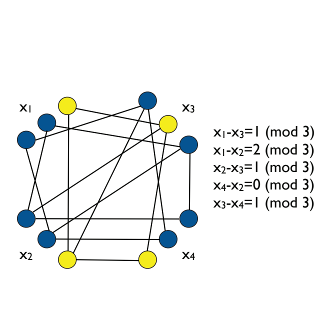
American Mathematical Society

CURRENT EVENTS BULLETIN

Saturday, January 8, 2011, 1:00 PM to 5:00 PM
Joint Mathematics Meetings, New Orleans, LA

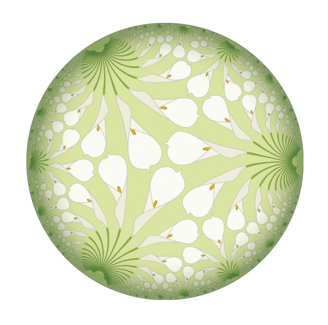
Organized by David Eisenbud, University of California, Berkeley



1:00 PM

Luca Trevisan, Khot's unique games conjecture: its consequences and the evidence for and against

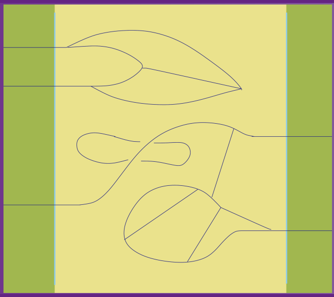
A conjecture that makes fundamental new connections between some classical mathematics and the difficulty of computing



2:00 PM

Thomas Scanlon, Counting special points: logic, Diophantine geometry and transcendence theory

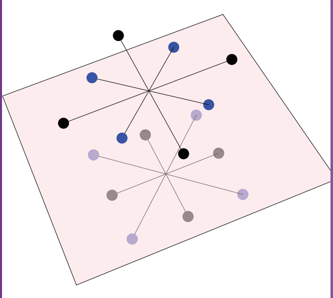
Another beautiful use of logic to prove a deep theorem about rational numbers



3:00 PM

Ulrike Tillmann, Spaces of graphs and surfaces

Topological tools lead to new ways of distinguishing families of graphs and surfaces



4:00 PM

David Nadler, The geometric nature of the Fundamental Lemma

The geometry behind the number theory behind one of this year's Fields Medals.

Introduction to the Current Events Bulletin

Will the Riemann Hypothesis be proved this week? What is the Geometric Langlands Conjecture about? How could you best exploit a stream of data flowing by too fast to capture? I love the idea of having an expert explain such things to me in a brief, accessible way. I think we mathematicians are provoked to ask such questions by our sense that underneath the vastness of mathematics is a fundamental unity allowing us to look into many different corners -- though we couldn't possibly work in all of them. And I, like most of us, love common-room gossip.

The Current Events Bulletin Session at the Joint Mathematics Meetings, begun in 2003, is an event where the speakers do not report on their own work, but survey some of the most interesting current developments in mathematics, pure and applied. The wonderful tradition of the Bourbaki Seminar is an inspiration, but we aim for more accessible treatments and a wider range of subjects. I've been the organizer of these sessions since they started, but a broadly constituted advisory committee helps select the topics and speakers. Excellence in exposition is a prime consideration.

A written exposition greatly increases the number of people who can enjoy the product of the sessions, so speakers are asked to do the hard work of producing such articles. These are made into a booklet distributed at the meeting. Speakers are then invited to submit papers based on them to the *Bulletin of the AMS*, and this has led to many fine publications.

I hope you'll enjoy the papers produced from these sessions, but there's nothing like being at the talks -- don't miss them!

David Eisenbud, Organizer
University of California, Berkeley
de@msri.org

For PDF files of talks given in prior years, see
http://www.ams.org/ams/current-events-bulletin.html.
The list of speakers/titles from prior years may be found at the end of this booklet.

ON KHOT'S UNIQUE GAMES CONJECTURE

LUCA TREVISAN

Abstract. In 2002, Subhash Khot formulated the *Unique Games Conjecture* a conjecture about the complexity of certain optimization problems.

The conjecture has inspired a remarkable body of work, which has clarified the computational complexity of several optimization problems and the effectiveness of "semidefinite programming" convex relaxations.

In this paper, which assumes no prior knowledge of computational complexity, we describe the context and statement of the conjecture, and we discuss in some detail one specific line of work motivated by the conjecture.

1. Introduction

Khot formulated the *Unique Games Conjecture* in a very influential 2002 paper [23]. In the subsequent eight years, the conjecture has motivated and enabled a large body of work on the computational complexity of approximating combinatorial optimization problems (the original context of the conjecture) and on the quality of approximation provided by "semidefinite programming" convex relaxations (a somewhat unexpected byproduct). Old and new questions in analysis, probability and geometry have played a key role in this development.

Khot has recently written an excellent and broad survey paper on this body of work [24]. In this paper we take a complementary approach and, after providing the context and the statement of the conjecture, we focus mostly on one specific line of work, which is representative of the style of research in this area. The interested reader who would like to learn more about this research area is referred to Khot's paper [24].

2. The Statement of the Unique Games Conjecture

In this section we give the context and the statement of the Unique Games Conjecture. The Unique Games Conjecture is a statement about the computational complexity of certain computational problems, so we begin with a very informal discussion of the computational complexity of "search problems," the notion of NP-completeness, and the application of NP-completeness to study the complexity of optimization problems. The reader can find a rigorous treatment in any introductory textbooks on the theory of computation, such as Sipser's [33]. We then explain the difficulties that arise in studying the complexity of approximations problems, and introduce the notion of Probabilistically Checkable Proofs and the PCP Theorem. An excellent introductory treatment of these notions can be found in the textbook of Arora and Barak [5]. Finally, we state the Unique Games Conjecture, as a variant of the PCP Theorem.

This material is based upon work supported by the National Science Foundation under Grant No. CCF-1017403.

2.1. **NP-Completeness.** To briefly introduce computational complexity theory, consider the *3-Coloring* problem. In this computational problem we are given as input an undirected graph[1] $G = (V, E)$, and the goal is to determine whether there is a *proper 3-coloring* of the vertices, that is, a function $c : V \to \{0, 1, 2\}$ such that $c(u) \neq c(v)$ for every $\{u, v\} \in E$. (If such proper colorings exist, we are also interested in finding one.)

The 3-coloring problem is easily solvable in finite time: just consider in some order all possible $3^{|V|}$ functions $c : V \to \{0, 1, 2\}$ and check each of them to see if it is a proper coloring. It is easy to improve the running time to about $2^{|V|}$, and there are non-trivial ways to achieve further speed-ups, but all the known algorithms have a worst-case running time that grows like $c^{|V|}$ for a constant $c > 1$, and they are unfeasible even on graphs with a few hundred vertices. Is there are an algorithm whose worst-case running time is bounded above by a polynomial function of $|V|$?

This is an open question equivalent to the P versus NP problem, one of the six unsolved Millenium Prize Problem. One thing that we know, however, is that the 3-Coloring problem is *NP-complete*, a notion introduced by Cook, Karp and Levin [14, 28, 22]. Informally, NP is the collection of all computational problems that, like 3-coloring, involve searching an exponentially large list in order to find an element that satisfies a given property; P is the collection of all computational problems that can be solved by algorithms whose running time is bounded from above by a polynomial function of the input length.

A computational problem A *reduces* to a computational problem B if, informally, there is a way to "encode" instances of problem A as instances of problem B, so that the existence of polynomial-time algorithms for B automatically implies the existence of a polynomial time algorithm for A.

A computational problem B is *NP-hard* if every problem $A \in NP$ reduces to B. If an NP-hard problem B is in NP, then B is said *NP-complete.* By the above discussion we see that an NP-complete problem can be solved in polynomial time if and only if $P = NP$.

It is not difficult to prove that NP-complete problems exist, and that simple and natural problems such as 3-Coloring are NP-complete. The implications, however, are quite surprising when seen for the first time. For example, searching for a bounded-length proof of a given mathematical statement is a NP problem, and so we have the following consequence of the NP-completeness of 3-Coloring:

Example 1. *There is an algorithm that given an integer n runs in time polynomial in n and constructs a graph with $O(n^2)$ vertices such that the graph has a proper 3-Coloring if and only if the Riemann Hypothesis has a proof of at most n pages.*

In fact, there is nothing special about the Riemann Hypothesis: for every mathematical statement S and every integer n it is possible to efficiently construct a graph of size polynomial in the length of S and in n such that the graph is 3-colorable if and only if S has a proof of at most n pages.

The consequences of $P = NP$, such as the ability to find mathematical proofs in time polynomial in the length of the proof, seem very implausible, and it is generally assumed that the resolution of the P versus NP question is that $P \neq NP$.

[1] An undirected graph is a pair (V, E) where V is a finite set and E is a set of unordered pairs of elements of V. The elements of V are called *vertices* and the elements of E are called *edges*. The vertices u and v are called the *endpoints* of the edge $\{u, v\}$, and the edge $\{u, v\}$ is said to be *incident* on the vertices u and v.

Assuming $P \neq NP$ as a conjecture, then the proof that a problem is NP-complete can be seen as conditional proofs that the problem cannot be solved in polynomial time.

2.2. **Optimization Problems and NP-Completeness.** The theory of NP-completeness can also be used to reason about *combinatorial optimization problems*, that is problems in which one wants to pick from an exponentially long list an item that *maximizes* or *minimizes* a given cost function. In this paper we will mainly consider the following problems:

Max Cut.: In the Maximum Cut (abbreviated Max Cut) problem we are given as input an undirected graph $G = (V, E)$ and we want to find a bipartition $(S, V - S)$ of the set of vertices maximizing the number of edges that are *cut* by the partition, that is the number of edges that have one endpoint in S and one endpoint in $V - S$. Equivalently, we want to find a 2-coloring of the vertices that maximizes the number of edges that have endpoints of different colors.

Given a graph G and a number c, it is an NP-complete problem to determine if there is a solution that cuts at least c edges.

Sparsest Cut.: In the Sparsest Cut problem we are given a d-regular graph $G = (V, E)$ (that is, a graph in which each vertex is an endpoint of precisely d edges), and we again want to find a bipartition $(S, V - S)$, but this time we want to *minimize* the number of edges cut by the partition relative to how balanced the partition is. Namely, we want to find the partition that minimizes the ratio

$$\phi(S) := \frac{|E(S, V - S)|}{d \cdot |S| \cdot |V - S|/|V|}$$

where $E(S, V - S)$ is the set of cut edges. The normalization is chosen so that the ratio in the optimal solution is always between 0 and 1.

It is an NP-complete problem to decide, given a graph $G = (V, S)$ and a number r, whether there is a set $S \subseteq V$ such that $\phi(S) \leq r$.

Such NP-completeness results rule out (assuming $P \neq NP$) the possibility of algorithms of polynomial running time computing optimal solutions for the above problems. What about the computational complexity of finding *approximate* solutions?

2.3. **Approximation Algorithms and PCPs.** The reductions that establish the above NP-completeness results do not offer much insight into the complexity of computing approximations. For example, the NP-completeness result for the Max Cut problem, relating it again to the task of finding a proof of the Riemann Hypothesis, gives the following implication:

Example 2. *There is an algorithm that, given an integer n, runs in time polynomial in n and outputs a graph $G = (V, E)$ and a number c such that:*

- *If there is a proof of the Riemann Hypothesis of at most n pages then there is a bipartition of V that cuts $\geq c$ edges;*
- *Every bipartition of G that cuts $\geq c$ edges can be efficiently converted to a valid proof of the Riemann Hypothesis of at most n pages.*

Looking more carefully into the argument, however, one sees that the transformation has the following "robustness" property with respect to approximations:

Example 3. *There is an algorithm that, given an integer n, runs in time polynomial in n and outputs a graph $G = (V, E)$ and a number c such that:*

- *If there is a proof of the Riemann Hypothesis of at most n pages then there is a bipartition of V that cuts $\geq c$ edges;*
- *Every bipartition of G that cuts $\geq c - k$ edges can be efficiently converted to a valid proof of the Riemann Hypothesis of at most n pages with at most k mistakes.*

This means that if we had, for example, an algorithm that finds in polynomial time solutions to the Max Cut problem that are at most 1% worse than the optimal, we would have that we could find an n-page "proof" such that at most 1% of the steps are wrong. Since it is always easy to come up with a proof that contains at most one mistake ("trivially, we have $0 = 1$, hence ..."), this doesn't cause any contradiction.

This doesn't mean that approximating the Max Cut problem is easy: it just means that the instance produced by the NP-completeness proof are easy to approximate, and if one wants to prove a statement of the form "if there is a polynomial time algorithm for the Max Cut problem that finds solutions at most 1% worse than the optimum, then $P = NP$," then such a result requires reductions of a rather different form from the ones employed in the classical theory of NP-completeness.

Indeed, with few exceptions, proving intractability results for approximation problems remained an open question for two decades, until the proof of the *PCP Theorem* in the early 1990s by Arora, Lund, Motwani, Safra, Sudan and Szegedy [4, 3]. The PCP Theorem (PCP stands for *Probabilistically Checkable Proofs*) can be thought of as describing a format for writing mathematical proofs such that even a "proof" in which up to, say, 1% of the steps are erroneous implies the validity of the statement that it is supposed to prove.

Theorem 1 (The PCP Theorem)**.** *There is a constant ϵ_0 and a polynomial time algorithm that on input a graph $G = (V, E)$ outputs a graph $G' = (V', E')$ such that*

- *If G has a proper 3-coloring then so does G'*
- *If there is a coloring $c' : V' \to \{1, 2, 3\}$ such that at least a $1 - \epsilon_0$ fraction of the edges of G' are properly colored by G', then G has a proper 3-coloring, and a proper 3-coloring can be efficiently constructed from c'.*

The contrapositive of the second property is that if G is not a 3-colorable graph then G' is a graph that is *not even approximately* 3-colorable, that is, G' is a graph such that, in every 3-coloring of the vertices, at least an ϵ_0 fraction of the edges have endpoints of the same color.

To see how this leads to "probabilistically checkable proofs," let us return to our running example of whether, for a given n, there is an n-page proof of the Riemann Hypothesis. For a given n, we can construct in time polynomial in n a graph G such that an n-page exists if and only if there is a proper 3-coloring of G. From G we can construct, again in time polynomial in n, a graph G' as in the PCP theorem. Now, an n-page proofs of the Riemann hypothesis can be encoded (at the cost of a polynomial blow-up in size) as a proper colorings of G'. Given a candidate proof, presented as a coloring of G', we can think of it as having $|E'|$

"steps," each being the verification that one of the edges of G' has indeed endpoints of different colors. If an n-page proof of the Riemann Hypothesis exists, then there is a proof, in this format, all whose "steps" are correct; if there is no n-page proof of the Riemann Hypothesis, however, every "proof" is now such that at least an ϵ_0 fraction of the "steps" are wrong. If we sample at random $100/\epsilon_0$ edges of G', and check the validity of the given coloring just on those edges, we will find a mistake with extremely high probability. Thus the PCP theorem gives a way to write down mathematical proofs, and a probabilistic verification procedure to check the validity of alleged proofs that *reads only a constant number of bits of the proof* and such that valid proofs pass the probabilistic test with probability 1, and if the test passes with probability higher than $(1-\epsilon_0)^{100/\epsilon_0} \approx e^{-100}$, then a valid proof exists.

While this application to proof checking is mostly interesting to visualize the meaning of the result, it might have applications to cryptographic protocols. In any case, the main application and motivation of the PCP Theorem is the study of the complexity of finding approximations to combinatorial optimization problems.

2.4. **Label Cover.** Various forms of the PCP Theorems are known, which are tailored to the study of specific optimization problems. A very versatile form of the Theorem, which was proved by Ran Raz [31] (solving a question raised by the work of Bellare et al. [10, 9]), refers to the *Label Cover* problem.

Definition 2 (Label Cover)**.** *An input to the label cover problem with range Σ is a set of equations of the form*

$$X_i = \sigma_{i,j}(Y_j)$$

where $\sigma_{i,j} : \Sigma \to \Sigma$ are functions specified as part of the input.

The goal is to find an assignment to the variables X_i and Y_j that satisfies as many equations as possible.

For example, the following is an instance of label cover with range $\mathbb{Z}/5\mathbb{Z}$:

$$\begin{aligned} X_1 &= Y_1^2 - 1 \bmod 5 \\ X_2 &= Y_1 - 1 \bmod 5 \\ X_1 &= Y_2^4 + 1 \bmod 5 \end{aligned}$$

The first and third equation are not simultaneously satisfiable, and so an optimal solution to the above instance is to satisfy two of the equations, for example the first and the second with the assignment $X_1 := 4$, $X_2 := 4$, $Y_1 := 0$, $Y_2 := 0$.

Notice that while the equations were of an algebraic nature in the example above, any function can be used to construct an equation.

Theorem 3 (Raz [31])**.** *For every $\epsilon > 0$ there is a Σ, $|\Sigma| \leq 1/\epsilon^{O(1)}$ and a polynomial time algorithm that on input a graph G outputs an instance C of label cover with range Σ such that*

- *If G has a proper 3-coloring then in C there is an assignment to the variables that satisfies all constraints;*
- *If G is not properly 3-colorable, then every assignment to the variables of C satisfies at most an ϵ fraction of the equations.*

This form of the PCP Theorem is particularly well suited as a starting point for reductions, because in the second case we have the very strong guarantee that it is impossible to satisfy even just an ϵ fraction of the equation. For technical reasons, it is also very useful that each equation involves only two variables.

The approach to derive intractability, for example for a graph problem, from Theorem 3 is to encode each variable as a small graph, and to lay out edges in such a way that the only way to have a good solution in the graph problem is to have it so that it defines a good solution for the label cover problem. If we are studying a cut problem, for example, and we have collection of vertices $v_{X,1}, \ldots, v_{X,k}$ corresponding to each variables X in the label cover instance, then a cut $(S, V - S)$ in the graph gives a k-bit string $(b_{X,1}, \ldots, b_{X,k})$ for every variable X of label cover, corresponding to which of the k vertices does or does not belong to S.

The problem then becomes:

(1) To make sure that only bit strings close to a valid codeword can occur in a near-optimal solution;
(2) To make sure that in near optimal solutions the decodings satisfy a large number of equations.

Task (2) is typically much harder than task (1), especially in reductions to graph problems. Indeed most NP-completeness results for approximating graph optimization problems have proceeded by first reducing label cover to an intermediate simpler problem, and then reducing the intermediate problem to the graph problem, but at the cost of weaker intractability results than the conjectured ones.

In 2002, Khot [23] formulated a conjecture that considerably simplifies (2), essentially making it of difficulty comparable to (1).

2.5. **The Unique Games Conjecture.**

Definition 4 (Unique Game)**.** *A* unique game *with range Σ is a set of equations of the form*

$$X_i = \sigma_{i,j}(Y_j)$$

where $\sigma_{i,j} : \Sigma \to \Sigma$ are bijective *functions specified as part of the input.*

The goal is to find an assignment to the variables X_i and Y_j that satisfies as many equations as possible.

For example, the following is a unique game with range $\mathbb{Z}/5\mathbb{Z}$:

$$\begin{aligned} X_1 &= Y_1 + 3 \bmod 5 \\ X_2 &= Y_1 + 1 \bmod 5 \\ X_1 &= Y_2 - 1 \bmod 5 \\ X_2 &= Y_2 - 1 \bmod 5 \end{aligned}$$

In the above example, it is not possible to satisfy all four equations, but the optimal solution $X_1 := 3, X_2 := 1, Y_1 := 0, Y_2 = 2$ satisfies three of the equations.

Notice that the only difference between a Label Cover instance and a Unique Game is that, in a Unique Game, the functions that define the equations have to be bijective. This is, however, a substantial difference.

In particular, given a Unique Game that has a solution that satisfies all equations, such a solution can be found very quickly in time linear in the number of equations, while in a satisfiable Label Cover instance it is an NP-hard problem to even find a solution that satisfies a small fraction of equations.

But what if we are given a Unique Game in which there is a solution that satisfies, say, a 99% fraction of the equation?

Conjecture 1 (Unique Games Intractability Conjecture)**.** *For every $1/2 > \epsilon > 0$, there is a Σ such that there is no polynomial time algorithm that, given an instance* *of Unique Games with range Σ in which it is possible to satisfy at least a $1-\epsilon$ fraction of equations, finds a solution that satisfies at least an ϵ fraction of equations.*

If $P = NP$ then Conjecture 1 is false; this means that proving Conjecture 1 would require first proving $P \neq NP$, which is beyond the reach of current techniques. The strongest evidence that we can currently hope to prove in favor of Conjecture 1 is:

Conjecture 2 (Unique Games NP-Hardness Conjecture)**.** *For every $1/2 > \epsilon > 0$ there is a Σ and a polynomial time algorithm that, on input a graph G outputs a unique games instance U with range Σ, such that*

- *If G is properly 3-colorable then there is an assignment that satisfies at least a $1-\epsilon$ fraction of equations in U;*
- *If G is not properly 3-colorable then every assignment to the variables of U satisfies at most an ϵ fraction of equations.*

If Conjecture 2 is true, then every inapproximability result proved via a reduction from Unique Games establishes an NP-hardness of approximation, in the same way as a reduction starting from label cover.

2.5.1. *Consequence for Max Cut.* In the following we let

$$\alpha_{GW} := \min_{1/2<\rho<1} \frac{\frac{1}{\pi}\cdot \arccos 1 - 2\rho}{\rho} \approx 0.878567 \tag{1}$$

And we let ρ_{GW} be the value of ρ that minimizes the above expression. The above constant comes up in the remarkable algorithm of Goemans and Williamson [20].

Theorem 5 (Goemans and Williamson [20])**.** *There is a polynomial time algorithm that on input a graph $G = (V,E)$ in which the optimal bipartition cuts opt edges finds a bipartition that cuts at least $\alpha_{GW} \cdot opt$ edges.*

Furthermore, if $opt/|E| = 1 - \epsilon \geq \rho_{GW}$, then the bipartition found by the algorithm cuts at least

$$\frac{1}{\pi}\cdot \arccos(-1+2\epsilon)\cdot |E| \tag{2}$$

edges., which is approximately $\left(1 - \frac{2}{\pi}\sqrt{\epsilon}\right)\cdot |E|$ edges

The value of Expression (2) is approximately

$$\left(1 - \frac{2}{\pi}\sqrt{\epsilon}\right)\cdot |E|$$

when ϵ is small.

It is known that an approximation better than 16/17 implies that $P = NP$ [34, 21], but no NP-hardness result is known in the range between $\alpha \approx .878$ and $16/17 \approx .941$, and there has been no progress on this problem since 1997.

Work of Khot, Kindler, Mossel and O'Donnel [25], together with later work of Mossel, O'Donnel and Oleszkiewicz [29], proves that no improvement is possible over the Goemans-Williamson algorithm assuming the Unique Games Intractability Conjecture.

Theorem 6 ([25, 29]). *Suppose that there is a $\delta > 0$, a $\rho > 0$ and a polynomial time algorithm that given a graph $G = (V, E)$ in which an optimal cut cuts $\rho \cdot |E|$ vertices finds a solution that cuts at least $\frac{1}{\pi} \cdot (\arccos(1 - 2\rho) + \delta) \cdot |E|$ edges.*

Then the Unique Games Intractability Conjecture is false.

In particular, by taking $\rho = \rho_{GW}$ we have that, for every $\delta > 0$ the existence of a polynomial time algorithm that, on input a graph in which the optimum is c finds a solution that cuts more than $(\alpha_{GW} + \delta) \cdot c$ edges would contradict the Unique Games Intractability Conjecture. So, assuming the conjecture, the constant α_{GW} is precisely the best achievable ratio between the value of polynomial-time constructible solutions and optimal solutions in the Max Cut problem.

In Section 3 we will present an overview of the proof of Theorem 6.

2.5.2. *Consequence for Sparsest Cut.* The algorithm achieving the best ratio between the quality of an optimal solution and the quality of the solution found in polynomial time is due to Arora, Rao and Vazirani [8].

Theorem 7 ([8]). *There is a polynomial time algorithm that given a graph $G = (V, E)$ finds a set C such that*

$$\phi(C) \leq O(\sqrt{\log |V|}) \cdot \phi(C^*)$$

where C^* is an optimal solution to the sparsest cut problem.

A classical algorithm based on spectral graph theory achieves a better approximation in graphs in which the optimum is large.

Theorem 8 (Spectral Partitioning [1, 2]). *There is a polynomial time algorithm that given a graph $G = (V, E)$ finds a set C such that*

$$\phi(C) \leq O(\sqrt{\phi(C^*)})$$

where C^* is an optimal solution to the sparsest cut problem.

Theorem 9 ([25, 29]). *There is an absolute constant $c > 0$ such that the following is true.*

Suppose that there is a $\delta > 0$, an $\epsilon > 0$ and a polynomial time algorithm that given a graph $G = (V, E)$ in which the sparsest cut C^* satisfies $\phi(C^*) \leq \epsilon$ finds a cut C such that

$$\phi(C) \leq c \cdot \sqrt{\epsilon} - \delta \ ;$$

then the Unique Games Intractability Conjecture is false.

In particular, assuming the conjecture, the trade-off between optimum and approximation in the spectral partitioning algorithm cannot be improved, and the approximation ratio in the Arora-Rao-Vazirani algorithm cannot be improved to a constant.

3. The Maximum Cut Problem

A general approach to reduce Unique Games (and, in general, Label Cover) with range Σ to other problems is to ensure that a solution in the target problem associates to each variable X of the unique game a function $f_X : \{-1, 1\}^\Sigma \to \{-1, 1\}$. Then we define a way to "decode" a function $f_X : \{-1, 1\}^\Sigma \to \{-1, 1\}$ to a value $a_X \in \Sigma$, and we aim to prove that if we have a good solution in the target problem, then the assignment $X := a_X$ to each variable X defines a good solution in the Unique Games instance. The general idea is that if a function "essentially depends" one of its variables, then we decode it to the index of the variable that it depends on.

3.1. **The Reduction from Unique Games to Max Cut.** We outline this method by showing how it works to prove Theorem 6. To prove the theorem, we start from a Unique Games Instance U with range Σ such that a $1 - \epsilon'$ fraction of equations can be satisfied. We show how to use the assumption of the Theorem to find a solution that satisfies at least an ϵ' fraction of equations. We do so by constructing a graph G, applying the algorithm to find a good approximation to Max Cut in the graph, and then converting the cut into a good solution for the Unique Games instance.

If U has N variables, then G has $N \cdot 2^{\Sigma}$ vertices, a vertex $v_{X,a}$ for every variable X of U and every value $a \in \{0,1\}^{\Sigma}$.

We define G as a *weighted* graph, that is a graph in which edges have a positive real-value weight. In such a case, the value of a cut is the total weight (rather than the number) of edges that are cut. There is a known reduction from Max Cut in weighted graph to Max Cut in unweighted simple graphs [15], so there is no loss of generality in working with weights.

We introduce the following action of the symmetric group of Σ on the vertices of G. If $x \in \{-1,1\}^{\Sigma}$ is an vector of $|\Sigma|$ bits indexed by the elements of Σ, and $\sigma : \Sigma \to \Sigma$ is a bijection, we denote by $x \circ \sigma$ the vector $x \circ \sigma \in \{0,1\}^{\Sigma}$ such that $(x \circ \sigma)_i := x_{\sigma(i)}$.

We also define the *noise operator* N_ρ as follows: if $x \in \{0,1\}^{\Sigma}$ is a boolean vector, then $N_\rho(x)$ is the random variable generated by changing each coordinate of x independently with probability ρ, and leaving it unchanged with probability $1 - \rho$.

The edge set of G is defined so that its total weight is 1, and we describe it as a probability distribution:

- Pick two random equations $X = \sigma(Y)$ and $X = \sigma'(Y')$ in U conditioned on having the same left-hand side.
- Pick a random element $a \in \{0,1\}^{\Sigma}$ and pick an element $b \in N_\rho(a)$
- Generate the edge $(v_{Y,a\circ\sigma}, v_{Y',b\circ\sigma'})$

Let A be an optimal assignment for the Unique Games instance U. Consider the cut of G in which $S = \{v_{Y,a} : a_{A(Y)} = 1\}$. This vertex bipartition cuts edges of total weight at least $\rho - 2\epsilon'$. From our assumption, we can find in polynomial time a cut S that cuts a $\frac{1}{\pi} \cdot (\arccos 1 - 2\rho) + \delta$ fraction of edges. We want to show how to extract from S an assignment for the Unique Games that satisfies a reasonably large number of equations.

First we not that S assigns a bit to each variable X and to each $a \in \{-1,1\}^{\Sigma}$. Let us call

$$f_Y(a) = 1 \text{ if } a \in S$$

and

$$f_Y(a) = -1 \text{ if } a \notin S$$

We went to decode each of these functions $f_X : \{-1,1\}^{\Sigma} \to \{-1,1\}$ into an index $i \in \Sigma$. We describe a probabilistic decoding process $Dec(\cdot)$ later.

Some calculations show that the functions that we derive in such a way have the property that

$$\mathop{\mathbb{E}}_{X,Y,Y',a,b}[f_Y(a\circ\sigma)\neq f_{Y'}(b\circ\sigma')]\geq\frac{1}{\pi}\cdot(\arccos 1-2\rho)+\delta$$

and from this we want to derive that

$$\mathop{\mathbb{E}}_{X,Y,Y'}[\sigma(Dec(f_Y))=\sigma'(Dec(f_{Y'}))]\geq\Omega_{\rho,\delta}(1)$$

from which it is easy to see that from the decodings $Dec(f_Y)$ we can reconstruct an assignment for all variables that satisfies at least an ϵ' fraction of equations in the unique game.

Some manipulations show that, essentially, it is sufficient to prove the following lemma:

Lemma 10 (Main)**.** *There is a probabilistic symmetric algorithm $Dec(\cdot)$ that on input a function $f:\{-1,1\}^\Sigma\to\{-1,1\}$ outputs an element $i\in\Sigma$, and such that the following is true.*

Suppose that $f:\{-1,1\}^\Sigma\to\{-1,1\}$ is such that

$$\mathbb{P}[f(x)\neq f(N_\rho x)]\geq\frac{1}{\pi}\cdot\arccos(1-2\rho)+\delta \tag{3}$$

Then there is an index $i\in\Sigma$ such that

$$\mathbb{P}[Dec(f)=i]\geq\Omega_{\delta,\rho}(1)$$

We say that the decoding is *symmetric* if the distribution of $Dec(f(\sigma(\cdot)))$ is the same as the distribution $\sigma(Dec(f(\cdot)))$ for every bijection $\sigma:\Sigma\to\Sigma$.

(Technically, the Main Lemma is not sufficient as stated. An extension that deals with all bounded real-valued functions is necessary. The boolean case, which is simpler to state and visualize, captures all the technical difficulties of the general case.)

3.2. The Proof of the Main Lemma.

Before discussing the proof of the Main Lemma, we show that it is tight, in the sense that from a weaker assumption in Equation (3) it is not possible to recover the conclusion.

Consider the majority function $Maj:\{-1,1\}^\Sigma\to\{-1,1\}$ such that $Maj(x)=1$ if and only if x has at least $|\Sigma|/2$ ones. (That is, $Maj(x):=sign(\sum_i x_i)$.) Then Maj is a *symmetric* function, in the sense that $Maj(x\circ\sigma)=Maj(x)$ for every bijection σ. This implies that for every symmetric decoding algorithm Dec we have that $Dec(Maj)$ is the uniform distribution over Σ, and so every index i has probability $1/|\Sigma|$ which goes to zero even when the other parameters in the Main Lemma are fixed. A standard calculation shows that, for large Σ,

$$\mathbb{E}[Maj(x)\neq Maj(N_\rho(x))]\approx\frac{1}{\pi}\arccos(1-2\rho)$$

so we have an example in which Equation (3) is nearly satisfied but the conclusion of the Main Lemma fails.

This example suggest that, if the Main Lemma is true, then the functions that satisfy Equation (3) must be non-symmetric, that is, it must not depend equally on all the input variables, and that the decoding procedure $Dec(\cdot)$ must pick up certain input variables that the function depends in a special way on.

Another example to consider is that of the functions arising in the bipartitions that are derived from an optimal solution in the unique game instance U. In that case, for every variable Y the corresponding function f_Y is of the form $f_Y(x) := x_i$ where i is the value assigned to Y in the optimal solution. In this case, we would expect the decoding algorithm to output the index i. In general, if f depends only on a small number of variables, we would expect Dec to only output the indices of those variables.

These observations suggest the use of the notion of *influence* of input variables. If $f : \{-1, 1\}^\Sigma \to \{-1, 1\}$ is a boolean function, then the influence of variable $i \in \Sigma$ on f is the probability

$$Inf_i(f) := \mathop{\mathbb{P}}_{x \in \{0,1\}^\Sigma}[f(x_1, \ldots, x_k) \neq f(x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_k)]$$

where we identified, for simplicity, Σ with $\{1, \ldots, k\}$, where $k := |\Sigma|$.

It is natural to consider the decoding algorithm that picks an index i with probability proportional to $Inf_i(f)$; note that this process is symmetric.

There is, unfortunately, a counterexample. Consider the function

$$f(x_1, \ldots, x_k) := Maj(x_1, x_2, Maj(x_3, \ldots, x_k))$$

and take $\rho = 1 - \epsilon$. Then $\frac{1}{\pi} \cdot \arccos(1 - 2\rho) \approx 1 - \frac{2}{\pi}\sqrt{\epsilon}$ and one can compute that

$$\mathbb{P}[f(x) \neq f(N_\rho(x))] \approx 1 - \epsilon - \frac{1}{\pi}\sqrt{\epsilon} > 1 - \frac{2}{\pi}\sqrt{\epsilon} + \Omega_\epsilon(1)$$

This means that we expect the decoding algorithm to select some index with a probability that is at least a fixed constant for every fixed ϵ.

When we compute the influence of the variables of f, however, we find out that x_1 and x_2 have constant influence 1/2, while the variables $x_3, \ldots, x_k$ have influence approximately $1/\sqrt{k}$. This means that the sum of the influences is about $\sqrt{k}$, and so x_1 and x_2 would be selected with probability about $1/\sqrt{k}$, and the remaining variables with probability about $1/n$. In particular, all probabilities go to zero with $k = |\Sigma|$, and so a decoding algorithm based only on influence does not satisfy the conditions of the Main Lemma.

In order to introduce the correct definition, it helps to introduce discrete Fourier analysis over the Hamming cube. For our purposes, only the following facts will be used. One is that if $g : \{-1, 1\}^\Sigma \to \mathbb{R}$ is a real-valued function, then there is a unique set of real values $\hat{g}(S)$, one for each subset $S \subseteq \Sigma$, such that

$$g(x) = \sum_S \hat{g}(S) \cdot \prod_{i \in S} x_i$$

The values $\hat{g}(S)$ are called the Fourier coefficients of g.

In particular, if $f : \{-1, 1\}^\Sigma \to \{-1, 1\}$ is a boolean function, then $\sum_S \hat{f}^2(S) = 1$.

It is easy to see that

$$Inf_i(f) = \sum_{S:i\in S} \hat{f}^2(S)$$

The fact that $\sum_i \hat{f}^2(S) = 1$ suggests that $\hat{f}$ naturally defines a probability distribution. Unfortunately, it is a probability distribution over subsets of Σ, rather than a probability distribution over *elements* of Σ. A natural step is to consider the algorithm *Dec* defined as follows: sample a set $S \subseteq \Sigma$ with probability equal to $\hat{f}^2(S)$, then output a random element of S. In particular, we have

$$\mathbb{P}[Dec(f) = i] = \sum_{S:\in S} \frac{\hat{f}^2(S)}{|S|} \tag{4}$$

which is similar to the expression for the influence of i, but weighted to give more emphasis of the Fourier coefficients corresponding to smaller sets.[2]

If we go back to the function $Maj(x_1, x_2, Maj(x_3, \ldots, x_k))$, we see that the algorithm defined in (4) has a probability of generating x_1 and x_2 which is at least an absolute constant, and that doesn't go to zero with k.

The decoding algorithm described in Equation (4) turns out to be the correct one. Proving the main lemma reduces now to proving the following result.

Lemma 11 (Main Lemma – Restated)**.** *Suppose that $f : \{0,1\}^\Sigma \to \{0,1\}$ is such that*

$$\mathbb{P}[f(x) \neq f(N_\rho x)] \geq \frac{1}{\pi} \cdot \arccos(1 - 2\rho) + \delta \tag{5}$$

Then there is an index $i \in \Sigma$ such that

$$\sum_{S:i\in S} \frac{\hat{f}^2(S)}{|S|} \geq \Omega_{\delta,\rho}(1)$$

The proof has two parts:

- An *invariance theorem* due to Mossel, O'Donnell and Oleszkiewicz [29] showing that the Main Lemma is true in the boolean setting provided that a "Gaussian version" of the Lemma hods for functions taking real inputs with Gaussian distribution is true;
- A theorem of Borell [11] establishing the Gaussian version of the Lemma

3.3. The Invariance Theorem and Borell's Theorem.

A starting point to gain intuition about the Invariance Theorem is to consider the Central Limit Theorem. Suppose that $X_1, \ldots, X_n$ is a collection of independent boolean random variables, each uniform over $\{-1, 1\}$, and suppose that $a_1, \ldots, a_n$ are arbitrary coefficients. Then the random variable

$$\sum_i a_i X_i$$

[2] An important point is that, with probability $\hat{f}(\emptyset)^2$ we generate the empty set, and the operation of "selecting a random element" of the empty set is undefined. In such a case, the decoding algorithm outputs a special failure symbol $\perp$ not in Σ.

is going to be close to a Gaussian distribution of average zero and variance $\sum_i a_i^2$, provided that the coefficients are reasonably smooth. (It is enough that if we scale them so that $\sum_i a_i^2 = 1$, then $\sum_i a_i^3$ is small.)

Suppose now that, instead of considering a sum, that is, a degree-1 function, we take an n-variate low-degree polynomial p and we consider the random variable

$$p(X_1, \ldots, X_n)$$

We cannot say any more that it has a distribution close to a Gaussian and, in fact, it does not seem that we can say anything at all. Looking back at the Central Limit Theorem, however, we can note that the "right" way of formulating it is to consider a collection $X_1, \ldots, X_n$ of independent boolean random variables each uniform over $\{-1, 1\}$, and also a collection of independent Gaussian random variables $Z_1, \ldots, Z_n$ each with mean zero and variance 1. Then we have that the two random variables

$$\sum_i a_i X_i \text{ and } \sum_i a_i Z_i$$

are close provided that the a_i are smooth.

This is exactly the same statement as before, because the distribution $\sum_i a_i Z_i$ happens to be a Gaussian distribution of mean zero and variance $\sum_i a_i^2$.

This formulation, however, as a natural analog to the case of low-degree polynomials. The Invariance Theorem states that if p is a sufficiently "smooth" low degree polynomial then the random variables

$$p(X_1, \ldots, X_n) \text{ and } p(Z_1, \ldots, Z_n)$$

are close. A result of this nature was first proved by Rotar [32].

When we apply the Invariance Theorem to a smoothed and truncated version of the Fourier transform of the function f in the Main Lemma, we have that either such a function is a "smooth polynomial" to which the Invariance Theorem applies, or else the conclusion holds and there is a coordinate with noticeably high probability of being output by the decoding algorithm. If the Invariance Theorem applies, then the probability that f changes value on anti-correlated boolean inputs is approximately the probability that a function changes its value on anti-correlated Gaussian inputs. The latter is given by a Theorem of Borrel

Theorem 12 (Borrel)**.** *Suppose $f : \mathbb{R}^n \to [-1, 1]$ is a measurable function according to the standard Gaussian measure in $\mathbb{R}^n$ and such that $\mathbb{E} f = 0$. For an element $x \in \mathbb{R}^n$ and for $0 \leq \rho \leq 1/2$, let $N_\rho(x)$ be the random variable $(1 - 2\rho) \cdot x + \sqrt{1 - (1 - 2\rho)^2} z$ where z is a standard Gaussian random variable.*

Then

$$\mathbb{P}[f(x) \neq f(N_\rho(x))] \geq \frac{1}{\pi} \arccos(1 - 2\rho)$$

There are a few ways in which Borrel's theorem is not the "Gaussian analog" of the Main Lemma. Notably, there is a condition on the expectation of f , there is a lower bound, rather than an upper bound, to the probability that f changes value, and the theorem applies to the range $\rho \in [0, 1/2]$, while we are interested in the "anti-correlation" case of $\rho \in [1/2, 1]$. There is a simple trick (consider only the "odd part" of the Fourier expansion of the boolean function f – that is only the terms corresponding to sets S of odd size) that takes care of all these differences.

3.4. **How Did We Use the Unique Games Conjecture?** When we stated the Unique Games Conjecture, we made the following informal claim, here rephrased in abbreviated form:

> *To reduce Label Cover to a graph optimization problem like Max Cut, we map variables to collections of vertices and we map equations to collections of edges; then we show how to "encode" assignments to variables as 2-colorings of vertices which cut a $\geq c_1$ fraction of edges, and finally (this is the hardest part of the argument) we show that given a 2-coloring that cuts a $\geq c_2$ fraction of edges, then*
>
> *(1) the given 2-coloring must be somewhat "close" to a 2-coloring coming from the encoding of an assignment and*
>
> *(2) if we "decode" the given 2-coloring to an assignment to the variables, such an assignment satisfies a noticeable fraction of equations.*
>
> *Starting our reduction from a Unique Game instead of a Label Cover problem, we only need to prove (1) above, and (2) more or less follows for free.*

To verify this claim, we "axiomatize" the properties of a reduction that only achieves (1): we describe a reduction mapping a single variable to a graph, such that assignments to the variable are mapped to good cuts, and somewhat good cuts can be mapped back to assignments to the variable. The reader can then go back to our analysis of the Max Cut inapproximability proof in the previous post, and see that almost all the work went into establishing the existence of a family of graphs satisfying the properties below.

Definition 13 ((c_1, c_2)-Graph Family). *A (c_1, c_2) graph family is a collection of graphs $G_m = (V_m, E_m)$, for each positive integer m, together with an encoding function $Enc_m : \{1, \ldots, m\} \to 2^{V_m}$ and a randomized decoding process $Dec_m : 2^{V_m} \to \{1, \ldots, m\}$ such that*

- *For every m and every $i \in m$, let $S_i := Enc_m(i)$. Then the partition $(S_i, V_m - S_i)$ cuts at least a c_1 fraction of the edges of G_m;*
- *If $(S, V_m - S)$ is a partition of the vertices of G_m that cuts at least a $c_2 + \delta$ fraction of the edges, then there is an index $i \in \{1, \ldots, m\}$ such that the probability*

$$\mathbb{P}[Dec_m(S) = i] \geq p(\delta) > 0$$

  is at least a positive quantity $p(\delta)$ independent of m;
- *The encoding and decoding procedures are* symmetric. *That is, it is possible to define an action of the symmetric group of $\{1, \ldots, m\}$ on V_m such that for every $i \in m$ and every bijection $\sigma : \{1, \ldots, m\} \to \{1, \ldots, m\}$ we have*

$$Enc_m(\sigma(i)) = \sigma(Enc_m(i))$$

  and

$$Dec_m(\sigma(S)) \approx \sigma(Dec_m(S))$$

where $D_1 \approx D_2$ means that D_1 and D_2 have the same distribution, and $\sigma(S) := \{x \circ \sigma : x \in S\}$, where $x \circ \sigma$ is the action of σ on x.

We claim that, in the previous post, we defined a $(1-\epsilon, 1-\frac{2}{\pi}\sqrt{\epsilon})$ graph family. The graph family is the following. For a given m:

(1) The vertex set is $V_m := \{-1,1\}^m$;
(2) The graph is weighted complete graph with edges of total weight 1. The weight of edge (x,y) is the probability of generating the pair (x,y) by sampling x at random and sampling y from the distribution $N_{1-\epsilon}(x)$;
(3) $Enc_m(i)$ defines the bipartition $(S_i, V_m - S_i)$ in which S_i is the set of all vertices x such that $x_i = 1$
(4) $Dec_m(S)$ proceeds as follows. Define $f(x) := -1$ if $x \in S$ and $f(x) := 1$ if $x \notin S$. Compute the Fourier expansion

$$f(x) = \sum_R \hat{f}(R)(-1)^{\sum_i \in R x_i}$$

Sample a set R with probability proportional to $\hat{f}^2(R)$, and then output a random element of R

4. Semidefinite Programming and Unique Games

Solving an instance of a combinatorial optimization problem of minimization type is a task of the form

$$\begin{array}{l} \max cost(z) \\ \text{subject to} \\ z \in Sol \end{array} \tag{6}$$

where Sol is the set of admissible solutions and $cost(z)$ is the cost of solution z. For example the problem of finding the maximum cut in a graph $G = (V,E)$ is a problem of the above type where Sol is the collection of all subsets $S \subseteq V$, and $cost(S)$ is the number of edges cut by the vertex partition $(S, V-S)$.

