f has bounded combinatorics. Under iteration of f, all but countably many points in I are attracted to the postcritical Cantor set

$$P(f) = \overline{\bigcup_{n>0} f^n(c_0(f))} \subset I.$$

THEOREM 4.2 *Let f and g be infinitely renormalizable maps with the same bounded combinatorics. Then $f|P(f)$ and $g|P(g)$ are $C^{1+\alpha}$-conjugate.*

Thus quantitative features of the attractor $P(f)$ (such as its Hausdorff dimension) are determined by the combinatorics $\tau(f)$.

These universal properties of quadratic-like maps were observed experimentally and linked to renormalization by Feigenbaum and Coullet-Tresser in the late 1970s. A program for applying complex quadratic-like maps to renormalization was formulated by Douady and Hubbard in the early 1980s. Sullivan introduced a wealth of new ideas and established the convergence $\mathcal{R}^{nq}(f) \to F$ [Sul3], [Sul4]. The inflexibility theory gives a new proof yielding, in addition, exponential speed of convergence and $C^{1+\alpha}$-smoothness of conjugacies.

Our approach to renormalization is via *towers* [Mc2]. For simplicity we treat the case of the Feigenbaum polynomial f. By Sullivan's *a priori* bounds, the sequence of renormalizations $\langle\mathcal{R}^n(f)\rangle$ is compact, and all limits are complex quadratic-like maps with definite moduli. Passing to a subsequence we can arrange that $\mathcal{R}^{n+i}(f) \to f_i$ and obtain a tower

$$\mathcal{T} = \langle f_i \,:\, i \in \mathbb{Z}\rangle \quad \text{such that } f_{i+1} = f_i \circ f_i \;\forall i.$$

The Julia set $J(\mathcal{T}) = \bigcup J(f_i)$ is dense in $\mathbb{C}$, and we deduce that $\mathcal{T}$ is rigid — it admits no quasiconformal deformations. Convergence of renormalization, $\mathcal{R}^n(f) \to F$, then easily follows.

The rapid speed of convergence of renormalization comes from inflexibility of the one-sided tower $\mathcal{T} = \langle f, f^2, f^4, \ldots \rangle$. To establish this inflexibility, we first show the full dynamical system $\mathcal{F}(f) = \{f^{-i} \circ f^j\}$ contains copies of f^{2^n} near every $z \in J(f)$ and at every scale. Thus $(\mathcal{F}(f), J(f))$ is *uniformly twisting.* Next we use expansion in the hyperbolic metric on $\mathbb{C} - P(f)$ to show $c_0(f)$ is a *deep point* of $J(f)$. Finally by Theorem 3.1, a quasiconformal conjugacy ϕ from f to $\mathcal{R}(f) = f \circ f$ is actually $C^{1+\alpha}$-conformal at the critical point. At small scales ϕ provides a nearly linear conjugacy from $\mathcal{R}^n(f)$ to $\mathcal{R}^{n+1}(f)$, and exponential convergence follows.

The second frame of Figure 1 depicts the Julia set of the infinitely renormalizable Feigenbaum polynomial f, centered at its critical point. The Julia set $J(f)$ is locally connected [JH], [LS]; it is still unknown if $\operatorname{area}(J(f)) = 0$.

Milnor has observed that the Mandelbrot set M is quite dense at the Feigenbaum point $c = -1.4101155\ldots \in \partial M$ and at other fixed-points of tuning [Mil], and it is reasonable to expect that c is a deep point of M. Lyubich has recently given an elegant proof of the hyperbolicity of renormalization at its fixed-points, including a new proof of exponential convergence of $\mathcal{R}^n(f)$ via the Banach space Schwarz lemma, and a proof of Milnor's conjecture that blowups of M around the Feigenbaum point converge to the whole plane in the Hausdorff topology [Lyu].

5 Critical circle maps

A *critical circle map* $f : S^1 \to S^1$ is a real-analytic homeomorphism with a single cubic critical point $c_0(f) \in S^1$. A typical example is the *standard map*

$$f(x) = x + \Omega + K \sin(x), \qquad x \in \mathbb{R}/2\pi\mathbb{Z}, \quad \Omega \in \mathbb{R}$$

with $K = -1$ and $c_0 = 0$. These maps arise in KAM theory and model the disappearance of invariant circles [FKS], [Lan], [Rand], [Mak], [DGK]. Another class of examples are the rational maps

$$f(z) = \lambda z^2 \frac{z-3}{1-3z}, \quad |\lambda| = 1, \tag{5.1}$$

acting on $S^1 = \{z : |z| = 1\}$ with $c_0(f) = 1$.

If $f : S^1 \to S^1$ has no periodic points, then it is topologically conjugate to a rigid rotation by angle $2\pi\rho(f)$, where the *rotation number* $\rho(f)$ is irrational [Y]. The behavior of f is strongly influenced by the continued fraction of its rotation number,

$$\rho(f) = 1/(a_1 + 1/(a_2 + 1/(a_3 + \cdots))), \quad a_i \in \mathbb{N}.$$

By truncating the continued fraction we obtain rational numbers $p_n/q_n \to \rho(f)$. We say $\rho(f)$ is of *bounded type* if $\sup a_i < \infty$.

Theorem 5.1 (de Faria-de Melo) *Let f_1, f_2 be two critical circle maps with equal irrational rotation numbers of bounded type. Then f_1 and f_2 are $C^{1+\alpha}$-conjugate.*

We sketch the proof from [dFdM]. Consider a complex analytic extension of $f(z)$ to a neighborhood of S^1. Let the *Julia set* $J(f)$ be the closure of the set of periodic points of f. As for maps of the interval, one finds the critical point $c_0(f)$ is a *deep point* of $J(f)$, and the full dynamical system $(\mathcal{F}(f), J(f))$ is *uniformly twisting*. Because of the good arithmetic of $\rho(f)$, the forward orbit of the critical point is spread evenly along S^1, so in fact the Julia set is deep at *every* point on the circle. To complete the proof, one constructs a quasiconformal conjugacy between f_1 and f_2, and then applies the inflexibility Theorem 3.1 to deduce that $\phi|S^1$ is $C^{1+\alpha}$.

To bring renormalization into the picture, it is useful to work on the universal cover $\mathbb{R}$ of $S^1 = \mathbb{R}/2\pi\mathbb{Z}$. One can then treat the lifted map $f : \mathbb{R} \to \mathbb{R}$ and the deck transformation $g(x) = x + 2\pi$ on an equal footing. The maps (f, g) form a basis for a subgroup $\mathbb{Z}^2 \subset \mathrm{Diff}(\mathbb{R})$, and any matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in GL_2(\mathbb{Z})$ determines a *renormalization operator* by

$$\mathcal{R}(f, g) = (f^a g^b, f^c g^d).$$

When the continued fraction of $\rho(f)$ is periodic, one can choose $\mathcal{R}$ such that $\mathcal{R}^n(f, g)$ converges exponentially fast to a fixed-point of renormalization (F, G). For the more general case where $\rho(f)$ is of bounded type, a finite number of renormalization operators suffice to relate any two adjacent levels of the tower $\mathcal{T} = \langle f^{q_n} \rangle$.

The third frame in Figure 1 depicts the Julia set of the rational map $f(z)$ given by equation (5.1), with $\lambda \approx -0.7557 - 0.6549i$ chosen so $\rho(f)$ is the golden ratio. The picture is centered at the deep point $c_0(f) \in J(f)$. Petersen has shown $J(f)$ is locally connected [Pet]; it is an open problem to determine if $\mathrm{area}(J(f)) = 0$.

Levin has proposed a similar theory for critical circle *endomorphisms* such as $f(z) = \lambda z^3(z - 2)/(1 - 2z)$ [Lev].

6 The golden-ratio Siegel disk

Let $f(z) = \lambda z + z^2$, where $\lambda = e^{2\pi i\theta}$.

Siegel showed that f is analytically conjugate to the rotation $z \mapsto \lambda z$ on a neighborhood of the origin when θ is Diophantine ($|\theta - p/q| > C/q^n$). The *Siegel disk* D for f is the maximal domain on which f can be linearized. For θ of bounded type, Herman and Świątek proved that ∂D is a quasicircle passing through the critical point $c_0(f) = -\lambda/2$ [Dou1], [Sw]. In particular, the critical point provides the only obstruction to linearization.

Now suppose θ is a quadratic rational such as the golden ratio:

$$\theta = \frac{\sqrt{5} - 1}{2} = 1/(1 + 1/(1 + 1/(1 + \cdots))).$$

Then the continued fraction of θ is preperiodic; there is an $s > 0$ such that $a_{n+s} = a_n$ for all $n \gg 0$. Experimentally, a universal structure emerges at the transition from linear to nonlinear behavior at ∂D [MN] [Wid]. In [Mc4] we prove:

Theorem 6.1 *If θ is a quadratic irrational, then the boundary of the Siegel disk D for f is self-similar about the critical point $c_0(f) \in \partial D$.*

More precisely, there is a map $\phi : (\overline{D}, c_0) \to (\overline{D}, c_0)$ which is a $C^{1+\alpha}$-conformal contraction at the critical point, and locally conjugates f^{q_n} to $f^{q_{n+s}}$.

Theorem 6.2 *Let f and g be quadratic-like maps with Siegel disks having the same rotation number of bounded type. Then $f|\overline{D}_f$ and $g|\overline{D}_g$ are $C^{1+\alpha}$ conjugate.*

For instance, let D_a be the Siegel disk for $f_a(z) = \lambda z + z^2 + az^3$. Then the Hausdorff dimension of ∂D_a is *constant* for small values of a. As for the Julia set we have:

Theorem 6.3 *If θ has bounded type, then the Hausdorff dimension of the Julia set of $f(z) = e^{2\pi i\theta}z + z^2$ is strictly less than two.*

A blowup of the golden ratio Siegel disk, centered at the critical point $c_0(f) \in \partial D$, is shown in the final frame of Figure 1. The picture is self-similar with a universal scaling factor $1.8166\ldots$ depending only on the rotation number. The Julia set of f is locally connected [Pet]. Recently Buff and Henriksen have shown that the golden Siegel disk contains a Euclidean triangle with vertex resting on the critical point [BH]; empirically, an angle of approximately 120° will fit.

The mechanism of rigidity for Siegel disks is visible in the geometry of the *filled* Julia set $K(f) = \{z \ : \ f^n(z) \text{ remains bounded for all } n > 0\}$. Under iteration, every point in the interior of $K(f)$ eventually lands in the Siegel disk, and $\partial K(f) = J(f)$. The gray cauliflower forming the interior of $K(f)$ in Figure 1 is visibly dense at the critical point. In fact $c_0(f)$ is a *measurable* deep point of $K(f)$, meaning

$$\frac{\text{area}(K(f) \cap B(c_0, r))}{\text{area}(B(c_0, r))} = 1 - O(r^\beta), \quad \beta > 0. \tag{6.1}$$

For the proof of Theorem 6.2, one starts with a quasiconformal conjugacy ϕ from f to g furnished by the theory of polynomial-like maps [DH]. Since f and g have the same linearization on their Siegel disks, we can assume ϕ is conformal on D_f. But then ϕ is conformal throughout $\text{int}\, K(f)$. By (6.1) the conformal behavior dominates near $c_0(f)$, and we conclude ϕ is $C^{1+\alpha}$-conformal at the critical point. This smoothness is spread to all points of ∂D_f using the good arithmetic of θ.

