Artur Avila · Michael Th. Rassias
Yakov Sinai *Editors*

Analysis at Large

Dedicated to the Life and Work of Jean Bourgain

Springer

Analysis at Large

Artur Avila • Michael Th. Rassias • Yakov Sinai
Editors

Analysis at Large

Dedicated to the Life and Work of Jean Bourgain

Springer

Editors

Artur Avila
Institute of Mathematics
University of Zurich
Zürich, Switzerland

Instituto Nacional de Matemática
Pura e Aplicada (IMPA)
Rio de Janeiro, Brasil

Yakov Sinai
Department of Mathematics
Princeton University
Princeton, NJ, USA

Michael Th. Rassias
Department of Mathematics and
Engineering Sciences
Hellenic Military Academy
Athens, Greece

Institute for Advanced Study
Program in Interdisciplinary Studies
Princeton, USA

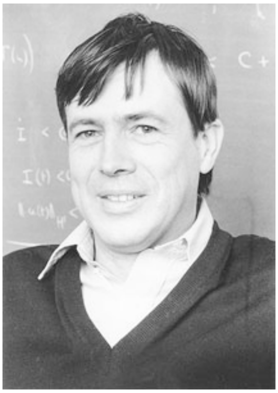
ISBN 978-3-031-05330-6 ISBN 978-3-031-05331-3 (eBook)
https://doi.org/10.1007/978-3-031-05331-3

Mathematics Subject Classification: 11-XX, 26-XX, 28-XX, 30-XX, 31-XX, 32-XX, 33-XX, 34-XX, 35-XX, 37-XX, 39-XX, 40-XX, 41-XX, 42-XX, 43-XX, 44-XX, 45-XX, 46-XX, 47-XX, 49-XX, 52-XX, 53-XX, 57-XX, 58-XX, 60-XX, 65-XX, 68-XX, 70-XX, 76-XX, 81-XX, 83-XX, 90-XX, 91-XX, 93-XX

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Fig. 1 Photograph of Jean Bourgain by Randall Hagadorn (1994), Institute for Advanced Study Public Affairs photographs. From the Shelby White and Leon Levy Archives Center, Institute for Advanced Study, Princeton (N.J.).



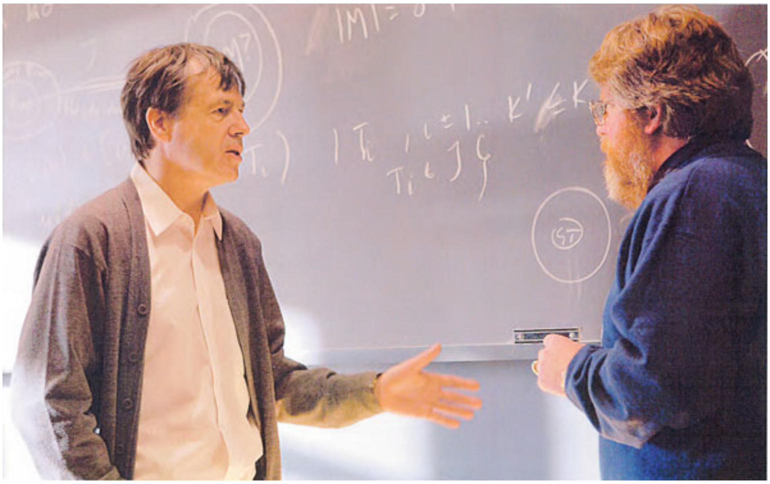


Fig. 2 Photograph of Jean Bourgain and Russell Impagliazzo by Andrea Kane (circa 2010), Institute for Advanced Study Public Affairs photographs. From the Shelby White and Leon Levy Archives Center, Institute for Advanced Study, Princeton (N.J.)

Preface

Analysis at Large is a book dedicated to the great mathematician Jean Bourgain, who passed away on December 22, 2018. His profound research has deeply influenced an array of mathematical areas, with its main focus being in mathematical analysis and its various facets and interconnections with other fields.

The present book publishes a highly selective collection of research and survey papers in a wide spectrum of subjects which have been deeply influenced by Bourgain's monumental contributions and have led to celebrated breakthroughs in mathematics. More specifically, topics investigated within this book include Bourgain's discretized sum-product theorem, Bourgain's work in nonlinear dispersive equations, the slicing problem by Bourgain, harmonious sets, the joint spectral radius, equidistribution of affine random walks, Cartan covers and doubling Bernstein type inequalities, a weighted Prékopa-Leindler inequality and sumsets with quasicubes, the fractal uncertainty principle for the Walsh-Fourier transform, the continuous formulation of shallow neural networks as Wasserstein-type gradient flows, logarithmic quantum dynamical bounds for arithmetically defined ergodic Schrödinger operators, polynomial equations in subgroups, trace sets of restricted continued fraction semigroups, exponential sums, twisted multiplicativity and moments, the ternary Goldbach problem, as well as the multiplicative group generated by two primes in $\mathbf{Z}/Q\mathbf{Z}$.

The papers have been contributed by leading experts in the corresponding topics and present the state of the art in the problems treated, each paying homage to the life and work of this pioneer in mathematics.

We are grateful to the mathematicians who have participated in this publication, for contributing their valuable works in honor of J. Bourgain.

We also wish to thank the staff at Springer for their help throughout the publication process of this book.

Zurich, Switzerland — Artur Avila
Athens, Greece — Michael Th. Rassias
Princeton, NJ, USA — Yakov Sinai

Contents

On the Joint Spectral Radius

Emmanuel Breuillard

Dedicated to the memory of Jean Bourgain

Abstract For a bounded subset S of $d \times d$ complex matrices, the Berger-Wang theorem and Bochi's inequality allow to approximate the joint spectral radius of S from below by the spectral radius of a short product of elements from S. Our goal is twofold: we review these results, providing self-contained proofs, and we derive an improved version with explicit bounds that are polynomial in d. We also discuss other complete valued fields.

1 Introduction

We denote by $\|\cdot\|$ a norm on $\mathbb{C}^d$ and its associated operator norm on the ring of $d \times d$ matrices $M_d(\mathbb{C})$. For a bounded subset $S \subset M_d(\mathbb{C})$, we let $\|S\| := \sup_{s \in S} \|s\|$. The *joint spectral radius* [3, 10, 15, 27, 29] is defined by:

$$\rho(S) := \lim_{n \to +\infty} \|S^n\|^{\frac{1}{n}} \tag{1}$$

where $S^n := \{s_1 \cdot \ldots \cdot s_n, s_i \in S\}$ is the n-th fold product set. From the submultiplicativity of the operator norm, it is clear that the limit exists, is independent of the choice of norm, and coincides with the infimum of all $\|S^n\|^{\frac{1}{n}}$, $n \geq 1$. A straightforward consequence is that $S \mapsto \rho(S)$ is upper-semicontinuous for the Hausdorff topology. Moreover $\rho(S^k) = \rho(S)^k$ for every $k \in \mathbb{N}$. It is also clear that $\rho(gSg^{-1}) = \rho(S)$ for every $g \in \mathrm{GL}_d(\mathbb{C})$. Rota and Strang [27] observed that $\rho(S)$ is equal to the infimum of $\|S\|$ as the norm varies among all possible norms on $\mathbb{C}^d$. Combined with John's ellipsoid theorem, this easily yields:

E. Breuillard (✉)
DPMMS, University of Cambridge, Cambridge, UK
e-mail: Emmanuel.Breuillard@maths.ox.ac.uk

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_1

Lemma 1 *Given a norm $\|\cdot\|$ on $\mathbb{C}^d$, for any bounded subset $S \subset M_d(\mathbb{C})$, we have:*

$$\rho(S) \leq \inf_{g \in \mathrm{GL}_d(\mathbb{C})} \|gSg^{-1}\| \leq d \cdot \rho(S).$$

When S is irreducible (i.e., does not preserve a proper subspace of $\mathbb{C}^d$), it turns out that there is norm such that $\rho(S) = \|S\|$. The existence of such *extremal norms* will be reviewed in Sect. 2 along with related known facts. It also follows easily from this that $\rho(S) = 0$ if and only if the subalgebra $\mathbb{C}[S]$ generated by S is nilpotent.

It turns out that $\rho(S)$ can also be approximated from below by eigenvalues. Let $\Lambda(s)$ be the largest modulus of an eigenvalue of $s \in M_d(\mathbb{C})$ and

$$\Lambda(S) := \max_{s \in S} \Lambda(s).$$

It is clear that $\Lambda(S) \leq \rho(S)$ and thus $\Lambda(S^n)^{\frac{1}{n}} \leq \rho(S)$ for all n. When S is a singleton, the classical Gelfand formula asserts that $\Lambda(s) = \rho(\{s\})$. For several matrices, the key fact is as follows:

Theorem 1 (Berger-Wang [3])

$$\rho(S) = \limsup_{n \to +\infty} \Lambda(S^n)^{\frac{1}{n}}.$$

An immediate consequence is that $S \mapsto \rho(S)$ is also lower-semicontinuous and hence continuous for the Hausdorff topology. Theorem 1 had been conjectured by Daubechies and Lagarias [10]. Elsner [11] gave a simple proof of it. In this article we will be interested in giving explicit estimates quantifying this convergence. Our first observation is that in fact the following slightly stronger result holds:

Theorem 2 *Let $S \subset M_d(\mathbb{C})$ be a bounded subset with $\rho(S) > 0$. Then*

$$\limsup_{n \to +\infty} \frac{\Lambda(S^n)}{\rho(S)^n} = 1.$$

The question of the speed of convergence in Theorem 1 or 2 is an interesting one and goes back at least to the Lagarias-Wang finiteness conjecture [18], which posited that the limsup should be attained at a certain finite n. This has been disproved by Bousch and Mairesse [7] for 2 × 2 matrices (see also [13, 14, 21]), and Morris (see [22, Thm 2.7]) gave an example with $S = \{a, b\} \subset \mathrm{SL}_2(\mathbb{R})$. In general, counterexamples are thought to be rare.

Elsner's proof of Theorem 1 is based on a pigeonhole argument, which we will revisit in this note and can roughly be described as follows under the assumption that S is irreducible: if $\rho(S) = 1$, then given a unit vector $x \in \mathbb{C}^d$, we may always find $s \in S$ such that sx is also a unit vector (this follows from the existence of a Barabanov norm, see Sect. 2), so iterating this construction, we eventually find a short product $w = s_n \cdot \ldots \cdot s_k$ with wy close to $y = s_{k-1} \cdot \ldots \cdot s_1 x$, implying that w has an eigenvalue close to 1. This idea also leads to a proof of Theorem 2 and to the following quantitative and explicit version:

Theorem 3 *Let $S \subset M_d(\mathbb{C})$ be a bounded subset with $\rho(S) = 1$. Set $n_0(d) = 3^d 4^{d^2}$ and let $\varepsilon > 0$. If $n \geq \varepsilon^{-d^2} n_0(d)$, then*

$$\max_{k \leq n} \Lambda(S^k) \geq 1 - \varepsilon.$$

This yields a polynomial decay of the form $|\sup_{k \leq n} \Lambda(S^k) - 1| = O_{S,d}(n^{-1/d^2})$ in Theorem 2 when $\rho(S) = 1$. In [20] Morris proved a much stronger super-polynomial upper bound on the speed of convergence: that is, $|\sup_{k \leq n} \Lambda(S^k) - 1| = O_{A,S}(n^{-A})$ for every $A \geq 1$, provided S is finite and $\rho(S) = 1$. However the implied constant is not explicit. He also points out that his argument fails when S is infinite.

In this note we will be interested in the d aspect. The bound on n in Theorem 3 is super-exponential in d. If we aim to approximate the joint spectral radius no longer up to a small error, but only up to a constant multiple, we can expect polynomial bounds in d. In this vein, Bochi [5, Theorem B] established the following general inequality:

Theorem 4 (Bochi [5]) *There are constants $c(d) > 0, N(d) > 0$ such that for every bounded set $S \subset M_d(\mathbb{C})$ we have:*

$$\max_{1 \leq k \leq N(d)} \Lambda(S^k)^{\frac{1}{k}} \geq c(d) \cdot \rho(S). \tag{2}$$

Note that Theorem 1 (but not Theorem 2) follows immediately from Bochi's inequality: indeed apply the inequality to S^n and let n tends to infinity. On the other hand, Theorem 3 implies Bochi's inequality with $N(d) = 3^d 8^{d^2}$ and $c(d) = \frac{1}{2}$ say. We are interested in quantifying the constants $c(d)$ and $N(d)$ in terms of the dimension d. Example 1 (2) below shows that $N(d) \geq d$. Bochi's proof gave $N(d) = 2^d - 1$, but a non-constructive $c(d)$ obtained via a topological argument involving some geometric invariant theory.

In [8, 2.7, 2.9], another non-constructive proof was given with $N(d) = d^2$. This proof actually allows to take for $N(d) = \ell(d)$ the smallest upper bound on the integer k such that for any $S \subset M_d(\mathbb{C})$ the powers $S, \ldots, S^k$ span linearly the matrix algebra $\mathbb{C}[S]$ generated by S. It is immediate that $\ell(d) \leq d^2$, but in a recent breakthrough, Shitov [28] has proved that $\ell(d) \leq 2d(\log_2 d + 2)$ greatly improving an earlier bound in $O(d^{3/2})$ due to Pappacena [24].

In order to motivate our main result and since it is very short, we give now a direct proof of Theorem 4 using the following slight variant of the argument from [8]:

Claim 1 *There is $c'(d) > 0$ such that for every bounded subset S of $M_d(\mathbb{C})$ with $\rho(S) = 1$ there is a non-zero idempotent $p \in M_d(\mathbb{C})$ (i.e., $p^2 = p$) such that $c'(d)p$ belongs to the complex convex hull of $S, \dots, S^{\ell(d)}$.*

Proof By the complex convex hull $Conv(Q)$ of $Q \subset M_d(\mathbb{C})$, we mean the set of linear combinations $\alpha_1 q_1 + \dots + \alpha_n q_n$ with $q_i \in Q$ and $|\alpha_1| + \dots + |\alpha_n| = 1$. Since the problem is invariant under conjugation, in view of Lemma 1, we may assume that S is confined to a bounded region of $M_d(\mathbb{C})$, allowing us to pass to a Hausdorff limit of potential counterexamples to the claim. By compactness and upper semi-continuity of the joint spectral radius, we get a bounded subset S with $\rho(S) \geq 1$, but such that $Conv(S \cup \dots \cup S^{\ell(d)})$ contains no scalar multiple of an idempotent. In particular, $\mathbb{C}[S]$ contains no idempotent. By the Artin-Wedderburn theorem, this means that $\mathbb{C}[S]$ is a nilpotent subalgebra of $M_d(\mathbb{C})$. In particular, $S^d = 0$, which is in contradiction with $\rho(S^d) = \rho(S)^d \geq 1$. □

Proof of Theorem 4 If $\rho(S) = 0$, there is nothing to prove. Otherwise, rescaling we may assume $\rho(S) = 1$. If the left-hand side in (2) is at most c (and we may assume $c \leq 1$, so $c^k \leq c$), then the trace of any element in S^k, $k \leq \ell(d)$, is at most cd in modulus. The trace of the idempotent element p found in Claim 1 is a non-zero integer. So $c'(d) \leq c'(d)|\mathrm{Tr}(p)| \leq cd$. So setting $c(d) = c'(d)/d$ and $N(d) = \ell(d)$ yields (2) as desired. □

As with Bochi's original argument, this one does not give any explicit estimate on the constant $c(d)$. It is however possible to "effectivize" the argument just given: this requires effectivizing the proof of Wedderburn's theorem and, after a fairly painstaking analysis, the details of which we will spare the reader, yields a rather poor lower bound on $c(d)$ of doubly exponential type in d. Another route is to use an idea appearing in the work of Oregon-Reyes [23, Rk. 4.5], which consists in using the effective arithmetic nullstellensatz by making explicit the implication $\{\mathrm{Tr}(S^k) = 0 \text{ for all } k = 1, \dots, \ell(d)\} \Rightarrow \{S^d = 0\}$. This also yields an effective bound on $c(d)$, which is again unfortunately rather poor, at least doubly exponential in d.

The following result, which is the main contribution of this note, gives explicit polynomial bounds on both $c(d)$ and $N(d)$.

Theorem 5 *For every bounded set $S \subset M_d(\mathbb{C})$, we have:*

$$\max_{1 \leq k \leq 2d^3} \Lambda(S^k)^{\frac{1}{k}} \geq \frac{1}{2^8 d^5} \cdot \rho(S).$$

In particular, applying this to S^m for a suitable integer m, we also have:

$$\max_{1 \leq k \leq N_2(d)} \Lambda(S^k)^{\frac{1}{k}} \geq \frac{1}{2} \cdot \rho(S),$$

where $N_2(d) = 2d^3 \lceil 8 + 5 \log_2 d \rceil$.

By the same trick of replacing S by S^m for suitable m, the factor $\frac{1}{2}$ can of course be replaced by any number $\kappa < 1$ provided $N_2(d)$ is replaced by $N_{\kappa^{-1}}(d) := 2d^3\lceil \log_{\kappa^{-1}}(2^8d^5)\rceil$. The proof exploits a different kind of pigeonhole argument, where one argues, as in the classical Siegel lemma in number theory, that some non-zero linear combination with small integer coefficients of the iterates $s_n \cdot \ldots \cdot s_1 x$ will vanish or be very small. In turn, this forces one of the products to have a spectral radius bounded away from zero.

The following natural questions then arise:

Questions How sharp is the bound $d^{3+o(1)}$ on $N_2(d)$? We only know that $N_2(d)$ must be at least d. Is there a polynomial bound on $c'(d)$ in Claim 1 above?

In [5, Theorem A], Bochi proves another inequality, giving this time a lower bound on $\rho(S)$ in terms of the norms of S^n, which, when iterated, gives a speed of convergence for (1); see [16]. Given any norm $\|\cdot\|$ on $\mathbb{C}^d$,

$$\|S^d\| \leq C_0(d)\rho(S)\|S\|^{d-1}. \tag{3}$$

While no explicit bound on $C_0(d)$ was given in [5], his proof gives a super-polynomial bound in $d^{3d/2}$ (see [16, Section 4]). It turns out that the pigeonhole argument for our Theorem 5 gives a polynomial bound for (3) at the expense of increasing the power of S:

Theorem 6 *Let $S \subset M_d(\mathbb{C})$ be a bounded subset and set $n_1 = 2d^2$. Then*

$$\|S^{n_1}\| \leq 2^7 d^4 \rho(S)\|S\|^{n_1-1}. \tag{4}$$

Iterating (4) yields an explicit estimate quantifying the convergence in (1) improving the bounds obtained in [16, Theorem 1].

Finally we examine what happens when the field $\mathbb{C}$ is replaced by an arbitrary algebraically closed complete valued field $(K, |\cdot|)$. By Ostrowski's theorem, if K is not $\mathbb{C}$, it must be non-Archimedean (for instance, $\mathbb{C}_p$ the completion of the algebraic closure of the field of p-adic numbers $\mathbb{Q}_p$, or the completion of the field of Laurent series over the algebraic closure of $\mathbb{F}_p$). All of the above makes sense of course, and the joint spectral radius is defined in the same way. As it turns out, the analogues of the results above are much simpler for such K, the Lagarias-Wang finiteness conjecture holds in a uniform way, and in fact:

Theorem 7 *Let K be an algebraically closed non-archimedean complete valued field. Consider an ultrametric norm $\|\cdot\|_0$ on K^d and a bounded subset S of $M_d(K)$. Then*

$$\max_{1\leq k\leq \ell(d)} \Lambda(S^k)^{\frac{1}{k}} = \rho(S) = \inf_{g\in \mathrm{GL}_d(K)} \|gSg^{-1}\|_0. \tag{5}$$

Moreover, $\rho(S) > 0$ if and only if the subalgebra generated by S is not nilpotent, in which case there is an ultrametric norm $\|\cdot\|$ on K^d with $\|S\| = \rho(S)$.

Recall that $\ell(d)$ denotes the smallest integer k such that for any field F and any $S \subset M_d(F)$ the power sets $S, \ldots, S^k$ span linearly the algebra $F[S]$. Obviously $\ell(d) \leq d^2$ and recall that in fact $\ell(d) \leq 2d \log_2 d + 4d - 4$ by [28].

If K is not algebraically closed, then (5) still holds with K replaced by its algebraic closure $\overline{K}$ (indeed, the absolute value extends uniquely to $\overline{K}$ and the completion of $\overline{K}$ will remain algebraically closed by Kürschák's theorem [26, 5.J.]).

In the special case, when K is a local field and S a compact subgroup of $\mathrm{GL}_d(K)$, the last assertion of the theorem recovers the well-known Bruhat-Tits fixed point theorem: the norm $\| \cdot \|$ will be preserved by S and thus be a fixed point in the Bruhat-Tits building of ultrametric norms [12].

Theorem 7 was proved in [8] for $K = \mathbb{C}_p$. We will give a slightly more direct proof of the general case.