If $Sol \subseteq Rel$, and $cost' : Rel \to \mathbb{R}$ is a function that agrees with $cost()$ on Sol, then we call the problem

$$\begin{array}{l} \max cost'(z) \\ \text{subject to} \\ z \in Rel \end{array} \tag{7}$$

a *relaxation* of the problem in (6). The interest in this notion is that combinatorial optimization problems in which the solution space is discrete are often NP-hard, while there are general classes of optimization problems defined over a continuous *convex* solution space that can be solved in polynomial time. A fruitful approach to approximating combinatorial optimization problems is thus to consider relaxations to tractable convex optimization problems, and then argue that the optimum of the relaxation is close to the optimum of the original discrete problem. See the book of Vazirani [35] for several appications of this approach.

The Unique Games Intractability Conjecture is deeply related to the approximation quality of *Semidefinite Programming* relaxations of combinatorial optimization problems.

4.1. **Semidefinite Programming.** A symmetric matrix A is positive semidefinite, written $A \succeq \mathbf{0}$, if all its eigenvalues are non-negative. We write $A \succeq B$ if $A - B$ if positive semidefinite. We quote without proof the following facts:

- A matrix $A \in \mathbb{R}^{n\times n}$ is positive semidefinite if and only if there are vectors $v^1, \ldots, v^n \in \mathbb{R}^m$ such that for every i, j we have $A_{ij} = \langle v^i, v^j\rangle$. Furthermore, there is an algorithm of running time polynomial in n that, given a matrix A, tests whether A is positive semidefinite and, if so, finds vectors $v^1, \ldots, v^n$ as above.
- The set of positive semidefinite matrices is a convex subset of $\mathbb{R}^{n\times n}$. More generally, it is a *convex cone*, that is, for every two positive semidefinite matrices A, B and non-negative scalars α, β, the matrix $\alpha A+\beta B$ is positive semidefinite.

It often the case the optimizing a linear function over a convex subset of $\mathbb{R}^N$ is a polynomial time solvable problem, and indeed there are polynomial time algorithms for the following problem:

Definition 14 (Semidefinite Programming)**.** *The Semidefinite Programming problem is the following computational program: given matrices $C, A^1, \ldots, A^m \in \mathbb{R}^{n\times n}$ and scalars $b_1, \ldots, b_m \in \mathbb{R}$, find a matrix X that solves the following optimization problem (called a semidefinite program):*

$$\begin{array}{l} \max C \bullet X \\ \textit{subject to} \\ A^1 \bullet X \leq b_1 \\ A^2 \bullet X \leq b_2 \\ \cdots \\ A^m \bullet X \leq b_m \\ X \succeq \mathbf{0} \end{array} \tag{8}$$

where we use the notation $A \bullet B := \sum_{ij} A_{ij} \cdot B_{ij}$.

In light of the characterization of positive semidefinite matrices described above, the semidefinite program (8) can be equivalently written as

$$\begin{array}{l} \max \sum_{ij} C_{ij} \cdot \langle v^i, v^j\rangle \\ \text{subject to} \\ \sum_{ij} A^1_{ij} \cdot \langle v^i, v^j\rangle \leq b_1 \\ \sum_{ij} A^2_{ij} \cdot \langle v^i, v^j\rangle \leq b_2 \\ \cdots \\ \sum_{ij} A^m_{ij} \cdot \langle v^i, v^j\rangle \leq b_m \\ v^1, \ldots, v^n \in \mathbb{R}^n \end{array} \tag{9}$$

That is, as an optimization problem in which we are looking for a collection $v^1, \ldots, v^n$ of vectors that optimize a linear function of their inner products subject to linear inequalities about their inner products.

4.2. **Semidefinite Programming and Approximation Algorithms.** A *quadratic program* is an optimization problem in which we are looking for reals $x_1, \ldots, x_n$ that optimize a quadratic form subject to quadratic inequalities, that is an optimization problem that can be written as

$$\begin{array}{l} \max \sum_{ij} C_{ij} \cdot x_i \cdot x_j \\ \text{subject to} \\ \sum_{ij} A^1_{ij} \cdot x_i \cdot x_j \leq b_1 \\ \sum_{ij} A^2_{ij} \cdot x_i \cdot x_j \leq b_2 \\ \cdots \\ \sum_{ij} A^m_{ij} x_i \cdot x_j \leq b_m \\ x_1, \ldots, x_n \in \mathbb{R} \end{array} \tag{10}$$

Since the quadratic condition $x \cdot x = 1$ can only be satisfied if $x \in \{-1, 1\}$, quadratic programs can express discrete optimization problems. For example, the Max Cut problem in a graph $G = (V, E)$, where $V = \{1, \ldots, n\}$ can be written as a quadratic program in the following way

$$\begin{array}{l} \max \sum_{ij \in E} \frac{1}{2} - \frac{1}{2} x_i \cdot x_j \\ \text{subject to} \\ x_1^2 = 1 \\ \cdots \\ x_n^2 = 1 \\ x_1, \ldots, x_n \in \mathbb{R} \end{array} \tag{11}$$

Every quadratic program has a natural Semidefinite Programming relaxation in which we replace reals x_i with vectors v^i and we replace products $x_i \cdot x_j$ with inner products $\langle v^i, v^j \rangle$. Applying this generic transformation to the quadratic programming formulation of Max Cut we obtain the following semidefinite programming formulation of Max Cut

$$\begin{array}{l} \max \sum_{ij \in E} \frac{1}{2} - \frac{1}{2} \langle v^i, v^j \rangle \\ \text{subject to} \\ \langle v^1, v^1 \rangle = 1 \\ \cdots \\ \langle v^n, v^n \rangle = 1 \\ v^1, \ldots, v^n \in \mathbb{R}^n \end{array} \tag{12}$$

The Max Cut relaxation (12), first studied by Delorme and Poljak [17, 16] is the one used by Goemans and Williamson.

Algorithms based on semidefinite programming provide the best known polynomial-time approximation guarantees for a number of other graph optimization problems and of constraint satisfaction problem.

4.3. **Semidefinite Programming and Unique Games.** The quality of the approximation of Relaxation (12) for the Max Cut problem exactly matches the intractability results proved assuming the Unique Games Intractability Assumptions. This has been true for a number of other optimization problems.

Remarkably, Prasad Raghavendra has shown [30] that for a class of problems (which includes Max Cut as well as boolean and non-boolean constraint satisfaction problems), there is a semidefinite programming relaxation such that, assuming the Unique Games Intractabiltiy Conjecture, no other polynomial time algorithm can provide a better approximation than that relaxation.

If one believes the conjecture, this means that the approximability of all such problems has been resolved, and a best-possible polynomial time approximation algorithm has been identified for each such problem. An alternative view is that, in order to contradict the Unique Games Intractabiltiy Conjecture, it is enough to find a new algorithmic approximation techniques that works better than semidefinite programming for any of the problems that fall into Raghavendra's framework, or maybe find a different semidefinite programming relaxation that works better than the one considered in Raghavendra's work.

4.4. **Sparsest Cut, Semidefinite Programming, and Metric Embeddings.**

If, at some point in the future, the Unique Games Intractability Conjecture will be refuted, then some of the theorems that we have discussed will become vacuous. There are, however, a number of unconditional results that have been discovered because of the research program that originated from the conjecture, and that would survive a refutation.

First of all, the analytic techniques developed to study reductions from Unique Games could become part of future reductions from Label Cover or from other variants of the PCP Theorem. As discussed above, reductions from Unique Games give ways of encoding values of variables of a Label Cover instance as good feasible solutions in the target optimization problems, and ways of decoding good feasible solutions in the target optimization problems as values for the variables of the Label Cover instance.

It is also worth noting that some of the analytic techniques developed within the research program of Unique Games have broader applicability. For example the impetus to prove the Invariance Theorem of Mossel, O'Donnell and Oleszkiewicz came from its implications for conditional inapproximability results, but it settles a number of open questions in social choice theory.

Perhaps the most remarkable unconditional theorems motivated by Unique Games regard *integrality gaps* of Semidefinite Programming relaxations. The integrality gap of a relaxation of a combinatorial optimization problem is the worst-case (over all instances) ratio between the optimum of the combinatorial problem and the optimum of the relaxation. The integrality gap defines how good is the optimum of the relaxation as a numerical approximation of the true optimum, and it is usually a bottleneck to the quality of approximation algorithms that are based on the relaxation.

The integrality gap of Relaxation (12) is $.8785\cdots$, the same as the hardness of approximation result proved assuming the Unique Games Intractabiltiy Conjecture. Indeed, the graph that exhibits the $.8785\cdots$ gap is related to the graph used in the reduction from Unique Games to Max Cut. (The integrality gap instance was discovered by Feige and Schechtman [19].)This is part of the larger pattern discovered by Raghavendra (cited above), who shows that, for a certain class of optimization problems, every integrality gap instance for certain semidefinite programming relaxations can be turned into a conditional inapproximability result assuming the Unique Games Intractability Conjecture.

The Sparsest Cut problem has a Semidefinite Programming relaxation, first studied by Goemans and Linial, whose analysis is of interest even outside of the area of approximation algorithms. A metric space (X, d) is of *negative type* if $(X, \sqrt{d})$ is also a metric space and is isometrically embeddable in Euclidean space. If every n-point metric space of negative type can be embedded into $L1$ with distortion at most $c(n)$, then the Semidefinite Programming relaxation of Goemans and Linial can be used to provide a $c(n)$-approximate algorithm for sparsest cut, where n is the number of vertices, and the integrality gap of the relaxation is at most $c(n)$. Equivalently, if there is an n-vertex instance of Sparsest Cut exhibiting an integrality gap at least $c(n)$, then there is an n-point negative-type metric space that cannot be embedded into $L1$ without incurring distortion at least $c(n)$.

Interestingly, there is a generalization of the Sparsest Cut problem, the Non-uniform Sparsest Cut problem, for which the converse is also true, that is, the integrality gap of the Goemans-Linial Semidefinite Programming relaxation of the Non-uniform Sparsest Cut problem for graphs with n vertices is $\leq c(n)$ if and only if every n-point negative-type metric space can be embedded into $L1$ with distortion at most $c(n)$.

It had been conjectured by Goemans and Linial that the integrality gap of the semidefinite relaxations of Sparsest Cut and Non-Uniform Sparsest Cut was at most a constant. Arora, Rao and Vazirani [8] proved in 2004 that the Sparsest Cut relaxation had integrality gap $O(\sqrt{\log n})$, and Arora, Lee and Naor [7] proved in 2005 that Non-Uniform Sparsest Cut relaxation had integrality gap $O(\sqrt{\log n} \cdot \log\log n)$, results that were considered partial progress toward the Goemans-Linial conjecture.

Later in 2005, however, Khot and Vishnoi [26] proved that the relaxation of Non-Uniform Sparsest Cut has an integrality gap $(\log\log n)^{\Omega(1)}$ that goes to infinity with n. Their approach was to:

(1) Prove that the Non-Uniform Sparsest Cut problem does not have a constant-factor approximation, assuming the Unique Games Intractability Conjecture, via a reduction from unique games to non-uniform sparsest cut;
(2) Prove that a natural Semidefinite Programming relaxation of Unique Games has integrality gap $(\log\log n)^{\Omega(1)}$;
(3) Show that applying the reduction in (1) to the Unique Games instance in (2) produces an integrality gap instance for the Goemans-Linial Semidefinite Programming relaxation of Non-Uniform Sparsest Cut.

In particular, Khot and Vishnoi exhibit an n-point negative-type metric space that requires distortion $(\log\log n)^{\Omega(1)}$ to be embedded into $L1$. This has been a rather unique approach to the construction of counterexamples in metric geometry. The lower bound was improved to $\Omega(\log\log n)$ by Krauthgamer and Rabani [27], and shortly afterward Devanur, Khot, Saket and Vishnoi [18] showed that even the Sparsest Cut relaxation has an integrality gap $\Omega(\log\log n)$.

Cheeger, Kleiner and Naor [13] have recently exhibited a $(\log n)^{\Omega(1)}$ integrality gap for Non-Uniform Sparsest Cut, via very different techniques.

5. Algorithms for Unique Games

When Khot introduced the Unique Games Conjecture, he also introduced a Semidefinite Programming relaxation. Charikar, Makarychev and Makarychev [12] provide a tight analysis of the approximation guarantee of that Semidefinite Program, showing that, given a unique game with range Σ in which a $1-\epsilon$ fraction of the equations can be satisfied, it is possible to find in polynomial time a solution that satisfies at least a $1/\Sigma^{O(\epsilon)}$ fraction of constraints.

This is about as good as can be expected, because earlier work had shown that if the Unique Games Intractability Conjecture holds, then there is no polynomial time algorithm able to satisfy a $1/\Sigma^{o_\Sigma(\epsilon)}$ fraction of constraints in a unique game with range Σ in which a $(1-\epsilon)$ fraction of equations is satisfiable. Furthermore, the analysis of Charikar, Makarychev and Makarychev [12] is (unconditionally) known to be tight for the specific Semidefinite Programming relaxation used in their algorithm because of the integrality gap result of Khot and Vishnoi [26] discussed in the previous section.

Recently, Arora, Barak and Steurer [6] have devised an algorithm that satisfies in time $2^{n^{\epsilon^{\Omega(1)}}}$ a constant fraction of the equations in an instance of unique games in which it is possible to satisfy a $1-\epsilon$ fraction of equations. Although this result is far from refuting the Unique Games Intractability Conjecture, it casts some doubts on the Unique Games NP-hardness Conjecture. The following stronger form of the $P \neq NP$ conjecture is generally considered to be very likely: that for every NP-hard problem that is a $c > 0$ such that the problem cannot be solved with worst-case running time faster than 2^{n^c}, where n is the size of the input. This means that if the running time of the Arora-Barak-Steurer algorithm could be improved to $2^{n^{o(1)}}$ for a fixed ϵ, the Unique Games NP-hardness Conjecture would be in disagreement with the above conjecture about NP-hard problems, and would have to be considered unlikely.

References

[1] N. Alon and V.D. Milman, λ_1*, isoperimetric inequalities for graphs, and superconcentrators*, Journal of Combinatorial Theory, Series B **38** (1985), no. 1, 73–88.

[2] Noga Alon, *Eigenvalues and expanders*, Combinatorica **6** (1986), no. 2, 83–96.

[3] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy, *Proof verification and hardness of approximation problems*, Journal of the ACM **45** (1998), no. 3, 501–555, Preliminary version in *Proc. of FOCS'92.*

[4] S. Arora and S. Safra, *Probabilistic checking of proofs: A new characterization of NP*, Journal of the ACM **45** (1998), no. 1, 70–122, Preliminary version in *Proc. of FOCS'92.*

[5] Sanjeev Arora and Boaz Barak, *Computational complexity: A modern approach*, Cambridge University Press, 2009.

[6] Sanjeev Arora, Boaz Barak, and David Steurer, *Subexponential algorithms for unique games and related problems*, Proceedings of the 51st IEEE Symposium on Foundations of Computer Science, 2010.

[7] Sanjeev Arora, James Lee, and Assaf Naor, *Euclidean distortion and the sparsest cut*, Proceedings of the 37th ACM Symposium on Theory of Computing, 2005, pp. 553–562.

[8] Sanjeev Arora, Satish Rao, and Umesh Vazirani, *Expander flows and a* $\sqrt{\log n}$*-approximation to sparsest cut*, Proceedings of the 36th ACM Symposium on Theory of Computing, 2004.

[9] M. Bellare, O. Goldreich, and M. Sudan, *Free bits, PCP's and non-approximability – towards tight results*, SIAM Journal on Computing **27** (1998), no. 3, 804–915, Preliminary version in *Proc. of FOCS'95.*

[10] M Bellare, S. Goldwasser, C. Lund, and A. Russell, *Efficient probabilistically checkable proofs and applications to approximation*, Proceedings of the 25th ACM Symposium on Theory of Computing, 1993, See also the errata sheet in *Proc of STOC'94*, pp. 294–304.

[11] Christer Borell, *Geometric bounds on the Ornstein-Uhlenbeck velocity process*, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **70** (1985), 1–13.

[12] Moses Charikar, Konstantin Makarychev, and Yury Makarychev, *Near-optimal algorithms for unique games*, Proceedings of the 38th ACM Symposium on Theory of Computing, 2006, pp. 205–214.

[13] Jeff Cheeger, Bruce Kleiner, and Assaf Naor, *A* $(\log n)^{\Omega(1)}$ *integrality gap for the sparsest cut SDP*, arxiv:0910.2024, 2009.

[14] S.A. Cook, *The complexity of theorem proving procedures*, Proceedings of the 3rd ACM Symposium on Theory of Computing, 1971, pp. 151–158.

[15] Pierluigi Crescenzi, Riccardo Silvestri, and Luca Trevisan, *On weighted vs unweighted versions of combinatorial optimization problems*, Information and Compututation **167** (2001), no. 1, 10–26.

[16] Charles Delorme and Svatopluk Poljak, *Combinatorial properties and the complexity of a max-cut approximation*, European J. of Combinatorics **14** (1993), no. 4, 313–333.

[17] ———, *Laplacian eigenvalues and the maximum cut problem*, Mathematical Programing **62** (1993), 557–574.

[18] Nikhil Devanur, Subhash Khot, Rishi Saket, and Nisheeth Vishnoi, *Integrality gaps for sparsest cut and minimum linear arrangement problems*, Proceedings of the 38th ACM Symposium on Theory of Computing, 2006, pp. 537–546.

[19] Uriel Feige and Gideon Schechtman, *On the optimality of the random hyperplane rounding technique for MAX CUT*, Random Structures and Algorithms **20** (2002), no. 3, 403–440.

[20] Michel X. Goemans and David P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, Journal of the ACM **42** (1995), no. 6, 1115–1145, Preliminary version in *Proc. of STOC'94*.

[21] Johan Håstad, *Some optimal inapproximability results*, Journal of the ACM **48** (2001), no. 4, 798–859.

[22] R.M. Karp, *Reducibility among combinatorial problems*, Complexity of Computer Computations (R.E. Miller and J.W. Thatcher, eds.), Plenum Press, 1972, pp. 85–103.

[23] Subhash Khot, *On the power of unique 2-prover 1-round games*, Proceedings of the 34th ACM Symposium on Theory of Computing, 2002, pp. 767–775.

[24] ———, *Inapproximability of np-complete problems, discrete fourier analysis, and geometry*, Proceedings of the International Congress of Mathematicians, 2010.

[25] Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O'Donnell, *Optimal inapproximability results for MAX-CUT and other two-variable CSPs?*, Proceedings of the 45th IEEE Symposium on Foundations of Computer Science, 2004, pp. 146–154.

[26] Subhash Khot and Nisheeth Vishnoi, *The unique games conjecture, integrality gap for cut problems and the embeddability of negative type metrics into ℓ_1*, Proceedings of the 46th IEEE Symposium on Foundations of Computer Science, 2005, pp. 53–63.

[27] Robert Krauthgamer and Yuval Rabani, *Improved lower bounds for embeddings intoL1*, SIAM Journal on Computing **38** (2009), no. 6, 2487–2498.

[28] Leonid A. Levin, *Universal search problems*, Problemi Peredachi Informatsii **9** (1973), 265–266.

[29] Elchanan Mossel, Ryan OÕDonnelland, and Krzysztof Oleszkiewicz, *Noise stability of functions with low influences: Invariance and optimality*, Annals of Mathematics **171** (2010), no. 1, 295–341.

[30] Prasad Raghavendra, *Optimal algorithms and inapproximability results for every CSP?*, Proceedings of the 40th ACM Symposium on Theory of Computing, 2008.

[31] Ran Raz, *A parallel repetition theorem*, SIAM Journal on Computing **27** (1998), no. 3, 763–803, Preliminary version in *Proc. of STOC'95.*

[32] V. I. Rotar, *Limit theorems for polylinear forms*, J of Multivariate Analysis **9** (1979), no. 4, 511Ð530.

[33] Michael Sipser, *Introduction to the theory of computation*, Thomson, 2005, Second Editon.

[34] L. Trevisan, G.B. Sorkin, M. Sudan, and D.P. Williamson, *Gadgets, approximation, and linear programming*, SIAM Journal on Computing **29** (2000), no. 6, 2074–2097.

[35] Vijay Vazirani, *Approximation algorithms*, Springer, 2001.

E-mail address: `trevisan@stanford.edu`

Computer Science Department, Stanford University, 353 Serra Mall Stanford CA 94305-9025

COUNTING SPECIAL POINTS: LOGIC, DIOPHANTINE GEOMETRY AND TRANSCENDENCE THEORY

THOMAS SCANLON

ABSTRACT. We expose a theorem of Pila and Wilkie on counting rational points in sets definable in o-minimal structures and some applications of this theorem to problems in diophantine geometry due to Masser, Peterzil, Pila, Starchenko, and Zannier.

1. INTRODUCTION

Over the past decade and a half starting with Hrushovski's proof of the function field Mordell-Lang conjecture [12], some of the more refined theorems from model theory in the sense of mathematical logic have been applied to problems in diophantine geometry. In most of these cases, the technical results underlying the applications concern the model theory of fields considered with some additional distinguished structure and the model theoretic ideas fuse algebraic model theory, the study of algebraic structures with a special emphasis on questions of definability, and stability theory, the development of abstract notions of dimension, dependence, classification, *et cetera* for the purpose of analyzing the class of models of a theory. Over this period, there has been a parallel development of the model theory of theories more suited for real analysis carried out under the rubric of o-minimality, but this theory did not appear to have much to say about number theory. Some spectacular recent theorems demonstrate the error of this impression.

In the paper [28], Pila presents an unconditional proof of a version of the so-called André-Oort conjecture about algebraic relations amongst the j-invariants of elliptic curves with complex multiplication using a novel technique coming from model theory. This proof of the André-Oort conjecture comes on the heels of reproofs of the Manin-Mumford conjecture (or, Raynaud's theorem [32]) by Pila and Zannier [29] (and then extended by Peterzil and Starchenko [22]) and a proof of a remarkable theorem due to Masser and Zannier [19] about simultaneous torsion for pairs of points on families of elliptic curves all of which employ Pila's method.

Each of these theorems is a beautiful and precise instantiation of the vague principle that algebraic relations on sets of arithmetically interesting points of geometric origin must be explained geometrically. As such, these results deserve their own survey independent from their proofs. However, while I agree that these theorems and their ilk are of intrinsic interest, there are two reasons why I shall focus on their proofs. First, the proof strategy employed coming as it does from the model theory of real geometry is the most striking common feature of these results. Secondly, the algebro-geometric overhead required for an accurate explication of how these problems fit into an overarching network of theorems and conjectures in diophantine geometry would overwhelm the remainder of this account and mask the fundamentally classical nature of Pila's approach to these problems.

With this preamble about what I shall not discuss, let me say what will appear in this paper. We begin in Section 2 with a sketch of the Pila-Zannier proof in a simplified case of algebraic relations on roots of unity. This sketch will not reveal the full strength of the method as the theorem in this special case has been known for many decades and has innumerable proofs. However, this proof shares the architecture of the proofs of the other geometrically more complicated theorems. In Section 3 we discuss the general theory of o-minimality. Section 4, the technical heart of this survey, concerns the Pila-Wilkie theorem on counting rational points in definable sets. With Section 5 we return to the diophantine problems, completing the proof sketch from Section 2 and then discussing some of the other theorems mentioned in the abstract.

On the same day that I was asked to prepare these notes for the Current Events Bulletin, I was asked to speak in the Bourbaki seminar about Pila's proof of the André-Oort conjecture. Shortly thereafter, I was asked to give a lecture series to the experts on André-Oort about the same topic. Foolhardily, I concluded that due to the similarity of these presentations, I would need only prepare one set of notes and accepted all three invitations. As a matter of fact, while there are some points of contact, these three sets of notes are radically different. The reader interested in an exposition of Pila's proof of the André-Oort conjecture pitched to the general mathematician should consult my notes for the Bourbaki seminar [35] while the reader who wishes to read a detailed précis of these proofs should read my notes for the Luminy lectures [34]. Even better, because the original papers [24] and [28] are well written and contain extensive introductions, the reader should go straight to the source.

2. Pila-Zannier argument for multiplicative group

We shall go into more detail about the proofs of the more sophisticated theorems announced in the introduction later in the paper, but let us sketch the method in a simple example for which, admittedly, many other proofs are known (see, for example, [32, 13]). The following theorem is a special case of the Manin-Mumford conjecture and was already proven by Mann [18] before the Manin-Mumford conjecture proper was enunciated.

Theorem 2.1. *Let $n \in \mathbb{Z}_+$ be a positive integer and let $\mathbb{G} = (\mathbb{C}^\times)^n$ be the n^{th} Cartesian power of the multiplicative group of the complex numbers. Let $G(x_1, \ldots, x_n) \in \mathbb{C}[x_1, \ldots, x_n]$ be a polynomial in n variables. Then the set*

$$\{(\zeta_1, \ldots, \zeta_n) \in \mathbb{G} : \text{ each } \zeta_i \text{ is a root of unity and } G(\zeta_1, \ldots, \zeta_n) = 0\}$$

is a finite union of cosets of subgroups of $\mathbb{G}$.

The Pila-Zannier argument in this case proceeds by observing that we have an analytic covering map $E : \mathbb{C}^n \to \mathbb{G}$ given by $(z_1, \ldots, z_n) \mapsto (e^{2\pi i z_1}, \ldots, e^{2\pi i z_n})$ and that relative to this covering, $\zeta = (\zeta_1, \ldots, \zeta_n)$ is tuple of roots of unity if and only if there is some $a \in \mathbb{Q}^n$ for which $E(a) = \zeta$. Thus, we may convert the problem of studying algebraic equations in roots of unity into the problem of understanding rational solutions to the transcendental equation $G(E(z)) = 0$.

In this form, we have not achieved much yet as sets defined by general complex analytic equations can be arbitrarily complicated. However, it is not necessary to consider E on all of $\mathbb{C}^n$. We could restrict E to a fundamental domain

$$D := \{z = (z_1, \ldots, z_n) \in \mathbb{C}^n : 0 \leq \operatorname{Re}(z_i) < 1 \text{ for each } i\}$$

obtaining a function $\widetilde{E} := E \upharpoonright D : D \to \mathbb{G}$, then it is still the case that $\zeta \in G$ is a tuple of roots of unity if and only if there is some $a \in \mathbb{Q}^n \cap D$ with $\tilde{E}(a) = \zeta$. The advantage of this move is that $\widetilde{E}$ lives in logically well-behaved structure while E does not. That is, even though from the point of view of complex analysis, the map E is just about the best function one could hope to study, from the point of view of mathematical logic it has a very complicated theory as the kernel of E is the set $\mathbb{Z}^n$ and with the ring structure inherited from $\mathbb{C}^n$ the theory of this structure suffers from Gödel incompleteness phenomena. On the other hand, using the real and imaginary part functions to identify $\mathbb{C}$ with $\mathbb{R}^2$, the function $\widetilde{E}$ is definable in the structure $\mathbb{R}_{\exp} := (\mathbb{R}, +, \times, \exp, \leq, 0, 1)$ of the real field considered together with the real exponential function.

Why is this important? The theory of $\mathbb{R}_{\exp}$ is *o-minimal*, which, technically, means that every definable subset of the universe is a finite union of points and intervals, but which means in practice that the definable sets in any number of variables admit a geometric structure theory. In particular, the set

$$\widetilde{X} := \{z \in D : G(\widetilde{E}(z)) = 0\}$$

is such a definable set. We have transformed the problem of describing those n-tuples of roots of unity ζ for which $G(\zeta) = 0$ to the problem of describing the intersection $\mathbb{Q}^n \cap \widetilde{X}$ which at this level of generality may appear to be even more hopeless than the original problem as we know the problem of describing the rational solutions to algebraic equations is notoriously intractable (and is conjecturally impossible algorithmically [20]). However, if we punt on the problem for algebraic equations, then for the remaining transcendental equation we can give numerical bounds.

More precisely, for any set $Y \subseteq \mathbb{R}^m$ definable in some o-minimal expansion of the real field, we define the algebraic part of Y, Y^{alg}, to be the union of all connected, positive dimensional semi-algebraic sets (that, definable using Boolean combinations of polynomial inequalities) contained in Y. The counting theorem of Pila and Wilkie asserts that there are sub-exponentially many rational points in the transcendental part of Y, $Y \smallsetminus Y^{\text{alg}}$. That is, if we define

$$N(Y, t) := \#\{(\frac{a_1}{b_1}, \ldots, \frac{a_n}{b_n}) \in Y \smallsetminus Y^{\text{alg}} : (\forall i \leq n)|a_i| \leq t, 0 < b_i < t, a_i \in \mathbb{Z}, b_i \in \mathbb{Z}\}$$

then for each $\epsilon > 0$ there is a constant $C = C_\epsilon$ so that $N(Y, t) \leq Ct^\epsilon$ for all $t \geq 1$.

To use the Pila-Wilkie bound one must understand the set Y^{alg} and while this is far from a trivial problem, it has a geometric character and can be solved in cases of interest. In our proof sketch of Theorem 2.1 it follows from a theorem of Ax on a differential algebraic version of Schanuel's conjecture [3] that $\widetilde{X}^{\text{alg}}$ is the union of finitely sets each defined by affine equations with rational coefficients which under $\widetilde{E}$ are transformed into translates of algebraic subgroups of $\mathbb{G}$. We complete the argument by playing the Pila-Wilkie bound against lower bounds from Galois theory, but we delay the details until Section 5.

Each of the theorems in diophantine geometry proven using this method follows the general outline sketched above, though, of course, the individual steps tend to be more complicated as there is work involved in proving that the requisite covering map is definable in some o-minimal structure, the determination of the algebraic part of the relevant definable sets may be difficult, and one needs appropriate Galois theoretic or analytic number theoretic results for the lower bounds. While each step implicates some beautiful mathematics, it is the invocation of the counting principle for definable sets in o-minimal structures which gives this method its special character. As such, we shall focus this exposé on the ideas of definability in o-minimal structures and the counting theorem.

3. Introduction to o-minimality

O-minimality is not well-known to the general mathematician for at least a couple of reasons. First, it owes its existence and most of its development to mathematical logic and for sociological reasons having to do with logic's place at the boundary between mathematics and philosophy, the basics of first-order logic are not as widely known amongst mathematicians as are the basics of algebra, analysis and geometry. Secondly, the name itself, while technically accurate in that it expresses that all of the one-dimensional structure is reducible to the order and ties the subject to other parts of model theory, masks the fundamental nature of the subject which is a general but tame and geometric theory of real analysis. In his text [39], van den Dries argues that o-minimality may be a realization of the theory of topologie moderée proposed by Grothendieck in [11] in which topology and real analysis follow geometric intuitions. While I do not subscribe to this thesis in the strong form that o-minimality is *the* realization of topologie moderée, o-minimality certainly fits the bill for a geometric theory of real analysis.

What follows is a condensed introduction to the theory of o-minimality. The book [39] develops the general theory especially as it relates to o-minimal structures on the real numbers from a geometric point of view. The reader may wish to consult the lecture notes from the recent thematic program on o-minimality at the Fields Institute for a more recent account or Wilkie's Bourbaki notes [43] for a fuller survey. The foundational papers by Pillay, Steinhorn and Knight [30, 15, 31] remain vital.

Definition 3.1. By an *o-minimal structure* we mean a structure in the sense of first-order logic $(R, <, \ldots)$ where $<$ is a total order on R and the ellipses refer to some extra relations, functions and constants so that each definable (using parameters) subset of R is a finite union of points and intervals.

Remark 3.2. While in the applications we have in mind, the underlying ordered set is the set of real numbers with its usual ordering, the proofs of the counting principles pass through an analysis of parametrizations of definable sets in more general o-minimal structures. That is, the compactness theorem of first-order logic when used in the context of an o-minimal theory allows for a kind of nonstandard analysis which converts simple existence and finiteness results into uniformity theorems.

As noted in Remark 3.2 there are good reasons beyond the historical accident that logicians isolated the notion of o-minimality for treating o-minimal structures as structures in the sense of first-order logic, but it is possible to make sense of o-minimality without explicit reference to logic. Let us give an alternate definition of an o-minimal structure.

Definition 3.3. By an *o-minimal structure* we mean a nonempty totally ordered set $(R, <)$ given together with a Boolean algebras $\mathcal{D}_n$ of subsets of R^n for each $n \in \mathbb{Z}_+$ so that

1. each $\mathcal{D}_n$ is closed under the natural action of $\mathrm{Sym}(n)$ induced by permutations of coordinates,
2. each singleton set $\{a\}$ belongs to $\mathcal{D}_1$ for $a \in R$,
3. if $X \in \mathcal{D}_n$ and $Y \in \mathcal{D}_m$, then $X \times Y \in \mathcal{D}_{n+m}$,
4. if $\pi : R^{n+1} \to R^n$ is the projection onto the first n coordinates and $X \in \mathcal{D}_{n+1}$, then the image of X under π belongs to $\mathcal{D}_n$,
5. $\{\langle a, b\rangle \in R^2 : a < b\} \in \mathcal{D}_2$,
6. $\{\langle a, b\rangle \in R^2 : a = b\} \in \mathcal{D}_2$, and
7. every set in $\mathcal{D}_1$ is a finite union of singletons and intervals. That is, sets of the form $(-\infty, a) := \{x \in R : x < a\}$, $(a, b) := \{x \in R : a < x < b\}$, and $(b, \infty) := \{x \in R : b < x\}$ for some $a, b \in R$.

We refer to the sets in $\mathcal{D}_n$ as the *definable subsets of* R^n.

It is a routine matter to check that these two definitions of o-minimality are essentially the same. The closure conditions 1. - 4. on the class of definable sets in Definitions 3.3 and the initial requirement that each $\mathcal{D}_n$ be a Boolean algebra correspond to the syntactic operations of logical Boolean operations, permutation of variables, naming of parameters, conjunction of formulae with disjoint variables, and existential quantification. Condition 5. corresponds to definability of the ordering while condition 6. asserts the definability of equality. It is with condition 7. that we insist upon o-minimality. The choice of a first-order signature in Definition 3.1 corresponds to specifying a set of generators for the class of definable sets in the sense of Definition 3.3.

While this second presentation permits one to work with o-minimal structures without ever thinking about first-order logic, I contend that it is a mistake to do so. In practice, one establishes the definability of specific sets or conditions by exhibiting a definition. For example, one can show that if $X \subseteq R^n$ is a definable set in some ordered structure $(R, <, \ldots)$, then so is the closure $\overline{X}$ of X. Of course, this can be done by manipulating definable sets using projections and setwise Boolean operations, but the first-order formula describing $\overline{X}$ is transparently the usual definition of the closure:

$$\begin{aligned}(a_1, \ldots, a_n) \in \overline{X} \iff & (\forall x_1) \cdots (\forall x_n)(\forall y_1) \ldots (\forall y_n)[\bigwedge_{i \leq n} x_i < a_i < y_i \\ & \to (\exists z_1) \cdots (\exists z_n)((z_1, \ldots, z_n) \in X \ \& \bigwedge_{i \leq n} x_i < z_i < y_i)]\end{aligned}$$

One of the characteristic features of mathematical arguments using model theory is the way in which properties of definable sets in one structure may be deduced from arguments performed in logically equivalent structures through a kind of transfer principle. While these arguments are possible without the logical formalism, they are much more natural with it.

While there are some degenerate o-minimal structures whose underlying orders are discrete, we shall insist that an o-minimal structure be densely ordered without endpoints and for our applications the underlying ordered set is the set of real numbers with its usual ordering.

It is not hard to see that $(\mathbb{R}, <)$, the set of reals just with its order, is an o-minimal structure. It takes a little more work to demonstrate the $(\mathbb{R}, <, +)$, the set of real numbers considered as an ordered group, is an o-minimal structure. This latter structure is the basis of piecewise linear geometry and of tropical geometry. While the finiteness and uniformity properties of PL-geometry and tropical geometry are easy enough to demonstrate directly, the o-minimality of this underlying structure puts these results into context. That $(\mathbb{R}, +, \times, <, 0, 1)$, the set of real numbers considered as an ordered field, is o-minimal is a consequence of Tarski's theorem on elementary geometry (first proved in 1929, but only published in 1948 [38]) and is the basis of semi-algebraic geometry.

Wilkie proved that $\mathbb{R}_{\text{exp}} := (\mathbb{R}, <, +, \times, \exp)$, the real field considered together with the real exponential function is o-minimal [42]. Indeed, he proved a stronger theorem: the real field considered together with all *Pfaffian functions* is o-minimal where we say that a function $f : \mathbb{R} \to \mathbb{R}$ is Pfaffian if there is a finite sequence of functions $f_1, \ldots, f_n = f$ so that for each $i \leq n$ we have $f_i' = G_i(x, f_1, \ldots, f_n)$ for some polynomial $G_i(y_0, y_1, \ldots, y_i) \in \mathbb{R}[y_0, y_1, \ldots, y_n]$. The order one linear differential equation satisfied by exp exhibits the exponential function as a Pfaffian function. In later work, Speissegger [37] showed that one may take any o-minimal structure and adjoin all Pfaffian functions relative to that structure and thereby obtain a new o-minimal structure. In fact, Speissegger's theorem is more general in that he allows for the adjunction of so-called Rolle leaves to definable vector fields, but as this generalization is not germane to our applications, we omit the details.

In another direction, as a complement to their theorems on p-adic analytic functions, Denef and van den Dries [7] proved the o-minimality of $\mathbb{R}_{\text{an}}$, the real field considered together with all *restricted analytic functions*. That is, for each $n \in \mathbb{Z}_+$ and each power series

$$f := \sum_{\alpha \in \mathbb{N}^n} f_\alpha x_1^{\alpha_1} \cdots x_n^{\alpha_n} \in \mathbb{R}[[x_1, \ldots, x_n]]$$

which converges on the unit box we are given a function symbol $\widetilde{f}$ to be interpreted as

$$\widetilde{f}(a_1, \ldots, a_n) := \begin{cases} \sum f_\alpha a_1^{\alpha_1} \cdots a_n^{\alpha_n} \text{ if } -1 \leq a_1 \leq 1 \text{ for each } i \leq n \\ 0 \qquad\qquad\qquad\qquad\qquad \text{otherwise} \end{cases}$$

Combining these two expansions to form $\mathbb{R}_{\text{an,exp}} := \mathbb{R}_{an}(\exp)$, the real field considered with all real analytic functions and the real exponential function. The o-minimality of this structure was first established by van den Dries and Miller [41] and then the structure of its definable sets was more thoroughly explored by van den Dries, Macintyre, and Marker [40]. It is this structure which is relevant to the diophantine applications mentioned in the abstract.

For some perspective, one should note that many expansions of the real field by global real analytic functions, for example $(\mathbb{R}, +, \times, <, \sin)$, are *not* o-minimal as the zero set $\{x \in \mathbb{R} : \sin(x) = 0\}$ is infinite but discrete. Likewise, for a sufficiently general smooth function $f : \mathbb{R} \to \mathbb{R}$ the expansion $(\mathbb{R}, +, \times, <, f \restriction$

$[0,1])$ is not o-minimal. Since o-minimal structures have a tame geometry and we know that functions arising from real analysis tend to be wild, these examples ought not to be so surprising. On the other hand, it follows from work of Rolin, Speissegger and Wilkie [33] that there is no ultimate o-minimal structure on the real numbers. Indeed, for any $\mathcal{C}^\infty$ function $f : \mathbb{R} \to \mathbb{R}$ one can find two other functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ so that $f = g + h$ but both structures $(\mathbb{R}, +, \times, \leq, g)$ and $(\mathbb{R}, +, \times, \leq, h)$ are o-minimal. Consequently, if we understand o-minimality as the formalization of the concept of topologie moderée, then there is not a single structure which captures all of tame geometry.

From the fact that a specific structure is o-minimal, one may immediately deduce strong finiteness properties of the definable subsets of the line. For example, if $f : \mathbb{R} \to \mathbb{R}$ is definable in some o-minimal structure on the real numbers, then the zero set of f consists of finitely many points and finitely many intervals. If f falls into some natural class for which an identity principle holds, as when f is real analytic, then we see that if f is not identically zero, it has only finitely many zeros. This applies, for instance, to the case that f is built via finitely many applications of sums, differences, products and compositions from the identity function, scalar multiplication and the real exponential function.

The simplicity of definable sets coming explicitly from o-minimality's definition, while often unexpected, is the basis for a far deeper structure theory of definable sets in any number of variables. For example, it follows very easily from o-minimality that in a sufficiently rich o-minimal structure, if one has a definable family of definable nonempty sets, then this family admits a definable choice function. Let us make this statement precise and prove it in detail.

Definition 3.4. We say that the o-minimal structure $(R, <, \ldots)$ is *sufficiently rich* if it has at least the structure of an ordered abelian group $(R, +, -, 0, 1, <, \cdots)$ with at least one positive element called 1 named by a constant.

It is an easy exercise to show that if an o-minimal structure is sufficiently rich, then the underlying group is divisible.

Definition 3.5. Let $(R, <, \ldots)$ be an o-minimal structure. By a *definable family* of definable subsets of R^n we mean a definable set $X \subseteq R^n \times R^m$ for some positive natural number m where we regard R^m as parametrizing the family where to $b \in R^m$ we associate the set $X_b := \{a \in R^n : \langle a, b \rangle \in X\}$. Sometimes, we write a definable family of definable sets as $\{X_b\}_{b \in B}$ where $B \subseteq R^n$ is a definable set containing the projection of X to R^m.

Definition 3.6. By a *definable choice function* for the definable family $\{X_b\}_{b \in B}$ we mean a definable function $f : B \to R^n$ for which $f(b) \in X_b$ for each $b \in B$.

With these definitions in place, let us prove the existence of definable choice functions.

Proposition 3.7. *Let $(R, +, -, 0, 1, <, \cdots)$ be a sufficiently rich o-minimal structure and $\{X_b\}_{b \in B}$ a definable family of non-empty definable subsets of R^n. Then there is a definable choice function $f : B \to R^n$ for this family.*

Proof. We work by induction with the case of $n = 0$ being trivial. For the inductive case of $n + 1$, let $b \in B$ and consider the set

$$A(b) := \{y \in R : (\exists x \in R^n) \langle x, y \rangle \in X_b\}$$

The family $\{A(b)\}_{b \in B}$ is a definable family of nonempty definable subsets of R. Let us define a choice function h for this family. For $b \in B$, if $A(b) = R$, define $h(b) := 0$. If $A(b) \neq R$, then by o-minimality it has a finite non-empty boundary. Consider the least boundary point y. If $y \in A(b)$, define $h(b) := y$. Otherwise, one of three situations must obtain: $(-\infty, y) \subseteq A(b)$, $(y, \infty) = A(b)$ or $(y, z) \subseteq A(y)$ for z the second boundary point of $A(b)$. Define $h(b) := y - 1$ in the first case, $A(b) := y + 1$ in the second case, and $h(b) := \frac{y+z}{2}$ in the last case.

For $b \in B$ define

$$Z_b := \{x \in R^n \ : \ \langle x, h(b) \rangle \in X_b\}$$

Then $\{Z_b\}_{b \in B}$ is a definable family of nonempty subsets of R^n which has a definable choice function $g : B \to R^n$ by induction. Our desired choice function $f : B \to R^{n+1}$ is then given by $b \mapsto \langle g(b), h(b) \rangle$. □

The fundamental theorem of o-minimality is the cell decomposition theorem which asserts, roughly, that every definable set may be partitioned into finitely many *cells*, definable sets which are definably homeomorphic to balls possibly in a lower dimensional space. Since this theorem is so important for all of the work in o-minimality and is invoked if only implicitly throughout the proof of the Pila-Wilkie theorem on counting rational points, we shall go into detail.

Let us begin by giving a precise definition of *cell*. As we define this notion, we shall define the dimension of a cell.