The self-similarity of ∂D is established similarly, using a conjugacy from f^{q_n} to $f^{q_{n+s}}$.

The dictionary. Table 3 summarizes the parallels which emerge between hyperbolic manifolds, quadratic-like maps on the interval, critical circle maps and Siegel disks. This table can be seen as a contribution to Sullivan's dictionary between conformal dynamical systems [Sul2], [Mc1].

7 Surface groups and their geometric limits

For a complete classification of conformal dynamical systems, one must go beyond the bounded geometry of the preceding examples, and confront short geodesics,

HYPERBOLIC MANIFOLDS	INTERVAL MAPS	SIEGEL DISKS/ CIRCLE MAPS
Discrete surface group $\Gamma \subset PSL_2(\mathbb{C})$ $M = \mathbb{H}^3/\Gamma$	$\mathbb{R}$-quadratic polynomial $f(z) = z^2 + c$	Nonlinear rotation $f(z) = \lambda z + z^2$ or $\lambda z^2(z-3)/(1-3z)$
Representation $\rho : \pi_1(S) \to \Gamma$	Quadratic-like map $f : U \to V$	Holomorphic commuting pair (f, g)
Ending lamination $\epsilon(M) \in \mathcal{GL}(S)$	Tuning invariant $\tau(f) = \langle \sigma(\mathcal{R}^n(f)) \rangle$	Continued fraction $\theta = [a_1, a_2, \cdots]$, $\lambda = e^{2\pi i\theta}$
Inj. radius$(M) > r > 0$	Bounded combinatorics	Bounded type
Cut points in Λ $= \bigcup_1^\infty$ (Cantor sets)	Postcritical set $P(f) = \overline{\bigcup f^n(c)}$, $f'(c) = 0$ $=$ (Cantor set)	$=$ (circle or quasi-circle)
($\mathbb{R}$-tree of $\epsilon(M), \pi_1(S)$)	(proj $\lim \mathbb{Z}/p_i, x \mapsto x+1$)	$(\mathbb{R}/\mathbb{Z}, x \mapsto x + \theta)$
$\Lambda(\Gamma)$ is locally connected	$J(f)$ is locally connected	$J(f)$ is locally connected
area $\Lambda(\Gamma) = 0$	area$(J(f)) = 0$?	
Inj. radius $\in [r, R]$ in core(M)	$(\mathcal{F}(f), J(f))$ is uniformly twisting	
Mapping class $\psi \in \mathrm{Mod}(S)$	Kneading permutation	Automorphism $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ of $\mathbb{Z}^2$
Renormalization Operators		
$\mathcal{R}(\rho) = \rho \circ \psi^{-1}$	$\mathcal{R}(f) = f^p(z)$	$\mathcal{R}(f, g) = (f^a g^b, f^c g^d)$
Stable Manifold of Renormalization		
$M =$ asymptotic fiber	$f =$ limit of doublings	$\theta =$ golden ratio
Elliptic points deep in $\Lambda(\Gamma)$	Critical point $c_0(f)$ deep in $J(f)$ or $K(f)$	
$\rho \circ \psi^{-n}$, $n = 1, 2, 3 \ldots$	f^n, $n = 1, 2, 4, 8, 16, \ldots$	f^n, $n = 1, 2, 3, 5, 8, \ldots$
Geometric limit of $\mathcal{R}^n(\rho)$	Quadratic-like tower $\langle f_i : i \in \mathbb{Z} \rangle$; $f_{i+1} = f_i \circ f_i$	Tower of commuting pairs
Hyperbolic 3-manifold $S \times [0,1]/\psi$ fibering over the circle	Fixed-points of Renormalization	
Conformal structure is $C^{1+\alpha}$-rigid at deep points $\Longrightarrow$ Renormalization converges exponentially fast		
M is asymptotically rigid	$J(f)$ is self-similar at the critical point $c_0(f)$	

Table 3.

unbounded renormalization periods and Liouville rotation numbers. We conclude with an example of such a complete classification in the setting of hyperbolic geometry.

Let S be the compact surface obtained by removing a disk from a torus. Let $AH(S) \subset V(S)$ be the set of discrete faithful representations such that $\rho(\pi_1(\partial S))$ is parabolic. A representation $\rho : \pi_1(S) \to \Gamma$ in $AH(S)$ gives a hyperbolic manifold $M = \mathbb{H}^3/\Gamma$ homeomorphic to $\mathrm{int}(S) \times \mathbb{R}$. To each end of M one can associate an *end invariant*

$$E^{\pm}(M) = \begin{cases} \partial^{\pm}(M) \in \mathrm{Teich}(S) & \text{or} \\ \epsilon^{\pm}(M) \in \mathbb{P}\mathcal{ML}(S). \end{cases}$$

In the first case the end is naturally completed by a hyperbolic punctured torus $\partial^{\pm}(M)$; in the second case the end is pinched along a simple curve or lamination $\epsilon^{\pm}(M)$.

Identifying $\mathrm{Teich}(S) \cup \mathbb{P}\mathcal{ML}(S)$ with $\overline{\mathbb{H}} = \mathbb{H} \cup \mathbb{R} \cup \infty$, we may now state:

Theorem 7.1 (Minsky) *The pair of end invariants establishes a bijection*

$$E : AH(S) \to \overline{\mathbb{H}} \times \overline{\mathbb{H}} - \overline{\mathbb{R}} \times \overline{\mathbb{R}}$$

with E^{-1} continuous.

Corollary 7.2 *Each Bers' slice of $AH(S)$ is bounded by a Jordan curve naturally parameterized by $\mathbb{R} \cup \infty$, with rational points corresponding to cusps.*

Corollary 7.3 *Geometrically finite manifolds are dense in $AH(S)$.*

Theorem 7.1 establishes a special case of Thurston's *ending lamination conjecture* [Mc1, §4]. We remark that E is *not* a homeomorphism, and indeed $AH(S)$ is not even a topological manifold with boundary [Mc3, Appendix].

The proof of Theorem 7.1 from [Min] can be illustrated in the case $E(M) = (\tau, \lambda)$, with $\tau \in \mathbb{H}$ and $\lambda \in \mathbb{R}$ an irrational number with continued fraction $[a_1, a_2, \dots]$. By rigidity of manifolds in $\partial AH(S)$, it suffices to construct a quasi-isometry

$$\phi : M \to M(a_1, a_2, \dots)$$

from M to a model Riemannian manifold explicitly constructed from the ending invariant. The quasi-isometry is constructed piece by piece, over blocks M_i of M corresponding to terms a_i in the continued fraction.

The construction yields a description not only of manifolds in $AH(S)$, but also of their geometric limits, which we formulate as follows.

Theorem 7.4 *Every geometric limit $M = \lim M_n$, $M_n \in AH(S)$, is determined up to isometry by a sequence $\langle a_i, i \in I \rangle$, where*

- *$I \subset \mathbb{Z}$ is a possibly infinite interval,*
- *$a_i \in \mathrm{Teich}(S) \cup \{*\}$ if i is an endpoint of I; and*
- *$a_i \in \{1, 2, 3, \dots, \infty\}$ otherwise.*

Here $\langle a_i \rangle$ should be thought of as a generalized continued fraction, augmented by Riemann surface data for the geometrically finite ends of M. (The special point $\{*\}$ is used for the triply-punctured sphere.)

For example, the sequence $\langle a_i \rangle = \langle \ldots, \infty, \infty, \infty, \ldots \rangle$ determines the periodic manifold

$$M_\infty \cong \mathrm{int}(S) \times \mathbb{R} - \left(\bigcup_{\mathbb{Z}} \gamma_i \times \{i\} \right),$$

where $\gamma_i \subset S$ are simple closed curves and $i(\gamma_i, \gamma_{i+1}) = 1$. These curves enumerate the rank two cusps of M_∞. Geometrically, M_∞ is obtained from the Borromean rings complement $S^3 - B$ (itself a hyperbolic manifold) by taking the $\mathbb{Z}$-covering induced by the linking number with one component of B.

In general the coefficients $\langle a_i \rangle$ in Theorem 7.4 specify how to obtain M by Dehn filling the cusps of M_∞. Compare [Th2, §7].

Corollaries 7.2 and 7.3 are reminiscent of two open conjectures in dynamics: the local connectivity of the Mandelbrot set, and the density of hyperbolicity for complex quadratic polynomials.

Quadratic polynomials, however, present an infinite variety of parabolic bifurcations, in contrast to the single basic type occurring for punctured tori. This extra diversity is reflected in the topological complexity of the boundary of the Mandelbrot set, versus the simple Jordan curve bounding a Bers slice.

Parabolic bifurcations can be analyzed by Ecalle cylinders [Dou2] and parabolic towers [Hin], both instances of geometric limits as in §3. A complete understanding of complex quadratic polynomials will likely entail a classification of all their geometric limits as well.

References

[BJ1] C. J. Bishop and P. W. Jones. Hausdorff dimension and Kleinian groups. *Acta Math.* 179(1997), 1–39.

[BJ2] C. J. Bishop and P. W. Jones. The law of the iterated logarithm for Kleinian groups. In *Lipa's Legacy (New York, 1995)*, volume 211 of *Contemp. Math.*, pages 17–50. Amer. Math. Soc., 1997.

[BH] X. Buff and C. Henriksen. Scaling ratios and triangles in Siegel disks. *Preprint, Cornell, 1998.*

[CaTh] J. W. Cannon and W. P. Thurston. Group invariant Peano curves. *Preprint, Princeton, 1985.*

[dFdM] E. de Faria and W. de Melo. Rigidity of critical circle mappings I, II. *SUNY IMS Preprints 1997/16,17.*

[DGK] T. W. Dixon, T. Gherghetta, and B. G. Kenny. Universality in the quasiperiodic route to chaos. *Chaos* 6(1996), 32–42.

[Dou1] A. Douady. Disques de Siegel et anneaux de Herman. In *Séminaire Bourbaki, 1986/87*, pages 151–172. Astérisque, volume 152-153, 1987.

[Dou2] A. Douady. Does a Julia set depend continuously on the polynomial? In R. Devaney, editor, *Complex Analytic Dynamics*. AMS Proc. Symp. Appl. Math. 49, 1994.

