American Mathematical Society

CURRENT EVENTS BULLETIN

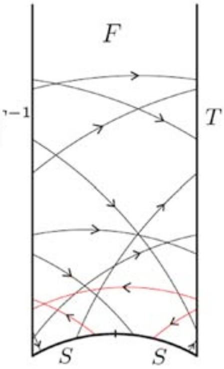
Sunday, January 7, 2007 • 1:00 – 5:00 PM

Joint Mathematics Meetings, New Orleans

1:00 PM

Robert Ghrist

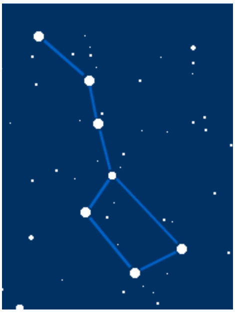
Barcodes: The Persistent Topology of Data



2:00 PM

Akshay Venkatesh

Flows on the Space of Lattices: Work of Einsiedler, Katok and Lindenstrauss



3:00 PM

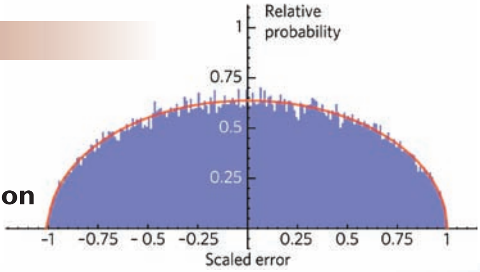
Izabella Laba

From Harmonic Analysis to Arithmetic Combinatorics

4:00 PM

Barry Mazur

The Structure of Error Terms in Number Theory and an Introduction to the Sato-Tate Conjecture



Organized by David Eisenbud, Mathematical Sciences Research Institute

Introduction

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
Mathematical Sciences Research Institute
de@msri.org

For the PDF files of the talks given in other 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.

BARCODES: THE PERSISTENT TOPOLOGY OF DATA

ROBERT GHRIST

Abstract. This article surveys recent work of Carlsson and collaborators on applications of computational algebraic topology to problems of feature detection and shape recognition in high-dimensional data. The primary mathematical tool considered is a homology theory for point-cloud data sets — **persistent homology** — and a novel representation of this algebraic characterization — **barcodes**. We sketch an application of these techniques to the classification of natural images.

1. The shape of data

When a topologist is asked, "How do you visualize a four-dimensional object?" the appropriate response is a Socratic rejoinder: "How do you visualize a three-dimensional object?" We do not see in three spatial dimensions directly, but rather via sequences of planar projections integrated in a manner that is sensed if not comprehended. We spend a significant portion of our first year of life learning how to infer three-dimensional spatial data from paired planar projections. Years of practice have tuned a remarkable ability to extract global structure from representations in a strictly lower dimension.

The inference of global structure occurs on much finer scales as well, with regards to converting discrete data into continuous images. Dot-matrix printers, scrolling LED tickers, televisions, and computer displays all communicate images via arrays of discrete points which are integrated into coherent, global objects. This also is a skill we have practiced from childhood. No adult does a dot-to-dot puzzle with anything approaching anticipation.

1.1. **Topological data analysis.** Problems of data analysis share many features with these two fundamental integration tasks: (1) how does one infer high dimensional structure from low dimensional representations; and (2) how does one assemble discrete points into global structure.

The principal themes of this survey of the work of Carlsson, de Silva, Edelsbrunner, Harer, Zomorodian, and others are the following:

1. It is beneficial to replace a set of data points with a family of **simplicial complexes**, indexed by a proximity parameter. This converts the data set into global topological objects.
2. It is beneficial to view these topological complexes through the lens of algebraic topology — specifically, via a novel theory of **persistent homology** adapted to parameterized families.
3. It is beneficial to encode the persistent homology of a data set in the form of a parameterized version of a Betti number: a **barcode**.

The author gratefully acknowledges the support of DARPA # HR0011-07-1-0002. The work reviewed in this article is funded by the DARPA program *TDA: Topological Data Analysis*.

This review will introduce these themes and survey an example of these techniques applied to a high-dimensional data set derived from natural images.

1.2. **Clouds of data.** Very often, data is represented as an unordered sequence of points in a Euclidean n-dimensional space $\mathbb{E}^n$. Data coming from an array of sensor readings in an engineering testbed, from questionnaire responses in a psychology experiment, or from population sizes in a complex ecosystem all reside in a space of potentially high dimension. The global 'shape' of the data may often provide important information about the underlying phenomena which the data represents.

One type of data set for which global features are present and significant is the so-called **point cloud data** coming from physical objects in 3-d. Touch probes, point lasers, or line lasers sweep a suspended body and sample the surface, recording coordinates of anchor points on the surface of the body. The cloud of such points can be quickly obtained and used in a computer representation of the object. A temporal version of this situation is to be found in motion-capture data, where geometric points are recorded as time series. In both of these settings, it is important to identify and recognize global features: where is the index finger, the keyhole, the fracture?

FIGURE 1. Determining the global structure of a noisy point cloud is not difficult when the points are in $\mathbb{E}^2$, but for clouds in higher dimensions, a planar projection is not always easy to decipher.

Following common usage, we denote by point cloud data any collection of points in $\mathbb{E}^n$, though the connotation is that of a (perhaps noisy) sample of points on a lower-dimensional subset. For point clouds residing in a low-dimensional ambient space, there are numerous approaches for inferring features based on planar projections: reconstruction techniques in the computer graphics and statistics literatures are manifold. From a naive point of view, planar projections would appear to be of limited value in the context of data which is inherently high dimensional or sufficiently 'twisted' so as to preclude a faithful planar projection (Figure 1[right]).

A more global and intrinsic approach to high-dimensional data clouds has recently appeared in the work of Carlsson and collaborators. This body of ideas applies tools from algebraic topology to extract coarse features from high-dimensional data sets. This survey is a brief overview of some of their work. As a result of our focus on techniques from algebraic topology, we neglect the large body of relevant work in nonlinear statistics (which is rarely topological) and in computer graphics (which is rarely high-dimensional).

1.3. **From clouds to complexes.** The most obvious way to convert a collection of points $\{x_\alpha\}$ in a metric space into a global object is to use the point cloud as the vertices of a combinatorial graph whose edges are determined by proximity (vertices within some specified distance ϵ). Such a graph, while capturing connectivity data, ignores a wealth of higher order features beyond clustering. These features can be accurately discerned by thinking of the graph as a scaffold for a higher-dimensional object. Specifically, one completes the graph to a **simplicial complex** — a space built from simple pieces (simplicies) identified combinatorially along faces. The choice of how to fill in the higher dimensional simplices of the proximity graph allows for different global representations. Two of the most natural methods for so doing are as follow:

Definition 1.1. Given a collection of points $\{x_\alpha\}$ in Euclidean space $\mathbb{E}^n$, the **Čech complex**[1], $\mathcal{C}_\epsilon$, is the abstract simplicial complex whose k-simplices are determined by unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ whose closed $\epsilon/2$-ball neighborhoods have a point of common intersection.

Definition 1.2. Given a collection of points $\{x_\alpha\}$ in Euclidean space $\mathbb{E}^n$, the **Rips complex**,[2] $\mathcal{R}_\epsilon$, is the abstract simplicial complex whose k-simplices correspond to unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ which are pairwise within distance ϵ.

The **Čech theorem** (or, equivalently, the "nerve theorem") states that $\mathcal{C}_\epsilon$ has the homotopy type of the union of closed radius $\epsilon/2$ balls about the point set $\{x_\alpha\}$. This means that $\mathcal{C}$, though an abstract simplicial complex of potentially high dimension, behaves exactly like a subset of $\mathbb{E}^n$ (see Figure 2). The Čech complex is a topologically faithful simplicial model for the topology of a point cloud fattened by balls. However, the Čech complex and various topologically equivalent subcomplexes (e.g., the **alpha complex** of [13]) are delicate objects to compute, relying on the precise distances between the nodes in $\mathbb{E}^n$.

From a computational point of view, the Rips complex is less expensive that the corresponding Čech complex, even though the Rips complex has more simplices (in general). The reason is that the Rips complex is a **flag complex**: it is maximal among all simplicial complexes with the given 1-skeleton. Thus, the combinatorics of the 1-skeleton completely determines the complex, and the Rips complex can be stored as a graph and reconstituted instead of storing the entire boundary operator

[1] Also known as the **nerve**.

[2] A more appropriate name would be the Vietoris-Rips complex, in recognition of Vietoris' original use of these objects in the early days of homology theory [21]. For brevity we use the term "Rips complex."

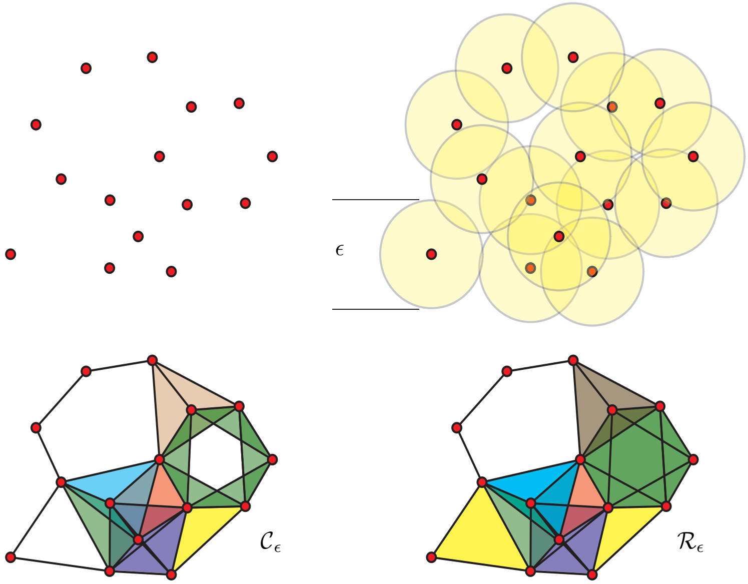


FIGURE 2. A fixed set of points [upper left] can be completed to a a Čech complex $\mathcal{C}_\epsilon$ [lower left] or to a Rips complex $\mathcal{R}_\epsilon$ [lower right] based on a proximity parameter ϵ [upper right]. This Čech complex has the homotopy type of the $\epsilon/2$ cover $(S^1 \vee S^1 \vee S^1)$, while the Rips complex has a wholly different homotopy type $(S^1 \vee S^2)$.

needed for a Čech complex. This virtue — that coarse proximity data on pairs of nodes determines the Rips complex — is not without cost. The penalty for this simplicity is that it is not immediately clear what is encoded in the homotopy type of $\mathcal{R}$. In general, it is neither a subcomplex of $\mathbb{E}^n$ nor does it necessarily behave like an n-dimensional space at all (Figure 2).

1.4. **Which ϵ?** Converting a point cloud data set into a global complex (whether Rips, Čech, or other) requires a choice of parameter ϵ. For ϵ sufficiently small, the complex is a discrete set; for ϵ sufficiently large, the complex is a single high-dimensional simplex. Is there an optimal choice for ϵ which best captures the topology of the data set? Consider the point cloud data set and a sequence of Rips complexes as illustrated in Figure 3. This point cloud is a sampling of points on a planar annulus. Can this be deduced? From the figure, it certainly appears as though an ideal choice of ϵ, if it exists, is rare: by the time ϵ is increased so as to remove small holes from within the annulus, the large hole distinguishing the annulus from the disk is filled in.

2. Algebraic topology for data

Algebraic topology offers a mature set of tools for counting and collating holes and other topological features in spaces and maps between them. In the context of

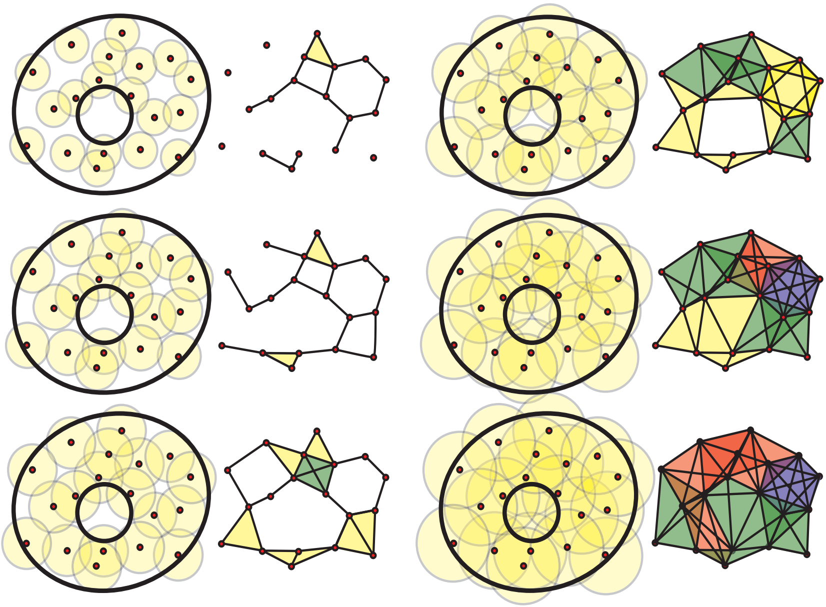


FIGURE 3. A sequence of Rips complexes for a point cloud data set representing an annulus. Upon increasing ϵ, holes appear and disappear. Which holes are real and which are noise?

high-dimensional data, algebraic topology works like a telescope, revealing objects and features not visible to the naked eye. In what follows, we concentrate on homology for its balance between ease of computation and topological resolution. We assume a rudimentary knowledge of homology, as is to be found in, say, Chapter 2 of [15].

Despite being both computable and insightful, the homology of a complex associated to a point cloud at a particular ϵ is insufficient: it is a mistake to ask which value of ϵ is optimal. Nor does it suffice to know a simple 'count' of the number and types of holes appearing at each parameter value ϵ. Betti numbers are not enough. One requires a means of declaring which holes are essential and which can be safely ignored. The standard topological constructs of homology and homotopy offer no such slack in their strident rigidity: a hole is a hole no matter how fragile or fine.

2.1. **Persistence.** Persistence, as introduced by Edelsbrunner, Letscher, and Zomorodian [12] and refined by Carlsson and Zomorodian [22], is a rigorous response to this problem. Given a parameterized family of spaces, those topological features which persist over a significant parameter range are to be considered as signal with short-lived features as noise. For a concrete example, assume that $\mathsf{R} = (\mathcal{R}_i)_1^N$ is a sequence of Rips complexes associated to a fixed point cloud for an increasing sequence of parameter values $(\epsilon_i)_1^N$. There are natural inclusion maps

$$\mathcal{R}_1 \overset{\iota}{\hookrightarrow} \mathcal{R}_2 \overset{\iota}{\hookrightarrow} \cdots \overset{\iota}{\hookrightarrow} \mathcal{R}_{N-1} \overset{\iota}{\hookrightarrow} \mathcal{R}_N \tag{2.1}$$

Instead of examining the homology of the individual terms $\mathcal{R}_i$, one examines the homology of the iterated inclusions $\iota : H_*\mathcal{R}_i \to H_*\mathcal{R}_j$ for all $i < j$. These maps reveal which features persist.

As a simple example, persistence explains why Rips complexes are an acceptable approximation to Čech complexes. Although no single Rips complex is an especially faithful approximation to a single Čech complex, pairs of Rips complexes 'squeeze' the appropriate Čech complex into a manageable hole.

Lemma 2.1 (de Silva [10]). *For any $\epsilon > 0$, there is a chain of inclusion maps*

$$\mathcal{R}_\epsilon \hookrightarrow \mathcal{C}_{\epsilon\sqrt{2}} \hookrightarrow \mathcal{R}_{\epsilon\sqrt{2}}. \tag{2.2}$$

(See [10] for the tight dimension-dependent bound smaller than $\sqrt{2}$.) This implies that any topological feature which persists under the inclusion $\mathcal{R}_\epsilon \hookrightarrow \mathcal{R}_{\epsilon'}$ is in fact a topological feature of the Čech complex $\mathcal{C}_{\epsilon'}$ when $\epsilon'/\epsilon \geq \sqrt{2}$. *Moral:* the homology of the inclusion $\iota_* : H_*\mathcal{R}_\epsilon \to H_*\mathcal{R}_{\epsilon'}$ reveals information that is not visible from $H_*\mathcal{R}_\epsilon$ and $H_*\mathcal{R}_{\epsilon'}$ unadorned. This is a foreshadowing of the broader idea of persistence arising in an arbitrary sequence of chain complexes.

2.2. **Persistent homology.** One begins with a **persistence complex**: a sequence of chain complexes $\mathsf{C} = (C_*^i)_i$ together with chain maps $x : C_*^i \longrightarrow C_*^{i+1}$. (For notational simplicity, we do not index the chain maps x.) This is motivated by having a sequence of Rips or Čech complexes of increasing ϵ sampled at an increasing sequence of parameters $\{\epsilon_i\}$. Since Rips or Čech complexes grow with ϵ, the chain maps x are naturally identified with inclusions.

Definition 2.2. For $i < j$, the (i,j)-persistent homology of C, denoted $H_*^{i\to j}(\mathsf{C})$ is defined to be the image of the induced homomorphism $x_* : H_*(C_*^i) \to H_*(C_*^j)$.

As an example, consider the filtration $\mathsf{R} = (\mathcal{R}_i)$ of Rips complexes parameterized by proximities ϵ_i. Lemma 2.1 implies that if $\epsilon_j/\epsilon_i \geq \sqrt{2}$, then $H_k^{i\to j}(\mathsf{R}) \neq 0$ implies $H_k(\mathcal{C}_{\epsilon_j}) \neq 0$. Holes in the Čech complex are inferred by the persistent homology of the Rips filtration.

There is a good deal more algebraic structure in the interleaving of persistent homology groups, as explained in the work of Carlsson and Zomorodian. Fix a PID of coefficients R and place a graded $R[x]$-module structure on C with x acting as a shift map. That is, a unit monomial $x^n \in R[x]$ sends C_*^i to C_*^{i+n} via n applications of x. One assumes a finite-type condition that each C_*^i is finitely generated as an $R[x]$-module and that the sequence stabilizes in i (in the case of an infinite sequence of chain complexes).

As the filtering of C is via chain maps x (*cf.* the setting of Rips complexes — simplices are added but never removed as ϵ increases), C is free as an $R[x]$-module. The resulting homology $H_*(\mathsf{C})$ retains the structure of an $R[x]$-module, but, unlike the chain module, is not necessarily free. Nor is it easily classified: the Artin-Rees theory from commutative algebra implies that the problem of classifying (finite-type) persistence modules such as $H_*(\mathsf{C})$ is equivalent to classifying finitely-generated non-negatively graded $R[x]$-modules. This is known to be very difficult in, say, $\mathbb{Z}[x]$.

However, for coefficients in a field F, the classification of $F[x]$-modules follows from the Structure Theorem for PID's, since the only graded ideals of $F[x]$ are of the form $x^n \cdot F[x]$. This implies the following:

Theorem 2.3 ([22])**.** *For a finite persistence module* C *with field* F *coefficients,*

$$H_*(\mathsf{C};F) \cong \bigoplus_i x^{t_i}\cdot F[x] \;\oplus\; \left(\bigoplus_j x^{r_j}\cdot (F[x]/(x^{s_j}\cdot F[x]))\right). \tag{2.3}$$

This classification theorem has a natural interpretation. The free portions of Equation (2.3) are in bijective correspondence with those homology generators which come into existence at parameter t_i and which persist for all future parameter values. The torsional elements correspond to those homology generators which appear at parameter r_j and disappear at parameter $r_j + s_j$. At the chain level, the Structure Theorem provides a birth-death pairing of generators of C (excepting those that persist to infinity).

2.3. **Barcodes.** The parameter intervals arising from the basis for $H_*(\mathsf{C};F)$ in Equation (2.3) inspire a visual snapshot of $H_k(\mathsf{C};F)$ in the form of a **barcode**. A barcode is a graphical representation of $H_k(\mathsf{C};F)$ as a collection of horizontal line segments in a plane whose horizontal axis corresponds to the parameter and whose vertical axis represents an (arbitrary) ordering of homology generators. Figure 4 gives an example of barcode representations of the homology of the sampling of points in an annulus from Figure 3 (illustrated in the case of a large number of parameter values ϵ_i).

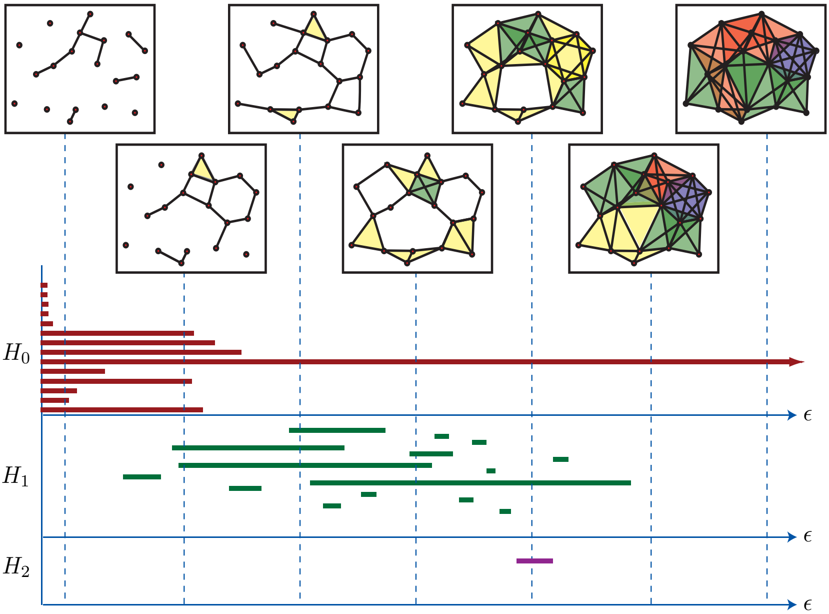


FIGURE 4. [bottom] An example of the barcodes for $H_*(\mathsf{R})$ in the example of Figure 3. [top] The rank of $H_k(\mathcal{R}_{\epsilon_i})$ equals the number of intervals in the barcode for $H_k(\mathsf{R})$ intersecting the (dashed) line $\epsilon = \epsilon_i$.

Theorem 2.3 yields the fundamental characterization of barcodes:

Theorem 2.4 ([22]). *The rank of the persistent homology group $H_k^{i\to j}(\mathsf{C};F)$ is equal to the number of intervals in the barcode of $H_k(\mathsf{C};F)$ spanning the parameter interval $[i,j]$. In particular, $H_*(C_*^i;F)$ is equal to the number of intervals which contain i.*

A barcode is best thought of as the persistence analogue of a Betti number. Recall that the k^{th} Betti number of a complex, $\beta_k := \text{rank}H_k$, acts as a coarse numerical measure of H_k. As with β_k, the barcode for H_k does not give any information about the finer structure of the homology, but merely a continuously parameterized rank. The genius of a barcode representation is the ability to qualitatively filter out topological noise and capture significant features. Indeed, as shown in [7], barcodes are stable in the presence of noise added to a [Morse] filtration. For example, in Figure 4, one sees (from a *very* coarse sampling) that the point cloud likely represents a connected object with one or two significant 'holes' as measured by H_1 and no significant higher homology.

2.4. **Computation.** Most invariants in modern algebraic topology are not known for their ease of computation. Homology (in its simplest manifestations) appears exceptional in that the invariants arise as quotients of finite-dimensional vector spaces. In the context of applications, 'finite' may exceed reasonable bounds. There is no recourse to chanting *"Homology is just linear algebra,"* when faced with millions of simplices: one needs good algorithms. Fortunately, such exist with increasing scope and speed. The text [16] gives a comprehensive introduction to issues of and algorithms available for computing homology for realistic problems in application domains.

More fortunate still, there is an excellent algorithm available for the computation of persistent homology groups and barcodes. The algorithm takes as its argument the filtered simplicial complex consisting of pairs (σ_i, τ_i), where σ_i is a simplex and τ_i is the time at which that simplex appears in the filtration. This algorithm first appears in the paper of Edelsbruner, Letscher, and Zomorodian [12] for simplicial subcomplexes of $\mathbb{E}^3$ with $\mathbb{Z}_2$ coefficients and in that of Carlsson and Zomorodian [22] for general persistence complexes with field coefficients. The Matlab-based front end `Plex` by de Silva and Perry [11] incorporates the C++ persistent homology library of Kettner and Zomorodian with tools for inputting and manipulating simplicial complexes.

It is worth noting that for chain filtrations arising from realistic data sets, the Rips complexes are of an unmanageable size. This necessitates efficient sampling or reduction of the complex with accurate topology. The **witness complex** of de Silva [8, 9, 14] is one solution to this problem.