Definition 3.8. Let $(R, <, \cdots)$ be an o-minimal structure. We define the class of cells in R^n by recursion on n and for each cell X we define $\dim(X)$. The set R^0 is a singleton which is the only cell in R^0. We define $\dim(R^0) := 0$. If $X \subseteq R^n$ is a cell and $f : X \to R$ is a definable, continuous function, then the graph of f,

$$\Gamma(f)_X := \{(x, y) \in R^n \times R \ : \ x \in X \ \& \ f(x) = y\}$$

is a cell with $\dim(\Gamma(f)) := \dim(X)$. If $g : X \to R$ is another definable, continuous function for which $(\forall x \in X) f(x) < g(x)$, then the interval

$$(f, g)_X := \{(x, y) \in R^n \times R \ : \ x \in X \ \& \ f(x) < y < g(x)\}$$

is a cell and $\dim((f, g)_X) := \dim(X) + 1$. Likewise,

$$(-\infty, f)_X := \{(x, y) \in R^n \times R \ : \ x \in X \ \& \ y < f(x)\}$$

and

$$(g, \infty)_X := \{(x, y) \in R^n \times R \mid x \in X \ \& \ g(x) < y\}$$

are cells of dimension $\dim(X) + 1$.

Let us specialize to the case of cells in $R = R^1$ for a moment. A definable function (indeed, any function) $f : R^0 \to R$ is given by choosing a single point $a \in R$. Via the natural identification $R^0 \times R^1 = R$, we see that $\Gamma(f)_{R^0} = \{a\}$. Likewise, if $g : R^0 \to R$ is another definable function for which $(\forall x \in R^0) f(x) < g(x)$, then taking b to be the sole value of g we have $a < b$ and again with respect to the natural identification of $R^0 \times R^1$ with R, the interval $(f, g)_{R^0}$ is simply the usual interval (a, b). That is, cells in R are singletons and open intervals, possibly unbounded. To say that $(R, <, \ldots)$ is o-minimal is precisely the same as to say that every definable set $X \subseteq R$ may be expressed as a finite union of cells. The cell decomposition theorem asserts that this property of definable sets in one space generalizes to any dimension.

Theorem 3.9. *Let $(R, <, \cdots)$ be an o-minimal structure. Given any $m \in \mathbb{Z}_+$ and any finite sequence $X_1, \ldots, X_n \subseteq R^m$ of definable subsets of R^m there is a finite sequence of cells $C_1, \ldots, C_k \subseteq R^m$ which partitions R^m and whose restriction to each X_ℓ also partitions X_ℓ. That is, $C_i \cap C_j = \varnothing$ for $i \neq j$, $R^m = \bigcup_{j=1}^k C_j$, and for each $\ell \leq m$ we have $X_\ell = \bigcup_{\{j \leq k \,:\, C_j \cap X_\ell \neq \varnothing\}} C_j$.*

In the course of the proof of Theorem 3.9 one shows that definable functions are very regular. The monotonicity theorem says that for any definable function $f : R \to R$ possibly after removing finitely many points one may partition the domain into finitely many intervals so that on each interval f is continuous and either constant or strictly monotone. In general, the piecewise continuity theorem says that if $f : R^n \to R$ is a definable function, then one may decompose the domain into finitely many cells so that the restriction of f to each cell is continuous. If the structure $(R, <, \ldots)$ includes at least the structure of an ordered field, then it makes sense to speak of the derivative of a function. In this case, for any $k \in \mathbb{Z}_+$ one may choose a cell decomposition of the domain so that f is C^k on each cell. Likewise, in the cell decomposition theorem itself, one may take the functions defining the cells to have any prescribed degree of smoothness. Unfortunately, it is not the case that one may always take the cells to be defined by analytic functions, but in many cases of interest, for example in $\mathbb{R}_{\text{an,exp}}$, one may take the defining functions to be real analytic.

It it hard to overstate the strength of the geometric consequences of the cell decomposition theorem and its refinements. For example, it implies a kind of infinitesimal rigidity on the topology of definable sets living in a definable family.

It is a fairly easy consequence of the cell decomposition theorem applied to the total space of a definable family that given a definable family $\{X_b\}_{b \in B}$ of definable sets, the cells required for the cell decompositions of the various fibres also vary in definable families. It follows from this uniformity theorem that at least when the underlying ordered set is the set of real numbers with its usual ordering that the topology of the sets in a definable family is rigid.

Proposition 3.10. *If $\{X_b\}_{b \in B}$ is a definable family of definable sets in some o-minimal structure on the real numbers (with the usual ordering) then there are only finitely many homemorphism types represented in the family.*

As a corollary of Proposition 3.10 we obtain a theorem of Khovanski [14] on fewnomials. To be fair, while the theorem on fewnomials which we shall discuss is logically a consequence of Proposition 3.10 both temporally and intellectually it is prior. Khovanski's work on fewnomials inspired much of the development of theory of o-minimality and many of his specific results underly Wilkie's proof of the o-minimality of $\mathbb{R}_{\text{exp}}$. Moreover, the argument we outline below is patterned on Khovanski's own proof through the passage from polynomials of indeterminate degree to exponential polynomials.

Theorem 3.11. *For fixed integers k and n there are only finitely many homemorphism types amongst the following sets*

$$\{(a_1, \ldots, a_n) \in (\mathbb{R}_+)^n \;:\; \sum_{i=1}^k f_i a_1^{m_{i,1}} \cdots a_n^{m_{i,n}} = 0\}$$

as $(f_1, \ldots, f_k)$ ranges through $\mathbb{R}^k$ and m ranges through the k by n matrices with natural number coordinates.

To prove Theorem 3.11 we observe that it suffices show that there are only finitely many homemorphism types even if we allow m to range through $M_{k\times n}(\mathbb{R})$ rather than merely $M_{k\times n}(\mathbb{N})$. The above family of semialgebraic sets may be embedded into the following $\mathbb{R}_{\exp}$-definable family.

$$\{(a, f, m) \in (\mathbb{R}_+)^n \times (\mathbb{R}^k \times (\mathbb{R}^n)^k) \ : \ \sum_{i=1}^k f_i \prod_{j=1}^n \exp(m_{i,j} \ln(a_i)) = 0\}$$

The finiteness of the number of homemorphism types is now a special case of Proposition 3.10.

4. Counting rational points in o-minimal definable sets

The key technical result behind the Pila-Wilkie theorem on counting rational points may be seen as, in some sense, a dual version of the cell decomposition theorem in that rather than concluding that a general definable set may be pieced together from definable subsets each defined in a very simple way it is shown that a general definable set may be covered by finitely many definable sets each of which is parametrized by a unit ball via functions with small derivatives. Before we go into detail about this technical result on parametrizations, let us go into some more detail about the counting theorem itself and then explain how an appropriate parametrization theorem could yield these bounds.

Definition 4.1. The usual multiplicative height function $H : \mathbb{Q} \to \mathbb{N}$ is defined by $H(0) := 0$ and $H(\frac{a}{b}) := \max\{|a|, |b|\}$ when $a, b \in \mathbb{Z} \smallsetminus \{0\}$ and $\gcd(a, b) = 1$. We extend H to a function, still denoted by H, on $\mathbb{Q}^n$ by $H(x_1, \ldots, x_n) := \max\{H(x_i) : i \leq n\}$. This is not the standard projective height, but it works well for the purposes of our counting problems. Given any subset $X \subseteq \mathbb{R}^n$ and a number $t \geq 1$ we define

$$X(\mathbb{Q}, t) := \{a \in \mathbb{Q}^n : a \in X \ \& \ H(a) \leq t\}$$

and define $N(X, t) := \#X(\mathbb{Q}, t)$ to be the number of points in $X(\mathbb{Q}, t)$.

If X happens to contain all of $\mathbb{Q}^n$, then $N(X, t)$ is asymptotic to a constant times t^{2n}. This simple observation combined with the even simpler remark that $\mathbb{R}^n$ itself is definable in any o-minimal structure on $\mathbb{R}$ shows that one cannot hope to show that $N(X, t)$ grows more slowly than any power of t for a general set $X \subseteq \mathbb{R}^n$ definable in some o-minimal expansion of $(\mathbb{R}, <)$. Somewhat less trivial considerations show that some further restrictions are required. For example, if $X \subseteq \mathbb{R}^n$ happens to contain the graph of a polynomial with integer coefficients $f : \mathbb{R}^{n-1} \to \mathbb{R}$, then $N(X, t)$ will grow at least on the order of $t^{2(n-1)/d}$ where d is the degree of f. Of course, there are more general algebraic varieties which have many rational points and if X should contain one of these, it, too, will have many rational points. Thus, to have any hope of proving a bound on $N(X, t)$ for general X definable in some o-minimal structure on $\mathbb{R}$, we must exclude those algebraic sets which have too many rational points. We achieve this by excluding *all* semi-algebraic sets.

Definition 4.2. We say that a set $X \subseteq \mathbb{R}^n$ is *semi-algebraic* if it is definable in the structure $(\mathbb{R}, +, \times, <)$ of the real numbers considered as an ordered field. More concretely, a set is semi-algebraic if it is a finite Boolean combination of sets defined by conditions of the form $f(x_1, \ldots, x_n) > 0$ where $f \in \mathbb{R}[X_1, \ldots, X_n]$ is polynomial with real coefficients in n variables. Given a set $X \subseteq \mathbb{R}^n$ we define X^{alg}, the *algebraic part* of X, to be the union of all sets Z where $Z \subseteq X$ is semi-algebraic, connected, and has dimension at least one. We define X^{trans}, the transcendental part of X, to be $X \smallsetminus X^{\mathrm{alg}}$.

There is no reason to expect the algebraic part of X to be semi-algebraic itself even if X is definable in some particularly well behaved o-minimal structure on the real numbers. For example, consider the following set which definable in $\mathbb{R}_{\exp}$.

$$X := \{(x,y,z) \in \mathbb{R}^3 \ : \ x > 0 \,\&\, z = x^y = \exp(y \ln(x))\}$$

The set X has dimension two, being the graph of a definable continuous function on $\mathbb{R}_+ \times \mathbb{R}$, but it does not contain any two dimensional semi-algebraic sets. On the other hand, its algebraic part consists of the union of the following countably infinite collection of one-dimensional semi-algebraic sets

$$\{(x,y,z) \in \mathbb{R}^3 \ : \ x > 0 \,\&\, y = \frac{n}{m} \,\&\, x^n = z^m\}$$

as $\frac{n}{m}$ ranges through the rational numbers. Thus, in this case X^{alg} is a properly infinite union of semi-algebraic sets.

With these definitions in place we may state the counting theorem from [24].

Theorem 4.3. *Let $X \subseteq \mathbb{R}^n$ be a subset of some Cartesian power of the real numbers which is definable in some o-minimal structure on $\mathbb{R}$. Then for each $\epsilon > 0$ there is some constant $C = C_\epsilon$ so that for $t \geq 1$ one has $N(X^{trans}, t) \leq Ct^\epsilon$.*

Remark 4.4. Strengthenings of Theorem 4.3 are known. The proof of Theorem 4.3 passes through a proof a uniform version in which X is allowed to vary in a definable family and the bound is shown to hold for a bigger set than simply X^{trans}. For purposes of the application of Theorem 4.3 to the André-Oort conjecture and some related problems it is necessary to count algebraic points of small degree rather than merely rational points. These bounds may be deduced from the bounds for rational points [26].

Remark 4.5. Examples have been constructed showing that in general one cannot hope for better universal bounds. However, one might hope that if X is definable in a particularly nice way, then the bounds may be strengthened. In the strongest forms, these strengthenings assert that certain transcendental equations have no non-obvious algebraic solutions. I have in mind conjectures along the lines of Schanuel's conjecture on the transcendence of the exponential function that if $\alpha_1, \ldots, \alpha_n$ are complex numbers which are $\mathbb{Q}$-linearly independent then the transcendence degree of the field $\mathbb{Q}(\alpha_1, \ldots, \alpha_n, e^{\alpha_1}, \ldots, e^{\alpha_n})$ is at least n [17], André's conjectures on G-functions [1], and the Kontsevich-Zagier conjectures on periods [16]. While these conjectures are inaccessible to contemporary techniques, Wilkie's conjecture about sets definable using the real exponential function may be within reach.

Conjecture 4.6. *Let $X \subseteq \mathbb{R}^n$ be a subset of a Cartesian power of $\mathbb{R}$ definable in $\mathbb{R}_{\exp}$. There there are constants C and K depending only on X so that for $t \geq 1$ we have $N(X^{trans}, t) \leq C(\log t)^K$.*

The proof Theorem 4.3 and the proofs [27] of partial results towards Conjecture 4.6 rely on a general geometric theorem about parametrizations of definable sets, referred to as a dual form of cell decomposition in the first paragraph of this section, and a linear algebraic argument to show that the rational points on the image of a ball under a sufficiently smooth function must lie in a small number of algebraic hypersurfaces.

Definition 4.7. Let $(R, +, \cdot, <, \ldots)$ be an o-minimal structure on an ordered field, $X \subseteq R^n$ be a definable set of dimension k in some Cartesian power of R and $r \in \mathbb{Z}_+$ be a positive integer. We say that $\phi = (\phi_1, \ldots, \phi_n) : (0,1)^k \to R^n$ is a partial r-parametrization of X if

- ϕ is definable,
- the range of ϕ is contained in X, and
- for each $i \leq n$ and multi-index $\alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{N}^k$ with $|\alpha| = \sum \alpha_i \leq r$ we have $|\frac{\partial^{|\alpha|} f_i}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}(x)| \leq 1$ for all $x \in (0,1)^n$.

By an *r-parametrization* of X we mean a finite set S of partial r-parametrizations of X for which X is covered by the ranges of the functions in S.

The main theorem about parametrizations is that r-parametrizations exist for every $r \in \mathbb{Z}_+$ and every sufficiently bounded definable set in an o-minimal structure.

Theorem 4.8. *Let $(R, +, \times, <, \ldots)$ be an o-minimal structure expanding an ordered field, $X \subseteq [-1,1]^n$ be a definable subset of the unit n-cube in R for some $n \in \mathbb{Z}_+$, and $r \in \mathbb{Z}_+$ be a positive integer. Then X admits an r-parametrization.*

For the intended applications, we work in an o-minimal structure on $\mathbb{R}$. What, then, is the point of the greater generality? The proof of Theorem 4.8, even in the case that the underlying ordered field is the field of real numbers, makes essential use of parametrizations of definable sets in more general o-minimal structures through a kind of nonstandard analysis. That is, by proving even individual instances of the parametrization theorem, say, in general o-minimal structures, one may deduce via the compactness theorem of first-order logic that the theorem holds uniformly in definable families. From the point of view of the ultimate theorem, these compactness arguments are hidden, but they are crucial for the proofs.

Theorem 4.8 generalizes a theorem of Yomdin on the existence of r-parametrizations for real semi-algebraic sets [45, 44] and its proof follows Gromov's version of the proof in the semi-algebraic setting [10]. Of course, for purposes of Theorem 4.3 in which we count only rational points in the transcendental part of X, the parametrization theorem for semialgebraic sets on its own does not help.

The logical structure of the proof of Theorem 4.8 is similar to that of the cell decomposition theorem (Theorem 3.9). For both of these theorems, the one-dimensional case itself is an immediate consequence of the definition of o-minimality, but to carry out the induction using the definable choice functions given by Proposition 3.7 one performs a concurrent induction showing that definable functions have strong regularity properties. For the cell decomposition theorem, this takes the form of the monotonicity theorem in dimension one and the piecewise continuity theorem in higher dimensions. For the parametrization theorem, this takes the form of a *reparameterization* theorem in all dimensions and a strengthening of the reparameterization theorem in dimension one.

To say what is meant by reparameterization, we need the definition of *strongly bounded*.

Definition 4.9. Let $(R, +, \cdot, 0, 1, <, \ldots)$ be an o-minimal expansion of an ordered field. We say that the set $X \subseteq R^n$ is *strongly bounded* if there is a natural number $N \in \mathbb{N}$ for which $X \subseteq [-N, N]^n$. We say that a function $f : Y \to R^n$ is strongly bounded if its range is strongly bounded.

Definition 4.10. Let $(R, +, \cdot, 0, 1, <, \cdots)$ be an o-minimal expansion of an ordered field. Given a positive integer $r \in \mathbb{Z}_+$ and a definable function $f = (f_1, \ldots, f_n) : (0,1)^m \to R^n$ we say that an r-parameterization S of the open unit box $(0,1)^m$ is an *r-reparameterization of f* if for each $\alpha \in \mathbb{N}^m$ with $|\alpha| \leq r$, $\phi \in S$ and $j \leq n$ the function $f_j \circ \phi$ is r-times differentiable and $\frac{\partial^{|\alpha|}(f_j \circ \phi)}{\partial x_1^{\alpha_1} \cdots \partial x_m^{\alpha_m}}$ is strongly bounded.

Remark 4.11. In the case that the underlying field is simply $\mathbb{R}$, then bounded and strongly bounded have the same meaning. The meanings diverge only for non-archimedian fields. In the real case, if $f : (0,1)^m \to \mathbb{R}^n$ is bounded and sufficiently smooth, then any r-parameterization of $(0,1)^m$ is an r-reparameterization. The utility of the concept of a reparameterization is only seen through the nonstandard analytic arguments.

The key auxiliary result in the proof of Theorem 4.8 is the reparameterization theorem.

Theorem 4.12. *Let $(R, +, \cdot, 0, 1, <, \cdots)$ be an o-minimal expansion of an ordered field, $m \in \mathbb{Z}_+$, $n \in \mathbb{Z}_+$, and $r \in \mathbb{Z}_+$ be three positive integers. If $f : (0,1)^m \to R^n$ is strongly bounded, then there exists an r-reparameterization of f. Moreover, if $m = n = 1$, the reparameterization S may be chosen so that for each $\phi \in S$ either ϕ or $\phi \circ f$ is a polynomial with strongly bounded coefficients.*

Theorem 4.12 and thereby the full Theorem 4.8 are established via a constructive and concrete argument in the base case and then via an inductive argument relying heavily upon definable choice and the cell decomposition theorems in higher dimensions.

Let us explain now how Theorem 4.3 follows from Theorem 4.8. We start with a set $X \subseteq \mathbb{R}^n$ definable in some o-minimal expansion of the real numbers and a number $\epsilon > 0$. Breaking X into the 2^n pieces $X_\tau := \{(a_1, \ldots, a_n) \in X : |a_i^{\tau_i}| \leq 1 \text{ for all } i \leq n\}$ as τ ranges over $\{\pm 1\}^n$, it suffices to estimate $N(X_\tau, t)$ for each such τ. Since the map $x \mapsto \frac{1}{x}$ does not effect our height function, we may replace each X_τ with its image under $(x_1, \ldots, x_n) \mapsto (x_1^{\tau_1}, \ldots, x_n^{\tau_n})$ and thereby we may assume that $X \subseteq [-1, 1]^n$. From Theorem 4.8 we know that for each $r \in \mathbb{Z}_+$ the set X admits an r-parameterization. We shall choose r depending on ϵ so that the existence of such an r-parameterization implies that $N(X, t) \leq Ct^\epsilon$.

At this point, the argument follows the lines of other constructive arguments bounding numbers of rational solutions and is similar in spirit to Bombieri's proof of the Mordell conjecture [5]. The key result is the following proposition whose proof is ultimately embedded in the paper [4].

Proposition 4.13. *For $m, n, d \in \mathbb{N}$ with $m < n$ there are numbers $r \in \mathbb{Z}_+$ and $\epsilon = \epsilon(m, n, d)$ and $C = C(m, n, d)$ in $\mathbb{R}_+$ so that for any C^r function $\phi : (0,1)^m \to \mathbb{R}^n$ with range X and $t \geq 1$ the set $X(\mathbb{Q}, t)$ is contained in at most Ct^ϵ hypersurfaces of degree d and $\epsilon(m, n, d) \to \infty$ as $d \to \infty$.*

The proof of Proposition 4.13 requires some nontrivial but elementary combinatorial estimates and some careful but again elementary analytic considerations, but ultimately it is based on a simple, but ubiquitous in the theory of diophantine approximations, observation: if an integer has absolute value less than one, it is zero.

With Proposition 4.13 in place, Theorem 4.3 follows by induction: those exceptional hypersurfaces which have full dimension intersection with X are part of X^{alg} and those which intersect X in a lower dimensional set contribute little to $N(X^{\mathrm{trans}}, t)$ by induction.

5. Diophantine applications

In general, it is not true that if $X \subseteq \mathbb{R}^n$ is definable in some o-minimal structure on the real numbers that $(X^{\mathrm{trans}})(\mathbb{Q})$ is finite. For example, if X is the graph of the function $x \mapsto 2^x$, then its algebraic part is empty, but for each integer a we have $\langle a, 2^a \rangle \in X(\mathbb{Q})$. However, in some cases of independent number theoretic interest the upper bounds of Theorem 4.3 may be played against lower bounds coming from Galois theory. In the introduction to this paper, we sketched a version of this argument due to Pila and Zannier to show that algebraic relations amongst roots of unity may always be explained by multiplicative dependencies. In this section we shall explain some of the more sophisticated results proven using variations of this method

Let us return to the argument sketched in the introduction giving a few more details to complete the proof. Recall that we wish to prove that for $G(x_1, \ldots, x_n) \in \mathbb{C}[x_1, \ldots, x_n]$ a polynomial over the complex numbers in n variables the set

$$X := \{(\zeta_1, \ldots, \zeta_n) \in (\mathbb{C}^\times)^n \ : \ G(\zeta_1, \ldots, \zeta_n) = 0 \ \& \ \text{ each } \zeta_i \text{ is a root of unity } \}$$

is a finite union of cosets of groups. We observed that if we define

$$D := \{z \in \mathbb{C} : 0 \leq \mathrm{Re}(z) < 1\}$$

and $\widetilde{E} : D^n \to (\mathbb{C}^\times)^n$ by

$$(z_1, \ldots, z_n) \mapsto (e^{2\pi i z_1}, \ldots, e^{2\pi i z_n})$$

then via the usual interpretation of $\mathbb{C}$ as $\mathbb{R}^2$ using the real and imaginary part functions, the function $\widetilde{E}$ is definable in $\mathbb{R}_{\exp}$ and the set X is the image under $\widetilde{E}$ of the set $\widetilde{X}(\mathbb{Q})$ where $\widetilde{X} := \{(z_1, \ldots, z_n) \in D^n : G(e^{2\pi i z_1}, \ldots, e^{2\pi i z_n}) = 0\}$ is an $\mathbb{R}_{\exp}$-definable set.

Before we can apply Theorem 4.3 to give even numerical bounds on the distributions of the points in $\widetilde{X}(\mathbb{Q})$, we need to compute $\widetilde{X}^{\mathrm{alg}}$. The key to this computation is Ax's function field version of the Schanuel Conjecture [3].

Theorem 5.1. *If $\gamma_1(t), \ldots, \gamma_n(t) \in t\mathbb{C}[[t]]$ are power series over the complex numbers with no constant term which are linearly independent over $\mathbb{Q}$, then the transcendence degree over $\mathbb{C}(t)$ of the field $\mathbb{C}(t, \gamma_1(t), \ldots, \gamma_n(t), \exp(\gamma_1(t)), \ldots, \exp(\gamma_n(t)))$ is at least n.*

It follows from Theorem 5.1 that if $\gamma : (0,1) \to \widetilde{X}$ were a semi-algebraic curve, then the components of γ would satisfy a nontrivial linear dependence over $\mathbb{Q}$.

This alone is not enough for the determination of $\widetilde{X}^{\mathrm{alg}}$. Rather, we now use another property of definable sets in o-minimal structures on the real numbers: every countable definable set is finite. For each positive $k \leq n$ we consider the set M_k of maximal k-dimensional affine spaces: affine spaces V (translates of vector subspaces) of dimension k for which $\dim(V \cap \widetilde{X}) = k$ but for which there is some point $a \in (V \cap \widetilde{X})$ for which is no $k+1$-dimensional affine space W which meets $\widetilde{X}$ near a in dimension $k+1$ set. It is not hard to see that relative to the usual representations of affine spaces via affine equations, each of the sets M_k is definable. It takes a little more work using properties of the covering map $\widetilde{E}$ to see that every element of M_k is defined over $\mathbb{Q}$. Hence, each M_k is countable and therefore finite. Combining this argument with Theorem 5.1, we see that $\widetilde{X}^{\mathrm{alg}} = \bigcup_{k=1}^{n} \bigcup_{H \in M_k} H$. For the algebraic part, it is now clear that $\widetilde{X}^{\mathrm{alg}}(\mathbb{Q})$ is a finite union of cosets of groups (intersected with $[0,1)^n$).

To complete the proof, we must show that $(\widetilde{X}^{\mathrm{trans}})(\mathbb{Q})$ is finite using the bounds from Theorem 4.3. For this we need a reduction: we may assume that G is an irreducible polynomial defined over some a number field K. If you are comfortable with basic algebraic geometry, this reduction is standard and quite easy and will be given in the parenthetical sentences to follow. Otherwise, take this point as given or just follow the argument in the case when G is in fact a polynomial over a number field.

(For the reduction, observe that Theorem 2.1 is equivalent to the apparent generalization where the hypersurface defined by G is replaced by a general subvariety. For a variety $Y \subseteq \mathbb{A}^n_{\mathbb{C}}$, if

$$Z := \overline{Y(\mathbb{C}) \cap \{(\zeta_1, \ldots, \zeta_n) \in (\mathbb{C}^\times)^n \ : \ \text{each } \zeta_i \text{ is a root of unity } \}}$$

then $Y(\mathbb{C})$ and $Z(\mathbb{C})$ meet the n-tuples of roots of unity in the same set. So, we may assume $Y = Z$. It then follows from Lagrange interpolation that Y is defined over the algebraic numbers as it contains a Zariski dense set of algebraic points. Since only finitely many equations are required to define Y, it is, in fact, defined over a number field.)

Let us now estimate $N(\widetilde{X}, t)$. Suppose that $z \in \widetilde{X}(\mathbb{Q}, t)$. We can write $z = (\frac{a_1}{b_1}, \ldots, \frac{a_n}{b_n})$ where each a_i and b_i is an integer, $0 \leq a_i < b_i \leq t$ and $(a_i, b_i) = 1$ (and $b_i = 1$ if $a_i = 0$). Exponentiating, we have $G(e^{2\pi i \frac{a_1}{b_1}}, \ldots, e^{2\pi i \frac{a_n}{b_n}}) = 0$. Let $L := K((e^{2\pi i \frac{a_1}{b_1}}, \ldots, e^{2\pi i \frac{a_n}{b_n}})$ be the field obtained by adjoining the coordinates of $\widetilde{E}(z)$ to K. Since G has coefficients in K, for any automorphism $\sigma : L \to L$ over K, we have $G(\sigma(\widetilde{E}(z))) = 0$. Since $e^{2\pi i \frac{a_j}{b_j}}$ is a primitive b_j^{th} root of unity, we know that $\sigma(e^{2\pi i \frac{a_j}{b_j}}) = e^{2\pi i \frac{a'}{b_j}}$ for some integer a' with $0 \leq a' < b_i$. Moreover, it follows from basic Galois theory that the orbit of $e^{2\pi i \frac{a_j}{b_j}}$ under the Galois group of L over K has cardinality at least $\varphi(b_j)/[K : \mathbb{Q}]$ where φ is Euler's totient function given by $\varphi(n) := \#(\mathbb{Z}/n\mathbb{Z})^\times$. A simple computation shows that for any constant $C > 0$ and number $\epsilon < 1$ we have $\varphi(n) > Cn^\epsilon$ for $n \gg 0$. Putting all these observations together, we see that if $t \in \mathbb{Z}_+$ and $N(\widetilde{X}^{\mathrm{trans}}, t) > N(\widetilde{X}^{\mathrm{trans}}, t-1)$, then $N(\widetilde{X}, t) \geq \frac{1}{[K:\mathbb{Q}]} \varphi(t)$. For $t \gg 0$ this would violate Theorem 4.3 with $\epsilon < 1$. Hence, there must be some t for which every element of $(\widetilde{X}^{\text{trans}})(\mathbb{Q})$ has height at most t. That is, this set is finite.

For the other applications we have mentioned, the statements are intrinsically more complicated as they refer to more sophisticated geometries and the proofs are correspondingly more involved. However, the fundamental strategies are the same.

Let us consider the theorem of Masser and Zannier [19] about torsion on elliptic curves. They consider the family of elliptic curves presented in their affine Legendre form where E_λ is defined by the affine planar equation $y^2 = x(x-1)(x-\lambda)$ for $\lambda \in \mathbb{C} \smallsetminus \{0,1\}$. From theory of elliptic curves, E_λ considered together with the point at infinity has a unique structure of an algebraic group with that point at infinity as the identity. For a fixed complex number a we might consider the set of λ for which the point $(a, \sqrt{a(a-1)(a-\lambda)})$ is torsion in the group $E_\lambda(\mathbb{C})$. It is not hard to see that for $a = 0$ or $a = 1$, then these points are always torsion. On the other hand, for every other a there are only countably many λ for which this point is torsion in $E_\lambda(\mathbb{C})$. Nevertheless, computing the rational functions which define the multiplication by n map on E_λ it is fairly easy to show that for any such a there will be infinitely many λ for which $(a, \sqrt{a(a-1)(a-\lambda)})$ is torsion. Masser and Zannier address the question: if we consider two number a and b, for how many λ are $(a, \sqrt{a(a-1)(a-\lambda)})$ and $(b, \sqrt{b(b-1)(b-\lambda)})$ both torsion in $E_\lambda(\mathbb{C})$? In the special case of $a = 2$ and $b = 3$ they given an answer.

Theorem 5.2. *There are only finitely many complex numbers λ for which*

$$P_\lambda := (2, \sqrt{2(2-\lambda)})$$

and

$$Q_\lambda := (3, \sqrt{6(3-\lambda)})$$

are torsion in $E_\lambda(\mathbb{C})$.

Remark 5.3. The proof of Theorem 5.2 applies perfectly well to any two numbers a and b for which the points

$$P_\lambda^a := (a, \sqrt{a(a-1)(a-\lambda)})$$

and

$$Q_\lambda^b := (b, \sqrt{b(b-1)(b-\lambda)})$$

are linearly independent over $\mathbb{Z}$ in the group $E_\lambda(\mathbb{Q}(\lambda))$.

The proof of Theorem 5.2 follows the pattern of the proof of Theorem 2.1 we have outlined above. For each elliptic curve E_λ, the the theory of analytic uniformizations gives a complex analytic covering map $\pi_\lambda : \mathbb{C} \to E_\lambda(\mathbb{C})$. As with the usual exponential function, this covering is not definable in any o-minimal expansion of the real numbers. However, if we restrict π_λ to a fundamental domain, it is. Moreover, at the cost of treating π_λ as simply a real analytic function, we may normalize the fundamental domain so that the domain of π_λ is the square $[0,1) \times [0,1)$ and the map π_λ is a group homomorphism when $[0,1)$ is given the usual wrap around additive structure. With some work, one can show that the two variable (or, really, four real variable) function $(\lambda, z) \mapsto \pi_\lambda(z)$ is definable in

$\mathbb{R}_{an,\exp}$ relative to the usual interpretation of $\mathbb{C}$ in $\mathbb{R}$ and when z is restricted to $[0,1) \times [0,1)$. Masser and Zannier then study the set

$$\begin{aligned} \widetilde{X} \ := \ & \{(x_1, y_1, x_2, y_2) \in [0,1)^4 : (\exists \lambda)\pi_\lambda(x_1, y_1) = (2, \sqrt{2(2-\lambda)}) \\ & \& \ \pi_\lambda(x_2, y_2) = (3, \sqrt{6(3-\lambda)})\} \end{aligned}$$

Visibly, $\widetilde{X}$ is definable in $\mathbb{R}_{\mathrm{an,exp}}$ and it is not hard to see that the rational points on $\widetilde{X}$ all come from λ for which P_λ and Q_λ are simultaneously torsion. Transcendence results about the Weierstraß $\wp$-function are used in place of Ax's theorem to show that $\widetilde{X}^{\mathrm{alg}}$ is empty and a theorem of David [6] about the degree of the field extension required to define elliptic curves with elements of specified order plays the rôle of the calculation of the degree of a cylcotomic extensions.

It bears noting that the published sketch of Theorem 5.2 avoids an explicit reference to definability in o-minimal structures as Pila had proved a provisional version of Theorem 4.3 for subanalytic surfaces without invoking the theory of o-minimality [25]. On the other hand, due to the work of Peterzil and Starchenko [23] on the uniform definability of theta functions in $\mathbb{R}_{\mathrm{an,exp}}$, it follows that the question of simultaneous torsion in families of higher dimensional abelian varieties may be analyzed via these methods.

Finally, let us close with Pila's proof of the André-Oort conjecture for modular curves. We shall introduce the André-Oort conjecture via the classical theory of complex elliptic curves. Unlike most other approaches to this problem where one might (or might not) define the terms using complex analysis but then address the questions with a more number theoretic theory, Pila's proof appeals directly to the complex analytic presentation of the problem.

As we observed above, for every elliptic curve E over the complex numbers, one can find a complex analytic surjective group homomorphism $\pi : \mathbb{C} \to E(\mathbb{C})$. The kernel of $\mathbb{C}$ is a lattice which after making a linear change of variables we may express as $\ker \pi = \mathbb{Z} \oplus \mathbb{Z}\tau$ for some complex number $\tau \in \mathfrak{h} := \{z \in \mathbb{C} : \operatorname{Im}(z) > 0\}$. Conversely, for any $\tau \in \mathfrak{h}$, the complex analytic group $E_\tau(\mathbb{C}) := \mathbb{C}/(\mathbb{Z} + \mathbb{Z}\tau)$ is complex analytically isomorphic to a complex algebraic curve with an algebraic group structure, which we shall continue to denote by E_τ. From the general theory of covering spaces, it is not hard to see that the endomorphisms of the elliptic curve E_τ correspond to complex numbers μ for which $\mu(\mathbb{Z} + \mathbb{Z}\tau) \leq \mathbb{Z} + \mathbb{Z}\tau$. A short computation shows that for most choices of τ, the number μ gives an endomorphism only when μ is an integer. On the other hand, if τ satisfies a quadratic equation over $\mathbb{Q}$, then there will be some endomorphisms not coming from $\mathbb{Z}$. This is the reason why elliptic curves whose endomorphism rings are strictly larger than $\mathbb{Z}$ are said to have *complex multiplication* or to be *CM*.

There is an analytic function $j : \mathfrak{h} \to \mathbb{C}$ having the property that $E_\tau(\mathbb{C})$ and $E_\sigma(\mathbb{C})$ are isomorphic as elliptic curves if and only if $j(\tau) = j(\sigma)$. We refer to the value $j(\tau)$ as the *j-invariant* of the elliptic curve E_τ. Let us say that a complex number ζ is a *special point* if it is the j-invariant of an elliptic curve with complex multiplication. By the above discussion, we see that a number is special if and only if it is the value of the analytic j-function on a quadratic imaginary number. The André-Oort conjecture in this case predicts the form of the algebraic subvarieties $X \subseteq \mathbb{A}^n_{\mathbb{C}}$ of affine n-space which contain a Zariski dense set of n-tuples of special points. Specializing to the case of $n = 2$, it proposes a solution to the question of for which polynomials $g(x,y) \in \mathbb{C}[x,y]$ are there infinitely many pairs (ξ,ζ) of special points for which $g(\xi,\zeta) = 0$? This case was solved early in the investigations around the André-Oort conjecture, first assuming the Riemann Hypothesis by Edixhoven [8] and then unconditionally by André [2].

Clearly, if ξ is a special point, then the set algebraic varieties $\{\xi\} \times \mathbb{A}^1_{\mathbb{C}}$ and $\mathbb{A}^1 \times \{\xi\}$ contain Zariski dense sets of special points as does the whole plane $\mathbb{A}^2_{\mathbb{C}}$. It follows from the general theory of coverings, that for each $n \in \mathbb{Z}_+$ there is a polynomial $P_n(x,y) \in \mathbb{C}[x,y]$ for which the function $\tau \mapsto P_n(j(n\tau), j(\tau))$ is identically zero. From this presentation, it is clear that the curve defined by the vanishing of P_n contains a Zariski dense set of special points for if τ is quadratic imaginary, then so is $n\tau$ and *vice versa*. The André-Oort conjecture (for the j-line) says that these are the only algebraic varieties other than points which can contain a Zariski dense set of special points.

Theorem 5.4 (Pila). *Let $X \subseteq \mathbb{A}^n_{\mathbb{C}}$ be an irreducible algebraic subvariety of affine n-space over the complex numbers. Suppose that the set*

$$\{(\xi_1,\ldots,\xi_n) \in X(\mathbb{C}) \ : \ \textit{each } \xi_i \textit{ is the j-invariant of a CM-elliptic curve}\}$$

is Zariski dense in X, then X is defined by equations of the form $P_m(x_i,x_j) = 0$ and $x_k = \xi$ for ξ a special point.

The proof of Theorem 5.4 follows our by now familiar pattern. First, Pila observes that j restricted to a fundamental domain is definable in $\mathbb{R}_{\text{an,exp}}$ by work of Peterzil and Starchenko [21]. He then moves from a study of X to that of $\widetilde{X}$, the inverse image of $X(\mathbb{C})$ via j (or really, the map $(z_1,\ldots,z_n) \mapsto (j(z_1),\ldots,j(z_n)))$ restricted to its fundamental domain, which is a definable set in $\mathbb{R}_{\text{an,exp}}$. He then must determine $\widetilde{X}^{\text{alg}}$ and does so using considerations of the action of the modular group showing that the algebraic part comes from the pre-images of finitely many varieties of the desired form. At this point, the goal is to show that if X does not already have the desired form, then there are only finitely many quadratic imaginary points in $\widetilde{X}^{\text{trans}}$. The counting theorem, Theorem 4.3, applied to rational points, but Pila deduces the same kinds of bounds for algebraic points of bounded degree [26]. Thus, for any $\epsilon > 0$ there is some constant C for which the number of quadratic imaginary points of height at most t in $\widetilde{X}^{\text{trans}}$ is at most Ct^{ϵ}. As in the proof of Theorem 2.1, he reduces to the case that X is defined over a number field and observes that if there are special points coming from $\widetilde{X}^{\text{trans}}$, then all of their Galois conjugates are also in this set. At this point, he estimates the size of these orbits from below using Siegel's theorem on the growth of the class number [36] to find that for $\epsilon < 2$ one has a lower bound of Ct^{ϵ} thus contradicting the upper bound from the counting theorem.

Remark 5.5. Theorem 5.4 had been proven previously by Edixhoven and Yafaev [9] under the assumption of the Generalized Riemann Hypothesis for quadratic imaginary fields. Their proof shares the same kind strategy at the end: find upper bounds geometrically and lower bounds via Galois theory and analytic number theory.

Remark 5.6. The paper in which the proof of Theorem 5.4 appears [28] includes proofs of theorems in the direction of the Pink-Zilber conjectures. On the other hand, while many parts of this argument succeed when applied to other Shimura varieties, some steps are incomplete. For example, it is known that the analytic covering maps for the moduli spaces of principally polarized abelian varieties are definable in $\mathbb{R}_{\text{an,exp}}$ (again, after suitable restriction) [23] and it seems plausible that the arguments employed to determine the algebraic parts of inverse images of algebraic varieties by Cartesian powers of j should work for these maps, too, but to date no one has carried out the details. More importantly, the lower bounds on the size of the Galois orbits of the special points are not yet known unconditionally.

References

[1] Yves André. *G-functions and geometry*. Aspects of Mathematics, E13. Friedr. Vieweg & Sohn, Braunschweig, 1989.

[2] Yves André. Finitude des couples d'invariants modulaires singuliers sur une courbe algébrique plane non modulaire. *J. Reine Angew. Math.*, 505:203–208, 1998.

[3] James Ax. On Schanuel's conjectures. *Ann. of Math. (2)*, 93:252–268, 1971.

[4] E. Bombieri and J. Pila. The number of integral points on arcs and ovals. *Duke Math. J.*, 59(2):337–357, 1989.

[5] Enrico Bombieri. The Mordell conjecture revisited. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 17(4):615–640, 1990.

[6] Sinnou David. Points de petite hauteur sur les courbes elliptiques. *J. Number Theory*, 64(1):104–129, 1997.

[7] J. Denef and L. van den Dries. p-adic and real subanalytic sets. *Ann. of Math. (2)*, 128(1):79–138, 1988.

[8] Bas Edixhoven. Special points on the product of two modular curves. *Compositio Math.*, 114(3):315–328, 1998.

[9] Bas Edixhoven and Andrei Yafaev. Subvarieties of Shimura varieties. *Ann. of Math. (2)*, 157(2):621–645, 2003.

[10] M. Gromov. Entropy, homology and semialgebraic geometry. *Astérisque*, (145-146):5, 225–240, 1987. Séminaire Bourbaki, Vol. 1985/86.

[11] Alexandre Grothendieck. Esquisse d'un programme. In *Geometric Galois actions, 1*, volume 242 of *London Math. Soc. Lecture Note Ser.*, pages 5–48. Cambridge Univ. Press, Cambridge, 1997. With an English translation on pp. 243–283.

[12] Ehud Hrushovski. The Mordell-Lang conjecture for function fields. *J. Amer. Math. Soc.*, 9(3):667–690, 1996.

[13] Ehud Hrushovski. The Manin-Mumford conjecture and the model theory of difference fields. *Ann. Pure Appl. Logic*, 112(1):43–115, 2001.

[14] A. G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991. Translated from the Russian by Smilka Zdravkovska.

[15] Julia F. Knight, Anand Pillay, and Charles Steinhorn. Definable sets in ordered structures. II. *Trans. Amer. Math. Soc.*, 295(2):593–605, 1986.

[16] Maxim Kontsevich and Don Zagier. Periods. In *Mathematics unlimited—2001 and beyond*, pages 771–808. Springer, Berlin, 2001.

[17] Serge Lang. *Introduction to transcendental numbers*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1966.

[18] Henry B. Mann. On linear relations between roots of unity. *Mathematika*, 12:107–117, 1965.

[19] David Masser and Umberto Zannier. Torsion anomalous points and families of elliptic curves. *C. R. Math. Acad. Sci. Paris*, 346(9-10):491–494, 2008.

[20] B. Mazur. Questions of decidability and undecidability in number theory. *J. Symbolic Logic*, 59(2):353–371, 1994.

[21] Ya'acov Peterzil and Sergei Starchenko. Uniform definability of the Weierstrass $\wp$ functions and generalized tori of dimension one. *Selecta Math. (N.S.)*, 10(4):525–550, 2004.

[22] Ya'acov Peterzil and Sergei Starchenko. Around pila-zannier: the semiabelian case. *preprint*, 2009.

[23] Ya'acov Peterzil and Sergei Starchenko. Definability of restricted theta functions and families of abelian varieties. *preprint*, 2010.

[24] J. Pila and A. J. Wilkie. The rational points of a definable set. *Duke Math. J.*, 133(3):591–616, 2006.

[25] Jonathan Pila. Rational points on a subanalytic surface. *Ann. Inst. Fourier (Grenoble)*, 55(5):1501–1516, 2005.

[26] Jonathan Pila. On the algebraic points of a definable set. *Selecta Math. (N.S.)*, 15(1):151–170, 2009.

[27] Jonathan Pila. Counting rational points on a certain exponential-algebraic surface. *Ann. Inst. Fourier (Grenoble)*, 60(2):489–514, 2010.

[28] Jonathan Pila. O-minimality and the André-Oort conjecture for $\mathbb{C}^n$. *Ann. of Math. (2)*, (to appear).

[29] Jonathan Pila and Umberto Zannier. Rational points in periodic analytic sets and the Manin-Mumford conjecture. *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl.*, 19(2):149–162, 2008.

[30] Anand Pillay and Charles Steinhorn. Definable sets in ordered structures. I. *Trans. Amer. Math. Soc.*, 295(2):565–592, 1986.

[31] Anand Pillay and Charles Steinhorn. Definable sets in ordered structures. III. *Trans. Amer. Math. Soc.*, 309(2):469–476, 1988.

[32] M. Raynaud. Sous-variétés d'une variété abélienne et points de torsion. In *Arithmetic and geometry, Vol. I*, volume 35 of *Progr. Math.*, pages 327–352. Birkhäuser Boston, Boston, MA, 1983.

[33] J.-P. Rolin, P. Speissegger, and A. J. Wilkie. Quasianalytic Denjoy-Carleman classes and o-minimality. *J. Amer. Math. Soc.*, 16(4):751–777 (electronic), 2003.