[DH] A. Douady and J. Hubbard. On the dynamics of polynomial-like mappings. *Ann. Sci. Éc. Norm. Sup.* 18(1985), 287–344.

[Hin] B. Hinkle. Parabolic limits of renormalization. *SUNY IMS Preprint 1997/7*.

[Lan] O. E. Lanford III. Renormalization group methods for circle mappings with general rotation number. In *VIIIth International Congress on Mathematics Physics (Marseille, 1986)*, pages 532–536. World Scientific, 1987.

[JH] Y. Jiang and J. Hu. The Julia set of the Feigenbaum quadratic polynomial is locally connected. *Preprint, 1993*.

[Kap] M. Kapovich. On dynamics of pseudo-Anosov homeomorphisms on the representation variety of surface groups. *Ann. Acad. Sci. Fenn. Math.* 23(1998), 83–100.

[Lev] G. Levin. Bounds for maps of an interval with one reflecting critical point. I. *Fund. Math., To appear*.

[LS] G. Levin and S. van Strien. Local connectivity of the Julia set of real polynomials. *SUNY IMS Preprint 1995/5; Annals of Math., To appear*.

[Lyu] M. Lyubich. Feigenbaum-Coullet-Tresser universality and Milnor's hairiness conjecture. *Annals of Math., To appear*.

[FKS] L. P. Kadanoff M. J. Feigenbaum and S. J. Shenker. Quasiperiodicity in dissipative systems: a renormalization group analysis. *Phys. D* 5(1982), 370–386.

[Mak] R. S. MacKay. *Renormalisation in Area-Preserving Maps*. World Scientific, 1993.

[MN] N. S. Manton and M. Nauenberg. Universal scaling behavior for iterated maps in the complex plane. *Commun. Math. Phys.* 89(1983), 555–570.

[Mc1] C. McMullen. The classification of conformal dynamical systems. In *Current Developments in Mathematics, 1995*, pages 323–360. International Press, 1995.

[Mc2] C. McMullen. *Renormalization and 3-Manifolds which Fiber over the Circle*. Annals of Math Studies 142, Princeton University Press, 1996.

[Mc3] C. McMullen. Complex earthquakes and Teichmüller theory. *J. Amer. Math. Soc.* 11(1998), 283–320.

[Mc4] C. McMullen. Self-similarity of Siegel disks and the Hausdorff dimension of Julia sets. *Acta Math., To appear, 1998.*

[Mil] J. Milnor. Self-similarity and hairiness in the Mandelbrot set. In M. C. Tangora, editor, *Computers in Geometry and Topology*, Lect. Notes Pure Appl. Math., pages 211–259. Dekker, 1989.

[Min] Y. Minsky. The classification of punctured torus groups. *SUNY IMS Preprint 1997/6; Annals of Math., To appear.*

[Pet] C. L. Petersen. Local connectivity of some Julia sets containing a circle with an irrational rotation. *Acta Math.* 177(1996), 163–224.

[Rand] D. Rand. Universality and renormalisation in dynamical systems. In *New Directions in Dynamical Systems*, pages 1–56. Cambridge University Press, 1988.

[Sul1] D. Sullivan. Growth of positive harmonic functions and Kleinian group limit sets of zero planar measure and Hausdorff dimension two. In *Geometry at Utrecht*, Lecture Notes in Math. 894, pages 127–144. Springer-Verlag, 1981.

[Sul2] D. Sullivan. Conformal dynamical systems. In *Geometric Dynamics*, Lecture Notes in Math 1007, pages 725–752. Springer-Verlag, 1983.

[Sul3] D. Sullivan. Quasiconformal homeomorphisms in dynamics, topology and geometry. In *Proceedings of the International Conference of Mathematicians*, pages 1216–1228. Amer. Math. Soc., 1986.

[Sul4] D. Sullivan. Bounds, quadratic differentials and renormalization conjectures. In F. Browder, editor, *Mathematics into the Twenty-first Century: 1988 Centennial Symposium, August 8-12*, pages 417–466. Amer. Math. Soc., 1992.

[Sw] G. Świątek. On critical circle homeomorphisms. *Preprint, 1997.*

[Th1] W. P. Thurston. *Geometry and Topology of Three-Manifolds.* Lecture Notes, Princeton University, 1979.

[Th2] W. P. Thurston. Hyperbolic structures on 3-manifolds II: Surface groups and 3-manifolds which fiber over the circle. *Preprint.*

[Wid] M. Widom. Renormalization group analysis of quasi-periodicity in analytic maps. *Commun. Math. Phys.* 92(1983), 121–136.

[Y] J.-C. Yoccoz. Il n'y a pas de contre-exemple de Denjoy analytique. *C. R. Acad. Sci. Paris Sér. I Math.* 298(1984), 141–144.

Curtis T. McMullen
Mathematics Department
Harvard University
1 Oxford St
Cambridge, MA 02138

Induced Hyperbolicity for One-Dimensional Maps

Grzegorz Świątek[1]

Abstract. We present a review of Yoccoz partitions and their relation with induced dynamics. A new way of interpreting the construction is shown, based on external Yoccoz partitions. These are governed by linear dynamics and hence much easier to handle. As an example of the usefulness of this method we prove the following result: For almost every parameter c on the boundary of the Mandelbrot set, in the sense of the harmonic measure, the map z^2+c satisfies the Collet-Eckmann condition.

1991 Mathematics Subject Classification: 30C10, 30D05
Keywords and Phrases: Yoccoz partition, Collet-Eckmann condition

1 Yoccoz Partitions

1.1 Historical outline

In the early 1980s M. Jakobson proved a theorem which asserted that in the logistic family $x \to ax(1-x)$, the mapping has a probabilistic absolutely continuous invariant measure for a set of parameters a with positive measure, see [10]. The crucial step of the proof was the construction of an expanding map Φ, defined on the union of countably many intervals, so that on each connected component of its domain Φ was an iterate of the original quadratic map, and the range of this restriction was a fixed interval. Hence, from the original clearly non-expanding transformation, a uniformly expanding one was constructed by taking iterates in a piecewise fashion.

In the early 1990s J.-Ch. Yoccoz proved a theorem about local connectivity of Julia sets of some quadratic polynomials, including many Julia sets which contained the critical point. He also showed that for this class of polynomials certain combinatorial data, which for real maps reduces to the kneading sequence, determines that polynomial uniquely. The proof is based on the construction of partitions of the phase space which are not far from being Markov: most pieces are mapped in a univalent way onto other pieces, except for the pieces which contain the critical point. The partitions were subjected to a process of infinite inductive refinement under which the critical pieces shrank to the point.

It was then observed that a powerful tool for studying both unimodal maps of the interval and complex polynomials in the plane is obtained when these ideas

[1]Partially supported by NSF Grant DMS-9704368

are applied jointly. When appropriate iterations of the original polynomial are applied on pieces of Yoccoz partitions, the result is a map which is expanding on all pieces except for the one which contains the critical point. Such transformations are called *induced mappings*. When a sequence of increasingly refined partitions is considered, the critical branches disappear in the limit and an expanding map of Jakobson's type is obtained. The original construction of Jakobson required exclusion of unsuitable parameters without regard for their topological dynamics, although a set of positive measure was left at the end. The new approach based on Yoccoz' discovery works for all polynomials without neutral or attracting periodic orbits, with the only exception of a well defined class of infinitely tunable ones.

This, or a closely related approach, played the key role in the solution of several important problems. For real quadratic polynomials, and by extension for unimodal maps with non-degenerate critical point, the infinitely tunable case turned out to be quite manageable, due to the phenomenon of the so-called a priori bounds, discovered by D. Sullivan. The combination of inducing and a priori bounds was the basis of the result about local connectivity of Julia sets for all real unimodal polynomials, see [14]. The same ingredients, and another phenomenon related to inducing and known as the *decay of geometry*, were used in [7] to prove that periodic windows are dense in the logistic family. Alternative proofs of both results can be found in [15].

Finally, there is a new result we wish to present which use inducing applied to complex polynomials. Recall that the *Collet-Eckmann* condition for a quadratic polynomial $f_c(z) := z^2 + c$ is that

$$\liminf_{n\to\infty} \frac{\log |Df_c^n(c)|}{n} > 0 \,.$$

Let χ denote the harmonic measure on the boundary of the Mandelbrot set.

THEOREM 1.1 *For χ-almost every point c, $z^2 + c$ satisfies the Collet-Eckmann condition.*

Theorem 1.1 is joint with J. Graczyk and implies the result earlier announced by Graczyk, Smirnov and the author, that for χ-almost every c and every $\alpha > 0$

$$\sum_{i=0}^{\infty} |Df_c^i(c)|^{-\alpha} < \infty \,.$$

The Collet-Eckmann condition for rational maps has been studied, see [5] and [16]. One of the results of these papers is that a map which satisfies the Collet-Eckmann condition has the Julia set with Hausdorff dimension less than 2. Thus, we get a corollary complementary to a theorem of which asserts that a complex polynomial in the residual subset of the boundary of the Mandelbrot set has the Julia set with Hausdorff dimension equal to 2, see [17].

STRUCTURE OF THE PAPER. Section 1 is a review of results about Yoccoz partitions in the context of their connections with inducing. Section 2 contains some new material, including an outline of the proof of Theorem 1.1.

1.2 Refining of partitions

Suppose that f is a rational mapping of the Riemann sphere. Consider a set B_0 which is a union of disjoint Jordan domains $B_0^1, \cdots, B_0^k$ chosen is such a way that each B_0^i contains exactly one critical point of f. Furthermore, assume that $f^n(\partial B_0) \cap B_0 = \emptyset$ for all $n > 0$. The question of how B_0 can be found will be adressed in Section 2. For now, let us describe how this initial partition can be refined into an infinite sequence of partitions.

We will regard B_0^i, $i = 0, \cdots, k$ as *Yoccoz pieces* of order 0. Next, we proceed recursively. If B_m^i is a Yoccoz piece of order m, then any connected component of $f^{-1}(B_m^i)$ is a Yoccoz piece of order $m+1$.

Trivially, all Yoccoz pieces of order m are disjoint and each is mapped by f^m onto some B_0^i as a proper holomorphic map. A more interesting property is that if B_m^i and $B_{m'}^{i'}$ have a non-empty intersection, then one of these pieces is contained in the other. To see this, assume first that $m = 0$. Then, if the claim were violated, we would have $m' > 0$ and $B_{m'}^{i'} \cap \partial B_0^i \neq \emptyset$. But then after m' iterations that part of the boundary of B_0^i would come back to B_0, contrary to our hypothesis. In general, the claim follows by induction with respect to m.