Similarly, the analogue of Theorem 6 reads:

Theorem 8 *For any ultrametric norm $\| \cdot \|_0$ on K^d and $S \subset M_d(K)$ bounded*

$$\|S^d\|_0 \leq \rho(S) \|S\|_0^{d-1}. \tag{6}$$

2 Extremal Norms and Barabanov Norms

In this section we recall some well-known facts about the joint spectral radius and extremal norms providing complete and self-contained proofs. Most of the material can be found in the first chapters of the book [15]. We then prove Theorem 2.

We begin by the observation of Rota and Strang mentioned in the introduction. Recall that $\| \cdot \|$ denotes both a norm on $\mathbb{C}^d$ and its associated operator norm and that for some subset Q (in either $\mathbb{C}^d$ or $M_d(\mathbb{C})$), we set $\|Q\| := \sup_{q \in Q} \|q\|$.

Lemma 2 (Rota-Strang) *Let $S \subset M_d(\mathbb{C})$ be a bounded subset.*

$$\rho(S) = \inf_{\|\cdot\|} \|S\|, \tag{7}$$

where the infimum is over all norms on $\mathbb{C}^d$.

Proof Let $r > 0$ with $r\rho(S) < 1$ and consider the norm $v_r(x) := \sum_{n \geq 0} \|S^n x\| r^n$. Clearly $v_r(sx) \leq \frac{1}{r} v_r(x)$ for all $s \in S$. So $v_r(S) \leq r^{-1}$. Letting r^{-1} tend to $\rho(S)$ yields the result. □

Lemma 1 follows immediately by combining Lemma 2 with the following well-known fact:

Lemma 3 (John's Ellipsoid) *If v is a norm on $\mathbb{C}^d$ and $\| \cdot \|_2$ the standard hermitian norm, then there is $g \in \mathrm{GL}_d(\mathbb{C})$ such that for all $x \in \mathbb{C}^d$*

$$\|gx\|_2 \leq v(x) \leq \sqrt{d} \cdot \|gx\|_2.$$

In particular if $w(x)$ is any another norm, then for some $h \in \mathrm{GL}_d(\mathbb{C})$

$$w(hx) \leq v(x) \leq d \cdot w(hx).$$

Proof According to John's ellipsoid theorem (e.g., [1]), every symmetric convex body K in $\mathbb{R}^k$ contains a unique ellipsoid E of maximal volume and E moreover satisfies $K \subset \sqrt{d}E$. If K is the ball of radius 1 of the complex norm v in $\mathbb{C}^d = \mathbb{R}^{2d}$, then the uniqueness implies that the norm associated to E is hermitian, hence of the form $\|gx\|_2$ for some $g \in \mathrm{GL}_d(\mathbb{C})$. □

Remark 1 This argument shows that the constant d in Lemma 1 can be replaced by $\sqrt{d}$ if the norm is $\|\cdot\|_2$. In fact a more subtle argument (see, e.g., [4]) shows that it can be replaced with $\sqrt{\min\{k, d\}}$ in case S has k elements.

One says that S is *irreducible* if it does not preserve a non-trivial proper subspace of $\mathbb{C}^d$. It is said to be *product bounded* if the semigroup it generates $T := \bigcup_{n \geq 1} S^n$ is bounded. The following is also classical (see [2, 3, 11, 29]):

Lemma 4 (Extremal Norms) *Suppose S is irreducible. Then $\rho(S) > 0$, $S/\rho(S)$ is product bounded and the infimum in (7) is attained.*

Norms realizing the infimum in (7) are called *extremal norms*.

Proof By Burnside's theorem, the subalgebra $\mathbb{C}[S]$ generated by S is all of $M_d(\mathbb{C})$. Since $\mathbb{C}[S]$ is linearly spanned by $S \cup \ldots \cup S^{d^2}$, we may express each element of the canonical basis E_{ij} of $M_d(\mathbb{C})$ as a linear combination of elements from T. Given that $\mathrm{Tr}(E_{ii}) = 1$, this means that at least one element of T has non-zero trace, which clearly forces $\rho(S) > 0$. Rescaling, we may assume without loss of generality that $\rho(S) = 1$. In particular $|\mathrm{Tr}(t)| \leq d$ for all $t \in T$ and thus $|\mathrm{Tr}(tE_{ij})|$ is bounded independently of $t \in T$, which means that T is bounded. Finally, given any norm $\|\cdot\|$ on $\mathbb{C}^d$ and setting $v(x) := \|Tx\|$, we get a well-defined norm such that $v(sx) \leq v(x)$ for all $s \in S$. Hence v is an extremal norm. □

The example of a single non-trivial unipotent matrix shows that the infimum in (7) is not attained in general. If S is not irreducible, it can be put in block triangular form in some basis of $\mathbb{C}^d$. Therefore the following is an immediate consequence of the previous lemma (recall that an algebra N is nilpotent if there is an integer n such that $N^n = 0$).

Corollary 1 *Let S be a bounded subset of $M_d(\mathbb{C})$. Then $\rho(S) = 0$ if and only if $\mathbb{C}[S]$ is a nilpotent subalgebra of $M_d(\mathbb{C})$.*

If S is irreducible and $\rho(S) = 1$, T is bounded, and we may define

$$v(x) := \limsup_{n \to +\infty} \|S^n x\|. \tag{8}$$

Then v is a norm, because $v(x) = 0$ for some $x \neq 0$ implies $v(S^n x) = 0$ for all n, which implies by irreducibility that v is identically zero and hence that $\rho(S) = 0$. In particular:

Lemma 5 (Barabanov Norms) *Let S be an irreducible bounded subset of $M_d(\mathbb{C})$, then there is a complex norm v on $\mathbb{C}^d$ such that for all $x \in \mathbb{C}^d$,*

$$\max_{s \in S} v(sx) = \rho(S) \cdot v(x). \tag{9}$$

Proof Indeed we may define v as in (8) for S replaced by $S/\rho(S)$. □

A norm satisfying (9) is a special kind of extremal norm called a *Barabanov norm* (see [2, 17, 25, 29]). Such norms are not unique in general (e.g., in Example 1 4. below any norm $\|\cdot\|$ on $\mathbb{C}^d$ with $\varepsilon\|x\|_2 \leq \|x\| \leq \|x\|_2$ is a Barabanov norm for S), but they can be in some situations [19].

Another object is naturally associated to S when $\rho(S) = 1$; it is the *attractor semigroup* [2, 29]

$$T_\infty := \bigcap_{n \geq 1} \overline{S^n T}.$$

In other words, this is the set of limit points of finite products $s_1 \cdot \ldots \cdot s_n$ whose length n tends to infinity. It is clearly compact, and for every operator norm, it contains an element of norm at least 1. Indeed otherwise we would have $\|S^n\| < 1$ for some n and thus $\rho(S^n) < 1$, which is impossible as $\rho(S^n) = \rho(S)^n = 1$. By construction, the Barabanov norm (8) is also equal to $v(x) = \max_{t \in T_\infty} \|tx\|$. Furthermore, it is straightforward that $T_\infty = T_\infty S = S T_\infty$ and $T_\infty^2 = T_\infty$ and that:

Lemma 6 *Suppose S is irreducible with $\rho(S) = 1$. Then T_∞ is also irreducible and $\rho(T_\infty) = 1$.*

Proof For every non-zero $x \in \mathbb{C}^d$, the linear span $\langle T_\infty \rangle x$ contains $\langle T_\infty \rangle S^k x$ for each k and hence $\langle T_\infty \rangle \mathbb{C}^d$ by irreducibility of S. So this must be 0 or $\mathbb{C}^d$. The former is impossible, because $T_\infty \neq \{0\}$ by the above discussion. So T_∞ is irreducible. Finally by construction $v(T_\infty) = 1$ and $T_\infty^k = T_\infty$ for every k. Hence $\rho(T_\infty) = 1$. □

We are now in a position to prove Theorem 2.

Lemma 7 (Existence of an Idempotent) *Suppose S is a bounded irreducible subset of $M_d(\mathbb{C})$ with $\rho(S) = 1$. Then the attractor semigroup T_∞ contains a non-zero idempotent.*

Proof Let K be the subset of T_∞ made of elements with operator norm 1 for the Barabanov norm (8). We have already seen that K is non-empty. If ab has norm one and $a, b \in T_\infty$, then both a and b have norm one. So $K \subset K^2$. Starting from some $t_0 \in K$, we may write $t_0 = t_1 s_1$ with $t_1, s_1 \in K$, and then similarly $t_1 = t_2 s_2$, etc. For each n we have $t_0 = t_n s_n \cdot \ldots \cdot s_1$. By compactness of K, there is a subsequence n_i

such that $s_{n_i} \cdot \ldots \cdot s_1$ converges, say towards $k \in K$. Passing to a further subsequence, we may assume that $s_{n_{i+1}} \cdot \ldots \cdot s_{n_i+1}$ also converges, say towards $u \in K$. At the limit we have $k = uk$. But there is a unit vector x such that $y := kx$ has norm 1. Hence $y = uy$ and u has 1 as an eigenvalue. So T_∞ contains an element u with eigenvalue 1. Now looking at u in Jordan normal form and considering large powers of u, we see that the Jordan blocks with eigenvalue of modulus 1 must be of size 1, because powers of non-trivial unipotents are unbounded. Therefore $\{u^n\}_{n \geq 1}$ contains an idempotent in its closure. □

Note that T_∞ may contain 0, so the lemma does not follow from a general result guaranteeing the existence of idempotents in compact semigroups such as the Ellis-Numakura lemma.

Proof of Theorem 2 We first assume that S is irreducible. Rescaling, we may assume that $\rho(S) = 1$. By Lemma 7, T_∞ contains an idempotent. In particular, $\Lambda(T_\infty) = 1$, which implies what we want. The general case follows from the irreducible one. Indeed if S is not irreducible, it can be put in block triangular form, and if S_{ii} denotes the i-th diagonal block, then it is straightforward to check (either from the definition or more directly from Theorem 1) that $\rho(S) = \max_i \rho(S_{ii})$. □

Example 1 The following are examples of irreducible subsets of $M_d(\mathbb{C})$ with joint spectral radius equal to 1.

1. $S = \{E_{ij}\}_{ij}$ the elementary matrices in $M_d(\mathbb{C})$. Note that S is made of rank 1 elements and $T_\infty = S \cup \{0\}$.
2. $S = \{E_{i,i+1}\}_{1 \leq i < d} \cup \{E_{d1}\}$. Note that $T = T_\infty = \{0\} \cup \{E_{ij}\}_{ij}$.
3. $S = U_d(\mathbb{C}) \cup \{t\}$, where $U_d(\mathbb{C})$ is the group of unitary matrices and $t = \mathrm{diag}(\alpha_1, \ldots, \alpha_d)$ with $|\alpha_i| < 1$. Then $T_\infty = T \cup \{0\}$.
4. $S = \{\mathrm{id}\} \cup \varepsilon U_d(\mathbb{C})$ for $\varepsilon < 1$. Then $T_\infty = T \cup \{0\}$.

3 Explicit Bounds for Theorem 2

In this section we prove Theorem 3. We need a basic lemma.

Lemma 8 *Let $\|\cdot\|$ be a norm on $\mathbb{C}^d$. Let $A \in M_d(\mathbb{C})$ and $x \in \mathbb{C}^d$ with $\|A\| \leq 1$ and $\|x\| = 1$. Let $\varepsilon > 0$ and $\lambda \in \mathbb{C}$ with $|\lambda| \leq 2$. Assume that $\|Ax - \lambda x\| \leq (\varepsilon|\lambda|)^d$. Then the spectral radius $\Lambda(A)$ of A satisfies $\Lambda(A) \geq |\lambda|(1 - 4\varepsilon)$.*

Proof Writing $A^k x - \lambda^k x = A^{k-1}(Ax - \lambda x) + \ldots + \lambda^{k-1}(Ax - \lambda x)$ and using that $\|A\| \leq 1$, we obtain for $k \leq d$

$$\|A^k x - \lambda^k x\| \leq (\varepsilon|\lambda|)^d (1 + |\lambda| + \ldots + |\lambda|^{k-1}) \leq (2\varepsilon|\lambda|)^d.$$

If $\chi_A(t) = t^d + a_{d-1}t^{d-1} + \ldots + a_0$ is the characteristic polynomial of A, then $\|\chi_A(A)x - \chi_A(\lambda)x\| \leq \sum |a_k| \|A^k x - \lambda^k x\|$, and $\chi_A(A) = 0$ by Cayley-Hamilton.

But $|a_{d-k}| \leq \binom{d}{k}$, and so $|\chi_A(\lambda)| \leq 2^d(2\varepsilon|\lambda|)^d$. To prove the claim, we may assume that $|\lambda| \geq \Lambda(A)$. Writing α_i for the roots of χ_A, the claim then follows from

$$||\lambda| - \Lambda(A)|^d \leq \prod_1^d ||\lambda| - |\alpha_i|| \leq |\chi_A(\lambda)|.$$

□

Proof of Theorem 3 We may put S in block triangular form. Since $\rho(S) = \max_i \rho(S_{ii})$, at least one of the irreducible diagonal blocks S_{ii} has $\rho(S_{ii}) = 1$. Hence, without loss of generality, we may assume that S is irreducible. Let v be a Barabanov norm for S as in Lemma 5. Pick a unit vector $x_0 \in \mathbb{C}^d$ and find recursively $s_1, s_2, \dots$ such that $x_n = s_n \cdot \dots \cdot s_1 x_0$ satisfies $v(x_n) = 1$ for all $n \geq 0$. Let $\delta = (\varepsilon/4)^d$. Note that the cardinality of a δ-separated set lying in the unit ball for v is at most $(1+\delta/2)^d/(\delta/2)^d = (1+\frac{2}{\delta})^d \leq n_0(d)\varepsilon^{-d^2}$, because the v-balls of radius $\frac{\delta}{2}$ centered at these points are disjoint and contained in the v-ball of radius $1+\frac{\delta}{2}$ around the origin. By pigeonhole, there are $0 \leq n < n'$ both smaller than $n_0(d)\varepsilon^{-d^2}$ such that $v(x_n - x_{n'}) < \delta$. In other words, $v(Ax_n - x_n) < \delta$, where $A := s_{n'} \cdot \dots \cdot s_{n+1}$. By Lemma 8, it follows that $\Lambda(A) \geq 1 - 4\delta^{1/d} = 1 - \varepsilon$. □

4 Explicit Bounds for Bochi's Inequalities

In this section we prove Theorems 5 and 6. We begin by the Siegel-type lemma already mentioned.

Lemma 9 (Siegel-Type Lemma) *Let $\|\cdot\|$ be a norm on $\mathbb{C}^d$. Let $\varepsilon \in (0,1)$ and $T, n \in \mathbb{N}$ with $(1+T)^n > (1+2nT\varepsilon^{-1})^d$. Pick $x_1, \dots, x_n$ vectors in $\mathbb{C}^d$ with $\|x_i\| \leq 1$. Then there are integers $c_1, \dots, c_n$, not all zero, such that $|c_i| \leq T$ for all i and*

$$\|\sum_1^n c_i x_i\| \leq \varepsilon.$$

Proof Consider the sums $\sum_1^n d_i x_i$ for integers $d_i \in [0, T]$. They have norm at most Tn. If all $\frac{\varepsilon}{2}$-balls around them were disjoint, then the ball of radius $Tn+\frac{\varepsilon}{2}$ around the origin would contain at least $(1+T)^n$ disjoint balls of radius $\frac{\varepsilon}{2}$. Comparing volumes we would have $(1+T)^n \leq (Tn+\varepsilon/2)^d/(\varepsilon/2)^d$, contrary to our assumption. Hence, two of these balls, corresponding to $(d_i)_i$ and $(d_i')_i$, say, must intersect. Setting $c_i = d_i' - d_i$ we get what we want. □

Lemma 10 *Let $\varepsilon > 0$. Let $A \in M_d(\mathbb{C})$ such that $|\mathrm{Tr}(A^k)| \leq \varepsilon^k$ for $k = 1, \dots, d$. Then the spectral radius of A satisfies $\Lambda(A) \leq 2\varepsilon$.*

Proof Let $s_k = \lambda_1^k + \ldots + \lambda_d^k$, where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of A. The Newton relations read $s_k + a_{d-1}s_{k-1} + \ldots + a_{d-k+1}s_1 = -ka_{d-k}$, where $t^d + a_{d-1}t^{d-1} + \ldots + a_0$ is the characteristic polynomial χ_A of A. We deduce from them that $|a_{d-k}| \leq \varepsilon^k$ for each $k = 1, \ldots, d$. If λ is an eigenvalue of A, then $\chi_A(\lambda) = 0$ and thus

$$|\lambda|^d \leq \varepsilon|\lambda|^{d-1} + \ldots + \varepsilon^k|\lambda|^{d-k} + \ldots + \varepsilon^d.$$

Setting $x = \varepsilon/|\lambda|$, we obtain $1 \leq x + \ldots + x^d$. But this implies $x \geq 1/2$. □

Lemma 11 *Let $n \in \mathbb{N}$ and $S \subset M_d(\mathbb{C})$ be a bounded set such that $\varepsilon := \max_{k \leq nd} \Lambda(S^k)^{\frac{1}{k}} \leq 1$. Let Q be the complex convex hull of $S \cup \ldots \cup S^n$. Then $\Lambda(Q) \leq 2d\varepsilon$.*

Proof Note that $Conv(A)Conv(B) \subset Conv(AB)$ for any two sets $A, B \subset M_d(\mathbb{C})$. So if $a \in Q$, then a^k belongs to the convex hull of $\bigcup_{k \leq i \leq nk} S^i$. In particular

$$|\mathrm{Tr}(a^k)| \leq d\varepsilon^k \leq (d\varepsilon)^k$$

for each $k = 1, \ldots, d$. The conclusion now follows from Lemma 10. □

We now prove Theorem 5. Rescaling and triangularizing S if necessary, we may assume without loss of generality that $\rho(S) = 1$ and that S is irreducible. As in the proof of Theorem 3, take a Barabanov norm $\| \cdot \|$ for S. Pick a unit vector $x \in \mathbb{C}^d$ and find $s_1, \ldots, s_n, \ldots$ in S such that $\|x_n\| = 1$ for all n, where $x_n := s_n \cdot \ldots \cdot s_1 x$. For T and $\varepsilon > 0$ as in Lemma 9, we obtain integers c_i not all zero such that $|c_i| \leq T$ and $\| \sum_1^n c_i x_i \| \leq \varepsilon$. Let i_0 be the smallest index i with $c_i \neq 0$ and set $y = x_{i_0}$. Hence, we may write:

$$\|c_{i_0} y + \sum_{i > i_0} c_i s_i \cdot \ldots \cdot s_{i_0+1} y\| \leq \varepsilon.$$

In other words:

$$\|\lambda y - Ay\| \leq \frac{\varepsilon}{N}, \tag{10}$$

where $A := \frac{1}{N} \sum_{i > i_0} -c_i s_i \cdot \ldots \cdot s_{i_0+1}$, $\lambda = \frac{c_{i_0}}{N}$ and $N := \sum_{i > i_0} |c_i|$. Note that $N \neq 0$, because $\varepsilon < 1$ and $\|x_i\| = 1$ for all i. Note further that $\|A\| \leq 1$ because $\|s\| \leq 1$ for all $s \in S$. And that $|\lambda| \geq \frac{1}{N} \geq \frac{1}{Tn}$, while $|\lambda| \leq \|Ay\| + \frac{\varepsilon}{N} \leq 1 + \frac{\varepsilon}{N} \leq 2$.

We can then apply Lemma 8 to A and λ and get

$$\Lambda(A) \geq |\lambda| - 4(\frac{\varepsilon}{N})^{\frac{1}{d}} \geq \frac{1}{2N} \geq \frac{1}{2nT}.$$

provided $4(\frac{\varepsilon}{N})^{\frac{1}{d}} \leq 1/2N$. The conditions for Lemma 8 require that $|\lambda| \leq 2$, while those for Lemma 9 require $(1+T)^n > (1+2nT\varepsilon^{-1})^d$. These conditions will be fulfilled if we set $T = 32d^2$, $n = 2d^2$, and $\varepsilon^{-1} = 8^d(nT)^{d-1}$. We conclude that

$$\Lambda(A) \geq \frac{1}{2^7 d^4}.$$

However, A belongs to the convex hull of $S \cup \ldots \cup S^n$. Therefore Lemma 11 implies that

$$\max_{k \leq nd} \Lambda(S^k)^{\frac{1}{k}} \geq \frac{1}{2^8 d^5}.$$

This yields the first inequality in Theorem 5. The second follows by applying the first to S^m for $m = \lceil 8 + 5\log_2 d \rceil$.

Proof of Theorem 6 This is very similar. Suppose $\|S\| = 1$ and let $\delta = \|S^{n_1}\|$. Pick a unit vector x and $s_1, \ldots, s_{n_1}$ such that $\|s_{n_1} \cdot \ldots \cdot s_1 x\| = \delta$. Arguing as in the above proof of Theorem 5, we get a y with $\|y\| \geq \delta$ such that (10) holds. Lemma 8 gives $\Lambda(A) \geq \frac{1}{2n_1 T}$ if ε is chosen so that $4(\varepsilon\delta^{-1}/n_1 T)^{1/d} = 1/2n_1 T$. Then setting $n_1 = 2d^2$, $n_1 T = M\delta^{-1}$, we see that the condition for Lemma 9 is fulfilled if $M \geq 2^6 d^4$. But $\rho(S) \geq \Lambda(A) \geq \delta/2M$, proving the claim. □

5 Ultrametric Complete Valued Fields

In this section we consider the analogue of the above for an algebraically closed complete and non-Archimedean valued field K and prove Theorem 7.