[34] Thomas Scanlon. The André-Oort conjecture via counting rational points in definable sets (after J. Pila). (in preparation). Mini-courses around the Pink-Zilber conjecture at Luminy, May 2011.

[35] Thomas Scanlon. The André-Oort conjecture via o-minimality (after J. Pila). (in preparation). Séminaire Bourbaki. Vol. 2010/2011.

[36] C. L. Siegel. Über die Classenzahl quadratischer Zahlkörper. *Acta Arith.*, (1):83–86, 1935.

[37] Patrick Speissegger. The Pfaffian closure of an o-minimal structure. *J. Reine Angew. Math.*, 508:189–211, 1999.

[38] Alfred Tarski. *A Decision Method for Elementary Algebra and Geometry*. RAND Corporation, Santa Monica, Calif., 1948.

[39] Lou van den Dries. *Tame topology and o-minimal structures*, volume 248 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 1998.

[40] Lou van den Dries, Angus Macintyre, and David Marker. The elementary theory of restricted analytic fields with exponentiation. *Ann. of Math. (2)*, 140(1):183–205, 1994.

[41] Lou van den Dries and Chris Miller. On the real exponential field with restricted analytic functions. *Israel J. Math.*, 85(1-3):19–56, 1994.

[42] A. J. Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. *J. Amer. Math. Soc.*, 9(4):1051–1094, 1996.

[43] Alex J. Wilkie. o-minimal structures. *Astérisque*, (326):Exp. No. 985, vii, 131–142 (2010), 2009. Séminaire Bourbaki. Vol. 2007/2008.

[44] Y. Yomdin. C^k-resolution of semialgebraic mappings. Addendum to: "Volume growth and entropy". *Israel J. Math.*, 57(3):301–317, 1987.

[45] Y. Yomdin. Volume growth and entropy. *Israel J. Math.*, 57(3):285–300, 1987.

E-mail address: `scanlon@math.berkeley.edu`

UNIVERSITY OF CALIFORNIA, BERKELEY, DEPARTMENT OF MATHEMATICS, EVANS HALL, BERKELEY, CA 94720-3840, USA

SPACES OF GRAPHS AND SURFACES
- ON THE WORK OF SØREN GALATIUS

Ulrike Tillmann

21.11.2010

1. Introduction

Galatius' most striking result is easy enough to state. Let Σ_n be the symmetric group on n letters and F_n be the free (non-abelian) group with n generators. The symmetric group Σ_n acts naturally by permutation on the n generators of F_n, and every permutation gives thus rise to an automorphism of the free group. Galatius proves that the map $\Sigma_n \hookrightarrow \mathrm{Aut}F_n$ in homology induces an isomorphism in degrees less than $(n-1)/2$. The homology of the symmetric groups in these ranges is well understood. In particular, in common with all finite groups, it has no non-trivial rational homology. By Galatius' theorem, in low degrees this then is also true for $\mathrm{Aut}F_n$:

$$H_*(AutF_n) \otimes \mathbb{Q} = 0 \quad \text{for} \quad 0 < * < 2n/3,$$

as had been conjectured by Hatcher and Vogtmann.

In this lecture I will put this result in context and explain the connection with previous work on the mapping class group of surfaces and the homotopy theoretic approach to a conjecture by Mumford on its rational, stable cohomology. Galatius' proof in [G] is inspired by this and at the same time improves the methods significantly. This in turn has led to further deep insights into the topology of moduli spaces of manifolds also in higher dimensions.

2. Groups and their (co)homology

I will first step back and say a bit more about the groups mentioned above and discuss their (co)homology in essentially algebraic terms. There are many parallels between mapping class groups and automorphisms of free groups. Indeed, much of the work on $\mathrm{Aut}F_n$ has been inspired by the work on the mapping class group as these groups show very similar behavior.

2.1. Groups of primary interest. I will first introduce the discrete groups that we will mainly be interested in.

2.1.1. The symmetric group Σ_n on n letters, is finite of size $n!$ and hardly needs further introduction. It has a presentation with generators the transpositions $\sigma_1, \dots, \sigma_{n-1}$ that swop two adjacent letters, and relations $\sigma_i^2 = 1$ and the braid relations $\sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1}$ and $\sigma_i\sigma_j = \sigma_j\sigma_i$ for $|i-j| > 1$.

2.1.2. The automorphism group $\mathrm{Aut}F_n$ is the group of invertible homomorphisms of the (non-abelian) free group F_n on n generators to itself. Closely related is the *outer automorphism group* $\mathrm{Out}F_n$ which is a quotient of $\mathrm{Aut}F_n$ by the normal subgroup $\mathrm{Inn}F_n$ of inner automorphisms given by conjugation by a fixed element of F_n. Both groups are infinite (for $n > 1$) and contain the symmetric group Σ_n as a subgroup.

The canonical map from the free group F_n to the free abelian group $\mathbb{Z}^n$ induces a surjective homomorphism

$$L : \mathrm{Aut}F_n \longrightarrow \mathrm{GL}(n, \mathbb{Z}),$$

to the general linear group. The inverse of the special linear group defines a subgroup $S\mathrm{Aut}F_n$ of index 2. A set of generators for this subgroup are the Nielsen transformations λ_{ij} and ρ_{ij} which multiply the ith generator of F_n by the jth on the left and right respectively, and leave all other generators fixed. A nice presentation of this subgroup is given by [Ge84]. To get a full set of generators, one needs to add an automorphism of determinant -1 such as the map that sends the first generator to its inverse and leaves all other generators fixed.

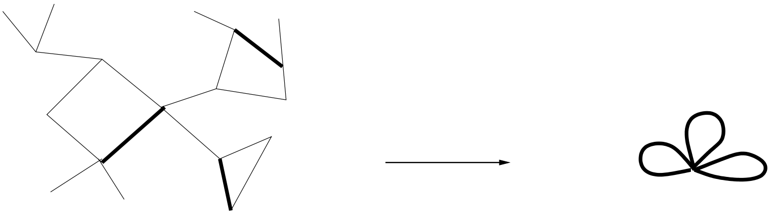


Figure 1: *Collapsing a maximal tree defines a homotopy equivalence.*

F_n is the fundamental group of a bouquet of n circles, or any graph G_n with Euler characteristic $1-n$ more generally. Let $\mathrm{HtEq}(G_n)$ denote the space of homotopy equivalences of G_n and $\mathrm{HtEq}(G_n; *)$ the subspace of homotopy equivalences that fix a basepoint. Their groups of components are $\mathrm{Out}F_n$ and $\mathrm{Aut}F_n$ respectively. Furthermore, each connected component is contractible. We thus have homotopy equivalences

$$\mathrm{HtEq}(G_n; *) \simeq \mathrm{Aut}F_n \quad \text{and} \quad \mathrm{HtEq}(G_n) \simeq \mathrm{Out}F_n.$$

2.1.3. The mapping class group $\Gamma_{g,1}$ *of an oriented surface* $S_{g,1}$ of genus g with one boundary component is the group $\pi_0(\mathrm{Diff}^+(S_{g,1}; \partial))$ of connected components of the group of orientation preserving diffeomorphisms of $S_{g,1}$ that fix the boundary point wise. Closely related is *the mapping class group* Γ_g *of an oriented, closed* *surface* S_g of genus g. They are generated by Dehn twists around simple closed curves defined by the following procedure: cut the surface along the given curve, twist one side by a full turn, and glue it back. A useful presentation was found by Wajnryb [W83]. It is not difficult to see that when two curves intersect once the associated Dehn twists satisfy the braid relation relation $aba = bab$; when two curves don't intersect their associated Dehn twist commute.

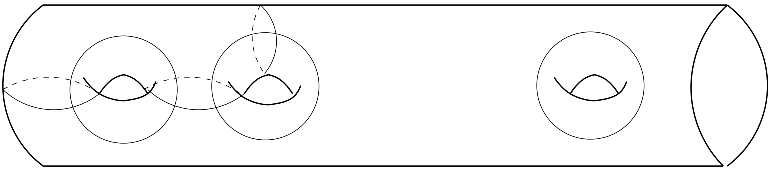


Figure 2: *Dehn twist around the* $2g+1$ *indicated curves generate* $\Gamma_{g,1}$.

Diffeomorphisms act on the first homology $H_1(S_g) = \mathbb{Z}^{2g}$ of the underlying surface, and when they are orientation preserving they preserve the intersection form. This defines a surjective representation

$$\Gamma_g \longrightarrow \mathrm{SP}(2g, \mathbb{Z}).$$

When the Euler characteristic of the underlying surfaces is negative, Earle and Eells [EE69] showed that the connected components of the diffeomorphism groups are contractible. We thus also have in this case homotopy equivalences

$$\mathrm{Diff}^+(S_{g,1}) \simeq \Gamma_{g,1} \quad \text{and} \quad \mathrm{Diff}^+(S_g) \simeq \Gamma_g.$$

2.1.4. Natural homomorphisms between these groups. For the first three groups we have $\Sigma_n \to \mathrm{Aut}F_n \to \mathrm{Out}F_n$ where the first map is induced by the permutation action on a set of generators for F_n and the second map is the quotient map. By gluing a disc to the boundary of the surface $S_{g,1}$ and extending diffeomorphisms by the identity, we get a natural map $\Gamma_{g,1} \to \Gamma_g$. Finally, every diffeomorphism of $S_{g,1}$ induces an automorphism of the fundamental group $\pi_1 S_{g,1} = F_{2g}$. This defines a map

$$\rho^+ : \Gamma_{g,1} \longrightarrow \mathrm{Aut}F_{2g}.$$

2.2. Group (co)homology. One way to study a discrete group G is through its homology $H_n(G)$ and cohomology $H^n(G)$ groups. These groups can be defined purely algebraically or as the homology and cohomology of a space BG. The space BG is determined (up to homotopy) by the fact that its fundamental group is G and its universal cover is contractible. In practice one constructs such spaces by finding a contractible space EG with a free G action. BG is then the orbit space EG/G. For an easy example, consider the integers acting by translations on the real line. A model for the space $B\mathbb{Z}$ is then given by the circle $S^1 = \mathbb{R}/\mathbb{Z}$.

The first homology group $H_1(G)$ is always the abelianisation $G/[G,G]$ of G. Hopf also found a purely algebraic formula for the second homology group $H_2(G)$. Both these groups can generally be computed when one has a presentation of G (indeed, often a subset of the relations suffices). Presentations for all the groups mentioned above are known by now.[1] For $n > 4$ and $g > 4$, the first two homology groups are as follows:

$$\begin{aligned}
&H_1(\Sigma_n) = H_2(\Sigma_n) = \mathbb{Z}/2\mathbb{Z}\\
&H_1(\mathrm{Aut}F_n) = H_2(\mathrm{Aut}F_n) = \mathbb{Z}/2\mathbb{Z}\\
&H_1(\Gamma_{g,1}) = 0; \qquad H_2(\Gamma_{g,1}) = \mathbb{Z}.
\end{aligned}$$

By work of Culler-Vogtmann [CV86] and Harer [H86] we know that both $\mathrm{Out}F_n$ and Γ_g have finite virtual cohomological dimensions:

$$vcd(\mathrm{Out}F_n) = 2n-3 \quad \text{and} \quad vcd(\Gamma_g) = 4g-5.$$

In particular this implies that the (co)homology in degrees above these dimensions is all torsion for both groups. Note that the virtual cohomological dimensions depend on n and g. In contrast, in what follows we will only be interested in the (co)homology that is independent of n and g.

2.3. Stable (co)homology and limit groups. The groups that we introduced in Section 2.1 come in families $\{G_n\}_{n\geq 0}$ indexed by the natural numbers. For the symmetric groups Σ_n, the automorphisms of free groups $\mathrm{Aut}F_n$ and the mapping class groups $\Gamma_{g,1}$ there are canonical inclusions $G_n \hookrightarrow G_{n+1}$. Indeed, $\Sigma_n \hookrightarrow \Sigma_{n+1}$ identifies the smaller group with those permutations that leave the $(n+1)$st letter fixed; $\mathrm{Aut}F_n \hookrightarrow \mathrm{Aut}F_{n+1}$ with those automorphisms of F_{n+1} that leave the $(n+1)$st generator fixed and send the first n generators to words not involving the $(n+1)$st; and $\Gamma_{g,1} \hookrightarrow \Gamma_{g+1,1}$ with those mapping classes that come from diffeomorphisms that restrict to the identity on $S_{g+1,1} \setminus S_{g,1}$, a torus with two boundary components. In each case we define the limit groups as $G_\infty := \lim_{n\to\infty} G_n$.

It is natural to ask how the homology of G_n is related to that of G_{n+1} or G_∞. In each of the three cases, the groups satisfy homology stability which means that for a fixed degree the homology does not change once n is large enough. This is the *stable* homology, or equivalently the homology of the group G_∞. For the symmetric groups this was first studied by Nakaoka [Na60], for the mapping class groups by Harer [H85] with improved ranges given by [I89] [B] [RW], and for the automorphism group of free groups by Hatcher and Vogtmann [HV98.C] (see also [HV04] [HVW], and [HW]). Homology stability theorems are generally quite tricky and difficult theorems to prove with the main techniques go back to Quillen, who studied the question for general linear groups. The following holds:

$H_*(\Sigma_n) \to H_*(\Sigma_{n+1})$ is an isomorphism in degrees $* < (n+1)/2$.

[1] Nielsen and McCool had given a presentations of $\mathrm{Aut}F_n$. Simplifications allowed Gersten [G84] to compute the second homology group. For the mapping class group the first presentation was given by Thurston and Hatcher. Building on this Harer determined the second homology group. Further simplifications led to Wajnryb's convenient presentation [W83][BW94].

$H_*(\mathrm{Aut}F_n) \to H_*(\mathrm{Aut}F_{n+1})$ is an isomorphism in degrees $* < (n-1)/2$.
$H_*(\mathrm{Aut}F_n) \to H_*(\mathrm{Out}F_n)$ is an isomorphism in degrees $* < (n-3)/2$.

$H_*(\Gamma_{g,1}) \to H_*(\Gamma_{g+1,1})$ is an isomorphism in degrees $* < 2g/3$.
$H_*(\Gamma_{g,1}) \to H_*(\Gamma_g)$ is an isomorphism in degrees $* < (2g+1)/3$.

These results are crucial for us because it is the homology of the limit group G_∞ that can be computed in each of the cases. Through homology stability also some information on the homology of each G_n can be obtained. For rational homology these stability ranges can often be improved. Indeed, for $\mathrm{Aut}F_n$ this is $2n/3$ as quoted in the introduction, see [HV98.C].

2.4. Products and group completion. At first it is counter-intuitive that the homology of the larger group G_∞ should be more easily determined than that of G_n. But note that we have natural product maps

$$\begin{aligned} \Sigma_n \times \Sigma_m &\longrightarrow \Sigma_{n+m}, \\ \mathrm{Aut}F_n \times \mathrm{Aut}F_m &\longrightarrow \mathrm{Aut}F_{n+m}, \\ \Gamma_{g,1} \times \Gamma_{h,1} &\longrightarrow \Gamma_{g+h,1}. \end{aligned}$$

The first two are given by having the first factor act on the first n points or generators and the second factor on the last m. In case of the mapping class group the product map is induced by gluing $S_{g,1}$ and $S_{h,1}$ to the legs of a pair of pants surface and extending the diffeomorphisms via the identity. Furthermore, the products are commutative up to conjugation by an element in the target group G_{n+m}.[2] It is a standard fact that conjugation induces the identity on group homology. So we see that the products on homology are graded commutative.

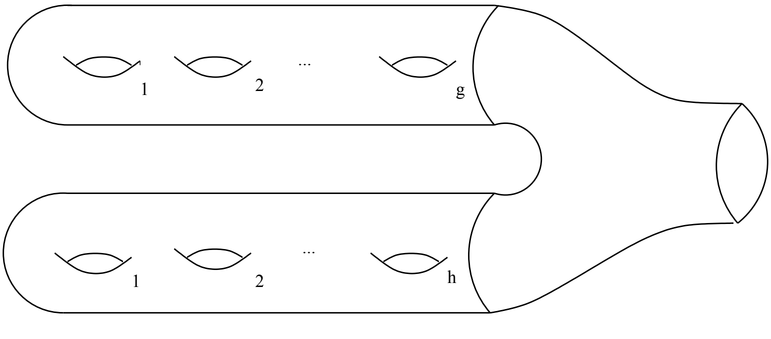


Figure 3: *Pair of pants product for surfaces.*

On the space level the above maps of groups induce a product on the disjoint union $M = \bigsqcup_n BG_n$ making it into a topological monoid. We will need to consider

[2] For the first two groups this is an element of the symmetric group Σ_{n+m} and its square is the identity; but for the mapping class group this corresponds to a twist of the glued on pair of pants surface, a braiding which is of infinite order.

its group completion. Group completion is a powerful tool but also one of the more mysterious constructions.

Just as discrete monoids have a group completion, so do topological monoids. It is a homotopy theoretic construction which associates to a topological monoid M the loop space ΩBM. (Here ΩX denotes the space of maps from a circle to X that send a base point in S^1 to a base point in X.) When $M = G$ is a discrete group, BG is the space mentioned above and the group completion is homotopic to G as it ought to be: $\Omega BG \simeq G$. In general, however, the homotopy of the monoid can be very different from that of its group completion. In our examples, the connected components of M have non-trivial and non-commutative fundamental groups but no higher homotopy while the group completion ΩBM has a small abelian fundamental group but highly non-trivial higher homotopy groups. Nevertheless, the homology groups are related very nicely. The group completion theorem (first instances of which are proved in [BP72] and [Q]) says that in general the homology of a connected component $\Omega_0 BM$ is the limit of the homology groups of the components of M.

To come back to our examples, $M = \bigsqcup_n BG_n$ has product that is commutative on homology. The group completion theorem can therefore be applied, and gives

$$H_*(G_\infty) = H_*(\Omega_0 B(\bigsqcup_n BG_n)).$$

In particular, the limit group G_∞ has the homology of a loop space. Indeed, a much stronger statement is true for our examples. We will see that G_∞ has actually the homology of an infinite loop space. While for the first two groups this has long been known, for the mapping class group it came as a surprise[3] [T97]. Below we will see how to identify these infinite loop spaces. Indeed for the symmetric group, a classical theorem in homotopy theory, the Barratt-Priddy-Quillen Theorem, asserts that the following two spaces are homotopic

$$\Omega B(\bigsqcup_n B\Sigma_n) \simeq \Omega^\infty S^\infty,$$

where $\Omega^\infty S^\infty = \lim_{N\to\infty} \Omega^N S^N$ is the limit space of maps from S^N to itself that fix a chosen basepoint. Its homotopy groups are the notoriously hard to compute stable homotopy groups of spheres. However, for every prime p, the homology with $\mathbb{Z}/p\mathbb{Z}$ coefficients has been computed for $\Omega^\infty S^\infty$, and hence for Σ_∞, see [AK56] and [DL62].

3. Moduli spaces and their (co)homology

We now switch from the algebraic to a more geometric point of view. The groups generally have now a natural topology and we need to distinguish between the homology of the group G as a topological space and that of BG. So the group (co)homology of G will always be thought of and written as the (co)homology of

[3] This is because the twisting for the mapping class group, as explained in the previous footnote, does not square to the identity; it is only a braiding.

the space BG which, as for discrete groups, is the quotient EG/G of a contractible space EG with a continuous, free G action.

3.1. Moduli spaces and characteristic classes. Our interest in groups and their (co)homology comes most often from an interpretation of the group G as the automorphism group $\mathrm{Aut}(W)$ of some geometric object W. We like the model for $B\mathrm{Aut}(W)$ to be a (topological) moduli space in the sense that

(i) the points in $B\mathrm{Aut}(W)$ are representing objects isomorphic to W, and

(ii) any family of objects isomorphic to W and indexed by a space X corresponds to a continuous map $f : X \to B\mathrm{Aut}(W)$.

Such a family E_f is called a *W-bundle* over X. The simplest example of such a bundle is the Cartesian product $W \times X$ which corresponds to the trivial map that sends every point in X to the point in $B\mathrm{Aut}(W)$ that represents W. If f is the trivial map, then so is the map f^* in cohomology. More generally, however, the elements in $f^*(H^*(BAut(W)))$, which we call the *characteristic classes* of E_f, will give information about how twisted the family E_f. It may be helpful to recall the well-known theory of characteristic classes for vector bundles.

Example: To be concrete, take $W = \mathbb{C}^n$. Then $\mathrm{Aut}(W) = \mathrm{GL}(n, \mathbb{C})$ is the group of invertible, linear maps. A good model for $B\mathrm{GL}(n, \mathbb{C})$ is the complex Grassmannian manifold $\mathrm{Gr}^{\mathbb{C}}(n, \infty)$ of n dimensional $\mathbb{C}$- linear subspaces of $\mathbb{C}^\infty := \bigcup_n \mathbb{C}^n$. This is a moduli space in the above sense. Its cohomology is well-known to be

$$H^*(B\mathrm{GL}(n, \mathbb{C})) = \mathbb{Z}[c_1, c_2, \dots, c_n]$$

where the c_i are the universal Chern classes of degree $2i$. For real vector spaces we have (the historically earlier) Pontryagin classes and Stiefel-Whitney classes.

These characteristic classes for vector bundles, which were discovered in the first half of the last century, have played a central role in the development of topology and geometry ever since. It is natural to ask,

What are the characteristic classes of bundles for more general W?

We might like to take W to be a compact manifold and its group of diffeomorphisms, or a finite simplicial complex and its group (up to homotopy) of homotopy equivalences. When W is a circle, every orientation preserving diffeomorphism is homotopic to a rotation which in turn is homotopic to $\mathrm{GL}(1, \mathbb{C})$. Thus the ring of characteristic classes for S^1-bundles is $\mathbb{Z}[c_1]$. With the proof of Mumford's conjecture and Galatius' theorem, we now also understand the characteristic classes, at least in the stable range, for W an oriented surface and W a simplicial complex of dimension one, as we will now explain.

3.2. Characteristic classes for manifold bundles. We take W to be an oriented, compact, smooth surface and its automorphism group to be the topological group $\mathrm{Diff}^+(W; \partial)$ of orientation preserving diffeomorphisms (which fix the boundary pointwise, if not empty). To construct a topological moduli space consider the space $\mathrm{Emb}(W, \mathbb{R}^\infty)$ of smooth embeddings of W in infinite dimensional Euclidean space. One may think of this as the space of embedded and parameterized surfaces of type W in $\mathbb{R}^\infty$. It is a consequence of Whitney's embedding theorem that this space is contractible. The diffeomorphism group $\mathrm{Diff}^+(W;\partial)$ acts freely on it by precomposing an embedding by a diffeomorphism. The quotient space is the (topological) moduli space for W

$$\mathcal{M}^{top}(W) := \mathrm{Emb}(W,\mathbb{R}^\infty)/\mathrm{Diff}^+(W;\partial) = B\mathrm{Diff}^+(W;\partial).$$

A point in it is a surface in $\mathbb{R}^\infty$. As we already mentioned, if $W = S_{g,1}$ or S_g and $g > 1$ then the diffeomorphism group is homotopic to the mapping class group and hence we have homotopy equivalences

$$\mathcal{M}^{top}(S_{g,1}) \simeq B\Gamma_{g,1} \quad \text{and} \quad \mathcal{M}^{top}(S_g) \simeq B\Gamma_g.$$

3.2.1. Relation to moduli spaces of Riemann surfaces. The homology of the spaces above are also of particular interest to algebraic geometers. The moduli space of Riemann surfaces $\mathcal{M}_g$ of genus g is the quotient of Teichmüller space, which is well-known to be homeomorphic to $\mathbb{R}^{6g-6}$, by the action of the mapping class group Γ_g. Mumford showed that it is a projective variety and moduli space for complex curves. The points of Teichmüller space are Riemann surfaces (with a homotopy class of a homeomorphism to a fixed surface S_g). As a Riemann surface can have at most finitely many automorphisms, the action of the mapping class group has at most finite stabilizers. It follows that rationally the (co)homology of $\mathcal{M}_g$ is the same as $B\Gamma_g$, and hence

$$H_*(\mathcal{M}_g) \otimes \mathbb{Q} \simeq H_*(\mathcal{M}_g^{top}) \otimes \mathbb{Q}.$$

In the early 1980's Mumford [Mu83] constructed characteristic classes κ_i for the $\mathcal{M}_g$ and initiated the systematic study of its cohomology ring. These classes were later also studied by Miller (and independently Morita) in the topological setting who showed that

$$H^*(B\Gamma_\infty) \otimes \mathbb{Q} \supset \mathbb{Q}[\kappa_1, \kappa_2, \dots].$$

The proof in [Mi86] uses Harer's homology stability as well as the commutativity of the product structure on the homology as describe above in Section 2.4. In the light of this, Mumford conjectured that the inclusion above is indeed an isomorphism. This is now a theorem by Madsen and Weiss [MW07]. Indeed, they prove a much stronger statement which was first conjectured in [MT01].

Theorem [MW07]. $\Omega B(\coprod_g BDiff^+(S_{g,1};\partial)) \simeq \Omega^\infty \mathbf{MTSO}(2)$.

Stringing several homotopy equivalences together, we see that the space on the left hand side has the homology of $\mathbb{Z} \times B\Gamma_\infty$ by the group completion theorem. We will now define the space on the right hand side and determine its rational cohomology.

3.2.2. Thom spaces and their rational cohomology. Let $\mathrm{Gr}^+(d,n)$ be the Grassmannian manifold of oriented $\mathbb{R}$-linear d planes in $\mathbb{R}^{d+n}$. There are two canonical vector bundles over $\mathrm{Gr}^+(d,n)$: the canonical d-bundle $\gamma_{d,n}$ with fibers over a plane $P \in \mathrm{Gr}^+(d,n)$ the vectors in P and its orthogonal complement $\gamma_{d,n}^{\perp}$. We will only use $\gamma_{d,n}^{\perp}$ and its one-point compactification $(\gamma_{d,n}^{\perp})^c$, also known as the Thom space of $\gamma_{d,n}^{\perp}$. Using the embeddings $\mathrm{Gr}^+(d,n) \to \mathrm{Gr}^+(d,n+1)$ we can form a limit space

$$\Omega^\infty \mathbf{MTSO}(d) := \lim_{n\to\infty} \Omega^{d+n}(\gamma_{d,n}^{\perp})^c;$$

here $\Omega^k X$ denotes the space of maps from S^k to X that take the point at infinity of $S^k = (\mathbb{R}^k)^c$ to the base point in X. The rational (co)homology of these spaces are well-understood and can be computed by standard methods in algebraic topology. For a connected component (they are all homotopic), we have

$$H^*(\Omega_0^\infty \mathbf{MTSO}(d)) \otimes \mathbb{Q} = \Lambda(H^{>d}(BSO(d))[-d] \otimes \mathbb{Q});$$

here $\Lambda(V^*)$ for a graded vector space V^* denotes the free graded commutative algebra on V^*. The V^* in question here is given by $V^n = H^{d+n}(BSO(d)) \otimes \mathbb{Q}$. As the rational classes of $H^*(BSO(d))$ are all even, this is just a polynomial algebra. Mumford's conjecture thus follows immediately.

Corollary. $H^*(B\Gamma_\infty) \otimes \mathbb{Q} = \mathbb{Q}[\kappa_1, \kappa_2, \dots]$.

Madsen and Weiss' theorem above however gives much more information. Thus one is able to determine the divisibility of the κ_i in the integral lattice in $H^*(B\Gamma_\infty)$. In [GMT06] we show that the maximal divisor of κ_{2i} is 2 and that of κ_{2i-1} is the denominator of the $B_i/2i$ where B_i is the ith Bernoulli number.

$\Omega^\infty \mathbf{MTSO}(2)$ has also a vast number of torsion classes, see [MT97], [G04]. Indeed, for every even integer there is an infinite family of torsion homology classes (each essentially a copy of $H_*(B\Sigma_\infty)$). These had not been detected before except for the first family, see [CL84].

3.3. Characteristic classes for graphs. In analogy to the above, Galatius [G] considers a moduli space $\mathcal{G}_n(\mathbb{R}^\infty)$ of embedded finite graphs in $\mathbb{R}^\infty$ that have fundamental group F_n for a fixed n. Its topology is such that the collapse of a (non-loop) edge can be achieved by a continuous path. Thus $\mathcal{G}_n(\mathbb{R}^\infty)$ is connected.[4] Similarly, one can define a based version $\mathcal{G}_n(\mathbb{R}^\infty;*)$ where each graph has a vertex at the origin. Using ideas from Igusa [I02] and Culler-Vogtmann's Outer space [CV86], Galatius proves

$$\mathcal{G}_n(\mathbb{R}^\infty;*) \simeq B\mathrm{Aut}F_n \quad \text{and} \quad \mathcal{G}_n(\mathbb{R}^\infty) \simeq B\mathrm{Out}F_n.$$

We will state now Galatius' theorem in analogue to Madsen and Weiss' theorem on the space level.

[4] The topology is somewhat delicate. In particular one wants the graph to be imbedded in such a way that every point in $\mathbb{R}^\infty$ has a neighborhood which either does not intersect the graph, or intersects it in a small interval (part of an edge), or contains a neighborhood of a vertex (and no more). Collapsing an edge has to be done in such a way that this is always satisfied.

Theorem [G]. $\Omega B(\coprod_{n\geq 0} BAutF_n) \simeq \Omega^\infty S^\infty$.

Previous evidence for this came in the form of a remark by Hatcher in [H95] who proved that $\Omega^\infty S^\infty$ is a direct factor of the left hand side. Hatcher already raises the question whether it could be an equivalence and in particular whether the rational homology of $\mathrm{Aut}F_\infty$ is trivial. Hatcher and Vogtmann in [HV98] prove that $H_*(\mathrm{Aut}F_n) \otimes \mathbb{Q} = 0$ for $0 < * < 7$ with the exception of $H_4(\mathrm{Aut}F_4) \otimes \mathbb{Q} = \mathbb{Q}$ providing strong evidence for the cohomological conjecture. Further evidence for the conjecture on the space level is given by a theorem by Igusa [I02]. It says that the linearisation map $L : \mathrm{Aut}F_n \to \mathrm{GL}(n, \mathbb{Z})$ on classifying spaces and after group completion (i.e. application of ΩB) factors through $\Omega^\infty S^\infty$.

With the Barratt-Priddy-Quillen theorem recalled in 2.4, we see that $\Sigma_\infty \hookrightarrow \mathrm{Aut}F_\infty$ induces an isomorphism in homology and in particular

Corollary. $H_*(BAutF_\infty) \otimes \mathbb{Q} = \mathbb{Q}$ *concentrated in degree 0.*

4. Towards a proof

Our very rough sketch here treats simultaneously the Barratt-Priddy-Quillen theorem, the Madsen-Weiss theorem as well as Galatius' theorem. We mainly follow [G] (and in parts the conceptually closely related [GMTW09]). In particular, we emphasize the role of the scanning map.

4.1. Scanning. In abstract terms, the scanning map can be applied to topological moduli spaces where the points are representing an object W embedded in $\mathbb{R}^N$, for N large. The idea is that the scanning map at the point $W \subset \mathbb{R}^N$ records the local, microscopic picture as a magnifying glass sweeps through $\mathbb{R}^N$.

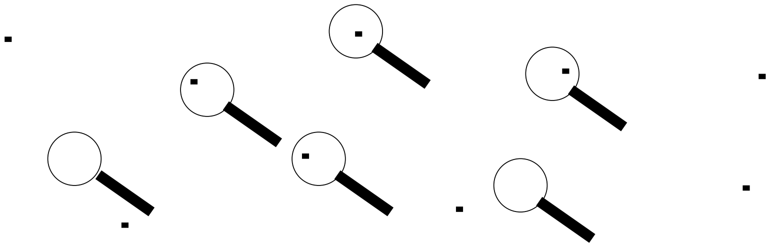


Figure 4: *A configuration of points in $\mathbb{R}^N$ sampled by a magnifying glass.*

We first consider the simplest and well known case when W is a set of n points. The ideas here go back to Segal and McDuff. The associated moduli space is the configuration space $\mathcal{C}_n(\mathbb{R}^N)$ of n distinct, unordered points in $\mathbb{R}^N$. When scanning, the lens of the magnifying glass can be taken small enough such that it only sees at most one point. Identifying the lens with a ball B^N in $\mathbb{R}^N$ we see that the space of all local pictures is just the sphere $S^N = (B^N)^c$ where the point at infinity corresponds to the lens giving us the view of the empty set. As the lens moves across $\mathbb{R}^N$ for a given configuration, we thus get a map from $\mathbb{R}^N$ to S^N. But away from a compact set containing the n points the lens sees nothing and the map is constant. We may therefore extend the map to $S^N = (\mathbb{R}^N)^c$ by sending the point at infinity to the empty set. Thus scanning defines a map

$$\bigsqcup_n \mathcal{C}_n(\mathbb{R}^N) \longrightarrow \Omega^N S^N.$$

Similarly, scanning can be applied to the moduli space $\mathcal{M}^{top}(S_g)^N$ of surfaces of type S_g embedded in $\mathbb{R}^N$ This time, unless the lens sees nothing, it will see an oriented 2-plane intersecting B^N, which is the tangent plane T_x of the nearest point x on the surface to the center of the lens. Identifying B^N with $\mathbb{R}^N$, this defines a 2-dimensional subspace $T_x - x$ of $\mathbb{R}^N$ and a vector x perpendicular to it. Thus we see that the Thom space $(\gamma_{2,N-2}^{\perp})^c$ is the space of all local data with the point at infinity corresponding again to the empty lens. Thus for every point in $\mathbb{R}^N$, we get a point in $(\gamma_{2,N-2}^{\perp})^c$, and again we can extend this map continuously to the compactification $S^N = (\mathbb{R}^N)^c$. Thus scanning defines a map

$$\bigsqcup_g \mathcal{M}^{top}(S_g)^N \longrightarrow \Omega^N(\gamma_{2,N-2}^{\perp})^c.$$

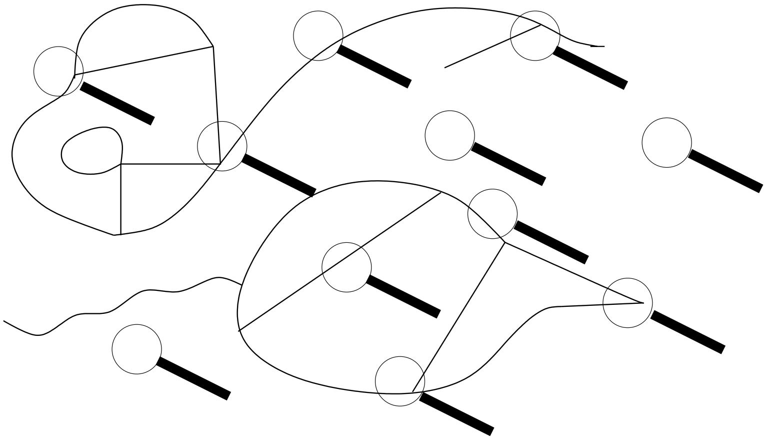


Figure 5: *A graph in $\mathbb{R}^N$ sampled by a magnifying glass.*

The case of graphs is similar, only that the space of all local data is much harder to identify. Indeed, Galatius spends considerable effort to show that a map from the N-sphere S^N to the space of local data in dimension N induces an isomorphism on homotopy groups in degrees $2N - c$ for some constant c. Thus in the limit as

$N \to \infty$ scanning defines (up to homotopy[5]) a map

$$\coprod_n \mathcal{G}_n(\mathbb{R}^\infty) \to \Omega^\infty S^\infty.$$

4.2. Spaces of manifolds and graphs. The above maps can of course not be homotopy equivalences. Not all path-components in the target are hit and the maps are not injective on fundamental groups. We need to enlarge our moduli spaces. Instead of considering only compact objects in $\mathbb{R}^N$ consider manifolds and graphs in $\mathbb{R}^N$ that may also be non-compact. So let $\Phi^{N,N}$ be the space of all manifolds W of a given dimension d or graphs G in $\mathbb{R}^N$. Each W respectively G has to be a closed subset of $\mathbb{R}^N$ but may extend to infinity. It also need not be connected. $\Phi^{N,N}$ is topologized in such a way that manifolds or graphs can be pushed continuously to infinity. The empty set is the basepoint. We have a filtration

$$\Phi^{N,N} \supset \cdots \supset \Phi^{N,1} \supset \Phi^{N,0} \simeq \coprod_n \mathcal{G}_n(\mathbb{R}^N),$$

where $\Phi^{N,i}$ contains only those W or G that are subsets of $\mathbb{R}^i \times (0,1)^{N-i}$. In particular, they are compact in the last $N-i$ coordinate directions.

To prove the theorems by Barratt-Priddy-Quillen, Madsen-Weiss or Galatius one would like to complete three steps. For configuration spaces this is an argument that essentially goes back to Segal [S73].

Step 1. $\Phi^{N,N}$ *is homotopic to the space of local data of the scanning map.*

This is relatively easy. The key here is that the topology of $\Phi^{N,N}$ allows us to push radially away from the origin. At the end of the homotopy what is left is the local data (at the origin).

Step 2. $\Phi^{N,k} \to \Omega\Phi^{N,k+1}$ *is a homotopy equivalence for* $k > 0$.[6]

We can construct a map as follows. For each t one can define a map $\Phi^{N,k} \to \Phi^{N,k+1}$ by sending W to its translate $W - te_{k+1}$ by t in the $(k+1)$st direction. So as t goes to infinity, W gets pushed out of sight and we can extend the map to $S^1 = \mathbb{R}^c$ by sending the point at infinity to the empty set. The case for graphs is just the same.

In the case of configuration spaces, it is straight forward to prove that this is a homotopy equivalence. But some extra argument is required in the case of higher dimensional manifolds and graphs.

Step 3. $\lim_{N\to\infty} \Phi^{N,1}$ *is homotopic to the classifying space of* $\coprod_n BAutW_n$.

Just as Step 2, this is straight forward for configuration spaces. More generally it is not difficult to see that in the manifold case, $\lim_{N\to\infty} \Phi^{N,1}$ is the classifying space

[5] More precisely, the target of the map is really something weakly homotopic to $\Omega^\infty S^\infty$.

[6] This statement is equivalent to saying that $\Phi^{N,k+1}$ is homotopic to $B\Phi^{N,k}$ and the connected components of $\Phi^{N,k}$ form a group. The product in $\Phi^{N,k}$ can be defined as follows: given two graphs W_1 and W_2, move W_2 to its translate $W_2 + e_{k+1}$. The resulting manifolds are disjoint and one can take the disjoint union. Using a homotopy $[0,2] \simeq [0,1]$ one can move the manifold back into $\Phi^{N,k}$. To show that $\Phi^{N,k+1} \simeq B\Phi^{N,k}$ is not too difficult. To show that the connected components from a group requres some argument. Note that it is here that we require the condition on k as the connected components certainly do not form a group when $k = 0$.

of the d-dimensional cobordism category as studied in topological field theory (see for example [GMTW09]; indeed this reproves the Main Theorem in that paper). Similarly, in the graph case, one sees that $\lim_{N\to\infty} \Phi^{N,1}$ is the classifying space of a cobordism category of graphs.

Though interesting in themselves, these identifications of $\lim_{N\to\infty} \Phi^{N,1}$ do not yet allow one to make deductions for $B\mathrm{Aut}W_\infty$ or $B\mathrm{Aut}W_n$. In order to do so, one wants to apply a group completion theorem following the arguments in [T97]. Two things are needed in order to apply it. First one needs to show that the classifying space of the cobordism category is homotopic to that of its subcategory in which the cobordisms are such that each component has non-empty out-going boundary. This is done in [GMTW09] for manifolds (of all dimensions ≥ 2) and in [G] for graphs. Secondly, one needs homology stability – which of course we have for graphs and surfaces.

We note here, that in [GRW] a proof of Madsen and Weiss' theorem in the form stated above is given that no longer uses homology stability. Indeed, Galatius and Randal-Williams show that the inclusion of the monoid $\coprod_n B\mathrm{Aut}W_n$ to the whole category induces a homotopy equivalence on classifying spaces and after group completion (i.e. applying ΩB).

5. Survey of further results

Madsen-Weiss' theorem as well as Galatius' theorem have been generalized in several directions. It is convenient to summarize some of these results in a table.

Σ_n	$(n+1)/2$	$\mathrm{Diff}(n \text{ pts})$	$\Omega^\infty S^\infty$
$\Gamma_{g,1}$	$(2g)/3$	$\mathrm{Diff}^+(S_{g,1};\partial)$	$\Omega^\infty \mathbf{MTSO}(2)$
$\mathcal{N}_{g,1}$	$(n-3)/3$	$\mathrm{Diff}(N_{g,1};\partial)$	$\Omega^\infty \mathbf{MTO}(2)$
.	.	$\mathrm{Diff}^+(\#_n S^k \times S^k \setminus \mathring{B}^{2k};\partial)$	$\Omega^\infty \mathbf{MTSO}(2k)^{\langle k\rangle}$
$\mathrm{Aut}F_n$	$(n-1)/2$	$\mathrm{HtEq}(G_n;*)$	$\Omega^\infty S^\infty$
$\mathcal{H}_{g,1}$	$(g-1)/2$	$\mathrm{Diff}^+(\#_g S^1 \times D^2; D^2 \subset \partial)$	$\Omega^\infty S^\infty BSO(3)_+$
.	.	$\mathrm{Diff}^+(\#_n S^1 \times S^2 \setminus \mathring{B}^3;\partial)$	$\Omega^\infty S^\infty BSO(4)_+$

The first column of the table gives the discrete group G_n to be considered; the second column lists the integer k so that the map $G_n \to G_{n+1}$ induces a homology isomorphisms in degrees less than k; the third column gives the automorphism group $\mathrm{Aut}W_n$ of the underlying geometric object; and finally, the fourth column contains a space homotopic to the group completion $\Omega B(\coprod_{n\geq 0} B\mathrm{Aut}W_n)$ and with the same homology as $\mathbb{Z} \times B\mathrm{Aut}W_\infty$.

We discuss now briefly the new entries .

5.1. Non-orientable surfaces. Let $N_{g,1}$ be an non-orientable surface of genus g (i.e. a connected sum of g copies of $\mathbb{R}P^2$s) with a boundary component. The group $\mathrm{Diff}(N_{g,1};\partial)$ has the same homotopy type as the associated mapping class group $\mathcal{N}_g$ by [EE69]. As for oriented surfaces, extending diffeomorphisms by the identity map over a glued on pair of pants surface induces a product $\mathcal{N}_{g,1} \times \mathcal{N}_{h,1} \to \mathcal{N}_{g+h,1}$. Wahl [W08] showed that $\mathcal{N}_{g,1}$ satisfies homology stability. Randal-Williams [RW] improved her range to $(n-3)/3$. (In this paper he also proves homology stability for surfaces with more exotic tangential structures (such as framed, spin and pin) which we have not listed in the above table.) The space $\Omega^\infty \mathbf{MTO}(2)$ is constructed just as in the oriented case only that the Grassmannian of non-oriented planes is considered. This gives the stable homology of the non-oriented mapping class group as

$$H_*(\mathcal{N}_\infty)\otimes\mathbb{Q}=\mathbb{Q}[\xi_i] \quad \text{with } \deg \xi_i = 4i.$$

5.2. A $(k-1)$-connected $2k$-manifold. The connected sum $\#_n S^k\times S^k$ of n copies of $S^k\times S^k$ is a higher dimensional analogue of a surface $S_g=\#S^1\times S^1$. We cut an open ball $\overset{\circ}{B}{}^{2k}$ out of the manifold and demand that diffeomorphisms fix the boundary. Then a product can be constructed by gluing as for surfaces. At the moment we do not know whether the classifying spaces of the diffeomorphism groups satisfy homology stability. Nor do we know whether the connected components are contractible. Nevertheless, for $k>2$ Galatius and Randal-Williams have identified the group completion of the associated monoid with a slight modification of the space $\Omega^\infty\mathbf{MTSO}(2k)$. To define this space, consider the k-connected cover $\pi:\mathrm{Gr}^+(2k,n)\langle k\rangle\to\mathrm{Gr}^+(2k,n)$, and instead of the universal bundle $\gamma^\perp_{2k,n}$ use its pulled back. (For a space X, its 1-connected cover $X\langle 1\rangle\to X$ is just its universal cover. The k-connected cover $X\langle k\rangle\to X$ is just a generalization of this in that $X\langle k\rangle$ has trivial homotopy groups for $*\leq k$ and the same homotopy groups as X for $*>k$.) The space we are looking for is $\Omega^\infty\mathbf{MTSO}(2k)^{\langle k\rangle}:=\lim_{n\to\infty}\Omega^{2k+n}(\pi^*(\gamma^\perp_{2k,n}))^c$.