The disjoint union of all Yoccoz pieces is often called a *Yoccoz puzzle*. The discovery of Yoccoz was proving that puzzles derived from a suitably chosen B_0 for quadratic polynomials are often generating. By "generating" we mean that for every point z of the Julia set of f, whenever the ω-limit set of z contains some critical points, one can find a nesting sequence of Yoccoz pieces which intersect down to $\{z\}$. For quadratic polynomials, this will happen whenever the polynomial is non-tunable with all periodic orbits repelling. In some cases, a single Yoccoz puzzle is not enough to be generating, but one can construct a sequence of Yoccoz puzzles and show that they are jointly generating. This happens in the proof of local connectivity of Julia sets for real unimodal polynomials, see [14].

Induced dynamics. The Yoccoz puzzle does not have canonical dynamics. We might try the natural map which sends $z \in B_m^i$ to $f(z) \in f(B_m^i)$ which is usually well defined since $f(B_m^i)$ is a Yoccoz piece of order $m-1$. This breaks down, however, when $m = 0$. The only way to continue is to "drop" z to some Yoccoz piece of higher order, and since a point typically belongs to infinitely many pieces, one faces a difficult choice. A general scheme for endowing Yoccoz pieces with dynamics is as follows.

Let us call a *Yoccoz partition* any collection $\mathcal{B}$ of pairwise disjoint Yoccoz pieces. On any element B_m^i choose an integer t_m^i subject to the condition $0 \leq t_m^i \leq m$. Then the map which is defined on the union of elements of the partition by

$$\phi(z) = f^{t_m^i}(z)$$

for $z \in B_m^i$ will be described as *induced* by f on the Yoccoz partition $\mathcal{B}$. The restriction of ϕ to any connected component of its domain will be called a *branch* of ϕ. This scheme is quite general. In particular, it incorporates various inducing constructions for unimodal maps, see [10] or [9]. If one applies these construc-

tions to a unimodal real quadratic polynomial, one gets induced maps on Yoccoz partitions restricted to the real line.

Suppose now that ζ is a branch induced by f on some Yoccoz piece, and ϕ is an unrelated map induced on some Yoccoz partition. Then $\phi \circ \zeta$, considered on the set of all points where it is well defined, is automatically an induced map defined on another Yoccoz partition. The consequence of this remark is that once a single induced map on some Yoccoz partition was obtained, we can use it as a starting point for an "inducing construction" in which we "refine" branches by composing them with other induced maps. Everything which we can obtain in this way will still be induced on Yoccoz partitions. Authors often don't talk about a Yoccoz puzzle directly, but instead proceed to define an inducing process. If such a process begins with a map induced on a Yoccoz partition, it implicitly belongs to our framework.

GEOMETRY OF CRITICAL PIECES. Yoccoz pieces which contain a critical point of f will be designated as *critical*. The geometry of critical Yoccoz pieces has been particularly carefully investigated. A curious discovery in this area was the so-called *decay of geometry* for quadratic polynomials.

To explain the setting, recall the mapping f with critical points $Z_1, \cdots, Z_k$ and the corresponding system of order 0 Yoccoz pieces B_0^i, with $Z_i \in B_0^i$. We will regard B_0^i as the 0-th generation of critical pieces. Recursively, suppose that the p-th generation of critical Yoccoz pieces D_p^i with $D_p^i \ni Z_i$ has been defined. Then D_{p+1}^i is the largest piece inside D_p^i, not equal to D_p^i, which contains Z_i and is mapped onto some D_p^j by an iterate of f. This iterate will be denoted with $f^{r_{p+1}^i}$. For each $1 \le i \le k$ we choose a subsequence $p_q(i)$ with $p_1(i) = 1$ and

$$p_q(i) = \min\{p : \ r_p^i > r_{p_{q-1}(i)}^i\}$$

for $q > 1$. This set could be empty, in which case $p_q(i)$ are all undefined from some point on. A trivial reason why it might happen is when D_p^i are defined for only finitely many p. But D_p^i may also be defined for all p, and $p_q(i)$ may still be undefined from some q on if the map is tunable.

DECAY OF GEOMETRY. *Suppose that f is a quadratic polynomial and consider the sequence, perhaps finite, of critical pieces D_{p_q}. For every $\alpha > 0$ there is $C > 0$ for which the following estimate holds. If $mod(D_0 \setminus \overline{D}_1) \ge \alpha$, then for every $q \ge 1$ for which p_q is well defined*

$$mod(D_{p_q - 1} \setminus \overline{D}_{p_q}) \ge Cq \ .$$

This fact depends on f having only one simple critical point. It is may not be true when the critical point is degenerate, see [12], and is generally not true in the presence of more than one critical point, see [18]. In both counterexamples the sequence of moduli not only does not increase at a linear rate, but remains bounded for all q, which are infinitely many. The decay of geometry has been

established by a sequence of works. In the form in which we state it, or equivalent, it was shown in [6] for Yoccoz pieces of real quadratic polynomials, in [15] for complex polynomials, and in [8] in a more general context of holomorphic box mappings. The decay of geometry plays a role in the proof the periodic windows are dense in the logistic family, see [7] and [15].

In his original approach Yoccoz relied on a weaker version of the decay of geometry, which simply states that whenever the sequence of critical pieces D_{p_q} is infinite, intersect down to $\{0\}$. Even this relies on exponent 2. While Yoccoz' work remains unpublished, a similar reasoning which shows the importance of exponent 2 is presented in [1] for cubic polynomials.

2 External Induced Maps

2.1 Construction

Invariant curves in the phase space. Let us go back to the question which was pushed aside in the previous section, and namely how to obtain the Yoccoz pieces of level 0. The construction will be described for polynomials of the form $f(z) = z^d + c$ with $d > 1$. Since there is only one piece of order 0, we will denote it with B_0 rather than B_0^1 which would in keeping with the notation of Section 1. A trivial solution would be to choose as B_0 the interior of a geometric disk centered at 0 with sufficiently large radius. Yoccoz partitions obtained in this way would not be interesting in the case when the Julia set is connected, since they would simply form a sequence of topological disks nesting down to the filled-in Julia set. A better construction, due to Yoccoz, uses external rays of the Julia set.

Historically, the idea of constructing dynamical partitions using some invariant curves had certainly appeared before Yoccoz. Back in [3], one finds a construction in which a chain of preimages of a point is considered. When these points are joined by arcs, and invariant curve can be obtained, which will often land at a periodic repelling point. External rays were considered more recently, [2]. In addition to providing an invariant family of curves, they show a connection with the Riemann map of the complement of the connectedness locus in the parameter space.

Construction of the initial partition. A fixed point of the map $z^d + c$ is called *admissible* if it is repelling and is the landing point of finitely many external rays, all of which are smooth and have non-zero external arguments. The parameter c is *admissible* provided that $z^d + c$ has an admissible fixed point. It should be emphasized that the orbit of c may escape to ∞; all that is needed for the construction to work is that c be admissible. Note that an admissible fixed point persists under a small perturbation of parameters, and that the external arguments of rays which converge at such a point don't change. Any repelling fixed point which is not the landing point of the ray with external argument 0 is going to be admissible unless a ray which lands there is not smooth. For $d = 2$, this restricts non-admissible parameters to those which lie on external rays of the Mandelbrot set landing on the boundary of the main hyperbolic component. The

set of such external arguments is of zero measure, see [4], Remark C.5. For the definition of external rays and basic facts about them, see [13].

The rays which converge at an admissible fixed point divide the plane into several sectors, one of which, say S_0 contains 0. Choose a level curve of the Green function of the Julia set, of a level L large enough so that it separates 0 from ∞. The intersection of the interior connected component of the complement of this curve with S_0 is the initial Yoccoz piece B_0. It is trivial to verify that B_0 has all properties assumed in Section 1.

BOUNDARIES OF YOCCOZ PIECES. The boundary of every Yoccoz piece B_m^i is a finite union of analytic arcs, some of which are preimages of the rays while others are arcs of the level curve of level L/d^m. Assuming that the level of the critical point is less than L/d^m, each arc of the level curve intercepts rays with external arguments from an open interval. In this way, with every Yoccoz piece B_m^i we can associate its "external" piece b_m^i defined as the interior of the set of external arguments for which the corresponding rays hit the piece. Every external Yoccoz piece is a finite union of intervals. The set of external Yoccoz pieces still has the property that two of them are either disjoint, or one is contained in the other. Even more interesting is the connection between the dynamics induced by f on the Yoccoz pieces and the dynamics induced by the map $T(x) = dx \bmod 1$ on external Yoccoz pieces. If $f^{m'}(B_m^i) = B_{m-m'}^{i'}$, then $T^{m'}(b_m^i) = b_{m-m'}^{i'}$.

INTRINSIC CONSTRUCTION OF EXTERNAL YOCCOZ PIECES. What is most striking is that external Yoccoz pieces can be constructed without reference do the dynamics of f, provided that B_0 has been chosen, which determines b_0, and that an external argument of 0 is known, say γ. When c is not in the connectedness locus, then 0 in fact has d external rays that meet it. The point γ can be chosen to be any of them and the choice will not affect the construction, since it is only $T(\gamma)$ that matters.

Assume first that c is not in the connectedness locus, at level $\lambda > 0$ with respect to the Green function of its Julia set, and choose $L >> \lambda$ to construct the initial Yoccoz piece B_0. Then b_0 establishes the set of external Yoccoz pieces of level 0. Now suppose that pieces b_m^i of level m have been defined. The critical piece of level $m+1$, which contains γ, is simply $T^{-1}(b_m^{i_0})$ where $b_m^{i_0} \ni T(\gamma)$, provided that such $b_m^{i_0}$ exists. Any other piece is $T_j^{-1}(b_m^i)$ where T_j^{-1} is an inverse branch of T mapping onto $S^1 \setminus \{\gamma\}$. By induction, we check that this is the "right" construction in the following sense. For $m < \frac{\log(L/\lambda)}{\log d}$ each b_m^i is the set of external rays intercepted by some B_m^i, and the piece b_m^i which contains γ corresponds in this way to the critical piece of level m.

Now let c tend to the connectedness locus along a fixed external ray in the parameter plane. This means that the external argument of c in the *phase* plane of $z^d + c$ remains fixed, see [2], and hence γ can be chosen fixed. At the same time, λ tends to 0. Since γ and b_0 are fixed, external Yoccoz pieces and the dynamics induced on them by T don't change at all. The correspondence with regular Yoccoz pieces holds for larger and larger levels m. In the limit, for c in the closure of the

ray and on the boundary of the connectedness locus, the correspondence holds for all m, unless Yoccoz pieces B_m^i undergo a discontinuous change (in Hausdorff topology). Such a change is only possible if the orbit of 0 hits the boundary of some Yoccoz piece, and hence of B_0. But now the limit value c_0 is in the connectedness locus, so the only point on the boundary of B_0 which can be met by the critical orbit is the fixed point. As we see, the correspondence holds except for a set of external rays of the connectedness locus with rational external arguments (recall that by custom external arguments are parametrized by the circle of *circumference* 1.)

Summary of the correspondence. The foregoing discussion can be summarized as follows.