Let $\mathcal{O} := \{x \in K, |x| \leq 1\}$ be the ring of integers, $\mathfrak{m} := \{x \in K, |x| < 1\}$ its maximal ideal, and $\overline{k} = \mathcal{O}/\mathfrak{m}$ the residue field. Recall that the value group of K is dense in $\mathbb{R}_{>0}$ since K is algebraically closed. By an ultrametric norm on K^d, we mean a map $\|\cdot\| : K \to \mathbb{R}_{\geq 0}$ such that $\|\lambda x\| = |\lambda|\|x\|$, $\|x+y\| \leq \max\{\|x\|, \|y\|\}$, and $\|x\| = 0$ if and only if $x = 0$, for all $x, y \in K^d$, $\lambda \in K$.

An *orthogonal basis* for an ultrametric norm is a basis $(e_i)_1^d$ of K^d such that $\|x\| = \max\{c_i|x_i|\}$ for some positive reals c_i, if $x = x_1e_1 + \ldots + x_de_d$. We say that it is orthonormal if $c_i = 1$ for all i. If K is locally compact, or just spherically complete [6, 2.4.4], all ultrametric norms admit an orthogonal basis, but in general we only have:

Lemma 12 *Let v and w be two ultrametric norms on K^d and $\alpha > 1$ a real. Then there is $g \in \mathrm{GL}_d(K)$ such that $w(x) \leq v(gx) \leq \alpha w(x)$ for all $x \in K^d$.*

Proof This is well-known and follows from the existence [6, 2.6.2 Prop. 3] of almost orthogonal bases for ultrametric norms on K^d and the density in $\mathbb{R}^+$ of the value group of K. □

We begin by pointing out that the Rota-Strang observation, Lemma 2, and its proof remain valid in the ultrametric setting. Combined with Lemma 12, this yields the right-hand side of (5). It turns out that the infimum in (7) is realized under some mild conditions (milder than in the complex case):

Lemma 13 *Suppose that the value group of K is all of $\mathbb{R}_{>0}$ and $S \subset M_d(K)$ is a bounded set. If $\rho(S) = 0$, then $S^d = 0$, while if $\rho(S) > 0$, then there is an ultrametric norm $\|\cdot\|$ on K^d with $\|S\| = \rho(S)$.*

Proof The first assertion follows from the same argument as in Lemma 4. If $\rho(S) > 0$, we may rescale and assume that $\rho(S) = 1$, because we can pick $\lambda \in K$ with $|\lambda| = \rho(S)$. If S is irreducible, then the proof of Lemma 4 works verbatim and yields the desired norm. In general, we may choose a basis of K^d for which S is in block triangular form with irreducible blocks and define the norm $\|x\| = \max_i \|x_i\|_i$, where $\|\cdot\|_i$ is a norm on the i-th block with $\rho(S_{ii}) = \|S_{ii}\|_i$, provided $S_{ii} \neq 0$ and arbitrary otherwise. We may further conjugate S by a block diagonal matrix g, where the i-th block is the scalar matrix t^i for some $t \in K$ with $0 \neq |t| < 1/\|S\|$. Then, because of the ultrametric property, $\|gSg^{-1}\| \leq 1$. Thus $\|g \cdot g^{-1}\|$ is the desired norm. □

We now proceed to the proof of Theorem 7. It follows the same idea as in the proof of Claim 1 from the introduction, but we will need to palliate the lack of compactness and the fact that the value group may not be all of $\mathbb{R}_{>0}$ by the use of an ultrapower construction. The gist of the proof is in the following lemma:

Lemma 14 *Suppose that $\|\cdot\|$ is an ultrametric norm admitting an orthonormal basis. If $S \subset M_d(K)$ is such that $\|S\| = \rho(S) = 1$, then*

$$\max_{k \leq \ell(d)} \Lambda(S^k) = 1.$$

Proof Let $(e_i)_1^d$ be the orthonormal basis, i.e., $\|x\| = \max_1^d |x_i|$ if $x = x_1e_1 + \ldots + x_de_d$. In this basis, $S \subset M_d(\mathscr{O})$. Consider the convex hull Q of $S, \ldots, S^{\ell(d)}$, that is, the $\mathscr{O}$-module they span. If $\Lambda(S^k) < 1$ for each $k = 1, \ldots, \ell(d)$, the characteristic polynomial of a matrix in S^k will be t^d modulo $\mathfrak{m}$. So the image of Q modulo $\mathfrak{m}$ in $M_d(\overline{k})$ will consist of nilpotent matrices, and it will be a subalgebra of $M_d(\overline{k})$ by definition of $\ell(d)$. By Wedderburn's theorem, it will therefore be a nilpotent algebra, and we conclude that $S^d \subset M_d(\mathfrak{m})$. In particular, $\|S^d\| < 1$, contradicting our assumption that $\rho(S) = 1$. □

Proof of Theorem 7 Suppose first that the value group of K is all of $\mathbb{R}_{>0}$ and that all ultrametric norms on K^d admit an orthonormal basis. Then the theorem follows from the combination of the two previous lemmas by renormalizing S. So to handle the general case, it is enough to show that K can be embedded in another such field with the above properties. Any ultralimit $\mathbf{K} = \ell_\infty(K)/\equiv$ of K with respect to some non-principal ultrafilter $\mathscr{U}$ on $\mathbb{N}$ will do. Here $\ell_\infty(K)$ is the space of bounded sequences in K and $(x_n)_n \equiv (y_n)_n$ if and only if $\lim_{\mathscr{U}} |x_n - y_n| = 0$. Indeed, by the countable saturation property of ultraproducts (e.g., [9, 2.25]), an ultralimit $\mathbf{K}$ will again be complete and algebraically closed, its value group will be $\mathbb{R}_{>0}$, and, because of Lemma 12, all norms will admit an orthonormal basis. This shows Theorem 7 in full. □

Proof of Theorem 8 This follows from Bochi's original argument [5, Theorem A] suitably adapted to the ultrametric setting. First, up to passing to a suitable field extension as in the proof of Theorem 7, we may assume that all norms admit an orthonormal basis. Pick one so that $\|x\|_0 = \max_i |x_i|$. Then observe the following: for every invertible diagonal matrix a, we have:

$$\|aS^d a^{-1}\|_0 \leq \|S\|_0 \cdot \|aSa^{-1}\|_0^{d-1}. \tag{11}$$

Indeed every matrix entry of an element of $aS^d a^{-1}$ is a sum of monomials of the form $a_{i_1} s^{(1)}_{i_1 i_2} \cdot \ldots \cdot s^{(d)}_{i_d i_{d+1}} a^{-1}_{i_{d+1}}$ for matrices $s^{(i)} \in S$. We may write it as $a_{i_1} s^{(1)}_{i_1 i_2} a^{-1}_{i_2} a_{i_2} \cdot \ldots \cdot a^{-1}_{i_d} a_{i_d} s^{(d)}_{i_d i_{d+1}} a^{-1}_{i_{d+1}}$, a product of d factors each bounded by $\|aSa^{-1}\|_0$. However at least one of the d factors is bounded by $\|S\|_0$, because for at least one $j \in [1, d]$, $|a^{-1}_{i_j} a_{i_{j+1}}| \leq 1$, proving (11). Now we claim that (11) holds for an arbitrary matrix $a \in \mathrm{GL}_d(K)$, no longer assumed diagonal. Indeed this follows from the fact that $\|\cdot\|_0$ is invariant under $\mathrm{GL}_d(\mathcal{O})$ and that any matrix in $\mathrm{GL}_d(K)$ can be written as a product $k_1 a k_2$, with k_1, k_2 in $\mathrm{GL}_d(\mathcal{O})$ and a diagonal, as can be easily checked using operations on rows and columns as in Gaussian elimination. Finally, the theorem is proved taking the infimum in overall $a \in \mathrm{GL}_d(K)$ in view of (7). □

Finally we record one last observation.

Proposition 1 *If $S \subset M_d(K)$ is bounded and irreducible, then it admits a Barabanov norm, i.e., an ultrametric norm $\|\cdot\|$ such that* $\max_{s \in S} \|sx\| = \rho(S)\|x\|$ *for all $x \in K^d$.*

Proof By the proof of Theorem 7, we may embed K into a complete algebraically closed valued field $\mathbf{K}$ whose value group is all of $\mathbb{R}_{>0}$. Pick $\lambda \in \mathbf{K}$ with $|\lambda| = \rho(S)$. Then Lemma 13 implies that $\lambda \neq 0$ and that $\mathbf{S} := S/\lambda \subset M_d(\mathbf{K})$ is product bounded and admits an extremal norm $\|\cdot\|$. We may define the Barabanov norm of S by the same formula (8) applied to $\mathbf{S}$. Irreducibility forces this semi-norm to be a genuine norm. □

Acknowledgments I am grateful to Cagri Sert for many interesting discussions about the joint spectral radius and for his remarks on this article. I am also indebted to Eduardo Oregon-Reyes for his comments and to Ian Morris for bringing my attention to Bochi's other inequality (3) and to related references. Finally I wish to thank the anonymous referee for their comments and helpful suggestions.

References

1. Ball, K.: An elementary introduction to modern convex geometry. In: Flavors of Geometry, vol. 31 of Math. Sci. Res. Inst. Publ., pp. 1–58. Cambridge Univ. Press, Cambridge (1997)
2. Barabanov, N.E.: On the Lyapunov exponent of discrete inclusions. I–III. Avtomat. i Telemekh. (2), 40–46 (1988)
3. Berger, M.A., Wang, Y.: Bounded semigroups of matrices. Linear Algebra Appl. **166**, 21–27 (1992)
4. Blondel, V.D., Nesterov, Y.: Computationally efficient approximations of the joint spectral radius. SIAM J. Matrix Anal. Appl. **27**(1), 256–272 (2005)
5. Bochi, J.: Inequalities for numerical invariants of sets of matrices. Linear Algebra Appl. **368**, 71–81 (2003)
6. Bosch, S., Güntzer, U., Remmert, R.: Non-Archimedean analysis, vol.261. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Berlin (1984). A systematic approach to rigid analytic geometry
7. Bousch, T., Mairesse, J.: Asymptotic height optimization for topical IFS, Tetris heaps, and the finiteness conjecture. J. Am. Math. Soc. **15**(1), 77–111 (2002)
8. Breuillard, E.: A height gap theorem for finite subsets of $\mathrm{GL}_d(\overline{\mathbb{Q}})$ and nonamenable subgroups. Ann. Math. (2) **174**(2), 1057–1110 (2011)
9. Chatzidakis, Z.: Motivic integration and its interactions with model theory and non-Archimedean geometry. Vol.I, vol.383. London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge (2011)
10. Daubechies, I., Lagarias, J.C.: Sets of matrices all infinite products of which converge. Linear Algebra Appl. **161**, 227–263 (1992)
11. Elsner, L.: The generalized spectral-radius theorem: an analytic-geometric proof. In: Proceedings of the Workshop “Nonnegative Matrices, Applications and Generalizations” and the Eighth Haifa Matrix Theory Conference (Haifa, 1993), vol. 220, pp. 151–159 (1995)
12. Goldman, O., Iwahori, N.: The space of $\mathfrak{p}$-adic norms. Acta Math. **109**, 137–177 (1963)
13. Hare, K.G., Morris, I.D., Sidorov, N., Theys, J.: An explicit counterexample to the Lagarias-Wang finiteness conjecture. Adv. Math. **226**(6), 4667–4701 (2011)
14. Jenkinson, O., Pollicott, M.: Joint spectral radius, Sturmian measures and the finiteness conjecture. Ergodic Theory Dynam. Syst. **38**(8), 3062–3100 (2018)
15. Jungers, R.: The joint spectral radius, vol. 385. Lecture Notes in Control and Information Sciences. Springer, Berlin (2009). Theory and Applications
16. Kozyakin, V.: On accuracy of approximation of the spectral radius by the Gelfand formula. Linear Algebra Appl. **431**(11), 2134–2141 (2009)
17. Kozyakin, V.: Iterative building of Barabanov norms and computation of the joint spectral radius for matrix sets. Discrete Contin. Dyn. Syst. Ser. B **14**(1), 143–158 (2010)
18. Lagarias, J.C., Wang, Y.: The finiteness conjecture for the generalized spectral radius of a set of matrices. Linear Algebra Appl. **214**, 17–42 (1995)
19. Morris, I.D.: Criteria for the stability of the finiteness property and for the uniqueness of Barabanov norms. Linear Algebra Appl. **433**(7), 1301–1311 (2010)
20. Morris, I.D.: A rapidly-converging lower bound for the joint spectral radius via multiplicative ergodic theory. Adv. Math. **225**(6), 3425–3445 (2010)
21. Morris, I.D., Sidorov, N.: On a devil’s staircase associated to the joint spectral radii of a family of pairs of matrices. J. Eur. Math. Soc. (JEMS) **15**(5), 1747–1782 (2013)
22. Oregón-Reyes, E.: Properties of sets of isometries of Gromov hyperbolic spaces. Groups Geom. Dyn. **12**(3), 889–910 (2018)
23. Oregón-Reyes, E.: A new inequality about matrix products and a Berger-Wang formula. J. Éc. polytech. Math. **7**, 185–200 (2020)
24. Pappacena, C.J.: An upper bound for the length of a finite-dimensional algebra. J. Algebra **197**(2), 535–545 (1997)
25. Protasov, V.Yu.: A generalized joint spectral radius. A geometric approach. Izv. Ross. Akad. Nauk Ser. Mat. **61**(5), 99–136 (1997)

26. Ribenboim, P.: The Theory of Classical Valuations. Springer Monographs in Mathematics. Springer, New York (1999)
27. Rota, G.-C., Strang, G.: A note on the joint spectral radius. Nederl. Akad. Wetensch. Proc. Ser. A 63 = Indag. Math. **22**, 379–381 (1960)
28. Shitov, Y.: An improved bound for the lengths of matrix algebras. Algebra Number Theory **13**(6), 1501–1507 (2019)
29. Wirth, F.: The generalized spectral radius and extremal norms. Linear Algebra Appl. **342**, 17–40 (2002)

The Failure of the Fractal Uncertainty Principle for the Walsh–Fourier Transform

Ciprian Demeter

To the memory of Jean Bourgain

Abstract We construct δ-regular sets with $\delta \geq \frac{1}{2}$ for which the analog of the Bourgain–Dyatlov fractal uncertainty principle fails for the Walsh–Fourier transform.

1 The Fractal Uncertainty Principle for the Fourier Transform

This note explores the so-called fractal uncertainty principle, one of the last significant results of Jean Bourgain. The principle is a fundamental result in Fourier analysis with far-reaching consequences in the spectral theory of hyperbolic surfaces.

Definition 1.1 Let $X \subset \mathbb{R}$ be a nonempty closed set. Consider the constants $\delta \in [0, 1)$, $C_R \geq 1$, and $0 \leq \alpha_0 \leq \alpha_1 \leq \infty$. We say that X is δ-regular with constant C_R on scales α_0 to α_1 if there is a Borel measure μ_X supported on X such that

- For each interval I of size $|I| \in [\alpha_0, \alpha_1]$, we have $\mu_X(I) \leq C_R|I|^\delta$
- If additionally I is centered at a point in X, then $\mu_X(I) \geq C_R^{-1}|I|^\delta$.

We will denote by $|X|$ the Lebesgue measure of X.

Examples of regular sets will be discussed in Sect. 3. At this point, we only mention that δ-regular sets need to have small Lebesgue measure, more precisely

The author is partially supported by the Research NSF grant DMS-1800305.

C. Demeter (✉)
Department of Mathematics, Indiana University, Bloomington, IN, USA
e-mail: demeterc@indiana.edu

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_2

(see Lemma 2.9 in [2])

$$|X| \leq 24C_R^2\alpha_1^\delta\alpha_0^{1-\delta}. \tag{1}$$

The following fractal uncertainty principle for the Fourier transform

$$\widehat{f}(\xi) = \int_{\mathbb{R}} f(x)e^{-2\pi i x\xi}dx$$

was proved in [2]. It refines earlier versions due to Dyatlov–Zahl [4] and Bourgain–Dyatlov [1].

Theorem 1.2 *Let $\delta \in [0, 1)$, $C_R \geq 1$, and $N \geq 1$. Assume that*

- *$X \subset [0, 1]$ is δ-regular with constant C_R on scales $\frac{1}{N}$ to 1*
- *$Y \subset [0, N]$ is δ-regular with constant C_R on scales 1 to N*

Then there exist constants $\beta > 0$ and C, both depending only on δ and C_R, such that for each $f \in L^2(\mathbb{R})$ with Fourier transform supported on Y, we have

$$\|f\|_{L^2(X)} \leq CN^{-\beta}\|f\|_{L^2(\mathbb{R})}. \tag{2}$$

When $\delta < \frac{1}{2}$, this theorem has an easy proof that also provides an explicit value for β. For reader's convenience, we recall this argument below. If $\widehat{f}$ is supported on Y, we have

$$\begin{aligned}
\|f\|_{L^2(X)} &\leq |X|^{1/2}\|f\|_{L^\infty(\mathbb{R})}\\
&\leq |X|^{1/2}\|\widehat{f}\|_{L^1(\mathbb{R})}\\
&= |X|^{1/2}\|\widehat{f}\|_{L^1(Y)}\\
&\leq |X|^{1/2}|Y|^{1/2}\|\widehat{f}\|_{L^2(\mathbb{R})}\\
&= |X|^{1/2}|Y|^{1/2}\|f\|_{L^2(\mathbb{R})}.
\end{aligned}$$

If X and Y are as in the theorem, then (1) implies that $|X|^{1/2}|Y|^{1/2} \leq CN^{-\beta}$, $\beta = \frac{1}{2} - \delta$.

On the other hand, the proof from [2] in the case $\delta \geq \frac{1}{2}$ is very involved. At its heart, it relies both on the multiscale structure of regular sets and on the following unique continuation result (Lemma 3.2 in [2]).

Lemma 1.3 *Let $\mathcal{I}$ be a non-overlapping collection of intervals of size 1 and let $c_0 > 0$. For each $I \in \mathcal{I}$, let $I'' \subset I$ be an interval of size c_0. Then there exists a constant C depending only on c_0 such that for all $r \in (0, 1)$, $0 < \kappa \leq e^{-C/r}$, and $f \in L^2(\mathbb{R})$ with $\widehat{f}$ compactly supported, we have*

$$\sum_{I\in\mathcal{I}} \|f\|^2_{L^2(I)} \le \frac{C}{r}\Big(\sum_{I\in\mathcal{I}} \|f\|^2_{L^2(I'')}\Big)^{\kappa} \|e^{2\pi r|\xi|}\widehat{f}(\xi)\|^{2(1-\kappa)}_{L^2(\mathbb{R})}.$$

In the next section, we recall the details about the Walsh transform, a closely related, though technically simpler analog of the Fourier transform. We will construct sets X and Y as in Theorem 1.2 with regularity $\delta \ge \frac{1}{2}$, such that the fractal uncertainty principle fails when the Walsh transform replaces the Fourier transform. This fundamental difference between the behavior of the two transforms explains why the proof in [2] is so complicated. The argument in [2] must necessarily rely not just on the fine structure of the regular sets but also on the stronger form of the uncertainty principle that governs the Fourier world. This has to do with the fact that there is no (nontrivial) compactly supported function whose Fourier transform is also compactly supported. Lemma 1.3 is a manifestation of this principle.

In the next section, we will see that there are compactly supported L^2 functions whose Walsh transforms are also compactly supported. This easily shows the failure of Lemma 1.3, and ultimately of Theorem 1.2, in the Walsh framework. Our main result, Theorem 3.1, is proved in the last section.

2 The Walsh Transform

For more details on the material in this section, the reader may consult the original paper of Walsh [11], or the modern reference [8].

Let $\mathbb{Z}_2 = \{0, 1\}$ with addition modulo 2 and Haar measure splitting the mass evenly between 0 and 1. We consider the infinite product group $G = \prod_1^\infty \mathbb{Z}_2$ equipped with the product Haar measure. This is sometimes referred to as the *Cantor group*.

Let $\mathcal{D} = \{j2^{-i} : 0 \le j \le 2^i\}$ be the dyadic numbers in [0, 1]. They have zero Lebesgue measure. The map

$$\Phi : G \to [0, 1], \quad \Phi(a_{-1}, a_{-2}, \ldots) = \sum_{k\le -1} a_k 2^k$$

is almost bijective—if $x \in [0, 1] \setminus \mathcal{D}$, $\Phi^{-1}(\{x\})$ consists of one point—measurable and maps the Haar measure on G to the Lebesgue measure $|\cdot|$ on [0,1]. This suggests a natural way to identify G with $([0, 1], \oplus, |\cdot|)$, where $\oplus$ is defined as follows. Given $x, y \in [0, 1] \setminus \mathcal{D}$, $x = \sum_{k\le -1} x_k 2^k$, $y = \sum_{k\le -1} y_k 2^k$, we write

$$x \oplus y = \sum_{k\le -1} c_k 2^k, \quad c_k = x_k + y_k \pmod 2.$$

See Sec 2.2 in [6] for details.