5.3. Handlebody in dimension 3. Diffeomorphism of the 3-dimensional handlebody $\#_g S^1\times D^2$ of genus g restrict to diffeomorphisms of the boundary surface. Furthermore, its connected components are contractible. This is still the case when we fix a disk $D^2\subset\partial$ on the boundary. Thus its mapping class group $\mathcal{H}_{g,1}$ may be identified with a subgroup of $\Gamma_{g,1}$. Hatcher and Wahl [HW] proved that these groups satisfy homology stability. Very recently, Hatcher showed that the group completion in this case is

$$\Omega^\infty S^\infty BSO(3)_+ = \lim_{k\to\infty}\Omega^k S^k(BSO(3)_+)$$

by thinking of the handlebody as a thickened graph and adopting Galatius' proof. The proof uses another ingredient, the Smale conjecture (see [H83]) which states that $\mathrm{Diff}(D^3)\simeq \mathrm{O}(3)$. Here X_+ denotes X with a disjoint base point. In particular, this gives the stable homology of the handlebody mapping class group as

$$H_*(\mathcal{H}_\infty)\otimes\mathbb{Q}=\mathbb{Q}[\kappa_{2i}] \quad \text{with } \deg\kappa_{2i}=4i.$$

5.4. A simple 3-dimensional manifold. Finally, the bottom line is also work by Hatcher, recently announced. It concerns the connected sum of n copies of $S^1\times S^2$ with a ball removed so that a product can be defined. Again Hatcher uses a modification of Galatius' argument for graphs and the Smale conjecture. Note however, that in this case the connected components of the diffeomorphism groups are not contractible and we do not know whether the classifying spaces of these groups satisfy homology stability.

6. Conclusion and future directions

We have seen that the method of scanning can be applied to topological moduli spaces $\mathcal{M}^{top}(W)$ of objects isomorphic to W embedded in $\mathbb{R}^\infty$. The target T of the scanning map is a highly structured space, an infinite loop space. In the case of zero dimensional manifolds, graphs and two dimensional manifolds the target of the scanning map T, in the presence of homology stability gives a better and better approximation to the homology of the moduli space $\mathcal{M}^{top}(W)$ as the complexity of W grows. In the previous section we encountered manifolds W of higher dimensions for which the scanning map induces a homotopy equivalence from the group completion of the associated monoid to the target T but for which we do not (yet) have homology stability.

One is naturally led to ask the following questions. Can the table above be completed and homology stability results be found for certain types of manifolds? Are there any other families of manifolds that can be added to the table? Indeed, are there other geometric objects, to which the scanning method can be applied. Galatius considered finite 1-dimensional complexes. Can the methods be pushed to higher dimensional finite complexes?

We emphasized the point of view of moduli spaces and characteristic classes for manifold bundles. Indeed, for every oriented, closed, d-dimensional manifolds W, scanning gives a map $\alpha : \mathcal{M}^{top}(W) \to \Omega^\infty\mathbf{MTSO}(d)$. So the cohomology of $\Omega^\infty\mathbf{MTSO}(d)$ provides characteristic classes for all oriented d-manifolds simultaneously. Ebert has shown that for d even every rational cohomology class c is detected by some manifold, i.e. $\alpha^*(c)$ is non-zero for some W. But this fails for d odd, see [E1] and [E2]. This suggests that the scanning map for d odd is not optimal and should factor through a space $X(d)$. For $d = 1$, $X(1) = \Omega^\infty S^\infty BSO(2)_+$ is the optimal space. Hatcher's last example suggests a similar solution for $d = 3$.

We have tried to give here a glimpse into an active area of research that uses new techniques to study basic questions in geometry and topology. Some of the questions above are already pursued, and we look forward to seeing the theory develop.

Acknowledgements

I am grateful to Søren Galatius, Allen Hatcher, and Oscar Randal-Williams for their comments on an earlier draft of this paper.

References

[AK56] S. Araki, T. Kudo, *Topology of H_n-spaces and H_n-squaring operations*, Mem. Fac. Sci. Kyushu Univ. Ser A **10** (1956), 85–120.

[BP72] M. Barratt, S. Priddy, *On the homology of non-connected monoids and their associated groups*, Comment. Math. Helv. **47** (1972), 1–14.

[BW94] J. Birman, B. Wajnryb, *Presentations of the mapping class group. Errata*, Israel J. Math. **88** (1994), no. 1-3, 425–427.

[B] S.K. Boldsen, *Improved homological stability for the mapping class group with integral or twisted coefficients*, arXiv:0904.3269

[CL87] R. Charney, R. Lee, *An application of homotopy theory to mapping class groups*, J. Pure Appl. Algebra **44** (1987), no. 1-3, 127–135.

[CV86] M. Culler, K. Vogtmann, *Moduli of graphs and automorphisms of free groups*, Invent. Math.**84** (1986), no. 1, 91–119.

[EE69] Clifford J. Earle and James Eells, *A fibre bundle description of Teichmüller theory*, J. Differential Geometry **3** (1969), 19–43.

[DL62] E. Dyer, R. Lashof, *Homology of iterated loop spaces.*, Amer. J. Math. **84** (1962), 35–88.

[E1] J. Ebert, *A vanishing theorem for characteristic classes of odd-dimensional manifold bundles*, arXiv:0902.4719

[E2] J. Ebert, *Algebraic independence of generalized Morita-Miller-Mumford classes*, arXiv:0910.1030

[G04] S. Galatius, *Mod p homology of the stable mapping class group*, Topology **43** (2004), no. 5, 1105–1132.

[GMT06] S. Galatius, I. Madsen, U. Tillmann, *Divisibility of the stable Miller-Morita-Mumford classes*, J. Amer. Math. Soc. **19** (2006), no. 4, 759–779.

[G] S. Galatius, *Stable homology of automorphism groups of free groups*, to appear in Ann. of Math., math/0610216

[GMTW09] S. Galatius, I. Madsen, U. Tillmann, and M. Weiss, *The homotopy type of the cobordism category*, Acta Math. **202** (2009), no. 2, 195–239.

[GRW10] S. Galatuis and O. Randal-Williams, *Monoids of moduli spaces of manifolds*, Geom. Topol. 14 (2010), no. 3, 1243–1302.

[Ge84] S.M. Gersten, *A presentation for the special automorphism group of a free group*, J. Pure Appl. Algebra **33** (1984), no. 3, 269–279.

[Ha85] J.L. Harer, *Stability of the homology of the mapping class groups of orientable surfaces*, Ann. of Math. (2) **121** (1985), no. 2, 215–249.

[Ha86] J.L. Harer, *The virtual cohomological dimension of the mapping class group of an orientable surface*, Invent. Math. **84** (1986), no. 1, 157–176.

[H83] A. Hatcher, *A proof of the Smale conjecture, $Diff(S^3) \simeq SO(3)$*, Ann. of Math. **117** (1983), 553–607.

[H95] A. Hatcher, *Homological stability for automorphism groups of free groups*, Comment. Math. Helv. **70** (1995), no. 1, 39–62.

[HV98] A. Hatcher, K. Vogtmann, *Rational homology of $Aut(F_n)$*, Math. Res. Lett. **5** (1998), no. 6, 759–780.

[HV98.C] A. Hatcher, K. Vogtmann, *Cerf theory for graphs*, J. London Math Soc. **58** (1998), 633–655.

[HV04] A. Hatcher, K. Vogtmann, *Homology stability for outer automorphism groups of free groups*, Algebr. Geom. Topol. **4** (2004), 1253–1272.

[HVW06] A. Hatcher, K. Vogtmann, N. Wahl, *Erratum to: Homology stability for outer automorphism groups of free groups*, Algebr. Geom. Topol.**6** (2006), 573–579.

[HW] A. Hatcher, N. Wahl, *Stabilization for mapping class groups of 3-manifolds*, to appear in Duke Math Journal. arXiv:0709.2173

[I02] K. Igusa, Higher Franz-Reidemeister torsion. AMS/IP Studies in Advanced Mathematics, **31**. American Mathematical Society, Providence, RI; International Press, Somerville, MA, 2002. xxii+370 pp

[I89] N.V. Ivanov, *Stabilization of the homology of Teichmüller modular groups*, (Russian) Algebra i Analiz 1 (1989), no. **3**, 110–126; translation in Leningrad Math. J. 1 (1990), no. **3**, 675–691.

[MT01] I. Madsen and U. Tillmann, *The stable mapping class group and* $Q(CP^\infty_+)$, Invent. Math. **145** (2001), no. 3, 509–544.

[MW07] I. Madsen and M. Weiss, *The stable moduli space of Riemann surfaces: Mumford's conjecture*, Ann. of Math. **165** (2007), 843–941.

[Mi86] E.Y. Miller, *The homology of the mapping class group*, J. Differential Geom. **24** (1986), no. 1, 1–14.

[Mu83] D. Mumford, *Towards an enumerative geometry of the moduli space of curves*, Arithmetic and geometry, Vol. II, 271–328, Progr. Math., **36**, Birkhuser Boston, Boston, MA, 1983.

[N60] M.Nakaoka, *Decomposition theorem for homology groups of symmetric groups*, Ann. of Math. **2** 71 (1960) 16–42.

[Q] E. Friedlander, B. Mazur, Filtrations on the homology of algebraic varieties. With an appendix by Daniel Quillen. Mem. Amer. Math. Soc. **110** (1994), no. 529, x+110 pp.

[RW] O. Randal-Williams, *Resolutions of moduli spaces and homological stability*, arXiv:0909.4278

[S73] G. Segal, *Configuration-spaces and iterated loop-spaces*, Invent. Math. **21** (1973), 213–221.

[T97] U. Tillmann, *On the homotopy of the stable mapping class group*, Invent. Math. **130** (1997), no. 2, 257–275.

[W08] N. Wahl, *Homological stability for the mapping class groups of non-orientable surfaces*, Invent. Math. **171** (2008), no. 2, 389–424.

[W83] B. Wajnryb, *A simple presentation for the mapping class group of an orientable surface*, Israel J. Math. **45** (1983), no. 2-3, 157–174.

Mathematical Institute
Oxford University
24-29 St Giles St.
Oxford OX1 3LB, UK

tillmann@maths.ox.ac.uk

THE GEOMETRIC NATURE OF THE FUNDAMENTAL LEMMA

DAVID NADLER

ABSTRACT. The Fundamental Lemma is a somewhat obscure combinatorial identity introduced by Robert P. Langlands [L79] as an ingredient in the theory of automorphic representations. After many years of deep contributions by mathematicians working in representation theory, number theory, algebraic geometry, and algebraic topology, a proof of the Fundamental Lemma was recently completed by Ngô Bao Châu [N08], for which he was awarded a Fields Medal. Our aim here is to touch on some of the beautiful ideas contributing to the Fundamental Lemma and its proof. We highlight the geometric nature of the problem which allows one to attack a question in p-adic analysis with the tools of algebraic geometry.

CONTENTS

1. INTRODUCTION

Introduced by Robert P. Langlands in his lectures [L79], the Fundamental Lemma is a combinatorial identity which just as well could have achieved no notoriety. Here is Langlands commenting on [L79] on the IAS website [L1]:

> "...the fundamental lemma which is introduced in these notes, is a precise and purely combinatorial statement that I thought must therefore of necessity yield to a straightforward analysis. This has turned out differently than I foresaw."

Instead, the Fundamental Lemma has taken on a life of its own. Its original scope involves distributions on groups over local fields (p-adic and real Lie groups). Such distributions naturally arise as the characters of representations, and are more than worthy of study on their own merit. But with the immense impact of Langlands' theory of automorphic and Galois representations, many potential advances turn out to be downstream of the Fundamental Lemma. In particular, in the

Received by the editors October 12, 2010.

2010 *Mathematics Subject Classification.* Primary 20G35.

absence of proof, it became "the bottleneck limiting progress on a host of arithmetic questions." [H] It is rare that popular culture recognizes the significance of a mathematical result, much less an esoteric lemma, but the recent proof of the Fundamental Lemma by Ngô Bao Châu [N08], for which he was awarded a Fields Medal, ranked seventh on *Time* magazine's *Top 10 Scientific Discoveries of 2009* list.[1]

Before continuing, it might be useful to have in mind a cartoon of the problem which the Fundamental Lemma solves. In fact, what we present is an example of the case of real Lie groups resolved long ago by D. Shelstad [S82]. Figure 1 depicts representative orbits for the real Lie group $SL(2,\mathbb{R})$ acting by conjugation on its Lie algebra $\mathfrak{sl}(2,\mathbb{R}) \simeq \mathbb{R}^3$ of traceless 2×2 real matrices A.

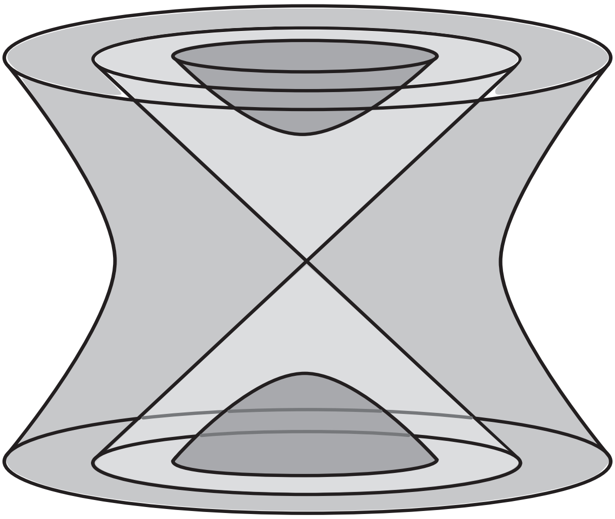


FIGURE 1. Orbits of $SL(2,\mathbb{R})$ acting on its Lie algebra $\mathfrak{sl}(2,\mathbb{R}) \simeq \mathbb{R}^3$.

Reading Figure 1 from outside to inside, one encounters three types of orbits (hyperbolic, nilpotent, and elliptic) classified by the respective values of the determinant ($\det(A) < 0$, $\det(A) = 0$, and $\det(A) > 0$). We will focus on the two elliptic orbits $\mathcal{O}_{A_+}, \mathcal{O}_{A_-} \subset \mathfrak{sl}(2,\mathbb{R})$ through the elements

$$A_+ = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \qquad A_- = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

For a smooth compactly supported function $\varphi : \mathfrak{sl}(2,\mathbb{R}) \to \mathbb{C}$, consider the distributions given by integrating over the elliptic orbits

$$\mathcal{O}_{A_+}(\varphi) = \int_{\mathcal{O}_+} \varphi, \qquad \mathcal{O}_{A_-}(\varphi) = \int_{\mathcal{O}_-} \varphi$$

with respect to an invariant measure.

Observe that the two-dimensional complex vector space spanned by these distributions admits the alternative basis

$$\mathcal{O}_{st} = \mathcal{O}_{A_+} + \mathcal{O}_{A_-}, \qquad \mathcal{O}_{tw} = \mathcal{O}_{A_+} - \mathcal{O}_{A_-}.$$

[1] Water on the moon was sixth, and Teleportation was eighth.

The first $\mathcal{O}_{st}$ is nothing more than the integral over the union $\mathcal{O}_{A_+} \sqcup \mathcal{O}_{A_-}$, which is the algebraic variety given by the equation $\det(A) = 1$. It is called a *stable distribution* since the equation $\det(A) = 1$ makes no reference to the field of real numbers $\mathbb{R}$. Over the algebraic closure of complex numbers $\mathbb{C}$, the equation $\det(A) = 1$ cuts out a single conjugacy class. In particular, A_+ and A_- are both conjugate to the matrix

$$\begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} \in \mathfrak{sl}(2, \mathbb{C}).$$

Thus the stable distribution can be thought of as an object of algebraic geometry rather than harmonic analysis on a real Lie algebra.

Unfortunately, there is no obvious geometric interpretation for the second $\mathcal{O}_{tw}$. (And one might wonder whether such a geometric interpretation could exist: the symmetry of switching the terms of $\mathcal{O}_{tw}$ gives its negative.) It is called a *twisted distribution* since it is a sum of $\mathcal{O}_{A_+}$ and $\mathcal{O}_{A_-}$ with nonconstant coefficients. By its very definition, $\mathcal{O}_{tw}$ distinguishes between the orbits $\mathcal{O}_{A_+}$ and $\mathcal{O}_{A_-}$ though there is no invariant polynomial function separating them. Indeed, as discussed above, over the complex numbers, they coalesce into a single orbit.

Langlands's theory of endoscopy, and in particular, the Fundamental Lemma at its heart, confirms that indeed one can systematically express such twisted distributions in terms of stable distributions. A hint more precisely: to any twisted distribution, there is assigned a stable distribution, and to any test function, a transferred test function, such that the twisted distribution evaluated on the original test function is equal to the stable distribution evaluated on the transferred test function. Detailed conjectures organizing the intricacies of the transfer of test functions first appear in Langlands's joint work with D. Shelstad [LS87]. The shape of the conjectures for p-adic groups, and in particular the until recently missing Fundamental Lemma in the non-Archimedean case, were decisively influenced by what could be more readily understood for real groups (ultimately building on work of Harish Chandra). As Langlands and Shelstad note, "if it were not that [transfer factors] had been proved to exist over the real field [S82], it would have been difficult to maintain confidence in the possibility of transfer or in the usefulness of endoscopy."

The extraordinary difficulty of the Fundamental Lemma, and also its mystical power, emanates from the fact that the sought-after stable distributions live on the Lie algebras of groups with little apparent relation to the original group. Applied to the example at hand, the general theory relates the twisted distribution $\mathcal{O}_{tw}$ to a stable distribution on the Lie algebra $\mathfrak{so}(2, \mathbb{R}) \simeq \mathbb{R}$ of the rotation subgroup $SO(2, \mathbb{R}) \subset SL(2, \mathbb{R})$ which stabilizes A_+ or equivalently A_-. Outside of bookkeeping, this is empty of content since $SO(2, \mathbb{R})$ is abelian, and so its orbits in $\mathfrak{so}(2, \mathbb{R})$ are single points. But the general theory is deep and elaborate and leads to surprising identities of which the Fundamental Lemma is the most basic and important.

It should be pointed out that in the absence of a general proof, many special cases of the Fundamental Lemma were established to spectacular effect. To name the most prominent applications without attempting to explain any of the terms involved, the proof of Fermat's Last Theorem due to Wiles and Taylor-Wiles depends upon base change for $GL(2)$, and ultimately the Fundamental Lemma for cyclic base change [L80]. The proof of the local Langlands conjecture for $GL(n)$, a parameterization of the representations of the group of invertible matrices with p-adic entries, due to Harris-Taylor and independently Henniart depends upon automorphic induction, and ultimately the Fundamental Lemma for $SL(n)$ established by Waldspurger [W91].

If the Fundamental Lemma had admitted an easy proof, it would have merited little mention in a discussion of these results. But for the general theory of automorphic representations and Shimura varieties, "...its absence rendered progress almost impossible for more than twenty years." [L2] Its recent proof has opened the door to many further advances. Arthur [A97, A05, A09] has outlined a program to obtain the local Langlands correspondence, for quasi-split classical groups from that of $GL(n)$ via twisted endoscopy (generalizing work of Rogawski [R90] on the unitary group $U(3)$). In particular, this provides a parameterization of the representations of orthogonal and symplectic matrix groups with p-adic entries. Furthermore, Arthur also describes how to reduce the automorphic spectrum of such groups to the cuspidal automorphic spectrum of $GL(n)$. Shin [S] has constructed Galois representations corresponding to automorphic representations in the cohomology of compact Shimura varieties, establishing the Ramanujan-Petersson conjecture for such representations. This completes earlier work of Harris-Taylor, Kottwitz and Clozel, and the Fundamental Lemma plays a key role in this advance. As Shin notes,

> "One of the most conspicuous obstacles was the fundamental lemma, which had only been known in some special cases. Thanks to the recent work of Laumon-Ngô ([LaN04]), Waldspurger ([W97], [W06], [W09]) and Ngô ([N08]) the fundamental lemma (and the transfer conjecture of Langlands and Shelstad) are now fully established. This opened up a possibility for our work."

Morel [M08] has obtained similar results via a comprehensive study of the cohomology of noncompact Shimura varieties.

Independently of its applications, the peculiar challenge of the Fundamental Lemma has spurred many ingenious advances of standalone interest. Its recent proof, completed by Ngô, spans many areas, appealing to remarkable new ideas in representation theory, model theory, algebraic geometry, and algebraic topology. A striking aspect of the story is its diverse settings. The motivation for and proof of the Fundamental Lemma sequentially cover the following algebraic terrain:

$$\boxed{\text{number fields} \to p\text{-adic fields} \to \text{Laurent series fields} \to \text{function fields}}$$

One begins with a number field and the arithmetic problem of comparing the anisotropic part of the Arthur-Selberg trace formula for different groups. This leads to the combinatorial question of the Fundamental Lemma about integrals over p-adic groups. Now not only do the integrals make sense for any local field, but it turns out that they are independent of the specific local field, and in particular its characteristic. Thus one can work in the geometric setting of Laurent series with integrals over loop groups (or synonymously, affine Kac-Moody groups). Finally, one returns to a global setting and performs analogous integrals along the Hitchin fibration for groups over the function field of a projective curve. In fact, one can interpret the ultimate results as precise geometric analogues of the stabilization of the original trace formula. To summarize, within the above algebraic terrain, the main quantities to be calculated and compared for different groups are the following:

global orbital integrals → local orbital integrals →
cohomology of affine Springer fibers → cohomology of Hitchin fibers

Over the last several decades, geometry has accrued a heavy debt to harmonic analysis and number theory: much of the representation theory of Lie groups and quantum algebras, as well as gauge theory on Riemann surfaces and complex surfaces, is best viewed through a collection of analogies (called the Geometric Langlands program, and pioneered by Beilinson and Drinfeld) with Langlands's theory of automorphic and Galois representations. Now with Ngô's proof of the Fundamental Lemma, and its essential use of loop groups and the Hitchin fibration, geometry has finally paid back some of this debt.

In retrospect, one is led to the question: *Why does geometry play a role in the Fundamental Lemma?* Part of the answer is implicit in the fact that the p-adic integrals involved turn out to be characteristic independent. They are truly of motivic origin, reflecting universal polynomial equations rather than analysis on the p-adic points of groups. Naively, one could think of the comparison between counting matrices over a finite field with a prescribed characteristic polynomial versus counting those with prescribed eigenvalues. More substantively, one could keep in mind Lusztig's remarkable construction of the characters of finite simple groups of Lie type (which encompass "almost all" finite simple groups). Though such groups have no given geometric structure, their characters can be uniformly constructed by recognizing the groups are the solutions to algebraic equations. For example, though we may care primarily about characters of $SL(n, \mathbb{F}_p)$, it is important to think not only about the set of such matrices, but also the determinant equation $\det(A) = 1$ which cuts them out.

In the case of the Fundamental Lemma, the Weil conjectures ultimately imply that the p-adic integrals to be evaluated are shadows of the cohomology of algebraic varieties, specifically the affine Springer fibers of Kazhdan-Lusztig. Therefore one could hope to apply the powerful tools of mid-to-late 20th century algebraic geometry – such as Hodge theory, Lefschetz techniques, sheaf theory, and homological algebra – in the tradition pioneered by Weil, Serre, Grothendieck, and Deligne. One of Ngô's crucial, and possibly indispensable, contributions is to recognize that the technical structure needed to proceed, in particular the purity of the Decomposition Theorem of Beilinson-Bernstein-Deligne-Gabber, could be found in a return to the global setting of the Hitchin fibration.

Our aim in what follows is to sketch some of the beautiful ideas contributing to the Fundamental Lemma and its proof. The target audience is not experts in any of the subjects discussed but rather mathematicians interested in having some sense of the essential seeds from which a deep and intricate theory flowers. We hope that in a subject with great cross-pollination of ideas, an often metaphorical account of important structures could prove useful.

Here is an outline.

In Section 2 immediately below, we recall some basics about characters of representations and number fields leading to a very rough account of the Arthur-Selberg trace formula for compact quotient.

In Section 3, we introduce the problem of stability of conjugacy classes, the twisted orbital integrals and endoscopic groups which arise, and finally arrive at a statement of the Fundamental Lemma.

In the remaining sections, we highlight some of the beautiful mathematics of broad appeal which either contribute to the proof of the Fundamental Lemma or were spurred by its study. Some were invented specifically to attack the Fundamental Lemma, while others have their own pre-history but are now inextricably linked to it.[2]

In Section 4, we introduce the affine Springer fibers and their cohomology which are the motivic avatars of the orbital integrals of the Fundamental Lemma. We then discuss the equivariant localization theory of Goresky-Kottwitz-MacPherson developed to attack the Fundamental Lemma. Strictly speaking, it is not needed for Ngô's ultimate proof, but it both set the scene for much of Laumon and Ngô's further successes, and has inspired an entire industry of combinatorial geometry.

In Section 5, we summarize and interpret several key aspects of Ngô's proof of the Fundamental Lemma. In particular, we discuss Laumon's initial forays to a global setting, and Ngô's Support Theorem which ultimately provides the main technical input.

Finally, in Section 6, we discuss some directions for further study.

There are many precise and readable, though of necessity lengthy, accounts of the mathematics we will discuss. In particular, we recommend the reader read everything by Langlands, Kottwitz, and Arthur, and time permitting, read all of Drinfeld and Laumon's lecture notes such as [D, La1, La2]. For the Fundamental Lemma and its immediate neighborhood, there are Ngô's original paper [N08], and the long list of references therein, Hales's beautifully concise paper [H05], along with DeBacker's timely update [De05], and the immensely useful book project organized by Harris [H].

1.1. **Acknowledgements.** This document is intended as a writeup of my anticipated talk at the Current Events Bulletin Session at the Joint Mathematics Meetings in New Orleans, January 2011. I would like to thank the committee and organizers for the impetus to prepare this document.

I am particularly indebted to D. Ben-Zvi for his generosity in sharing ideas during many far-ranging discussions.

I am grateful to D. Jordan for creating Figures 1 and 2 which beautifully evoke the mystery of endoscopy.

For their many insights, I would like to thank S. Gunningham, B. Hannigan-Daley, I. Le, Y. Sakallaridis, T. Schedler, M. Skirvin, X. Zhu, and all the participants in the 2010 Duntroon Workshop, especially its organizers J. Kamnitzer and C. Mautner.

I would also like to thank B. Conrad, M. Emerton, M. Harris, B. Kra, and M. Strauch for their detailed comments on earlier drafts.

Finally, I gratefully acknowledge the support of NSF grant DMS-0901114 and a Sloan Research Fellowship.

[2] Like Tang to NASA.

2. Characters and conjugacy classes

To begin to approach the Fundamental Lemma, let's listen once again to Langlands [L2]:

> "Nevertheless, it is not the fundamental lemma as such that is critical for the analytic theory of automorphic forms and for the arithmetic of Shimura varieties; it is the stabilized (or stable) trace formula, the reduction of the trace formula itself to the stable trace formula for a group and its endoscopic groups, and the stabilization of the Grothendieck-Lefschetz formula.

In this section, we will give a rough impression of the trace formula, and in the next section, explain what the term stable is all about.

2.1. **Warmup: finite groups.**

To get in the spirit, we begin our discussion with the well known character theory of a finite group G. There are many references for the material in this section, for example [FH91], [S77].

Definition 2.1. By a *representation* of G, we will mean a finite-dimensional complex vector space V and a group homomorphism $\pi : G \to GL(V)$.

Equivalently, we can form the group algebra $\mathbb{C}[G] = \{\varphi : G \to \mathbb{C}\}$ equipped with convolution

$$(\varphi_1 * \varphi_2)(g) = \textstyle\sum_{g_1 g_2 = g} \varphi_1(g_1)\varphi_2(g_2), \qquad \varphi_1, \varphi_2 \in \mathbb{C}[G],$$

and consider finite-dimensional $\mathbb{C}[G]$-modules.

Example 2.2. (1) Trivial representation: take $V = \mathbb{C}$ with the trivial action.

(2) Regular representation: take $V = \mathbb{C}[G]$ with the action of left-translation.

Definition 2.3. The *character* of a representation $\pi : G \to GL(V)$ is the function

$$\chi_\pi : G \to \mathbb{C} \qquad \chi_\pi(g) = \mathrm{Trace}(\pi(g))$$

Definition 2.4. Consider the action of G on itself by conjugation.

We denote the resulting quotient set by G/G and refer to it as the *adjoint quotient.* Its elements are conjugacy classes $[g] \subset G$.

A *class function* on G is a function $f : G/G \to \mathbb{C}$, or equivalently a conjugation-invariant function $f : G \to \mathbb{C}$. We denote the ring of class functions by $\mathbb{C}[G/G]$.

Lemma 2.5. *(1) Each character χ_π is a class function.*

(2) Compatibility with direct sums: $\chi_{\pi_1 \oplus \pi_2} = \chi_{\pi_1} + \chi_{\pi_2}$.

(3) Compatibility with tensor products: $\chi_{\pi_1 \otimes \pi_2} = \chi_{\pi_1}\chi_{\pi_2}$.

(4) Trivial representation: $\chi_{triv}(g) = 1$, for all g.

(5) Regular representation:

$$\chi_{reg}(g) = \begin{cases} |G|, & g = e \\ 0, & g \neq e \end{cases}$$

Class functions have a natural Hermitian inner product

$$\langle \alpha, \beta \rangle = \tfrac{1}{|G|} \textstyle\sum_{g \in G} \overline{\alpha(g)}\beta(g), \qquad \alpha, \beta \in \mathbb{C}[G/G].$$

Proposition 2.6. *The characters of irreducible representations of G form an orthonormal basis of the class functions $\mathbb{C}[G/G]$.*

Thus we have two canonical bases of class functions. On the one hand, there is the *geometric basis* of characteristic functions

$$\mathcal{O}_{[g]} : G/G \longrightarrow \mathbb{C} \qquad \mathcal{O}_{[g]}(h) = \begin{cases} 1, & h \in [g] \\ 0, & \text{else} \end{cases}$$

of conjugacy classes $[g] \subset G$. These are pairwise orthogonal though not orthonormal since $\langle \mathcal{O}_{[g]}, \mathcal{O}_{[g]} \rangle$ is the volume of the conjugacy class $[g] \subset G$. On the other hand, there is the *spectral basis* of characters

$$\chi_\iota : G/G \longrightarrow \mathbb{C} \qquad \chi_\iota(h) = \mathrm{Trace}(\pi_\iota(h))$$

of irreducible representations π_ι of G. The geometric basis is something one has on any finite set (though the volumes contain extra information). The spectral basis is a reflection of the group structure of the original set G.

Remark 2.7. One uses the term *spectral* with the following analogy in mind. Given a diagonalizable operator A on a complex vector space V, the traditional spectral decomposition of V into eigenspaces can be interpreted as the decomposition of V into irreducible modules for the algebra $\mathbb{C}[A]$.

Given an arbitrary representation (π, V), we can expand its character χ_π in the two natural bases to obtain an identity of class functions. Though V might be completely mysterious, it nevertheless admits an expansion into irreducible representations

$$V \simeq \bigoplus_{\iota \in I} V_\iota^{\oplus m_\iota(\pi)}$$

where I denotes the set of irreducible representations. Thus we obtain an identity of class functions

$$\sum_{[g] \in G/G} \mathrm{Trace}(\pi(g))\mathcal{O}_{[g]} = \sum_{\iota \in I} m_\iota(\pi)\chi_\iota$$

The left hand side is geometric and easy to calculate. The right hand side is spectral both mathematically speaking and in the sense that like a ghost we know it exists though we may not be able to see it clearly. The formula gives us a starting point to understand the right hand side in terms of the left hand side.

Remark 2.8. It is very useful to view the above character formula as an identity of distributions. Namely, given any test function $\varphi : G \to \mathbb{C}$, we can write $\varphi = \sum_{g \in G} \varphi(g)\delta_g$, where $\delta_g : G \to \mathbb{C}$ is the characteristic function of the group element g. Then since everything in sight is linear, we can evaluate the character formula on φ. This is very natural from the perspective of representations as modules over the group algebra $\mathbb{C}[G]$.

In general, it is difficult to construct representations. Outside of the trivial and regular representations, the only others that appear immediately from the group structure of G are induced representations.

Definition 2.9. Fix a subgroup $\Gamma \subset G$.

For a representation $\pi : \Gamma \to GL(W)$, the corresponding *induced representation* $\pi_{ind} : G \to GL(V)$ is defined by

$$V = \{f : G \to W | f(\gamma x) = \pi(\gamma) f(x)\} \qquad \pi_{ind}(g) f(x) = f(xg)$$

We denote by χ_{ind} the character of π_{ind}.

Calculating the character χ_{ind} is a particularly simple but salient calculation from Mackey theory. Since we have no other starting point, we will focus on the induction of the trivial representation. (Exercise: the induction of the regular representation of Γ is the regular representation of G.) When we start with the trivial representation, the induced representation

$$V = \{f : \Gamma\backslash G \to \mathbb{C}\}$$

is simply the vector space of functions on the finite set $\Gamma\backslash G$. It has a natural basis given by the characteristic functions of the cosets. Thus every element of G, or more generally, the group algebra $\mathbb{C}[G]$, acts on the vector space V by a matrix with an entry for each pair of cosets $x, y \in \Gamma\backslash G$.

Lemma 2.10. *An element $\varphi : G \to \mathbb{C}$ of the group algebra $\mathbb{C}[G]$ acts on the induced representation $V = \{f : \Gamma\backslash G \to \mathbb{C}\}$ by the matrix*

$$K_\varphi(x, y) = \sum_{\gamma\in\Gamma} \varphi(x^{-1}\gamma y), \qquad x, y \in \Gamma\backslash G.$$

Now to calculate the character χ_{ind}, we need only take the traces of the above matrices, or in other words, the sum of their entries when $x = y \in \Gamma\backslash G$.

Corollary 2.11. *The character χ_{ind} is given by the formula*

$$\chi_{ind}(\varphi) = \sum_{\gamma\in\Gamma/\Gamma} a_\gamma \int_{[\gamma]\subset G} \varphi,$$

where a_γ denotes the volume, or number of elements, of the quotient of centralizers $\Gamma_\gamma\backslash G_\gamma$, and the integral denotes the sum

$$\int_{[\gamma]\subset G} \varphi = \sum_{x\in G_\gamma\backslash G} \varphi(x^{-1}\gamma x)$$

over the G-conjugacy class of γ.

Remark 2.12. Suppose we equip the quotients G/G, Γ/Γ with the natural quotient measures, and let $p : \Gamma/\Gamma \to G/G$ denote the natural projection. Then the above corollary can be concisely rephrased that χ_{ind} is the pushforward along p of the quotient measure on Γ/Γ.

Definition 2.13. For $\gamma \in \Gamma$, the distribution on G given on a test function $\varphi : G \to \mathbb{C}$ by the integral over the conjugacy class

$$\mathcal{O}_\gamma(\varphi) = \int_{[\gamma]\subset G} \varphi = \sum_{x\in G_\gamma\backslash G} \varphi(x^{-1}\gamma x)$$

is called an *orbital integral.*

We have arrived at the *Frobenius character formula* for an induced representation

$$\sum_{\gamma\in\Gamma/\Gamma} a_\gamma \mathcal{O}_\gamma(\varphi) = \sum_{\iota\in I} m_\iota(\pi_{ind})\chi_\iota(\varphi) \tag{2.1}$$

This is the most naive form of the Arthur-Selberg trace formula. Observe that the right hand side remains mysterious, but the left hand side is now a concrete geometric expression involving volumes and orbital integrals.

2.2. **Poisson Summation Formula.** Now let us leave finite groups behind, and consider generalizations of the Frobenius character formula (2.1). We will begin by sacrificing complicated group theory and restrict to the simplest commutative Lie group $G = \mathbb{R}$.

A deceptive advantage of working with a commutative group is that we can explicitly calculate its spectrum.

Lemma 2.14. *The irreducible representations of* $\mathbb{R}$ *are the characters*

$$\chi_\lambda : \mathbb{R} \to \mathbb{C}^\times \qquad \chi_\lambda(x) = \exp(2\pi i \lambda x)$$

with $\lambda \in \mathbb{C}$. *In particular, the irreducible unitary representations are the characters* χ_λ, *with* $\lambda \in \mathbb{R}$.

Now let us consider the analogue of the Frobenius character formula 2.1 for the group $G = \mathbb{R}$. In order for the formula (not to mention our derivation of it) to make immediate sense, we should restrict to a subgroup $\Gamma \subset G$ which is discrete with compact quotient $\Gamma\backslash G$. Thus we are led to the subgroup of integers $\Gamma = \mathbb{Z}$ with quotient the circle $S^1 \simeq \mathbb{Z}\backslash\mathbb{R}$.

Let us calculate the various terms in the formula 2.1 for the induced Hilbert representation $L^2(S^1)$ of square-integrable complex-valued functions. For the geometric side, since $\mathbb{R}$ is commutative, the conjugacy class of $n \in \mathbb{Z}$ is simply n itself with volume 1. For the spectral side, Fourier analysis confirms that the representation $L^2(S^1)$ is a Hilbert space direct sum of the irreducible characters χ_λ, with $\lambda \in 2\pi i\mathbb{Z}$. Furthermore, a compactly supported test function $\varphi \in C_c^\infty(\mathbb{R})$ acts on the summand χ_λ by multiplication by its Fourier transform

$$\widehat{\varphi}(\lambda) = \int_{\mathbb{R}} \varphi(x)\chi_\lambda(x)dx.$$

Theorem 2.15 (Poisson Summation Formula)**.** *For a test function* $\varphi \in C_c^\infty(\mathbb{R})$, *one has the equality*

$$\sum_{n\in\mathbb{Z}} \varphi(n) = \sum_{\lambda\in\mathbb{Z}} \widehat{\varphi}(\lambda)$$

With this success in hand, one could seek other commutative groups and attempt a similar analysis. ¿From the vantage point of number theory, number fields provide a natural source of locally compact commutative groups.

By definition, a number field F is finite extension of the rational numbers $\mathbb{Q}$. There is a deep and pervasive analogy between number fields and the function fields $k(X)$ of algebraic curves X. For example, the fundamental example of a number field is $\mathbb{Q}$, and by definition, all others are finite extensions of it. The fundamental example of an algebraic curve is the projective line $\mathbb{P}^1$, and all other algebraic curves are finite covers of it. The history of the analogy is long with many refinements by celebrated mathematicians (Dedekind, Artin, Artin-Whaples, Weil,...). As we will recount below, one of the most intriguing aspects of the (currently known) proof of the Fundamental Lemma is its essential use of function fields and the extensive analogy between them and number fields.

Throughout what follows, we will need the number field analogue of the most basic construction of Calculus: the Taylor series expansion of a function around a point. Given a curve X and a (rational) point $x \in X$, we can choose a local coordinate t with a simple zero at x. Then for any non-zero rational function

$f \in k(X)$, we have its Laurent series expansion

$$\sum_{i=j}^{\infty} a_i t^i \in k((t)), \text{ with } a_j \neq 0.$$

Since rational functions are locally determined, this provides an embedding of fields $k(X) \subset k((t))$. The embedding realizes $k((t))$ as the completion of $k(X)$ with respect to the valuation $v_x(f) = k$.

Let us illustrate the form this takes for number fields with the fundamental example of the rational numbers $\mathbb{Q}$. The local expansion of an element of $\mathbb{Q}$ should take values in a completion of $\mathbb{Q}$. Ostrowski's Theorem confirms that the completions are precisely the p-adic numbers $\mathbb{Q}_p$, for all primes p, along with the real numbers $\mathbb{R}$. The real numbers are of course complete with respect to the usual Euclidean absolute value. The p-adic numbers are complete with respect to the absolute value $|f|_p = p^{-k}$, where $f = p^k a/b$, with $(a,p) = (b,p) = 1$. It satisfies the non-Archimedean property $|f+g|_p \leq \max\{|f|_p, |g|_p\}$, and so the compact unit ball of p-adic integers

$$\mathbb{Z}_p = \{f \in \mathbb{Q}_p | |f|_p \leq 1\} \subset \mathbb{Q}_p$$

is in fact a subring.

It is an elementary but immensely useful idea to keep track of all of the local expansions of a rational number at the same time. Observe that for a rational function on a curve, the points where it has a pole are finite in number. Similarly, only finitely many primes divide the denominator of a rational number. This leads one to form the ring of adeles

$$\mathbb{A}_{\mathbb{Q}} = \prod_{p \text{ prime}}^{rest} \mathbb{Q}_p \times \mathbb{R}$$

where the superscript "rest" denotes that we take the restricted product where all but finitely many terms lie in the compact unit ball of p-adic integers $\mathbb{Z}_p$. The simultaneous local expansion of a rational number provides an embedding of rings $\mathbb{Q} \subset \mathbb{A}_{\mathbb{Q}}$ with discrete image.

Let us justify the above somewhat technical construction with a well known result of number theory. To solve an equation in $\mathbb{Q}$, it is clearly necessary to provide solutions in $\mathbb{Q}_p$, for all primes p, and also $\mathbb{R}$. The Hasse principle asserts that to find solutions in $\mathbb{Q}$, one should start with such a solution in the adeles $\mathbb{A}_{\mathbb{Q}}$, or in other words, a collection of possibly unrelated solutions, and attempt to glue them together. Here is an example of the success of this approach.

Theorem 2.16 (Hasse-Minkowski). *Given a quadratic form*

$$Q(x_1, \ldots, x_n) = \textstyle\sum_{i \leq j} a_{ij} x_i x_j, \qquad a_{ij} \in \mathbb{Q},$$

the equation

$$Q(x_1, \ldots, x_n) = 0$$

has a solution in the rational numbers $\mathbb{Q}$ if and only if it has a solution in the adeles $\mathbb{A}_{\mathbb{Q}}$, or equivalently, solutions in the p-adic numbers $\mathbb{Q}_p$, for all primes p, and the real numbers $\mathbb{R}$.

A similar constructions of adeles make sense for arbitrary number fields F. The completions of F will be finite extensions of the completions of $\mathbb{Q}$, so finite extensions $F_{\mathfrak{p}}$ of the p-adic numbers $\mathbb{Q}_p$, along with possibly the real numbers $\mathbb{R}$ or complex numbers $\mathbb{C}$. The former are non-Archimedean so the compact unit balls of integers $\mathcal{O}_{\mathfrak{p}} \subset F_{\mathfrak{p}}$ are in fact subrings. One similarly forms the ring of adeles

$$\mathbb{A}_F = \prod_{\mathfrak{p}}^{rest} F_{\mathfrak{p}} \times \mathbb{R}^r \times \mathbb{C}^c$$

where $\mathfrak{p}$ runs over all non-Archimedean completions of F, and the superscript "rest" denotes that we take the restricted product where all but finitely many terms lie in the compact unit balls $\mathcal{O}_p$. The simultaneous local expansion of elements provides an embedding $F \subset \mathbb{A}_F$ with discrete image.

In his celebrated thesis, Tate generalized the Poisson Summation Formula to the pair of the locally compact group $\mathbb{A}_F$ and its discrete subgroup F. The resulting formula is an exact analogue of the classical Poisson Summation Formula

$$\sum_{x \in F} \varphi(x) = \sum_{\lambda \in F} \widehat{\varphi}(\lambda)$$

This is an essential part of Tate's interpretation of Class Field Theory in terms of harmonic analysis. One could approach all that follows as an attempt to explore the generalization of Class Field Theory to a noncommutative setting.

2.3. **Arthur-Selberg Trace Formula.** The Arthur-Selberg Trace Formula is a vast generalization of the Frobenius character formula for finite groups and the Poisson summation formula for number fields.

The starting point is an algebraic group G defined over a number field F. One can always realize G as a subgroup of $GL(n)$ defined by polynomial equations with coefficients in F. We will be most interested in *reductive* G, which means that we can realize G as a subgroup of $GL(n)$ preserved by the transpose of matrices, or equivalently, that the unipotent radical of G is trivial. Without further comment, we will also assume that G is connected in the sense that G is not the union of two proper subvarieties. Of course, $GL(n)$ itself is a fundamental example of a reductive algebraic group. For many important questions, the reader would lose nothing considering only $GL(n)$. But as we shall see, the role of the Fundamental Lemma is to help us compare different groups, and in particular, reduce questions about complicated groups to simpler ones.