Proposition 1 *Let γ be irrational and c be an admissible parameter in the intersection of the boundary of the connectedness locus with the closure of the external ray with argument γ. Consider an admissible fixed point q of $z^d + c$ and let b_0 denote the arc of external arguments of all rays contained in the sector which is cut off by two adjacent rays landing at q and contains 0. Then there is a one-to-one correspondence between Yoccoz pieces constructed from the rays converging at q and some equipotential curve and external Yoccoz pieces of the map $T(x) = dx \bmod 1$ with distinguished point γ and with the order 0 piece b_0. Key properties of this correspondence are:*

- *it conjugates the action on Yoccoz pieces by f to the action on external Yoccoz pieces by T,*
- *it respects inclusion and inclusion with closure,*
- *critical Yoccoz pieces correspond exactly to the external pieces which contain γ.*

The usefulness of this correspondence lies in the fact that the dynamics of T is much simpler and hence external Yoccoz pieces can be kept track of more easily than Yoccoz pieces. Properties of external Yoccoz puzzles which hold typically with respect to the Lebesgue measure on γ translate to properties which are typically true with respect to the harmonic measure on the boundary of the Mandelbrot set. This is the main line of the proof of Theorem 1.1.

2.2 Collet-Eckmann condition

We will now sketch main steps of the proof of Theorem 1.1. Consider a connected component A of the set of admissible parameters. In the quadratic case, the initial external Yoccoz piece b_0 remains fixed throughout A. Since the whole boundary of the Mandelbrot set, except for the boundary of main cardioid which is of zero harmonic measure, is in the admissible set, it will be enough to restrict our considerations once and for all to A. In b_0, pick a point γ. This corresponds to choosing an external ray of the Mandelbrot set. There is a map $\Gamma : b_0 \to A \cap \partial M$ well defined almost everywhere, which associates to γ the landing point of the

external ray with argument $T(\gamma)$. The harmonic measure on $A \cap \partial M$ is simply $\Gamma_*(\lambda)$, where λ is the Lebesgue measure on the circle.

THE MAIN CONSTRUCTION. For any $\gamma \in b_0$ and irrational consider the sequence d_q of all external pieces which contain γ, in the order of inclusion. Hence, $d_0 := b_0$. For c in the Mandelbrot set and in the closure of the external ray with argument γ, note the corresponding sequence of all critical pieces D_q. Let $k(q)$ be the smallest positive k for which $T^k(d_q) \ni \gamma$. Clearly, $k(q)$ form a non-decreasing sequence. An external Yoccoz piece b will be called *nested* if whenever $b \subset d_q$ for some q, then $\overline{b} \subset d_q$. An analogous definition applies to Yoccoz pieces induced by f and by Proposition 1 the correspondence between external and regular pieces observes the property of being nested.

PROPOSITION 2 *There is a set $A' \subset A$, with $A \setminus A'$ of zero measure, with the following properties. For every $\gamma \in A'$, infinitely many critical pieces are defined. Furthermore, for every such γ there are a positive integer m_0, a finite Yoccoz partition $\mathcal{B}$ with all pieces nested, and a sequence n_j with the following properties:*

1. *for every j, $T^{k(n_j)}$ maps d_{n_j} onto some d_q, with $q \le m_0$; moreover, $T^{k(n_j)}(\gamma)$ belongs to some element of $\mathcal{B}$,*
2. $$\lim_{j\to\infty} \frac{k(n_{j+1})}{k(n_j)} = 1 \,,$$
3. $$\limsup_{j\to\infty} \frac{k(n_j)}{j} < \infty \,.$$

DERIVATION OF THE COLLET-ECKMANN CONDITION. These conditions are seen to imply the Collet-Eckmann condition at all points c in the boundary of the Mandelbrot set intersected with the closure of the external ray with argument γ. The first condition is easily understood in terms of the corresponding maps induced by f. It says that for each j the critical piece D_{n_j} is mapped by $f_c^{k(n_j)}$ onto a large D_q, while 0 is mapped into a piece which corresponds to an element of $\mathcal{B}$ in the sense of Proposition 1. Since $\mathcal{B}$ consists of nested pieces and is finite, it follows that $f^{k(n_j)}(0)$ is separated from the boundary of D_q by some positive distance independent of j. This plays a double role. First, the distortion at the critical value c is bounded, so that

$$\log |Df^{k(n_j)-1}(c)| \ge K_1 \bmod (D_{m_0} \setminus \overline{D}_{n_j}) \tag{1}$$

with $K_1 > 0$. Secondly, since $f^{k(n_j)}(0)$ gets into a nested piece of some fixed Yoccoz partition, the modulus surrounding this piece in D_q will be pulled back, resulting in

$$\bmod (D_{n_j} \setminus \overline{D}_{n_j+q_0}) \ge K_2$$

for all j, K_2 positive and q_0 constant. Hence,

$$\liminf \frac{\mathrm{mod}\,(D_{m_0} \setminus \overline{D}_{n_j})}{j} > 0\,.$$

Together with the last assertion of Proposition 2 this implies

$$\liminf \frac{\mathrm{mod}\,(D_{m_0} \setminus \overline{D}_{n_j})}{k(n_j)} > 0\,.$$

From this and estimate (1) we see that

$$\liminf \frac{\log |Df^{k(n_j)-1}(c)|}{k(n_j)} > 0\,.$$

The second claim of Proposition 2 easily shows that the same condition in fact holds for all k, not just the subsequence $k(n_j)$, and hence the Collet-Eckmann condition is established.

A probabilistic reasoning. As to the proof of Proposition 2, it is useful to think of the "excursion times" from the large scale, $k(n_{j+1})-k(n_j)$, as independent random variables. Let us also forget at first that γ has to be mapped into an element of $\mathcal{B}$, instead just try to make sure that $T^{k(n_j)}$ maps to the large scale. The probabilistic interpretation is intuitively natural, because the next excursion depends mostly on $T^{k(n_j)}(\gamma)$, and that it is independent of the length of previous excursions, since they all end with γ mapped somewhere in the large box with uniform distribution. With that interpretation, it is enough to see that these random variables have finite expectation and variance and use the strong law of large numbers. These estimates follows rather easily once $\mathcal{B}$ is chosen carefully. The interpretation of the excursion times as independent random variables requires a careful construction. Details will be provided in a forthcoming paper.

The role of exponent 2. Theorem 1.1 is stated only for the quadratic family. However, the proof does not rely on exponent 2 is an essential way. In particular, the decay of geometry of critical Yoccoz pieces is never used a priori, it simply follows from the conditions of Proposition 2. The reason why the theorem is not stated for the family $z^d + c$ is that certain basic facts remain to be checked in this more general case. For example, it should be proved that almost all parameters in the complement of the connectedness locus are admissible.

References

[1] Branner, B. & Hubbard, J.H.: *The iteration of cubic polynomials. Part II: patterns and parapatterns*, Acta Math. 169 (1992), 229-325

[2] Douady, A.: *Systemès Dynamiques Holomorphes*, Astérisque 105 (1982), 39-63

[3] Fatou, P.: *Sur les équations fonctionelles*, Bull. Soc. Math. Fr. 47 (1919), 161-271; 48 (1920), 33-94; 208-314

[4] Goldberg, L. & Milnor, J.: *Fixed points of polynomial maps. Part II. Fixed point portraits*, Ann. Sc. Éc. Norm. Sup. 26 (1993), 51-98

[5] Graczyk, J. & Smirnov, S.: *Collet, Eckmann, & Hölder*, Invent. Math. 133 (1998), 69-96

[6] Graczyk, J. & Świątek, G.: *Induced expansion for quadratic polynomials*, Ann. Scient. Éc. Norm. Sup. 29 (1996), 399-482

[7] Graczyk, J. & Świątek, G.: *Generic hyperbolicity in the logistic family*, Ann. of Math 146 (1997), 1-56

[8] Graczyk, J. & Świątek, G.: *Holomorphic box mappings*, preprint IHES (1996), to appear in Astérisque

[9] Guckenheimer, J. & Johnson, S.: *Distortion of S-unimodal maps*, Ann. of Math. 132 (1990), 71-130

[10] Jakobson, M.: *Absolutely continuous invariant measures for one-parameter families of one-dimensional maps*, Commun. Math. Phys. 81 (1981), 39-88

[11] Jakobson, M. & Świątek, G.: *Metric properties of non-renormalizable S-unimodal maps*, Ergod. Th. Dyn. Sys. 14 (1994), 721-755

[12] Keller, G. & Nowicki, T.: *Fibonacci maps re(al) visited*, Erg. Th. & Dyn. Sys. 15 (1995), 97-130

[13] Levin, G. & Przytycki, F.: *External rays to periodic points*, Israel Jour. Math. 94 (1995), 29-57

[14] Levin, G. & Van Strien, S.: *Local connectivity of the Julia set of real polynomials*, manuscript (1994), to appear in Ann. of Math.

[15] Lyubich, M.: *Dynamics of complex polynomials, part I-II*, Acta Math., 178 (1997), 185-297

[16] Przytycki, F. & Rohde, S.: *Porosity of Collet-Eckmann Julia sets*, Fund. Math. 155 (1998), 189-199

[17] Shishikura, M.: *The Hausdorff dimension of the boundary of the Mandelbrot set and Julia sets*, Ann. of Math. 147 (1998), 225-267

[18] Świątek, G. & Vargas, E.: *Decay of geometry in the cubic family*, Penn State preprint (1996), to appear in Erg. Th. Dyn. Sys.

Grzegorz Świątek
The Pennsylvania State University
Mathematics Department
209 Mc Allister
University Park, PA 16802, USA

Arnold Diffusion: A Variational Construction

Zhihong Xia[1]

Abstract. We use variational method to study Arnold diffusion and instabilities in high dimensional Hamiltonian systems. Our method is based on a generalization of Mather's theory on twist maps and their connecting orbits to a higher dimensional setting. Under some generic nondegeneracy conditions, we can construct transition chains of arbitrary fixed length, crossing gaps of any size between invariant KAM (lower dimensional) tori. One of notable features of our result is that, instead of using transition tori alone for diffusion as in Arnold's construction, we also use cantori from Aubry-Mather theory in our mechanism for diffusion. Other results, such as shadowing properties, symbolic dynamics and transitivity, etc., can also be obtained by our method. Our nondegeneracy condition is a condition on the splitting of separatrix and in the so-called *a priori* unstable systems, this condition can be verified by the so-called Poincaré-Melnikov integrals.

In Arnold's original example for the instability, the perturbation is carefully chosen so that it does not touch any invariant tori on the normally hyperbolic invariant manifold. As an application of our results, we can choose arbitrary perturbations and are able to conclude the same results (in fact stronger), as long as the Poincaré-Melnikov integrals are in some sense non-degenerate.