The characters on G are the so-called Walsh functions. For $n \geq 0$, the $n-$th Walsh function $W_n : [0, 1) \to \{-1, 1\}$ is defined recursively by the formula

$$W_0 = 1_{[0,1)}$$

$$W_{2n}(x) = W_n(2x) + W_n(2x - 1)$$

$$W_{2n+1}(x) = W_n(2x) - W_n(2x - 1).$$

In particular,

$$W_1(x) = \begin{cases} 1, & 0 \leq x < \frac{1}{2} \\ -1, & \frac{1}{2} \leq x < 1 \end{cases},$$

$$W_2(x) = \begin{cases} 1, & x \in [0, \frac{1}{4}) \cup [\frac{1}{2}, \frac{3}{4}) \\ -1, & x \in [\frac{1}{4}, \frac{1}{2}) \cup [\frac{3}{4}, 1) \end{cases}$$

$$W_3(x) = \begin{cases} 1, & x \in [0, \frac{1}{4}) \cup [\frac{3}{4}, 1), \\ -1, & x \in [\frac{1}{4}, \frac{3}{4}) \end{cases}.$$

In many ways, the functions W_n resemble the (Fourier) system of exponentials $e^{2\pi inx}$. For example, the functions $(W_n)_{n\geq 0}$ form an orthonormal basis for $L^2([0, 1])$. See Sec 4.1 [6] for more details.

The Walsh–Fourier coefficients of a function $f : [0, 1] \to \mathbb{C}$ are given by

$$\mathcal{F}_W f(n) = \int f(x)W_n(x)dx, \quad n \geq 0.$$

To get a greater perspective on the role of the Walsh system and its closeness to the Fourier system of exponentials, we introduce a new operation. For $x, y \in [0, \infty)$ having unique representations (i.e., for Lebesgue almost all pairs (x, y))

$$x = \sum_{k=-\infty}^{\infty} x_k 2^k, \quad y = \sum_{k=-\infty}^{\infty} y_k 2^k,$$

we define

$$x \otimes y := \sum_{k=-\infty}^{\infty} c_k 2^k$$

where

$$c_k = \sum_{j \in \mathbb{Z}} x_j y_{k-j} \pmod 2.$$

We note that this sum is always finite. From now on, we will implicitly ignore the zero measure dyadic points.

Define the function $e_W : [0, \infty) \to \{-1, 1\}$ such that $e_W(x) = 1$ when $x_{-1} = 0$ and $e_W(x) = -1$ when $x_{-1} = 1$. This 1-periodic function is the Walsh analog of $e^{2\pi i x}$. It is easy to check that

$$W_n(x) = e_W(x \otimes n) 1_{[0,1]}(x). \tag{3}$$

We may introduce the Walsh (also called Walsh–Fourier) transform of a compactly supported function $f : [0, \infty) \to \mathbb{C}$ to be the function

$$\mathcal{F}_W f : [0, \infty) \to \mathbb{C}, \quad \mathcal{F}_W f(y) := \int_{[0,\infty)} e_W(x \otimes y) f(x) dx.$$

The Walsh–Fourier inversion formula takes the form $\mathcal{F}_W \circ \mathcal{F}_W = id$.

It is worth noting that

$$e_W(x \otimes y) = e_W(x \otimes z)$$

whenever $x \in [0, 1)$ and $n \le y, z < n+1$. Consequently, if f is supported on $[0, 1]$, then $\mathcal{F}_W f$ is constant on intervals $[n, n+1)$. This explains why for such functions the Walsh–Fourier coefficients completely characterize the function f.

While the Walsh transform behaves very similar to the Fourier transform, it has one notable feature that makes it easier to work with. This has to do with the fact that there are (plenty of) compactly supported functions whose Walsh transforms are also compactly supported. A quick computation shows that for each dyadic interval $I = [l2^k, (l+1)2^k)$, we have

$$\mathcal{F}_W 1_I(y) = |I| 1_{[0,|I|^{-1}]}(y) e(x_I \otimes y), \tag{4}$$

where x_I is an arbitrary element of I. Because of this feature, typically the results that hold in the Fourier case are expected to also hold in the Walsh setting, with the argument in the latter case being cleaner, less technical. The approach of first proving results in the Walsh setting and then "transferring" them to the Fourier world was successfully employed in the time-frequency analysis of modulation invariant operators, starting with [10]. The interested reader may consult the survey paper [3], which explores a few different arguments for the Walsh analog of Carleson's Theorem and contains some relevant references.

In this paper we present an example that goes against the aforementioned philosophy. We show that a fundamental result that holds for the Fourier transform is in fact false for the Walsh transform.

3 The Main Result

The "textbook" example of regular sets can be constructed as follows. Fix integers $0 < M < L$. Let $\mathcal{S}$ be a collection of subsets S of $\{0, 1, \dots, L-1\}$ with cardinality M. We create a collection of nested sets $X_1, X_2, \dots$ as follows. Pick $S_1 \in \mathcal{S}$ and let

$$A_1 = L^{-1}S_1, \quad X_1 = A_1 + [0, L^{-1}].$$

Next, for each $a \in A_1$, choose some $S_{2,a} \in \mathcal{S}$ and define

$$A_{2,a} = a + L^{-2}S_{2,a}, \quad A_2 = \cup_{a \in A_1} A_{2,a}, \quad X_2 = A_2 + [0, L^{-2}].$$

The rest of the construction is recursive. Assume we have constructed A_j and X_j for $1 \le j \le n-1$. For each $a \in A_{n-1}$, choose some $S_{n,a} \in \mathcal{S}$ and define

$$A_{n,a} = a + L^{-n}S_{n,a}, \quad A_n = \cup_{a \in A_{n-1}} A_{n,a}, \quad X_n = A_n + [0, L^{-n}].$$

Note that $X_n \subset [0,1]$ consists of M^n intervals $I \in \mathcal{I}_{X_n}$ of length L^{-n}. Also, X_n is $\frac{\log M}{\log L}$−regular on scales $\frac{1}{L^n}$ to 1, with constant C_n satisfying the uniform bound $C_n \le C(M, L)$, where $C(M, L)$ depends only on M, L. The reader may check that Definition 1.1 is satisfied with the measure μ_{X_n} given by $\mu_{X_n}(I) = \frac{1}{M^n}$, for each $I \in \mathcal{I}_{X_n}$.

We specialize this construction as follows. Fix the positive integers m_1 and $m_2 \ge m_1$. We consider a set as above with $M = 2^{m_2}$ and $L = 2^{m_1+m_2}$. The collection $\mathcal{S}$ will consist of only the set $S = \{k2^{m_1},\ 0 \le k \le 2^{m_2} - 1\}$.

More precisely, define

$$A_n = \{\sum_{i=1}^{n} \frac{k_{n-i+1}2^{m_1}}{L^i}\ :\ 0 \le k_1, \dots, k_n \le 2^{m_2} - 1\}$$

and

$$X_n = A_n + [0, L^{-n}]. \tag{5}$$

Then $X_n \subset [0,1]$ is $\frac{m_2}{m_1+m_2}$−regular on scales L^{-n} to 1, with constant C_n uniformly bounded in n.

Define also the dilate

$$Y_n = L^n X_n = \{L^n x \,:\, x \in X_n\}.$$

Note that Y_n is the union of intervals of length 1 and $Y_n \subset [0, L^n]$. It is $\frac{m_2}{m_1+m_2}$−regular on scales 1 to L^n, with the same constant C_n as X_n.

Theorem 3.1 *The (real) vector space $\mathcal{V}_{X_n,Y_n}$ of all L^2 functions*

$$f : [0, 1] \to \mathbb{R}, \quad \operatorname{supp} f \subset X_n, \quad \operatorname{supp} \mathcal{F}_W f \subset Y_n$$

has dimension at least $2^{n(m_2-m_1)}$. In particular, for each $n \geq 1$, there is a function f_n (other than the zero function) with $\mathcal{F}_W f_n$ supported on Y_n such that

$$\|f_n\|_{L^2(X_n)} = \|f_n\|_{L^2([0,1])}.$$

Fixing m_1, m_2 and letting $n \to \infty$ shows that the Walsh analog of (2) fails to hold for any $\beta > 0$, when $\delta \geq \frac{1}{2}$.

We remark that the restriction $m_2 \geq m_1$ is needed in Theorem 3.1, as it is equivalent with the lower bound $\delta \geq \frac{1}{2}$ for the regularity of X_n, Y_n. When $\delta < \frac{1}{2}$, Theorem 1.2 remains true in the Walsh framework, and the argument from the first section for the Fourier case translates to the Walsh case, too.

4 Proofs

We start by proving a sequence of lemmas.

Lemma 4.1 *For $x, y \in [0, \infty)$ and $l \in \mathbb{Z}$, we have*

$$(2^l x) \otimes y = x \otimes (2^l y).$$

Proof If

$$x = \sum_{k\in\mathbb{Z}} x_k 2^k, \quad y = \sum_{k\in\mathbb{Z}} y_k 2^k$$

then

$$2^l x = \sum_{k\in\mathbb{Z}} x_{k-l} 2^k, \quad 2^l y = \sum_{k\in\mathbb{Z}} y_{k-l} 2^k$$

and

$$((2^l x)\otimes y)_k=\sum_{j\in\mathbb{Z}}(2^l x)_j y_{k-j}=\sum_{j\in\mathbb{Z}}x_{j-l}y_{k-j}=\sum_{j\in\mathbb{Z}}x_j y_{k-j-l}=(x\otimes(2^l y))_k.$$

□

Combining this lemma with (3) and (4) reveals that if $I=[\frac{k}{L^n},\frac{k+1}{L^n}]\subset[0,1]$ then

$$\mathcal{F}_W 1_I(y)=L^{-n}W_k(\frac{y}{L^n}). \tag{6}$$

Lemma 4.2 *The functions $W_0,W_1,\dots,W_{2^m-1}$ span the vector space*

$$\mathcal{C}_m=\{f:[0,1]\to\mathbb{R}:\ f \textit{ constant on dyadic intervals of length } 2^{-m}\}.$$

Proof An easy induction argument based on the recursive formula for W_n shows that $W_0,W_1,\dots,W_{2^m-1}\in\mathcal{C}_m$. The vector space $\mathcal{C}_m$ has dimension 2^m, and since $W_0,W_1,\dots,W_{2^m-1}$ are linearly independent (being orthogonal), they form a basis for this space.

□

The recursive definition of W_n also immediately implies the following periodicity property.

Lemma 4.3 *The function W_{k2^l} is 2^{-l} periodic, if k,l are positive integers. Moreover, when $x\in[0,2^{-l}]$, we have*

$$W_{k2^l}(x)=W_k(x2^l).$$

The combination of the last two lemmas yields the following result.

Proposition 4.4 *Consider the (real) vector space of all $F:[0,1]\to\mathbb{R}$ having the following two properties for some positive integers l,m*

(P1): *F is 2^{-l} periodic.*
(P2): *F is constant on dyadic intervals of length 2^{-l-m}.*

Then this vector space coincides with the span of the Walsh functions W_{k2^l}, for $0\le k\le 2^m-1$.

Let us recall that $L=2^{m_1+m_2}$. Rescaling the above result gives:

Corollary 4.5 *For $1\le i\le n$, consider the (real) vector space $\mathcal{V}_{i,n}$ of all functions $F_i:[0,L^n]\to\mathbb{R}$ such that*

(P1): *F_i is $\frac{L^i}{2^{m_1}}$ periodic.*
(P2): *F_i is constant on dyadic intervals of length L^{i-1}.*

Then $\mathcal{V}_{i,n}$ coincides with the span of the rescaled Walsh functions $W_{kL^{n-i}2^{m_1}}(\frac{y}{L^n})$, for $0\le k\le 2^{m_2}-1$.

Let $\mathcal{V}_{X_n}$ be the (real) vector space spanned by the Walsh transforms $\mathcal{F}_W 1_I$ of all intervals I of length L^{-n} in X_n. According to (5) and (6), this is the same as the vector space spanned by the rescaled Walsh functions

$$W_{\sum_{i=1}^n k_{n-i+1} 2^{m_1} L^{n-i}}(\frac{y}{L^n}) : \ 0 \leq k_1, \ldots, k_n \leq 2^{m_2} - 1. \tag{7}$$

Note that $\mathcal{V}_{X_n}$ is a proper subset of the family of Walsh transforms of functions supported on X_n. We are going to search for functions in $\mathcal{V}_{X_n}$ that are supported on Y_n.

Lemma 4.6 *For each $k, k' \in \mathbb{Z}$*

$$W_k W_{k'} = W_{k \oplus k'}.$$

Proof

$$W_k(x) W_{k'}(x) = (-1)^{(x \oplus k)_{-1}} (-1)^{(x \oplus k')_{-1}} = (-1)^{\sum_j x_j k_{-1-j}} (-1)^{\sum_j x_j k'_{-1-j}}$$

$$= (-1)^{\sum_j x_j (k \oplus k')_{-1-j}} = W_{k \oplus k'}(x).$$

□

Combining the last lemma and corollary, we get:

Proposition 4.7 *The space $\mathcal{V}_{X_n}$ coincides with the collection of arbitrary finite sums of products (over i) of functions $F_i \in \mathcal{V}_{i,n}$.*

Proof Note that since $k_{n-i+1} 2^{m_1} < L$, we have

$$\sum_{i=1}^n k_{n-i+1} 2^{m_1} L^{n-i} = \oplus_{i=1}^n k_{n-i+1} 2^{m_1} L^{n-i},$$

where the factors on the right-hand side are summed using $\oplus$ rather than $+$. Thus

$$W_{\sum_{i=1}^n k_{n-i+1} 2^{m_1} L^{n-i}}(\frac{y}{L^n}) = \prod_{i=1}^n W_{k_{n-i+1} 2^{m_1} L^{n-i}}(\frac{y}{L^n}).$$

□

Lemma 4.8 *Assume that $f_1, \ldots, f_N : \mathbb{R} \to \mathbb{R}$ are linearly independent, $g_1, \ldots, g_M : \mathbb{R} \to \mathbb{R}$ are linearly independent, and $\{f_n g_m : 1 \leq n \leq N, \ 1 \leq m \leq M\}$ are linearly independent. Let $\mathbb{V}_1$ be a linear subspace of* $\text{span}(f_1, \ldots, f_N)$

of dimension $d_1 \le N$*, and let* $\mathbb{V}_2$ *be a linear subspace of* $\operatorname{span}(g_1, \dots, g_M)$ *of dimension* $d_2 \le M$*.*

Then the linear space $\mathbb{V}$ *spanned by the functions* fg *with* $f \in \mathbb{V}_1$ *and* $g \in \mathbb{V}_2$ *has dimension* $d_1 d_2$*.*

Proof It is clear that $\dim(\mathbb{V}) \le \dim(\mathbb{V}_1)\dim(\mathbb{V}_2)$, so it remains to prove the reverse inequality. Assume that the functions $f^{(k)} = \sum_{n=1}^{N} a_{k,n} f_n$ with $1 \le k \le d_1$ form a basis for $\mathbb{V}_1$. It follows that the (d_1, N) matrix $A = (a_{k,n})$ contains a nonsingular (d_1, d_1) minor A'. Assume that the functions $g^{(l)} = \sum_{m=1}^{M} b_{l,m} g_m$ with $1 \le l \le d_2$ form a basis for $\mathbb{V}_2$. It follows that the (d_2, M) matrix $B = (b_{l,m})$ contains a nonsingular (d_2, d_2) minor B'.

We will show that the functions $f^{(k)}g^{(l)}$, $1 \le k \le d_1$, and $1 \le l \le d_2$ are linearly independent. We order the functions $f_n g_m$ using the lexicographic order for pairs (n, m), that is, $f_1g_1, \dots, f_1g_M, f_2g_1, \dots, f_2g_M, \dots, f_Ng_1, \dots, f_Ng_M$. We similarly order the functions $f^{(k)}g^{(l)}$ lexicographically with respect to the pairs (k, l). We construct the $(d_1 \times d_2, N \times M)$ matrix C as follows. The (i, j) entry is the coefficient of the ith function $f^{(k)}g^{(l)}$ with respect to the jth function $f_n g_m$. We denote this matrix by $A \otimes B$. One easy way to visualize it is to start with the matrix A and replace each entry $a_{k,n}$ with the matrix $a_{k,n}B$

$$C = \begin{bmatrix} a_{1,1}B & \dots & a_{1,N}B \\ \dots & \dots & \dots \\ a_{d_1,1}B & \dots & a_{d_1,N}B \end{bmatrix}.$$

We need to prove that C contains a nonsingular (d_1d_2, d_1d_2) minor. We claim that this minor is $C' = A' \otimes B'$, with the tensor operation described above. It is immediate that C' is a minor of C. Also, it is well known that

$$\det(C') = \det(A')^{d_1} \det(B')^{d_2}.$$

See, for example, [9]. In particular, $\det(C') \neq 0$, as desired.

□

We now prove Theorem 3.1 by induction. It suffices to show that the vector space of those F supported on Y_n, that are in the span of the rescaled Walsh functions in (7), has dimension at least $2^{n(m_2-m_1)}$.

Let us start with the base case $n = 1$. Using the characterization from Proposition 4.7, it suffices to prove that the vector space

$$\{F \in \mathcal{V}_{1,1} : \operatorname{supp} F \subset Y_1\}$$

has dimension $2^{m_2-m_1}$. The functions F in this space are 2^{m_2} periodic and constant on all intervals $[l, l+1)$. Since Y_1 contains exactly $2^{m_2-m_1}$ unit intervals in $[0, 2^{m_2}]$ (these are $I_k = [k2^{m_1}, k2^{m_1} + 1]$, $0 \le k \le 2^{m_2-m_1} - 1$), and since

$$Y_1 = \bigcup_{0 \le k' \le 2^{m_1}-1} ((I_0 \cup I_1 \cup \ldots \cup I_{2^{m_2-m_1}-1}) + k'2^{m_2}),$$

it is immediate that the values of F on $I_0, \ldots, I_{2^{m_2-m_1}-1}$ may be chosen arbitrarily. This verifies the base case of the induction. Note that by choosing the values of F to be 1 on all these intervals, we get the function $F = 1_{Y_1}$. This shows that $\mathcal{F}_W 1_{Y_1}$ is in $\mathcal{V}_{X_1,Y_1}$.

Next, let us prove the theorem for $n \ge 2$, assuming its validity for $n-1$. We write

$$Y_n = LY_{n-1} \cap Z_n, \quad Z_n = \bigcup_{k \le \frac{L^n}{2^{m_1}}} [k2^{m_1}, k2^{m_1}+1]. \tag{8}$$

Let $\mathcal{V}_{1,n}(Z_n)$ be the vector space of those $F_1 \in \mathcal{V}_{1,n}$ that are supported on Z_n. Note first that this has dimension $2^{m_2-m_1}$, since there are $2^{m_2-m_1}$ unit intervals in Z_n that lie in the periodicity interval $[0, 2^{m_2}]$ associated with $\mathcal{V}_{1,n}$. Pick $2^{m_2-m_1}$ functions H in the span of $W_{k_n L^{n-1} 2^{m_1}}$, with $0 \le k_n \le 2^{m_2}-1$, such that the rescaled functions $H(\frac{y}{L^n})$ form a basis for $\mathcal{V}_{1,n}(Z_n)$.

By the induction hypothesis, we may find a subset consisting of $2^{(n-1)(m_2-m_1)}$ linear independent functions G in the span of

$$W_{\sum_{i=1}^{n-1} k_{n-i} 2^{m_1} L^{n-1-i}} : \ 0 \le k_1, \ldots, k_{n-1} \le 2^{m_2}-1$$

such that each $G(\frac{y}{L^{n-1}})$ is supported on Y_{n-1}. So $G(\frac{y}{L^n})$ is supported on LY_{n-1}.

Because of Lemma 4.8, (8) and since

$$W_{\sum_{i=1}^{n} k_{n-i+1} 2^{m_1} L^{n-i}}(\frac{y}{L^n}) = W_{k_n L^{n-1} 2^{m_1}}(\frac{y}{L^n}) W_{\sum_{i=1}^{n-1} k_{n-i} 2^{m_1} L^{n-1-i}}(\frac{y}{L^n})$$

is supported on Y_n, we conclude that there are at least $2^{n(m_2-m_1)}$ linearly independent functions in $\mathcal{V}_{X_n}$ (recall that these are functions spanned by the functions in (7)) that are supported on Y_n. We thus have

$$\dim \mathcal{V}_{X_n,Y_n} \ge 2^{n(m_2-m_1)}.$$

Remark 4.9 The inductive argument from above shows that in fact $F = \mathcal{F}_W 1_{Y_n}$ is in $\mathcal{V}_{X_n,Y_n}$. Indeed, we observed that this is true for $n = 1$. The case $n > 1$ follows since

$$1_{Y_n} = 1_{LY_{n-1}} 1_{Z_n}$$

and since $1_{Z_n} \in \mathcal{V}_{1,n}(Z_n)$.

Acknowledgments I am grateful to Semyon Dyatlov for pointing out to me the following facts. The Walsh transform appears in applications to the toy model of open quantum baker's maps. In that case, the maps can be Walsh-quantized, and the spectral gap results do sometimes fail on these.