Suppose we are given an algebraic group G defined over a number field F. Then it makes sense to consider the solutions $G(R)$ to the equations defining G in any commutative ring R containing F. Such solutions are called the *R-points of G* and form a group in the traditional sense of being a set equipped with a group law.

Less naively, although more trivially, for a field K containing F, we can also regard the coefficients of the equations defining G as elements of K. Hence we can consider G as an algebraic group defined over K. To keep things straight, we will write G_K to denote G thought of as an algebraic group defined over K. We will refer to G_K as the *base change of G* since all we have done is change the base field.

The only difference between the base change G_K and the original group G is that we are only allowed to form the R-points $G_K(R)$ of the base change for commutative rings R containing K. Experience tells us that over algebraically closed fields, there is little difference between equations and their solutions. In practice, this is true for algebraic groups: for an algebraic closure $\overline{F}$, one can go back and forth between the $\overline{F}$-points $G(\overline{F})$ and the base change $G_{\overline{F}}$.

Here is the most important class of reductive groups.

Definition 2.17. (1) A *torus* T defined over F is an algebraic group defined over F such that the base change $T_{\overline{F}}$ is isomorphic to the product $GL(1)^k$, for some k.

(2) A torus T defined over F is said to be *split* if it is isomorphic to the product $GL(1)^k$, for some k, without any base change necessary.

(3) A torus T defined over F is said to be *anisotropic*, or synonymously *elliptic*, if all of its characters are trivial

$$\mathrm{Hom}_F(T, GL(1)) = \langle 0 \rangle$$

where Hom_F denotes homomorphisms of algebraic groups defined over F.

Example 2.18. There are two types of one-dimensional tori defined over the real numbers $\mathbb{R}$. There is the split torus

$$GL(1) \simeq \{ab = 1\}$$

with $\mathbb{R}$-points $\mathbb{R}^\times$, and the anisotropic torus

$$SO(2) = \{g \in GL(2) | g^{-1} = g^\tau\} \simeq \{x^2 + y^2 = 1\}$$

with $\mathbb{R}$-points the circle S^1. Over the complex numbers $\mathbb{C}$, the equations $ab = 1$ and $x^2 + y^2 = 1$ become equivalent via the transformation $a = x + iy$, $b = x - iy$.

It is often best to think of an algebraic group G defined over F as comprising roughly two pieces of structure:

(1) the base change $G_{\overline{F}}$ or group of $\overline{F}$-points $G(\overline{F})$ for an algebraic closure $\overline{F}$,
(2) the Galois descent data needed to recover the original equations of G itself.

To understand the result of the first step, we recall the following definition.

Definition 2.19. (1) A torus $T \subset G$ is said to be *maximal* if it is not a proper subgroup of another torus in G.

(2) A reductive algebraic group G is said to be *split* if it contains a maximal torus which is split.

Proposition 2.20. *Let G be a reductive algebraic group defined over a number field F. Then there is a unique split reductive algebraic group G^{spl} defined over $\mathbb{Q}$ such that*

$$G_{\overline{F}} \simeq G^{spl}_{\overline{F}}, \qquad \text{and in particular} \qquad G(\overline{F}) \simeq G^{spl}(\overline{F}).$$

In other words, over algebraically closed fields, all reductive groups are split.

There is a highly developed structure theory of reductive algebraic groups, but the subject is truly example oriented. There is the well known Cartan classification of split reductive algebraic groups.

Example 2.21 (Split classical groups)**.** There are four series of automorphism groups of familiar linear geometry.

(A_n) The special linear group

$$SL(n+1) = \{A \in GL(n+1) | \det(A) = 1\}.$$

(B_n) The odd special orthogonal group

$$SO(2n+1) = \{A \in GL(2n+1) | A^\tau Q_{2n+1} A = Q_{2n+1}, \det(A) = 1\},$$

$$Q_{2n+1} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & Q_{2n-1} & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

(C_n) The symplectic group

$$Sp(2n) = \{A \in GL(2n) | A^\tau \Omega_{2n} A = \Omega_{2n}, \det(A) = 1\},$$

$$\Omega_{2n} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & \Omega_{2n-2} & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

(D_n) The even special orthogonal group

$$SO(2n) = \{A \in GL(2n) | A^\tau Q_{2n} A = Q_{2n}, \det(A) = 1\},$$

$$Q_{2n} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & Q_{2n-2} & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

Each is simple in the sense that (it is not a torus and) any normal subgroup is finite, unlike for example $GL(n)$ which has center $GL(1)$ realized as diagonal invertible matrices.

Each is also split with maximal torus its diagonal matrices. Outside of finitely many exceptional groups, all simple split reductive algebraic groups are isomorphic to one of the finitely many finite central extensions or finite quotients of the above classical groups.

To illustrate the Galois descent data involved in recovering a reductive group G defined over F from its base change $G_{\overline{F}}$, let us consider the simple example of a torus T. The base change $T_{\overline{F}}$ is split and hence isomorphic to $GL(1)^k$, for some k. Its characters form a lattice

$$X^*(T_{\overline{F}}) = \operatorname{Hom}(T_{\overline{F}}, GL(1)) \simeq \mathbb{Z}^k$$

from which we can recover $T_{\overline{F}}$ as the spectrum

$$T_{\overline{F}} = \operatorname{Spec} \overline{F}[X^*(T_{\overline{F}})].$$

The Galois group $\Gamma = Gal(\overline{F}/F)$ naturally acts on the character lattice $X^*(T_{\overline{F}})$ by a finite group of automorphisms. This induces an action on the ring $\overline{F}[X^*(T_{\overline{F}})]$, and we can recover T as the spectrum of the ring of invariants

$$T = \operatorname{Spec} \overline{F}[X^*(T_{\overline{F}})]^\Gamma.$$

In the extreme cases, T is split if and only if the Galois action on $X^*(T_{\overline{F}})$ is trivial, and by definition, T is anisotropic if and only if the invariant characters $X^*(T_{\overline{F}})^\Gamma = \operatorname{Hom}_F(T, GL(1))$ are trivial.

There is a large class of non-split groups which contains all tori and is particularly easy to describe by Galois descent. All groups directly relevant to the Fundamental Lemma will come from this class. We give the definition and a favorite example here, but defer discussion of the Galois descent until Section 3.3

Definition 2.22. (1) A *Borel subgroup* $B \subset G$ is an algebraic subgroup such that the base change $B_{\overline{F}} \subset G_{\overline{F}}$ is a maximal solvable algebraic subgroup.

(2) A reductive algebraic group G is said to be *quasi-split* if it contains a Borel subgroup $B \subset G$.

Example 2.23 (Unitary groups)**.** Suppose E/F is a separable degree 2 extension of fields. Then there is a unique nontrivial involution of E fixing F which one calls conjugation and denotes by $a \mapsto \overline{a}$, for $a \in E$. The unitary group is the matrix group

$$U(n, J, E/F) = \{A \in GL(n, E) | \overline{A}^{\tau} JA = J, \det(A) = 1\}$$

where $\overline{J}^{\tau} = J$ is nondegenerate. We can view $U(n, J, E/F)$ as a subgroup of $GL(2n, F)$ cut out by equations defined over F, and hence $U(n, J, E/F)$ is a reductive algebraic group defined over F.

If we take J to be antidiagonal, then $U(n, J, E/F)$ is quasi-split with Borel subgroup its upper-triangular matrices. But if we take J to be the identity matrix, then for example $U(n, J, \mathbb{C}/\mathbb{R})$ is the familiar compact unitary group which is as far from quasi-split as possible. (All of its connected algebraic subgroups are reductive, and so have trivial unipotent radical.)

Suppose we are given an algebraic group G defined over a number field F, and also a commutative ring R with a locally compact topology. Then the group of R-points $G(R)$ is a locally compact topological group, and so amenable to the techniques of harmonic analysis. Most prominently, $G(R)$ admits a bi-invariant Haar measure, and we can study representations such as $L^2(\Gamma\backslash G(R))$, for discrete subgroups $\Gamma \subset G(R)$. As harmonic analysts, our hope is to classify the irreducible representations of $G(R)$, and decompose complicated representations such as $L^2(\Gamma\backslash G(R))$ into irreducibles. Our wildest dreams are to find precise analogues of the successes of Fourier analysis where the initial group G is commutative.

For example, suppose we are given a reductive algebraic group G defined over the integers $\mathbb{Z}$, so in particular, the rational numbers $\mathbb{Q}$. Then we can take K to be either the local field of p-adic numbers $\mathbb{Q}_p$ or real numbers $\mathbb{R}$. We obtain the p-adic groups $G(\mathbb{Q}_p)$ and the real Lie group $G(\mathbb{R})$. They are locally compact with respective maximal compact subgroups $G(\mathbb{Z}_p)$ where $\mathbb{Z}_p \subset \mathbb{Q}_p$ is the compact unit ball of p-adic integers, and $K(\mathbb{R})$, where $K \subset G$ is the fixed points of the involution which takes a matrix to its inverse transpose. Finally, we can consider them all simultaneously by forming the locally compact adèlic group

$$G(\mathbb{A}_{\mathbb{Q}}) \simeq \prod_{p \text{ prime}}^{rest} G(\mathbb{Q}_p) \times G(\mathbb{R}).$$

It is a fundamental observation that the inclusion $\mathbb{Q} \subset \mathbb{A}_{\mathbb{Q}}$ induces an inclusion $G(\mathbb{Q}) \subset G(\mathbb{A}_{\mathbb{Q}})$ with discrete image, and so the space of automorphic functions $L^2(G(\mathbb{Q})\backslash G(\mathbb{A}_{\mathbb{Q}}))$ presents a natural representation of $G(\mathbb{A}_{\mathbb{Q}})$, and in particular of the p-adic groups $G(\mathbb{Q}_p)$ and Lie group $G(\mathbb{R})$, to approach via harmonic analysis.

In general, for a reductive algebraic group G defined over an arbitrary number field F, by passing to all of the completions of F, we obtain the locally compact p-adic groups $G(F_{\mathfrak{p}})$ and possibly the Lie groups $G(\mathbb{R})$ and $G(\mathbb{C})$, depending on whether $\mathbb{R}$ and $\mathbb{C}$ occur as completions. They are locally compact with respective maximal compact subgroups $G(\mathcal{O}_{\mathfrak{p}}) \subset G(F_{\mathfrak{p}})$ where $\mathcal{O}_{\mathfrak{p}} \subset F_{\mathfrak{p}}$ is the ring of integers, $K(\mathbb{R}) \subset G(\mathbb{R})$, where $K \subset G$ is the fixed points of the involution which takes a matrix to its inverse transpose, and $U(\mathbb{C}) \subset G(\mathbb{C})$, where $U \subset G$ is the fixed points of the involution which takes a matrix to its conjugate inverse transpose. We can form the locally compact adèlic group

$$G(\mathbb{A}_F) \simeq \prod_{\mathfrak{p}}^{rest} G(F_{\mathfrak{p}}) \times G(\mathbb{R})^r \times G(\mathbb{C})^c$$

where $\mathfrak{p}$ runs over all non-Archimedean completions of F. As with the rational numbers, the inclusion $F \subset \mathbb{A}_F$ induces an inclusion $G(F) \subset G(\mathbb{A}_F)$ with discrete image, and so the space of automorphic functions $L^2(G(F)\backslash G(\mathbb{A}_F))$ presents a natural representation of $G(\mathbb{A}_F)$ to approach via harmonic analysis.

Example 2.24. The automorphic representation $L^2(G(F)\backslash G(\mathbb{A}_F))$ is far less abstract than might initially appear. Rather than recalling the general statement of strong approximation, we will focus on the classical case when our number field is the rational numbers $F = \mathbb{Q}$ and our group is $G = SL(2)$. Inside of the adèlic group $G(\mathbb{A}_\mathbb{Q})$, consider the product of maximal compact subgroups

$$K = \prod_{p \text{ prime}} SL(2, \mathbb{Z}_p) \times SO(2, \mathbb{R}).$$

Then with $\mathbb{H} \subset \mathbb{C}$ denoting the open upper halfplane, there is a canonical identification

$$\begin{array}{rcl} SL(2,\mathbb{Q})\backslash SL(2,\mathbb{A}_\mathbb{Q})/K & \simeq & SL(2,\mathbb{Z})\backslash SL(2,\mathbb{R})/SO(2,\mathbb{R}) \\ & \simeq & SL(2,\mathbb{Z})\backslash \mathbb{H}, \end{array}$$

and the latter is the moduli of elliptic curves. By passing to smaller and smaller subgroups of K, we obtain the moduli of elliptic curves with level structure. This classical realization opens up the study of the original automorphic representation to the more familiar techniques (Laplace-Beltrami operators, Hecke integral operators) of harmonic analysis.

Remark 2.25. It is beyond the scope of this article to explain, but suffice to say, the deepest secrets of the universe are contained in the spectrum of the automorphic representation $L^2(G(F)\backslash G(\mathbb{A}_F))$. The *Langlands correspondence* is a conjectural description of the spectrum, with the most prominent ingredient being representations of the Galois group $Gal(\overline{F}/F)$. Thanks to the symmetry of the situation, one can turn things around and attempt to understand $Gal(\overline{F}/F)$ in terms of $L^2(G(F)\backslash G(\mathbb{A}_F))$. When one shows that a Galois representation is automorphic, or in other words, occurs in the spectrum, this leads to many deep structural implications.

Not only in general, but even in specific cases, it is extremely difficult to confirm that a given Galois representation is automorphic. Often the only hope is to bootstrap off of the precious few historical successes by concrete techniques such as induction and less obviously justified approaches such as prayer. But the prospect of success is at least supported by *Langlands's functoriality* which conjectures that whenever there is an obvious relation between Galois representations, there should be a parallel relation between automorphic representations. In particular, there are often highly surprising relations between automorphic representations for different groups corresponding to much more prosaic relations of Galois representations. It is in this context that the Fundamental Lemma plays an essential role.

Now we arrive at the Arthur-Selberg Trace Formula which is the primary tool in the study of automorphic representations. For simplicity, let us restrict for the moment to the far more elementary setting where the quotient $G(F)\backslash G(\mathbb{A}_F)$ is compact. The group algebra $C_c^\infty(G(\mathbb{A}_F))$ of smooth, compactly supported functions on the adèlic group acts on the automorphic representation $L^2(G(F)\backslash G(\mathbb{A}_F))$

by compact operators

$$R(\varphi)f(g) = \int_{G(\mathbb{A}_F)} \varphi(gh)f(h)dh.$$

It follows that the representation decomposes as a Hilbert space direct sum of irreducible unitary representations

$$L^2(G(F)\backslash G(\mathbb{A}_F)) \simeq \bigoplus_{\iota} m_\iota \pi_\iota$$

We can form the character Trace R as a distribution on $G(\mathbb{A}_F)$. The formal analogue of the Frobenius character formula 2.1 is an instance of the Selberg Trace Formula.

Theorem 2.26 (Selberg Trace Formula for compact quotient)**.** *Suppose $G(F)\backslash G(\mathbb{A}_F)$ is compact. Then for any test function $\varphi \in C_c^\infty(G(\mathbb{A}_F))$, we have an identity*

$$\sum_{\gamma \in G(F)/G(F)} a_\gamma \mathcal{O}_\gamma(\varphi) = \sum_\iota m_\iota \chi_\iota(\varphi)$$

where a_γ is the volume of the quotient $G_\gamma(F)\backslash G_\gamma(\mathbb{A}_F)$, and the distribution $\mathcal{O}_\gamma$ is the orbital integral

$$\mathcal{O}_\gamma(\varphi) = \int_{[\gamma]\subset G} \varphi$$

over the $G(\mathbb{A}_F)$-conjugacy class $[\gamma] \subset G(\mathbb{A}_F)$.

For modern theory and applications, one needs Arthur's generalizations of the Selberg Trace Formula for very general quotients. The technical details are formidable and Arthur's expositions can not be improved upon. But its formal structure and application is the same. On the geometric side, we have a formal sum involving volumes and explicit orbital integrals in the adèlic group. On the spectral side, we have the character of the automorphic representation expressed as a formal integral over the characters of irreducible representations. The identification of the two sides gives us a starting point to attempt to understand the spectrum in terms of geometry.

Although there are important and difficult issues in making this formal picture rigorous, there is an immediately accessible piece of it which can be isolated. On the spectral side, there is the discrete part of the automorphic spectrum consisting of irreducibles which occur on their own with positive measure. On the geometric side, there are orbital integrals for elements $\gamma \in G(F)$ whose centralizers G_γ are anisotropic tori.

The Fundamental Lemma is needed for the comparison of the anisotropic terms of the geometric side of the trace formula for different groups. We can leave for another time the thorny complications of other aspects of the trace formula. From hereon, we can focus on orbital integrals over anisotropic conjugacy classes. Moreover, we can expand each anisotropic orbital integral around each adèlic place to obtain

$$\mathcal{O}_\gamma(\varphi) = \prod_{\mathfrak{p}} \mathcal{O}_\gamma(\varphi_{\mathfrak{p}}) \tag{2.2}$$

where $\mathfrak{p}$ runs over all completions of F, we expand γ at each place, $\mathcal{O}_\gamma(\varphi_{\mathfrak{p}})$ denotes the orbital integral along the conjugacy class $[\gamma] \subset G(F_{\mathfrak{p}})$, and without sacrificing too much, we work with a product test function $\varphi = (\varphi_{\mathfrak{p}})$. Thus from hereon, leaving global motivations behind, we can focus on orbital integrals over conjugacy classes in local groups.

3. Eigenvalues versus characteristic polynomials

Our discussion of the previous section is a success if the reader comes away with the impression that outside of the formidable technical issues in play, the basic idea of the trace formula is a kind of formal tautology. The great importance and magical applications of Arthur's generalizations to arbitrary adèlic groups are found in comparing trace formulas for different groups. This is the primary approach to realizing instances of Langlands's functoriality conjectures on the relation of automorphic forms on different groups. The general strategy is to compare the geometric sides where traces are expressed in concrete terms, and thus arrive at conclusions about the mysterious spectral sides. By instances of Langlands's reciprocity conjectures, the spectral side involves Galois theory, and eventually leads to deep implications in number theory.

Now an immediate obstruction arises when one attempts to compare the geometric sides of the trace formulas for different groups. Orbital integrals over conjugacy classes in different groups have no evident relation with each other. Why should we expect conjugacy classes of say symplectic matrices and orthogonal matrices to have anything to talk about? If we diagonalize them, their eigenvalues live in completely different places. But here is the key observation that gives one hope: *the equations describing their eigenvalues are in fact intimately related.* In other words, if we pass to an algebraic closure, where equations and their solutions are more closely tied, then we find a systematic relation between conjugacy classes. To explain this further, we will start with some elementary linear algebra, then build to Langlands's theory of endoscopy, and in the end, arrive at the Fundamental Lemma.

3.1. **The problem of Jordan canonical form.** Suppose we consider a field k, and a finite-dimensional k-vector space $V \simeq k^n$. Given an endomorphism $A \in \operatorname{End}_k(V) \simeq M_{n\times n}(k)$, form the characteristic polynomial

$$p_A(t) = \det(t \operatorname{Id}_V - A) = a_0 + a_1 t + \cdots + a_{n-1}t^{n-1} + t^n \in k[t].$$

For simplicity, we will assume that the roots of $p_A(t)$, or equivalently, the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A, are all distinct. Of course, if k is not algebraically closed, or more generally, does not contain the roots of $p_A(t)$, we will need to pass to an extension of k to speak concretely of the roots.

Let's review the two "canonical" ways to view the endomorphism A. On the one hand, we can take the coefficients of $p_A(t)$ and form the companion matrix

$$C_A = \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & -a_{n-1} \end{bmatrix}$$

Since we assume that $p_A(t)$ has distinct roots, and hence is equal to its minimal polynomial, C is the rational normal form of A, and hence A and C_A will be conjugate. We think of this as the naive *geometric* form of A.

On the other hand, we can try to find a basis of V in which A is as close to diagonal as possible. If k is algebraically closed, or more generally, contains the eigenvalues of A, then we will be able to conjugate A into Jordan canonical form.

In particular, since we assume that A has distinct eigenvalues, A will be conjugate to the diagonal matrix

$$D_A = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

We think of this as the sophisticated *spectral* form of A. It is worth noting that the most naive "trace formula" is found in the identity

$$\mathrm{Trace}(A) = -a_{n-1} = \lambda_1 + \cdots + \lambda_n$$

which expresses the spectral eigenvalues of A in terms of the geometric sum along the diagonal of A.

When k is not algebraically closed, or more specifically, does not contain the eigenvalues of A, understanding the structure of A is more difficult. It is always possible to conjugate A into rational normal form, but not necessarily Jordan canonical form. One natural solution is to fix an algebraic closure $\bar{k}$, and regard A as an endomorphism of the extended vector space $\overline{V} = V \otimes_k \bar{k} \simeq \bar{k}^n$. Then we can find a basis of $\overline{V}$ for which A is in Jordan canonical form. Equivalently, we can conjugate A into Jordan canonical form by an element of the automorphism group $\mathrm{Aut}_{\bar{k}}(\overline{V}) \simeq GL(n, \bar{k})$. This is particularly satisfying since Jordan canonical forms of matrices completely characterize their structure.

Lemma 3.1. *If two matrices $A, A' \in M_{n\times n}(k)$ are conjugate by an element of $GL(n, \bar{k})$, they are in fact conjugate by an element of $GL(n, k)$.*

All of the subtlety of the Fundamental Lemma emanates from the difficulty that when we consider subgroups of $GL(n)$, the above lemma consistently fails. For example, suppose we restrict the automorphism group of our vector space V to be the special linear group $SL(n)$. In other words, we impose that the symmetries of V be not all invertible linear maps, but only those preserving volume. Then Jordan canonical form is no longer a complete invariant for the equivalence classes of matrices.

Example 3.2. Take $k = \mathbb{R}$. Consider the rotations of the real plane $V = \mathbb{R}^2$ given by the matrices

$$A(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \qquad A'(\theta) = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}.$$

Observe that both $A(\theta), A'(\theta)$ lie in $SL(2,\mathbb{R})$, and they are conjugate by the matrix

$$M = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} \in SL(2, \mathbb{C}).$$

Furthermore, when $\theta \notin \pi\mathbb{Z}$, there is no element in $SL(2,\mathbb{R})$ which conjugates one into the other. When we view $A(\theta), A'(\theta)$ as endomorphisms of the complex plane $\overline{V} = \mathbb{C}^2$, they both are conjugate to the diagonal matrix

$$D(\theta) = \begin{bmatrix} \cos(\theta) + i\sin(\theta) & 0 \\ 0 & \cos(\theta) - i\sin(\theta) \end{bmatrix}.$$

Let us introduce some Lie theory to help us think about the preceding phenomenon. For simplicity, we will work with a split reductive group G whose derived group $G_{der} = [G, G]$ is simply connected. For example, the split classical groups of Example 2.21 are all simple, hence equal to their derived groups. The special linear and symplectic groups are simply-connected, but for the special orthogonal group, one needs to pass to the spin two-fold cover.

Fix a split maximal torus $T \subset G$, and recall that the Weyl group of G is the finite group $W = N_T/T$, where $N_T \subset G$ denotes the normalizer of T. All split tori are conjugate by $G(k)$ and the choice of $T \subset G$ is primarily for convenience.

To begin, let us recall the generalization of Jordan canonical form. Recall that to diagonalize matrices with distinct eigenvalues, in general, we have to pass to an algebraically closed field $\overline{k}$.

Definition 3.3. For an element $\gamma \in G(k)$, let $G_\gamma \subset G$ denote its centralizer.

(1) The element γ is said to be *regular* if G_γ is commutative.

(2) The element γ is said to be *semisimple* if G_γ is connected and reductive.

(3) The element γ is said to be *regular semisimple* if it is regular and semisimple, or equivalently G_γ is a torus.

(3) The element γ is said to be *anisotropic* if G_γ is an anisotropic torus.

Example 3.4. Take $k = \mathbb{R}$ and $G = SL(2)$. Consider the elements

$$r = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \qquad s = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad h = \begin{bmatrix} 2 & 0 \\ 0 & 1/2 \end{bmatrix} \qquad a = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

with respective centralizers $T_r = \mathbb{Z}/2\mathbb{Z} \times \mathbb{A}^1$, $T_s = SL(2)$, $T_h = GL(1)$, $T_a = SO(2)$. Thus r is regular, s is semisimple, h and a are regular semisimple, and a is anisotropic. To see the latter fact, observe that there are no nontrivial homomorphisms $S^1 \to \mathbb{R}^\times$ of their groups of $\mathbb{R}$-points.

Remark 3.5. Some prefer the phrase *strongly regular semisimple* for an element $\gamma \in G(k)$ whose centralizer G_γ is a torus and not a possibly disconnected commutative reductive group. When G_{der} is simply-connected, if the centralizer G_γ is a reductive group then it will be connected.

Remark 3.6. Some might prefer to define anisotropic to be slightly more general. Let us for the moment call a regular semisimple element $\gamma \in G(F)$ *anistropic modulo center* if the quotient $T/Z(G)$ of the centralizer $T = G_\gamma$ by the center $Z(G)$ is anisotropic.

A group with split center such as $GL(n)$ will not have anisotropic elements, but will have elements anisotropic modulo center. A regular semisimple element $\gamma \in GL(n)$ will be anisotropic modulo center if and only if its characteristic polynomial is irreducible (and separable).

For the Fundamental Lemma, we will be able to focus on anisotropic elements. Somewhat surprisingly, it is not needed for $GL(n)$ where there is no elliptic endoscopy.

The following justifies the idea that semisimple elements are "diagonalizable" and regular semisimple elements are "diagonalizable with distinct eigenvalues".

Proposition 3.7. *(1) Every semisimple element of $G(\overline{k})$ can be conjugated into $T(\overline{k})$.*

(2) Two semisimple elements of $T(\overline{k})$ are conjugate in $G(\overline{k})$ if and only if they are conjugate by the Weyl group W_G.

(3) An element of $T(\overline{k})$ is regular semisimple if and only the Weyl group W_G acts on it with trivial stabilizer.

Second, let us generalize the notion of characteristic polynomial. Recall that the coefficients of the characteristic polynomial are precisely the conjugation invariant polynomial functions on matrices.

Theorem 3.8 (Chevalley Restriction Theorem)**.** *The G-conjugation invariant polynomial functions on G are isomorphic to the W-invariant polynomial functions on T. More precisely, restriction along the inclusion $T \subset G$ induces an isomorphism*

$$k[G]^G \xrightarrow{\sim} k[T]^W.$$

Passing from polynomial functions to algebraic varieties, we obtain the G-invariant Chevalley morphism

$$\chi : G \longrightarrow T/W = \operatorname{Spec} k[T]^W.$$

It assigns to a group element its "unordered set of eigenvalues", or in other words its characteristic polynomial.

Finally, let us mention the generalization of rational canonical form for split reductive groups. Recall that a *pinning* of a split reductive group G consists of a Borel subgroup $B \subset G$, split maximal torus $T \subset B$, and basis vectors for the resulting simple positive root spaces. The main consequence of a pinning is that only central conjugations preserve it, and so it does away with ambiguities coming from inner automorphisms. (For slightly more discussion, including the example of $G = SL(n)$, see Section 3.3 below where we discuss root data.)

Theorem 3.9 (Steinberg section)**.** *Given a pinning of the split reductive group G with split maximal torus $T \subset G$, there is a canonical section*

$$\sigma : T/W \longrightarrow G$$

to the Chevalley morphism χ. In other words, $\chi \circ \sigma$ is the identity.

Thus to each "unordered set of eigenvalues", we can assign a group element with those eigenvalues.

With the above results in hand, we can now introduce the notion of stable conjugacy. Recall that given $\gamma \in G(k)$, we denote by $[\gamma] \subset G(k)$ the conjugacy class through γ.

Definition 3.10. Let G be a simply-connected reductive algebraic group defined over a field k.

We say two regular semisimple elements $\gamma, \gamma' \in G(k)$ are *stable conjugate* and write $\gamma \sim_{st} \gamma'$ if they satisfy one of the following equivalent conditions:

(1) γ and γ' are conjugate by an element of $G(\overline{k})$,

(2) γ and γ' share the same characteristic polynomial $\chi(\gamma) = \chi(\gamma')$.

Given $\gamma \in G(k)$, the *stable conjugacy class* $[\gamma]_{st} \subset G(k)$ through γ consists of $\gamma' \in G(k)$ stably conjugate to γ.

Remark 3.11. Experts in algebraic groups over fields of finite characteristic will note the useful fact that in Proposition 3.7 and Definition 3.10, one need only go to the separable closure k_s.

Recall that the geometric side of the trace formula leads to orbital integrals over conjugacy classes of regular semisimple elements in groups over local fields. The theory of canonical forms for elements is intricate, and conjugacy classes are not characterized by their Jordan canonical forms. The complication sketched above for $SL(2,\mathbb{R})$ is quite ubiquitous, and one will also encounter it for the classical groups of Example 2.21. Any hope to understand conjugacy classes in concise terms must involve passage to the stable conjugacy classes found over the algebraic closure. In simpler terms, we must convert constructions depending on eigenvalues, such as orbital integrals, into constructions depending on characteristic polynomials.

3.2. **Fourier theory on conjugacy classes.** Imbued with a proper fear of the intricacy of conjugacy classes over non-algebraically closed fields, one dreams that the geometric side of the trace formula could be rewritten in terms of stable conjugacy classes which are independent of the field.

One might not expect that the set of conjugacy classes in a given stable conjugacy class would be highly structured. But it turns out there is extra symmetry governing the situation.

To simplify the discussion, it will be useful to make the standing assumption that the reductive group G is simply connected, or more generally, its derived group $G_{der} = [G, G]$ is simply-connected.

Proposition 3.12. *Let G be a reductive algebraic group defined over a local field F.*

Let $\gamma_0 \in G(F)$ be a regular semisimple element with centralizer the torus $T = G_{\gamma_0}$.

Then the set of conjugacy classes in the stable conjugacy class $[\gamma_0]_{st}$ is naturally a finite abelian group given by the kernel of the Galois cohomology map

$$\mathcal{A}_{\gamma_0} \simeq \ker\{H^1(F,T) \longrightarrow H^1(F,G)\}.$$

In particular, it only depends on the centralizing torus T as a subgroup of G.

Remark 3.13. (1) When G is simply connected, $H^1(F,G)$ is trivial.

(2) One can view the Galois cohomology $H^1(F,T)$ as parameterizing principal T-bundles over $\operatorname{Spec} F$. Since T is abelian, this is naturally a group. Under the isomorphism of the proposition, the trivial bundle corresponds to γ_0.

(3) One can view the quotient $[\gamma_0]_{st}/G(F)$ of the stable conjugacy class by conjugation as a discrete collection of classifiying spaces for stabilizers. Each of the classifiying spaces is noncanonically isomorphic to the classifying space of $T = G_{\gamma_0}$. The possible isomorphisms form a principal T-bundle giving the corresponding class in $H^1(F,T)$.

Now suppose we have a $G(F)$-invariant distribution on the stable conjugacy class containing γ_0. In other words, we have a distribution of the form

$$\delta(\varphi) = \sum_{\gamma \in \mathcal{A}_{\gamma_0}} c_\gamma \mathcal{O}_\gamma(\varphi)$$

where as usual $\mathcal{O}_\gamma(\varphi)$ denotes the orbital integral over the $G(F)$-conjugacy class through γ. In what sense could we demand the distribution δ be invariant along the entire stable conjugacy class? Requiring the coefficients c_γ are all equal is a lot to ask for, but there is a reasonable generalization presented by Fourier theory.

Consider the Pontryagin dual group of characters

$$\mathcal{A}_{\gamma_0}^D = \mathrm{Hom}(\mathcal{A}_{\gamma_0}, \mathbb{C}^\times).$$

Definition 3.14. Let G be a reductive algebraic group defined over a local field F. Let $\gamma_0 \in G(F)$ be an anisotropic element.

Given $\kappa \in \mathcal{A}_{\gamma_0}^D$, the *$\kappa$-orbital integral* through γ_0 is the distribution

$$\mathcal{O}_{\gamma_0}^\kappa(\varphi) = \sum_{\gamma \in \mathcal{A}_{\gamma_0}} \kappa(\gamma)\mathcal{O}_\gamma(\varphi)$$

In particular, when $\kappa = e \in \mathcal{A}_{\gamma_0}^D$ is the trivial character, the stable orbital integral is the distribution

$$\mathcal{SO}_{\gamma_0}(\varphi) = \mathcal{O}_{\gamma_0}^e(\varphi).$$

Remark 3.15. Observe that the stable orbital integral $\mathcal{SO}_{\gamma_0}(\varphi)$ is independent of the choice of base point γ_0 in the stable conjugacy class. Thus it is truly associated to the characteristic polynomoial $\chi(\gamma_0)$.

On the other hand, the dependence of the κ-orbital integral $\mathcal{O}_{\gamma_0}^\kappa(\varphi)$ on the base point γ_0 is modest but nontrivial. If one chooses some other $\gamma_0' \in \mathcal{A}_{\gamma_0}$, the resulting expression will scale by $\kappa(\gamma_0')$. For groups with simply-connected derived groups, there is the base point, which is canonical up to a choice of pinning, given by the image of the Steinberg section $\sigma(\chi(\gamma_0))$.

Now by Fourier theory, we can write our original distribution δ as a finite sum

$$\delta(\varphi) = \sum_{\kappa \in \mathcal{A}_{\gamma_0}^D} c_\kappa \mathcal{O}_{\gamma_0}^\kappa(\varphi).$$

Hence while δ might not have been stable, it can always be written as a linear combination of distributions which vary along the stable conjugacy class by a character.

Now to proceed any further, we must understand the character group

$$\mathcal{A}_{\gamma_0}^D = \mathrm{Hom}(\mathcal{A}_{\gamma_0}, \mathbb{C}^\times).$$

A closer examination of the possible characters will reveal the possibility of a deep reinterpretation of the κ-orbital integrals.

Suppose the local field F is non-Archimedean, and fix a torus T defined over F. Recall that we can think of T as the information of a split torus $T_{\overline{F}} \simeq GL(1)^k$ over the algebraic closure $\overline{F}$, together with the Galois descent data needed to recover the original equations cutting out T. The descent is captured by the finite action of the Galois group $\Gamma = Gal(\overline{F}/F)$ on the cocharacter lattice

$$X_*(T_{\overline{F}}) = \mathrm{Hom}(GL(1), T_{\overline{F}}) \simeq \mathbb{Z}^k.$$

Consider the dual complex torus

$$T^\vee = \mathrm{Spec}\, \mathbb{C}[X_*(T_{\overline{F}})] \simeq GL(1)^k$$

whose monomial functions are the cocharacter lattice. The Γ-action on $X_*(T_{\overline{F}})$ induces a corresponding Γ-action on $T^\vee$.

Proposition 3.16 (Local Tate-Nakayama duality)**.** *Assume F is a non-Archimedean local field. There is a canonical identification of abelian groups*

$$H^1(F, T)^D \simeq \pi_0((T^\vee)^\Gamma)$$

between the Pontryagin dual of the Galois cohomology of T, and the component group of the Γ-invariants in the dual torus $T^\vee$.

Remark 3.17. When G is simply connected, $H^1(F, G)$ is trivial, and so we have calculated $\mathcal{A}^D_{\gamma_0}$.

When G is not simply-connected, elements of $\pi_0((T^\vee)^\Gamma)$ nonetheless restrict to characters of $\mathcal{A}^D_{\gamma_0}$. It is an exercise to relate the kernel of this restriction to $\pi_1(G)$.

Thus a regular semisimple element $\gamma_0 \in G(F)$ provides a centralizing torus $T = G_{\gamma_0}$ which in turn determines a Galois action on the dual torus $T^\vee$. To each element $\kappa \in (T^\vee)^\Gamma$ in the Galois-fixed locus, we can associate the κ-orbital integral $\mathcal{O}^\kappa_{\gamma_0}(\varphi)$ defined by the image of κ in the component group $\pi_0((T^\vee)^\Gamma)$.

3.3. **Endoscopic groups and the Fundamental Lemma.** We have reached a pivotal point in our discussion. Let's step back for a moment and take measure of its successes and shortcomings.

Given a number field F, and a reductive algebraic group G defined over F, we aim to understand the automorphic representation $L^2(G(F)\backslash G(\mathbb{A}_F))$. Our main tool is the Arthur-Selberg Trace Formula which provides the character of the representation in terms of orbital integrals over conjugacy classes of the adèlic group. Furthermore, we have focused on the anisotropic conjugacy classes and expressed their orbital integrals in terms of κ-twisted orbital integrals over stable conjugacy classes in p-adic groups.

It is not too much of a stretch to argue that the κ-stable orbital integrals $\mathcal{O}^\kappa_{\gamma_0}(\varphi)$ are more appealing than the basic orbital integrals $\mathcal{O}_\gamma(\varphi)$ since their dependence on the conjugacy classes within a stable conjugacy class is through a character rather than a specific choice of conjugacy class. This is an early manifestation of the motivic, or universal algebraic, nature of the κ-orbital integrals. But of course, aesthetics aside, Fourier inversion tells us we can go back and forth between the two, and so in some sense we have not accomplished very much.

Thus perhaps we have made a Faustian bargain: we have traded the evident geometric structure of basic orbital integrals $\mathcal{O}_\gamma(\varphi)$ for the representation theoretic structure of κ-orbital integrals $\mathcal{O}^\kappa_{\gamma_0}(\varphi)$. With our original aim to compare trace formulas for different groups, one could even worry that we have made things more difficult rather than less. Indeed, one could argue that what we have done "is obviously useless, because the term $\mathcal{O}^\kappa_{\gamma_0}(\varphi)$ is still defined in terms of G" rather than some other group [H].

But now we have arrived in the neighborhood of the Fundamental Lemma. It is the lynchpin in Langlands's theory of endoscopy which relates κ-orbital integrals to stable orbital integrals on other groups. The theory of endoscopy (for which we recommend the original papers of Kottwitz [K84, K86]) has many facets, but at its center is the following question:

> *For a local field F, given an element $\gamma_0 \in G(F)$, and a compatible character $\kappa \in T^\vee$, on what group H should we try to express the κ-orbital integral $\mathcal{O}^\kappa_{\gamma_0}(\varphi)$ as a stable orbital integral?*

The answer is what ones calls the *endoscopic group* associated to the given data. At first pass, it is a very strange beast, neither fish nor fowl. But the Fundamental Lemma is what confirms it is the correct notion.

There is a great distance between the intuitive idea of an endoscopic group and the minimal notions one needs to at least spell out the Fundamental Lemma. Most of the technical complications devolve from the intricacy of Galois descent for quasi-split groups. So it seems useful, though less efficient, to first explain the basic notions assuming all groups are split (Definition 2.19), and then add in the necessary bells and whistles for quasi-split groups (Definition 2.22).

3.3.1. *Split groups.* We begin with a reminder of the "combinatorial skeleton" of a split reductive group given by its root datum. We will always equip all split reductive groups G with a *pinning* consisting of a Borel subgroup $B \subset G$, split maximal torus $T \subset B$, and basis vectors for the resulting simple positive root spaces. This has the effect of providing a canonical splitting

$$\begin{tikzcd} 1 \arrow[r] & \mathrm{Inn}(G) = G/Z(G) \arrow[r] & \mathrm{Aut}(G) \arrow[r] & \mathrm{Out}(G) \arrow[r] \arrow[l, bend right] & 1 \end{tikzcd}$$

since the automorphisms of G preserving the pinning map isomorphically to the outer automorphisms of G.

Example 3.18. Take $G = SL(n)$ and $T \subset SL(n)$ the split maximal torus of diagonal matrices of determinant one.

Then the symmetric group Σ_n acts simply transitively on the possible Borel subgroups $B \subset G$ satisfying $T \subset B$. Let us choose $B \subset SL(n)$ to consist of upper-triangular matrices of determinant one.

The resulting simple positive root spaces can be identified with the $n-1$ superdiagonal matrix entries (directly above the diagonal). Let us choose the basis given by taking the element 1 in each simple positive root space.

The outer automorphisms $\mathrm{Out}(SL(2))$ are trivial, but when $n > 2$, the outer automorphisms $\mathrm{Out}(SL(n))$ are the group $\mathbb{Z}/2\mathbb{Z}$. The above pinning realizes the nontrivial outer automorphism as the automorphism given by

$$A \longmapsto M(A^{-1})^{\tau} M^{-1}, \qquad A \in SL(n),$$

where τ denotes the transpose operation, and M is the antidiagonal matrix with $M_{i,n-i+1} = 1$ when $i < n/2$, and $M_{i,n-i+1} = -1$ when $i \geq n/2$.

Definition 3.19. (1) A (reduced) *root datum* is an ordered quadruple

$$\Psi = (X, \Phi, X^\vee, \Phi^\vee)$$

of the following data:

(1) $X, X^\vee$ are finite rank free $\mathbb{Z}$-modules in duality by a pairing

$$\langle , \rangle : X \times X^\vee \longrightarrow \mathbb{Z}$$

(2) $\Phi, \Phi^\vee$ are finite subsets of $X, X^\vee$ respectively in fixed bijection

$$\alpha \longleftrightarrow \alpha^\vee$$

We will always assume that our root data are reduced in the sense that if $\alpha \in \Phi$, then $c\alpha \in \Phi$ if and only if $c = \pm 1$.

The data must satisfy the following properties:

(a) $\langle \alpha, \alpha^\vee \rangle = 2$,

(b) $s_\alpha(\Phi) \subset \Phi, s^\vee_\alpha(\Phi^\vee) \subset \Phi^\vee$, where

$$\begin{aligned} s_\alpha(x) &= x - \langle x, \alpha^\vee \rangle \alpha, \quad && x \in X, \alpha \in \Phi, \\ s^\vee_\alpha(y) &= y - \langle \alpha, y \rangle \alpha^\vee, \quad && y \in X^\vee, \alpha \in \Phi. \end{aligned}$$

The *Weyl group* W_Ψ of the root datum is the finite subgroup of $GL(X)$ generated by the reflections s_α, for $\alpha \in \Phi$.

(2) A *based root datum* is an ordered sextuple

$$\Psi = (X, \Phi, \Delta, X^\vee, \Phi^\vee, \Delta^\vee)$$

consisting of a root datum $(X, \Phi, X^\vee, \Phi^\vee)$ together with a choice of subsets

$$\Delta \subset \Phi, \Delta^\vee \subset \Phi^\vee$$

satisfying the following properties:

(a) the bijection $\Phi \longleftrightarrow \Phi^\vee$ restricts to a bijection $\Delta \longleftrightarrow \Delta^\vee$,

(b) there exists an element $v \in X$ with trivial stabilizer in W_Ψ, for which we have
$$\Delta^\vee = \{\alpha^\vee \in \Phi^\vee | \langle v, \alpha^\vee \rangle > 0\}$$

To a split reductive group G with a Borel subgroup $B \subset G$, and maximal torus $T \subset B$, one associates the based root datum

$$\Psi(G) = (X_*, \Phi_G, \Delta_G, X^*, \Phi_G^\vee, \Delta_G^\vee)$$

consisting of the following:

- $X_* = X_*(T) = \mathrm{Hom}(GL(1), T)$ the cocharacter lattice,
- $X^* = X^*(T) = \mathrm{Hom}(T, GL(1))$ the character lattice,
- $\Phi_G \subset X_*$ the coroots,
- $\Phi_G^\vee \subset X^*$ the roots,
- $\Delta_G \subset \Phi_G$ the simple coroots, and
- $\Delta_G^\vee \subset \Phi_G^\vee$ the simple roots.

The Weyl group $W_{\Psi(G)}$ coincides with the usual Weyl group $W_G = N_T/T$.

Here is a key motivation for the notion of based root data.

Theorem 3.20. *Fix a field k.*

(1) Every based root datum Ψ is isomorphic to the based root datum $\Psi(G)$ of some split reductive group G, defined over k, and equipped with a pinning.

(2) The automorphisms of the based root datum $\Psi(G)$ are isomorphic to the outer automorphisms of G, or equivalently, the automorphisms of G, as an algebraic group defined over k, that preserve its pinning.

The combinatorial classification of groups finds ubiquitous application, and is further justified by many natural occurrences of related structures such as Dynkin diagrams. But one of its initially naive but ultimately deep implications is the evident duality for reductive groups coming from the duality of root data. It generalizes the very concrete duality for tori we have seen earlier which assigns to a split torus $T = \mathrm{Spec}\, k[X^*(T)]$ the dual complex torus $T^\vee = \mathrm{Spec}\, \mathbb{C}[X_*(T)]$.