1 Introduction

Perhaps one of the most important problems in Hamiltonian dynamics, after the celebrated KAM theory, is the topological stability of near integrable Hamiltonian systems. KAM theory completely answered the problem for two-degree of freedom autonomous Hamiltonian systems, where we have generic stability. However, the higher dimensional situation is much more complicated. A standing conjecture, due to Arnold (cf. [2], [6]), is that genericly we have topological instability. To support his conjecture, Arnold [1] gives an example of a two-degree of freedom, time-periodically forced Hamiltonian system where arbitrary small perturbation produces orbits whose action variables changes arbitrarily in size, resulting in a phenomenon known as *Arnold diffusion.*

[1]Research is supported in part by National Science Foundation.

The mechanism in Arnold's example for diffusion uses transition tori, which are lower dimensional (weakly) hyperbolic KAM tori. In typical situations, some of these lower dimensional invariant tori (near resonances) break, hence breaking the transition chain, which severely limits the size of diffusion. This difficulty is termed as the gap problem and it is one of the main problems in Arnold diffusion. To avoid this difficulty in his example, Arnold chooses a very special perturbation so that all the invariant tori (resonant or non-resonant) preserve under the perturbation, hence achieving diffusion of arbitrary size. Subsequent results on Arnold diffusion all ignored the gap problem, resulting in much limited results and leaving much to be desired. In some cases ([7], [12], [15], [16]) one obtains diffusion in a weaker sense of size depending on perturbations, while in other cases [5], relying on the density of surviving invariant tori in certain restricted region in the space, one obtains stronger diffusion of length order one (independent of small perturbation), but very limited physical size.

We solve this gap problem for the most interesting and common cases where the invariant tori are two-dimensional for the Hamiltonian flow (one-dimensional for symplectic map). We use variational method and it is based on a generalization to a higher dimensional setting of the Aubry-Mather theory on twist maps (cf. [3]) and Mather's theory on connecting orbits of action-minimizing sets [10]. One of the most prominent feature of our method is that, instead of using transition tori alone for diffusion as in Arnold's mechanism, we also use the cantori, which are cantor sets and the remains of the broken invariant tori, for transition and diffusion. We call this mechanism of diffusion Mather's mechanism, which can be thought of as a generalization of Arnold's mechanism.

We think that the variation method and Mather's mechanism is more natural in studying Arnold diffusion and various instability problems in Hamiltonian dynamics. This is because that the hyperbolicity, often required in geometric method, is lacking in diffusion problems. However, variational method requires much weaker hyperbolicity and much weaker smoothness assumptions. Many results that are known to be difficult to obtain with geometric method can be obtained in our settings with relative ease. Recently, Mather was able to construct orbits with infinite energy for Lagrangian systems on a torus using variational method. Our results in this paper can be regarded also as a generalization of this remarkable work.

We illustrate our method and ideas in the so-called *a priori* unstable systems. We consider a near integrable Hamiltonian system of the form:

$$H_\epsilon(p, q, I, \theta, t) = F(x) + G(I) + \epsilon H^1(x, I, \theta, t) \tag{1}$$

Where we assume that $x \in M$ for some symplectic manifold M^{2n} with symplectic form ω_M, $I \in \mathbb{R}$ and $\theta \in S^1$ are a pair of action angle variables and H_ϵ is periodic in t with period one. We also assume the non-degeneracy condition

$$G'(I) > 0 \quad \text{for} \quad I \in \mathbb{R}$$

For appropriate change of coordinates, we may assume that $G(I) = I^2/2$.

For $\epsilon = 0$, the Hamiltonian $H_0(x, I, \theta, t) = F(x) + G(I)$ is time independent and it defines a decoupled flow on $M \times (\mathbb{R} \times S^1)$. The Hamiltonian flow on M is given by the vector field X_F, where X_F is defined by $\omega_M(X_F, \cdot) = dF(\cdot)$. We assume that the Hamiltonian vector field X_F has a hyperbolic periodic point at $p \in M$ and p is connected to itself by its stable and unstable manifolds, or in other words, $W^s(p) \equiv W^u(p)$.

The flow on $\mathbb{R} \times S^1$ defined by the Hamiltonian $G(I)$ is completely integrable. For each fixed constant c, the circle $I = c$ is left invariant by the flow.

Since the perturbation of the Hamiltonian system is time periodic, it is convenient to reduce the system to a symplectic map. For this purpose, we fix a cross section Σ^{t_0}, for some $t_0 \in [0, 1)$, define by

$$\Sigma^{t_0} = \left\{(x, I, \theta, t) \in M \times \mathbb{R} \times S^1 \times S^1 \mid t = t_0\right\}$$

Let P_ϵ be the Poincaré map, of the Hamiltonian H_ϵ, defined on Σ^{t_0}. P_ϵ preserves the symplectic form $\omega_M + dI \wedge d\theta$.

For simplicity of notations, we identify points on Σ^{t_0} with the points on $M \times \mathbb{R} \times S^1$. In terms of the Poincaré map P_ϵ for $\epsilon = 0$, the invariant set $A_0 = \{p\} \times \mathbb{R} \times S^1$ is normally hyperbolic. This normally hyperbolic invariant set is foliated by invariant circles of the form $T_c = \{p\} \times \{I = c\} \times S^1 \subset \Sigma^{t_0}$. Each invariant circle is connected to itself by a P_0 invariant $n+1$ dimensional manifold which serves as both the stable manifold and the unstable manifold of T_c.

We are interested in what happens when we perturb the map P_0 to P_ϵ for $\epsilon \neq 0$. Restricting to a bounded domain, say $I \in [a, b] \subset \mathbb{R}$, the normally hyperbolic invariant manifold A_0 persists under small perturbations. Let A_ϵ be the new P_ϵ invariant normally hyperbolic manifold. Then P_ϵ restricted to A_ϵ is area preserving and with the nondegeneracy assumption, KAM theory states that all the invariant curves with diophantine rotation numbers of fixed diophantine constants survive under small perturbations. Let $T_{c,\epsilon}$ be one surviving invariant torus with a diophantine rotation number c. Before the perturbation, $W^s(T_{c,\epsilon}) \equiv W^u(T_{c,\epsilon})$ for $\epsilon = 0$. For $\epsilon \neq 0$ small, $W^s(T_{c,\epsilon})$ typically intersects $W^u(T_{c,\epsilon})$ transversally at some point. This type of transversal intersections results in some very complicated dynamics.

A P_ϵ-invariant torus $T_{c,\epsilon}$ (invariant circle in this particular case) on the normally hyperbolic invariant set A_ϵ is said to be a transition torus if the stable manifold of $T_{c,\epsilon}$ intersects transversally the unstable manifold of $T_{c,\epsilon}$ at some point. A sequence of transition tori $T_{c_1,\epsilon}$, $T_{c_2,\epsilon}$, $\ldots$, $T_{c_k,\epsilon}$ is said to form a transition chain if the stable manifold of $T_{c_i,\epsilon}$ intersect transversally the unstable manifold of $T_{c_j,\epsilon}$ for all $|i - j| = 1$. We define the length of the chain to be $\max_{1 \leq i,j \leq k}\{|c_i - c_j|\}$.

The concept of transition chain was introduced by Arnold to construct unstable near-integrable Hamiltonian systems. By introducing the concept of obstructing sets, Arnold was able to show that for any two small neighborhoods of two transition tori in a transition chain, one can find an orbit that connects these two neighborhoods.

Since any given KAM torus $T_{c,\epsilon}$ is usually non-isolated, at least for small perturbations, it is easy to construct transition chains of small lengths as long as one can find just one transition torus. Hence one can easily find complicated

dynamics associated to transition chain. However, to apply the above mechanism to show instabilities in near integrable Hamiltonian systems, one needs to construct transition chain of arbitrary fixed length. A detailed estimate shows that the resonant gaps between invariant KAM tori can be as large as of order $O(\sqrt{\epsilon})$, while the transversality in the stable manifold and the unstable manifold, as best as one can hope, is only of order $O(\epsilon)$, smaller than the gap size for small ϵ. Hence one can only obtain transition chain of limited size, due to breaking down of invariant tori in resonance gap.

The main result of this paper is that we can overcome the above apparent difficulties by introducing a new approach: the variational method. The variational approach has been successfully applied to the study of the twist maps in the so-called Aubry-Mather theory. In the twist maps, one obtains a collection of action-minimizing orbits, known as Aubry-Mather sets, enjoying many interesting properties. Mather was able to further obtain the connecting orbits among these action minimizing orbits whenever there is no obvious topological obstruction.

We first construct a local variational principle near the homoclinic loop of p, using the fact that the stable manifold and unstable manifold of the hyperbolic fixed point $p \in M$ are Lagrangian submanifolds in the symplectic manifold M. All of orbits we construct are action minimizing in the local sense. To construct the connecting orbits between the action minimizing orbits, we generalize Peierls' Barrier functions to high dimensions. It turns out that the barrier function measures the splitting of the stable and unstable manifold of these normally hyperbolic invariant tori and cantori. The gradient of this Barrier function, in first order approximation, is precisely the so-called Poincaré-Melnikov function (vector). This enable us to verify our barrier conditions in specific systems using the Poincaré-Melnikov functions.

The variational method also enables us to obtain some fine structure and to unfold the underlying complicated dynamics. These structure are known to be hard to obtain with the traditional geometric method. As an example, we can prove that transition chains are transitive and stable manifold of any transition torus intersects transversally the unstable manifold of any other transition torus at some point. In fact, we can construct orbit that "shadows" any sequence of transition tori in a transition chain. Thus diffusion can be observed and understood without the help of the obstructing sets, as introduced by Arnold. Existence of a large number of different types of periodic points, symbolic dynamics, etc., can all be obtained with relative ease.

In the following sections, we will introduce our main ideas and outlines of the proofs of our main results. Complete results and proofs will appear elsewhere.

The author is grateful to Professor John Mather for his interest and patience in listening and discussing with the author on this and other works, and for making useful suggestions.

2 A variational principle

In this section, we construct a local variational principle near homoclinic loops of the hyperbolic periodic point p. The construction relies on the fact that stable

and unstable manifolds of any periodic points are lagrangian submanifolds of the ambient symplectic manifold.

We begin with the integrable case, $\epsilon = 0$. In this case, the system is decoupled, the second component of the system in (I, θ) is a integerable twist map and it has a natural variational principle $h_2 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with $h_2(x_1, x_2) = (x_1 - x_2)^2/2$.