This was observed by Nonnenmacher and Zworski in [7] (Section 5, in particular Remark 5.2). In [5] (at the end of the Introduction), Dyatlov and Jin briefly interpreted this phenomenon as special instances of the failure of the fractal uncertainty principle for the Walsh–Fourier transform.

References

1. Bourgain, J., Dyatlov, S.: Fourier dimension and spectral gaps for hyperbolic surfaces. Geom. Funct. Anal. **27**(4), 744–771 (2017)
2. Bourgain, J., Dyatlov, S.: Spectral gaps without the pressure condition. Ann. Math. (2) **187**(3), 825–867 (2018)
3. Demeter, C.: A guide to Carleson's theorem. Rocky Mountain J. Math. **45**(1), 169–212 (2015)
4. Dyatlov, S., Zahl, J.: Spectral gaps, additive energy, and a fractal uncertainty principle. Geom. Funct. Anal. **26**(4), 1011–1094 (2016)
5. Dyatlov, S., Jin, L.: Resonances for open quantum maps and a fractal uncertainty principle. Comm. Math. Phys. **354**(1), 269–316 (2017)
6. Folland, G.B.: A course in abstract harmonic analysis. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL (1995). x+276 pp.
7. Nonnenmacher, S., Zworski, M.: Distribution of resonances for open quantum maps. Comm. Math. Phys. **269**(2), 311–365 (2007)
8. Schipp, F., Wade, W.R., Simon, P.: Walsh series. An introduction to dyadic harmonic analysis. With the Collaboration of J. Pal. Adam Hilger, Bristol (1990)
9. Silvester, J.: Determinants of Block Matrices. Mathematical Gazette, vol. 84(501), pp. 460–467. The Mathematical Association, (2000)
10. Thiele, C.: The quartile operator and pointwise convergence of Walsh series. Trans. Am. Math. Soc. **352**(12), 5745–5766 (2000)
11. Walsh, J.L.: A closed set of normal orthogonal functions. Am. J. Math. **45**, 5–24 (1923)

The Continuous Formulation of Shallow Neural Networks as Wasserstein-Type Gradient Flows

Xavier Fernández-Real and Alessio Figalli

Dedicated to the memory of Jean Bourgain

Abstract It has been recently observed that the training of a single hidden layer artificial neural network can be reinterpreted as a Wasserstein gradient flow for the weights for the error functional. In the limit, as the number of parameters tends to infinity, this gives rise to a family of parabolic equations. This survey aims to discuss this relation, focusing on the associated theoretical aspects appealing to the mathematical community and providing a list of interesting open problems.

1991 Mathematics Subject Classification 35Q49, 49Q22, 68T07

1 Introduction

The extensive and successful use of machine learning in recent years has been remarkable. However, from a mathematical viewpoint, an adequate theoretical understanding of its primary governing principles is still missing in many situations. Often, each problem needs to be studied individually, even within the application of the same technique, to obtain the desired visible result.

This work has received funding from the European Research Council (ERC) under the Grant Agreement No 721675. In addition, X. F. was supported by the SNF grant 200021_182565.

X. Fernández-Real (✉)
EPFL, Lausanne, Switzerland
e-mail: xavier.fernandez-real@epfl.ch

A. Figalli
Department of Mathematics, ETH Zürich, Zürich, Switzerland
e-mail: alessio.figalli@math.ethz.ch

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_3

Recently, a new continuous viewpoint of artificial neural networks has risen, intending to shine some light on this computing system's understanding. This theory has already been developed and shown important results and, roughly speaking, consists in viewing the gradient descent used to optimize parameters in a neural network as a gradient flow in the Wasserstein distance for their own empirical measure.

More precisely, training neural networks can be thought of as discretizations of a gradient flow with the appropriate metric and functional. This observation has opened the door to studying (at a theoretical level) the general convergence properties of such methods deducing properties of the corresponding continuous limit. Most of this study has been conducted from a numerical point of view, and there are still many open questions that are also interesting from a purely theoretical perspective.

In this framework, new mathematical problems and PDE systems have arisen, which have not yet been fully adopted by the mathematical community. This short survey aims to bridge this gap to present this fascinating problem in the gradient flows community's language.

We refer the interested reader to [4, 7, 9, 11, 15] and references therein for an in-depth introduction to the topic and also to [5, 6] for an approach more focused on dynamical systems and optimal control problems.

2 Shallow Neural Network and Gradient Flows

Given a domain $D \subset \mathbb{R}^n$, and a function $f : D \to \mathbb{R}$, training a single hidden layer artificial neural network (or shallow neural network) consists in approximating f with expressions of the form

$$f_N(x) = f_N(x, w_1, \dots, w_N, \theta_1, \dots, \theta_N) = \frac{1}{N}\sum_{i=1}^{N} w_i h(\theta_i, x), \tag{1}$$

where $w_i \in \mathbb{R}$ and $\theta_i \in \Theta \subset \mathbb{R}^d$ are parameters to be optimized (usually taken in pairs (w_i, θ_i)), and $h : \Theta \times D \to \mathbb{R}$ is called the activation function, which is nonlinear. Such construction of approximating functions is often graphically represented as seen in Fig. 1, and when the number of layer increases, the number of interconnections between the neurons increases as well, very loosely resembling a biological neural network.

In applications, it is usual to assume that

$$d = n + 1 \qquad \text{and} \qquad h(\theta, x) = \sigma(\theta' \cdot x + \theta^{(d)}), \tag{2}$$

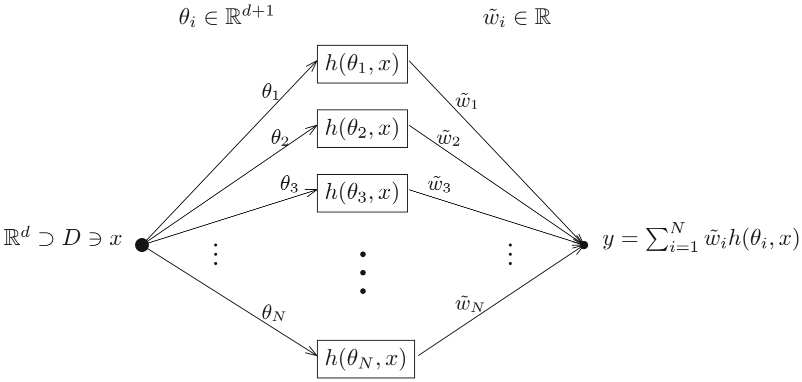


Fig. 1 Graphic representation of the approximating functions given by what is known as a single hidden layer *artificial neural network*. The variables are $\tilde{w}_i = \frac{1}{N} w_i$ according to the notation in (1)

where $\theta = (\theta', \theta^{(d)}) \in \mathbb{R}^n \times \mathbb{R}$, for a suitable nonlinearity σ.[1] Thus, neural networks try to approximate a given function with linear combinations of nonlinearities. However, for the sake of generality, here we will not consider a specific form of $h(\theta, x)$, and we focus instead on the general formulation where $h(\theta, x)$ can be arbitrary.

The number N of parameters $(w, \theta) \in \mathbb{R}^{d+1}$ used to in (1) corresponds to the number of neurons or hidden units. When training a neural network, one tries to minimize the expected error, sometimes called *risk* or *generalization error*, obtained from approximating f by f_N. To do so, one needs to define a loss function ℓ, that we consider to be

$$\ell\big(f, f_N\big) = \frac{1}{2}\int_D |f(x) - f_N(x)|^2 dx.$$

Let us denote by $\mathcal{H}_N$ the class of f_N that can be obtained as (1). Then, one wants to solve the minimization problem

$$\min_{f_N \in \mathcal{H}_N} \ell\big(f, f_N\big), \tag{3}$$

where ℓ is as above. The standard approach nowadays is to start from some choice of weights $\bar{w} = (\bar{w}_1, \ldots, \bar{w}_N)$ and $\bar{\theta} = (\bar{\theta}_1, \ldots, \bar{\theta}_N)$ and perform gradient descent on these parameters (w, θ) in order to (possibly) achieve the minimizer to (3):[2]

[1] A typical nonlinearity is the sigmoid function. Namely, if we denote $\sigma(t) = \frac{1}{1+e^{-t}}$, we consider $h(\theta, x) = \sigma(\theta \cdot x)$. However, nowadays, the most frequently used activation function in applications is not smooth nor bounded: the ReLU function $\sigma(t) = \max\{t, 0\}$.

[2] In fact, in reality, one uses stochastic gradient descent, by considering random samples $(x_i, f(x_i))$ of our data or training set.

$$\begin{cases} \frac{d}{dt}(w(t),\theta(t)) = -N\,\nabla_{w,\theta}\ell\big(f, f_N(\cdot, w(t),\theta(t))\big), \\ \\ (w(t),\theta(t)) = (\bar{w},\bar{\theta}), \end{cases} \tag{4}$$

with $w(t) = (w_1(t),\dots,w_N(t))$ and $\theta(t) = (\theta_1(t),\dots,\theta_N(t))$.[3] Unfortunately, given the structure of the approximating functions (1), this problem is non-convex, and thus one does not expect to arrive to the minimizer in general.

Because of this degeneracy, a recent approach has been to consider a continuum model where one lets the number N of neurons go to infinity. The general hope is that this limit problem can be studied with PDE techniques, and then one may try to extract informations also on the original problem (with N fixed) provided N is sufficiently large. This latter step has been studied, for instance, in [4], although many questions are still open (see Sect. 5 for more details).

In this note we shall not discuss the consistency of the approximation as $N \to \infty$, but we instead focus on the analysis of the continuum interpretation. As we shall see, there is more than one way to interpret the limit as $N \to \infty$, and more than one possible formulation exists. In the next sections, we first present the continuum energy functionals that one can obtain by taking the limit of $\ell\big(f, f_N\big)$ as $N \to \infty$, and then we shall analyze the possible gradient flows that can arise from this model.

2.1 The μ Formulation

We start with the most commonly used interpretation of a neural network, when the number of neurons is allowed to go to infinity. In this case, we want to treat the two variables w and θ in the same way. For that, let us slightly reformulate the previous problem.

Set $\xi := (w,\theta) \in \mathbb{R}\times\Theta$, $\Omega := \mathbb{R}\times\Theta \subset \mathbb{R}^{d+1}$, and let us define $\Phi(\xi, x) := w\,h(\theta, x)$, so that we can deal with both parameters simultaneously. Thus, (1) can be written as

$$f_N(x) = f_N(x, \xi_1,\dots,\xi_N) = \frac{1}{N}\sum_{i=1}^{N}\Phi(\xi_i, x).$$

Let μ_N denote the empirical distribution of $\{\xi_i\}_{1\le i\le N}$, namely,

$$\mu_N(\xi) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\xi_i}(\xi).$$

[3] Actually, to avoid overfitting, it is usual to add to the loss function ℓ a convex potential on the parameters; see (7) or (13).

Then the function f_N can be expressed in terms of μ_N as

$$f_N(x) = \int_\Omega \Phi(\xi, x)\mu_N(d\xi),$$

and the gradient descent (4) can be rewritten only in terms of the empirical measure at time t, that is, $\mu_N(t) = \frac{1}{N}\sum_{i=1}^N \delta_{\xi_i(t)}(\xi)$ with $\xi_i(t) = (w_i(t), \theta_i(t))$.

Letting $N \to \infty$, the space of empirical measures can approximate any probability measure $\mu \in \mathscr{P}(\Omega)$. Hence, this suggests the study of approximating functions defined as

$$f_\mu(x) := \int_\Omega \Phi(\xi, x)\mu(d\xi) \qquad \forall\, \mu \in \mathscr{P}(\Omega). \tag{5}$$

Then, our minimization problem consists in minimizing

$$F(\mu) := \frac{1}{2}\int_D (f - f_\mu)^2 dx$$

among probability measures $\mu \in \mathscr{P}(\Omega)$. That is,

$$\min_{\mu\in\mathscr{P}(\Omega)} F(\mu) = \min_{\mu\in\mathscr{P}(\Omega)} \frac{1}{2}\int_D \left(f - \int_\Omega \Phi(\xi, x)\mu(d\xi)\right)^2 dx. \tag{6}$$

In other words, we are looking at the best way of approximating f in $L^2(D)$ using functions of the form (5).

Note that, for many choices of Φ, the set of functions of the form (5) may be dense in $L^2(D)$, so that the minimum may be zero (and we want to study ways to attain or approximate it). Moreover, oftentimes, to avoid overfitting in the training space, it is common to add a potential term used as a renormalization in the optimization of the neural networks. Therefore, the energy that we want to minimize over $\mu \in \mathscr{P}(\Omega)$ becomes

$$F(\mu) = \frac{1}{2}\int_D \left(f - \int_\Omega \Phi(\xi, x)\mu(d\xi)\right)^2 + \int_\Omega V(\xi)\mu(d\xi), \tag{7}$$

for some fixed function $V : \Omega \to \mathbb{R}$. A natural choice of V is given by the quadratic potential

$$V(\xi) = \frac{\lambda}{2}|\xi|^2, \qquad \text{with } \lambda > 0. \tag{8}$$

Notice that, with this additional term, the minimum of our functional will not be zero anymore.

We remark that, by considering probability measures instead of discrete parameters, we are not losing information. Indeed, if we restrict our problem to the set of atomic measures with N atoms, then we go back to formulation (1).

Remark 2.1 One might benefit from the convex structure of the functional F with respect to the classical linear structure of $\mathscr{P}(\Omega)$, namely,

$$F\left(\alpha\mu_1 + (1-\alpha)\mu_2\right) \le \alpha F(\mu_1) + (1-\alpha)F(\mu_2) \qquad \forall \alpha \in [0,1].$$

In particular, from here one can show that if μ_1 and μ_2 are two local minimizers, then $\int_D \Phi(\xi, x)\mu_1(d\xi) = \int_D \Phi(\xi, x)\mu_2(d\xi)$ for all $x \in D$ and their potential energy is the same, i.e., $\int_\Omega V(\xi)\mu_1(d\xi) = \int_\Omega V(\xi)\mu_2(d\xi)$.[4] In particular, local minimizers are unique under Φ.

An advantage of the continuous formulation is that the invariance with respect to permutations of neurons is included in the model. Also, assuming that one already knows symmetries for the objective function (e.g., rotational symmetry to identify certain images), they can be incorporated directly into the minimization problem, much more easily than in the discrete case.

2.2 Comparison Between the Continuous and Discrete Model

At the discrete level, adding a potential term corresponds to considering the minimization of the loss functional

$$F_N(f_N) = \frac{1}{2}\int_D |f(x) - f_N(x)|^2\,dx + \sum_{i=1}^{N} V(\xi_i), \tag{9}$$

[4] Indeed, suppose that μ_1 and μ_2 are two local minimizers, and for $\alpha \in [0,1]$, consider $\mu_\alpha := (1-\alpha)\mu_0 + \alpha\mu_1$. Then, we can compute $\frac{d}{d\alpha}F(\mu_\alpha)$, which equals

$$\frac{d}{d\alpha}F(\mu_\alpha) = \alpha\int_D |f - f_1|^2 - (1-\alpha)\int_D |f - f_0|^2 + (1-2\alpha)$$
$$\times \int_D (f - f_0)(f - f_1) - \int V(\mu_0 - \mu_1).$$

Since μ_0 and μ_1 are local minimizers, we have $\frac{d}{d\alpha}\big|_{\alpha=0}F(\mu_\alpha) \ge 0$ and $\frac{d}{d\alpha}\big|_{\alpha=1}F(\mu_\alpha) \le 0$, and therefore

$$0 \ge \frac{d}{d\alpha}\big|_{\alpha=1}F(\mu_\alpha) - \frac{d}{d\alpha}\big|_{\alpha=0}F(\mu_\alpha) = \int_D |f_0 - f_1|^2$$

thus $f_0 = f_1$. This implies that $\frac{d}{d\alpha}F(\mu_\alpha) = -\int V(\mu_0 - \mu_1)$, so it follows from $\frac{d}{d\alpha}\big|_{\alpha=0}F(\mu_\alpha) \ge 0$ and $\frac{d}{d\alpha}\big|_{\alpha=1}F(\mu_\alpha) \le 0$ that $\int V\mu_0 = \int V\mu_1$.

for some convex function V. Equivalently, we are considering the discrete minimization problem

$$\min_{\xi_i \in \Omega} F_N(f_N) \qquad \text{for} \qquad f_N(x) = \frac{1}{N} \sum_{i=1}^{N} \Phi(\xi_i, x). \tag{10}$$

We have seen that this minimization problem can be interpreted as a particular case of the more general problem for probability measures. Namely, if we consider F given by (7), then Problem (10) generalizes to

$$\min_{\mu \in \mathscr{P}(\Omega)} F(\mu). \tag{11}$$

Notice also that, while Problem (10) is heavily non-convex, Problem (11) has a convex structure (see Remark 2.1).

Consistency

The consistency between Problems (10) and (11) has generated some research in the recent years. These are some results:

(i) If μ_N is the empirical distribution of a minimizer of F_N and μ is a minimizer of F, then $F_N(\mu_N) = F(\mu) + O(N^{-1})$. In addition, if V is coercive, then μ_N converges weakly* to a minimizer μ of F (up to subsequences).

(ii) The Wasserstein gradient flow of F with initialization μ_N is the same as the corresponding gradient descent of the discretized problem, cf. (4) (see [4]).

(iii) As shown in [11] (see also [14]), the stochastic gradient descent for (10) (cf. (4)) converges to the gradient flow of (11) with its own initialization. More precisely, if one denotes by $\mu_N^{(k)}$ the empirical distribution of the parameters $(\xi_i^k)_{1 \le i \le N}$ in the stochastic gradient descent for F_N at step k, then one can prove quantitative convergence of $\mu_N^{(t/\varepsilon)}$ to μ_t as $N \to \infty$ and $\varepsilon \downarrow 0$, where μ_t is the gradient flow in the Wasserstein metric for the functional F.

(iv) In [4] the authors proved that if one approximates an initial measure μ_0 by N atoms, the corresponding gradient descents converge, under some conditions on the initial measure, to the gradient flow for F with initial measure μ_0, also as $t \to \infty$. Thus, they showed that one does actually benefit from the convex structure in (11): given a nice enough initial measure (initial configuration of weights, with enough neurons), its gradient descent will converge to a configuration of parameters very close to a minimizer for F. This is currently a non-quantitative result that checks the consistency of the problem posed.

All these facts show that studying the minimization problem (11) could be very useful in trying to derive properties for the discrete problem (10).

On the other hand, one should keep in mind that, in general, the Wasserstein gradient flow of (6) may not converge to a global minimizer but just to a stationary point. For example, given an initial configuration with a fixed number of deltas, the corresponding gradient flow never increases the amount of deltas, and thus it converges to some measure that has at most the same number of deltas as the initial configuration. In particular, this limit will not generally be a global minimizer of F. Still, the result in [4] says that such limiting configuration will approximate a minimizer, under suitable conditions.

2.3 The (ρ, H) Formulation

An alternative approach to the previous generalization (what we called "the μ formulation") consists in taking advantage of the structure of Φ, where the weights w and positions θ have asymmetric roles. One can think of this approach as a charged particles system, where we can discretize in θ (positions of the particle) assigning a coefficient w to each atomic measure of the discretization (charge of the particle). We refer to some examples in [7].

While these continuous methods a priori do not necessarily arise from a discrete gradient descent, they yield other evolution equations whose discretization could benefit from additional properties. As we will see, some of these associated PDE systems also dissipate energy, suggesting that alternative gradient flow formulations are possible and interesting.

Recall that we have $\theta \in \Theta \subset \mathbb{R}^d$ and $w \in \mathbb{R}$. Consider the measure in θ given by

$$\rho_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} w_i \delta_{\theta_i}(\theta)$$

(observe that now ρ_N is not necessarily a probability measure, and not even a positive measure, since the weights w_i may be negative). Then the function f_N in (1) can be expressed as

$$f_N(x) = \int_\Theta h(\theta, x)\rho_N(d\theta).$$

This suggests considering functions of the form

$$f_m(x) = \int_\Theta h(\theta, x)m(d\theta),$$

where now $m \in \mathcal{M}$ is a finite (signed) measure.