Definition 3.21. Let G be a split reductive group with based root datum

$$\Psi(G) = (X_*, \Phi_G, X^*, \Phi_G^\vee, \Delta_G, \Delta_G^\vee).$$

The Langlands dual group $G^\vee$ is the split reductive complex algebraic group with dual based root datum

$$\Psi(G^\vee) = (X^*, \Phi_G^\vee, \Delta_G^\vee, X_*, \Phi_G, \Delta_G).$$

Remark 3.22. We have stated the duality asymmetrically, where G is defined over some field k, but the dual group $G^\vee$ is always a complex algebraic group. Observe that such asymmetry arose for tori when we described complex characters. In general, it stems from the fact that our automorphic representations are complex vector spaces.

For a group G with Langlands dual group $G^\vee$, the maximal torus $T^\vee \subset G^\vee$ is the dual of the maximal torus $T \subset G$, the Weyl groups W_G and $W_{G^\vee}$ coincide, the outer automorphisms $\mathrm{Out}(G)$ and $\mathrm{Out}(G^\vee)$ coincide, and the roots of G are the coroots of $G^\vee$ and vice-versa. If G is simple, then so is $G^\vee$, and if in addition G is complex, $Z(G) \simeq \pi_1(G^\vee)$ and vice-versa (generalizations of the last assertion are possible but one has to be careful to compare potentially different kinds of groups).

Example 3.23. The following are pairs of Langlands dual groups: $GL(n) \longleftrightarrow GL(n)$, $SL(n) \longleftrightarrow PGL(n)$, $SO(2n+1) \longleftrightarrow Sp(2n)$, $SO(2n) \longleftrightarrow SO(2n)$.

Although the above definition is concrete, there is a deep mystery in passing from a group to root data, dual root data, and then back to a group again. Commutative and combinatorial structures are the only things which can easily cross the divide.[3]

Now we arrive at the notion of endoscopic group in the context of split groups.

Definition 3.24. Let G be a split reductive algebraic group with split maximal torus $T \subset G$.

(1) *Split endoscopic data* is an element $\kappa \in T^\vee \subset G^\vee$.

(2) Given split endoscopic data $\kappa \in T^\vee$, the associated *split endoscopic group* of G is the split reductive algebraic group H whose Langlands dual group $H^\vee$ is the connected component of the centralizer $G^\vee_\kappa \subset G^\vee$ of the element κ.

It follows immediately that T is also a maximal torus of H and the coroots Φ_H are a subset of the coroots Φ_G. More precisely, the element $\kappa \in T^\vee = \mathrm{Hom}(X_*(T), GL(1))$ can be evaluated on $X_*(T)$, and in particular on $\Phi_G \subset X_*(T)$, and the coroots Φ_H are given by the kernel

$$\Phi_H = \{\alpha \in \Phi_G | \kappa(\alpha) = 1\}.$$

This immediately implies that the roots $\Phi_H^\vee$ are the corresponding subset of the roots Φ_G, and the Weyl group W_H is a subgroup of the Weyl group W_G. *But this by no means implies that H is anything close to a subgroup of G.*

Example 3.25. We will work with the split groups defined in Example 2.21.

Take $G = Sp(2n)$ the symplectic group so that $G^\vee = SO(2n+1)$ the odd orthogonal group. Recall that the diagonal matrices inside of $SO(2n+1)$ furnish a split maximal torus $T^\vee$. Take the element

$$\kappa = \begin{bmatrix} 1 & 0 \\ 0 & -I_n \end{bmatrix} \in T^\vee$$

with centralizer $O(2n)$ which is disconnected with connected component $H^\vee = SO(2n)$. Taking the Langlands dual of $H^\vee$ gives the endoscopic group $H = SO(2n)$.

[3] There are many hints in quantum field theory of "missing" higher-dimensional objects which can be specialized on the one hand to reductive groups, and on the other hand to their root data. But until they or their mathematical analogues are understood in some form, the relation of group to root data and hence to dual group will likely remain mysterious.

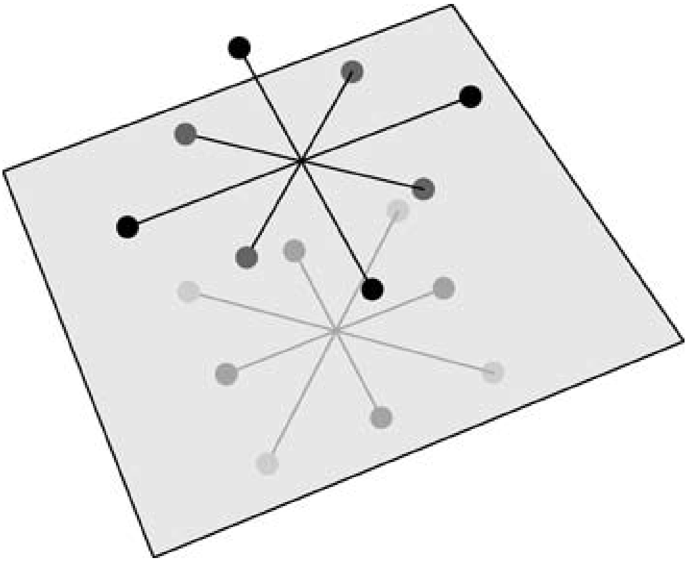


FIGURE 2. The root system of the group $Sp(4)$ with the roots of the endoscopic group $SO(4)$ highlighted. Reflected in the background is the root system for the Langlands dual group $SO(5)$ with the roots of the subgroup $SO(4)$ highlighted.

One can check that there is no nontrivial homomorphism from $SO(2n)$ to $Sp(2n)$ when $n \geq 3$. (When $n = 2$, there are homomorphisms, but they do not induce the correct maps on root data.)

3.3.2. *General story: quasi-split groups.* Now we will pass to the general setting of quasi-split groups. The inexperienced reader could skip this material and still find plenty of interesting instances of the Fundamental Lemma to ponder. Throughout what follows, let F be a non-Archimedean local field with residue field f.

Definition 3.26. (1)A field extension E/F with residue field e/f is said to be *unramified* if the degrees satisfy $[E : F] = [e : f]$.

(2) A reductive algebraic group G defined over a local field F is said to be *unramified* if the following hold:

(a) G is quasi-split,
(b) $G_{F^{un}}$ is split for the maximal unramified extension F^{un}/F.

Unramified groups G are combinatorial objects, classified by the split reductive group $G_{F^{un}}$ together with the Galois descent homomorphism

$$\rho_G : Gal(F^{un}/F) \longrightarrow \mathrm{Out}(G)$$

which is completely determined by its value on the Frobenius automorphism.

To state the general notion of unramified endoscopic group, we will need the following standard constructions. Given a split reductive group G, and an element $\kappa \in T^\vee \subset G^\vee$, recall that we write $H^\vee$ for the connected component of the centralizer $G^\vee_\kappa \subset G^\vee$ of the element κ. We will write $\pi_0(\kappa)$ for the component group of the centralizer of the element (κ, e) of the semi-direct product $G^\vee \rtimes \mathrm{Out}(G^\vee)$. There are canonical maps

$$\begin{tikzcd} \mathrm{Out}(H^\vee) & \pi_0(\kappa) \arrow[l, "\pi_{H^\vee}"'] \arrow[r, "\pi_{G^\vee}"] & \mathrm{Out}(G^\vee) \end{tikzcd}$$

$$\begin{tikzcd} W_{H^\vee} \rtimes \mathrm{Out}(H^\vee) \arrow[r] & W_{G^\vee} \rtimes \mathrm{Out}(G^\vee) \end{tikzcd}$$

where the latter is compatible with the actions on $T^\vee$ and projections to $\pi_0(\kappa)$, and extends the canonical inclusion $W_{H^\vee} \to W_{G^\vee}$.

Definition 3.27. Let G be an unramified reductive group, so in particular quasi-split with Borel subgroup $B \subset G$, and (not necessarily split) maximal torus $T \subset B$.

(1) Unramified *endoscopic data* is a pair (κ, ρ_κ) consisting of an element $\kappa \in T^\vee \subset G^\vee$, and a homomorphism

$$\rho_\kappa : Gal(F^{un}/F) \longrightarrow \pi_0(\kappa) \qquad \text{such that } \rho_G = \pi_{G^\vee} \circ \rho_\kappa.$$

(2) Given unramified endoscopic data (κ, ρ_κ), the associated unramified *endoscopic group* of G is the unramified reductive group H defined over F constructed as follows. Recall that the split endoscopic group H^{spl} associated to the element κ is the connected component of the centralizer $G^\vee_\kappa \subset G^\vee$. The endoscopic group H is the form of H^{spl} defined over F by the the Galois descent homomorphism

$$\rho_H = \pi_{H^\vee} \circ \rho_\kappa : Gal(F^{un}/F) \longrightarrow \mathrm{Out}(H^{spl}).$$

Example 3.28 (Endoscopic groups for $SL(2)$)**.** There are three unramified endoscopic groups for $SL(2)$. Recall that the dual group is $PGL(2)$.

1. $SL(2)$ itself with $\kappa = e \in T^\vee$ the identity, and ρ_κ trivial.
2. $GL(1)$ with $\kappa \neq e \in T^\vee$ any nontrivial element, and ρ_κ trivial.
3. $U(1, E/F)$ with $\kappa = -e \in T^\vee$ the square-root of the identity, and ρ_κ the non-trivial map to $\pi_0(\kappa) \simeq \mathbb{Z}/2\mathbb{Z}$.

The first two are split, but the last is not.

3.3.3. *Statement of Fundamental Lemma.* Finally, we arrive at our destination.

The Fundamental Lemma relates the κ-orbital integral over a stable conjugacy class in the group G with the stable orbital integral over a stable conjugacy class in an endoscopic group H. Such an idea should lead to some immediate confusion: *the orbital integrals to be compared are distributions on different groups, so to compare them we must also have some correspondence of test functions.*

There is a deep and intricate theory of transferring test functions of which the Fundamental Lemma is in some sense the simplest and thus most important instance. It states that in the most hospitable situation, the most simple-minded transfer of the simplest test functions leads to a good comparison of orbital integrals. There are many variations (twisted, weighted, ...) of the Fundamental Lemma, but the most important are now understood thanks to reductions to the Fundamental Lemma or extensions of the ideas of its proof.

Fix an unramified group G defined over F. The fact that G is unramified implies that it is the localization of a smooth affine group scheme $\mathcal{G}$ defined over the ring of integers $\mathcal{O}_F$ whose special fiber over the residue field k is connected reductive.

Definition 3.29. A maximal compact subgroup $K \subset G(F)$ is said to be *hyperspecial* if there is a smooth affine group scheme $\mathcal{G}$ defined over $\mathcal{O}_F$ such that

(1) $\mathcal{G}_F = G$,
(2) $\mathcal{G}(\mathcal{O}_F) = K$, and
(3) $\mathcal{G}_k$ is connected reductive.

Lemma 3.30. *A reductive algebraic group G defined over a local field F is unramified if and only if $G(F)$ contains a hyperspecial maximal compact subgroup.*

Now an endoscopic group H is not a subgroup of G. Rather we must content ourselves with the relationship of characteristic polynomials

$$\begin{tikzcd} H \arrow[d, "\chi_H"'] & G \arrow[d, "\chi_G"] \\ T/W_H \arrow[r, "\nu"] & T/W_G \end{tikzcd}$$

This provides a relationship between stable conjugacy classes as follows. Given a parameter $a_H \in (T/W_H)(F)$, we can consider its transfer $a_G = \nu(a_H) \in (T/W_G)(F)$. Even if $a_H \in (T/W_H)(F)$ is regular, since the inclusion $W_H \subset W_G$ is not an isomorphism (except when $H = G$), the transfer $a_G \in (T/W_G)(F)$ might not be regular.

Definition 3.31. A parameter $a_H \in (T/W_H)(F)$ is said to be *G-regular* if it and its transfer $a_G = \nu(a_H) \in (T/W_G)(F)$ are both regular.

Now given a *G-regular* parameter $a_H \in (T/W_H)(F)$ with transfer $a_G = \nu(a_H) \in (T/W_G)(F)$, we have the stable conjugacy class

$$[\gamma_H]_{st} = \chi_H^{-1}(a_H) \subset H \quad \text{and its transfer} \quad [\gamma_G]_{st} = \chi_G^{-1}(a_G) \subset G.$$

Fundamental Lemma 3.1 ([N08]). *Let F be a local field.*

Let G be an unramified group defined over F.

Let H be an unramified endoscopic group for G associated to endoscopic data (κ, ρ_κ).

Let $K_G \subset G(F), K_H \subset H(F)$ be hyperspecial maximal compact subgroups.

Then for G-regular $a_H \in T/W_H$ and transfer $a_G = \nu(a_H) \in (T/W_G)(F)$, we have an equality

$$\mathcal{SO}_{\gamma_H}(1_{K_H}) = \Delta(\gamma_H, \gamma_G)\mathcal{O}^\kappa_{\gamma_G}(1_{K_G})$$

where $\gamma_H \in \chi_H^{-1}(a_H)$, $\gamma_G \in \chi_G^{-1}(a_G)$, and $\Delta(\gamma_H, \gamma_G)$ is the transfer factor (which shall not be defined here).

Remark 3.32. A precise formulation of transfer factors first appears in Langlands's joint work with D. Shelstad [LS87]. The transfer factor $\Delta(\gamma_H, \gamma_G)$ accounts for the ambiguity that the κ-orbital integral $\mathcal{O}^\kappa_{\gamma_G}(\varphi)$ depends on the choice of lift $\gamma_G \in \chi_G^{-1}(a_G)$. By definition, the stable orbital integral $\mathcal{SO}_{\gamma_H}(\varphi)$ is an invariant of $a_H = \chi_H(\gamma_H)$. It is worth mentioning that if G is split and the derived group $G_{der} = [G, G]$ is simply-connected, then the Steinberg section provides a distinguished lift $\gamma_G = \sigma(a_G)$.

Remark 3.33. There is an analogous "Fundamental Lemma" for Archimedean local fields, resolved long ago by D. Shelstad [S82], which one also needs for applications of the trace formula. The example at the beginning of the Introduction fits into this Archimedean part of the theory.

It should be apparent that one can formulate a Lie algebra variant of the Fundamental Lemma. Namely, let $\mathfrak{g}$ be the Lie algebra of G, and $\mathfrak{h}$ the Lie algebra of an endoscopic group. Then one can replace the stable conjugacy classes of the group elements $\gamma_H \in H(F)$ and $\gamma_G \in G(F)$ with those of Lie algebra elements $\xi_H \in \mathfrak{h}(F)$ and $\xi_G \in \mathfrak{g}(F)$.

In fact, the situation simplifies further in that each stable conjugacy class now has a canonical element. To see this, observe that stable conjugacy classes in the Lie algebra $\mathfrak{g}$ are the fibers of the Chevalley morphism

$$\chi : \mathfrak{g} \longrightarrow \mathfrak{t}/W = \operatorname{Spec} k[\mathfrak{t}]^W.$$

For the choice of a pinning, there is the Kostant section

$$\sigma : \mathfrak{t}/W \longrightarrow \mathfrak{g}.$$

This is completely general, unlike for the group G, where the Steinberg section could be defined only when the derived group $G_{der} = [G, G]$ is simply-connected.

Thus with the assumptions of the Fundamental Lemma stated above, the Lie algebra variant takes the form of an identity

$$\mathcal{SO}_{a_H}(1_{\mathfrak{h}(\mathcal{O}_F)}) = \mathcal{O}^{\kappa}_{a_G}(1_{\mathfrak{g}(\mathcal{O}_F)})$$

where we index the stable orbital integral by the parameter $a_H \in (\mathfrak{t}/W_H)(F)$, and we base the κ-orbital integral at the image of the Kostant section applied to the transfer $a_G = \nu(a_H) \in (\mathfrak{t}/W_G)(F)$. In particular, the distinguished base point obviates the need for any transfer factor.

An important theorem of Waldspurger [W08] asserts that the Lie algebra variant of the Fundamental Lemma implies the original statement. And it is the Lie algebra variant which Ngô proves, and which we will turn to in the next section.

4. Geometric interpretation of Fundamental Lemma

In retrospect, the search for a proof of the Fundamental Lemma turns out to be the search for a setting where the powerful tools of algebraic geometry – Hodge theory, Lefschetz techniques, sheaf theory, homological algebra – could be brought to bear on the problem. As we have seen, the Fundamental Lemma is an analytic assertion about integrals of characteristic functions on p-adic groups. But these functions turn out to be of motivic origin, and hence amenable to the deep mid-to-late 20th century synthesis of algebraic geometry and algebraic topology.

Here is a historical antecedent worth keeping in mind (which is deeply intertwined with the Fundamental Lemma and its proof). The classical Riemann hypothesis that all non-trivial zeros of the Riemann zeta function have real part $1/2$ is a difficult problem. Even more dauntingly, it admits a well-known generalization to any number field F for which the classical version is the case of the rational numbers $\mathbb{Q}$. Nevertheless, an analogous Riemann hypothesis for function fields of curves, formulated by Artin, was proved for elliptic curves by Hasse, and then for all genus curves by Weil. It involves counting the number of points of the curve over finite fields. The basic case of the projective line is completely trivial. The Weil conjectures are a vast generalization to all algebraic varieties. They were a prominent focus of Grothendieck and Serre, and eventually established by Deligne. At the heart of the proof is the interpretation of counting points in terms of cohomology with Galois actions. So in the end, the point counts have much more structure than one might have thought.

4.1. **Motivic origin.** Though the Fundamental Lemma is a local statement, its motivation comes from global questions about a number field. Without Langlands functoriality and the Arthur-Selberg Trace Formula in mind, it would be hard to arrive at the Fundamental Lemma as a reasonable assertion. Nevertheless, once we dispense with motivation, the concrete problem to be solved involves only local fields.

Remarkably, the analogy between number fields and function fields of curves leads to precise mathematical comparisons between their completions. It is hard to imagine that all of the intricate arithmetic of a number field could be found in the geometry of an algebraic curve. For example, the structure of the rational numbers $\mathbb{Q}$ is far more complicated than that of the projective line $\mathbb{P}^1$. But it turns out that important structures of the p-adic fields $\mathbb{Q}_p$ can be found in the Laurent series fields $\mathbb{F}_p((t))$. Naively, though $\mathbb{Q}_p$ is of characteristic zero and $\mathbb{F}_p((t))$ is of characteristic p, they share the formal structure that each is a complete local field with residue field $\mathbb{F}_p$. In particular, one can transport the statement of the Fundamental Lemma from the setting of p-adic fields to Laurent series fields.

The following liberating theorem of Waldspurger opens the door to geometric techniques.

Theorem 4.1 ([W06]). *The Lie algebra variant of the Fundamental Lemma in the equal characteristic or geometric case of* $\mathbb{F}_p((t))$ *implies the Lie algebra variant of the Fundamental Lemma in the unequal characteristic or arithmetic case of* $\mathbb{Q}_p$.

Remark 4.2. Waldspurger proves the above assertion for sufficiently large residual characteristic p. Thanks to previous work of Hales [H95], this suffices to establish the Fundamental Lemma for arbitrary residual characteristic.

Waldspurger's proof is a tour de force of representation theory. It provides a detailed analysis of constructions which are natural but specific to reductive groups.

Cluckers, Hales, and Loeser [CHL] have discovered a completely independent proof of the above result. The context of their arguments is mathematical logic, specifically model theory. While this is unfamiliar to many, it is very appealing in that it gets to the heart, or motivic truth, of the independence of characteristic. Very roughly speaking, the arguments take seriously the idea that integrals contain more information than their numerical values. Rather, they should be thought of as universal linear expressions for their cycles of integration weighted by their integrands. It applies generalizations of Cluckers-Loeser [CL05, CL], building on work of Denef-Loeser [DL01], of the classical Ax-Kochen-Ersov Theorem that given φ a first order sentence (a formula with no free variables) in the language of rings, for almost all prime numbers p, the sentence φ is true in $\mathbb{Q}_p$ if and only if it is true in $\mathbb{F}_p((t))$.

4.2. **Affine Springer fibers.** In the geometric setting of Laurent series fields, thanks to the Weil conjectures, the Fundamental Lemma takes on a topological form. It comes down to comparing the cohomology of affine Springer fibers for endoscopic groups. We will explain what affine Springer fibers are and some beautiful techniques of broad applicability developed to understand their cohomology.

Affine Springer fibers were introduced by Kazhdan-Lusztig [KL88] as natural generalizations of Grothendieck-Springer fibers. The latter play a fundamental role in Springer's theory of Weyl group representations, as well as Lusztig's theory of character sheaves. If we take the viewpoint that Grothendieck-Springer fibers are essential to the characters of groups over a finite field, then it is not surprising that affine Springer fibers figure prominently in the characters of p-adic groups.

For simplicity, and to appeal to topological methods, we will work here with base field the complex numbers $\mathbb{C}$. We write $\mathcal{K} = \mathbb{C}((t))$ for the Laurent series field, and $\mathcal{O} = \mathbb{C}[[t]]$ for its ring of integers.

Remark 4.3. Since $\mathbb{C}$ is algebraically closed, all unramified groups over $\mathcal{K}$ will be split, and so we will not be discussing the most general form of the Fundamental Lemma. We hope the broader intutions available in this setting will compensate for the sacrifice of generality.

Let G be a complex reductive group with Lie algebra $\mathfrak{g}$. Consider the so-called affine or loop group $G_\mathcal{K} = G(\mathbb{C}((t)))$, and its subgroup of arcs $G_\mathcal{O} = G(\mathbb{C}[[t]])$. The quotient $Gr_G = G_\mathcal{K}/G_\mathcal{O}$ is called the affine Grassmannian of G. It is an increasing union of projective varieties indexed by the natural numbers. The group $G_\mathcal{K}$ naturally acts on Gr_G by multiplication, and we can think of its Lie algebra $\mathfrak{g}_\mathcal{K}$ as acting infinitesimally by vector fields.

Definition 4.4. For an element $\xi \in \mathfrak{g}_\mathcal{K}$ of the Lie algebra of the loop group $G_\mathcal{K}$, the affine Springer fiber Gr_G^ξ is the fixed-points of the vector field ξ acting on the affine Grassmannian Gr_G. It admits the concrete description

$$Gr_G^\xi = \{g \in G_\mathcal{K} | \operatorname{Ad}_g(\xi) \in \mathfrak{g}_\mathcal{O}\}/G_\mathcal{O}$$

where $\mathfrak{g}_\mathcal{O}$ denotes the Lie algebra of $G_\mathcal{O}$.

Although affine Springer fibers are not finite-type schemes, when the parameter ξ is regular semisimple, their underlying reduced schemes are reasonable. We will see this by examining the natural symmetry groups acting upon them.

First, observe that the Springer fiber only depends upon the $G_\mathcal{K}$-conjugacy class of ξ, and if ξ is not conjugate to any $\xi' \in \mathfrak{g}_\mathcal{O}$ then the Springer fiber will be empty. So there is no loss in assuming $\xi \in \mathfrak{g}_\mathcal{O}$ from the start. Moreover, any regular semisimple $\xi \in \mathfrak{g}_\mathcal{O}$ is $G_\mathcal{K}$-conjugate to a unique $\xi_0 = \sigma(a) \in \mathfrak{g}_\mathcal{O}$ in the image of the Kostant slice

$$\sigma : (\mathfrak{t}/W)_\mathcal{O} \longrightarrow \mathfrak{g}_\mathcal{O}.$$

So in what follows, we will assume that $\xi = \sigma(a)$ for a fixed regular element $a \in (\mathfrak{t}/W)_\mathcal{O}$.

Now, let $T = (G_\mathcal{K})_\xi \subset G_\mathcal{K}$ denote the (possibly ramified) torus given by the centralizer of ξ. Let $S \subset T$ be its maximally unramified subtorus. In other words, if we write $X_*(T_{\overline{\mathcal{K}}})$ for the cocharacters of T, then the cocharacters of S are the Galois-invariants

$$X_*(S) = X_*(T_{\overline{\mathcal{K}}})^{Gal(\overline{\mathcal{K}}/\mathcal{K})}.$$

A key observation is that T canonically extends to a smooth commutative group-scheme J defined over $\mathcal{O}$. Namely, over the regular locus $\mathfrak{g}^{reg} \subset \mathfrak{g}$, we have the smooth commutative group-scheme of centralizers $I \to \mathfrak{g}^{reg}$, and we can form the fiber product

$$J = \operatorname{Spec} \mathcal{O} \times_\mathfrak{g} I$$

over the base point $\xi = \epsilon(a) : \operatorname{Spec} \mathcal{O} \to \mathfrak{g}^{reg} \subset \mathfrak{g}$.

We will write Λ_ξ for the coweight lattice $X_*(S)$, and $\mathcal{P}_\xi$ for the affine Grassmannian $J_\mathcal{K}/J_\mathcal{O}$. The latter is an increasing union of not necessarily projective varieties. Since J is commutative, $\mathcal{P}_\xi$ is naturally a commutative group. ¿From the constructions, $\mathcal{P}_\xi$ and hence also its subgroup Λ_ξ naturally act on Gr_G^ξ.

Proposition 4.5 ([KL88]). *Suppose $\xi \in \mathfrak{g}_\mathcal{K}$ is regular semisimple.*

Then the underlying reduced scheme $Gr_{G,red}^\xi$ of the affine Springer fiber is a countable union of projective irreducible components. Furthermore, the natural Λ_ξ-action on $Gr_{G,red}^\xi$ is free with quotient $Gr_{G,red}^\xi/\Lambda_\xi$ a projective variety.

Example 4.6. Take $G = SL(2)$, and consider the elements

$$\xi_1 = \begin{bmatrix} 0 & t \\ 1 & 0 \end{bmatrix}, \xi_2 = \begin{bmatrix} 0 & t^2 \\ 1 & 0 \end{bmatrix}, \xi_2' = \begin{bmatrix} t & 0 \\ 0 & -t \end{bmatrix} \in \mathfrak{g}_\mathcal{K}.$$

Then (at the level of reduced schemes) $Gr_G^{\xi_1}$ is a point, and $Gr_G^{\xi_2}$ is an infinite string of $\mathbb{P}^1$'s attached one after the other at nodes. The lattice Λ_{ξ_2} is rank one, and the quotient $Gr_G^{\xi_2}/\Lambda_{\xi_2}$ is a nodal elliptic curve.

To see things for ξ_2, it helps to note that ξ_2 and ξ_2' are $G_\mathcal{K}$-conjugate, and so one can calculate with ξ_2'. On the other hand, ξ_1 is anisotropic and can not be diagonalized.

Now given a character

$$\kappa : \mathcal{P}_\xi/\Lambda_\xi \to \mathbb{C}^\times,$$

we can consider the summand

$$H^*(Gr^\xi/\Lambda_\xi)_\kappa \subset H^*(Gr^\xi/\Lambda_\xi)$$

consisting of cocycles that transform by the character κ under the action of $\mathcal{P}_\xi/\Lambda_\xi$.

Thanks to the following application of the Weil conjectures, the Fundamental Lemma involves identifying this cohomology with the invariant cohomology of an affine Springer fiber for an endoscopic group.

Theorem 4.7. *Suppose $\xi \in \mathfrak{g}_\mathcal{K}$ is regular semisimple. Then the κ-orbital integral $\mathcal{O}_\xi^\kappa(1_{\mathfrak{g}_\mathcal{O}})$ can be recovered from the κ-summand $H^*(Gr^\xi/\Lambda_\xi)_\kappa$, or more precisely from its underlying motive.*

Remark 4.8. By underlying motive, we mean that one should properly work over a finite field with ℓ-adic cohomology, and use the Grothendieck-Lefschetz fixed point formalism to recover orbital integrals from traces of Frobenius.

Unfortunately, there is no evident geometric relationship between the affine Springer fibers for a group and an endoscopic group. And any attempt to explicitly calculate the cohomology must face the fact that already for the symplectic group $Sp(6)$, Bernstein and Kazhdan [KL88, Appendix] found an affine Springer fiber whose cohomology contains the motive of a hyperelliptic curve.

4.3. **Equivariant localization.** Goresky-Kottwitz-MacPherson proposed the intriguing idea that the combinatorics of endoscopic groups hint at a possible geometric mechanism for relating affine Springer fibers. Recall that the roots of an endoscopic group are an explicit subset of the roots of the original group. For an affine Springer fiber with large toric symmetry, Goresky-Kottwitz-MacPherson recognized that its toric one-skeleton of zero and one-dimensional toric orbits was in fact completely encoded by the roots. In turn, they also discovered that for a very general class of varieties with large toric symmetries, their cohomology could be read off from their toric one-skeleta.

Let X be a (possibly singular) projective variety equipped with an action of the torus $T = (\mathbb{C}^\times)^n$. For simplicity, we will fix an embedding $X \subset \mathbb{CP}^N$, and assume the T-action is induced by a linear T-action on $\mathbb{C}^{N+1}$. Then there is a moment map

$$\mu : X \hookrightarrow \mathbb{CP}^N \longrightarrow \mathfrak{t}^\vee$$

which induces the corresponding infinitesimal action of $\mathfrak{t}$. For a vector $v \in \mathfrak{t}$, regarded as a vector field on $\mathbb{CP}^N$, the one-form $\omega(v, -)$ obtained by contracting with the Kahler form is given by the pairing $\langle d\mu(-), v\rangle$ Furthermore, we will assume the images of the fixed points $X^T \subset X$ are all distinct, in particular, there are only finitely many.

Definition 4.9. Suppose $T = (\mathbb{C}^\times)^n$ acts on X with finitely many one-dimensional orbits. Let $X_1 \subset X$ be the one skeleton of fixed points and one-dimensional orbits. The moment graph $\Gamma_T(X)$ is defined to be the quotient

$$\Gamma_T(X) = X_1/T_c,$$

where $T_c = (S^1)^n$ is the compact torus inside T. The moment map descends to a canonical immersion

$$\mu : \Gamma_T(X) \longrightarrow \mathfrak{t}^\vee.$$

The structure of the moment graph is not only the abstract graph, but also its immersion into $\mathfrak{t}^\vee$. The image of any one-dimensional orbit $\mathcal{O} \subset X_1$ will be a line segment $\ell \subset \mathfrak{t}^\vee$ whose orthogonal in $\mathfrak{t}$ is the Lie algebra of the stabilizer of $\mathcal{O}$.

Example 4.10. Fix a maximal torus $T \subset G$. For $\xi \in \mathfrak{t}_\mathcal{O} \subset \mathfrak{g}_\mathcal{K}$, the centralizer $Z_{G_\mathcal{K}}(\xi)$ contains the affine torus $T_\mathcal{K}$, and hence the full coweight lattice $\Lambda = \mathrm{Hom}(\mathbb{C}^\times, T)$. If ξ is regular semsimple, the maximal torus $T \subset G$ acts on Gr_G^ξ/Λ with finitely many one-dimensional orbits. The resulting moment graph $\Gamma_T(Gr_G^\xi/\Lambda)$ is an invariant of the root data of G.

The following equivariant localization theorem is not difficult but gives a beautiful combinatorial picture of the cohomology of X. Its validity depends on the technical assumption that the T-action on X is equivariantly formal. We will not explain what this means, but only mention that it follows from more familiar conditions such as the vanishing of the odd degree cohomology of X, or if the mixed Hodge structure on the cohomology of X is in fact pure.

Theorem 4.11 ([GKM98]). *Suppose the T-action on X is equivariantly formal. Then the cohomology of X is an invariant of the moment graph $\Gamma_T(X)$.*

This general viewpoint on the cohomology of spaces with torus actions has been very fruitful. Foremost, there is the original application of Goresky-Kottwitz-MacPherson to the Fundamental Lemma.

Corollary 4.12 ([GKM04]). *For $\xi \in \mathfrak{t}_\mathcal{O} \subset \mathfrak{g}_\mathcal{K}$ regular semisimple, if the cohomology of Gr_G^ξ/Λ is pure, then the Fundamental Lemma holds for ξ.*

Remark 4.13. Note that the corollary assumes ξ lies in the loop algebra of $\mathfrak{t}$. This is a very restrictive condition on a regular semisimple element. In particular, it implies that its centralizer is an *unramified* torus. This starting point is what provides sufficient symmetries to apply equivariant localization to the affine Springer fiber.

Outside of the question of what to do with affine Springer fibers with less symmetry, the issue of purity is a formidable obstacle to further progress in the local setting. Some cases were directly settled by Goresky-Kottwitz-MacPherson [GKM06], but for a uniform understanding, new ideas appeared necessary.

5. Hitchin fibration and the search for purity

In this section, we will continue to work over the complex numbers $\mathbb{C}$ in order to appeal to topological methods of broad familiarity. We will adopt all of the notation of the previous section, so for example we write $\mathcal{K} = \mathbb{C}((t))$ for the Laurent series field, and $\mathcal{O} = \mathbb{C}[[t]]$ for its ring of integers. Given the reductive group G, we write $G_\mathcal{K}$ for its loop group, and $G_\mathcal{O}$ for the subgroup of arcs.

5.1. **Grothendieck-Springer resolution.**

Affine Springer fibers for the loop group $G_\mathcal{K}$ are a natural generalization of Springer fibers for the original group G.

For a nilpotent element $\xi \in \mathcal{N} \subset \mathfrak{g}$ of the Lie algebra of G, the Springer fiber $\mathcal{B}^\xi$ is the fixed-points of the vector field ξ acting on the flag variety $\mathcal{B}$ of all Borel subalgebras $\mathfrak{b} \subset \mathfrak{g}$. They naturally arise as the fibers of the Springer resolution of the nilpotent cone

$$\mu_\mathcal{N} : \tilde{\mathcal{N}} \simeq T^*(G/B) = \{(\mathfrak{b}, \xi) \in \mathcal{B} \times \mathcal{N} | \xi \in \mathfrak{b}\} \longrightarrow \mathcal{N} \qquad (\mathfrak{b}, \xi) \longmapsto \xi$$

Beginning with Springer's construction of Weyl group representations in their cohomology, Springer fibers are now ubiquitous in representation theory.

A vital observation is that the Springer resolution may be extended to the so-called Grothendieck-Springer resolution

$$\mu_\mathfrak{g} : \tilde{\mathfrak{g}} = \{(\mathfrak{b}, \xi) \in \mathcal{B} \times \mathfrak{g} | \xi \in \mathfrak{b}\} \longrightarrow \mathfrak{g} \qquad (\mathfrak{b}, \xi) \longmapsto \xi$$

While Springer fibers for nilpotent elements $\xi \in \mathcal{N}$ are difficult and important, Springer fibers for regular semisimple elements $\xi \in \mathfrak{g}^{reg,ss}$ are finite and on their own uninteresting. So what have we accomplished by introducing the Grothendieck-Springer resolution? *We can now reduce questions about interesting Springer fibers to questions about dull Springer fibers.*

What technique can we use to relate the cohomology of the fibers of a map such as the Grothendieck-Springer resolution? Sheaf theory. The cohomology of the fibers of a map are precisely the local invariants of the derived pushforward of the constant sheaf along the map. Global results about the pushforward will imply local results about the cohomology of the fibers. For example, much of Springer's theory of Weyl group representations is encoded in the following statement (see for example [G83, HoKa84]).

Theorem 5.1. *The restriction of the derived pushforward $R\mu_{\mathfrak{g}*}\mathbb{C}_{\tilde{\mathfrak{g}}}$ of the constant sheaf on $\tilde{\mathfrak{g}}$ to the open locus of regular semisimple elements $\mathfrak{g}^{reg,ss} \subset \mathfrak{g}$ is a local system with monodromies given by the regular representation of the Weyl group of $\mathfrak{g}$. The entire pushforward is the canonical intersection cohomology extension of this local system.*

Returning to the loop group $G_{\mathcal{K}}$, the affine Springer fibers Gr_G^{ξ} are also fibers of an analogous map. But here the target is the infinite-dimensional Lie algebra $\mathfrak{g}_{\mathcal{K}}$ and the fibers Gr_G^{ξ} are not projective varieties. The powerful topological methods of algebraic geometry do not (yet) extend to such a setting. Some new idea is needed to proceed.

5.2. **Compactified Jacobians.** Laumon introduced a beautiful way to begin to study the relation of affine Springer fibers Gr_G^{ξ} for varying parameters $\xi \in \mathfrak{g}_{\mathcal{K}}$. As we vary the parameter $\xi \in \mathfrak{g}_{\mathcal{K}}$, the behavior of the affine Springer fibers Gr_G^{ξ} is far wilder than we might expect.

Example 5.2. Take $G = SL(2)$, and consider the family of elements

$$\xi(\varepsilon) = \left[\begin{array}{cc} 0 & \varepsilon t^2 + t^3 \\ 1 & 0 \end{array} \right] \in \mathfrak{g}_{\mathcal{K}}, \text{ for } \varepsilon \in \mathbb{C}.$$

When $\epsilon = 0$, (at the level of reduced schemes) the affine Springer fiber $Gr_G^{\xi(\varepsilon)}$ is simply $\mathbb{P}^1$, but when $\varepsilon \neq 0$, it is an infinite string of $\mathbb{P}^1$'s attached one after the other at nodes with symmetry lattice $\Lambda_{\xi(\varepsilon)} \simeq \mathbb{Z}$.

Laumon recognized that the quotients Gr_G^{ξ}/Λ_{ξ} which interest us most in fact form reasonable families in a topological sense.

Proposition 5.3 ([La06]). *Suppose $G = GL(n)$, and let $\xi \in \mathfrak{g}_{\mathcal{K}}$ be regular semisimple. Then there is a complex projective curve C_{ξ} of genus zero, with a single singular point, such that the quotient Gr_G^{ξ}/Λ_{ξ} is homeomorphic to the compactified Jacobian*

$$\overline{Jac}(C_{\xi}) = \{\textit{degree } 0, \textit{ rank } 1 \textit{ torsion free sheaves on } C_{\xi}\}.$$

This is a powerful insight: such curves C_{ξ} form nice finite-dimensional families, and hence so do their compactificed Jacobians $\overline{Jac}(C_{\xi})$.

Example 5.4. (compare with Example 5.2). Take $G = SL(2)$, and consider the family of elements

$$\xi(\varepsilon) = \left[\begin{array}{cc} 0 & \varepsilon t^2 + t^3 \\ 1 & 0 \end{array} \right] \in \mathfrak{g}_{\mathcal{K}}, \text{ for } \varepsilon \in \mathbb{C}.$$

Then we can take $C_{\xi(\varepsilon)}$ to be the family of singular elliptic curves $y^2 = \varepsilon t^2 + t^3$. When $\epsilon = 0$, the curve $C_{\xi(0)}$ is a cuspidal elliptic curve, and when $\varepsilon \neq 0$, it is a nodal elliptic curve. The compactified Jacobian of any elliptic curve, smooth or singular, is isomorphic to the curve itself.

By deforming such curves, Laumon provides a natural geometric setting for the equivariant localization techniques of Goresky-Kottwitz-MacPherson. In particular, he was able to deduce the Fundamental Lemma for unitary groups contingent upon the purity of the affine Springer fibers involved [La]. The main obstruction for further progress remained the elusive purity on which all conclusions were conditional.

5.3. **Hitchin fibration.** Ngô recognized that Laumon's approach to affine Springer fibers via compactified Jacobians is a natural piece of the Hitchin fibration. Although it might appear complicated at first glance, the Hitchin fibration is nothing more than the natural generalization of the Chevalley morphism

$$\chi : \mathfrak{g} \longrightarrow \mathfrak{t}/W = \operatorname{Spec} k[\mathfrak{t}]^W.$$

to the setting of algebraic curves. Recall that χ descends to the quotient $\mathfrak{g}/G$, and is also equivariant for $GL(1)$ acting on $\mathfrak{g}$ by linear scaling, and on $(\mathfrak{t}/W)$ by the resulting weighted scaling.

Fix a smooth projective curve C. Although the constructions to follow are very general, for simplicity we will continue to work over the complex numbers $\mathbb{C}$, so in particular C is nothing more than a compact Riemann surface. Let us imagine the Lie algebra $\mathfrak{g}$ varying as a vector bundle along C. In order to preserve its natural structures, we should insist that the transition functions of the vector bundle take values in G acting by the adjoint representation. In other words, the twisting of the vector bundle should be encoded by a principal G-bundle $\mathcal{P}$ and the vector bundle should take the form of the associated bundle

$$\mathfrak{g}_{\mathcal{P}} = \mathcal{P} \times_G \mathfrak{g}.$$

We will write Bun_G for the moduli of principal G-bundles over C

Now suppose we are given a line bundle $\mathcal{L}$ over C. The $GL(1)$-action of linear scaling on $\mathfrak{g}$ together with the line bundle $\mathcal{L}$ provides the option to twist further and form the tensor product

$$\mathfrak{g}_{\mathcal{P},\mathcal{L}} = \mathfrak{g}_{\mathcal{P}} \otimes_{\mathcal{O}_C} \mathcal{L}.$$

Similarly, the $GL(1)$-action of weighted scaling on $\mathfrak{t}/W$ together with the line bundle $\mathcal{L}$ provides an affine bundle

$$(\mathfrak{t}/W)_{\mathcal{L}} = \mathcal{L} \times_{GL(1)} \mathfrak{t}/W.$$

Definition 5.5. Fix a smooth projective curve C equipped with a line bundle $\mathcal{L}$.

The total space $\mathcal{M}_G$ of the Hitchin fibation is the moduli of pairs called Higgs bundles of a principal G-bundle $\mathcal{P}$ over C, and a section of the twisted adjoint bundle

$$\varphi \in \Gamma(C, \mathfrak{g}_{\mathcal{P},\mathcal{L}})$$

called a Higgs field.

The base $\mathcal{A}_G$ of the Hitchin fibration is the space of possible eigenvalues

$$\mathcal{A}_G = \Gamma(C, (\mathfrak{t}/W)_{\mathcal{L}}).$$

The Hitchin fibration is the pointwise unordered eigenvalue map

$$\chi : \mathcal{M}_G \longrightarrow \mathcal{A}_G.$$

The Hitchin fibers are the inverse images $\mathcal{M}_G^a = \chi^{-1}(a)$ for parameters $a \in \mathcal{A}_G$.

Remark 5.6. Hitchin's original construction [Hi87] focused on the case where $\mathcal{L}$ is the canonical line bundle ω_C of one-forms on C. This is most natural from the perspective that then $\mathcal{M}_G$ is the cotangent bundle to the moduli Bun_G of principal G-bundles, and the Hitchin fibration is a complete integrable system. The choice of line bundle $\mathcal{L}$ provides useful technical freedom, since by choosing $\mathcal{L}$ ample enough, we can eliminate any constraint imposed by the global geometry of C.

Remark 5.7. Given a line bundle $\mathcal{L}$ and principal G-bundle $\mathcal{P}$, we can find finitely many points of C so that the restrictions of the bundles to their complement are trivializable. Thus any point of $\mathcal{M}_G$ gives an element of $\mathfrak{g}(F)$, well-defined up to dilation and conjugation, where F is the function field of C. Likewise any point of $\mathcal{A}_G$ gives an element of $(\mathfrak{t}/W)(F)$, well-defined up to dilation.

In this way, we can import abstract definitions for algebraic groups defined over a field F to the setting of the Hitchin fibration. For example, a point of $\mathcal{M}_G$ is said to be regular, semisimple, regular semisimple, or anisotropic if the generic value of its Higgs field is so as an element of $\mathfrak{g}(F)$. Likewise a point of $\mathcal{A}_G$ is said to be generically regular if its generic value is a regular value of $(\mathfrak{t}/W)(F)$.

Example 5.8 (Hitchin fibration for $GL(n)$)**.** Recall that for $\mathfrak{gl}(n) = \mathrm{End}(\mathbb{C}^n)$, the Chevalley morphism is simply the characteristic polynomial

$$\chi : \mathrm{End}(\mathbb{C}^n) \longrightarrow \textstyle\bigoplus_{k=1}^n L_k \qquad \chi(A) = \det(t\,\mathrm{Id} - A)$$

where $L_k \simeq \mathbb{C}$ denotes the one-dimensional vector space generated by the kth elementary symmetric polynomial.