The first component of the system in M is more complicated and needs a little more work. It may not have a global variation principle for the Poincar'e map. We can construct a local one in a small neighborhood of the homoclinic loop of p. Topologically, $W^(p)$ ($\equiv W^u(p)$) is homeomorphic to S^n with two points, both corresponding to p, identified. Lifting p to two distinct points and labeling these two points p^- and p^+, we obtain S^n topologically, with poincaré map moves the points near p^- to p^+. By a global Darboux theorem [14], a small neighborhood of S^n is symplecticly diffeomorphic to a small neighborhood of the zero section of the cotangent bundle T^*S^n. It turns that the Poincaré map in this small neighborhood can be obtained by a generating function $h_2 : S^n \times S^n \to \mathbb{R}$. This is based on the following two facts: (1). p is a hyperbolic periodic point, near $p^{\pm}$, there is a local coordinate system such that the map is given by, for example, a generating function of the type: $h_1 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with

$$h_1(\mathbf{x_1}, \mathbf{x_2}) = \sum_{i=1}^{n} \frac{1}{2}(x_1^i - x_2^i)^2 + \frac{1}{2}(x_1^i)^2$$

where we assume that all eigenvalues of p are simple and real. The cases with complex eigenvalues and multiple eigenvalues require more careful analysis; (2). Away from the fixed points $p^{\pm}$, S^n is invariant lagrangian submanifold and all orbits are non-recurrent.

We remark that the case $n = 1$ is easier and since there are two components in the stable (unstable) manifold, one need only to consider one branch.

Let

$$h = h_1 + h_2 : (S^n \times \mathbb{R}) \times (S^n \times \mathbb{R}) \to \mathbb{R}$$

then h gives a local variational principle for the map near the homoclinic loop for $\epsilon = 0$. When $\epsilon \neq 0$, h is slightly perturbed and we no longer have a decoupled system. However, the normally hyperbolic surfaces $\{p^{\pm}\} \times (\mathbb{R} \times S^1)$, when restricted to bounded domain, persists for small ϵ. Without losing generality, we may assume that these normally hyperbolic surface takes the same form: $\{p^{\pm}\} \times (\mathbb{R} \times S^1)$.

For simplicity of statements, from now on we assume that our map is extended to whole space $T^*(S^n \times S^1)$, keeping in mind that only these orbits that stay in our original domain give arises to true orbits of our system. Also we will drop the dependences in ϵ in our notations.

3 Action-Minimizing Orbits

We consider the space

$$(S^n \times \mathbb{R})^{\mathbb{Z}} = \{x \mid x : \mathbb{Z} \to (S^n \times \mathbb{R})\}$$

of bi-infinite sequences of points on $(S^n \times \mathbb{R})$ with the usual product topology. An element $x \in (S^n \times \mathbb{R})^{\mathbb{Z}}$ will also be denoted by $(x_i)_{i\in\mathbb{Z}}$ and will be called an orbit, or a configuration, in the configuration space. The orbits in the configuration space may or may not correspond to any true orbits in the phase space. Only the critical configurations have corresponding true orbits.

Given an action function $h : (S^n \times \mathbb{R}) \times (S^n \times \mathbb{R}) \to \mathbb{R}$, we extend h to finite segments $(x_j, \dots, x_k), j < k$ of an orbit x by

$$h(x_j, \dots, x_k) = \sum_{i=j}^{k-1} h(x_i, x_{i+1})$$

We say that the segment $(x_j, \dots, x_k)$ is minimal or action-minimizing with respect to h if

$$h(x_j, \dots, x_k) \le h(x_j^*, \dots, x_k^*)$$

for all $(x_j^*, \dots, x_k^*)$ with $x_j = x_j^*$ and $x_k = x_k^*$.

An orbit x in the configuration space is said to be minimal or action-minimizing, with respect to h if every finite segment of x is minimal. We denote the set of action-minimizing orbits with respect to h by M_h. It's easy to see that M_h is closed.

Now, we restrict ourselves to the normally hyperbolic surface and consider the minimal orbits in that surface. This is the situation where we have a monotone twist map. This problem has been well-studied and the results are collectively known as the Aubry-Mather theory. We recall some of the basic results.

Let $M_h^{\pm} \subset M_h$ be the set of all minimal orbits that are supported, in the phase space, on the two normally hyperbolic surfaces, say $N^{\pm}$, corresponding to $p^{\pm}$. In other words, if $x = (x_i)_{i\in\mathbb{Z}} \in M_h^{\pm}$, then $\pi_1(x_i) = \{p^{\pm}\} \in S^n$ for all $i \in \mathbb{Z}$. Where $\pi_1 : (S^n \times \mathbb{R}) \to S^n$ is the natural projection into the first component. For simplicity in the notations, we identify x_i for all $i \in \mathbb{Z}$ with its projection into the second component $\mathbb{R}$. M^+ and M^- are identical copies. The following results are well-known.

- For any $x = (x_i)_{i\in\mathbb{Z}} \in M_h^{\pm}$, let $\alpha(x) = \lim_{|i|\to\infty} x_i/i$. The limit always exists and it is called the rotation number for x. Moreover, $\alpha : M_h^{\pm} \to \mathbb{R}$ is a continuous function.

- For any $\alpha \in \mathbb{R}$ the set $M_\alpha^{\pm} = \{x \in M^{\pm} \mid \alpha(x) = \alpha\}$ is not empty.

- If C is an invariant curve for the twist map on $N^{\pm}$, with rotation number α, then the lift of all the orbits in C belongs to $M_\alpha^{\pm}$, i.e., all orbits in the invariant curve are minimal.

- If $\alpha \in \mathbb{Q}$, then $M_\alpha^{\pm}$ can be decomposed into three subsets: (1). the set of all minimal periodic points of period $\alpha = p/q$, still labeled as $M_\alpha^{\pm}$; (2). the set of all minimal orbits whose α-limit set is smaller that its ω-limit set. We label this set $M_{(p/q)^+}^{\pm}$ and (3). the set of all minimal orbits whose α-limit set is larger that its ω-limit set. We label this set $M_{(p/q)^-}^{\pm}$.

In various stability problems in Hamiltonian systems, it is very important and desirable to construct orbits that connect one region or invariant set to another set or region. It is often also important to construct orbits that visit prescribed sequences of regions in the phase space. In his remarkable works on the monotone twist maps, Mather was able to obtain various connecting orbits between minimal orbits whenever there is no obvious topological obstruction. Since Mather's work is very instrumental in our construction of Arnold diffusion, we need to recall his results first.

Consider the monotone twist map f on the cylinder N. N is either N^+ or N^-. Let Γ_1 and Γ_2 be two f-invariant homotopically non-trivial Jordan curve on N, $\alpha(\Gamma_1) < \alpha(\Gamma_2)$. Where $\alpha(\Gamma_1)$ and $\alpha(\Gamma_2)$ are rotation numbers of Γ_1 and Γ_2 respectively. Assume that there is no such invariant curve for any rotation number α, $\alpha(\Gamma_1) < \alpha < \alpha(\Gamma_2)$. The region bounded by Γ_1 and Γ_2 are called the *Birkhoff region of instability.*

Theorem 3.1 (Mather) *Suppose $\alpha(\Gamma_1) < \alpha_1, \alpha_2 < \alpha(\Gamma_2)$. Then there is an orbit of f whose α-limit set lies in M_{α_1} and whose ω-limit set lies in M_{α_2}. Furthermore, if $\alpha(\Gamma_1)$ (resp. $\alpha(\Gamma_2)$) is irrational, then this conclusion still holds with the weaker hypothesis $\alpha(\Gamma_1) \le \alpha_1, \alpha_2$ (resp. $\alpha_1, \alpha_2 \le \alpha(\Gamma_2)$).*

Moreover, for each $i \in \mathbb{Z}$ a real number $\alpha(\Gamma_1) \le \alpha_i \le \alpha(\Gamma_2)$ and a positive number ϵ_i, there exists an orbit in the phase space $(\dots, P_j, \dots)$ and an increasing bi-infinite sequence of integers $j(i)$ such that distance between $P_{j(i)}$ and M_{α_i} is smaller than ϵ_i.

The connecting orbits Mather constructed are constraint minima. The main technical difficulty is to construct the constraints so that the constraint minima do not bump up against the constraints. i.e., the constraint minima have to take place in the interior of the constraints rather than on the boundary. Therefore, certain *a priori* estimates on the boundary of the constraints are required. One of the important idea here is the introduction of the so-called *Peierl's energy barrier*. We shall discuss the energy barrier and it's generalizations in the next section.

4 Barrier functions

In this section, we define Peierl's energy barrier function. Our definition is different from that of Mather's. We choose this definition so that it works in high dimensions. When applying our definition to twist maps, ours is consistent with that of Mather's, even though it appears a little bit different. Mather also has given a generalization of the barrier function to high dimensions in terms of action-minimizing measure (cf. [8], [9], [10],[11]). Ours is different from that generalization.

Now we come back to the full system. Our construction of diffusion orbits are based on two types of action-minimizing orbits. The first type is the one we already discussed: for any given $\alpha \in \mathbb{R}$, we have the action-minimizing set $M_\alpha^\pm$. The second type of action-minimizing set is the set of connecting orbits between M_α^- and M_α^+. This is a set of action-minimizing orbit whose α-limit set is contained in M_α^- and whose ω-limit set is contained in M_α^+. We denote this set by $M_{(0^+,\alpha)}$, where 0^+ indicates the rotation number of the action-minimizing

orbit in its first component. In this notation, the rotation numbers for the sets $M_\alpha^\pm$ are both $(0,\alpha)$. We can also denote the union of the two sets $M_\alpha^\pm$ by $M_{(0,\alpha)}$. It is easy to show that the set $M_{(0^+,\alpha)}$ is non-empty.

For any rotation vector $\alpha = (\alpha_1, \alpha_2)$, where $\alpha_1 \in \{0, 0^+\}$ and $\alpha_2 \in \mathbb{R}$, we fix a minimal orbit $x^\alpha = (x_i^\alpha)_{i\in\mathbb{Z}} \in M_\alpha$. For any $a \in (S^n \times \mathbb{R})$, define

$$P_\alpha(a) = \inf_{x^*} \sum_{i\in\mathbb{Z}} \big(h(x_i^*, x_{i+1}^*) - h(x_i^\alpha, x_{i+1}^\alpha)\big)$$

where the infimum is taken among all $x^* \in (S^n \times \mathbb{R})^{\mathbb{Z}}$ such that (1). $x_0^* = a$ and (2). the α-limit set and ω-limit set of x^* are both contained in the closure of M_α.

The infinite series in the above definition may not necessarily be convergent in the usual sense. The above summation is taken in the sense of $(C,1)$ summation. Recall that an infinite series $\sum_{i=1}^\infty a_i$ is said to be $(C,1)$ summable to a real value s if

$$s = \lim_{n\to\infty} (s_1 + s_2 + \ldots + s_n)/n,$$

where s_1, s_2, ..., are partial summations $s_k = (a_1 + a_2 + \cdots + a_k)$. A convergent series is always $(C,1)$ summable with the same limit.

We remark that the infimum in the definition of $P_\alpha(a)$ can always be realized by an orbit $x^* = (x_i^*)_{i\in\mathbb{Z}}$ with $a = x_0^*$.