Notice that this is related to what we were doing before with probability measures μ defined on $\Omega = \mathbb{R} \times \Theta$. Indeed, to see the relation between the two formulations, given $\mu \in \mathcal{P}(\Omega)$, consider its disintegration with respect to θ. Namely, one can write

$$\mu(d\xi) = \nu_\theta(dw) \otimes \rho(d\theta),$$

where ρ and ν are formally defined as[5]

$$\rho(\theta) = \int_{\mathbb{R}} \mu(dw, \theta), \qquad \nu_\theta(w) = \frac{1}{\rho(\theta)} \mu(w, \theta).$$

Then, given $\Phi(\xi, x) = w\, h(\theta, x)$, we have

$$\int_\Omega \Phi(\xi, x)\mu(d\xi) = \int_\Theta \left(\int_{\mathbb{R}} w\, \nu_\theta(dw) \right) h(\theta, x)\rho(d\theta) = \int_\Theta h(\theta, x) m(d\theta)$$

where

$$m(d\theta) = \left(\int_{\mathbb{R}} w\, \nu_\theta(dw) \right) \rho(d\theta).$$

In other words, (6) is equivalent to the minimization problem

$$\min_{m \in \mathcal{M}(\Theta)} \frac{1}{2} \int_D \left(f - \int_\Theta h(\theta, x) m(d\theta) \right)^2 dx,$$

where $\mathcal{M}(\Theta)$ denotes the set of (finite) signed measures on Θ. Equivalently, if we define

$$H(\theta) = \int_{\mathbb{R}} w\, \nu_\theta(dw),$$

then our problem consists in finding the best approximation of f in $L^2(D)$ with functions of the form

$$f_{\rho,H}(x) = \int_\Theta H(\theta) h(\theta, x)\rho(d\theta). \tag{12}$$

In addition, keeping the same notation and assuming to introduce a potential term of the form $V(\xi) = \frac{\lambda}{2}|\xi|^2$ in the μ formulation, then by Jensen's inequality, we have

[5] This definition of the disintegration is correct if μ is absolutely continuous, and therefore can be identified as a function. Otherwise, the existence and uniqueness of such representation is provided by the disintegration theorem (see for instance [8, Theorem 1.4.10 and Appendix B]).

$$\int_\Omega |\xi|^2\mu(d\xi) = \int_\Theta\int_{\mathbb{R}} w^2\nu_\theta(dw)\rho(\theta) + \int_\Theta |\theta|^2\rho(d\theta) \geq \int_\Theta \left(H(\theta)^2 + |\theta|^2\right)\rho(d\theta).$$

In particular, if we were assuming $\int |\xi|^2\mu(d\xi) < +\infty$ (i.e., μ has bounded second moments) in the previous formulation, then it is natural to assume ρ to have bounded second moments as well, and $H \in L^2(\Theta, \rho)$.[6]

Notice that the expression (12) is similar to (5), the one appearing in the μ formulation. There are, however, two main differences: on the one hand, the number of parameters has been reduced (from $\xi = (w, \theta)$ to simply θ); on the other hand, we are optimizing not only over probability measures ρ but also over functions $H \in L^2(\Theta, \rho)$. Thus, by looking at the explicit expression $\Phi(\xi, x)$ had in the previous formulation, we are trading off the amount of parameters of our problem with a new variable to optimize. We can do so because, in reality, the freedom given to the measure $\nu_\theta(dw)$ was limited: since $\Phi(\xi, x) = w\,h(\theta, x)$, we only see it through its first moment. In particular, given $\mu = \nu_\theta(dw) \otimes \rho(d\theta)$, one can replace it with $\delta_{H(\theta)}(dw) \otimes \rho(dw)$, and the problem remains the same. This is the idea behind what we call "the (ρ, H) formulation."

In conclusion, in the (ρ, H) formulation, we are considering the functional

$$G(\rho, H) = \frac{1}{2}\int_D \left(f - \int_\Theta H(\theta)h(\theta, x)\rho(d\theta)\right)^2 dx + \int_\Theta \bar{V}(H, \theta)\rho(d\theta), \tag{13}$$

where now we have removed the dependence on the variable w, and we have added a potential term $\bar{V} : \mathbb{R} \times \Theta \to \mathbb{R}$. Note that, in this case, the L^2 regularization induced by (8) corresponds to $\bar{V}(H, \theta) = \frac{\lambda}{2}(H^2 + |\theta|^2)$.

3 PDE Formulations

In this section we first compute the Wasserstein gradient flow in the μ formulation (see Sect. 3.1). Then we discuss some evolution equations in the (ρ, H) formulation, as introduced in [7] (Sect. 3.2). Finally, in Sect. 3.3, we present a new original approach to the problem of defining a gradient flow (ρ, H) formulation, based on propagation of chaos.

[6] Similarly, if our potential term was given by the p-moments instead, i.e., $V(\xi) = \lambda|\xi|^p$ for some $p \geq 1$ and $\lambda > 0$, then it would be natural to assume $H \in L^p(\Theta, \rho)$.

3.1 Gradient Flow in the μ Formulation

Recall that $\Omega = \mathbb{R} \times \Theta \subset \mathbb{R}^{d+1}$, and $\xi = (w, \theta) \in \Omega$ denotes the parameters in this setting. Let $D \subset \mathbb{R}^n$, and let $h(\theta, x) : \Theta \times D \to \mathbb{R}$ be a given function, and let $\Phi(\xi, x) = w\, h(\theta, x)$.

We consider the minimization problem

$$\min_{\mu \in \mathscr{P}(\Omega)} F(\mu), \tag{14}$$

where

$$F(\mu) = \frac{1}{2} \int_D \left(\int_\Omega \Phi(\xi, x) \mu(d\xi) - f(x) \right)^2 dx + \int_\Omega V(\xi)\, \mu(d\xi). \tag{15}$$

Note that this expression can be rewritten as

$$F(\mu) = \bar{F} + \int_{\Omega \times \Omega} K(\xi, \bar{\xi}) \mu(d\xi) \mu(d\bar{\xi}) + \int_\Omega \mathcal{S}(\xi) \mu(d\xi) + \int_\Omega V(\xi) \mu(d\xi), \tag{16}$$

where

$$K(\xi, \bar{\xi}) = \frac{1}{2} \int_D \Phi(\xi, x) \Phi(\bar{\xi}, x) dx, \qquad \mathcal{S}(\xi) = -\int_D \Phi(\xi, x) f(x)\, dx, \tag{17}$$

and $\bar{F} = \frac{1}{2} \| f \|^2_{L^2(D)}$ is a constant. We remark that the smoothness of $\Phi(\xi, x)$ is related to the smoothness of $\mathcal{S}$ (in particular, if Φ is smooth, then $\mathcal{S}$ is a smooth).

The first variation of F with respect to μ at fixed measure $\mu_* \in \mathscr{P}(\Omega)$ is given by[7]

$$\begin{aligned} \frac{\delta F}{\delta \mu}(\mu_*) &= \int_D \Phi(\cdot, x) \left[\int_\Omega \Phi(\bar{\xi}, x)\, d\mu_*(\bar{\xi}) - f(x) \right] dx + V \\ &= 2 \int_\Omega K(\cdot, \bar{\xi}) \mu_*(d\bar{\xi}) + \mathcal{S} + V, \end{aligned} \tag{18}$$

so that the Wasserstein subdifferential on the support of μ_* is

[7] By definition, $\frac{\delta F}{\delta \mu}(\mu_*)$ is defined as the unique element such that

$$\left. \frac{d}{d\varepsilon} \right|_{\varepsilon = 0} F(\mu_* + \varepsilon \varphi) = \int_\Omega \frac{\delta F}{\delta \mu}(\mu_*)\, \varphi\, d\xi \qquad \forall \varphi \in C_c^\infty(\Omega).$$

$$\nabla\frac{\delta F}{\delta\mu}(\mu_*) = \int_D \nabla_\xi \Phi(\cdot, x)\left[\int_\Omega \Phi(\bar{\xi}, x)\, d\mu_*(\bar{\xi}) - f(x)\right] dx + \nabla V$$
$$= 2\int_\Omega \nabla_\xi K(\cdot, \bar{\xi})\mu_*(d\bar{\xi}) + \nabla\mathcal{S} + \nabla V$$

(see, for instance, [1, Chapter 10] or [8, Chapter 4.2]). Also, the Wasserstein gradient flow of F is by definition (see [8, Chapter 4.2]

$$\partial_t\mu_t = \mathrm{div}\left(\mu_t\nabla\frac{\delta F}{\delta\mu}(\mu_t)\right), \tag{19}$$

therefore the formulas above give us the following PDE:

$$\boxed{\partial_t\mu_t = \mathrm{div}\,(\mu_t\nabla\mathcal{L}(\mu_t)) + \mathrm{div}\,(\mu_t\nabla\mathcal{S}) + \mathrm{div}(\mu_t\nabla V),} \tag{20}$$

with

$$\mathcal{L}(\mu_t)(\xi) = 2\int_\Omega K(\xi, \bar{\xi})\mu_t(d\bar{\xi}). \tag{21}$$

Notice that $\mathcal{L}$ is an integral operator that is positive semi-definite.[8] Also, it can be checked by a direct computation that a solution $\mu_t(\xi)$ of the PDE satisfies an energy dissipation from the gradient flow structure, that is, the energy is monotone non-increasing along trajectories:

$$\frac{d}{dt}F(\mu_t) = -\int_\Omega \left|\nabla\frac{\delta F}{\delta\mu}(\mu_t)\right|^2 \mu_t(d\xi). \tag{22}$$

In particular, stationary points correspond to measures for which the derivative of the energy is zero. This motives the following:

Definition 3.1 We say that measure μ_* is a stationary point of our functional F (in the Wasserstein sense) if

$$\nabla\frac{\delta F}{\delta\mu}(\mu_*) = 0 \qquad \text{on} \quad \mathrm{supp}(\mu_*). \tag{23}$$

[8] Indeed, it follows by (17) that

$$\int_\Omega \mathcal{L}(\mu)(\xi)\,\mu(d\xi) = \frac{1}{2}\int_D\left(\int_\Omega \Phi(\xi, x)\mu(d\xi)\right)^2 dx \geq 0.$$

Notice that if we consider the natural potential term $V(\xi) = \frac{\lambda}{2}|\xi|^2$, then our PDE (20) becomes

$$\partial_t \mu_t = \operatorname{div}(\mu_t \nabla \mathcal{L}(\mu_t)) + \operatorname{div}(\mu_t \nabla \mathcal{S}) + \lambda \operatorname{div}(\mu_t\, \xi).$$

Let us conclude this subsection by observing that, in general, the previous PDEs are posed when the domain $\Omega = \mathbb{R}^{d+1}$. If, instead, one considers Θ a bounded smooth domain, an extra zero Neumann boundary condition (so that the mass cannot escape) needs to be imposed:

$$\boxed{\boldsymbol{\nu} \cdot \nabla\left(\mathcal{L}(\mu_t) + \mathcal{S} + V\right)\, \mu_t = 0 \quad \text{on } \partial\Omega,} \tag{24}$$

where $\boldsymbol{\nu}$ denotes the unit outer normal vector to $\partial\Omega$.

3.2 A First PDE Approach in the (ρ, H) Formulation

As discussed before, an alternative approach is based on the (ρ, H) formulation described in Sect. 2.3. So, it makes sense to design an appropriate evolution system of PDEs with good convergence properties, which could potentially lead to a nice particle method in the discrete case.

Let $\Theta \subset \mathbb{R}^d$ be the parameter space in this setting. Let $D \subset \mathbb{R}^n$ and let $h(\theta, x) : \Theta \times D \to \mathbb{R}$ be a fixed activation function.

We consider now the functional

$$G(\rho, H) = \frac{1}{2}\int_D \left(\int_\Theta H(\theta)h(\theta, x)\rho(d\theta) - f(x)\right)^2 dx + \int_\Theta \bar{V}(H, \theta)\rho(d\theta), \tag{25}$$

where, as before, $f \in L^2(D)$ is a given function.

As in (16), we can write

$$\begin{aligned} G(\rho, H) = \bar{G} + \int_{\Theta\times\Theta} &\bar{K}(\theta, \bar{\theta})H(\theta)\rho(d\theta)H(\bar{\theta})\rho(d\bar{\theta}) \\ &+ \int_\Theta \overline{\mathcal{S}}(\theta)H(\theta)\rho(d\theta) + \int_\Theta \bar{V}(H, \theta)\rho(d\theta), \end{aligned} \tag{26}$$

where

$$\bar{K}(\theta, \bar{\theta}) = \frac{1}{2}\int_D h(\theta, x)h(\bar{\theta}, x)dx, \qquad \overline{\mathcal{S}}(\theta) = -\int_D h(\theta, x)f(x)\,dx, \tag{27}$$

and $\bar{G} = \frac{1}{2}\|f\|^2_{L^2(D)}$ is a constant. Notice that, also as before, the function $\overline{\mathcal{S}}$ is smooth if h is smooth with respect to θ.

We shall directly focus on the quadratic potential

$$\bar{V}(H,\theta) = \bar{V}_\lambda(H,\theta) := \frac{\lambda}{2}\left(H^2 + |\theta|^2\right), \tag{28}$$

so that, as discussed in Sect. 2.3, the natural space for H is given by $L^2(\Theta,\rho)$, and our minimization problem is given by

$$\min_{\substack{\rho\in\mathscr{P}(\Theta)\\ H\in L^2(\Theta,\rho)}} G(\rho, H). \tag{29}$$

We now want to obtain an evolution system of PDEs for (ρ, H) with nice properties. As we shall see, this can be performed in more than one way.

We first start with the evolution of $\rho \in \mathscr{P}(\Theta)$. As before, it makes sense to make it evolve according to the Wasserstein gradient flow of G. Namely, if we denote (ρ_t, H_t) our evolution variables, we have

$$\partial_t\rho_t = \operatorname{div}\left(\rho_t\nabla\frac{\delta G}{\delta\rho}(\rho_t, H_t)\right),$$

where the first variation density of G with respect to ρ at $(\rho_*, H_*) \in \mathscr{P}(\Theta) \times L^2(\Theta,\rho_*)$ is given by

$$\frac{\delta G}{\delta\rho}(\rho_*, H_*) = 2H_*(\cdot)\int_\Theta \bar{K}(\cdot,\bar\theta)H_*(\bar\theta)\rho_*(d\bar\theta) + \overline{\mathcal{S}}H_* + \bar{V}_\lambda(\theta, H_*),$$

so that

$$\nabla\frac{\delta G}{\delta\rho}(\rho_*, H_*) = 2\nabla\left[H_*(\cdot)\int_\Theta \bar{K}(\cdot,\bar\theta)H_*(\bar\theta)\rho(d\bar\theta)\right] + \nabla\left(\overline{\mathcal{S}}H_*\right) + \partial_H\bar{V}_\lambda(\cdot, H_*)\nabla H_* + (\nabla_\theta\bar{V}_\lambda)(\cdot, H_*).$$

Thus, recalling (28), the evolution of ρ_t is given by

$$\boxed{\partial_t\rho_t = \operatorname{div}\left[\rho_t\nabla(H_t\overline{\mathcal{L}}(\rho_t, H_t))\right] + \operatorname{div}\left(\rho_t\nabla(\overline{\mathcal{S}}H_t)\right) + \lambda\operatorname{div}\left[\rho_t\left(H_t\nabla H_t + \theta\right)\right],} \tag{30}$$

where

$$\overline{\mathcal{L}}(\rho_t, H_t)(\theta) := 2\int_\Theta \bar{K}(\theta,\bar\theta)H_t(\bar\theta)\rho_t(d\bar\theta) \tag{31}$$

is a positive semi-definite integral operator (cf. (21)).

This gives the evolution of ρ_t, and one needs to couple it with an evolution for H_t. We shall present now two possible approaches.

Separating Variables

The first way to obtain an evolution for the non-conserved variable H_t is to disregard partially the interaction between H and ρ: one performs the Wasserstein gradient flow of ρ, on the one hand, and the $L^2(\Theta, \rho)$ gradient flow of H, on the other (see [7, Examples 1 and 2]).

Namely, one considers

$$\partial_t H_t = -\frac{\delta G}{\delta H}(\rho_t, H_t)$$

where, for a fixed ρ_t, $\frac{\delta G}{\delta H}(\rho_t, \cdot)$ denotes the variation of $G(\rho_t, \cdot)$ with respect to H in $L^2(\Theta, \rho_t)$.[9] This is

$$\frac{\delta G}{\delta H}(\rho_*, H_*) = 2\int_\Theta \bar{K}(\cdot, \bar{\theta}) H_*(\bar{\theta})\rho(d\bar{\theta}) + \overline{\mathcal{S}} + \partial_H \bar{V}_\lambda(\cdot, H_*)$$

on supp ρ_*, and therefore the evolution of (ρ_t, H_t) is given by

$$\boxed{\begin{aligned} \partial_t \rho_t = \operatorname{div}\big[\rho_t \nabla(H_t \overline{\mathcal{L}}(\rho_t, H_t))\big] + \operatorname{div}\big(\rho_t \nabla(\overline{\mathcal{S}} H_t)\big) + \lambda \operatorname{div}\big[\rho_t\big(H_t \nabla H_t + \theta\big)\big] \\ \partial_t H_t = -\overline{\mathcal{L}}(\rho_t, H_t) - \overline{\mathcal{S}} - \lambda H_t \end{aligned}} \tag{32}$$

(this is coupled with a Neumann boundary condition for ρ_t, analogous to (24), whenever the domain Θ is not $\mathbb{R}^d$).

Note that this evolution has some difficulties, since one needs to make sure that all the terms appearing in the above PDE are well defined, at least in a weak sense. For instance, one needs to ensure that $\rho_t H_t \nabla H_t$ is well defined. Giving a meaning to this expression may be delicate if ρ is a singular measure. However, at least in the smooth case, this PDE makes sense. In addition, there is dissipation of the energy G along the path (ρ_t, H_t), namely,

[9] Namely, for a fixed $\rho_* \in \mathscr{P}(\Theta)$, $\frac{\delta G}{\delta H}(\rho_*, H_*)$ is the unique function in $L^2(\Theta, \rho_*)$ such that

$$\left\langle \frac{\delta G}{\delta H}(\rho_*, H_*), \varphi \right\rangle_{L^2(\Theta, \rho_*)} = \frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} G(H_* + \varepsilon\varphi) \qquad \forall \varphi \in L^2(\Theta, \rho_*).$$

$$\frac{d}{dt}G(\rho_t, H_t) = -\int_\Theta \left|\nabla \frac{\delta G}{\delta \rho}(\rho_t, H_t)\right|^2 \rho_t(d\theta) - \int_\Theta \left|\frac{\delta G}{\delta H}(\rho_t, H_t)\right|^2 \rho_t(d\theta).$$

In particular, since $G(\rho_t, H_t)$ controls the $L^2(\Theta, \rho_t)$ norm of H_t, if one starts from a pair (ρ_0, H_0) with $H_0 \in L^2(\Theta, \rho_0)$, then $H_t \in L^2(\Theta, \rho_t)$ (whenever the evolution is well defined). Also, integrating the dissipation inequality above over any time interval implies that

$$\int_0^\infty \left[\int_\Theta \left|\nabla \frac{\delta G}{\delta \rho}(\rho_t, H_t)\right|^2 \rho_t(d\theta) + \int_\Theta \left|\frac{\delta G}{\delta H}(\rho_t, H_t)\right|^2 \rho_t(d\theta)\right] dt \le G(\rho_0, H_0),$$

which implies in particular that $\nabla \frac{\delta G}{\delta \rho}(\rho_t, H_t)$ and $\frac{\delta G}{\delta H}(\rho_t, H_t)$ belong to $L^2(\Theta, \rho_t)$ for a.e. t.

Transporting Along the Flow of ρ_t

Another way to describe the evolution of H_t is by incorporating the information that it is transported along the flow in the corresponding variable to be studied.

More precisely, note that the evolution of ρ_t in (32) can be written as a continuity equation (see [8, Eq. (4.6)]):

$$\partial_t \rho_t + \operatorname{div}(\rho_t \boldsymbol{v}_t) = 0, \qquad \text{where} \quad \boldsymbol{v}_t = -\nabla \frac{\delta G}{\delta \rho}(\rho_t, H_t).$$

Hence, if we define $X_t : \Theta \to \Theta$ as the flow of $\boldsymbol{v}_t$, namely,

$$\begin{cases} \dot{X}_t = \boldsymbol{v}_t \circ X_t \\ X_0 = \mathrm{Id}, \end{cases} \tag{33}$$

then $\rho_t = (X_t)_\# \rho_0$, where $(X_t)_\# \rho_0$ denotes the push-forward measure of ρ_0 through the map X_t.[10]

Thus, instead of considering simply H_t (which does not see the flow for ρ_t), an alternative option consists in rewriting the functional in terms of the variable $H_t \circ X_t$, which corresponds to transporting H_t along the flow of ρ_t. Hence, recalling that we are considering the potential $\bar{V}_\lambda$ from (28), one considers the evolution of $H_t \circ X_t$ given by

[10] That is,

$$\int_\Theta \varphi(\theta)[(X_t)_\# \rho_0](d\theta) = \int_\Theta \varphi(X_t(\theta))(\rho_0(d\theta))$$

for any Borel function $\varphi : \Theta \to \mathbb{R}$.

$$\partial_t(H_t \circ X_t) = -(\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}} - \lambda H_t) \circ X_t.$$

Noticing that $\partial_t(H_t \circ X_t) = [\partial_t H_t + \boldsymbol{v}_t \cdot \nabla H_t] \circ X_t$ (as a consequence of (33)), one obtains

$$\partial_t H_t + \boldsymbol{v}_t \cdot \nabla H_t = -\overline{\mathcal{L}}(\rho_t, H_t) - \overline{\mathcal{S}} - \lambda H_t.$$

Hence, the evolution system now becomes

$$\boxed{\begin{aligned} \partial_t \rho_t + \operatorname{div}(\rho_t \boldsymbol{v}_t) = 0 \\ \partial_t H_t + \boldsymbol{v}_t \cdot \nabla H_t = -\overline{\mathcal{L}}(\rho_t, H_t) - \overline{\mathcal{S}} - \lambda H_t, \end{aligned}} \tag{34}$$

with

$$\boxed{\boldsymbol{v}_t = -\nabla(H_t \overline{\mathcal{L}}(\rho_t, H_t) + H_t \overline{\mathcal{S}}) - \lambda H_t \nabla H_t - \lambda \theta,} \tag{35}$$

and again there is a zero Neumann boundary condition for ρ_t whenever Θ is not $\mathbb{R}^d$ (see (24)).