The total space $\mathcal{M}_{GL(n)}$ is the moduli of pairs of a rank n vector bundle $\mathcal{V}$ over C, and a twisted endomorphism

$$\varphi \in \Gamma(C, End_{\mathcal{O}_C}(\mathcal{V}) \otimes_{\mathcal{O}_C} \mathcal{L}).$$

The base $\mathcal{A}_{GL(n)}$ is the space of possible eigenvalues

$$\mathcal{A}_{GL(n)} = \bigoplus_{k=1}^{n} \Gamma(C, L_k \otimes_{\mathcal{O}_C} \mathcal{L}^{\otimes k})).$$

The Hitchin fibration is the pointwise unordered eigenvalue map

$$\chi : \mathcal{M}_{GL(n)} \longrightarrow \mathcal{A}_{GL(n)}.$$

A parameter $a \in \mathcal{A}_{GL(n)}$ assigns to each point $c \in C$ a collection of n unordered eigenvalues. The *spectral curve* C_a is the total space of this varying family over C. More precisely, it is the solution to the equation on the total space of $\mathcal{L}$ given by the characteristic polynomial corresponding to a. When the spectral curve C_a is reduced, the Hitchin fiber is isomorphic to the compactified Picard

$$\mathcal{M}^a_{GL(n)} \simeq \overline{Pic}(C_a)$$

of rank 1, torsion free sheaves on C_a.

Example 5.9 (Hitchin fibration for $SL(2)$)**.** For $\mathfrak{sl}(2) = \{A \in \mathfrak{gl}(2) | \mathrm{Trace}(A) = 0\}$, the Chevalley morphism is simply the determinant map

$$\chi : \mathfrak{sl}(2) \longrightarrow \mathbb{C} \qquad \chi(A) = \det(A)$$

The total space $\mathcal{M}_{SL(2)}$ is the moduli of pairs of a rank 2 vector bundle $\mathcal{V}$ over C with trivialized determinant $\wedge^2\mathcal{V} \simeq \mathcal{O}_C$, and a twisted traceless endomorphism

$$\varphi \in \Gamma(C, End_{\mathcal{O}_C}(\mathcal{V}) \otimes_{\mathcal{O}_C} \mathcal{L}), \qquad \mathrm{Trace}(\varphi) = 0 \in \Gamma(C, \mathcal{L}).$$

The base $\mathcal{A}_{SL(2)}$ is the space of possible determinants

$$\mathcal{A}_{SL(2)} = \Gamma(C, \mathcal{L}^{\otimes 2}).$$

The Hitchin fibration is the pointwise determinant map

$$\chi : \mathcal{M}_{SL(2)} \longrightarrow \mathcal{A}_{SL(2)}.$$

We can regard a parameter $a \in \mathcal{A}_{SL(2)}$ as an element of $\mathcal{A}_{GL(2)}$ and hence assign to it a spectral curve C_a. This will be nothing more than the two-fold cover $c_a : C_a \to C$ given by the equation $t^2 + a$ on the total space of $\mathcal{L}$. When a is not identically zero, C_a is reduced, and the Hitchin fiber is isomorphic to the moduli

$$\mathcal{M}^a_{SL(2)} \simeq \{\mathcal{F} \in \overline{Pic}(C_a) | \det(c_{a*}\mathcal{F}) \simeq \mathcal{O}_C\}.$$

The importance of the Hitchin fibration in gauge theory, low dimensional topology, and geometric representation theory can not be underestimated. Note that the moduli Bun_G of principal G-bundles is a precise analogue of the primary locally symmetric space of automorphic representation theory. Namely, if F denotes the function field of the smooth projective complex curve C, and $\mathbb{A}_F$ its adèles with ring of integers $\mathcal{O}_F$, then the moduli of principal G-bundles is isomorphic to the double coset quotient

$$\mathrm{Bun}_G \simeq G(F)\backslash G(\mathbb{A}_F)/G(\mathcal{O}_F).$$

So in some sense we have come full circle. Starting from questions about number fields, we arrived at p-adic local fields and the Fundamental Lemma. Then we translated the questions to Laurent series local fields, and finally we will appeal to the geometry of function fields.

Now instead of considering affine Springer fibers, Ngô proposes that we consider Hitchin fibers. To spell out more precisely the relation between the two, we will focus on their symmetries.

Recall that to a regular semisimple Kostant element $\xi \in \mathfrak{g}_\mathcal{O}$, we associate a commutative group-scheme J over $\mathcal{O}$, and the affine Grassmannian $\mathcal{P}_\xi = J_\mathcal{K}/J_\mathcal{O}$ naturally acts on the affine Springer fiber Gr_G^ξ.

Similarly, there is a global version of this construction which associates to a generically regular element $a \in \mathcal{A}_G$, a commutative group scheme J over the curve C, and the moduli $\mathcal{P}_a$ of principal J-bundles is a commutative group-stack which naturally acts on the Hitchin fiber $\mathcal{M}_G^a$. Here generically regular means that the value of the section a is regular except at possibly finitely many points of C.

The following result of Ngô is a direct analogue of the adèlic factorization appearing in Formula (2.2).

Theorem 5.10. *Suppose $a \in \mathcal{A}_G$ is generically regular.*

Let $c_i \in C$, for $i \in I$, be the finitely many points where a is not regular, and let $D_i = \operatorname{Spec} \mathcal{O}_i$ be the formal disk around c_i.

Consider the Kostant elements $\xi_i = \sigma(a|_{D_i}) \in \mathfrak{g}_{\mathcal{O}_i}$. Then there is a canonical map inducing a topological equivalence

$$\textstyle\prod_{i\in I} Gr_G^{\xi_i}/\mathcal{P}_{\xi_i} \xrightarrow{\sim} \mathcal{M}_G^a/\mathcal{P}_a.$$

Example 5.11 (Example 5.8 of $GL(n)$ continued)**.** When the spectral curve C_a of the parameter $a \in \mathcal{A}_{GL(n)}$ is reduced, the symmetry group $\mathcal{P}_a$ is precisely the Picard $Pic(C_a)$ of line bundles on C_a. Under the identification of the Hitchin fiber

$$\mathcal{M}_{GL(n)}^a \simeq \overline{Pic}(C_a)$$

with the compactified Picard, the action of $\mathcal{P}_a \simeq Pic(C_a)$ is simply tensor product.

Example 5.12 (Example 5.9 of $SL(2)$ continued)**.** When the parameter $a \in \mathcal{A}_{GL(n)}$ is nonzero so that C_a is reduced, the symmetry group $\mathcal{P}_a$ is the Prym variety given by the kernel of the norm map

$$\mathcal{P}_a \simeq \ker\{N : Pic(C_a) \longrightarrow Pic(C)\}$$

Under the identification of the Hitchin fiber explained above, the action of $\mathcal{P}_a$ is simply tensor product.

¿From the theorem and careful choices of the parameter $a \in \mathcal{A}_G$ and the line bundle $\mathcal{L}$, one can realize the cohomology of affine Springer fibers completely in terms of the cohomology of Hitchin fibers. Now we are in a finite-dimensional setting where the tools of algebraic geometry apply. Without developing any further theory, Laumon and Ngô [LaN04] were able to establish the necessary purity to deduce the Fundamental Lemma for unitary groups.

5.4. **Ngô's Support Theorem.** ¿From our preceding discussion, we conclude that we can replace the study of affine Springer fibers with that of Hitchin fibers. Unlike the somewhat discontinuous behavior of affine Springer fibers, the Hitchin fibration is a highly structured family. To understand its cohomology, Ngô introduces the following general notion of an abelian fibration. What he proves about them will no doubt find much further application. For simplicity, we will continue to work over the complex numbers $\mathbb{C}$, though the notions and results make sense in great generality.

Definition 5.13. A *weak abelian fibration* consists of a base variety S, a projective map $f : M \to S$, and a smooth abelian group-scheme $g : P \to S$, with connected fibers, acting on M. Thus for $s \in S$, the fiber $M_s = f^{-1}(s)$ is a projective variety, the fiber $P_s = f^{-1}(s)$ is a connected abelian group, and we have a P_s-action on M_s.

We require the following properties to hold:

(1) For each $s \in S$, the fibers M_s and P_s have the same dimension.
(2) For each $s \in S$, and $m \in M_s$, the stabilizer $Stab_{P_s}(m) \subset P_s$ is affine.
(3) P has a polarizable Tate module.

Remark 5.14. We will not attempt to explain the third technical condition other than the following brief remark. For each $s \in S$, there is a canonical Chevalley exact sequence

$$\begin{tikzcd} 1 \arrow[r] & R_s \arrow[r] & P_s \arrow[r] & A_s \arrow[r] & 1 \end{tikzcd}$$

where A_s is an abelian variety, and R_s is a connected abelian affine group. If R_s were always trivial, the third condition would assert that P is a polarizable family of abelian varieties.

As the name suggests, the above notion is quite weak. To strengthen it, let us cut up the base variety S into its subvarieties S_δ where the dimension $\delta(s) = \dim(R_s)$ of the affine part of P_s is precisely δ.

Definition 5.15. (1) A smooth connected abelian group scheme $g : P \to S$ is said to be *δ-regular* if it satisfies

$$\operatorname{codim}(S_\delta) \geq \delta.$$

(2) A *δ-regular abelian fibration* is a weak abelian fibration whose group scheme is δ-regular.

Remark 5.16. Given any δ-regular abelian fibration, over the generic locus $S_0 \subset S$, the group-schemes P_s are in fact abelian varieties, act on M_s with finite stabilizers, and hence M_s is a finite union of abelian varieties.

Example 5.17. Here are two good examples of δ-regular abelian fibrations to keep in mind:

(1) Any integrable system (though here we use that we are working over the complex numbers $\mathbb{C}$).

(2) For $X \to S$ a versal deformation of a curve with plane singularities, $M = \overline{Jac}(X/S)$ with its natural action of $P = Jac(X/S)$.

Now we arrive at the main new technical result underlying Ngô's proof of the Fundamental Lemma. It is a refinement of the celebrated Decomposition Theorem of Beilinson-Bernstein-Deligne-Gabber in the setting of abelian fibrations. In citing the Decomposition Theorem, we are invoking the full power of Hodge theory, in particular a very general form of the relative Hard Lefschetz Theorem. Let us recall it in a specific form sufficient for our current discussion.

Theorem 5.18 (Decomposition Theorem). *Let $f : M \to S$ be a projective map of varieties with M smooth. The pushforward $Rf_*\mathbb{C}_M$ is a direct sum of (shifted) intersection cohomology sheaves of local systems on subvarieties of S.*

Example 5.19. (1) Let $f : M \to S$ be a proper fibration with M and S smooth. Then the pushforward $Rf_*\mathbb{C}_M$ is a direct sum of (shifted) semisimple local systems whose fiber at $s \in S$ is the cohomology of the fiber $M_s = f^{-1}(s)$.

(2) Recall the Springer resolution $\mu_{\mathcal{N}} : \tilde{\mathcal{N}} \to \mathcal{N}$ discussed above. The pushforward $R\mu_{\mathcal{N}*}\mathbb{C}_{\tilde{\mathcal{N}}}$ is a direct sum of intersection cohomology sheaves supported on the various nilpotent orbits. For simplicity, let us restrict to $G = GL(n)$. Then the nilpotent orbits $\mathcal{O}_y \subset \mathcal{N}$ are indexed by Young diagrams y. Irreducible representations V_y of the symmetric group Σ_n are also indexed by Young diagrams y. On each orbit $\mathcal{O}_y \subset \mathcal{N}$, there is a local system $\mathcal{L}_y$ of rank $\dim V_y$ such that its intersection cohomology is the contribution to $R\mu_{\mathcal{N}*}\mathbb{C}_{\tilde{\mathcal{N}}}$ from the orbit $\mathcal{O}_y$. This gives a geometric decomposition of the regular representation of Σ_n.

The following support result significantly constrains the supports of the summands occurring in the Decomposition Theorem. Very roughly speaking, it says that in the case of δ-regular abelian fibrations, one can surprisingly calculate the pushforward by studying generic loci.

Theorem 5.20 (Ngô's Support Theorem). *Let $f : M \to S$, $g : P \to S$ be a δ-regular abelian fibration of relative dimension d. Assume the base S is connected and the total space M is smooth.*

Let $\mathcal{F}$ be an intersection cohomology sheaf occurring as a summand in the pushforward $Rf_*\mathbb{C}_M$, and let $Z \subset S$ be the support of $\mathcal{F}$.

Then there exists an open subset $U \subset S$ such that $U \cap Z \neq \emptyset$, and a non-zero local system $\mathcal{L}$ on the intersection $U \cap Z$ such that the tautological extension by zero of $\mathcal{L}$ to all of U is a summand of the restriction $R^{2d}f_*\mathbb{C}_M|_U$.

In particular, if the fibers of f are irreducible, then $Z = S$.

Example 5.21. Here is an example and then a non-example of the kind of phenomenon explained by the support theorem.

(1) Let $f : M \to S$ be a proper flat family of irreducible curves with M and S smooth. Let $S^{reg} \subset S$ be the open locus over which f is smooth, so in particular where the fibers are necessarily smooth. Then the pushforward $Rf_*\mathbb{C}_M$ is the intersection cohomology extension of the (shifted) local system $Rf_*\mathbb{C}_M|_{S^{reg}}$ whose fiber at $s \in S^{reg}$ is the cohomology $H^0 \oplus H^1[-1] \oplus H^2[-2]$ of the curve $M_s = f^{-1}(s)$.

This satisfies the conclusion of the support theorem, though strictly speaking it is not an application of it (except in special instances where the curves are of genus 1).

(2) Consider the proper flat family of irreducible surfaces

$$f : M = \{([x,y,z,w],s) \in \mathbb{P}^3 \times \mathbb{A}^1 | x^3 + y^3 + z^3 + sw^3 = 0\} \longrightarrow S = \mathbb{A}^1$$

with the obvious projection $f([x,y,z,w],s) = s$. It is smooth over $\mathbb{A}^1 \setminus \{0\}$, but singular over $\{0\}$. One can check that the pushforward $Rf_*\mathbb{C}_M$ contains summands which are skyscraper sheaves supported at $\{0\}$.

This does not satisfy the conclusion of the support theorem, though f is a proper flat map.

5.5. **Geometric elliptic endoscopy.** Now let us return to the Hitchin fibration along with its relative symmetry group

$$\chi : \mathcal{M}_G \longrightarrow \mathcal{A}_G \qquad \mathcal{P} \longrightarrow \mathcal{A}_G$$

The results to follow are strikingly parallel to the stabilization of the anisotropic part of the Arthur-Selberg trace formula. Recall that the adèlic factorization of Formula (2.2) expressed the anisotropic terms of the Arthur-Selberg trace formula in terms of local orbital integrals, and ultimately led to the Fundamental Lemma. Here we reverse the process and thanks to the adèlic factorization of Theorem 5.10, calculate the cohomology of affine Springer fibers in terms of anisotropic Hitchin fibers, and in this way ultimately prove the Fundamental Lemma.

Definition 5.22 (Anisotropic Hitchin fibration)**.** The anisotropic Hitchin base $\mathcal{A}_G^{ani} \subset \mathcal{A}_G$ consists of parameters $a \in \mathcal{A}_G$ such that any Higgs bundle $(\mathcal{P}, \varphi) \in \chi^{-1}(a)$ is generically anisotropic in the sense that for any generic trivialization of $\mathcal{P}$ (and the line bundle $\mathcal{L}$), the Higgs field φ is anisotropic as an element of $\mathfrak{g}(F)$, where F is the function field of C.

The anisotropic Hitchin fibration and relative symmetry group is the restriction of the Hitchin fibration and its relative symmetry group to the anisotropic Hitchin base

$$\chi^{ani} : \mathcal{M}_G^{ani} = \mathcal{M}_G \times_{\mathcal{A}_G} \mathcal{A}_G^{ani} \longrightarrow \mathcal{A}_G^{ani} \qquad \mathcal{P}^{ani} = \mathcal{P} \times_{\mathcal{A}_G} \mathcal{A}_G^{ani} \longrightarrow \mathcal{A}_G^{ani}$$

Remark 5.23. An equivalent characterization of the anisotropic Hitchin base $\mathcal{A}_G^{ani}$ is the complement of the image of the Hitchin bases for all Levi subgroups.

Remark 5.24. Since $GL(n)$ has nontrivial split center, it contains no anisotropic tori. Thus the anisotropic Hitchin fibration for $GL(n)$ is empty and all that follows is vacuous. This is not surprising: there is no nontrivial elliptic endoscopy for $GL(n)$. All of its endoscopic subgroups are in fact Levi subgroups.

The reader interested in examples is recommended to look at the end of this section where $SL(2)$ is discussed.

Theorem 5.25. *Over the anisotropic locus, the Hitchin fibration and its relative symmetry group form a δ-regular abelian fibration.*

Remark 5.26. (1) In fact, in general $\mathcal{M}_G^{ani}$ will be a Deligne-Mumford stack, but this does not obstruct any aspect of our discussion.

(2) The theorem is not known in characteristic p, but one can restrict further to where it holds and continue with the calculations to be performed. Then local-global compatibility can be used to extend the calculations further.

To shorten the notation, we will write

$$\mathcal{F}_G = R\chi_*^{ani}\mathbb{C}_{\mathcal{M}_G^{ani}}$$

for the pushforward of the constant sheaf along the Hitchin fibration.

Since cohomology is invariant under isotopy, the natural $\mathcal{P}^{ani}$-action on $\mathcal{F}_G$ factors through the component group $\pi_0(\mathcal{P}^{ani})$. Our aim consists of two steps:

(1) *Fourier theory*: decompose the pushforward $\mathcal{F}_G$ into eigenspaces for the natural $\pi_0(\mathcal{P}^{ani})$-action,
(2) *Endoscopic stabilization*: identify the eigenspaces for nontrivial characters with the invariant eigenspaces for endoscopic groups.

Because we are over the anisotropic locus, $\pi_0(\mathcal{P}^{ani})$ has finite fibers, and so the eigenspace decomposition is discrete.

For the trivial character of $\pi_0(\mathcal{P}^{ani})$, Ngô's Support Theorem gives a striking analogue of the central Theorem 5.1 of Springer theory.

Theorem 5.27 (Stable summand). *The $\pi_0(\mathcal{P}^{ani})$-invariant (shifted) simple perverse summands of the pushforward $\mathcal{F}_G$ are the canonical intersection cohomology extensions to all of $\mathcal{A}_G^{ani}$ of their restrictions to any non-empty open subset.*

Without loss of applicability, we can find an étale base change $\tilde{\mathcal{A}}_G^{ani} \to \mathcal{A}_G^{ani}$ over which there is a surjective homomorphism of relative groups

$$X_*(T) \longrightarrow \pi_0(\mathcal{P}^{ani}).$$

This is convenient in that we can think of a character

$$\kappa : \pi_0(\mathcal{P}^{ani}) \longrightarrow \mathbb{C}^\times$$

as in turn a character of $X_*(T)$, or in other words, an element of $T^\vee$.

On the one hand, we can consider the summand

$$\mathcal{F}_{G,\kappa} \subset \mathcal{F}_G$$

consisting of sections that transform by the character κ under the action of $\pi_0(\mathcal{P}^{ani})$. In particular, when $\kappa = e$ is the trivial character, we denote the stable summand of $\pi_0(\mathcal{P})$-invariants by

$$\mathcal{F}_{G,st} = \mathcal{F}_{G,e}.$$

On the other hand, we can consider endoscopic groups H associated to endoscopic data (κ, ρ_κ) where $\kappa \in T^\vee \subset G^\vee$ and $\rho_\kappa : \pi_1(C, c_0) \to \pi_0(\kappa)$. Though we are now in a global geometric setting, the constructions are completely parallel to those discussed in Section 3.3. The resulting endoscopic groups H are unramified quasi-split group schemes defined over C. It is straightforward to generalize the Hitchin total space, base, and fibration to such group schemes. In particular, the natural transfer map induces a closed immersion of Hitchin bases

$$\nu_H : \tilde{\mathcal{A}}_H^{ani} \hookrightarrow \tilde{\mathcal{A}}_G^{ani}.$$

Theorem 5.28 (Endoscopic summands). *The geometric Fundamental Lemma holds: the κ-eigenspace $\mathcal{F}_{G,\kappa}$ is a direct sum over all κ-endoscopic groups H of their (shifted) stable summands $\nu_{H*}\mathcal{F}_{H,st}$.*

The theorem is proved by an intricate application of Ngô's support theorem. Roughly speaking, it is straightforward to establish an identification over a generic open locus intersecting the Hitchin bases of all relevant endoscopic groups. Then one uses the support theorem to conclude that the identification extends over the entire anisotropic Hitchin base.

The identification of sheaves gives an identification of their fibers and hence an identification of the stable cohomology of Hitchin fibers for H with the corresponding endoscopic cohomology of Hitchin fibers for G. As we have discussed, such identifications imply analogous identifications for affine Springer fibers, hence also for p-adic orbital integrals, and ultimately confirm the Fundamental Lemma.

Example 5.29 (Geometric endoscopy for $SL(2)$)**.** Recall that to a parameter $a \in \mathcal{A}_{SL(2)}$, we can assign the spectral curve C_a given by the equation $t^2 + a$ on the total space of $\mathcal{L}$. As long as a is not identically zero, it is generically regular and C_a is reduced. In this case, we write $\tilde{C}_a$ for the normalization of C_a.

We can cut up $\mathcal{A}_{SL(2)} \setminus \{0\}$ into three natural pieces:

(1) $\mathcal{A}^{st}$: C_a is irreducible, $\tilde{C}_a \to C$ is ramified. Then $\pi_0(\mathcal{P}_a)$ is trivial.
(2) $\mathcal{A}^{U(1)}$: C_a is irreducible, $\tilde{C}_a \to C$ is unramified. Then $\pi_0(\mathcal{P}_a) \simeq \mathbb{Z}/2\mathbb{Z}$.
(3) $\mathcal{A}^{GL(1)}$: C_a is reducible. Then $\pi_0(\mathcal{P}_a) \simeq \mathbb{Z}$.

The anisotropic locus $\mathcal{A}^{ani}_{SL(2)}$ is the union $\mathcal{A}^{st} \sqcup \mathcal{A}^{U(1)}$.

Each component of $\mathcal{A}^{U(1)}$, $\mathcal{A}^{GL(1)}$ is the image of the Hitchin base for an endoscopic $U(1)$, $GL(1)$ respectively.

6. Some further directions

In this section, we briefly list some research directions related to Ngô's proof of the Fundamental Lemma. In particular, we focus on geometric questions, some solved, some open. It goes without saying that the list is idiosyncratic and far from comprehensive.

6.1. **Deeper singularities.**

From a geometric perspective, Ngô's endoscopic description of the cohomology of the anisotropic Hitchin fibration is only a first step. In the spirit of Springer theory, one should study the entire fibration, proceeding from the anisotropic locus to more complicated Higgs bundles.

Chadouard and Laumon [ChLa, ChLaI, ChLaII] have extended Ngô's picture to the locus of regular semisimple Higgs bundles in order to prove Arthur's weighted fundamental lemma. Recall that the anisotropic locus consists of Higgs bundles such that the generic value of the Higgs field is an anisotropic element of $\mathfrak{g}(F)$, where F is the function field of the curve C. A Higgs bundle is generically regular semisimple if the generic value of the Higgs field is a regular semisimple element of $\mathfrak{g}(F)$. The regular semisimple Hitchin fibers are no longer of finite type, but are increasing unions of finite-type schemes of bounded dimension. (One could compare with affine Springer fibers which display analogous behavior.) Chadouard and Laumon develop a beautiful truncation framework, directly inspired by Arthur's truncations of the trace formula, and exhibit the regular semisimple Hitchin fibration as an increasing union of proper maps. The truncations are governed by stability conditions on Higgs bundles, though the eventual contributions to the weighted fundamental lemma are independent of the stability parameters.

6.2. **Transfer for relative trace formulas.** The relative trace formula[4] (as described in [Lap]) is a yet to be fully developed framework, originally pioneered by Jacquet, for understanding period integrals of automorphic forms. It includes a transfer principle depending upon identities analogous to the Fundamental Lemma.

From a geometric perspective, the main object of study in the relative trace formula is a reductive group G equipped with two spherical subgroups $H_1, H_2 \subset G$, and ultimately the two-sided quotient $H_1 \backslash G / H_2$. The basic example is $G = H \times H$ with $H_1 = H_2 = H$ which gives the adjoint quotient G/G and the usual trace formula.

Given two such groups with spherical subgroups $H_1 \times H_2 \subset G, H'_1 \times H'_2 \subset G'$, the transfer principle is predicated on a relation between the $H_1 \times H_2$-invariant functions on G and $H'_1 \times H'_2$-invariant functions on G'. This induces a transfer of invariant distributions generalizing that between groups and their endoscopic groups mediated by the relation of their characteristic polynomials.

Geometric methods have been used by Ngô [N99] and Yun [Y] to prove fundamental lemmas for relative trace formulas. The latter employs direct analogues of Ngô's global proof of the Fundamental Lemma. There are also precursors to further such results coming from traditional Springer theory as in the work of Grinberg [Gr98].

6.3. **Geometric trace formulas.** A striking aspect of Ngô's proof of the Fundamental Lemma is that its final global calculations are direct geometric analogues of the stabilization of the trace formula. Recall that at the end of the day, the Fundamental Lemma is simply a tool in the analysis of the trace formula and its stabilization. With the Geometric Langlands program in mind, this naturally leads to the question: what is the geometric analogue of the trace formula itself?

The Geometric Langlands program, pioneered by Beilinson and Drinfeld, is a fruitful geometric analogue of the theory of automorphic forms and Galois representations. The basic automorphic objects are $\mathcal{D}$-modules on the moduli of principal G-bundles on a curve. The corresponding spectral objects are quasicoherent sheaves on the moduli of flat $G^\vee$-connections on the curve. Langlands's conjectural reciprocity takes the form of a conjectural equivalence of categories between $\mathcal{D}$-modules and quasicoherent sheaves. Broadly understood, the subject forms a substantial industry with motivations from mathematical physics and representation theory.

Frenkel, Langlands, and Ngô [FLN, FN] have taken the first steps towards a trace formula in this setting. They describe the intricate contours of the problem, and make serious progress towards the development of a precise formulation. Stepping back from the challenges of principal bundles on curves, one can ask what kind of math should a geometric trace formula involve? Or slightly more precisely, what kind of object is the character of a group acting on a category? Here algebraic topology provide precise answers involving loop spaces, and Hochschild and cyclic homology. For a realization of these ideas in the context of group actions, and in particular, a connection to Lusztig's character sheaves, see [BZN].

6.4. **Purity of Hitchin fibration.** The decisive geometric input to Ngô's proof of the Fundamental Lemma is the Decomposition Theorem applied to the Hitchin fibration. The purity of the pushforward along the Hitchin fibration leads to the endoscopic decomposition of the cohomology.

[4] I would like to thank Y. Sakellaridis for explaining to me what this world is all about.

If one restricts to equivalence classes of appropriate semistable Higgs bundles, the Hitchin space takes the more concrete form of a quasiprojective variety $\mathcal{M}_{Dol}$. Nonabelian Hodge theory provides a diffeomorphism between $\mathcal{M}_{Dol}$ and a corresponding affine variety $\mathcal{M}_B$ of representations of the fundamental group of the curve. The diffeomorphism is far from an isomorphism: the algebraic structure on $\mathcal{M}_{Dol}$ depends on the algebraic structure of the curve, while $\mathcal{M}_B$ depends only on the fundamental group of the curve. Though $\mathcal{M}_B$ is affine, the fibers of the Hitchin fibration for $\mathcal{M}_{Dol}$ are compact half-dimensional subvarieties.

De Cataldo, Hausel, and Migliorini [dCHM] study the weight and perverse filtrations on the cohomology of $\mathcal{M}_{Dol}$ and $\mathcal{M}_B$ induced by the Hitchin fibration. Their specific results and what they point towards should shed further light on the topological nature of the indispensable purity invoked in the proof of the Fundamental Lemma.

6.5. **Affine Springer theory.** The broad paradigms of Springer theory explain many aspects of Ngô's proof of the Fundamental Lemma. For example, it employs in an essential way the idea that the cohomology of complicated Springer fibers could be recovered from simpler fibers.

But conversely, there are many other aspects of Springer theory which would be worth pursuing in the setting of the Hitchin fibration. For example, Yun [YI, YII, YIII] studies a tamely ramified version of the Hitchin fibration consisting of Higgs bundles equipped with a compatible flag at a point of the curve. This leads to a generalization of the commutative symmetries appearing in the endoscopic decomposition of the cohomology. Namely, the cohomology becomes a module over the double affine Hecke algebra, and other intriguing relations with quantum algebra appear.

It would be interesting to find other important aspects of traditional Springer theory: the role of the nilpotent cone, a resolution of nilpotent Higgs bundles, and the role of the Fourier transform, to name a few.

References

[A97] J. Arthur, The problem of classifying automorphic representations of classical groups, Advances in mathematical sciences: CRMs 25 years (Montreal, PQ, 1994), CRM Proc. Lecture Notes, vol. 11, Amer. Math. Soc., Providence, RI, 1997, pp. 1–12.

[A05] J. Arthur, An introduction to the trace formula. in Harmonic analysis, the trace formula, and Shimura varieties, 1–263, Clay Math. Proc., 4, Amer. Math. Soc., Providence, RI, 2005.

[A09] J. Arthur, Report on the trace formula. Automorphic forms and L-functions I. Global aspects, 1–12, Contemp. Math., 488, Amer. Math. Soc., Providence, RI, 2009.

[BZN] D. Ben-Zvi, D. Nadler, The Character Theory of a Complex Group, arXiv:0904.1247.

[ChLa] P.-H. Chaudouard, G. Laumon, Sur l'homologie des fibres de Springer affines tronquées, arXiv:math/0702586.

[ChLaI] P.-H. Chaudouard, G. Laumon, Le lemme fondamental pondéré I : constructions géométriques, arXiv:math/0902.2684.

[ChLaII] P.-H. Chaudouard, G. Laumon, Le lemme fondamental pondéré. II. Énoncés cohomologiques, arXiv:math/0702586.

[CHL] R. Cluckers, T. Hales, F. Loeser, Transfer Principle for the Fundamental Lemma, arXiv:0712.0708.

[CL05] R. Cluckers, F. Loeser, Ax-Kochen-Ersov Theorems for p-adic integrals and motivic integration, in Geometric methods in algebra and number theory, edited by F. Bogomolov and Y. Tschinkel, Progress in Mathematics 235, 109–137 (2005), Birkhauser.

[CL] R. Cluckers, F. Loeser, Constructible exponential functions, motivic Fourier transform and transfer principle, math.AG/0512022, Ann. of Math., to appear.

[De05] S. DeBacker, The Fundamental Lemma: What is it and what do we know?, Current Developments in Mathematics Volume 2005 (2007), 151–171.

[DL01] J. Denef, F. Loeser, Definable sets, motives and p-adic integrals, J. Amer. Math. Soc. 14 (2001), 429–469.

[D] V. Drinfeld, Informal notes available at

`http://www.math.uchicago.edu/`

[FLN] E. Frenkel, R. Langlands, B. C. Ngô, Formule des Traces et Fonctorialité: le Début d'un Programme, arXiv:1003.4578.

[FN] E. Frenkel, B. C. Ngô, Geometrization of Trace Formulas, arXiv:1004.5323.

[FH91] W. Fulton, J. Harris. Representation theory. A first course. Graduate Texts in Mathematics, 129. Readings in Mathematics. Springer-Verlag, New York, 1991.

[G83] V. Ginsburg, "Intégrales sur les orbites nilpotentes et représentations des groupes de Weyl," C. R. Acad. Sci. Paris ér. I Math. 296 (1983), no. 5, 249–252.

[GKM98] M. Goresky, R. Kottwitz, R. MacPherson, Koszul duality, equivariant cohomology, and the localization theorem. Invent. Math. 131 (1998), 25–83.

[GKM04] M. Goresky, R. Kottwitz, R. MacPherson, Homology of affine Springer fiber in the unramified case. Duke Math. J. 121 (2004) 509–561.

[GKM06] M. Goresky, R. Kottwitz, R. MacPherson, Purity of equivalued affine Springer fibers. Represent. Theory 10 (2006), 130–146.

[Gr98] M. Grinberg, A generalization of Springer theory using nearby cycles. Represent. Theory 2 (1998), 410–431 (electronic).

[H95] T. Hales, On the fundamental lemma for standard endoscopy: reduction to unit elements, Canad. J. Math., 47 (1995) 974–994.

[H05] T. Hales, A statement of the fundamental lemma. Harmonic analysis, the trace formula, and Shimura varieties, 643–658, Clay Math. Proc., 4, Amer. Math. Soc., Providence, RI, 2005.

[H] M. Harris et. al., The Stable Trace Formula, Shimura Varieties, and Arithmetic Applications, book project available at

`http://fa.institut.math.jussieu.fr/node/29`

[dCHM] M. A. de Cataldo, T. Hausel, L. Migliorini, Topology of Hitchin systems and Hodge theory of character varieties, arXiv:1004.1420.

[Hi87] N. Hitchin, Stable bundles and integrable connections. Duke Math. J. 54 (1987) 91–114.

[HoKa84] R. Hotta and M. Kashiwara, "The invariant holonomic system on a semisimple Lie algebra," Invent. Math. 75 (1984), no. 2, 327–358.

[KL88] D. Kazhdan, G. Lusztig, Fixed point varieties on affine flag manifolds. Israel J. Math. 62 (1988), no. 2, 129–168.

[K84] R. Kottwitz, Stable trace formula: cuspidal tempered terms. Duke Math J. 1984 vol. 51 (3) pp. 611–650.

[K86] R. Kottwitz, Stable trace formula: elliptic singular terms. Math. Ann. 275 (1986), no. 3, 365–399.

[L79] R.P. Langlands, Les débuts d'une formule des traces stable, Publications mathématiques de l'Université Paris VII, 1979.

[L80] R. P. Langlands, Base change for GL(2). Annals of Mathematics Studies, 96. Princeton University Press, Princeton, 1980.

[L1] R. P. Langlands, Informal remarks available at

`http://publications.ias.edu/rpl/series.php?series=54`

[L2] R. P. Langlands, Informal remarks available at

`http://publications.ias.edu/rpl/series.php?series=56`

[LS87] R. Langlands and D. Shelstad, On the definition of transfer factors, Math. Ann. 278 (1987), 219–271.

[Lap] E. Lapid, The relative trace formula and its applications, Automorphic Forms and Automorphic L-Functions (Kyoto, 2005), Surikaisekikenkyusho Kokyuroku No. 1468 (2006), 76–87.

[La1] G. Laumon, The Fundamental Lemma for Unitary Groups, lecture at Clay Math. Inst., available at

`http://www.claymath.org/research_award/Laumon-Ngo/laumon.pdf`

[La2] G. Laumon, Fundamental Lemma and Hitchin Fibration, lecture at Newton Inst., available at
`http://www.newton.ac.uk/programmes/ALT/seminars/051316301.pdf`

[La06] G. Laumon, Fibres de Springer et Jacobiennes compactifiées in Algebraic geometry and number theory, 515–563, Progr. Math., 253, Birkhauser Boston, Boston, MA, 2006.

[La] G. Laumon, Sur le lemme fondamental pour les groupes unitaires, arXiv:math/0212245.

[LaN04] G. Laumon and B. C. Ngô, Le lemme fondamental pour les groupes unitaires, arXiv:math/0404454v2.

[M08] S. Morel, Étude de la cohomologie de certaines varietes de Shimura non compactes, arXiv:0802.4451.

[N99] Ngô Bao Châu, Le lemme fondamental de Jacquet et Ye en caractéristique positive. Duke Math. J. 96 (1999), no. 3, 473–520.

[N08] Ngô Bao Châu, Le lemme fondamental pour les algebres de Lie, arXiv:0801.0446.

[R90] J. Rogawski, Automorphic representations of unitary groups in three variables, Annals of Mathematics Studies, vol. 123, Princeton University Press, Princeton, NJ, 1990.

[S77] J.-P. Serre. Linear representations of finite groups. Translated from the second French edition by Leonard L. Scott. Graduate Texts in Mathematics, Vol. 42. Springer-Verlag, New York-Heidelberg, 1977.

[S82] D. Shelstad, L-indistinguishability for real groups. Math. Ann. 259 (1982) 385–430.

[S] S.-W. Shin, Galois representations arising from some compact Shimura varieties, to appear in Annals of Math.

[W91] J.-L. Waldspurger, Sur les intégrales orbitales tordues pour les groupes linéaires: un lemme fondamental. Can. J. Math. 43 (1991) 852–896.

[W97] J.-L. Waldspurger, Le lemme fondamental implique le transfert, Comp. Math. 105 (1997), no. 2, 153–236.

[W06] J.-L. Waldspurger, Endoscopie et changement de caractéristique, J. Inst. Math. Jussieu 5 (2006), no. 3, 423–525.

[W08] J.-L. Waldspurger, L'endoscopie tordue n'est pas si tordue, Memoirs of AMS 194 (2008), no. 908.

[W09] J.-L. Waldspurger, À propos du lemme fondamental pondéré tordu, Math. Ann. 343 (2009), no. 1, 103–174.

[YI] Z. Yun, Towards a Global Springer Theory I: The affine Weyl group action, arXiv:0810.2146.

[Y] Z. Yun, The fundamental lemma of Jacquet-Rallis in positive characteristics, arXiv:0901.0900.

[YII] Z. Yun, Towards a Global Springer Theory II: the double affine action, arXiv:0904.3371.

[YIII] Z. Yun, Towards a Global Springer Theory III: Endoscopy and Langlands duality, arXiv:0904.3372.

Department of Mathematics, Northwestern University, Evanston, IL 60208-2370
E-mail address: `nadler@math.northwestern.edu`

CURRENT EVENTS BULLETIN
Previous speakers and titles

For PDF files of talks, and links to *Bulletin of the AMS* articles, see http://www.ams.org/ams/current-events-bulletin.html.

January 15, 2010 (San Francisco, CA)

Ben Green, University of Cambridge
Approximate groups and their applications: work of Bourgain, Gamburd, Helfgott and Sarnak

David Wagner, University of Waterloo
Multivariate stable polynomials: theory and applications

Laura DeMarco, University of Illinois at Chicago
The conformal geometry of billiards

Michael Hopkins, Harvard University
On the Kervaire Invariant Problem

January 7, 2009 (Washington, DC)

Matthew James Emerton, Northwestern University
Topology, representation theory and arithmetic: Three-manifolds and the Langlands program

Olga Holtz, University of California, Berkeley
Compressive sensing: A paradigm shift in signal processing

Michael Hutchings, University of California, Berkeley
From Seiberg-Witten theory to closed orbits of vector fields: Taubes's proof of the Weinstein conjecture

Frank Sottile, Texas A & M University
Frontiers of reality in Schubert calculus

January 8, 2008 (San Diego, California)

Günther Uhlmann, University of Washington
Invisibility

Antonella Grassi, University of Pennsylvania
Birational Geometry: Old and New

Gregory F. Lawler, University of Chicago
Conformal Invariance and 2-d Statistical Physics

Terence C. Tao, University of California, Los Angeles
Why are Solitons Stable?

January 7, 2007 (New Orleans, Louisiana)

Robert Ghrist, University of Illinois, Urbana-Champaign
Barcodes: The persistent topology of data

Akshay Venkatesh, Courant Institute, New York University
Flows on the space of lattices: work of Einsiedler, Katok and Lindenstrauss

Izabella Laba, University of British Columbia
From harmonic analysis to arithmetic combinatorics

Barry Mazur, Harvard University
The structure of error terms in number theory and an introduction to the Sato-Tate Conjecture

January 14, 2006 (San Antonio, Texas)

Lauren Ancel Myers, University of Texas at Austin
Contact network epidemiology: Bond percolation applied to infectious disease prediction and control

Kannan Soundararajan, University of Michigan, Ann Arbor
Small gaps between prime numbers

Madhu Sudan, MIT
Probabilistically checkable proofs

Martin Golubitsky, University of Houston
Symmetry in neuroscience

January 7, 2005 (Atlanta, Georgia)

Bryna Kra, Northwestern University
The Green-Tao Theorem on primes in arithmetic progression: A dynamical point of view

Robert McEliece, California Institute of Technology
Achieving the Shannon Limit: A progress report

Dusa McDuff, SUNY at Stony Brook
Floer theory and low dimensional topology

Jerrold Marsden, Shane Ross, California Institute of Technology
New methods in celestial mechanics and mission design

László Lovász, Microsoft Corporation
Graph minors and the proof of Wagner's Conjecture

January 9, 2004 (Phoenix, Arizona)

Margaret H. Wright, Courant Institute of Mathematical Sciences, New York University
The interior-point revolution in optimization: History, recent developments and lasting consequences

Thomas C. Hales, University of Pittsburgh
What is motivic integration?

Andrew Granville, Université de Montréal
It is easy to determine whether or not a given integer is prime

John W. Morgan, Columbia University
Perelman's recent work on the classification of 3-manifolds

January 17, 2003 (Baltimore, Maryland)

Michael J. Hopkins, MIT
Homotopy theory of schemes

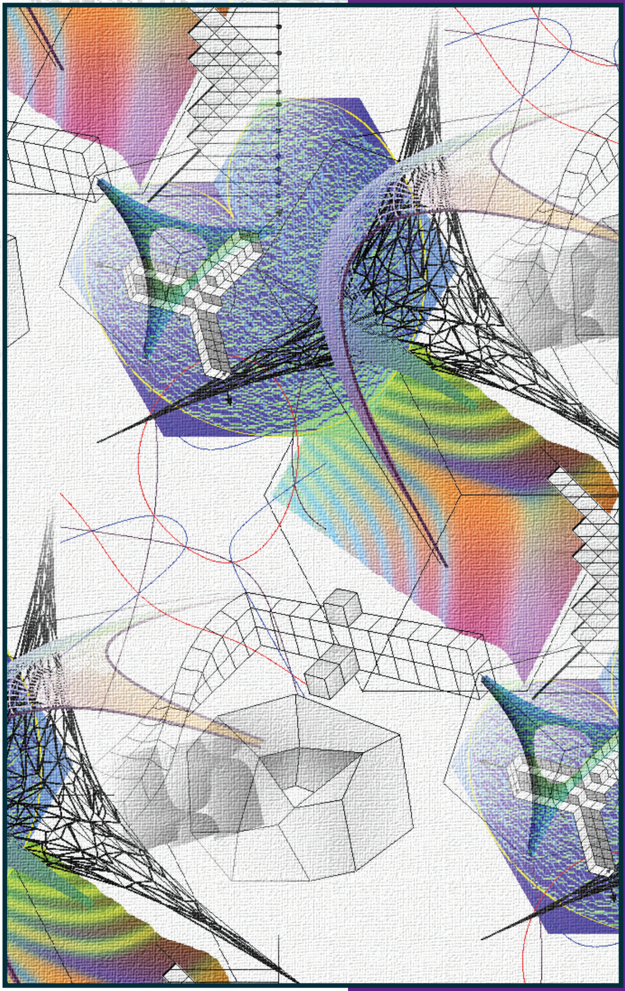
Ingrid Daubechies, Princeton University
Sublinear algorithms for sparse approximations with excellent odds

Edward Frenkel, University of California, Berkeley
Recent advances in the Langlands Program

Daniel Tataru, University of California, Berkeley
The wave maps equation

2011 CURRENT EVENTS BULLETIN *Committee*

David Eisenbud, *University of California, Berkeley, Chair*
David Aldous, *University of California, Berkeley*
David Auckly, *Mathematical Sciences Research Institute*
Hélène Barcelo, *Mathematical Sciences Research Institute*
Laura DeMarco, *University of Illinois at Chicago*
Matthew Emerton, *Northwestern University*
Susan Friedlander, *University of Southern California*
Robert Ghrist, *University of Pennsylvania*
Andrew Granville, *University of Montreal*
Ben Green, *Cambridge University*
Michael Hopkins, *Harvard University*
László Lovász, *Eotvos Lorand University*
John Morgan, *State University of NewYork, Stony Brook*
George Papanicolaou, *Stanford University*
Frank Sottile, *Texas A & M University*
Richard Taylor, *Harvard University*
Akshay Venkatesh, *Stanford University*
David Wagner, *University of Waterloo*
Avi Wigderson, *Institute for Advanced Study*
Carol Wood, *Wesleyan University*
Andrei Zelevinsky, *Northeastern University*



Cover graphic associated with Nadler's talk courtesy of David Jordan.

Cover graphic associated with Scanlon's talk by C. Goodman-Strauss from John H. Conway, Heidi Burgiel, and C. Goodman-Strauss. *The Symmetries of Things*. A K Peters, Ltd., 2008.

Cover graphic associated with Tillmann's talk courtesy of Ulrike Tillmann.

Cover graphic associated with Trevisan's talk courtesy of Luca Trevisan.

The back cover graphic is reprinted courtesy of Andrei Okounkov.