2.5. **Other directions.** We note that the above is the briefest of treatments of what quickly becomes a fascinating and very active sub-topic of computational topology. For those interested in the algebraic-topological aspects of the theory, we note the following recent developments:

- There are other filtrations besides those associated to Čech or Rips complexes which are natural settings in which to contemplate persistence. The **Morse filtration** of a space X outfitted with $f : X \to \mathbb{R}$ is a filtration of X by excursion sets $X_t = \left\{f^{-1}\left((-\infty, t]\right)\right\}$. This (or a discretized version thereof) is one commonly investigated setting [1, 7], as is filtration by means of curvature data [5].

- Our discussion of persistence is couched in the setting of chain complexes indexed by a single parameter. There are strong motivations for wanting to treat multi-parameter families of complexes. However, there are fundamental algebraic difficult in constructing an analogous theory of persistence modules in this setting [24].
- The computation of persistent relative homology is more subtle, since the ensuing parameterized chain complex C is no longer free as an $F[x]$-module. Bendich and Harer [in progress] have developed an algebraic construction for defining and computing persistent homology which has a particularly clean form in the setting of a Morse filtration. The analogue of Theorem 2.3 provides a perfect pairing of Morse critical points.
- The computation of persistent cohomology is not straightforward. As shown by de Silva [in progress], if you take the graded free $F[x]$-module chain complex C for the Morse filtration X_t of a space X, and dualize it as a graded free $F[x]$-module, i.e. if you construct $\mathrm{Hom}_{F[x]}(\mathsf{C}, F[x])$, then the homology of the resulting object as a graded $F[x]$-module is *not* the persistent cohomology of $H^*(X_t)$, but rather that of the relative cohomology $H^*(X, X_t)$. Computing absolute persistent cohomology necessitates a recourse to duality and the theory of Bendich-Harer above.

3. Example: natural images

One recent example of discovering topological structure in a high-dimensional data set comes from **natural images**. A collection of 4167 digital photographs of random outdoor scenes was assembled in the late 1990s by van Hateren and van der Schaaf [20]. Mumford and others have posed several fascinating questions about the structure and potential universality of the statistics of this and similar sets of images in the context of visual perception [17].

3.1. **"Round about the cauldron go".**

Mumford, Lee, and Pederson [18] construct a data set by choosing at random 5000 three-pixel by three-pixel squares within each digital image and retaining the top 20% of these with respect to contrast. Each such square is a matrix of grey-scale intensities. The full data set consists of roughly 8,000,000 points in $\mathbb{E}^9$. By normalizing with respect to mean intensity and restricting attention to high-contrast images (those away from the origin), the data set is projected to a set of points $\mathcal{M}$ on a topological seven-sphere $S^7 \subset \mathbb{E}^8$. The details of this data set construction requires a choice of natural basis with respect to a particular norm for values of contrast patches. We refer the interested reader to [18] for details.

3.2. **"Hover through the fog".**

So coarse a reduction of natural images (into three-by-three squares of grey-scale intensities) still leads to a point cloud of too high a dimension to visualize. Worse still, what structure is there is blurred and foggy: points appear at first to be distributed over the entire S^7. A resort to density considerations is thus in order. The subject of density filtration is a well-trod area of statistics: see, e.g., [19].

A **codensity** function is used in [3] as follows. Fix a positive integer $k > 0$. For any point x_α in the data set, define $\delta_k(x_\alpha)$ as the distance in $\mathbb{E}^n$ from x_α to k^{th} nearest neighbor of x_α in the data set. For a fixed value of k, δ_k is a positive distribution over the point cloud which measures the radius of the ball needed to enclose k neighbors. Values of δ_k are thus inversely related to the point cloud density. The larger a value of k used, the more averaging occurs among neighbors, blurring finer variations.

The codensity is used to filter the data as follows. Denote by $\mathcal{M}[k, T]$ the subset of $\mathcal{M}$ in the upper T-percent of density as measured by δ_k. This is a two-parameter subset of the point cloud which, for reasonable values of k and T, represent an appropriate core.

3.3. **"When shall we three meet again?"** The first interesting persistent homology computation on this data set occurs at the level of H_1: to what extent are there 'loops' in the data set along which the cloud is concentrated?

Taking a density threshold of $T = 25$ at neighbor parameter $k = 300$, with 5000 points sampled at random from $\mathcal{M}[k, T]$, computing the barcode for the first homology H_1 reveals a unique persistent generator [3]. See Figure 5. This indicates that the data set is diffused about a primary circle in the 7-sphere. The structure of the barcode is robust with respect to the random sampling of the points in $\mathcal{M}[k, T]$.

The goal of the homology computation is to discover a 'hidden' feature of a data set that is not discernable by clustering and connectivity alone. The simplest such feature would be, as indicated by the computation above, a primary circle about which the data is scattered. To what might this correspond? A close examination of the data point corresponding to the primary circle reveals a pattern of 3-by-3 patches with one light region and one dark region separated by a linear transition. This **nodal curve** between light and dark is linear and appears in a circular family parameterized by the angle of the nodal line, as shown in Figure 5.

As seen from the barcode, this generator is dominant at the threshold and codensity parameters chosen. An examination of the barcodes for the first homology group H_1 of the data set filtered by codensity parameter $k = 15$ and threshold $T = 25$ reveals a different persistent first homology. The reduction in k leads to less averaging and more localized density sensitivity. The barcode of Figure 6 reveals that the persistent H_1 of samples from $\mathcal{M}[k, T]$ has Betti number five. This does not connote the presence of five disjoint circles in the data set. Rather, by focusing on the generators and computing the barcode for H_0, it is observed [3] that, besides the primary circle from the high-k H_1 computation, there are two secondary circles which come into view at the lower density parameter.

A close examination of these three circles reveals that each intersects the high-k primary twice, yet the two secondary circles are disjoint. To what features in the data might these secondary circles correspond? As noted in [3], each secondary circle regulates images with three contrasting regions and interpolates between these states and the primary circle. The difference between the two secondary circles lies in their bias for horizontal and vertical stratification respectively. Figure 7 gives an interpretation of the meanings of the secondary circles.

3.4. **"Come like shadows, so depart!"** What is the good of temporary topological features which emerge and dissolve as a function of the parameter ϵ? Does this lead to anything more than a heuristic for high-dimensional data sets that are

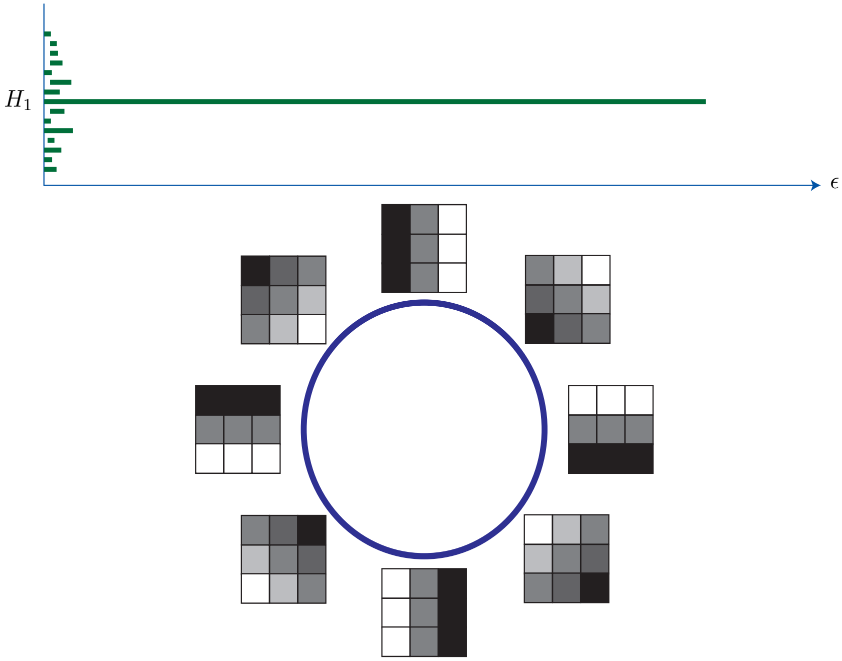


Figure 5. The H_1 barcode for a random sampling of 5000 points of $\mathcal{M}[300, 25]$ yields a single generator. This generator indicates the nodal line between a single light and single dark patch as being the dominant feature of the primary circle in $\mathcal{M}$.

hard to visualize? While the work of Carlsson et al. is very recent, there are several applications of the topological approach to data analysis which argue in favor of the proposition that homological structures in high dimensional data sets are of scientific significance. Besides the Mumford data set reviewed here, persistent homology computations are being applied to geometric features of curves (*e.g.,* optical character recognition) [5] and visual cortex data from primate experiments [4].

With regards to the natural image data, it is instructive to think of the persistent homology of $\mathcal{M}$ as something akin to a Taylor approximation of the true space. The reduction of the full data set to an S^7 via projection is really a normalization to eliminate the zero-order (or "single patch") terms in the data set. Following this analogy, the H_1 primary generator fills the role of a next term in the expansion of the homotopy type of the data set, collating the nodal curve between two contrasting patches. The secondary circles, interpolating between single and dual nodal curves, act as higher-order terms in the expansion, in which horizontal and vertical biases arise.

It is here that one gets deeper insight into the data set. Inspired by the meaning of the H_1 barcodes of $\mathcal{M}$, further investigation reveals what appears to be an intrinsic bias toward horizontal and vertical directions in the natural image data, as opposed to an artifact of the (right) angle at which the camera was held: [3] reports

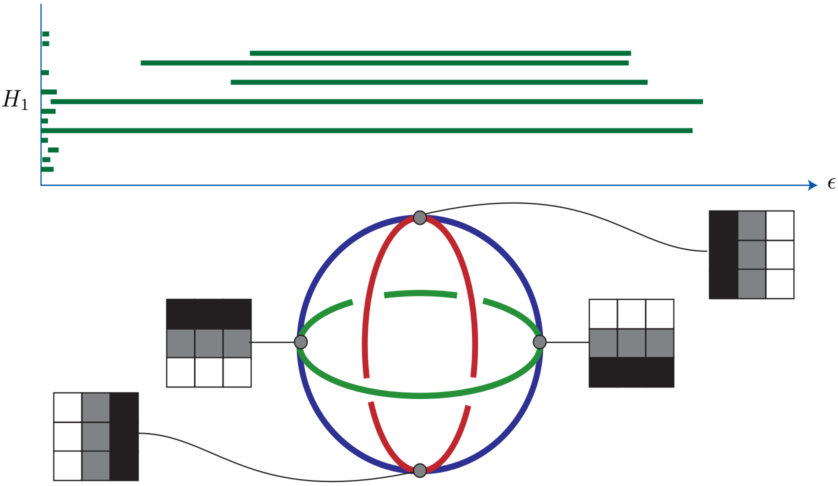


FIGURE 6. The H_1 barcode for $\mathcal{M}[15, 25]$ reveals five persistent generators. This implies the existence of two secondary circles, each of which intersects the third, large-k, primary circle twice.

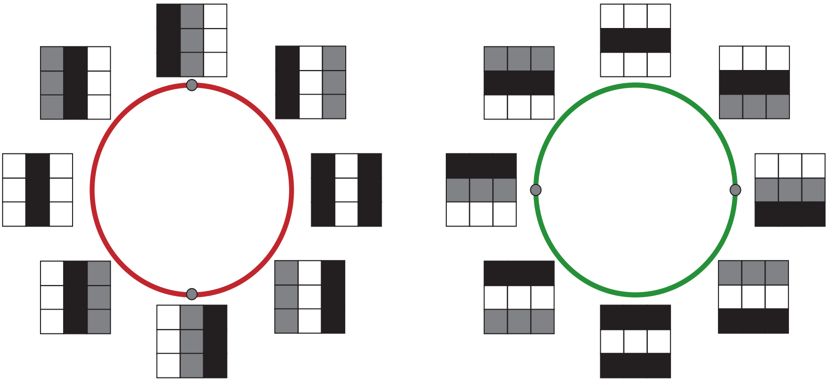


FIGURE 7. The secondary generators of H_1 for $\mathcal{M}[15, 25]$ have an interpretation as regulating changes from dual-patch to triple-patch high contrast regions in horizontal and vertical biases respectively.

that a repetition of the experiment with a camera held at a constant angle $\pi/4$ yields a data set whose secondary persistent H_1 generators exhibits a bias towards true vertical and true horizontal: the axis of pixellation appears less relevant than the axis of gravity in natural image data.

Is there any predictive power in the barcodes of the data set? Recent progress [3] demonstrates the insight that a persistent topology approach can yield. The barcodes for the second persistent homology H_2 are more volatile with respect to

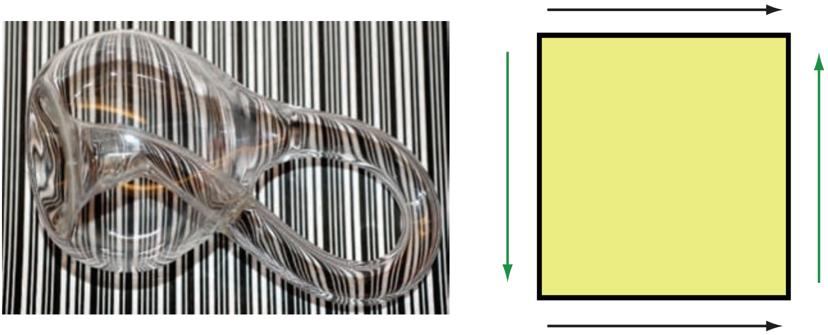


FIGURE 8. A Klein bottle (projected against its 'barcode') [left] is the non-orientable surface obtained by identifying opposite sides of a square as shown [right].

changes in density and thresholding. This is not surprising: the lowest order terms in any expansion are always most easily perceived. However, there is indication of a persistent H_2 generator (in $\mathbb{Z}_2$ coefficients) at certain settings of k and T. Combined with the basis of H_1 generators, one obtains predictive insight to the structure of the space of high-contrast patches. At certain density thresholds, the H_2 barcode, suitably trimmed with Occam's razor, suggests a two-dimensional completion of the low-k persistent H_1 basis into a **Klein bottle** (see Figure 8). Recall that this non-orientable surface can be realized as an identification space of a square as in the figure. Figure 9 illustrates an embedding of this surface in the space of pixellated images. One notes that this is a natural completion of the low-density persistent H_1 readings: the primary and secondary circles appear with the appropriate intersection properties. Fortunately, $\mathbb{Z}_2$ coefficients — most natural for computer implementation — is efficacious in detecting H_2 for a Klein bottle.

We emphasize that the point cloud data set $\mathcal{M}$ is vast, high-dimensional, and not at all concentrated sharply along distinct features. A cursory viewing of the data seems to indicate that the 7-sphere is filled densely with data points and that there is seemingly no coherent structure to be found. It is through the lens of persistent homology — suitably tuned and aimed — that cogent features emerge and fade with changing parameters. These persistent generators, upon close examination, do correspond to meaningful structures in the data, inspiring a sensible parametrization of the global structure of the data set. This is the type of explanatory power that any exemplar of good applied mathematics provides to a scientific challenge.

References

[1] P. Bubenik and P. Kim, "A statistical approach to persistent homology," preprint (2006).

[2] E. Carlsson, G. Carlsson, and V. de Silva, "An algebraic topological method for feature identification," *Intl. J. Computational Geometry and Applications*, 16:4 (2006), 291-314.

[3] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian, "On the local behavior of spaces of natural images," preprint, (2006).

[4] G. Carlsson, T. Ishkhanov, F. Mémoli, D. Ringach, G. Sapiro, "Topological analysis of the responses of neurons in V1.," in preparation (2006).

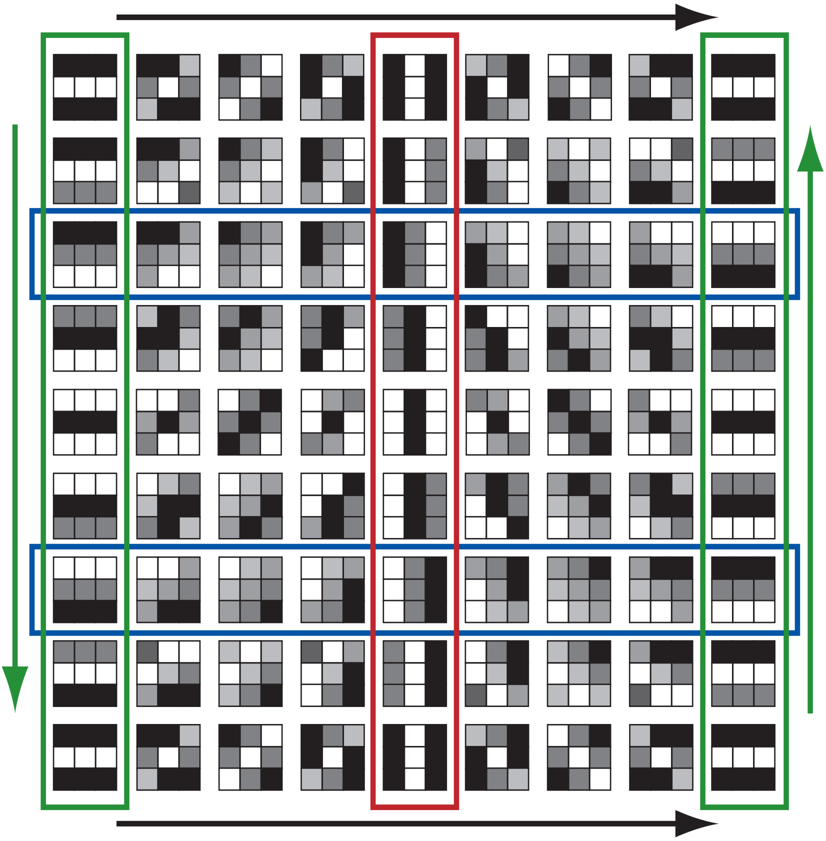


Figure 9. A Klein bottle embeds naturally in the parameter space as a completion of the 3-circle model. In the unfolded identification space shown, the primary circle wraps around the horizontal axis twice. The two secondary circles each wrap around the vertical axis once (note: the circle on the extreme left and right are glued together with opposite orientation). Each secondary circle intersects the primary circle twice.

[5] G. Carlsson, A. Zomorodian, A. Collins, and L. Guibas, "Persistence barcodes for shapes," *Intl. J. Shape Modeling*, 11 (2005), 149-187.

[6] F. Chazal and A. Lieutier, "Weak feature size and persistent homology: computing homology of solids in $\mathbb{R}^n$ from noisy data samples," in *Proc. 21st Sympos. Comput. Geom.* (2005).

[7] D. Cohen-Steiner, H. Edelsbrunner and J. Harer, "Stability of persistence diagrams," in *Proc. 21st Sympos. Comput. Geom.* (2005), 263–271.

[8] V. de Silva, "A weak definition of Delaunay triangulation," preprint (2003).

[9] V. de Silva and G. Carlsson. "Topological estimation using witness complexes," in *SPBG04 Symposium on Point-Based Graphics* (2004), 157-166.

[10] V. de Silva and R. Ghrist, "Coverage in sensor networks via persistent homology," to appear, *Alg. & Geom. Topology* (2006).

[11] V. de Silva and P. Perry, PLEX home page, http://math.stanford.edu/comptop/programs/plex/

[12] H. Edelsbrunner, D. Letscher, and A. Zomorodian, "Topological persistence and simplification," *Discrete Comput. Geom.*, 28:4 (2002), 511-533.

[13] H. Edelsbrunner and E.P. Mücke, "Three-dimensional alpha shapes," *ACM Transactions on Graphics*, 13:1, (1994), 43-72.

[14] L. Guibas and S. Oudot, "Reconstruction using witness complexes," in *Proc. 18th ACM-SIAM Sympos. on Discrete Algorithms*, (2007).

[15] A. Hatcher, *Algebraic Topology*, Cambridge University Press, (2002).

[16] T. Kaczynski, K. Mischaikow, and M. Mrozek, *Computational Homology,* Applied Mathematical Sciences 157, Springer-Verlag, (2004).

[17] D. Mumford, "Pattern Theory: the Mathematics of Perception," Proc. Intl. Congress of Mathematicians, Vol. III (2002), 1–21.

[18] D. Mumford, A. Lee, and K. Pedersen, "The nonlinear statistics of high-contrast patches in natural images," *Intl. J. Computer Vision*, Vol. 54 (2003), 83–103.

[19] B. Silverman, *Density Estimation for Statistics and Data Analysis.* Chapman & Hall/CRC (1986).

[20] J. van Hateren and A. van der Schaff, "Independent Component Filters of Natural Images Compared with Simple Cells in Primary Visual Cortex", *Proc. R. Soc. London*, vol B 265 (1998), 359–366.

[21] L. Vietoris, "Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen," *Math. Ann.* 97 (1927), 454–472.

[22] A. Zomorodian and G. Carlsson, "Computing Persistent Homology," *Discrete Comput. Geom.*, 33, (2005), 249–274.

[23] A. Zomorodian and G. Carlsson, "Localized homology," preprint (2005).

[24] A. Zomorodian and G. Carlsson, "The theory of multidimensional persistence," preprint (2006).

Department of Mathematics and Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801, USA

E-mail address: ghrist@math.uiuc.edu

THE WORK OF EINSIEDLER, KATOK AND LINDENSTRAUSS ON THE LITTLEWOOD CONJECTURE

AKSHAY VENKATESH

Contents

This document is intended as a (slightly expanded) writeup of my (anticipated) talk at the AMS Current Events Bulletin in New Orleans, January 2007. It is a brief report on the work of Einsiedler, Katok and Lindenstrauss on the Littlewood conjecture [5].

It is not intended in any sense for specialists and is, indeed, aimed at readers without any specific background either in measure theory, dynamics or number theory.

Any reader with any background in ergodic theory will be better served by consulting either the original paper, or one of the surveys written by those authors: see [7] and [12].

1. The Littlewood conjecture

1.1. For $x \in \mathbb{R}$, let $\|x\|$ denote distance from x to the nearest integer.

It is not difficult to check that, for any $\alpha \in \mathbb{R}$, there exists integers p, q with $1 \leq q \leq Q$ and $|\alpha - \frac{p}{q}| \leq \frac{1}{qQ}$. In other words, $\|q\alpha\| \leq 1/Q$. The behavior of $\|q\alpha\|$, as q varies through integers, thereby reflects *approximation of α by rational numbers.*

The Littlewood conjecture concerns simultaneous approximation of *two* numbers α, β by irrationals. It asserts that:

$$\liminf n.\|n\alpha\|\|n\beta\| = 0, \tag{1}$$

whatever be α, β. In words it asserts (in a somewhat peculiar-seeming way)

(2) α, β may be simultaneously approximated, moderately well,
by rationals with the same denominator

My goal is to discuss, and give some of the context around, the following theorem of M. Einsiedler, A. Katok and E. Lindenstrauss in [5]:

1.1**. Theorem.** *The set of α, β for which* (1) *fails, has Hausdorff dimension* 0*.*

The Theorem is proved using ideas from dynamics: namely, by studying the action of coordinate dilations (e.g. $(x, y, z) \mapsto (\frac{x}{2}, 2y, z)$) on the space of *lattices* in $\mathbb{R}^3$. It is not important solely as a result about simultaneous Diophantine approximation, but because of the techiques and results in dynamics that enter into its proof.

Several applications of this type of dynamics are surveyed in [7]. For now it is worth commenting on two rather different contexts where exactly the same dynamics arise:

- In the study of analytic behavior of automorphic forms (see [17] for discussion and historical context)
- In the study of the analytic behavior of ideal classes in number fields, see [8].

1.2. **This document.** I will try to stress:

(1) Dynamics arises from a (not immediately visible) symmetry group; see §1.3; I will then discuss some historical context for this type of connection (§2, §3).
(2) The dynamics that is needed is similar to the simultaneous action of $x \mapsto 2x, x \mapsto 3x$ on $\mathbb{R}/\mathbb{Z}$; see §4.3 for a description of these parallels.
(3) A sketch of just one of the beautiful ideas that enters in proving Theorem 1.1 (see §5), which is to study the picture transverse to the acting group.

A massive defect of the exposition is that I will make almost no mention of *entropy.* This is an egregious omission, because the intuition which comes from the study of entropy underpins much of the recent progress in the subject. However, any serious discussion of entropy this would require more space and time and competence than I have, and better references are available. So, instead, I have given a somewhat *ad hoc* discussion adapated to the cases under consideration.

I will not come even close to sketching a proof of the main result.