This corresponds the system introduced in [7, Section 5.4] in the zero potential case ($\lambda = 0$), where they also design a particle method for this "modified gradient flow." This is definitely a very interesting model. However, since this system does not seem to dissipate energy in general, the mathematical analysis becomes more complicated.

3.3 A Gradient Flow in the (ρ, H) Formulation via Propagation of Chaos

Let us give yet another possible evolution for (ρ_t, H_t) that produces a dissipative flow and does not rely on the smoothness of the measure. In this case, we do so by expressing the evolution in the μ formulation in the (ρ, H) variables, under a propagation of chaos assumption. As we shall explain below, the resulting system in this case is given by

$$\boxed{\begin{aligned} \partial_t \rho_t + \operatorname{div}(\rho_t \boldsymbol{w}_t) = 0 \\ \partial_t H_t + \boldsymbol{w}_t \cdot \nabla H_t = -\overline{\mathcal{L}}(\rho_t, H_t) - \overline{\mathcal{S}} - \partial_H \bar{V}(\cdot, H_t), \end{aligned}} \tag{36}$$

where now the vector $\boldsymbol{w}_t$ is (cf. (35))

$$\boxed{\boldsymbol{w}_t = -H_t\nabla(\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}}) - \nabla_\theta \bar{V}(\cdot, H_t),} \tag{37}$$

and with zero Neumann boundary conditions

$$\boxed{\boldsymbol{\nu}\cdot \boldsymbol{w}_t\, \rho_t = 0 \quad \text{on } \partial\Theta} \tag{38}$$

whenever Θ is not $\mathbb{R}^d$. In particular, the evolution of H_t is still given by the corresponding evolution along the flow transporting ρ_t, but differently from before, ρ_t is not the standard Wasserstein flow. As proved below, this system has the main advantage that it dissipates energy; see Proposition 3.3.

In order to motivate the previous evolution system for the pair (ρ_t, H_t), we start by rewriting the PDE (20) as a hierarchy system in the (w, θ) variables. This is an infinite non-closed system of PDEs that depends on higher moments for the disintegration ν_θ and for the first derivatives of the potential.

Lemma 3.2 *Let $\Phi(\xi, x) = w\, h(\theta, x)$ and $\Omega = \mathbb{R}\times\Theta$ with Θ a smooth domain.*

Consider μ_t a (smooth and fast decaying) solution to (20) and define the disintegration into probability measures

$$\mu_t(\xi) = \nu_{\theta,t}(w)\otimes\rho_t(\theta).$$

Define $H_{t,i}(\theta) := \int_{\mathbb{R}} w^i\nu_{\theta,t}(dw)$,

$$V^w_{t,i}(\theta) := \int_{\mathbb{R}} w^i\partial_w V(w,\theta)\nu_{\theta,t}(dw), \quad \boldsymbol{V}^\theta_{t,i}(\theta) := \int_{\mathbb{R}} w^i\nabla_\theta V(w,\theta)\nu_{\theta,t}(dw),$$

and consider

$$\overline{\mathcal{L}}(\rho_t, H_t)(\theta) := 2\int_\Theta \bar{K}(\theta,\bar\theta)H_t(\bar\theta)\rho_t(d\bar\theta), \tag{39}$$

and $\bar{K}$ and $\overline{\mathcal{S}}$ be given by (27). Then, we have

$$\begin{cases} \partial_t\rho_t = \operatorname{div}_\theta\left(\rho_t H_{t,1}\nabla_\theta\left[\overline{\mathcal{L}}(\rho_t, H_{t,1}) + \overline{\mathcal{S}}\right]\right) + \operatorname{div}_\theta(\rho_t \boldsymbol{V}^\theta_{t,0}) \\ \partial_t(H_{t,i}\rho_t) = \operatorname{div}_\theta\left(\rho_t H_{t,i+1}\nabla_\theta\left[\overline{\mathcal{L}}(\rho_t, H_{t,1}) + \overline{\mathcal{S}}\right]\right) + \operatorname{div}_\theta(\rho_t\boldsymbol{V}^\theta_{t,i}) \\ \qquad\qquad -iH_{t,i-1}\rho_t(\overline{\mathcal{L}}(\rho_t, H_{t,1}) + \overline{\mathcal{S}}) - i\,V^w_{t,i-1}\rho_t \qquad \forall i\ge 1, \end{cases} \tag{40}$$

with boundary conditions (whenever $\Theta\neq\mathbb{R}^d$)

$$\boldsymbol{\nu}\cdot\left\{\rho_t H_{t,i}\nabla_\theta\left[\overline{\mathcal{L}}(\rho_t, H_{t,1}) + \overline{\mathcal{S}}\right] + \rho_t\boldsymbol{V}^\theta_{t,i-1}\right\} = 0 \quad \textit{on } \partial\Theta. \tag{41}$$

Proof Notice that

$$\mathcal{L}(\mu_t)(\xi) = w\overline{\mathcal{L}}(\rho_t, H_{t,1})(\theta), \qquad \mathcal{S} = w\overline{\mathcal{S}} \tag{42}$$

(recall (17), (21), (27), and (31)). Integrating (20) with respect to w and recalling (42), we obtain the first equation

$$\partial_t \rho_t = \mathrm{div}_\theta(\rho_t H_{t,1} \nabla_\theta \overline{\mathcal{L}}(\rho_t, H_{t,1})) + \mathrm{div}_\theta(\rho_t H_{t,1} \nabla_\theta \overline{\mathcal{S}}) + \mathrm{div}_\theta(\rho_t \boldsymbol{V}^\theta_{t,0})$$

using that

$$\int_{\mathbb{R}} w\mu_t(dw, \theta) = \rho_t H_{t,1}, \qquad \int_{\mathbb{R}} \nabla_\theta V(w, \theta)\mu_t(dw, \theta) = \rho_t \boldsymbol{V}^\theta_{t,0},$$

and that $\mu_t w$ has sufficient decay in w so that the terms in ∂_w in the divergence disappear when integrating by parts.

Similarly, given $i \geq 1$, we multiply (20) by w^i and then integrate with respect to w, to obtain

$$\begin{aligned}\partial_t(H_{t,i}\rho_t) = {} & \mathrm{div}_\theta\left(\rho_t H_{t,i+1} \nabla_\theta\left[\overline{\mathcal{L}}(\rho_t, H_{t,1}) + \overline{\mathcal{S}}\right]\right) + \mathrm{div}_\theta(\rho_t \boldsymbol{V}^\theta_{t,i}) \\ & + \int_{\mathbb{R}} w^i \partial_w\left[\left(\overline{\mathcal{L}}(\rho_t, H_{t,1}) + \overline{\mathcal{S}} + \partial_w V\right)\nu_{\theta,t}(dw) \otimes \rho_t\right].\end{aligned}$$

Integrating by parts, we obtain the desired result.

The Neumann boundary conditions follow with the same procedure. □

As noticed above, the previous system (40) is not closed, as the i-th equation depends on H_{i+1}. Note however that the system could be closed if one knew that $\nu_{\theta,t}(w) = \delta_{H_t(\theta)}(w)$, since in that case

$$H_{t,i}(\theta) = H_t(\theta)^i = H^i_{t,1}(\theta) \qquad \forall i \geq 1.$$

This suggests a propagation of chaos assumption on the w variable in the previous expressions: by *assuming* that μ preserves being a delta in the w-variable (viz., $\nu_{\theta,t}(dw) = \delta_{H_t(\theta)}$ for some $H_t(\theta)$ for all $t \geq 0$), one gets a well-defined system of equations that now depends only on (ρ, H) and no longer sees the μ structure from before. In this case, if we denote $H_t = H_{t,1}$, we have that $H_{t,2} = H_t^2$, $V^w_{t,0} = \partial_w V(H_t, \theta)$, and $\boldsymbol{V}^\theta_{t,i} = H^i_t \nabla_\theta V(H_t, \theta)$. Also, since the equation for $H_{t,1}$ is already closed, one does not need to look at the other equations for $i \geq 2$.

Based on this discussion, our proposed new system is given by the following evolution equations:

$$\begin{cases} \partial_t \rho_t = \mathrm{div}\left(\rho_t H_t \nabla\left[\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}}\right]\right) + \mathrm{div}(\rho_t \nabla_\theta \bar{V}(\theta, H_t)) \\ \partial_t(H_t\rho_t) = \mathrm{div}\left(\rho_t H_t^2 \nabla\left[\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}}\right]\right) - \rho_t(\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}}) \\ \qquad\qquad +\mathrm{div}(\rho_t H_t \nabla_\theta \bar{V}(\theta, H_t)) - \rho_t \partial_H \bar{V}(\theta, H_t), \end{cases} \tag{43}$$

where $\overline{\mathcal{L}}$ is given by (39), and it is combined with the zero Neumann boundary condition

$$\boldsymbol{\nu}\cdot\left\{\rho_t H_t \nabla\left[\overline{\mathcal{L}}(\rho_t, H_t)+\overline{\mathcal{S}}\right]+\rho_t\nabla_\theta\bar{V}(\theta, H_t)\right\}=0 \quad \text{on } \partial\Theta \tag{44}$$

whenever Θ is not $\mathbb{R}^d$ (cf. (40)–(41)).

Note that (43)–(44) is exactly our proposed model (36)–(38). In order to see that this new system is a reasonable candidate for the minimization of the energy (26), we prove now that the energy decreases along this evolution.

Proposition 3.3 *Let* (ρ_t, H_t) *solve* (43)–(44)*, and let* G *be given by* (26)*. Then*

$$\begin{aligned}\frac{d}{dt}G(\rho_t, H_t) = &-\int_\Theta \left|H_t\nabla(\overline{\mathcal{L}}(\rho_t, H_t)+\overline{\mathcal{S}})+\nabla_\theta\bar{V}(\theta, H_t)\right|^2\rho_t\\ &-\int_\Theta\left(\overline{\mathcal{L}}(\rho_t, H_t)+\overline{\mathcal{S}}+\partial_H\bar{V}(\theta, H_t)\right)^2\rho_t.\end{aligned}$$

In particular, the energy G *is decreasing along* (ρ_t, H_t)*.*

Proof We compute the derivative of $G(\rho_t, H_t)$ starting from (26). We have

$$\begin{aligned}\frac{d}{dt}G(\rho_t, H_t) &= \int_\Theta(\overline{\mathcal{L}}(\rho_t, H_t)+\overline{\mathcal{S}})\,\partial_t(H_t\rho_t)\\ &\quad+\int_\Theta\partial_H\bar{V}(\theta, H_t)\rho_t\partial_t H_t+\int_\Theta\bar{V}(\theta, H_t)\partial_t\rho_t\\ &= I+II+III,\end{aligned}$$

so that we can use (43) to substitute the time derivatives by the corresponding expressions. In particular, using that $\partial_t(H_t\rho_t) = H_t\partial_t\rho_t+\rho_t\partial_t H_t$ and (43), we deduce that

$$\begin{aligned}\rho_t\partial_t H_t = \rho_t H_t\nabla H_t\cdot\nabla\left[\overline{\mathcal{L}}(\rho_t, H_t)+\overline{\mathcal{S}}\right]-\rho_t(\overline{\mathcal{L}}(\rho_t, H_t)+\overline{\mathcal{S}})\\ +\rho_t\nabla H_t\cdot\nabla_\theta\bar{V}(\theta, H_t)-\rho_t\partial_H\bar{V}(\theta.H_t).\end{aligned}$$

For the sake of readability, let us denote

$$\begin{aligned}\overline{\mathcal{N}}_t &:= \overline{\mathcal{L}}(\rho_t, H_t)+\overline{\mathcal{S}}, & \bar{V}_t &:= \bar{V}(\theta, H_t),\\ \partial_H\bar{V}_t &:= (\partial_H\bar{V})(\theta, H_t), & \nabla_\theta\bar{V}_t &:= (\nabla_\theta\bar{V})(\theta, H_t).\end{aligned}$$

Using these formulas, and integrating by parts using (44), we get

$$I = \int_\Theta\overline{\mathcal{N}}_t\,\partial_t(H_t\rho_t) = -\int_\Theta|\nabla\overline{\mathcal{N}}_t|^2H_t^2\rho_t-\int_\Theta\overline{\mathcal{N}}_t^2\rho_t$$

$$- \int_\Theta \nabla \overline{\mathcal{N}}_t \cdot \nabla_\theta \bar{V}_t H_t \rho_t - \int_\Theta \partial_H \bar{V}_t \overline{\mathcal{N}}_t \rho_t,$$

$$\begin{aligned} II = &\int_\Theta \partial_H \bar{V}_t \nabla H_t \cdot \nabla \overline{\mathcal{N}}_t H_t \rho_t - \int_\Theta \partial_H \bar{V}_t \overline{\mathcal{N}}_t \rho_t \\ &+ \int_\Theta \partial_H \bar{V}_t \nabla H_t \cdot \nabla_\theta \bar{V}_t \rho_t - \int_\Theta |\partial_H \bar{V}_t|^2 \rho_t, \end{aligned}$$

and

$$\begin{aligned} III = &- \int_\Theta \nabla_\theta \bar{V}_t \cdot \nabla \overline{\mathcal{N}}_t H_t \rho_t - \int_\Theta \partial_H \bar{V}_t \nabla H_t \cdot \nabla \overline{\mathcal{N}}_t H_t \rho_t \\ &- \int_\Theta |\nabla_\theta \bar{V}_t|^2 \rho_t - \int_\Theta \partial_H \bar{V}_t \nabla H_t \cdot \nabla_\theta \bar{V}_t \rho_t. \end{aligned}$$

Adding these identities, one finally gets

$$\begin{aligned} I + II + III = &- \int_\Theta |\nabla \overline{\mathcal{N}}_t|^2 H_t^2 \rho_t - \int_\Theta |\nabla_\theta \bar{V}_t|^2 \rho_t - 2 \int_\Theta \nabla \overline{\mathcal{N}}_t \cdot \nabla_\theta \bar{V}_t H_t \rho_t \\ &- \int_\Theta \overline{\mathcal{N}}_t^2 \rho_t - \int_\Theta |\partial_H \bar{V}_t|^2 \rho_t - 2 \int_\Theta \partial_H \bar{V}_t \overline{\mathcal{N}}_t \rho_t, \end{aligned}$$

from which we obtain the desired result. □

Remark 3.4 The fact that the system (36) dissipates energy suggests that there might be a gradient flow structure associated to it. We claim that this is the case.

Indeed, denote by $\Gamma(\rho^{(1)}, \rho^{(2)})$ the set of transport plans between $\rho^{(1)}$ and $\rho^{(2)}$, namely,

$$\Gamma(\rho^{(1)}, \rho^{(2)}) := \left\{ \gamma \in \mathscr{P}(\Theta \times \Theta) : \pi^j_\# \gamma = \rho^{(j)} \right\},$$

where $\pi^j : \Theta \times \Theta \to \Theta$, $j = 1, 2$, are the canonical projection onto the first and second factor, respectively. Then, we consider the distance between $(\rho^{(1)}, H^{(1)})$ and $(\rho^{(2)}, H^{(2)})$ given by

$$\begin{aligned} &\mathcal{D}^2((\rho^{(1)}, H^{(1)}), (\rho^{(2)}, H^{(2)})) \\ &\quad := \inf_{\gamma \in \Gamma(\rho^{(1)}, \rho^{(2)})} \int_{\Theta \times \Theta} \left(|\theta_1 - \theta_2|^2 + |H^{(1)}(\theta_1) - H^{(2)}(\theta_2)|^2 \right) d\gamma(\theta_1, \theta_2). \end{aligned}$$

A classical but tedious computation shows that, at least formally, the gradient flow of G with the distance $\mathcal{D}$ is given by (36).

It would be interesting to make this argument rigorous (perhaps using a scheme *à la* JKO [1, 10]) and to use this gradient flow interpretation to better study (36).

4 Regularized Problems

In order to study the behavior of solutions to the PDEs constructed in the previous sections (viz., (20), (32), (34)–(35), or (36)–(37)), it is sometimes convenient to regularize them by adding a small perturbation to the energy functional (or the PDE) that regularizes it.

For simplicity, we focus here on (20), although similar discussions could be done to the other PDEs. We present here two possible of such strategies, by converting the original PDE into a heat-type equation or a porous medium-type equation.

4.1 *Heat Regularization*

A natural way to control the degeneracy of critical points of our functional in (7) or (15) is to add a small entropy term in the minimization procedure. That is, for $\tau > 0$, consider the functional

$$F_\tau(\mu) = \frac{1}{2}\int_D \left(\int_\Omega \Phi(\xi, x)\mu(d\xi) - f(x)\right)^2 dx + \int_\Omega V(\xi)\,\mu(d\xi) + \tau\,\mathrm{Ent}(\mu) \tag{45}$$

where

$$\mathrm{Ent}(\mu) := \begin{cases} \int_\Omega \rho(\xi)\log(\rho(\xi))d\xi & \text{if } \mu(d\xi) = \rho(\xi)\,d\xi, \\ \\ +\infty & \text{if } \mu \not\ll d\xi. \end{cases}$$

Adding this entropy term corresponds to a variation in the stochastic gradient descent in which, when performing discrete in time approximations of (4), one adds a noisy diffusion term. Alternatively, in terms of the PDE describing the evolution of the gradient flow in the Wasserstein metric of F_τ, the addition of the entropy corresponds to adding a small diffusive term in the right-hand side of (20) (see, for instance, [10]). Thus, if μ_t is the Wasserstein gradient flow of F_τ, then

$$\partial_t \mu_t = \mathrm{div}\,(\mu_t \nabla \mathcal{L}(\mu_t)) + \mathrm{div}\,(\mu_t \nabla \mathcal{S}) + \mathrm{div}(\mu_t \nabla V) + \tau\Delta\mu_t, \tag{46}$$

where we are using the notation in (17) and (21).

As before, when Ω is not $\mathbb{R}^{d+1}$, we add zero Neumann boundary conditions:

$$\boldsymbol{\nu} \cdot \{\nabla\left(\mathcal{L}(\mu_t) + \mathcal{S} + V\right) \mu_t + \tau \nabla \mu_t\} = 0 \quad \text{on } \partial\Omega. \tag{47}$$

The PDE (46)–(47) presents a nicer structure than the original (20), and it has been studied in the context of training shallow neural networks. In particular, in [9, 11], this equation appears when approximating functions f by an increasing number of "bumps." There, the authors prove existence and uniqueness of solutions (even in domains with Neumann boundary conditions (47)), and they provide some regularity and convergence estimates for solutions. Observe that, in this case, one gets immediate smoothing (and also immediate full support) for μ_t.

It is interesting to rewrite (45) in a different way, in terms of stationary solutions. Indeed, let us denote by $\mu_* \in \mathscr{P}(\Omega)$ a stationary solution (viz., such that the corresponding dissipation vanishes; see Definition 3.1). Then, we can write

$$\begin{aligned} F_\tau(\mu) - F_\tau(\mu_*) = &\int_{\Omega\times\Omega} K(\xi,\bar{\xi})\mu(d\xi)\mu(d\bar{\xi}) - \int_{\Omega\times\Omega} K(\xi,\bar{\xi})\mu_*(d\xi)\mu_*(d\bar{\xi}) \\ &+ \int_\Omega \mathcal{S}(\xi)(\mu-\mu_*)(d\xi) + \int_\Omega V(\xi)(\mu-\mu_*)(d\xi) \\ &+ \tau\left(\mathrm{Ent}(\mu) - \mathrm{Ent}(\mu_*)\right). \end{aligned} \tag{48}$$

On the other hand, since μ_* is a stationary solution, it has full support, and the first variation density of F_τ must be constant everywhere. That is,

$$\frac{\delta F_\tau}{\delta\mu}(\mu_*) = 2\int_\Omega K(\xi,\bar{\xi})\mu_*(d\bar{\xi}) + \mathcal{S}(\xi) + V(\xi) + \tau\log(\mu_*) \equiv \lambda \quad \text{in } \Omega$$

for some $\lambda \in \mathbb{R}$. In particular, integrating with respect to both μ_* and μ, we get

$$\begin{aligned} 2\int_{\Omega\times\Omega} K(\xi,\bar{\xi})\mu_*(d\xi)\mu_*(d\bar{\xi}) + \int_\Omega \left(\mathcal{S}(\xi) + V(\xi)\right)\mu_*(d\xi) + \tau\int_\Omega \log(\mu_*)\mu_* = \lambda, \\ 2\int_{\Omega\times\Omega} K(\xi,\bar{\xi})\mu(d\xi)\mu_*(d\bar{\xi}) + \int_\Omega \left(\mathcal{S}(\xi) + V(\xi)\right)\mu(d\xi) + \tau\int_\Omega \log(\mu_*)\mu = \lambda. \end{aligned}$$

We can now subtract the previous two expressions and substitute in (48) to obtain

$$\begin{aligned} F_\tau(\mu) = F_\tau(\mu_*) + &\int_{\Omega\times\Omega} K(\xi,\bar{\xi})(\mu_t(d\xi) - \mu_*(d\xi))(\mu_t(d\bar{\xi}) \\ &- \mu_*(d\bar{\xi})) + \tau D_{KL}(\mu\|\mu_*), \end{aligned} \tag{49}$$

where

$$D_{KL}(\mu\|\mu_*) := \begin{cases} \int_\Omega \mu\log\left(\frac{\mu}{\mu_*}\right) & \text{if } \mu \ll \mu_*, \\ +\infty & \text{if } \mu \not\ll \mu_*, \end{cases}$$

is the relative entropy (also called Kullback-Leibler divergence) of μ with respect to μ_*. Note that the middle term in (49) is always non-negative (since K is positive semi-definite) and that $D_{KL}(\mu\|\nu) \geq 0$ with equality if and only if $\mu = \nu$. In particular, $F(\mu) \geq F(\mu_*)$ with equality if and only if $\mu = \mu_*$.