$P_\alpha(a)$ is called the energy barrier function. $P_\alpha(a)$ depends on the rotation number α, but it does not on the specific minimal orbit $x^\alpha \in M_\alpha$ used in the definition. $P_\alpha(a) \geq 0$ for all α and a. $P_\alpha(a) = 0$ if and only if $a = x_0$ for some $x = (x_i)_{i\in\mathbb{Z}} \in M_\alpha$.

It is easy to see that the barrier function $P_\alpha(a)$ is continuous with respect to $a \in (S^n \times \mathbb{R})$. Its regularity with respect to the rotation number is more complicated. For $\alpha = (0, \alpha_2)$ and $a = (p^\pm, a_2)$, it can be shown that the value of $P_\alpha(a)$ is the same as those defined by Mather and the barrier function $P_\alpha(a) = P_{(0,\alpha_2)}(p^\pm, a_2)$ is continuous at every point $\alpha_2 \notin \mathbb{Q}$ and continuous from one side for all $\alpha_2 = (\frac{p}{q})^\pm$ for integers p, q. The barrier function is typically discontinuous at the rational points $\alpha_2 = \frac{p}{q}$. For $\alpha = (0^+, \alpha_2)$, one can show that the function $P_\alpha(a)$ is continuous at every point $\alpha_2 \in \mathbb{R}$.

Fix a rotation vector α. The barrier function $P_\alpha(a)$ is said to have a nondegenerate local minimum at $a^* \in S^n \times \mathbb{R}$ if there exists a neighborhood U of a^*, contractible to a point, such that $P_\alpha(a^*) \leq P_\alpha(a)$ for all $a \in U$ and $P_\alpha(a^*) < P_\alpha(a)$ for all $a \in \partial U$, where ∂U is the non-empty boundary of U. The orbit that realizes $P_\alpha(a^*)$ is a local minimal and it gives arises to a true orbit in the phase space.

In order to construct long connecting orbits, we first construct orbits that connects nearby minimal orbits. For this purpose, we define the joint barrier function.

For any two rotation vectors, α and α', define

$$\begin{aligned} P_{(\alpha,\alpha')}(a) &= \inf_{x^*} \sum_{i=-\infty}^{0} \big(h(x_i^*, x_{i+1}^*) - h(x_i^\alpha, x_{i+1}^\alpha)\big) \\ &+ \inf_{x^*} \sum_{i=0}^{\infty} \Big(h(x_i^*, x_{i+1}^*) - h(x^{\alpha'_i}, x_{i+1}^{\alpha'})\Big) \end{aligned}$$

where the infimum is taken among all $x^* \in (S^n \times \mathbb{R})^{\mathbb{Z}}$ such that (1). $x_0^* = a$; (2). the α-limit set is contained in the closure of M_α and (3). the ω-limit set of x^* is contained in the closure of $M_{\alpha'}$.

Same as in the definition of the barrier function, the above summation is in the sense of $(C, 1)$. Unlike the barrier function, this joint barrier may take negative values.

Fix $\alpha = (0^+, \alpha_2)$. Let a^* be a nondegenerate local minimum for $P_\alpha(a)$ and let U be the open set such that $P_\alpha(a^*) < P_\alpha(a)$. By the continuity of $P_\alpha(a)$, for α_2' sufficiently close to α_2, $P_{\alpha'}(a)$ also has a local minimum in U, where $\alpha' = (0^+, \alpha_2')$. We can further show that the joint barrier function $P_{(\alpha,\alpha')}(a)$ has a nondegenerate local minimum in U too. This provides us with the existence of local minimum orbits that connect nearby action-minimizing sets, provided that the barrier function has a nondegenerate local minimum. We can summarize this in the following lemma:

LEMMA 4.1 *Let α be a real number. Assume that the barrier function $P_{(0^+,\alpha)}(a)$ has a nondegenerate local minimum in some contractible open set U, then there exists a positive number $\delta > 0$ such that if $|\alpha' - \alpha| \le \delta$ then there exists a local minimum orbit, through the interior of U, that connects $M_{\alpha_1}^-$ to $M_{\alpha_2}^+$.*

To obtain diffusion of arbitrary length, we need to join two or more connecting orbits of the above type. Here the idea is to put barriers very close to $M_\alpha^\pm$. However, a better setting for this construction perhaps would be to lift the phase space to infinite to one covering so that the preimage of p consists of $\cdots$, p_{-1}, p_0, p_1, $\cdots$, one then construct connecting orbits through the normal hyperbolic surfaces for each p_i. In the current setting, p has only two to one covering (p^+ and p^-). We state our main results as follows.

THEOREM 4.2 *Suppose that $P_{(0^+,\alpha)}(a)$ has a nondegenerate local minimum in some open set U_α for every $\alpha \in [A, B] \subset \mathbb{R}$, then for any $\alpha_1, \alpha_2 \in [A, B]$, there is an orbit in the phase space whose α-limit set lies in $M_{\alpha_1}^-$ and whose ω-limit set lies in $M_{\alpha_2}^+$.*

Moreover, for each $i \in \mathbb{Z}$ a real number $A \le \alpha_i \le B$ and a positive number ϵ_i, there exists an orbit in the phase space $(\dots, P_j, \dots)$ and an increasing bi-infinite sequence of integers $j(i)$ such that distance between $P_{j(i)}$ and $M_{(0,\alpha_i)}$ is smaller than ϵ_i.

We finish this section by making the following remarks:

(1). The condition that $P_{(0^+,\alpha)}(a)$ has a nondegenerate local minimum in some open set U_α for every $\alpha \in [A, B] \subset \mathbb{R}$ is an open and dense condition in any smooth or analytic topology.

(2). If $M_{(0,\alpha)}$ is an invariant torus, then for near integrable systems where the perturbation is small, $P_{(0^+,\alpha)}(a)$ measures the splitting of the stable manifold and the unstable manifold of $M_{(0,\alpha)}$. In fact, over a fixed compact neighborhood of a in S^n not containing $p^\pm$, $W^s(M_\alpha)$ and $W^u(M_\alpha)$ are horizontal lagrangian submanifolds, and thus are gradients of some potential functions. $P_{(0^+,\alpha)}(a)$ is precisely the difference of these two potential functions.

(3). The barrier function $P_{(0^+,\alpha)}(a)$ can be estimated, to its first order, by the so-called Poincaré-Melnikov integrals in this *a priori* unstable setting (cf. [13]). In *a priori* stable cases, estimating $P_{(0^+,\alpha)}(a)$ is a much more difficult problem, often requiring very delicate analysis.

5 Arnold's example

Arnold considered the following periodically forced two degree of freedom Hamiltonian system of the form $H = H_0 + \epsilon H_1$, where

$$\begin{aligned} H_0 &= \frac{1}{2}(I_1^2 + I_2^2) \\ H_1 &= (\cos\phi_1 - 1) + \mu P \\ P &= (\cos\phi_1 - 1)(\sin\phi_2 + \cos t) \end{aligned}$$

For $\mu = 0$ and $\epsilon > 0$, the system decouples into a pendulum and a rotor. One obtains a hyperbolic (weakly) periodic point from the pendulum. The normally hyperbolic invariant tori are in the surface $I_1 = \phi_1 = 0$ For $\mu \neq 0$, Arnold proved the following theorem:

THEOREM 5.1 (Arnold) *Assume $0 < A < B$. For every $\epsilon > 0$ we can find a $\mu_0 > 0$ such that for $0 < \mu < \mu_0$ the system is unstable: there exists a trajectory which connects the region $I_2 < A$ with the region $I_2 > B$.*

A notable feature of Arnold's example, which makes the system much easier to analyze, is that the perturbation term is specificly chosen so that, for $\mu > 0$, it has a factor $(1-\cos\phi_1)$ which vanishes on the normally hyperbolic surface $I_1 = \phi_1 = 0$. This implies that all invariant tori on the surface survive the perturbation. Hence one does not ncounter the difficulties associated with breaking of invariant tori and gaps in the resonant zones.

To apply our results to this setting, we may choose arbitrary perturbation function P, as long as the Melnikov potential has a non-degenerate local minima for every I_2, for $A \leq I_2 \leq B$. One easier example would be just taking $P = \sin\phi_2 + \cos t$. Indeed in this case, the Poincaré-Melnikov integrals have non-degenerate local minima (cf. [5]). Thus all the results in Arnold's theorem hold in this case too.

References

1. Arnold, V.I., *Instabilities of dynamical systems with several degrees of freedom*. Sov. Math. Dokl., 5, 581-585(1964).

2. Arnold, V.I. (ed), *Dynamical Systems III*, Encyclopaedia of Mathematical Sciences, Vol. 3, Springer-Verlag, New York/Berlin, 1998.

3. Bangert, V., *Mather sets for twist maps and geodesics on tori*, Dyn. Rep., 1, 1-56(1988).

4. Bessi, U., *An approach to Arnold diffusion through calculus of variations*, Nonlinear Analysis, 26(6), 1115-1135(1996).

5. Chierchia, L. & Gallavotti, G., *Drift and diffusion in the phase space*. Ann. de la Inst. H. Poincaré Phys. Th., bf 60(1), 1-144 (1994).

6. Douady, R., *Stabilité ou instabilité des points fixes elliptiques*, Ann. scient. Éc. norm. sup., 21(4), 1-46(1988).

7. Holmes, P., & Marsden, J., *Melnikov method and Arnold diffusion for perturbation of integrable Hamiltonian systems*, J. Math. Phys., 234, 669-675(1982).

8. Mañé, R., *Generic properties and problems of minimizing measures of Lagrangian systems*, Nonlinearity, 9, 273-310(1996).

9. Mather, J., *Action minimizing invariant measures for positive definite Lagrangian systems*, Math. Z., 207, 169-207(1991).

10. Mather, J., *Variational construction of orbits of twist diffeomorphisms*, J. Amer. Math. Soc., 4, 207-263(1991).

11. Mather, J., *Variational construction of connecting orbits*, Ann. Inst. Fourier, Grenoble, 435, 1349-1386(1993).

12. Moeckel, R., *Transition tori in five-body problem*, J. Diff. Eqns., 129, 290-314(1996).

13. Robinson, C., *Horseshoes for autonomous Hamiltonian systems using Melnikov integrals,* Ergodic Theory Dynamical Systems, 8, 395-409(1988).

14. Weinstein, A., *Lagrangian submanifolds and hamiltonian systems*, Ann. Math. 98, 377-410 (1973).

15. Xia, Z., *Arnold diffusion in elliptic restricted three-body problem*, J. Dyn. Diff. Eqns, bf 5(2), 219-240 (1993).

16. Xia, Z., *Arnold diffusion and oscillatory solutions in planar three-body problem*, J. Diff. Eqns., 110, 289-321 (1994).

Zhihong Xia
Department of Mathematic
Northwestern University
Evanston, Illinois 60208
xiamath.nwu.edu

Author Index for Volumes II, III