Let us make two notes before starting any serious discussion:

(1) The Littlewood conjecture, (1), is quite plausible. Here is a naive line of heuristic reasoning that supports it. A consequence of what we have said in §1.1 is that there exists a sequence $q_k \to \infty$ of positive integers so that $q_k \|q_k \alpha\| \leq 1$. Barring some conspiracy to the

contrary, one might expect that $\|q_k\beta\|$ should be small *sometimes.* The problem in implementing this argument is that we have rather little control over the q_k.[1]

(2) Despite all the progress that I shall report on, we do not know that the statement (1) is true even for $\alpha = \sqrt{2}, \beta = \sqrt{3}$. The question of removing the exceptional set in Theorem 1.1 is related to celebrated conjectures (see Conjecture 4.1 and Conjecture 4.2) of Furstenberg and Margulis.

1.3. **Symmetry.**

The next point is that the question (1) has a symmetry group that is not immediately apparent. This is responsible for our ability to apply dynamical techniques to it.

Pass to a general context for a moment. Let $f(x_1, \ldots, x_n)$ be an integral polynomial in several variables. An important concern of number theory has been to understand *Diophantine equalities:* solutions to $f(\mathbf{x}) = 0$ in integers $\mathbf{x} \in \mathbb{Z}^n$ (e.g. does $x^2 - y^2 - z^2 = 1$ have a solution? Does $x^3 + y^3 = z^3$ have a solution?)

A variant of this question, somewhat less visible but nonetheless (in my opinion) difficult and fascinating, concerns *Diophantine inequalities:* if f does not have rational coefficients, one may ask about the solvability of an equation such as $|f(\mathbf{x})| < \varepsilon$ for $\mathbf{x} \in \mathbb{Z}^n$ (e.g. does $|x^2 + y^2 - \sqrt{2}z^2| < 10^{-6}$ have a solution?)

In the most general context of an arbitrary f, our state of knowledge is somewhat limited. On the other hand, for special classes of f we know more: a typical class which is accessible to analytic methods is when the *degree of f is small compared to the number of variables.*

Another important class about which we have been able to make progress, consists of those f possessing symmetry groups. Both the examples $x^2 - y^2 - z^2 = 1$ and $x^2 + y^2 - \sqrt{2}z^2$ admit orthogonal groups in three variables as automorphisms.[2] The homogeneous equation $x^3 + y^3 = z^3$ has symmetry but not by a linear algebraic group (it defines an elliptic curve inside $\mathbb{P}^2$).

The Littlewood conjecture also has symmetry, although not immediately apparent. To see it, we note that $\|x\| = \inf_{m \in \mathbb{Z}} |x - m|$; consequently, we may rewrite (1) as the statement:

(3)

$$|n(n\alpha - m)(n\beta - \ell)| < \varepsilon \text{ is solvable, with } (n, m, \ell) \in \mathbb{Z}^3, n \neq 0, \text{ for all } \varepsilon > 0$$

[1] Amusingly, it is not even clear this heuristic argument will work. It may be shown that given a sequence q_k so that $\liminf q_{k+1}/q_k > 1$, there exists $\beta \in \mathbb{R}$ so that $\|q_k\beta\|$ is bounded away from 0. See [14] for this and more discussion.

[2] Although unimportant in the context of this paper, there is an important difference: while $x^2 + y^2 - \sqrt{2}z^2$ admits an action of the real Lie group $\mathrm{O}(2, 1)$, the analysis of the form $x^2 + y^2 + z^2$ involves studying the action of the much larger *adelic* Lie group of automorphisms. In particular, this adelic group is noncompact, even though the real group $\mathrm{O}(3)$ is compact, and this is a point that can be fruitfully exploited; see [2].

But the function $L(n,m,\ell) = n(n\alpha - m)(n\beta - \ell)$ is a product of three linear forms and admits a two-dimensional torus as group of automorphisms.[3]

2. The Oppenheim conjecture

Here we pause to put the developments that follow into their historical context. The reader may skip directly to §4.

2.1. **Statement of the Oppenheim conjecture.**

We briefly discussed above the form $x^2 + y^2 - \sqrt{2}z^2$. This is a particular case of a problem considered in the 1929: A. Oppenheim conjectured that if $Q(x_1, \dots, x_n) = \sum_{i,j} a_{ij}x_i x_j$ is an *indefinite* quadratic form in $n \geq 3$ variables which is not a multiple of a rational form, then Q takes values which are arbitrary small, in absolute value.

In other words – note the analogy with (3) –

$$|Q(\mathbf{x})| < \varepsilon \text{ is solvable, with } \mathbf{x} \in \mathbb{Z}^n, \text{ for all } \varepsilon > 0 \tag{4}$$

When n is sufficiently large his conjecture was solved by Davenport (in 1956) by analytic methods. His paper required $n \geq 74$. This is an example of the fact, noted in §1.3, that purely analytic methods can often handle cases when the number of variables is sufficiently large relative to the degree.

On the other hand, the complete resolution of the conjecture[4] had to wait until G. Margulis, in the early 1980s, gave a complete proof using dynamical methods that made critical use of the group of automorphisms of Q.

2.2. **Symmetry.**

Let $H = \mathrm{SO}(Q)$, the group of orientation-preserving linear transformations of $\mathbb{R}^n$ preserving Q. By definition $Q(\mathbf{x}) = Q(h.\mathbf{x})$. We wish to exploit[5] the fact that H is large.

In particular, in order to show (4), it suffices to show that Q takes values in $(-\varepsilon, \varepsilon)$ at a point of the form $h.\mathbf{x}\,(h \in H, \mathbf{x} \in \mathbb{Z}^n)$. *A priori*, this set might be much larger than $\mathbb{Z}^n$; certainly, if it were dense in $\mathbb{R}^n$, this would be enough to show (4).

For instance, if we could prove that

$$\text{The set } h.\mathbf{x} : h \in H,\ \mathbf{x} \in \mathbb{Z}^n \text{ contains } 0 \text{ in its closure} \tag{5}$$

then (4) would follow immediately.

[3] The n-dimensional version of the Littlewood conjecture takes n linear forms $\ell_1, \dots, \ell_n$ and asks: is the equation $0 < |\ell_1(\mathbf{x}) \dots \ell_n(\mathbf{x})| < \varepsilon$ solvable? Conjecturally, this is so if $n \geq 3$ and $\ell_1 \dots \ell_n$ is not a multiple of a rational polynomial. It is false for $n = 2$, see footnote 4.

[4] The analogous statement is false for $n = 2$: take, e.g. $Q(x,y) = (x - \sqrt{2}y)y$. To see that, write $Q(x,y) = \frac{(x^2-2y^2)y}{(x+\sqrt{2}y)}$.

[5] The idea that this should be exploitable was suggested by M. Raghunathan. It is also implicitly used in a paper of Cassels and Swinnerton-Dyer from the 1950s.

2.3. **Lattices.** (5) is rather nice, but a little unwieldy. We would rather deal with the H-orbit of a single point rather than an infinite collection. This can be done by "packaging" all $\mathbf{x} \in \mathbb{Z}^n$ into a single object: a *lattice.*

A *lattice* in $\mathbb{R}^n$ is simply a "grid containing the origin", i.e. a set of all *integral* combinations of n linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Every such lattice is of the form $g.\mathbb{Z}^n$ for some $g \in \mathrm{GL}(n, \mathbb{R})$.

Let $\widetilde{\mathcal{L}}_n$ be the set of lattices[6] and let $\widetilde{\mathcal{L}}_n[\varepsilon]$ be the set of lattices that contain $\mathbf{v} \in \mathbb{R}^n$ with Euclidean length $\|\mathbf{v}\| \leq \varepsilon$. So $\mathbb{Z}^n$ can be thought of as a point $[\mathbb{Z}^n] \in \widetilde{\mathcal{L}}_n$.

Then (5) would follow if:

$$H.[\mathbb{Z}^n] \cap \widetilde{\mathcal{L}}_n[\varepsilon] \neq \emptyset, \text{ for all } \varepsilon > 0 \tag{6}$$

This is a statement that fits cleanly into the context of dynamics: does the orbit of the point $\mathbb{Z}^n \in \widetilde{\mathcal{L}}_n$, under the group H, intersect the subset $\widetilde{\mathcal{L}}_n[\varepsilon]$? It is (6) which was proven by Margulis.

2.4. **Background on the space of lattices.** To each lattice we can assign a natural invariant, its *covolume.* This is the absolute value of the determinant of the matrix with rows $\mathbf{v}_1, \ldots, \mathbf{v}_n$; that is to say, the volume of a fundamental parallelpiped $\sum \lambda_i \mathbf{v}_i : \lambda_i \in [0, 1)$. For $g \in \mathrm{GL}(n, \mathbb{R})$ and $L \in \widetilde{\mathcal{L}}_n$, we observe that $\mathrm{covol}(g.L) = |\det g| \mathrm{covol}(L)$. In particular, because all $h \in H$ have determinant 1, all elements in $H.\mathbb{Z}^n$ have covolume 1. So $H.\mathbb{Z}^n$ belongs to the subset

$$\mathcal{L}_n = \{L \in \widetilde{\mathcal{L}}_n : L \text{ has covolume } 1.\} \tag{7}$$

The space $\mathcal{L}_n$ is more pleasant to work with than $\widetilde{\mathcal{L}}_n$. The map $g \mapsto g.[\mathbb{Z}^n]$ identifies $\widetilde{\mathcal{L}}_n$ with the quotient $\mathrm{GL}(n, \mathbb{R})/\mathrm{GL}(n, \mathbb{Z})$ and $\mathcal{L}_n$ with the quotient $\mathrm{SL}(n, \mathbb{R})/\mathrm{SL}(n, \mathbb{Z})$. These identifications give rise to topologies on $\widetilde{\mathcal{L}}_n$ and $\mathcal{L}_n$; indeed, they are given the structure of manifolds.

Although $\mathcal{L}_n$ is not compact, it admits a natural $\mathrm{SL}(n, \mathbb{R})$-invariant measure which has finite volume, which is a reasonable substitute for compactness. Moreover, Mahler's criterion gives a precise description of in what way $\mathcal{L}_n$ fails to be compact:

2.1. **Theorem.** *A subset $K \subset \mathcal{L}_n$ is bounded (=precompact) if and only if it does not intersect $\mathcal{L}_n[\varepsilon]$, some $\varepsilon > 0$.*

In words, it asserts that the only way that a sequence of lattices $L_1, L_2, \ldots$ in $\mathcal{L}_n$ can degenerate (leave any compact set in $\mathcal{L}_n$) is if there exist vectors $\mathbf{v}_1 \in L_1, \mathbf{v}_2 \in L_2, \ldots$ so that $\|\mathbf{v}_i\| \to 0$.

We may therefore rephrase (6): The Oppenheim conjecture would follow if

$$H.[\mathbb{Z}^n] \text{ is unbounded in } \mathcal{L}_n. \tag{8}$$

[6] We will later work almost exclusively with the subset of $\mathcal{L} \subset \widetilde{\mathcal{L}}_n$ consisting of lattices of volume 1; therefore, for notational simplicity, we prefer to put a tilde for the whole space of lattices and omit it for the subset $\mathcal{L}_n$.

3. Unipotents acting on lattices.

Obviously, the statement (4) is false for Q positive definite, and, as observed in footnote 4, (less obviously) false for Q in two variables. How are we to detect this difference when considering the problem from the dynamical viewpoint of (6) or (8)?

3.1. Unipotents from Margulis to Ratner.

3.1. Unipotents from Margulis to Ratner. An important difference is that the group H is isomorphic to $\mathrm{SO}(n) \subset \mathrm{GL}(n, \mathbb{R})$ in the first case, and $\mathrm{SO}(1,1) \subset GL(2, \mathbb{R})$ in the second case. In either case, *the group H consists entirely of semisimple elements.* Margulis' idea was to exploit the fact that, if Q is indefinite in $n \geq 3$ variables, the group H contains *unipotent* elements, i.e. $g \in \mathrm{GL}(n, \mathbb{R})$ for which all of the (generalized) eigenvalues of g are equal to 1.

At a vague level, the reason why these might be helpful is quite easy to state: if $u \in \mathrm{GL}(n, \mathbb{R})$ is unipotent, the matrix entries of u^n grow only *polynomially* in n. This contrasts sharply with the behavior of a "typical" element $g \in \mathrm{GL}(n, \mathbb{R})$, when these entries will grow expontially. This means that, when studying the trajectory $ux_0, u^2x_0, u^3x_0, \ldots$, we are able to "retain information" about it for much longer.

3.2. Ratner's theorem. We will not say anything about the specifics of Margulis' proof; see [1] for an elementary presentation. A far-reaching generalization of Margulis' result, which has been of fundamental importance for later work, is the following (special case of a) theorem of Ratner, see [15] and [16]: [7]

3.1. Theorem. *Let $H \subset \mathrm{SL}(n, \mathbb{R})$ be generated by one-parameter unipotent subgroups.*[8] *The closure of the orbit $\overline{H.[\mathbb{Z}^n]}$ inside $\mathcal{L}_n$ is of the form $H'.[\mathbb{Z}^n]$ for a closed subgroup $H' \geqslant H$. Moreover, there exists an H'-invariant probability measure on $H'.[\mathbb{Z}^n]$.*

This is a difficult theorem, which settled a conjecture of M. Raghunathan. The orbit $H.[\mathbb{Z}^n]$ can be extremely complicated. Ratner's theorem asserts that its closure is determined by a very simple piece of algebraic data: a subgroup intermediate between H and $\mathrm{SL}(n, \mathbb{R})$.

Let us see how this implies (6). The group $H = \mathrm{SO}(Q)$ is *maximal* inside $\mathrm{SL}(n, \mathbb{R})$. So Theorem 3.1 means that either $H.[\mathbb{Z}^n]$ is closed or $H.[\mathbb{Z}^n]$ is dense in $\mathcal{L}_n$. It may be seen that $H.[\mathbb{Z}^n]$ is closed only if the form Q is a multiple of a rational form. In this fashion, Theorem 3.1 implies the Oppenheim conjecture.

[7] Ratner's theorem is phrased not just about spaces like $\mathcal{L}_n = \mathrm{SL}(n, \mathbb{R})/\mathrm{SL}(n, \mathbb{Z})$, but more general quotients of Lie groups by discrete subgroups.

[8] i.e. of the form $\exp(tX)$ where X is a nilpotent matrix.

3.3. An idea from the proof of Theorem 3.1: Measures not sets.

Because our concern is not with unipotent dynamics here, we will not try to indicate any of the ideas of the proof of Theorem 3.1 that are specific to properties of the unipotent flows.

Instead, we will emphasize a more philosophical point from the proof of Theorem 3.1 that has been indispensable in later work.

(9) Measures are often easier to work with than sets.

To be a little more specific, let us comment on how Ratner's proof of Theorem 3.1 works. Let us take the simple case when H consists *entirely* of unipotent elements. (A comprehensive exposition of the proof is to be found in [18]).

Ratner begins by classifying the *probability measures* on $\mathcal{L}_n$ that are invariant under H. The topological statement of Theorem 3.1 is then *deduced* from the classification of H-invariant probability measures.

The relation between probability measures and invariant sets is quite simple: an invariant probability measure has a support, which is a closed H-invariant set. Conversely, $Y \subset \mathcal{L}_n$ is an H-invariant closed set, it must support an H-invariant probability measure (average your favorite measure under H – note that this requires H to be amenable.) This relation is a good deal more tenuous than one would like – the support of the measure constructed this way may be strictly smaller than Y – and the deduction of statements concerning invariant sets from statements about probability measures is not formal.

Nonetheless, what is gained by going through measures? Measures have much better formal properties than sets. A particularly important difference is that an H-invariant probability measure can be decomposed into "minimal" invariant measures (ergodic decomposition). [9] That property does not seem to have a clean analogy at the level of H-invariant closed sets. In particular, an H-invariant closed set always *contains* a minimal H-invariant closed set, but cannot be decomposed into minimal H-invariant closed sets in any obvious way.

This is not to say that it is necessarily impossible to prove Theorem 3.1 by purely topological methods. Indeed, Margulis' original proof of (6) was purely topological (and utilized a study of minimal H-invariant closed sets). But, to my knowledge, no such proof has been carried out in the general case.

[9] The set of H-invariant probability measures forms, clearly, a convex set in the space of all probability measures. Any point in this convex set can be expressed as a convex linear combination of extreme points. These extreme points are called *ergodic* measures for H and are "minimal", in the sense that they cannot be expressed nontrivially as an average of two other H-invariant probability measures.

4. The dynamics of coordinate dilations on lattices, I: conjectures and analogies.

.

We have seen that the assertion (4) about the values of the quadratic form $x^2 + y^2 - \sqrt{2}z^2$ can be converted to the assertion (6) about the orbit of $[\mathbb{Z}^n]$ under the group $H = \mathrm{SO}(Q)$. We now briefly carry through the corresponding reasoning in the case of the Littlewood conjecture. This will lead us to study the action of the diagonal group A_3 inside $\mathrm{GL}(3,\mathbb{R})$, on $\mathcal{L}_3$.

4.1. Reduction to dynamics.

Let $P(x_1,x_2,x_3) = x_1(\alpha x_1 - x_2)(\beta x_1 - x_3)$. We have seen (see (3)) that the Littlewood conjecture is (almost, with a constraint $x_1 \neq 0$) equivalent to the assertion that $|P(\mathbf{x})| < \varepsilon$ is solvable. Let T be the *automorphism group of* P, that is to say, the set of $g \in \mathrm{GL}(3,\mathbb{R})$ such that $P(g.\mathbf{x}) = P(\mathbf{x})$. T contains a conjugated copy of the group of diagonal matrices. [10]

So $P(a.\mathbf{x}) = P(\mathbf{x})$ for $a \in T$. It would appear to be enough to show that $\{a.\mathbf{x} : a \in T, \mathbf{x} \in \mathbb{Z}^n\}$ approaches arbitrarily close to 0; or, repeating the line of implications (4) $\Longleftarrow$ (6) $\Longleftarrow$ (8), it seems to be enough to show that $T.[\mathbb{Z}^n]$ is unbounded in $\mathcal{L}_n$.

This is not quite right, though: $T.[\mathbb{Z}^n]$ being unbounded in $\mathcal{L}_n$ indeed would produce solutions to $|x_1(x_1\alpha - x_2)(x_1\beta - x_3)| < \varepsilon$, but, regrettably, provides no guarantee that $x_1 \neq 0$.

However, this can be avoided by replacing T with a certain subsemigroup $T^+ \subset T$ engineered specifically to avoid this. Moreover, T contains a conjugate copy of A_3 as a finite index subgroup, we can rephrase this assertion in terms of the dynamics of A_3, not of T.

We will not go through the details, but rather will explicate the result of going through this process: if $L_{\alpha,\beta} \subset \mathbb{R}^3$ is the lattice spanned by $(1,\alpha,\beta), (0,1,0)$ and $(0,0,1)$, the Littlewood conjecture for (α,β) is equivalent to:

(10)

$$A_3^+.L_{\alpha,\beta} \text{ is unbounded in } \mathcal{L}_n, A_3^+ = \{\begin{pmatrix} x & 0 & 0 \\ 0 & y & 0 \\ 0 & 0 & z \end{pmatrix} : x \le 1, y \ge 1, z \ge 1\}$$

The reader can easily verify (10) directly.

We are led to study the action of A_n on $\mathcal{L}_n$, and in particular, to seek an analogue of Theorem 3.1. The obstacle will be that the analogue of Theorem 3.1 *totally fails* for (conjugates of) A_2 acting on $\mathcal{L}_2$. There exists a plethora of orbit closures that do not correspond to closed orbits of intermediate subgroups $A_2 \leqslant H \leqslant \mathrm{SL}(2,\mathbb{R})$. (This corresponds roughly to the fact

[10] In a suitable coordinate system, P becomes $P(x_1,x_2,x_3) = x_1x_2x_3$. But the set of linear transformations that preserve $(x_1,x_2,x_3) \mapsto x_1x_2x_3$ consist of all permutation matrices whose determinant is ± 1, according to the sign of the permutation.

that there are many α for which $\liminf n\|n\alpha\| > 0$, i.e. the "one variable" Littlewood conjecture is false.)

4.2. **An analogy with $\times 2 \times 3$ on S^1.** Let us reprise: we are studying the action of the group A_n (diagonal matrices of size n, with determinant 1) on the space $\mathcal{L}_n = \mathrm{SL}(n,\mathbb{R})/\mathrm{SL}(n,\mathbb{Z})$; or, geometrically, we are studying the action of *coordinate dilations* on grids in $\mathbb{R}^n$.

A very helpful analogy in studying the action of A_n on $\mathcal{L}_n$ is the following:

$$\text{Action of } A_2 \text{ on } \mathcal{L}_2 \text{ behaves like } x \mapsto 2x \text{ on } \mathbb{R}/\mathbb{Z}; \tag{11}$$

$$\text{Action of } A_3 \text{ on } \mathcal{L}_3 \text{ behaves like } x \mapsto 2x, x \mapsto 3x \text{ on } \mathbb{R}/\mathbb{Z} \tag{12}$$

Note that A_3 is a two-parameter (continuous) group, whereas $x \mapsto 2x, x \mapsto 3x$ generate a two-parameter (discrete) semigroup.

These analogies appear to be quite strong, although I do not know of any entirely satisfying "reason" for them. The analogy between $(A_2, \mathcal{L}_2)$ and $(\times 2, \mathbb{R}/\mathbb{Z})$ is particularly strong: in a fairly precise sense[11], the action of a suitable element $a \in A_2$ on $\mathcal{L}_2$ behaves like a shift on $\{0,1\}^{\mathbb{Z}}$, whereas the action of $x \mapsto 2x$ behaves like a shift on $\{0,1\}^{\mathbb{N}}$. We will list in the next section some results and questions in both the $\mathcal{L}$ and $\mathbb{R}/\mathbb{Z}$ cases and see they are quite analogous.

For the moment, let us just observe that the action of $x \mapsto 2x$ on $\mathbb{R}/\mathbb{Z}$ is fundamentally different to the simultaneous action of $x \mapsto 2x, x \mapsto 3x$. Indeed, the trajectory $\{2^n x\}$ of a point under $x \mapsto 2x$ essentially encodes the binary expansion of x, which can be arbitrarily strange (cf. Lemma 4.1). For instance, there exist uncountably many possibilities for the closure $\overline{\{2^n x\}}$. On the other hand, it is much more difficult to arrange that the binary and ternary expansions of a given x be *simultaneously* strange. This means it is much harder to arrange that the orbit of x under $x \mapsto 2^n 3^m x$ be strange, and indeed it is known that the possibilities for the closure $\overline{\{2^n 3^m x\}}$ are very simple (see Theorem 4.1).

Correspondingly, one might hope that the fact that Theorem 3.1 fails for $(A_2, \mathcal{L}_2)$, as commented at the end of §4.1, might be a phenomenon that vanishes when one passes to $(A_n, \mathcal{L}_n)$ for $n \geq 3$. Indeed, this is believed to be largely the case.

4.3. **Conjectures and results for $\times 2 \times 3$ and for A_n.** Recall that a probability measure ν invariant under a group G is said to be G-ergodic if any G-invariant measurable subset S has either $\nu(S) = 1$ or $\nu(S) = 0$. An equivalent definition is found in footnote 9. We observe that a classification of G-invariant ergodic probability measures is as good as a classification of G-invariant probability measures, for any G-invariant probability measure can be expressed as a convex combination of G-invariant ergodic probability measures.

[11] e.g. the systems are measure-theoretically isomorphic

We shall also make use in this section of the notion of *positive entropy*; for a definition see (19), but the reader might be better served by simply treating it as a black-box notion for the moment and reading on.

Formalizations of some of the intuitions we suggested in the previous section are to be found in the following results. They state, in that order, that:

- There are a huge number of closed invariant sets for $x \mapsto 2x$.
- There are very few closed invariant sets for $x \mapsto 2x, x \mapsto 3x$ simultaneously (and a clean classification)
- Conjecturally, there are very few invariant probability measures under $x \mapsto 2x, x \mapsto 3x$;
- One can prove the third assertion under an additional assumption on the measure, positive entropy.

4.1. **Lemma.** *There exists orbit closures $\overline{\{2^n x\}}_{n \geq 0}$ of any Hausdorff dimension between* 0 *and* 1*.*

Similarly, there exist "very many" probability measures on $\mathbb{R}/\mathbb{Z}$ invariant under $x \mapsto 2x$. (A measure is said to be invariant under $x \mapsto 2x$ if the integral of $f(x)$ and $f(2x)$ is the same, for f a continuous function).

4.1. **Theorem.** *(Furstenberg) The orbit closure $\overline{\{2^n 3^m x\}_{n,m \geq 0}}$ is $\mathbb{R}/\mathbb{Z}$ or finite, according to whether x is irrational or rational.*

4.1. **Conjecture.** *(Furstenberg) Let μ be a probability measure on $\mathbb{R}/\mathbb{Z}$ that is invariant under $x \mapsto 2x$ and $x \mapsto 3x$ and ergodic w.r.t. $x \mapsto 2x, x \mapsto 3x$. Then μ is either Lebesgue measure, or supported on a finite set of rationals.*

4.2. **Theorem.** *(Rudolph) Let μ be a probability measure on $\mathbb{R}/\mathbb{Z}$ that is invariant under $x \mapsto 2x$ and $x \mapsto 3x$ and ergodic w.r.t. $x \mapsto 2x, x \mapsto 3x$, and so that either $\times 2$ or $\times 3$ acts with positive entropy. Then μ is Lebesgue measure.*

Now let us enunciate the analogues of these statements for A_n acting on $\mathcal{L}_n$. They state, in this order, that:

- There are a huge number of orbit closures and invariant measures for A_2 acting on $\mathcal{L}_2$.
- Conjecturally, there are very few closed sets for A_n acting on $\mathcal{L}_n$, when $n \geq 3$. The statement here is not as satisfactory as in the $(\times 2 \times 3, \mathbb{R}/\mathbb{Z})$ case.
- Conjecturally, there are very few invariant probability measures on $\mathcal{L}_n$ under A_n, when $n \geq 3$.
- One can prove the third assertion under an additional assumption on the measure, positive entropy.

4.2. **Lemma.** *There exists orbit closures $\overline{A_2.x}$ of any Hausdorff dimension between* 1 *and* 3*.*

Similarly, there exists “very many” probability measures on $\mathcal{L}_2$ invariant by A_2.

The following conjectures are stated (in a considerably more general form) in [13].

4.2. **Conjecture.** *(Margulis) The orbit closure* $\overline{A_n.x}$ *(for* $n \geq 3$ *and* $x \in \mathcal{L}_n$*) is,* if compact*, a closed* A_n*-orbit.* [12]

4.3. **Conjecture.** *(Margulis) Let* μ *be a probability measure on* $\mathcal{L}_n$ *that is invariant under* A_n *and ergodic w.r.t.* A_n*. Then*[13] μ *coincides with the* H'*-invariant measure on a closed orbit* $H'x_0$*, for some subgroup* $A_n \leqslant H' \leqslant \mathrm{SL}(n, \mathbb{R})$*.*

4.3. **Theorem.** *(Einsiedler-Katok-Lindenstrauss) Let* μ *be a probability measure on* $\mathcal{L}_n$ *that is invariant under* A_n *and ergodic w.r.t.* A_n*, so that* some element of A_n acts with positive entropy*. Then* μ *is algebraic.*

Theorem 4.3 is the main theorem of [5]. The result concerning Littlewood's conjecture is deduced from it. It should be noted that, while Theorem 4.3 is closely analogous to Theorem 4.2, the technique of proof is quite different.

In the remainder of this article, we shall indicate one key idea that enters not only into the proof of 4.3, but into the proof of all results in that line proved so far, including [4] and [11].

5. Coordinate dilations acting on lattices, II: the product lemma of Einsiedler-Katok

The contents of this section are sketchy and impressionistic! For concreteness, we will primarily confine ourselves to the action of A_3 on $\mathcal{L}_3$.

The main thing which the reader might come away with is the importance and naturality of *conditional measures.* The study and usage of conditional measures is a formalization of the following natural idea: given an A_3-invariant measure μ on $\mathcal{L}_3$, study μ along slices transverse to A_3. Note that the action of A_3 contracts part of these slices and dilates other parts.

The ideas we will discuss in this section are contained in the important paper [4] of Einsiedler and Katok; we will not discuss the new ideas introduced in [5]. Those new ideas stem from [11] and are, indeed, essential to get the main result on the Littlewood conjecture. On the other hand, the ideas from [4] that we now discuss have been fundamental in all the later work in this topic.

[12] This is not quite as good as a complete classification of orbit closures, and, indeed, [13] posits a more precise classification. Conjecture 4.2 is just a simple clean statement that can be extracted from this classification.

[13] i.e. “the measure-theoretic analogue of Theorem 3.1 holds for A_n acting on $\mathcal{L}_n$.”

5.1. **Closed sets.** Before we embark on describing some of the ideas in [4], we begin by explaining how one might try to approach the analysis of A_3-invariant closed sets. We then explain – in the spirit of §3.3 – why it might be helpful to switch to measures.

Suppose $\sigma \subset \mathcal{L}_3$ is an A_3-invariant closed set.

We wish to study the behavior of σ in directions transverse to A_3. Let e_{ij} be the elementary matrix with a 1 in the (i, j) position and 0s everywhere else; for $i \neq j$ let $n_{ij}(x) = \exp(x.e_{ij})$. Then $N_{ij} = \{n_{ij}(x) : x \in \mathbb{R}\}$ is a subgroup of $SL_3(\mathbb{R})$.

A natural way of studying, then, of how σ behaves *transverse to* A_3 are the subsets:

$$\sigma_x^{ij} := \{t \in \mathbb{R} : n_{ij}(t)x \in \sigma\} \subset \mathbb{R}$$

This set is a closed subset of $\mathbb{R}$ and is defined for all $x \in \mathcal{L}_3$.

Now, we wish to use the fact that a typical element $a \in A_3$ can contract some N_{ij}s and expand others.

Let us take an explicit example: the matrix $a = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 4 \end{pmatrix}$. It centralizes N_{12} but it *shrinks* N_{23}, that is to say:

$$a.n_{23}(x).a^{-1} = n_{23}(x/8)$$

Now consider two points $x_1, x_2 \in \sigma$ which lie along the N_{23} direction from one another, i.e. $x_2 \in N_{23}x_1$. Let us compare $\sigma_{x_1}^{12}$ and $\sigma_{x_2}^{12}$. Because our element a centralizes N_{12},

$$\sigma_{x_1}^{12} = \sigma_{ax_1}^{12} = \sigma_{a^2x_1}^{12} = \dots \text{ AND } \sigma_{x_2}^{12} = \sigma_{ax_2}^{12} = \sigma_{a^2x_2}^{12} = \dots . \tag{13}$$

But $a^k x_1$ and $a^k x_2$ are becoming very close as $k \to \infty$ – because a shrinks the direction N_{23}. Therefore, if we had some version of the statement

$$\text{Wishful thinking: as } x \text{ approaches } y,\ \sigma_x^{12} \text{ approaches } \sigma_y^{12} \tag{14}$$

we could deduce from (13) – by considering $a^k x_1, a^k x_2$ as $k \to \infty$ – the following surprising fact:

$$\sigma_{x_1}^{12} = \sigma_{x_2}^{12} \quad \text{(NOT proved, based on wishful thinking!)} \tag{15}$$

In other words, were some version of (14) true: we would have a rather weak version of the following statement: the behavior of a closed set σ in the N_{12}-direction, is constant along the N_{23}-direction. It is not immediate how to *use* this, but nonetheless it is an important structural fact. (See discussion after Lemma 5.1 for an indication of how the measure-theoretic version of this fact is used). It is quite surprising, because we assumed nothing about the behavior of σ besides A_3-invariance.

In order to get any mileage, of course, we need to be able to find points x_1, x_2 which differ in the N_{23} directions; or equivalently, the sets $\sigma_x^{N_{23}}$ should have more than one point. So in order to have any hope of using this entire setup, we should also have:

(16) The sets $\sigma_x^{N_{ij}}$ should not always be singletons.

Let me emphasize that the above is, indeed, essentially wishful thinking, and is based on the rather baselessly optimistic (14). The surprising fact is that, by working with measures, we can salvage a version of (14).

5.1. **Example.** *Take a closed subset S of the square $[0,1]^2$. Let $\pi : [0,1]^2 \to [0,1]$ be the projection. For each $x \in [0,1]$, we can consider the set $S_x = \pi^{-1}(\{x\}) \cap S$. There is no reason that nearby xs should have similar S_xs; this is the failure of* (14).

However, a measure-theoretic version of this is valid. *If μ is a probability measure on $[0,1]^2$, we can disintegrate it along fibers: we can write $\mu = \int_{x\in[0,1]} \mu_x d\nu(x)$, where $\nu = \pi_* \mu$ is the pushed-down measure on $[0,1]$, and μ_x is a probability measure supported on the fiber $\pi^{-1}(\{x\})$. The μ_xs are the measure-theoretic analogue of S_x; and:*

(17) *On a set of measure* 0.999999 *the function $x \mapsto \mu_x$ is continuous.*

In other words, throwing away a set of small measure, we can think of the μ_xs as satisfying a version of (14).

5.2. **What comes next.** Einsiedler and Katok implement the strategy discussed in §5.1, but in the world of measures, not sets.

- Rather than an A_3-invariant *closed set* σ, we start with an A_3-invariant probability measure μ.
- The analogue of $\sigma_x^{N_{ij}} \subset \mathbb{R}$ is played by *conditional measures* $\mu_x^{ij} \in$ Measures($\mathbb{R}$) discussed in §5.3. (Note that these are *not* probability measures in general, and may have infinite mass.)
- The assumption (16) that σ_x^{ij} not be singletons is replaced by the assumption that μ_x^{ij} not be *atomic* (a multiple of a point mass), which will be needed in both Theorem 5.1 and Theorem 5.2.
- One can prove the analogue of (15): it is the *product-lemma*, Lemma 5.1.

5.3. **Conditional measures: the analogue of the σ_x^{ij} for measures.** Let a nice group G (e.g. $G = N_{ij}$) act on a nice space X (e.g. $X = \mathcal{L}_3$).

Given a closed subset $S \subset X$, we can define the sets $\sigma_x^G = \{g \in G : gx \in S\}$, which isolates behavior of S along the G-direction. Now we want to define a similar concept but with the set S replaced by a probability measure μ, and replace the closed subset $\sigma_x^G \subset G$ by a measure μ_x^G (or just μ_x) on G.

This can indeed be done in a canonical way, except that the measures μ_x are defined only *up to scaling by a positive number.* In other words, there exists an association $x \mapsto \mu_x$ from points of X to measures on G, referred to as *conditional measures* along G, with the following properties:

(1) The map $x \mapsto \mu_x$ (thought of as a map from X to measures on G) is itself measurable.
(2) For $g \in G$ and $x \in X$ so that both μ_{gx} and μ_x are defined, the measures $\mu_{g.x}$ and $g.\mu_x$ are proportional[14] (one would like to say "equal" but everything is defined only up to a positive scalar).
(3) Let B be any open ball containing the identity in G. Then $\mu_x(B) > 0$ for almost all $x \in X$.
(4) μ is invariant under the G-action if and only if μ_x is a Haar measure on G for almost all $x \in X$.

Let's briefly describe how to do this when X is general but G is finite. In that case, one can normalize the μ_x canonically by requiring them to be probability measures on the finite set G. We will just describe the function $x \mapsto \mu_x(\{1\})$; then (2) determines μ_x totally (in this case, after normalizing the μ_x, the $\propto$ of (2) becomes equality).

Average μ under G to get a measure ν, w.r.t. which μ is absolutely continuous. Therefore, by the theorem of Radon and Nikodym, there exists a function $f \in L^1(\nu)$ so that $\mu = f.\nu$, i.e. $\mu(S) = \int_S f d\nu$. Then $f(x) = \mu_x(\{1\})$ almost everywhere, when matters are normalized so that μ_x is a probability measure.

Returning to the context of an A_3-invariant measure μ on $\mathcal{L}_3$, we denote by μ_x^{ij} the measure on $N_{ij} \cong \mathbb{R}$ defined by the process described above, applied to the action of N_{ij} on $\mathcal{L}_3$.

5.4. **From product lemma to unipotent invariance.** Let μ be an A_3-invariant measure on $\mathcal{L}_3$. The following is established in [4], Corollary to Proposition 5.1.

5.1. **Lemma.** *[Product lemma] Let μ be an A_3-invariant measure on $\mathcal{L}_3$.*

Then, for $(k, \ell) \neq (i,j), (j,i)$ we have $\mu^{ij}_{n^{k\ell}(t)x} \propto \mu_x^{ij}$, for $\mu_x^{k\ell}$-almost all $t \in \mathbb{R}$, and for μ-almost every $x \in X$.

The reasoning is a measure-theoretic version of that already discussed in (5.1). Thus Lemma 5.1 is "just" a consequence of the fact that it is possible to "shrink" the $N_{k\ell}$ while leaving N_{ij} unchanged.

We say that μ_x^{ij} is *trivial* if it is proportional to the Dirac measure supported at 0, i.e. if $\mu_x^{ij}(f) \propto f(0)$ for every continuous function f on the real line. To make usage of the μ_x^{ij}s, one really needs them to be *nontrivial* for almost all x. This is the analogue of (16).

Now let us briefly – and very heuristically – indicate how one might use Lemma 5.1. The assertion (5.1) says, in particular, that the value of $x \mapsto \mu_x^{13}$ is "the same" (at least, proportional) at x and at $n_{12}(t)x$, except for a set of t of $\mu_x^{(12)}$-measure 0. If μ_x^{12} is far from being atomic, we can find plenty of $t \neq 0$ for which this will be true. Similarly, if μ_x^{23} is far from being atomic,

[14] Here $g.\mu_x$ is the measure defined as $g.\mu_x(S) = \mu_x(Sg)$ for a subset $S \subset G$.

we can find plenty of s for which the value of $x \mapsto \mu_x^{13}$ is the same at x and at $n_{23}(s)x$.

Applying this argument repeatedly, we may hope to find t, s so that μ^{13} takes proportional values at x and $n_{12}(t)n_{23}(s)n_{12}(-t)n_{23}(-s)x$. But the groups N_{12} and N_{23} do not commute: indeed $n_{12}(t)n_{23}(s)n_{12}(-t)n_{23}(t) = n_{13}(ts)$. This shows that $\mu_x^{(13)}$ is proportional at x_0, and at $n_{13}(ts)x$.

This says something quite strong: the measure $\mu_x^{(13)}$ on the real line is proportional to its translate under ts! A simple auxiliary argument shows that we can find enough (t, s) to force $\mu_x^{(13)}$ to be Lebesgue measure on $\mathbb{R}$; so (by property (4) of conditional measures) μ is *invariant by* N_{13}. At this point we have invariance in a *unipotent* direction; and one may apply the measure-theoretic version of Ratner's theorem (Theorem 3.1, see also discussion of §3.3) to classify possibilities for μ.[15]

In words, (5.1) combines with the noncommutativity of the subgroups N_{ij} to show that μ is invariant in a unipotent direction.

The conclusion of this line of reasoning is the following, part of [4, Theorem 4.2]:

5.1. **Theorem.** *Suppose μ is an A_3-ergodic measure on $\mathcal{L}_3$ so that, for every $i \neq j$ and for a positive measure set of $x \in X$, the measure μ_x^{ij} is nontrivial. Then μ is Haar measure.*

Here *Haar measure* refers to the unique $\mathrm{SL}_3(\mathbb{R})$-invariant probability measure on $\mathcal{L}_3$. New ideas introduced by Lindenstrauss (based on his earlier work [11]) allowed this to be refined to the following result, which is (as we briefly discuss in §5.6) equivalent up to rephrasing to Theorem 4.3.

5.2. **Theorem.** *Suppose μ is an A_3-ergodic measure on $\mathcal{L}_3$ so that, for* at least one pair *$i \neq j$ and for a positive measure set of $x \in X$, the measure μ_x^{ij} is nontrivial. Then μ is Haar measure.*

Suitable analogues of these theorems are true replacing $(A_3, \mathcal{L}_3)$ by $(A_n, \mathcal{L}_n)$. In that case there are, in general, more possibilities for μ besides Haar measure, as in the statement of Theorem 4.3.

The question of removing the assumption in Theorem 5.2 seems to be a very difficult and fundamental one. If one could do so, the Littlewood conjecture (without any set of exceptions) would follow.

5.5. **Back to Theorem 1.1.** Now let's return to Theorem 1.1, which can be attacked using Theorem 5.2 and the relation between sets and measures.

We claim that for any fixed positive δ,

$$\text{BoxDimension}\,\{(\alpha, \beta) : \inf n.\|n\alpha\|.\|n\beta\| \geq \delta\} = 0 \tag{18}$$

from this it is easy to deduce Theorem 1.1.

[15] In fact, in [4], the use of Ratner's theorem was avoided by applying this argument repeatedly, with 13 replaced by various ij.

We saw that in the discussion preceding (10) that the failure of the Littlewood conjecture for a fixed pair (α, β) would correspond to the A_3^+-orbit of a certain $L_{\alpha,\beta} \in \mathcal{L}_3$ being bounded. If (18) fails, indeed, there exists a set of lattices $L_{\alpha,\beta}$ of box dimension ≥ 0.01 (say) whose A_3^+-orbits all remain within a fixed bounded set inside $\mathcal{L}_3$.

So the closure Y of $A_3^+.\{L_{\alpha,\beta}\}$ is a bounded, A_3^+-invariant closed set on $\mathcal{L}_3$ with box dimension ≥ 2.01 ((the extra 2 comes from taking the A_3-orbit; in words it means that Y has thickness transverse to the A_3-direction).

In what follows, let us ignore the distinction between A_3^+ and A_3 for simplicity. The necessity of dealing with A_3^+ complicates the argument slightly. So, let us assume that Y was actually A_3-invariant.

We construct a A_3-invariant measure μ supported on Y. It turns out that the fact that Y has thickness transverse to the A_3-direction translates into the fact that it is possible to choose μ so that at least one of the conditional measures μ_{ij}^x is nontrivial for almost all x. But then Theorem 5.2 shows that μ has to be Haar measure. So the support of μ is all of $\mathcal{L}_3$ and μ cannot be supported on the bounded set Y – a contradiction.

We observe that the need to allow a set of exceptions in Theorem 1.1 arises from the condition in Theorem 5.2 concerning conditional measures (equivalently, the positive entropy condition – see below). Removing that condition would settle the Littlewood conjecture in whole.

5.6. **Positive entropy.** The theorems 5.1 and 5.2 are not useful without a reasonable way to verify the conditions on μ_x^{ij}. The utility of these results stem, in enormous part, from the fact that there *is* a very usable way to verify the conditions. This is provided by the theory of entropy, and, in many other applications, it is through entropy that these conditions have been verified. Indeed, even the discussion in §5.5 is made rigorous using entropy.

The importance of entropy justifies ending this paper with a brief discussion. For more, see [7, Section 3].

The theory of metric entropy assigns to a measure-preserving transformation T of a probability space (X, μ) a non-zero number, the *entropy* $h_\mu(T)$ of T. We briefly reprise the definition, which, of course, is rather little use without motivation. If $\mathcal{P}$ is a partition of the probability space (X, μ), the entropy of $\mathcal{P}$ is defined as $h_\mu(\mathcal{P}) := \sum_{S \in \mathcal{P}} -\mu(S) \log \mu(S)$. We define the entropy of T as:

$$h_\mu(T) = \sup_{\mathcal{P}} \lim_{n \to \infty} \frac{h_\mu(\mathcal{P} \vee T^{-1}\mathcal{P} \vee \cdots \vee T^{-(n-1)}\mathcal{P})}{n} \tag{19}$$

where the supremum is taken over all finite partitions of X.

Roughly speaking, this means the following. Suppose for simplicity that there exists a finite partition $\mathcal{P}$ attaining the supremum on the right-hand side of (19). The entropy measures, in a suitable average sense, the amount of extra information required to specify which part $x \in X$ belongs to, *given* that one knows which part $Tx, T^2x, T^3x, \dots$ belong to. One bit of information corresponds to entropy $\log 2$.

(Clearly the above is not a complete description, because, besides being very ill-defined, it made no mention of the measure μ!)

For example, if $X = \mathbb{R}/\mathbb{Z}, T(x) = 2x, \mathcal{P} = \{[0, 1/2), [1/2, 1)\}$, the knowledge of $T^k x$ specifies the $(k+1)$st binary digit of x. So it requires one extra bit to specify x given $\{Tx, T^2x, \dots, \}$, which corresponds to entropy $= \log 2$

On the other hand, if $X = \mathbb{R}/\mathbb{Z}, T(x) = x + \sqrt{2}, \mathcal{P} = \{[0, 1/2), [1/2, 1)\}$, the entropy is 0: if we know the first binary digit of $\{Tx, T^2x, \dots\}$, we also know the first binary digit of x.

Thus, to a very crude approximation, positive entropy arises from the possibility of different $x, x' \in X$ having "similar" forward trajectories $\{Tx, T^2x, T^3x, \dots\}$. But, in the context of $(A_n, \mathcal{L}_n)$ there is a simple reason this could happen: if $x = n_{ij}x'$ and $a \in A_n$ contracts n_{ij} (cf. discussion near (13)), then the points $a^k x, a^k x'$ become very close as $k \to \infty$.

Formalizing this reasoning gives:

5.3. **Theorem.** *Let μ be an A_3-invariant probability measure on $\mathcal{L}_3$. Then $h_\mu(a) = 0$ for all $a \in A_3$ if and only if, for almost all $x \in \mathcal{L}_3$, the conditional measures μ_x^{ij} are trivial.*

Thus Theorem 4.3 and Theorem 5.2 are indeed the same.

6. Acknowledgements

I would like to thank Einsiedler, Lindenstrauss and Gregory Margulis for very many fruitful discussions and explanations relating to their work on this topic. I would also like to thank Sinan Güntürk for his comments on this paper.

References

[1] S. G. Dani and G. A. Margulis. Limit distributions of orbits of unipotent flows and values of quadratic forms. In *I. M. Gelfand Seminar*, volume 16 of *Adv. Soviet Math.*, pages 91–137. Amer. Math. Soc., Providence, RI, 1993.

[2] J. Ellenberg and A. Venkatesh. Local-global principles for representations of quadratic forms. `arxiv: math.NT/0604232`.

[3] M. Einsiedler and A. Katok. Rigidity of measures – the high entropy case, and non-commuting foliations. *to appear in Israel J. Math.*

[4] M. Einsiedler and A. Katok. Invariant measures on G/Γ for split simple Lie-groups G. *Comm. Pure Appl. Math.*, 56(8):1184–1221, 2003.

[5] M. Einsiedler, A. Katok, and E. Lindenstrauss. Invariant measures and the set of exceptions to Littlewood's conjecture. to appear in Ann. of Math.

[6] Manfred Einsiedler and Elon Lindenstrauss. On measures invariant under maximal split tori for semisimple S-algebraic groups. in preparation, 2005.

[7] Manfred Einsiedler and Elon Lindenstrauss. Diagonal flows on locally homogeneous spaces and number theory. to appear in the Proceedings of the International Congress of Mathematicians 2006 (29 pages), 2006.

[8] Manfred Einsiedler, Elon Lindenstrauss, Philippe Michel and Akshay Venkatesh. The distribution of periodic torus orbits on homogeneous spaces. `arxiv: math.DS/0607815`.

[9] Y. Katznelson. Chromatic numbers of Cayley graphs on $\mathbb{Z}$ and recurrence. *Combinatorica* 21 (2001).

[10] Elon Lindenstrauss. Arithmetic quantum unique ergodicity and adelic dynamics. Proceedings of Current Developments in Mathematics conference (2004), to appear.

[11] Elon Lindenstrauss. Invariant measures and arithmetic quantum unique ergodicity. *Annals of Math*, 163 (2006).

[12] Elon Lindenstrauss. Rigidity of multiparameter actions. *Israel J. of Math*, 149 (2005)

[13] Gregory Margulis. Problems and conjectures in rigidity theory. In *Mathematics: frontiers and perspectives*, pages 161–174. Amer. Math. Soc., Providence, RI, 2000.

[14] Andrew Pollington and Sanju Velani. On a problem in simultaneous Diophantine approximation: Littlewood's conjecture. *Acta. Math* 185 (2000)

[15] Marina Ratner. On Raghunathan's measure conjecture. *Ann. of Math. (2)*, 134(3):545–607, 1991.

[16] Marina Ratner. Raghunathan's topological conjecture and distributions of unipotent flows. *Duke Math. J.*, 63(1):235–280, 1991.

[17] Lior Silberman and Akshay Venkatesh. Quantum unique ergodicity for locally symmetric spaces. `math.RT/0407413`, to appear, *GAFA*.

[18] David Witte. Ratner's theorems on unipotent flows. *Chicago Lectures in Mathematics Series.*

FROM HARMONIC ANALYSIS TO ARITHMETIC COMBINATORICS

IZABELLA ŁABA

ABSTRACT. We will describe a certain line of research connecting classical harmonic analysis to PDE regularity estimates, an old question in Euclidean geometry, a variety of deep combinatorial problems, recent advances in analytic number theory, and more.

Traditionally, restriction theory is a part of classical Fourier analysis that investigates the relationship between geometric and Fourier-analytic properties of singular measures. It became clear over the years that the theory would have to involve sophisticated geometric and combinatorial input. Two particularly important turning points were Fefferman's work in the 1970s invoking the "Kakeya problem" in this context, and Bourgain's application of Gowers's additive number theory techniques to the Kakeya problem almost 30 years later.

All this led harmonic analysts to explore areas previously foreign to them, such as combinatorial geometry, graph theory, and additive number theory. Although the Kakeya and restriction problems remain stubbornly open, the exchange of knowledge and ideas has led to breathtaking progress in other directions, including the Green-Tao theorem on arithmetic progressions in the primes. The level of interest in the subject has skyrocketed since then, and many exciting developments are sure to follow.

PROLOGUE

In April 2004, the mathematical world was jolted wide awake as Ben Green and Terence Tao announced their proof of the long-standing conjecture that primes contain arbitrarily long arithmetic progressions. Theirs was a stunning piece of work, not only in its originality and ingenuity, but also in the breadth of mathematical territory that it covered. The proof blended seamlessly a multitude of ideas from number theory, combinatorics, harmonic analysis and ergodic theory. The subsequent Green-Tao papers made it clear that their breakthrough result was only the first step in a far-reaching program of research, inspired by the Hardy-Littlewood conjecture in analytic number theory.

To say that many were taken by surprise would be an understatement. Green had just completed his Ph.D. degree less than a year earlier, and Tao was already known as an brilliant mathematician but he had never worked in

Received by the editors November 15, 2006.

2000 *Mathematics Subject Classification.* Primary 28A78, 42B15, 11P99; Secondary 11B25, 01A05.

Key words and phrases. Harmonic analysis, analytic and combinatorial number theory.

Supported in part by NSERC Discovery Grant.

analytic number theory until then. While they had been aware of each other's work much earlier, they did not meet and start to collaborate until early 2004. Their primes paper was then completed within just a few months.

This work, however, was not simply conjured out of thin air. It was built upon decades of research by many excellent mathematicians, working in rather diverse fields and not always concerned with any sort of arithmetic progressions. It then drew on the ideas and experience of the earlier contributors to this area, including Szeméredi, Furstenberg, Bourgain, Gowers, and others. Green and Tao studied their work in depth, molded and rearranged it, long before they embarked on a collaboration. They did truly stand on the shoulders of giants.

The ergodic-theoretic background of the Green-Tao work was surveyed in Bryna Kra's 2005 Current Events Bulletin talk and in the article [44]. Here we will focus mostly on harmonic analysis, but with some combinatorics and additive number theory also mixed in. It is far from my intentions to suggest that the work described here is merely a background for the Green-Tao theorem. On the contrary, the questions mentioned here and the areas of research that they represent are fascinating in their own right, and they still would be if Green and Tao had never met.

To keep this presentation reasonably short and coherent, I will limit it to a few problems in each area, selected with a view to showcasing the often unexpected paths between them. Even so, the list of references has repeatedly threatened to run out of control. I hope to expand this to a longer article in the future; meanwhile, I can only invite the reader to enjoy the story and, should he wish to learn more, refer him to the more thorough and specialized surveys cited in the text.

.

1. The Kakeya Problem

1.1. **Life during wartime.** By all accounts, Abram Samoilovitch Besicovitch (1891-1970) had an interesting life. He was born in 1891 in Berdyansk, in the south of Russia. Having demonstrated exceptional mathematical abilities at an early age, he went on to study under the direction of the famous probabilist A.A. Markov at the University of St. Petersburg, from which he graduated in 1912.

The University of Perm was established in October 1916, first as a branch of the University of St. Petersburg and then as an independent institution. Perm, located in the Ural Mountains, was closed off to foreign visitors from the 1920s until 1989, and the university, which remains the main intellectual center of the region, has seen difficult times. But in the hopeful early years (1916–1922), it managed to attract many brilliant and ambitious young academics. Besicovitch was appointed professor of mathematics at the University of Perm in 1917. Among his colleagues were the mathematician I.M. Vinogradov, of the three-primes theorem in analytic number theory, and the physicist A.A. Friedmann, best known for his mathematical models of the “big bang" and the expanding universe.

After several months of political unrest, the Bolshevik Revolution erupted in October 1917. Soon thereafter a civil war engulfed Russia. The White Army, led by former Tsarist officers, opposed the communist Red Army. Perm was controlled by the Red Army until December 1918, when the White Army took over. In August 1919 the Red Army returned. According to Friedmann, all the staff except Besicovitch left the university:

> *The only person who kept his head and saved the remaining property was Besicovitch, who is apparently A.A. Markov’s disciple not only in mathematics but also with regard to resolute, precise definite actions.*

In 1920 Besicovitch returned to St. Petersburg, which had been renamed Petrograd six years earlier, and accepted a position at Petrograd University. (Petrograd would change names twice more: it became Leningrad after Lenin’s death in 1924, and in 1991 it reverted to its original name St. Petersburg.) The war years had not been kind to Petrograd. The city lost its capital status to Moscow in 1918, the population dwindled to a third of its former size, and the economy was in tatters. This is how *Encyclopedia Britannica* describes the education reform in the newborn Soviet Union:

> *To destroy what they considered the “elitist” character of Russia’s educational system, the communists carried out revolutionary changes in its structure and curriculum. All schools,from the lowest to the highest, were nationalized and placed in charge of the Commissariat of Enlightenment. Teachers lost the authority to enforce discipline in the classroom. Open admission to institutions of higher learning was introduced to assure that anyone who desired, regardless of qualifications, could enroll. Tenure for university professors was abolished, and the universities lost their traditional right of self-government.*

Besicovitch was awarded a Rockefeller Fellowship in 1924, but was denied permission to leave Russia. He escaped illegally, along with his colleague J.D. Tamarkin, and took up his fellowship in Copenhagen, working with Harald Bohr. After a brief stay in Liverpool (1926-27), he finally settled down in Cambridge, where he spent the rest of his life. >From 1950 until his retirement in 1958, he was the Rouse Ball Professor of Mathematics; this is the same chair that was held by John Littlewood prior to Besicovitch’s tenure, and is currently being held by W.T. Gowers, whose work will play a major part later in this story.

Besicovitch will be remembered for his contributions in the theory of almost periodic functions (a subject to which Bohr introduced him in Copenhagen) and other areas of function theory, and especially for his pioneering work in geometric measure theory, where he established many of the fundamental results. He was a powerful problem solver who combined a mastery of weaving long and intricate arguments with a capacity to approach a question from completely unexpected angles. His solution of the Kakeya problem, to which we are about to turn, is a prime example of his ingenuity.

1.2. **Riemann integrals and rotating needles.** Sometime during his Perm period, between the comings and goings of the Red and White Armies, Besicovitch worked on a problem in Riemann integration:

> *Given a Riemann-integrable function f on $\mathbb{R}^2$, must there exist a rectangular coordinate system (x, y) such that $f(x, y)$ is Riemann-integrable as a function of x for each y, and that the two-dimensional integral of f is equal to the iterated integral $\int\int f(x, y)dxdy$?*

He observed that to answer the question in the negative it would suffice to construct a set of zero Lebesgue measure in $\mathbb{R}^2$ containing a line segment in every direction. Specifically, suppose that E is such a set, and fix a coordinate system in $\mathbb{R}^2$. Let f be the function such that $f(x, y) = 1$ if $(x, y) \in E$ and if at least one of x, y is rational, and $f(x, y) = 0$ otherwise. We may also assume, shifting E if necessary, that the x- and y-coordinates of the line segments parallel to the y- and x-axes, respectively, are irrational. Then for every direction in $\mathbb{R}^2$, there is at least one line segment in that direction along which f is not Riemann-integrable as a function of one variable. However, f is Riemann-integrable in two dimensions, as the set of its points of discontinuity has planar measure 0.

Besicovitch then proceeded to construct the requisite set E. This, along with the solution of the Riemann integration problem, was published in a Perm scientific journal in 1919 [2]. I wonder if any copies of that article have survived!

The construction is roughly as follows. We start with a triangle ABC, which contains line segments in all directions from AB to AC. We divide it into many long and this triangles with one vertex at A and the other two on the base line segment BC, then rearrange them by sliding them along the base. This can be done so that the rearranged set has area less than any small constant fixed in advance. Iterating the construction and then taking the limit, we obtain a set of measure 0. The details of the construction can be found in many books and articles, for example [18], [54], [67]. There have been many subsequent improvements and simplifications of Besicovitch's construction, by Perron, Schoenberg, and many other authors including Besicovitch himself.

Independently but around the same time (1917), the Japanese mathematician Soichi Kakeya proposed a similar question which became known as the *Kakeya needle problem*:

> *What is the smallest area of a planar region within which a unit line segment (a "needle") can be rotated continuously through 180 degrees, returning to its original position but with reversed orientation?*

Kakeya [39] and Fujiwara-Kakeya [23] conjectured that the smallest *convex* planar set with this property was the equilateral triangle of height 1, and mentioned that one could do better if the convexity assumption was dropped. For example, the region bounded by a three-cusped hypocycloid inscribed in a circle of diameter 1 has the required property and has area $\pi/8 \approx .39$, whereas the area of the triangle is $\sqrt{3}/3 \approx 0.58$. Kakeya's conjecture for the convex case was soon confirmed by Julius Pál (1921), but the more interesting non-convex problem remained open.

Due to the civil war, there was hardly any scientific communication between Russia and the Western world at the time. Both Besicovitch and Kakeya were unaware of each other's work. Besicovitch learned of Kakeya's problem after he left Russia, possibly from a 1925 book by G.D. Birkhoff which he mentions in [4], and realized that a modification of his earlier construction provided the unexpected answer:

> *For any $\epsilon > 0$, there is a planar region of area less then ϵ within which a needle can be rotated through 180 degrees.*

His solution was published in 1928 [3]. There are now many other such constructions, some with additional conditions on the planar region in question.

1.3. **The Kakeya conjecture.**

Definition 1.1. A Kakeya set, or a Besicovitch set, is a subset of $\mathbb{R}^d$ which contains a unit line segment in each direction.

Besicovitch's construction shows that Kakeya sets in dimension 2 can have measure 0. With this information, it is easy to see that the same is true in higher dimensions: let E be a planar Kakeya set of measure 0, then the set $E \times [0,1]^{d-2}$ in $\mathbb{R}^d$ is a Kakeya set and has d-dimensional measure 0.

The following conjecture, however, remains open for all $d \geq 3$:

Conjecture 1.2. *A Kakeya set in $\mathbb{R}^d$ must have Hausdorff dimension d.*

In dimension 2, this was first proved by Davies [16] in 1971; an important alternative argument was given later by Córdoba [14].

The current interest in the Kakeya conjecture is largely motivated by problems in harmonic analysis. Analysts quickly realized that Besicovitch's construction of Kakeya sets of measure zero, along with a closely related construction due to Nikodym (1927), could be used to produce counterintuitive examples involving maximal functions and differentiation of integrals (see e.g. [11]). However, it was not until the 1970s and 80s that substantial qualitative differences between the planar and higher-dimensional cases were brought to light, and it gradually became understood that Conjecture 1.2 (along with its stronger *maximal function* variant) is the key question to consider. This will be discussed in more detail in the next section, after which we will return to the Kakeya conjecture and the progress that has been made so far.

2. Questions in harmonic analysis

2.1. **The restriction problem.**

The Fourier transform of a function $f : \mathbb{R}^d \to \mathbb{C}$ is defined by

$$\widehat{f}(\xi) = \int f(x) e^{-2\pi i x \cdot \xi} dx.$$

This maps the Schwartz space of functions S to itself, and is clearly a bounded operator from $L^1(\mathbb{R}^d)$ to $L^\infty(\mathbb{R}^d)$. A basic result in harmonic analysis is that the Fourier transform extends to an isometry on $L^2(\mathbb{R}^d)$; furthermore, by the Hausdorff-Young inequality the Fourier transform is also bounded from $L^p(\mathbb{R}^d)$ to $L^{p'}(\mathbb{R}^d)$ if $1 < p < 2$ and $\frac{1}{p} + \frac{1}{p'} = 1$.

The following question has become known as the *restriction problem*:

Let μ be a non-zero measure on $\mathbb{R}^d$. For what values of p', q' does the Fourier transform, defined on S, extend to a bounded operator from $L^{q'}(\mathbb{R}^d)$ to $L^{p'}(d\mu)$? In other words, when do we have an estimate

$$\|\widehat{f}\|_{L^{p'}(d\mu)} \leq C\|f\|_{L^{q'}(\mathbb{R}^d)},\ f \in S? \tag{2.1}$$

We will usually assume that the measure μ is finite. Here and below, C and other constants may depend on the dimension d, the measure μ, and the exponents p, q, but not on f except where explicitly indicated otherwise.

In the classical version of the problem, μ is the Lebesgue measure on a $d-1$-dimensional hypersurface Γ in $\mathbb{R}^d$, e.g. a sphere or cone. Then the above question can be rephrased in terms of *restricting* the Fourier transform of an $L^{q'}$ function f to the hypersurface. This is trivial if $q' = 1$, since then $\widehat{f}$ is continuous and bounded everywhere, in particular on Γ. On the other hand, it is easy to see that no such result is possible if $q' = 2$. This is because the Fourier transform maps L^2 *onto* L^2, so that we are not able to say anything about the behaviour of $\widehat{f}$ on a set of measure 0. It is less clear what happens for $q' \in (1, 2)$. As it turns out, the answer here depends on the geometry of Γ: for example, there can be no estimates such as (2.1) with $q' > 1$ if Γ is a hyperplane, but we do have nontrivial restriction estimates for a variety of curved hypersurfaces, some of which will be discussed shortly.

The reason for the somewhat curious notation so far is that we reserved the exponents p, q for the dual formulation of the problem. We will write $\widehat{fd\mu}(\xi) = \int f(x)e^{-2\pi i x\cdot\xi}d\mu(x)$.

Let μ be a non-zero measure on $\mathbb{R}^d$. For what values of $p.q$ do we have an estimate

$$\|\widehat{fd\mu}\|_{L^q(\mathbb{R}^d)} \leq C\|f\|_{L^p(d\mu)},\ f \in S? \tag{2.2}$$

A reasonably simple argument shows that (2.2) and (2.1) are equivalent if p, p' and q, q' are pairs of dual exponents: $\frac{1}{p} + \frac{1}{p'} = \frac{1}{q} + \frac{1}{q'} = 1$. While the restriction problem took its name from the first formulation (2.1), the second one turns out to be much more useful in applications.

In the case when μ is the surface measure on a hypersurface Γ with non-vanishing Gaussian curvature, classical stationary phase estimates (e.g. [36]) yield asymptotic expressions for $\widehat{fd\mu}(\xi)$ if f is a smooth compactly supported function on Γ. In particular, we then have

$$|\widehat{fd\mu}(\xi)| = O((1 + |\xi|)^{-\frac{d-1}{2}}), \tag{2.3}$$

and it follows that $\widehat{fd\mu} \in L^q(\mathbb{R}^d)$ for $q > \frac{2d}{d-1}$. A wide variety of similar estimates has been obtained under weaker assumptions on the curvature of Γ, for example “finite type" surfaces and surfaces with less than $d-1$ nonvanishing principal curvatures are allowed. A comprehensive survey of such work up to 1993 is given in [54] (see also [37]).

The point of the restriction estimates is that we no longer expect our functions to be smooth, and that our estimates are intended to be uniform in L^q norms, regardless of the smoothness of the data. This is particularly useful in applications to PDE questions. Much as stationary phase estimates are ubiquitous in traditional linear PDE theory, restriction estimates can be used to prove regularity estimates if we only know that the initial data is in some L^p space and expect L^q or mixed-norm regularity, rather than smoothness, of the solution. For example, restriction estimates are very closely related to *Strichartz estimates* [55]. We will not attempt to survey this rich and complex area here, instead referring the reader to references such as e.g. [52], [54], [62], [58], [71]. The same references elaborate on many other problems in harmonic analysis, involving oscillatory integrals, maximal functions, averaging operators and Fourier integral operators, which bear close relations to restriction estimates as well as to one another.

2.2. **Restriction for the sphere and arrangements of needles.** We will now take a closer look at the restriction phenomenon for the sphere S^{d-1} in $\mathbb{R}^d$. Let σ be the normalized surface measure on S^{d-1}. The following conjecture is due to Elias M. Stein:

Conjecture 2.1. *For all $f \in L^\infty(S^{d-1})$, we have*

$$\|\widehat{fd\sigma}(\xi)\|_q \le C\|f\|_\infty,\ q > \frac{2d}{d-1}. \tag{2.4}$$

This is known for $d = 2$ (due to Fefferman and Stein [19]), but remains open for all $d > 2$. The range of q is suggested by stationary phase formulas such as (2.3). Plugging in $f \equiv 1$ shows that this range cannot be improved. Indeed, $\widehat{d\sigma}$ can be computed explicitly:

$$\widehat{d\sigma}(\xi) = 2|\xi|^{-\frac{d-1}{2}}\cos(2\pi(|\xi| - \frac{d-1}{8})) + O(|\xi|^{-\frac{d+1}{2}}),$$

which belongs to $L^q(\mathbb{R}^d)$ only for q exactly as indicated above.

If instead of assuming that $f \in L^\infty$ we make the weaker assumption that $f \in L^2(S^{d-1})$, then the best possible result is known [64], [65], [53]:

Theorem 2.2. *(Tomas-Stein) Let $f \in L^2(S^{d-1})$, then*

$$\|\widehat{fd\sigma}(\xi)\|_q \le C\|f\|_{L^2(S^{d-1})},\ q \ge \frac{2d+2}{d-1}. \tag{2.5}$$

This was first proved by Stein in 1967 (unpublished) for a smaller range of q. In 1975 P.A. Tomas extended the result to $q < \frac{2d+2}{d-1}$, and later that year the endpoint was settled by Stein. A simple construction known as the *Knapp counterexample* shows that the range of q in Theorem 2.2 is optimal.

The Tomas-Stein argument is very general and uses only limited information about the geometry of S^{d-1}, namely its dimensionality and the decay of

$\widehat{\sigma}$ at infinity. Large parts of the proof can be adapted to different or more general settings; in fact, later on we will see a very similar argument applied to a number-theoretic problem.

One can interpolate between Tomas-Stein and the trivial L^1-L^∞ estimate to get a range of intermediate estimates. Going beyond that, however, was much more difficult, and for many years, until Bourgain's breakthrough in 1991 [7], it was not even known whether this was possible at all. It turns out that a substantially new approach was required. While Theorem 2.2 is mostly based on analytic considerations, restriction estimates such as (2.2) with $p > 2$ require deeper geometrical information, and this is where we discover Kakeya sets lurking under the surface.

Our starting point is that the restriction conjecture (2.4) implies the Kakeya conjecture (Conjecture 1.2). This was perhaps first stated and proved formally by Bourgain in [7], but very similar arguments were used in the harmonic analysis literature throughout the 1970s and 80s, all inspired by the work of Fefferman [20] where Besicovitch sets were used to produce a counterexample to the (closely related) ball multiplier conjecture. Below is a rough summary of this argument, adapted to the restriction setting.

Let $f(x) = e^{2\pi i\eta x}\chi_a(x)$, where $\eta \in \mathbb{R}^d$, $a \in S^{d-1}$, and χ_a is the characteristic function of the spherical cap centered at a of radius δ for some very small $\delta > 0$. Scaling considerations, standard in harmonic analysis, show that $\widehat{f}$ is roughly constant on tubes of length δ^{-2} and radius δ^{-1}. Forgetting about mathematical rigour for a moment, we will in fact think of $\widehat{f}$ as the characteristic function of one such tube. Moreover, by adjusting the phase factor η we can place that tube at any desired point in the dual space $\mathbb{R}^d_\xi$.

Now cover the sphere by such δ-caps, and let $F(x)$ be the sum of the associated functions defined above. Then $\|F\|_\infty \leq C$, uniformly in δ. On the other hand, $\widehat{F}$ is the sum of a large number of characteristic functions of tubes as described above. If we now arrange these tubes as in the Besicovitch set construction, then the size of the support of $\widehat{F}$ will be very small compared to its L^1 norm, and an application of Hölder's inequality shows that this forces the L^p norms of $\widehat{F}$ to be large. This can be worked out quantitatively, taking into account the many technicalities that we conveniently brushed off here, and the result follows.

The truly groundbreaking contribution of [7] was the discovery that this reasoning was, to some extent, reversible. More precisely, Bourgain developed an analytic machinery to deduce restriction estimates from Kakeya-type geometric information. It is a difficult and analytically sophisticated argument. First of all, it does not quite suffice to have a dimension bound for Kakeya sets in $\mathbb{R}^d$ – a stronger result expressed in terms of *maximal functions* is needed. This is followed by simultaneous analysis on two different scales (*local restriction estimates*), combining the maximal function result just mentioned with Tomas-Stein type orthogonality arguments. The numerology produced here is complicated and unclear, and there is no simple way to explain where the resulting values of the exponents p come from.

Bourgain's work was continued by other authors: Wolff (1995), Moyua-Vargas-Vega (1996), Tao-Vargas-Vega (1998), Tao-Vargas (2000), Tao (2003). While Wolff improved on Bourgain's result by producing a better Kakeya bound, other authors tended to focus on the Kakeya-to-restriction conversion mechanism. It should be added, though, that Wolff has also made indirect but crucial contributions of the second kind, as the analytic tools developed by him in other related contexts were then used by other authors (notably Tao) to make progress here. The updated toolbox includes bilinear restriction estimates, induction on scales, wave packet decompositions, local restriction estimates, and more. A comprehensive review of the modern approach to the subject is given in [59].

The current best result belongs to Tao [58], and can be explained as follows. Interpolating between the Stein-Tomas theorem (2.5) and the conjectured estimate (2.4), we get a family of conjectured intermediate estimates of the form (2.2). The challenge is to improve the range of p for which such estimates are known. Tao's result is that (2.2) holds with $p > \frac{2(n+2)}{n}$, if q is the corresponding exponent from the interpolation. This is obtained as a consequence (via scaling) of a bilinear restriction estimate for paraboloids, proved also in [58] and largely inspired by Wolff's sharp bilinear restriction estimate for the light cone [69].

3. The Kakeya problem revisited

We now return to the Kakeya conjecture in dimensions $d \geq 3$. Although the conjecture remains open, partial results are available in the form of lower bounds on the Hausdorff dimension of Besicovitch sets in $\mathbb{R}^d$, and it is this question that will concern us in this section.

In addition to the Hausdorff dimension, we will also consider the related but somewhat different notion of *Minkowski dimension*, defined as follows. For a compact set $E \subset \mathbb{R}^d$, we let E_δ be the δ-neighbourhood of E, and consider the asymptotic behaviour of the d-dimensional volume of E_δ as $\delta \to 0$. We say that E has Minkowski dimension α if the limit

$$\lim_{\delta \to 0} \log_\delta |E_\delta| \tag{3.1}$$

exists and is equal to $n - \alpha$; in other words, if we have $|E_\delta| \approx \delta^{d-\alpha}$. If the limit in (3.1) does not exist, we instead use the lower and upper limit in (3.1) to define the upper and lower Minkowski dimension, denoted by $\overline{\dim}_M(E)$ and $\underline{\dim}_M(E)$. We also use $\dim_H(E)$ to denote the Hausdorff dimension of E.

For all compact sets $E \subset \mathbb{R}^d$ we have $\dim_H(E) \leq \underline{\dim}_M(E) \leq \overline{\dim}_M(E)$, so that any lower bound on the Hausdorff dimension of Kakeya sets implies the same bound on the Minkowski dimension. However, the converse does not hold, and there are several results concerning the Minkowski dimension of Kakeya sets that so far have not been replicated for the Hausdorff dimension.

The Minkowski dimension has many disadvantages compared to the Hausdorff dimension, for example it is not associated with any countably additive measure and there are countable sets that have positive Minkowski dimension. However, its use will allow us to simplify considerably the exposition while retaining the essence of the proofs. In the sequel we will therefore focus on Minkowski dimension arguments even where Hausdorff versions are also available.

Prior to 1991, it was known that the Hausdorff dimension of a Kakeya set in $\mathbb{R}^d$ must be at least $(d+1)/2$. I was not able to determine where this first appeared explicitly, but it certainly follows from the x-ray and k-plane transform estimates of Drury [17] and Christ [13]. Bourgain's work [7] started a race to improve the known Kakeya bounds. In the next two subsections we give an account of the developments so far and sketch a few key arguments. A summary of the best known bounds at this time is given at the end of the section.

3.1. **Geometric arguments.** We begin with an argument due to Bourgain [7], known in the harmonic analysis community as the "bush argument", which provides a geometric proof of the previously mentioned bound $(d+1)/2$. Suppose that E is a Kakeya set in $\mathbb{R}^d$, then for each $e \in S^{d-1}$ E contains a unit line segment T^e in the direction of e. Let $\mathcal{E}$ be a maximal δ-separated subset of S^{d-1}, so that $|\mathcal{E}| \approx \delta^{-(d-1)}$, and let T^e_δ be the δ-neighbourhood of T^e. Abusing notation only very slightly, we write $E_\delta = \bigcup_{e\in\mathcal{E}} T^e_\delta$. Suppose that $\dim_M(E) = \alpha$, so that $|E_\delta| \approx \delta^{n-\alpha}$. Since $\sum_{e\in\mathcal{E}} |T^e_\delta| \approx 1$, there must be at least one point, say x_0, which belongs to at least $\delta^{-(n-\alpha)}$ tubes T^e_δ. The key observation is that these tubes are essentially disjoint (more precisely, have finite overlap) away from a small neighbourhood of x_0. (Two straight lines can only intersect at one point.) Thus $|E_\delta|$ is bounded from below by a constant times the sum of volumes of the tubes through x_0:

$$|E_\delta| \geq C\delta^{-(d-\alpha)} \cdot \delta^{d-1} = \delta^{\alpha-1}.$$

But this is only possible if $\alpha - 1 \leq d - \alpha$, i.e. $\alpha \leq \frac{d+1}{2}$.

In [7], this is supplemented by an additional geometrical argument improving the dimension bound to $\frac{d+1}{2} + \epsilon_d$, with ϵ_d given by a recursive formula (for $d = 3$ this yields the bound $7/3$).

A more efficient geometrical argument, leading to the estimate $\dim_H(E) \geq \frac{d+2}{2}$, was given a few years later by Tom Wolff [67]. Wolff observes that in order for E_δ to have small volume, it is necessary for a large fraction of the set, not just one point, to have high multiplicity. In fact, many of the tubes T^e_δ must consist largely of high multiplicity points. Take one such tube, along with the union of all tubes that intersect it (this object is often called "hairbrush"). By combining Bourgain's "bush" construction above with an earlier planar estimate due to Córdoba [14], one can prove that the bristles of the hairbrush must be essentially disjoint. We then bound the volume of E_δ from below by the volume of the hairbrush, and the Minkowski dimension estimate again follows upon taking $\delta \to 0$.

This comes with a few caveats. The argument does not quite work as stated and requires some modifications if the tubes of E_δ tend to intersect at very low angles. More importantly, there are additional issues that arise in the calculation of the Hausdorff dimension (as opposed to Minkowski). We will not elaborate on this here, but we do want to mention the *two ends reduction* of [67], which was introduced to resolve that problem and may well have inspired some of the induction on scales techniques in restriction theory.

Wolff's argument, although more elaborate than Bourgain's, is still relatively simple in the sense that only very basic geometric information is being used, and it was tempting to try to improve on it by using more sophisticated combinatorial methods. This is how harmonic analysts were introduced to *combinatorial geometry*, an area of combinatorics which studies, among other things, arrangements of lines, planes and other geometric objects in Euclidean space. Of particular interest here are combinatorial bounds on the number of *incidences* between points and objects such as lines, curves or surfaces. (A curve is *incident* to a point if the point lies on the curve.) A classic result of this type is the Szemerédi-Trotter theorem giving a bound $O(n + m + n^{2/3}m^{2/3})$ on the number of incidences between n lines and m points in $\mathbb{R}^2$; we invite the reader to consult the review article [47] for an overview of this fascinating subject and many more examples of estimates of this type.

The use of incidence geometry in harmonic analysis – essentially, decomposing functions into “wave packets", then treating the latter as thin geometric objects and applying combinatorial methods to deduce information about their possible arrangements – was pioneered by Wolff in the 1990s. While the Kakeya problem resisted this approach, Wolff was much more successful with other questions, for example the local smoothing problem for the wave equation whose solution [70] required obtaining deep geometric information about arrangements of circles. Just as importantly, ongoing communication was gradually established between discrete geometers and harmonic analysts. Many more intriguing connections between the two areas have since been uncovered and continue to be pursued.

3.2. **Additive and hybrid arguments.** A radically different “arithmetic" approach to the problem was introduced by Bourgain in 1998 [10]. Let us forget about the hairbrush construction for a moment, and try to improve on the bush argument instead in another direction. Suppose that we are given a hypothetical Kakeya set $E \subset \mathbb{R}^d$ of dimension close to $(d+1)/2$. We perform a discretization procedure as in the last subsection, except that we will now ignore the distinction between a tube and a line. (This is cheating, but it is good for the exposition.) We will also restrict our attention to those lines which make an angle less than $\pi/100$ with the x_d-axis. Consider the intersections A, B, C of the discretized set E with the three parallel hyperplanes $x_d = 0$, $x_d = 1$, $x_d = 1/2$ (rescale and translate the set if necessary). We consider A, B, C as subsets of $\mathbb{R}^{d-1}$. Let $S = \{(a,b) : \text{there is a line from } a \text{ to } b\}$. Then

$$\{(a+b)/2 : (a,b) \in S\} \subset C.$$

The key result is the following lemma.

Lemma 3.1. *Let A, B be two subsets of $\mathbb{Z}^d$ of cardinality $\leq n$, and let $S \subset A \times B$. If $|\{a+b : (a,b) \in S\}| \leq Cn$, then*

$$|\{a-b : (a,b) \in S\}| \leq C' n^{2-\frac{1}{13}}.$$

We will say more about Lemma 3.1 later on, but first we will see how it applies to our setting. Due to multiplicity considerations similar to those in the last subsection, we have $|A|, |B|, |C| \leq n$ with n close to $\delta^{-(d-1)/2}$. The lemma then states that $|\{a-b : (a,b) \in S\}| \leq cn^{2-1/13}$. But the last set includes the set of “all" directions, hence it must have cardinality about $\delta^{-(d-1)}$, which is greater than the lemma allows if n is too close to $\delta^{-(d-1)/2}$.

Bourgain worked out a quantitative version of this in [10], obtaining a lower bound $(13d+12)/25$ for the dimension of the Kakeya sets in $\mathbb{R}^d$, which is better than Wolff’s result in high dimensions. The Minkowski dimension argument is more or less as described above, but the Hausdorff and maximal function version present many additional difficulties in arranging a setup in which the lemma can be applied, and one cannot help but admire Bourgain’s ingenuity in overcoming this.

The bounds in [10] have since been improved in various ways. The arithmetic approach was developed further by Katz and Tao [41], [42], first by improving the bound in Bourgain’s lemma and then by using more than three “slices". There are also hybrid arguments [40], [42], combining Wolff’s geometric combinatorics with Bourgain’s arithmetic method. We embarked on the work [40]. in two separate groups, with the expectations that Wolff’s hairbrush estimate could be improved by more sophisticated geometrical arguments... but we found that this was just not going to happen, at least not in three dimensions. Our collection of geometrical observations (many of which were due to Tom Wolff or inspired by him) was growing, but it still did not add up to an improved bound. That was only achieved when we turned to Bourgain’s approach, first using geometrical techniques to effectively factor out one dimension. (On the other hand, a similar result in higher dimensions [45] involves only geometry and no additive techniques.)

Finally, we present the somewhat complicated list of the current best lower bounds on the dimension of Besicovitch sets in $\mathbb{R}^d$. We start with the Minkowski results:

- $d = 3$: $5/2 + 10^{-10}$ (Katz-Łaba-Tao 1999)
- $d = 4$: 3 + 10-10 (Łaba-Tao 2000)
- $4 < d < 24$: $(2 - 21/2)(d-4) + 3$ (Katz-Tao 2001)
- $d \geq 24$: $(d+t-1)/t$, where $t = 1.67513...$ is the root of $t^3 - 4t + 2 = 0$ that lies between 1 and 2 (Katz-Tao 2001).

The Hausdorff list is shorter:

- $d = 3, 4$: $(d+2)/2$ (Wolff 1994)
- $d > 4$: $(2 - 21/2)(d-4) + 3$ (Katz-Tao 2001)

The reader may have forgotten by now that we still have not said anything about Bourgain's lemma. We will do that now, and this will take us into the very different realm of *additive number theory*. Lemma 3.1 is actually a modification of a result of Gowers [26], [27] which in turn is a quantitative version of a result known as the Balog-Szemerédi theorem. We will explain this in more detail in the next section.

This is a good moment to say that it was the connection between these questions and the Kakeya conjecture, via Lemma 3.1 and Bourgain's work in [10], that attracted many harmonic analysts to additive number theory and inspired us to work on its problems. The Green-Tao theorem and many other developments might have never happened, were it not for Bourgain's brilliant leap of thought in 1998.

4. Additive number theory

Additive number theory is a mixture of number theory, combinatorics, and discrete harmonic analysis, applied in various proportions to problems concerning additive properties of sets of numbers. The questions of interest are often stated in the language of first-grade arithmetic: addition, multiplication, and counting of integers. Yet, starting with those most basic ingredients, one weaves a surprisingly rich tapestry of techniques and results. We are actually interested in a certain subfield of additive number theory that can be hard to define, but is often thought to be closer to combinatorics than to the rest of number theory. Below we describe two results that are central to, and representative of, this field: Freiman's theorem and Szemerédi's theorem. There are excellent expositions and surveys of the area, for example [15], [28] or [63], where the interested reader will find more information.

4.1. **Freiman's theorem.** Let $A \subset \mathbb{Z}$ be a finite set, and let $A + A = \{a + b : a, b \in A\}$. It is easy to prove that $|A + A| \geq 2|A| - 1$, and that the equality is attained if and only if A is an arithmetic progression. But what if we only know that $|A + A| \leq C|A|$ for some (possibly large) constant C? Does this imply that A has arithmetic structure? Of course arithmetic progressions still qualify, but so do more general lattice-like sets of the form

$$A = \{a_0 + j_1 r_1 + \cdots + j_m r_m : 0 \leq j_i \leq J_i,\ i = 1, \ldots, m\}, \tag{4.1}$$

with m small enough depending on C. Such sets are called *generalized arithmetic progressions* of dimension m. Freiman's theorem [21], [22] asserts that all sets with small sumsets are essentially of this form:

Theorem 4.1. *Suppose that $A \subset \mathbb{Z}$ and that $|A + A| \leq C|A|$. Then A is contained in a generalized arithmetic progression (4.1) of size at most $C'|A|$ and dimension m, where C' and m depend only on C.*

Following Freiman's work, there have been several other proofs of Theorem 4.1, by Bilu [6], Ruzsa [49], [50], [51], and Chang [12], where the current best quantitative bounds were obtained.

Freiman's theorem has a variety of extensions and generalizations. It can be extended to more general abelian groups – the most general result of this type was recently obtained by Green and Ruzsa. In a different direction, the Balog-Szemerédi theorem [1] addresses the case when we do not know the size of the entire sumset $A+A$, assuming instead that the set $\{a+a' : (a,a') \in S\}$ is small for a large set $S \subset A \times A$. It was a quantitative version of this theorem that was required in Gowers's proof of Szemerédi's theorem (to be discussed in the next subsection), and then strengthened further by Bourgain to produce Lemma 3.1 in the last section.

We recommend the book [46] for more information regarding Freiman's theorem and other inverse problems in additive number theory.

4.2. **Szemerédi's theorem.** We will say that a set $A \subset \mathbb{N}$ has upper density δ if

$$\overline{\lim}_{N\to\infty} \frac{|A \cup [1,N]|}{N} = \delta.$$

Motivated by van der Waerden's theorem in Ramsey theory, Erdős and Turán conjectured in 1936 that any set of integers A of positive upper density must contain arithmetic progressions of length k for any k. This was indeed proved by Roth [48] for $k = 3$, then by Szemerédi [56], [57] for all k. Below is an equivalent statement of this result:

Theorem 4.2. *For any $\delta > 0$ and any integer k there is a $N(\delta,k)$ such that if $N > N(\delta,k)$ and A is a subset of $\{1,2,\dots,N\}$ of cardinality $|A| \geq \delta N$, then A must contain a non-trivial k-term arithmetic progression.*

As of now, Szemerédi's theorem has four remarkably distinct proofs, each of which was a milestone in combinatorics in its own right. The original combinatorial proof by Szemerédi [57], ingenious and complicated even by Szemerédi's standards, featured the *regularity lemma*, which has since become an invaluable tool in Ramsey theory. Furstenberg's ergodic-theoretic proof [24], based on the *multiple recurrence theorem*, has the advantage of admitting a variety of extensions to more general problems of similar type, for example the multidimensional Szemerédi theorem due to Furstenberg and Katznelson [25], or the polynomial Szemerédi theorem of Bergelson and Leibman [5]. Gowers's proof [26], [27] is often referred to as "harmonic analytic", more for its resemblance to Roth's proof for $k = 3$ than for its actual use of harmonic analysis. It yields the best available quantitative bounds, in terns of the dependence of $N(\delta,k)$ on k and δ, for $k \geq 4$ (but this is now being challenged by Green and Tao for $k = 4$). Finally, there is a very recent hypergraph proof, due independently to Gowers and Nagle-Rödl-Schacht-Skokan (2004).

All known proofs of Szemerédi's theorem rely on a certain dichotomy between randomness and structure. Roughly speaking, if the elements of A were chosen from $\{1,\dots,N\}$ independently at random, each with probability δ, then with high probability there would be about $\delta^k N^2$ k-term arithmetic progressions in A, as there are about N^2 k-term arithmetic progressions in

$\{1, \ldots, N\}$, and each one is contained in A with probability δ^k. The same is true if A imitates a random set closely enough, in a sense that needs to be made precise. On the other hand, a non-random set should have a certain amount of additive structure, reminiscent of that in Freiman's theorem but *much* weaker. We then use that structure to our advantage, for example by passing to a long arithmetic subprogression of $\{1, \ldots, N\}$ on which A has higher density and then iterating the argument. The challenge is to find a notion of randomness which is strong enough to guarantee existence of k-term arithmetic progressions, but also weak enough so that its failure implies useful structural properties.

We illustrate this by taking a brief look at Roth's proof for $k = 3$. We will identify $\{1, \ldots, N\}$ with the additive group $\mathbb{Z}_N$. The discrete Fourier transform on $\mathbb{Z}_N$ is defined by

$$\widehat{f}(\xi) = N^{-1} \sum_{x=1}^{N} f(x) e^{-2\pi i x \xi}.$$

Let $A(x)$ be the characteristic function of A. A short Fourier-analytic calculation shows that if A contains no non-trivial 3-term arithmetic progressions, then there is a $\xi \neq 0$ such that

$$|\widehat{A}(\xi)| \geq \delta^2. \tag{4.2}$$

In other words, a set whose Fourier coefficients $\widehat{A}(\xi)$ are small enough behaves like a random set and contains 3-term arithmetic progressions. It remains to consider the case when (4.2) holds for some $\xi \neq 0$. In this case, we use (4.2) to prove that A cannot be uniformly distributed among long arithmetic progressions of step r for some r “dual" to ξ (i.e. $|\xi \cdot r|$ is small modulo N). This allows us to choose a long subprogression of $\{1, \ldots, N\}$ on which A has increased density, and then continue the inductive argument.

In Gowers's proof for arbitrary k, randomness (or *uniformity*) of A is determined by the size of the *Gowers norms* of its characteristic function. This is equivalent to the above for $k = 3$, but more complicated for higher k. Again, if A is uniform then it contains many k-term arithmetic progressions, but now uniformity is a stronger notion and, unlike for $k = 3$, its failure does not imply linear structure. Instead one must first find more complicated polynomial patterns in A, then exploit them, eventually arriving again at a density increment on a subprogression. It is in this part of the proof that advanced tools from additive number theory, such as the theorems of Freiman and Balog-Szemerédi, become crucial.

While this offers a short glimpse at the outline of Roth's and Gowers's arguments, we are not really able to do justice to any of this work here. More specialized surveys, such as [61] or [63], offer a better look at Szemerédi's theorem, its context in combinatorics and number theory, and the wide diversity of techniques and ideas involved in its proofs.

5. The Green-Tao theorem

5.1. **Once in a lifetime.** We finally turn to the k-term arithmetic progressions in the primes. It has long been conjectured that such progressions should exist for any k, for example this would follow from a much more general conjecture of Hardy and Littlewood in [35]. Van der Corput proved in 1939, by an application of the circle method, that primes contain infinitely many 3-term arithmetic progressions. The conjecture was settled by Green and Tao in [33]:

Theorem 5.1. *For any $k \geq 3$, primes contain arithmetic progressions of length k.*

An earlier result is due to van der Corput, who proved in 1939 that primes contain infinitely many 3-term arithmetic progressions. Ben Green extended this in [29] to dense subsets of primes. Both proofs rely on the *circle method*, a classic Fourier-analytic technique in number theory.

By contrast, the Green-Tao proof employed ideas from all then-existing proofs of Szemerédi's theorem (combinatorics, ergodic theory, Fourier analysis), combined with further number-theoretic information. Their approach was to embed the primes in a sufficiently random background set in which they have positive density, then prove a “relative Szemerédi theorem" which applies in this setting.

We begin with the latter part. Instead of sets $A \subset \{1, \dots, N\}$ of positive relative density, we consider *functions* f and ν on $\{1, \dots, N\}$ such that $0 \leq f \leq \nu$ and $\sum_x f(x) \geq \delta \sum_x \nu(x)$. Here f is the target function (later on it will be supported on the primes), and ν is the background function. We assume ν to be random in the sense that it satisfies certain explicit correlation conditions (not easy to reproduce here). A key point is that both f and ν need not be bounded uniformly in N. We wish to prove a Szemerédi theorem in this setting; more precisely, we need to estimate from below the quantity

$$\sum_{x,r} f(x)f(x+r)f(x+2r)\dots f(x+(k-1)r), \tag{5.1}$$

which counts the number of $k - term$ arithmetic progressions in a set A if f is the characteristic function of it. The proof of this proceeds roughly along the lines of Furstenberg's ergodic proof of Szemerédi's theorem. An inductive procedure is used to decompose f into random and quasiperiodic parts. The contribution of the random part to the quantity (5.1) is negligible. On the other hand, the “usual" Szemerédi theorem gives a bound from below on the contribution of the quasiperiodic part, and the result follows.

We now have to find appropriate functions f and ν. The reader should be used by now to occasional cheating in this exposition, and we will do it again here. Let $f = \Lambda$ be the von Mangoldt function, i.e. $\Lambda(n) = \log p$ if $n = p^k$ and 0 otherwise. This is not quite supported on the primes, but it is close enough and we can pretend that prime powers do not exist. We also define ν to be a “truncated" von Mangoldt function, supported on the almost primes (roughly, numbers which do not have small divisors). Now we bless our good luck.

An almost identical function had been considered earlier by Goldston and Yildirim in their work on small gaps between prime numbers. In fact, they had obtained correlation estimates on ν that are very close to those we need to establish the randomness of ν. There is still some work to do, but much of it has already been done for us. The work of Goldston and Yildirim was first circulated in 2003, then a gap was found in the proof a few months later. The main claims were withdrawn, but the preprints remained available and they certainly turned out to be useful! Later on, Goldston and Yildirim, joined by Pintz, fixed the proof and they now hold results on small gaps between primes that far exceed anything previously known.

There are now many expositions and reviews of various aspects of the Green-Tao work, see e.g. [31], [32], [44], [60], [61]. The focus of this note will remain on connections to harmonic analysis, and thus we return to restriction theory for the last time.

5.2. **What goes around, comes around.** Restriction estimates for finite exponential sums, as opposed to continuous Fourier transforms, were first derived by Bourgain [9] in the context of proving Strichartz estimates for solutions of evolution equations (such as Schrödinger and KdV) on the torus $\mathbb{T}^d$. They were then revisited in 2003 by Green in [29], a paper that directly inspired the work in [33].

We will try to explain the approach of [29] in the framework of the last subsection. Define f and ν as before (again we will not quite make this precise). Our goal is to prove lower bounds on (5.1) for $k = 3$. In this context, the randomness of ν simply means that ν has small Fourier coefficients, as explained earlier in connection with Roth's theorem. Green, however, does not proceed further along the same lines as [33]. Instead, his main tool is the restriction estimate

$$\|\widehat{f d\nu}\|_p \leq C_p \|f\|_{L^2(d\nu)},\ p > 2. \tag{5.2}$$

This has exactly the same form as (2.2), if we interpret ν as the density of a probabilistic measure supported on the almost primes. Moreover, the proof of (5.2) follows the Tomas-Stein argument very closely, from the interpolation between endpoints down to such details as the use of dyadic decompositions. Does this mean that the almost primes have curvature? Or that they have a Hausdorff dimension? Some questions are perhaps best dismissed without a hearing.

Although this type of Fourier analysis is not directly applicable to Szemerédi-type problems for progressions of length 4 and more, it was reportedly a major source of ideas for Green and Tao. They are currently working to develop a “quadratic Fourier analysis" that could be applied to finding 4-term progressions, or more generally solutions to systems of 2 linear equations, in suitable sets such as the primes or their dense subsets. This is a rapidly developing area and many more exciting developments are sure to follow.

6. Notes and acknowledgements

I am grateful to David Eisenbud and Susan Friedlander for inviting me to speak in the Current Events Bulletin and encouraging me to write this survey. This article is also based in part on the expository talks I gave at the Canadian Mathematical Society Winter 2004 meeting in Montreal, the MSRI workshop "Women in Mathematics: the legacy of Ladyzhenskaya and Oleinik" in May 2006 (there is some overlap between this article and my extended abstract in the workshop proceedings), and the American Mathematical Society 2006 Fall Western Section meeting in Salt Lake City.

I have relied on a variety of sources in preparing the manuscript. In addition to the many references cited in the text, I have also consulted the wonderful Internet-based MacTutor History of Mathematics Archive, maintained at the University of St. Andrews (`http://www-history.mcs.st-andrews.ac.uk/history`). This is where some of the historical information in Section 1, including the quote in Subsection 1.1, came from, though I also found Kenneth Falconer's historical comments in [18] to be informative and reliable.

The Big Dipper image on the booklet cover illustrates a layman's version of the multidimensional Szemerédi theorem: if the stars in the night sky shine brightly enough so that sufficiently many can be seen, then any desired pattern can be found among them. Mathematicians, for example Benjamin Weiss and Terence Tao, have sometimes used this metaphor in their lectures. In the film "A Beautiful Mind", there is a scene where the hero and his fiancée watch the night sky together. He asks her to pick a pattern. She chooses an umbrella. He looks up for a few seconds. Then their joined hands trace the shape of an umbrella between the stars.

References

1. A. Balog, E. Szemerédi, *A statistical theorem of set addition*, Combinatorica **14** (1994), 263–268.
2. A.S. Besicovitch, *Sur deux questions d'intégrabilité des fonctions*, J. Soc. Phys.-Math. (Perm), **2** (1919), 105-123.
3. A.S. Besicovitch, *On Kakeya's problem and a similar one*, Math. Zeitschrift **27** (1928), 312-320.
4. A.S. Besicovitch, *The Kakeya problem*, Amer. Math. Monthly **70** (1963), 697-706.
5. V. Bergelson, A. Leibman, *Polynomial extensions of van der Waerden's and Szemerédi's theorems*, J. Amer. Math. Soc.*9* (1996), 725–753.
6. Y. Bilu, *Structure of sets with small sumset*, in *Structure Theory of Set Addition*, Astérisque **258** (1999), 77–108.
7. J. Bourgain, *Besicovitch type maximal operators, and applications to Fourier analysis*, Geom. Funct. Anal. **1** (1991), 147–187.
8. J. Bourgain, *L^p estimates for oscillatory integrals in several variables*, Geom. Funct. Anal. **1** (1991), 321-374.
9. J. Bourgain, *Fourier restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations*, I, Geom. Funct. Anal. **3** (1993), 107–156.

10. J. Bourgain, *On the dimension of Kakeya sets and related maximal inequalities*, Geom. Funct. Anal. **8** (1998), 256–282.
11. H. Busemann, W. Feller, *Differentiation der L-integrale*, Fund. Math. **22** (1934), 226-256.
12. M.-C. Chang, *A polynomial bound in Freiman's theorem*, Duke Math. J. **3** (2002), 399–419.
13. M. Christ, *Estimates for the k-plane transform*, Indiana Univ. Math. J. **33** (1984), 891–910.
14. A. Córdoba, *The Kakeya maximal function and spherical summation multipliers*, Amer. J. Math. **99** (1977), 1–22.
15. E. Croot, V. Lev, *Open problems in additive combinatorics*, preprint, 2006.
16. R.O. Davies, *Some remarks on the Kakeya problem*, Proc. Cambridge Phil. Soc. **69** (1971), 417–421.
17. S. Drury, *L^p estimates for the x-ray transform*, Ill. J. Math. **27** (1983), 125–129.
18. K.J. Falconer, *The Geometry of Fractal Sets*, Cambridge Univ. Press, Cambridge, 1985.
19. C. Fefferman, *Inequalities for strongly singular operators*, Acta Math. **124** (1970), 9–36.
20. C. Fefferman, *The multiplier problem for the ball*, Ann. Math. **94** (1971), 330–336.
21. G.A. Freiman, *On the addition of finite sets* (in Russian), Dokl, Akad. Nauk SSSR **158** (1964), 1038–1041.
22. G.A. Freiman, *Foundations of a structural theory of set addition* (translated from Russian), Translations of Mathematical Monographs, vol. 37, Amer. Math. Soc., 1973.
23. M. Fujiwara, S. Kakeya, *On some problems of maxima and minima for the curve of constant breadth and the in-revolvable curve of the equilateral triangle*, Tôhoku Mathematical Journal, **11** (1917), 92–110.
24. H. Furstenberg, *Ergodic behaviour of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Analyse Math. **31** (1977), 204–256.
25. H. Furstenberg, Y, Katznelson, *An ergodic Szemerédi theorem for commuting transformations*, J. Analyse Math. **34** (1979), 275–291.
26. W.T. Gowers, *A new proof of Szemerédi's theorem for arithmetic progressions of length four*, Geom. Funct. Anal. **8** (1998), 529–551.
27. W.T. Gowers, *A new proof of Szemerédi's theorem*, Geom. Funct. Anal. **11** (2001), 465–588.
28. W.T. Gowers, *Some unsolved problems in additive and combinatorial number theory*, preprint, 2001 (available at `http://www.dpmms.cam.ac.uk/ wtg10/papers.html`).
29. B. Green, *Roth's Theorem in the primes*, Ann. Math. **161** (2005), 1609-1636.
30. B. Green, *Structure theory of set addition*, unpublished, available at `http://www.dpmms.cam.ac.uk/ bjg23/papers/icmsnotes.pdf`.
31. B. Green, *Generalising the Hardy-Littlewood method for primes*, Proc. Intern. Congress. Math., Madrid 2006.
32. B. Green, *Long arithmetic progressions of primes*, submitted to Proceedings of the Gauss-Dirichlet conference, Gottingen 2005.
33. B. Green, T. Tao, *The primes contain arbitrarily long arithmetic progressions*, Ann. Math., to appear.
34. B. Green, T. Tao, *Restriction theory of the Selberg sieve, with applications*, Jour. Th. Nombres Bordeaux **18** (2006), 147–182.
35. G. Hardy, J. Littlewood, *Some problems of "partitio numerorum", III: On the expression of a number as a sum of primes*, Acta Math. **44** (1923), 1–70.
36. C.S. Herz, *Fourier transforms related to convex sets*, Ann. Math. **75** (1962), 81–92.

37. L. Hörmander, *The Analysis of Linear Partial Differential Operators*, volume 1, 2nd edition, Springer Verlag, 1990.
38. A. Iosevich, *Curvature, combinatorics and the Fourier transform*, Notices Amer. Math. Soc. **46** (2001), 577–583.
39. S. Kakeya, *Some problems on minima and maxima regarding ovals*, Tôhoku Science Reports, **6** (1917), 71–88.
40. N.H. Katz, I. Łaba, T. Tao, *An improved bound on the Minkowski dimension of Besicovitch sets in* $\mathbb{R}^3$, Ann. of Math. **152** (2000), 383-446.
41. N.H. Katz, T. Tao, *Bounds on arithmetic projections, and applications to the Kakeya conjecture*, Math. Res. Letters 6 (1999), 625-630.
42. N.H. Katz, T. Tao, *New bounds for Kakeya sets*, J. Anal. Math. **87** (2002), 231–263.
43. N.H. Katz, T. Tao, *Recent progress on the Kakeya conjecture*, Publicacions Matematiques, Proceedings of the 6th El Escorial International Conference on Harmonic Analysis and Partial Differential Equations, U. Barcelona 2002, 161-180.
44. B. Kra, *The Green-Tao theorem on arithmetic progressions in the primes: an ergodic point of view*, Bull. Amer. Math. Soc. **43** (2006), 3–23.
45. I. Łaba, T. Tao, *An improved bound for the Minkowski dimension of Besicovitch sets in medium dimension*, Geom. Funct. Anal. **11** (2001), 773-806.
46. M.B. Nathanson, *Additive Number Theory: Inverse Problems and the Geometry of Sumsets*, Graduate Texts in Mathematics 165, Springer-Verlag, 1996.
47. J. Pach, M. Sharir, *Geometric incidences*, in: Towards a Theory of Geometric Graphs (J. Pach, ed.), Contemporary Mathematics, vol. 342, Amer. Math. Soc. 2004.
48. K. Roth, *On certain sets of integers*, J. London Math. Soc. **28** (1953), 245–252.
49. I.Z. Ruzsa, *Arithmetic progressions and the number of sums*, Period. Math. Hung. **25** (1992), 105–111.
50. I.Z. Ruzsa, *An application of graph theory to additive number theory*, Scientia, Ser. A **3** (1989), 97–109.
51. I.Z. Ruzsa, *Generalized arithmetic progressions and sumsets*, Acta Math. Hungar. **65** (1994), 379–388.
52. C. Sogge, *Fourier integrals in Classical Analysis*, Cambridge University Press, Cambridge, 1993.
53. E.M. Stein, *Oscillatory integrals in Fourier analysis*, in *Beijing Lectures in Harmonic Analysis* (E.M. Stein, ed.), Ann. Math. Study # 112, Princeton Univ. Press, 1986, pp. 307-355.
54. E.M. Stein, *Harmonic Analysis*, Princeton Univ. Press, Princeton, 1993.
55. R. Strichartz, *Restriction of Fourier transform to quadratic surfaces and decay of solutions of wave equations*, Duke Math. J. **44** (1977), 705–774.
56. E. Szemerédi, *On sets of integers containing no four elements in arithmetic progression*, Acta Math. Acad. Sci. Hungar. **20** (1969), 89–104.
57. E. Szemerédi, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. **27** (1975), 299–345.
58. T. Tao, *A sharp bilinear restriction estimate for paraboloids*, Geom. Funct. Anal. **13** (2003), 1359-1384.
59. T. Tao, *Recent progress on the restriction conjecture*, to appear in Park City conference proceedings.
60. T. Tao, *Arithmetic progressions and the primes - El Escorial lectures*, Collectanea Mathematica (2006), Vol. Extra., 37-88 (Proceedings of the 7th International Conference on Harmonic Analysis and Partial Differential Equations, El Escorial).
61. T. Tao, *The dichotomy between structure and randomness, arithmetic progressions, and the primes*, Proc. Intern. Congress. Math., Madrid, 2006.
62. T. Tao, *Nonlinear dispersive equations: local and global analysis*, CBMS Regional Conference Series in Mathematics, Amer. Math. Soc., 2006.

63. T. Tao, V. Vu, *Additive Combinatorics*, Cambridge University Press, 2006
64. P.A. Tomas, *A restriction theorem for the Fourier transform*, Bull. Amer. Math. Soc. **81** (1975), 477–478.
65. P.A. Tomas, *Restriction theorems for the Fourier transform*, in *Harmonic Analysis in Euclidean Spaces*, G. Weiss and S. Wainger, eds., Proc. Symp. Pure Math. # 35, Amer. Math. Soc., 1979, vol, I, pp. 111-114.
66. T. Wolff, *Decay of circular means of Fourier transforms of measures*, Internat. Math. Res. Notices **10** (1999), 547-567.
67. T. Wolff, *An improved bound for Kakeya type maximal functions*, Rev. Mat. Iberoamericana **11** (1995), 651–674.
68. T. Wolff, *Recent progress connected with the Kakeya problem*, in *Prospects in Mathematics*, H. Rossi, ed., Amer. Math. Soc., Providence, R.I. (1999), 129–162.
69. T. Wolff, *A sharp bilinear cone restriction estimate*, Ann. Math. **153** (2001), 661–698.
70. T. Wolff, *Local smoothing estimates in L^p for large p*, Geom. Funct. Anal. **10** (2000), 1237–1288.
71. T. Wolff, *Lectures on Harmonic Analysis*, I. Łaba and C. Shubin, eds., Amer. Math. Soc., Providence, R.I. (2003).

Department of Mathematics, University of British Columbia, Vancouver, B.C. V6T 1Z2, Canada
E-mail address: `ilaba@math.ubc.ca`

THE STRUCTURE OF ERROR TERMS IN NUMBER THEORY AND AN INTRODUCTION TO THE SATO-TATE CONJECTURE.

BARRY MAZUR

CONTENTS

ABSTRACT. It is wonderful to see the individual strengths of otherwise separate mathematical sub-disciplines coming together and connecting with each other (in as startling a way as the theory of continental drift connects the shape of disparate continents) and then providing for us the resolution of a long-sought conjecture. This is indeed what happened last Spring, when a conjecture about certain important probability distributions in number theory, posed forty years ago by Mikio Sato and John Tate, was finally verified for a large number of cases as the culmination of three major works:

- *in the study of modular liftings and automorphic representation theory* (work of Laurent Clozel, Michael Harris, and Richard Taylor **[1]**)
- *in algebraic geometry and automorphic representations* (work of Michael Harris, Nicholas Shepherd-Barron, and Richard Taylor **[3]**)
- *in Galois deformation theory* (work of Richard Taylor **[12]**).

the last-mentioned breakthrough establishing the result.

My aim is just to discuss, in concrete terms, two "sample problems" — one still open, and one settled by the recent work—that are addressed by the Sato-Tate Conjecture.

1. Error Terms and the Sato-Tate Conjecture

1.1. **Why are there still unsolved problems in Number Theory?** Eratosthenes, to take an example—or other ancient Greek mathematicians—might have imagined that all they needed were a few powerful insights and then everything about numbers would be as plain, say, as facts about triangles in the setting of Euclid's *Elements of Geometry*. If Eratosthenes had felt this, and if he now—transported by some time machine—dropped in to visit us, I'm sure he would be quite surprised to see what has developed.

Of course, geometry has evolved splendidly but has expanded to higher realms and more profound structures. Nevertheless, there is hardly a question that Euclid could pose with his vocabulary about triangles that we don't know the answer to today. And, in stark contrast, many of the basic naive queries that Euclid or his contemporaries might have had about primes, perfect numbers, and the like, would still be open.

Sometimes, but not that often, in number theory, we get a complete answer to a question we have posed, an answer that finishes the problem off. Often something else happens: we—perhaps after some major effort—-manage to find a fine, simple, *good approximation* to the data, or phenomena, that interests us, and then we discover that yet deeper questions lie hidden in the error term—in the measure of how badly our approximation misses its mark.

A telling example of this, and of how in the error term lies richness, is the manner in which we study of $\pi(X)$:= the number of prime numbers less than X. The function $\pi(X)$ is shown below, in various ranges as step functions giving the “staircase" of numbers of primes.

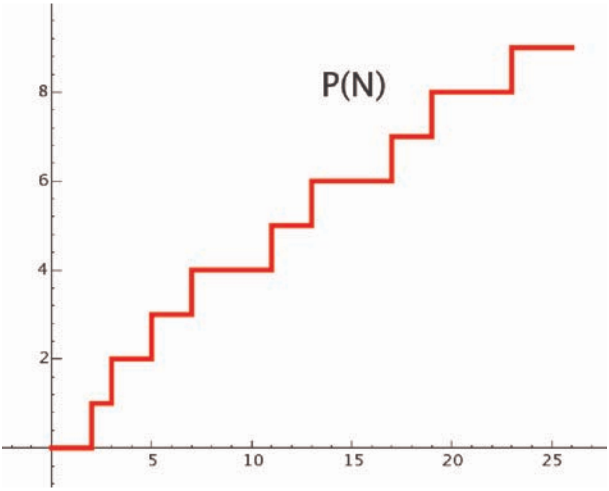


Figure 1.1. The step function $\pi(N)$ counts the number of primes up to N

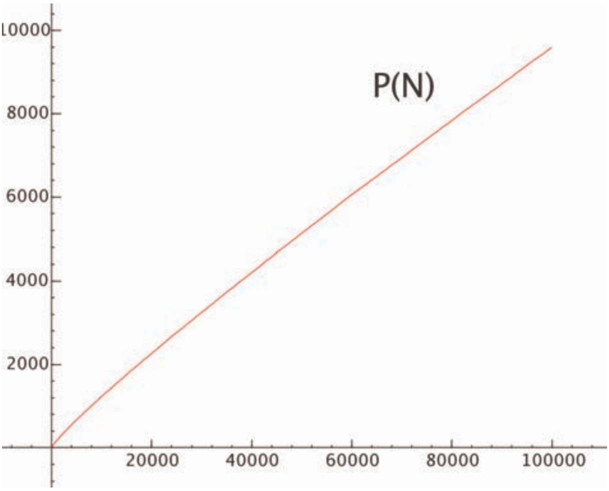


Figure 1.2. The step function $\pi(N)$ counts the number of primes up to N

As is well known, Carl Friedrich Gauss, two centuries ago, computed tables of $\pi(X)$ by hand, for X up to the millions, and offered us a probabilistic "first" guess for a nice smooth approximating curve for this data; a certain beautiful curve that, experimentally, seems to be an exceptionally good fit for the staircase of primes.

The data, as we clearly see, certainly cries out to us to guess a *good approximation.* If you make believe that the chances that a number N is a prime is inversely proportional to the number of digits of N you might well hit upon Gauss's guess, which produces indeed a very good fit. In a letter written in 1849 Gauss claimed that as early as 1792 or 1793 he had already observed that the density of prime numbers over intervals of numbers of a given rough magnitude X seemed to average $1/\log X$.

The Riemann Hypothesis is equivalent to saying that the integral $\int_2^X dx/\log x$ (i.e., the area under the graph of the function $1/\log x$ from 2 to X) is *essentially square root close* to $\pi(X)$. *Essentially square root close* by the way just means that for any given exponent greater than 1/2 (you choose it: 0.501, 0.5001, 0.50001 for example) and for large enough X—the size, here, depending on your choice of exponent—the difference between $\int_2^X dx/\log x$ and $\pi(X)$ in absolute value is less than X raised to that exponent (e.g. $X^{0.501}$ etc.).

1.2. **Much of the depth of the problem is hidden in the structure of the error term.** In a general context, once we make what we hope to be a good approximation to some numerical data, we can focus our attention to the *error term* that has thereby been created, namely:

$$\text{Error term} = \text{Exact Value} - \text{Our "good approximation."}$$

In our attempt to understand $\pi(x)$, i.e., the placement of primes in the sequence of natural numbers, we might choose, with Gauss, our *good approximation* to be $\int_2^X dx/\log x$. If so, then we will have focused our mind on the *error term* which so that in this instance we have

$$\mathrm{Error}(x) \;=\; \pi(x) \;-\; \int_2^X dx/\log x.$$

It is Riemann's analysis of—what is in effect— this error term that first showed us the immense world of structure packaged in it. For Riemann did what is, in effect, a Fourier analysis of $\pi(e^t)$ expressing $\mathrm{Error}(x)$ (or, more precisely, a closely related function that has the same information as $\mathrm{Error}(x)$) as an *exact* infinite sum of corrective terms, each of these corrective terms easily described in terms of the value of a *zero of the Remann zeta function*; all of these corrective terms are square root small if and only if "his" hypothesis holds[1].

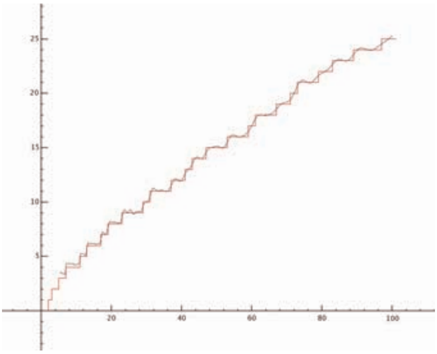


Figure 1.3. The smooth function slithering up the staircase of primes up to 100 **is Riemann's approximation that uses the "first"** 29 **zeroes of the Riemann zeta function**

[1] *All the data in figures appearing in this article have been tabulated by William Stein.*

1.3. **Strict square-root accuracy.** We will be considering a somewhat different class of number theoretic problem than the example that we have been discussing, and for those an even stronger notion of *square-root approximation* is relevant. We will be interested in situations where the *error term* is less than a *fixed constant* times the square root of the quantity being approximated; let us say that an approximation to numerical data has **strict square-root accuracy** if its error term has this property.

We have witnessed great successes in the last century in obtaining good approximations to important problems in Number theory, with error terms demonstrated to be strictly square-root accurate. Specifically, through the work of Helmut Hasse in the 1930s, André Weil in the 1940s and Pierre Deligne in the 1970s, a large class of major approximations were proved to have this kind of accuracy.

1.4. **Some Sample Arithmetic Problems.** It has known since the time of Fermat, and proved by Euler, that a prime p can be written as a sum of two square numbers if and only if $p \not\equiv 3$ modulo 4 and if it can be written as a sum of two squares, it can be done so in only one way (not counting the order of the two squares). For example:

$$401 = 1^2 + 20^2$$

is the only way (up to changing the order of the two summands) to express the prime number 401 as a sum of two squares. This result is, for many reasons, a much more central and important classical result than it may first appear to be. The problem, which seems to mix *prime numbers* with *geometry* (squares of distances to the origin of integral lattice points in the plane) has the virtue that its answer is equivalent to knowledge of the splitting properties of primes, and the validity of the unique factorization theorem, in the ring of gaussian integers.

In how many ways can the prime p be expressed as *a sum of the squares of three integers?* The answer for $p \geq 5$ —due to Gauss—can be given in terms of the function $h(-d)$ the class number of the quadratic imaginary field of discriminant $-d$. The number of ways that $p \geq 5$ be expressed as *a sum of the squares of three integers* is:

- $12h(-4p)$ if $p \equiv 1, 5$ modulo 8;
- $24h(-p)$ if $p \equiv 3$ modulo 8;
- 0 if $p \equiv 7$ modulo 8.

The rules of the game here is that the ordering of the summands, and the signs of the integers chosen, count in the tally so for $p = 2$ we have $2 = 0^2 + (\pm 1)^2 + (\pm 1)^2 = (\pm 1)^2 + 0^2 + (\pm 1)^2 = (\pm 1)^2 + (\pm 1)^2 + 0^2$ and therefore we have that 2 can be written “as a sum of three squares" in $3 \cdot 2^2 = 12$ ways.

These two problems are simply the first two of a series of companion questions that have a long history,

In many ways can the prime p be expressed as a sum of the squares of r integers?

To get some sample problems that drive home a point I want to make in this exposition—and for no other reason—of I'll restrict consideration to certain select values of r.

For $r = 4$ we have a simply stated, exact, solution: the prime p can be expressed as a sum of four squares in $8p + 8$ ways.

For $r = 8$, any odd prime number p can be expressed as a sum of eight squares in $16p^3 + 16$ ways.

In both of these cases the answer to our problem (at least for $p > 2$) is a polynomial in p of degree $r/2 - 1$ (i.e., of degree 1 and 3, respectively). Things, however, don't remain as simple, for larger values of r—probably for most larger values of r. To illustrate how things can change, let us focus on $r = 24$.

1.5. **Our first "sample problem.".** Define, then, $N(p)$ to be the number of ways in which p can be written as a sum of 24 squares of whole numbers.

Recall that squares of positive numbers, negative numbers and zero are all allowed, and the ordering of the squares of the numbers that occur in this summation also counts. Thus, the first prime number, 2, can already be written as a sum of 24 squares of whole numbers in $1,104$ ways. So:

$$N(2) = 1,104.$$

What about $N(p)$ for the other prime numbers $p = 3, 5, 7, 11, \ldots$? Here is some data.

2	1104
3	16192
5	1362336
7	44981376
11	6631997376
13	41469483552
17	793229226336
19	2697825744960
23	22063059606912
29	282507110257440
31	588326886375936
37	4119646755044256
41	12742799887509216
43	21517654506205632
47	57242599902057216
53	214623041906680992
59	698254765677746880
61	1007558483942335776
67	2827903926520931136
71	5351602023957373056
73	7264293802635839712
79	17319684851070915840
83	29819539398107307072
89	64258709626203556320
97	165626956557080594016

Eyeballing the data, it is already convincingly clear that $N(p)$ is growing less than exponentially, for otherwise the shadow of figures on the page would probably look triangular. Following the pattern we've seen for the smaller values of r we have considered we might expect that $N(p)$ be a polynomial in p of degree $r/2 - 1 = 11$. If we had enough data I imagine we might "curve-fit" a polynomial approximation. But happily, without having to lean on numerical experimentation, certain theoretical issues—which I don't want to get into—allow us to guess the following *good approximation* for the values $N(p)$; namely the polynomial in p of degree 11:

$$N_{\text{approx}}(p) := \frac{16}{691}(p^{11} + 1).$$

The difference, then, between the data and our good approximation is:

$$\text{Error}(p) := \ N(p) \ - \ N_{\text{approx}}(p) \ = \ N(p) \ - \ \frac{16}{691}(p^{11} + 1).$$

This error term been proven to be square-root small. And perhaps one should emphasize that this square-root smallness statement is hardly an elementary result: it is a consequence of deep work of Deligne. In fact, using the work of Deligne I am alluding to, you can show that:

$$|\mathrm{Error}(p)| \le \frac{66,304}{691}\sqrt{p^{11}}.$$

What with that hefty constant, $\frac{66,304}{691}$, the "smallness" of our error term here may not impress us for quite a while as we systematically tabulate the values of $N(p)$, but—of course— this result tells us that as we get into the high prime numbers our data will hug startlingly close to the simple smooth curve

$$f(x) = \frac{16}{691}(x^{11} + 1).$$

1.6. **The "next question".** Whenever some element of some theory is settled, or is considered settled, many of us mathematicians propose a subsequent plan of inquiry with that phrase: "So, the next question to ask is . . . "

Here too. Given the precise inequality

$$|\mathrm{Error}(p)| \le \frac{66,304}{691}\sqrt{p^{11}}$$

described in the previous section, and given the fact that this represents one consequence of what has been a great project that has spanned half a century of progress in number theory, some natural (and related) "next" questions arise. We might—for example—ask

- Is the bound on this error term (e.g., the constant $\frac{66,304}{691}$) is the best possible?
- Is $f(x) = \frac{16}{691}(x^{11} + 1)$ the *best* polynomial approximation to our data?
- Might we, more specifically, find another polynomial $g(x)$ which *beats* $f(x)$ in the sense that the absolute values of the corresponding error terms $|N(p) - g(p)|$ are $\le C\sqrt{p^{11}}$ with a constant C that is strictly less than $\frac{66,304}{691}$?
- For any given constant $C < \frac{66,304}{691}$ is there a positive proportion of prime numbers p for which

$$|N(p) - f(p)| \le C\sqrt{p^{11}}.$$

- We might ask what that proportion is, as a function of C.
- We might ask for the proportion of primes p for which the error term is positive, i.e., where our good approximation is an undercount.

To be sure, we would want to phrase such questions not only about our specific "sample problem" but about the full range of problems for which we have—thanks to Deligne et al— such good square-root close approximations.

It is the *Sato-Tate Conjecture* that addresses this "next," more delicate, tier of questions[2].

[2] *As is only to be expected, there are whole books of questions about this sample problem that one could ask, and mathematicians have asked—some of these questions being structurally important, and some at least traditionally of great interest. Eg., how often*

1.7. **The distribution of scaled error terms.** Given that in our sample problem we know the bound

$$|\mathrm{Error}(p)| \leq \frac{66,304}{691}\sqrt{p^{11}},$$

let us focus our microscope on the fluctuations here. Namely, consider the *scaled* error term

$$\mathrm{Scaled\,Error}(p) := \frac{\mathrm{Error}(p)}{\frac{66,304}{691}\sqrt{p^{11}}} = \frac{N(p) \ - \ \frac{16}{691}(p^{11}+1)}{\frac{66,304}{691}\sqrt{p^{11}}}$$

so that we have:

$$-1 \ \leq \ \mathrm{Scaled\,Error}(p) \ \leq \ +1.$$

About this type of *scaled error value distribution,* let me recall the words of Susan Holmes, a mathematician and statistician at Stanford, who—when I sent her some numerical computations related to a similar number theoretic problem for which I had some statistical questions—exclaimed: “what beautiful data!"

But what can we say further about this data? How do these scaled error values distribute themselves on the interval $[-1,+1]$? That is, what is the function $I \mapsto \mathcal{P}(I)$ that associates to any subinterval I contained in $[-1,+1]$ the *probability* $\mathcal{P}(I)$ that for a randomly chosen prime number p its scaled error term $\mathrm{Error}(p)$ lies in I?

In 1960, Mikio Sato (by studying numerical data) and John Tate (following a theoretical investigation) predicted—for a large class of number theoretic questions including many problems of current interest, of which our example is one—that the values of the scaled error terms for data in these problems conforms to a specific probability distribution, Usually the Sato-Tate conjecture predicts that this distribution is no more complicated than the elementary function $t \mapsto \frac{2}{\pi}\sqrt{1-t^2}$, i.e., the thing whose graph is a semi-circle of radius 1 centered at the origin, but squished vertically to have its integral equal to one. This makes it far from the Gaussian normal distribution! Indeed, Sato and Tate predict this type of behavior in our example problem, so that their conjecture would have it that

$$\mathcal{P}(I) = \frac{2}{\pi}\int_I \sqrt{1-t^2}dt.$$

This is still an open question, for our sample problem! Nevertheless, we have an impressive amount of data in support of it (see below).

To compute this data for primes up to 10^6 in reasonable time required much ingenuity on the part of William Stein. For a more complete description

is our approximate value $N_{\mathrm{approx}}(p)$ above exactly equal to the actual value $N(p)$? A conjecture of Lehmer would say that this never happens.

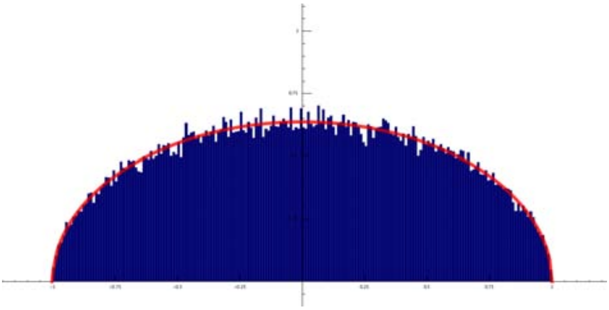


Figure 1.4. Probability distribution of error terms. The Sato-Tate distribution $\frac{2}{\pi}\sqrt{1-t^2}$, the smooth profile curve in this figure, can be compared with the probability distribution of *scaled error terms* **for the number of ways $N(p)$ in which a prime number p can be written as a sum of 24 squares ($p < 10^6$). These computations were made by William Stein.**

of what these computations entail, together with other background material, consult William Stein's: http://sage.math.washington.edu:8100/193

The great breakthrough last Spring was the resolution of the Sato-Tate conjecture for a large class of elliptic curves.

2. AN ELLIPTIC CURVE, AND A NEW "SAMPLE PROBLEM"

2.1. The number of points of an elliptic curve mod p; for varying p.

The example we will use is one of the favorites of many number theorists, namely the curve in the plane, call it E, cut out by the equation

$$y^2 + y = x^3 - x^2.$$

This is an elliptic curve that is something of a showcase for number theory, in that it has been extensively studied—much is known about it—and yet it continues to repay study, for—as with all other elliptic curves—its deeper features have yet to be understood.

This curve $E : y^2 + y = x^3 - x^2$ when extended to the projective plane has exactly one rational point on the line at infinity, and if you stipulate that that unique point "at infinity" be the *origin,* there is a unique algebraic group law on E, allowing us—for any field k of characteristic different from 11 (i.e., any field where $11 \neq 0$)—to endow the set consisting of ∞ and the points of E with values $(x, y) = (a, b) \in k$ with the structure of an abelian group. Let k be of characteristic different from 11 and let us denote by $E(k)$ this group of k-rational points of E. The reason why we have to exclude 11 is that the polynomial equation above modulo 11 has a singular point.

Every one of these groups $E(k)$ contains the five rational points

$$\{\infty,\ (0,0),\ (0,-1),\ (1,0),\ (1,-1),\}$$

and it isn’t difficulty to check that these five points comprise a cyclic subgroup of $E(k)$ of order five. The *data* we shall be focussing on, in this problem is *the number of rational points that E has over the prime field containing p elements* (excluding, again, $p = 11$). So, let p be a prime number (different from 11) and let $\mathbf{F}_p$ denote the field of integers modulo p, and define

$$N(p) := \text{the number of elements in the finite group } E(\mathbf{F}_p).$$

There is much that is surprising in this “data.’ That is the numerical function

$$p \longmapsto N(p)$$

or (essentially equivalently) the error term we are now concentrating on:

$$p \longmapsto Error(p) = N(p) - (p + 1)$$

and it can be expressed in a few quite different-sounding ways. Here is one: Expand the infinite product

$$q \prod_{n=1}^{\infty} (1 - q^n)^2 (1 - q^{11n})^2 = \sum a_n q^n$$

and we have that:

$$Error(p) = -a_p.$$

Here is what $N(p)$ looks like for small primes p:

p	2	3	5	7	13	17	19	23	29	31	37	41	43	47	53	59	61	67	71
$N(p)$	5	5	5	10	10	20	20	25	30	25	35	50	50	40	60	55	50	75	75

Since, from the first of the two definitions, $N(p)$ is the order of a finite group that contains a cyclic group of order five, we know, from Lagrange’s theorem of elementary group theory that $N(p)$ is divisible by 5, but what more can we say about the data

$$p \longmapsto N(p)?$$

This, now, will constitute our *sample problem* on which be focussing for the rest of this article.

For starters, following the format of the the previous sections of this article, we should look for a “good approximation" to $N(p)$. An old result due to Helmut Hasse tells us that a square-root accurate approximation to $N(p)$ is given by the simple expression: $p + 1$, which is, by the way, just the number of points on a line in the projective plane over $\mathbf{F}_p$.

It s a deep theorem (proved in the PhD thesis of Noam Elkies) that for this elliptic curve as well as any other elliptic curved defined over Q there are an infinite number of primes p such $N(p)$ is equal to *precisely* this simple expression $p+1$. But, it is generally true that the error term for this approximation is quite small. Explicitly, writing

$$\mathrm{Error}(p) := N(p) - (p+1)$$

Hasse proved the inequality

$$|\mathrm{Error}(p)| = |N(p) - (p+1)| \le 2\sqrt{p}.$$

Another way of saying this is that there is a conjugate pair of complex numbers $e^{i\theta_p}$ and $e^{-i\theta_p}$ for which the error term can be written as

$$\mathrm{Error}(p) := N(p) - (p+1) = \sqrt{p}(e^{i\theta_p} + e^{-i\theta_p}) = 2\sqrt{p}\cos(\theta_p).$$

Following, again, the format of our example-problem of the previous sections, we might ask for the distribution of error values, and here we can do this just by asking for the statistics of the rule that assigns to prime numbers p the conjugate-pair of complex numbers on the unit circle in the complex plane

$$p \longmapsto e^{\pm i\theta_p}.$$

Here is some data:

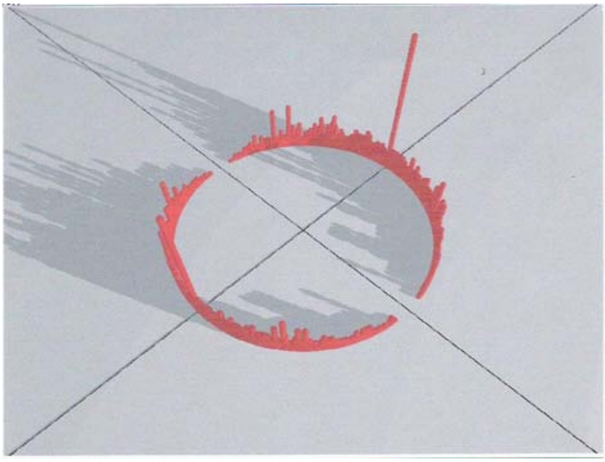


Figure 2.1. The accumulation of red dots on a position in this unit circle of the complex plane corresponds to the frequency of occurrences of θ_p in a small arc around that position for a good number of primes p. The two axis lines are the x-axis and y-axis; from the data—which conforms to the Sato-Tate statistics—you can guess which is which.

2.2. **The Sato-Tate distribution.** Thanks to the recent advance due to Taylor et al, the data

$$p \longmapsto \cos(\theta_p) = 1/2(e^{i\theta_p} + e^{-i\theta_p})$$

of the previous section conforms to the Sato-Tate distribution $\frac{2}{\pi}\sqrt{1-t^2}$. That is,

Theorem 2.1. (The Sato-Tate Conjecture for our sample case) *For any continuous function $F(t)$ on the interval $[-1, +1]$ we have that the limit*

$$\lim_{X\to\infty} \sum_{p\leq X} F(\cos\theta_p)/\pi(X)$$

exists and is equal to the integral

$$\frac{2}{\pi}\int_{-1}^{+1} F(t)\sqrt{1-t^2}dt.$$

To express our expected distribution in terms of the θ_p's, one could make the change of variables ($t \mapsto \cos\theta$)

$$\frac{2}{\pi}\int_{-1}^{+1} F(t)\sqrt{1-t^2}dt = \frac{1}{\pi}\int_{-\pi}^{+\pi} F(\cos\theta)\sin^2\theta d\theta = \frac{2}{\pi}\int_{0}^{+\pi} F(\cos\theta)\sin^2\theta d\theta,$$

i.e., expressing things in terms of θ we get a “sine-squared" distribution. Here is what the data looks like in these terms:

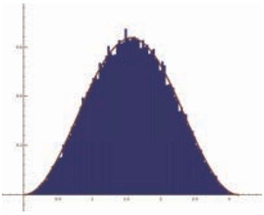


Figure 2.2.

In a sequel to these notes (a sequel yet to be written) I would like to say a few things for nonexperts about the actual proof of this theorem. But to conclude here let us see how the problem reduces to a study of L-functions.

To prove the above theorem, it would be enough to show that

$$\lim_{X\to\infty} \sum_{p\leq X} F(\cos\theta_p)/\pi(X) = \frac{2}{\pi}\int_{-1}^{+1} F(t)\sqrt{1-t^2}dt.$$

for all real-valued polynomial functions $F(t)$ by the Weierstrass approximation theorem, and therefore, since our task is linear, we could concentrate on proving this for $F(t) =$ all the powers of the variable t, i.e.,

$$1, t, t^2, t^3, \ldots$$

or, for that matter it would suffice to prove it for $F(t)$ = any other $\mathbf{R}$-basis of the ring of real-valued polynomials.

2.3. **Bases for the ring of polynomials.** Write the variable x as a sum $\alpha + \alpha^{-1}$ so that any polynomial in x (with, e.g., real coefficients) is a polynomial in α and α^{-1} invariant under the interchange $\alpha \leftrightarrow \alpha^{-1}$, and conversely: any polynomial in α and α^{-1} invariant under the above interchange is a polynomial in x. Consider then, these polynomials (let's call them *symmetric power polynomials*)

$$\begin{array}{rcl} s_0 & = & 1 \\ s_1 & = & \alpha + \alpha^{-1} \\ s_2 & = & \alpha^2 + 1 + \alpha^{-2} \\ s_3 & = & \alpha^3 + \alpha^1 + \alpha^{-1} + \alpha^{-3} \\ s_4 & = & \alpha^4 + \alpha^2 + 1 + \alpha^{-2} + \alpha^{-4} \\ s_5 & = & \alpha^5 + \alpha^3 + \alpha^1 + \alpha^{-1} + \alpha^{-3} + \alpha^{-5} \\ & \dots & \end{array} \tag{2.1}$$

which, when expressed as polynomials in x, look like

$$\begin{array}{rcl} s_0(x) & = & 1 \\ s_1(x) & = & x \\ s_2(x) & = & x^2 - 1 \\ s_3(x) & = & x^3 - 2x \\ s_4(x) & = & x^4 - 3x^2 + 1 \\ s_5(x) & = & x^5 - 4x^3 + 3x \\ & \dots & \end{array} \tag{2.2}$$

where $s_n(x)$ is a monic polynomial in x of degree n (these are also called the *Chebychev polynomials of the second kind*). They form a basis of the vector space of polynomials in the variable x. Any collection of products

$$\{s_m(2t)s_n(2t)\}_{(m,n)\in\mathcal{I}}$$

forms a basis of the vector space of polynomials in the variable t where $\mathcal{I}$ is a collection of a pairs of nonnegative integers such that the sums $m + n$ run through all nonegative numbers with no repeats.

Here is an elementary calculus exercise:

Proposition 2.2. *If $F(t) = s_m(2t)s_n(2t)$ with $m \neq n$ then*

$$\frac{2}{\pi}\int_{-1}^{+1} F(t)\sqrt{1-t^2}dt = 0.$$

Corollary 2.3. *Theorem 2.1 would follow if for every positive integer k there is a pair of distinct nonnegative integers (n, m) with $n + m = k$ and such that*

$$\lim_{X\to\infty} \sum_{p\le X} s_m(2\cos\theta_p)s_n(2\cos\theta_p)/\pi(X) = 0.$$

But how can we get that such limits vanish? A standard strategy—in fact, it seems, the only known strategy—is to invoke L functions. So we turn to:

2.4. L-**functions.** To study the data $\{p \longmapsto \pm\theta_p\}$ effectively it is a good idea to "package it" into complex analytic functions (Dirichlet series) whose behavior will tell us about the limits described in Corollary 2.3.

Let us do this. For any choice of prime number p different from 11 and for any pair of nonnegative numbers $0 \le m \le n$, define *the local factor at p of the L-function* $L_{m,n}(s)$ as follows:

$$L_{m,n}^{\{p\}}(s) := \prod_{j=0}^{m}\prod_{k=0}^{n}\left(1 - e^{i(m+n-2j-2k)\theta_p}p^{-s}\right)^{-1}.$$

If m (or n) is zero, the factors in "$\prod_{j=0}^{m}$" (or "$\prod_{k=0}^{n}$") don't occur, so, for example:

$$L_{0,n}^{\{p\}}(s) := \prod_{k=0}^{n}\left(1 - e^{i(n-2k)\theta_p}p^{-s}\right)^{-1}.$$

Now form the infinite product over all prime numbers p different from 11:

$$L_{m,n}(s) := \prod_{p} L_{m,n}^{\{p\}}(s)$$

and expand this to get a Dirichlet series

$$L_{m,n}(s) = \sum_{r=0}^{\infty} a_{m,n}(r)r^{-s}.$$

The terms $a_{m,n}(r)$ are easily computed: we have, for example, that $a_{m,n}(p) = s_m(2\cos\theta_p)s_n(2\cos\theta_p)$ for p a prime number different from 11, and for any positive integer r the term $a_{m,n}(r)$ is bounded from above in absolute value by a fixed polynomial (depending only on m and n) in $\log(r)$. This guarantees that the Dirichlet series $L_{m,n}(s)$ converges in the half-plane $Re(s) > 1$.

Here we rely on analytic number theory (in the form of a classical theorem of Wiener and Ikehara) which gives us that if we know enough further analytic facts about these Dirichlet series $\sum a_{m,n}(r)r^{-s}$ we can control limits of the form

$$\lim_{X\to\infty} \frac{\sum_{p<X} a_{m,n}(p)}{\pi(X)},$$

i.e., since $a_{m,n}(p) = s_m(2\cos\theta_p)s_n(2\cos\theta_p)$ $(p \ne 11)$ these are exactly the limits we are interested in.

Proposition 2.4. *Let* $m < n$. *If* $L_{m,n}(s)$ *extends to a meromorphic function on the entire complex plane, holomorphic on* $\mathrm{Re}(s) \geq 1$ *and nonzero on all points* $\mathrm{Re}(s) = 1$ *other than* $s = 1$ *then*

$$\lim_{X\to\infty} \sum_{p\leq X} s_m(2\cos\theta_p)s_n(2\cos\theta_p)/\pi(X) = 0.$$

If, by the way, $L_{m,n}(s)$ extended to a meromorphic function on the entire complex plane, holomorphic on $\mathrm{Re}(s) \geq 1$ except for having a pole of order k at $s = 1$ the analytic proposition above would tell us that the limit is k, rather than 0.

This analytic theorem follows from classical results due to Weiner and Ikehara. A beautiful discussion of these ideas and proofs can be found in the Appendix to Chapter 1 of Serre's monograph **[7]**. See also Tate's article **[11]**, Shahidi's article **[9]** and Serre's letter to Shahidi **[8]** that discusses in some detail the implications in the direction of the Sato-Tate conjecture that would follow if one assumes that the $L_{0,\nu}$'s satisfy the hypotheses of Proposition 2.4 for $\nu \leq d$. This knowledge is known for $\nu = 1$ (our "sample problem" comes from a modular form; indeed by the celebrated results regarding modularity of elliptic curves, it would be known for any elliptic curve defined over Q that we care to choose). It is also known for $\nu = 2$ using an integral representation due to Shimura **[10]**; see also Gelbart's and Jacquet's article **[2]**. It is known for $\nu = 3, 4$ by work of Shahidi; see the enlightening discussion in **[9]** about this)[3].

2.5. **The coming together of different mathematical viewpoints.** But how can we get that Dirichlet series such as $L_{m,n}(s)$ extend meromorphically to the entire complex plane for *enough* values of (m, n) to guarantee that we have computed all the moments of the distribution determined by our data? And how can we determine whether these meromorphic extensions have (or better: don't have) zeroes or poles on the line $Re(s) = 1$? A standard strategy—in fact, it seems, the only known strategy to get L-functions to have all the analytic properties that they need to have is to connect these L-functions with automorphic forms over Q or with pairs of automorphic forms on GL_m and GL_n over Q relying on ideas of Rankin-Selberg. For the problem we are interested in, it turns out that one gets sufficiently valuable information if one can make the analogous connection with automorphic forms over *some* number field F—not necessarily Q—so long as F is Galois over Q, and totally real.

Part of the beauty of the new theorem we are discussing—which applies, in fact, to all elliptic curves over Q that have at least one prime of multiplicative reduction—is how it pulls together work from significantly different viewpoints. There are three major pieces that go into it: work of Laurent Clozel, Michael Harris, and Richard Taylor) on modular lifting and *automorphic representation theory*; work of Michael Harris, Nicholas Shepherd-Barron, and

[3] *The corresponding symmetric cube and fourth power of the modular form of weight two (corresponding to our sample problem) are cuspidal automorphic forms; cf. the articles* **[4]**, **[5]** *by Kim and Shahidi.*

Richard Taylor bringing in an extraordinary piece of *algebraic geometry:* the pencil of Calabi-Yau varieties

$$X_0^{n+1} + X_1^{n+1} + \cdots + X_n^{n+1} = (n+1)tX_0X_1 \ldots X_n$$

for even values of n, parametrized by the variable t; and the last: Richard Taylor's major discovery in *Galois deformation theory* which, using ideas of Mark Kisin, improved dramatically the mechanism of modular lifting, allowing Richard Taylor to prove this extraordinary result.

2.6. **Expository accounts of this recent work.** Different audiences benefit from different shapes of exposition. I wrote a brief “news" article in the journal Nature **[6]** meant to give a hint of the nature of the Sato-Tate Conjecture and some related mathematical problems to scientists who are not necessarily familiar with much modern mathematics. For professional mathematicians, a number of excellent articles and videos—requiring different levels of prerequisites of their audiences—are devoted to exposing this material:

(1) Available through the MSRI website (http://www.msri.org/):
   (a) An introductory one hour lecture by Nicholas Katz emphasizing the background and the historical perspective of the work.
   (b) A series of lectures for a number theory workshop, by Richard Taylor; by Michael, Harris; and by Nicholas Shepherd-Barron, where an exposition of the proof itself is given.

(2) Two hours of expository lectures by Laurent Clozel on this topic which goes in considerable detail through the ideas of the proof, aimed at a general mathematical audience, delivered in the conference on Current Developments in Mathematics, at Harvard University. The notes for these should soon be available as well.

(3) An expository article by Michael Harris: “The Sato-Tate Conjecture: introduction to the proof." This will be submitted to the proceedings of the École d'été Franco-Asiatique, held at the IHES during the summer of 2006.

References

[1] Clozel,L., Harris, M., Taylor, R.: Automorphy for some ℓ-adic lifts of automorphic mod l representations (preprint) http://www.math.harvard.edu/ rtaylor/

[2] Gelbart, S., Jacquet, H.: A relation between automorphic representations of $G(2)$ and $GL(3)$, Ann. Sci. École Norm. Sup. (4) **11** (1978) 471-552

[3] Harris, M., Shepherd-Barron, N., Taylor, R.: Ihara's lemma and potential automorphy (preprint) http://www.math.harvard.edu/ rtaylor/

[4] Kim, H., Shahidi, F.: Cuspidality of symmetric powers with applications, Duke Math. J. **112**, no. 1 (2002), 177Ð197

[5] Kim, H., Shahidi, F.: Functorial products for $\mathrm{GL}(2) \times \mathrm{GL}(3)$ and the symmetric cube for $\mathrm{GL}(2)$, Annals of Math. **155** (2002), 837-893

[6] Mazur, B.: Controlling our Errors, Nature Vol **443**, **7** (2006) 38-40

[7] Serre, J.-P.: *Abelian ℓ-adic Representations* Benjamin (1968)

[8] Serre, J.-P.: Letter to F. Shahidi, January 24, 1992; Appendix (pp. 175-180) in reference 9 below,

[9] Shahidi, F.: Symmetric power L-functions for GL(2), pp. 159-182, in *Elliptic Curves and Related Topics,* Volume 4 of CRM Proceedings and Lecture Notes (Eds.: H. Kisilevsky, M.R. Murty), American Mathematical Society, (1994)

[10] Shimura, G.: On the holomorphy of certain Dirichlet series, Proc. London Math. Soc (3) **31** (1975) 79-98

[11] Tate, J.: Algebraic Cycles and Poles of Zeta Functions, pp.93-110 in *Arithmetic Algebraic Geometry*, Proceedings of a conference held in Purdue, Dec. 5-7 1963, Harpers (1965)

[12] Taylor, R.: Automorphy for some ℓ-adic lifts of automorphic mod ℓ representations. II (preprint) http://www.math.harvard.edu/ rtaylor/

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 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

2007 CURRENT EVENTS BULLETIN *Committee*

Alejandro Adem, *University of British Columbia*

Peter Bates, *Michigan State University*

David Eisenbud, *Mathematical Sciences Research Institute*, Chair

Daniel Freed, *University of Texas, Austin*

Edward Frenkel, *University of California, Berkeley*

Susan Friedlander, *University of Illinois, Chicago*

Andrew Granville, *University of Montreal*

Martin Golubitsky, *University of Houston*

Carlos Kenig, *University of Chicago*

László Lovász, *Microsoft Research*

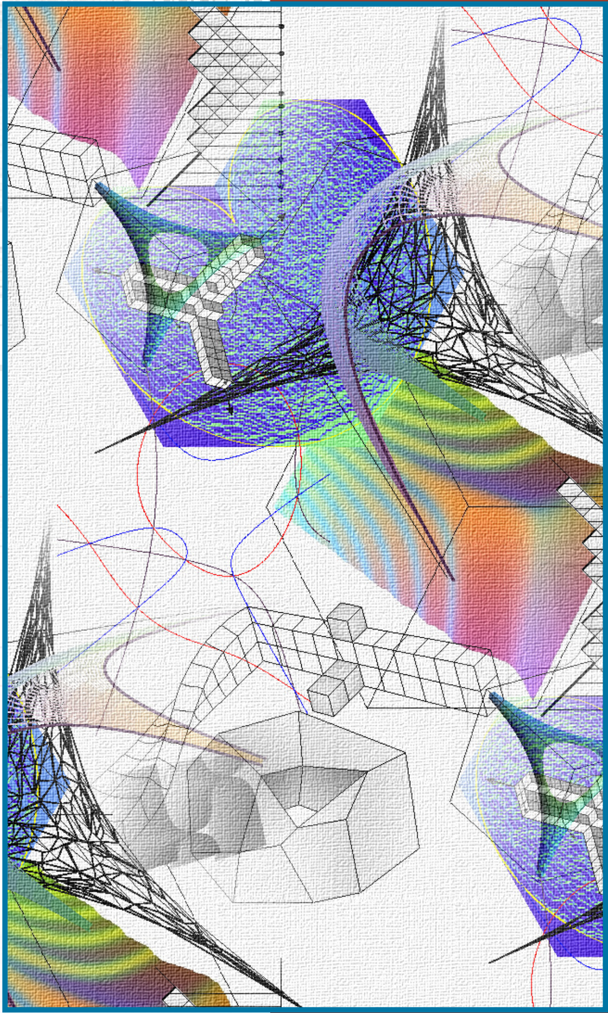
Lauren Ancel Meyers, *University of Texas, Austin*

Michael Singer, *North Carolina State University*

Madhu Sudan, *Massachusetts Institute of Technology*

Kannan Soundararajan, *University of Michigan*

Margaret Wright, *Courant Institute*



The back cover graphic is reprinted courtesy of Andrei Okounkov.

The cover graphic associated with Akshay Venkatesh's talk is reprinted courtesy of: Svetlana Katok and Ilie Ugarcovici, "Geometrically Markov Geodesics on the Modular Surface, "Mosc. Math. J. 5 (2005), 135-155.

The cover graphic associated with Barry Mazur's talk is reprinted courtesy of: William Stein. For details of the computation, see http://sage.math.washington.edu/sato-tate.