Hence, besides obtaining a nice expression for F_τ in terms of a stationary solution, (49) also shows that stationary solutions are unique and they coincide with the unique minimizer of the functional (45).

4.2 The Porous Medium Regularization

Another possible regularization, that has been much less studied in this context, is the one arising from the porous medium equation.

In this case, we consider the functional

$$F_\tau(\mu) = \frac{1}{2}\int_D \left(\int_\Omega \Phi(\xi, x)\mu(d\xi) - f(x)\right)^2 dx + \int_\Omega V(\xi)\,\mu(d\xi) + \frac{\tau}{2}\int_\Omega \mu^2 \tag{50}$$

for some small parameter $\tau > 0$ (again, $\int_\Omega \mu^2 = +\infty$ by definition if μ is not absolutely continuous). Then, the Wasserstein gradient flow is given by

$$\partial_t \mu_t = \operatorname{div}(\mu_t \nabla \mathcal{L}(\mu_t)) + \operatorname{div}(\mu_t \nabla \mathcal{S}) + \operatorname{div}(\mu_t \nabla V) + \tau \operatorname{div}(\mu_t \nabla \mu_t), \tag{51}$$

(with the analogous Neumann boundary condition when Ω is not $\mathbb{R}^{d+1}$, cf. (24) and (47)).

In this context, one still expects nice properties of the corresponding evolution of the gradient flow, consistent with those in the porous medium equation [17]. In particular, any stationary solution should have full support (since the support increases with time, up until covering the whole domain). And the same reasoning as in the case of the heat regularization (which was based on the full support of a stationary solution μ_*) applies, and we get

$$\begin{aligned} F_\tau(\mu) = F(\mu_*) + \int K(\xi, \bar{\xi})(\mu_t(d\xi) - \mu_*(d\xi))(\mu_t(d\bar{\xi}) - \mu_*(d\bar{\xi})) \\ + \frac{\tau}{2}\int_\Omega (\mu - \mu_*)^2, \end{aligned} \tag{52}$$

which is similar to (49), where the relative entropy is substituted by the L^2 distance. Thus, from (52), we also get the uniqueness of stationary solutions (and hence, they coincide with the unique minimizer).

Remark 4.1 The two previous regularizations also make sense in the (ρ, H) formulation setting. In particular, one could also add a Laplacian or porous medium term to the PDE transporting ρ in (32), (34), or (36), in order to obtain improved convergence properties.

4.3 An Observation Without Regularization

We can also rewrite the functional F in (16) in terms of a local minimizer (thus removing the explicit dependence on f in its expression), even in the case without regularization.

That is, let μ_* be a local minimizer for F. In particular, it is a stationary point, and it satisfies[11]

$$\frac{\delta F}{\delta \mu}(\mu_*) \equiv \lambda \qquad \text{in} \quad \operatorname{supp}(\mu_*). \tag{53}$$

Moreover, from the local minimality condition, we also have[12]

$$\frac{\delta F}{\delta \mu}(\mu_*) \geq \lambda \qquad \text{in} \quad \Omega. \tag{54}$$

So, combining (53)–(54), and proceeding as in the regularized cases, we obtain

$$F(\mu_t) - F(\mu_*) \geq \int \left(\int \Phi(\xi, x)(\mu_t(d\xi) - \mu_*(d\xi)) \right)^2 dx.$$

[11] To see this, take $\varphi \in C_c^\infty(\Omega)$ with $\int_\Omega \varphi(\xi)\mu_*(d\xi) = 0$, and for $|\varepsilon| \ll 1$, we consider the variation $\mu_\varepsilon := (1 + \varepsilon\varphi)\mu_* \in \mathscr{P}(\Omega)$. Then, by local minimality, we get

$$0 = \frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} F(\mu_\varepsilon) = \int_\Omega \frac{\delta F}{\delta \mu}(\mu_*)(\xi)\, \varphi(\xi)\, \mu_*(d\xi) \qquad \forall\, \varphi \in C_c^\infty(\Omega) \text{ s.t. } \int_\Omega \varphi(\xi)\mu_*(d\xi) = 0.$$

By the arbitrariness of φ, this implies that $\frac{\delta F}{\delta \mu}(\mu_*)$ is constant on $\operatorname{supp}(\mu_*)$.

[12] To see this, given $\nu \in \mathscr{P}(\Omega)$, for $\varepsilon \in [0, 1]$, we consider the variation $\mu_\varepsilon := (1 - \varepsilon)\mu_* + \varepsilon\nu \in \mathscr{P}(\Omega)$. Then, by local minimality, we get

$$\begin{aligned} 0 \leq \frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} F(\mu_\varepsilon) &= \int_\Omega \frac{\delta F}{\delta \mu}(\mu_*)(\xi)\nu(d\xi) - \int_\Omega \frac{\delta F}{\delta \mu}(\mu_*)(\xi)\mu_*(d\xi) \\ &= \int_\Omega \frac{\delta F}{\delta \mu}(\mu_*)(\xi)\nu(d\xi) - \lambda = \int_\Omega \left[\frac{\delta F}{\delta \mu}(\mu_*)(\xi) - \lambda \right] \nu(d\xi), \end{aligned}$$

where the second equality follows from (53). By the arbitrariness of $\nu \in \mathscr{P}(\Omega)$, this implies that $\frac{\delta F}{\delta \mu}(\mu_*)$ is everywhere greater than or equal to λ.

In particular we recover the uniqueness of local minimizers under Φ, that we already knew by Remark 2.1.

5 Open Questions

We conclude this manuscript by discussing some open questions that we believe to have a mathematical interest.

5.1 Regularity and Convergence

One of the main open questions is concerned to the convergence properties of our gradient flows and its relation to the discrete version of the gradient descent. The main currently known results in this direction can be found (and referenced) in [4], where the authors are able to prove the consistency between the many neuron limits and the Wasserstein gradient flow as time goes to infinity, whenever such limits exist. Nonetheless, many questions remain open in this setting, starting from a quantitative (uniform) convergence to the Wasserstein gradient flow, in the limits as $N \to \infty$ and $t \to \infty$. Furthermore, the results in [4] use the specific (homogeneous) structure of the activation function. Thus, the results included in [4] in more general settings remain open, even if one assumes discriminating smooth kernels.

Concerning the continuous formulation (20), this PDE pose a series of interesting challenges. For example:

(i) What are reasonable assumptions on μ_0 and the data, to expect a conservation of its smoothness over time? (That is, to avoid convergence in finite and/or infinite time to a singular measure.)

(ii) It looks likely to us that one can prove a qualitative rate of convergence, using, for instance, the approach in [3]. More challenging and relevant in this setting is to obtain quantitative convergence rates. Such quantification seems far from being easy in the μ formulation case, where one would need to find an "entropy-entropy dissipation inequality," showing that the dissipation (22) controls $F(\mu_t) - F(\mu_*)$, at least when the μ_t is close to the minimizer μ_*.

(iii) Even in the regularized cases (46) or (51), finding quantitative rates of convergence is an interesting open problem.[13]

[13] Consider for simplicity the PDE (46) with $V \equiv 0$. Then, assuming that for t large μ_t is close in some strong sense to the stationary state μ_* and that μ_* is smooth and has full support, then one can get an inequality of the form

$$\frac{d}{dt} F_\tau(\mu_t) \le -c \left(F_\tau(\mu_t) - F_\tau(\mu_*) + \tau \mathcal{F}(\mu_t, \mu_*)\right)^2$$

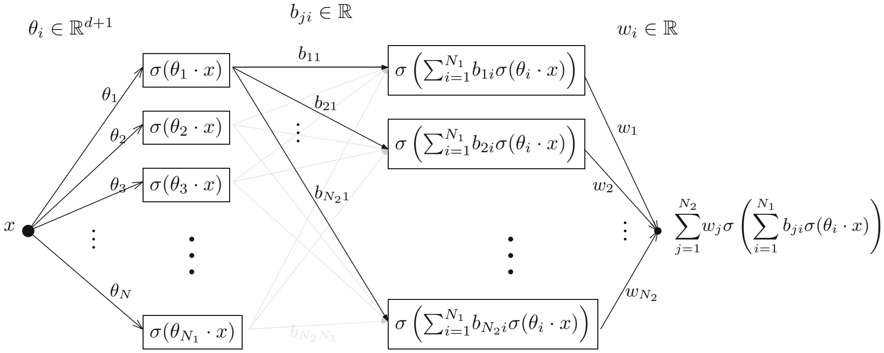

$\theta_i \in \mathbb{R}^{d+1}$
$b_{ji} \in \mathbb{R}$
$w_i \in \mathbb{R}$
x
θ_1
θ_2
θ_3
θ_N
$\sigma(\theta_1 \cdot x)$
$\sigma(\theta_2 \cdot x)$
$\sigma(\theta_3 \cdot x)$
$\sigma(\theta_{N_1} \cdot x)$
b_{11}
b_{21}
$b_{N_2 1}$
$b_{N_2 N_1}$
$\sigma\left(\sum_{i=1}^{N_1} b_{1i}\sigma(\theta_i \cdot x)\right)$
$\sigma\left(\sum_{i=1}^{N_1} b_{2i}\sigma(\theta_i \cdot x)\right)$
$\sigma\left(\sum_{i=1}^{N_1} b_{N_2 i}\sigma(\theta_i \cdot x)\right)$
w_1
w_2
w_{N_2}
$\sum_{j=1}^{N_2} w_j \sigma\left(\sum_{i=1}^{N_1} b_{ji}\sigma(\theta_i \cdot x)\right)$

Fig. 2 Graphic representation of the approximating functions given by what is known as a *two-layer neural network* (55)

5.2 *Multilayer Neural Networks*

Training a multilayer neural network corresponds to the approximation problem of a given function $f \in L^2(D)$, where the approximating functions are obtained by iterations of the construction in (1).

Assume for simplicity that $h(\theta, x) = \sigma(\theta \cdot x)$ and ignore the independent term (i.e., $n = d$ and $\theta^{(d)} = 0$; see (2)). Then, in the two-layer case, given an input $x \in \mathbb{R}^n$, we want to approximate a given output $f(x)$ through a neural network with two hidden layers, consisting of N_1 and N_2 neurons each. Let us denote the parameters in this case as $\{w_j\}_{1 \le j \le N_2}$ with $w_i \in \mathbb{R}$, $\{\theta_j\}_{1 \le j \le N_1}$ with $\theta_j \in \mathbb{R}^n$, and $\{b_{ji}\}_{1 \le j \le N_2, 1 \le i \le N_1}$ with $b_{ji} \in \mathbb{R}$. The corresponding approximating function is then given by

$$\sum_{j=1}^{N_2} w_j \sigma\left(\sum_{i=1}^{N_1} b_{ji} \sigma(\theta_i \cdot x)\right), \tag{55}$$

for some activation function $\sigma : \mathbb{R} \to \mathbb{R}$ (see Fig. 2 and compare with Fig. 1). Thus, we want to optimize the parameters in order to minimize a functional of the form

$$\frac{1}{2} \int_D \left(\sum_{j=1}^{N_2} w_j \sigma\left(\sum_{i=1}^{N_1} b_{ji} \sigma(\theta_i \cdot x) \right) - f(x) \right)^2 dx.$$

for some suitable function $\mathcal{F}(\mu_t, \mu_*)$ such that $\mathcal{F}(\mu_t, \mu_*) \to 0$ as $t \to \infty$. This suggests a rate of convergence of the form $F_\tau(\mu_t) - F_\tau(\mu_*) \sim \frac{1}{t}$, at least in the regularized case.

The corresponding expression of the previous functional in the (appropriate) limit $N_1, N_2 \to \infty$ is an interesting open problem, and some possible interpretations have recently been suggested in [2, 12, 13, 16]. However, a simple unified connection between multiple layers neural networks and Wasserstein gradient flows, as the one presented in this paper, seems to be missing.

In this direction, it might be worth mentioning that the (ρ, H)-approach seems more adequate when dealing with systems in which one needs to consider separately each of the layers: already in the single layer case, the (ρ, H)-formulation is the one that takes advantage of the structure of the activation functions $w\, h(\theta, x)$. Even there, however, one does not fully take advantage of the linear structure of $h(\theta, x) = \sigma(\theta \cdot x)$ inside the function σ.

References

1. Ambrosio, L., Gigli, N., Savare, G.: Gradient Flows in Metric Spaces and in the Space of Probability Measures. Lectures in Mathematics. Springer, Berlin (2008)
2. Araújo, D., Oliveira, R.I., Yukimura, D.: A mean-field limit for certain deep neural networks. Preprint arXiv https://arxiv.org/abs/1906.00193
3. Carrillo, J., Gvalani, R., Wu, J.: An invariance principle for gradient flows in the space of probability measures. Preprint arXiv https://arxiv.org/abs/2010.00424
4. Chizat, L., Bach, F.: On the global convergence of gradient descent for over-parameterized models using optimal transport. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
5. E, W.: A proposal on machine learning via dynamical systems. Commun. Math. Stat. **5**, 1–11 (2017)
6. E. W., Han, J., Li, Q.: A mean-field optimal control formulation of deep learning. Res. Math. Sci. **6**, 10 (2019)
7. E, W., Ma, C., Wu, L.: Machine learning from a continuous viewpoint, I. Sci. China Math. **63**, 2233–2266 (2020)
8. Figalli, A., Glaudo, F.: An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows, EMS Textbooks in Mathematics, 144 pp. European Mathematical Society (EMS), Zürich (2021)
9. Javanmard, A., Mondelli, M., Montanari, A.: Analysis of a twolayer neural network via displacement convexity. Ann. Statist. **48**, 3619–3642 (2020). ArXiv version: https://arxiv.org/abs/1901.01375
10. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker-Planck equation. SIAM. J. Math. Anal. **29**, 1–17 (1998)
11. Mei, S., Montanari, A., Nguyen, P.: A mean field view of the landscape of two-layer neural networks. PNAS **115**, 7665–7671 (2018). ArXiv version: https://arxiv.org/abs/1804.06561
12. Nguyen, P.-M.: Mean field limit of the learning dynamics of multilayer neural networks. Preprint arXiv https://arxiv.org/abs/1902.02880
13. Nguyen, P.-M., Pham, H.-T.: A rigorous framework for the mean field limit of multilayer neural networks. Preprint arXiv https://arxiv.org/abs/2001.11443
14. Rotskoff, G., Vanden-Eijnden, E.: Neural networks as interacting particle systems: asymptotic convexity of the loss landscape and universal scaling of the approximation error (2018). Preprint arXiv

15. Sirignano, J., Spiliopoulos, K.: Mean field analysis of neural networks: a law of large numbers. SIAM J. Appl. Math. **80**(2), 725–752 (2020)
16. Sirignano, J., Spiliopoulos, K.: Mean field analysis of deep neural networks. Math. Oper. Res. **47**(1), 120–152 (2022)
17. Vázquez, J.L.: The Porous Medium Equation: Mathematical Theory. Oxford Lecture Series in Mathematics and its Applications, vol. 33. Oxford University Press, Oxford (2006)

On the Origins, Nature, and Impact of Bourgain's Discretized Sum-Product Theorem

Alexander Gamburd

Abstract We discuss the origins, nature, and development of the discretized sum-product theorem, a result which Jean Bourgain viewed as one of his most significant.

> There are two labyrinths of the human mind: one concerns the composition of the continuum, and the other the nature of freedom, and both spring from the same source – the infinite.
>
> Baron von Leibniz

> During World War II, when von Neumann was working on the design of nuclear weapons, he came to the conclusion that analytical methods were inadequate to the task, and that the only way to deal with equations of continuum mechanics is to discretize them. . . . It is to this task that von Neumann devoted his energies after the war.
>
> Peter Lax

1 Overture

Baron Bourgain, the IBM von Neumann Professor in the School of Mathematics at the Institute for Advanced Study (IAS), is one of the most original, penetrating,

This essay, dedicated to the memory of Jean Bourgain, is an augmented and expanded version of *Singular Adventures of Baron Bourgain in the Labyrinth of the Continuum*, which appeared in the Notices of the AMS, **67**, 2020, 1716–1733. The opening image is coat of arms of Jean Bourgain, who was bestowed the title of Baron by King Philippe of Belgium in July 2015.

A. Gamburd (✉)
Department of Mathematics, CUNY Graduate Center, New York, NY, USA

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_4

and versatile analytical minds of our troubled times, justly celebrated[1] and revered without reservations.

IN SPEM CONTRA SPEM

While he rejected outright the suggestion of a sixtieth birthday conference, a proposal to have a gathering occasioned by the publication of his 500th paper was

[1] An excerpt from Bourgain's interview upon receiving the *2017 Breakthrough Prize in Mathematical Sciences* concludes this essay.

The following quote is from *The Work of Jean Bourgain* by Luis Caffarelli, Proceedings of ICM, 1994 [23] (the year Bourgain was awarded the Fields Medal): "Bourgain's work touches on several central topics of mathematical analysis: the geometry of Banach spaces, convexity in high dimensions, harmonic analysis, ergodic theory, and, finally, nonlinear partial differential equations from mathematical physics. In all of these areas, he made spectacular inroads into questions where progress has been blocked for a long time. This he did by simultaneously bringing into play different areas of mathematics: number theory, combinatorics, probability, and showing their relevance to the problem in the previously unforeseen fashion. . . . Some of the outstanding qualities of Bourgain are his power to use whatever it takes—number theory, probabilistic methods, covering techniques, sharp decompositions – to understand the problem at hand, and his versatility, which allowed him to deeply touch so many areas in such a short period of time."

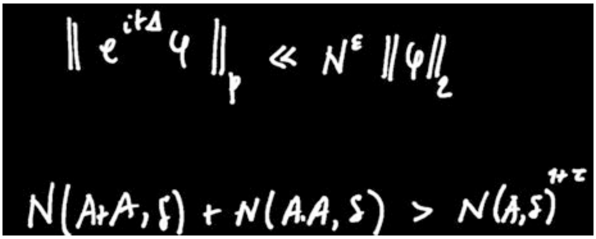

$\| e^{it\Delta} \varphi \|_p \ll N^\varepsilon \|\varphi\|_2$
$N(A+A, \delta) + N(A.A, \delta) > N(A,\delta)^{1+\tau}$

Fig. 1 Two of Jean Bourgain's signature results

not immediately dismissed—the conference *Analysis and Beyond: Celebrating Jean Bourgain's Work and Impact* took place at the IAS in Princeton on May 21–24, 2016. The conference talks (all of which were videotaped) are a tribute to the depth and breadth of Bourgain's work and its singular and transcendent impact on the whole of our discipline. The beauty and power of the first result highlighted by Jean's hand (Fig. 1) on the conference poster $\|e^{it\Delta}\varphi\|_p \ll N^{\varepsilon}\|\varphi\|_q$ is apparent from reading the splendid paper by Andrea Nahmod in the Bulletin of the American Mathematical Society (BAMS), [71]. The brief of this paper is to explicate the origins, nature, and development of the second result, *the discretized sum-product inequality*

$$\mathcal{N}(A + A, \delta) + \mathcal{N}(A \cdot A, \delta) > \mathcal{N}(A, \delta)^{1+\tau}, \tag{1}$$

in analysis and beyond.

***

> The three great branches of mathematics are, in historical order, Geometry, Algebra and Analysis. Geometry we owe essentially to Greek civilization, Algebra is of Indo-Arab origin and Analysis (or Calculus) was the creation of Newton and Leibniz, ushering in the modern era.
>
> Sir Michael Atiyah [1]

Von Zahlen und Figuren—"On Numbers and Shapes"[2] is the title of one of the most successful expositions of mathematics aimed at a broad audience, reflecting a common perception of our discipline as a marriage between Algebra and Geometry. This happy marriage, notwithstanding Count Tolstoy's contention ("All happy marriages are alike; each unhappy marriage is unhappy in its own way."), is not without tensions (as, perhaps, each happy marriage—including, possibly, bicameral mind—is in its own way). "In these days the angel of topology and the devil of abstract algebra fight for the soul of each individual mathematical domain" is the

[2] The book was written in 1933 by Hans Rademacher and Otto Toeplitz, two outstanding analysts of the past century, who made a deliberate decision not to refer in their exposition to the analysis (or Calculus) of Leibniz and Newton. The English translation is entitled "The Enjoyment of Math."

way Hermann Weyl[3] put it; three score and seven years later, in conversation at Google with the company's CEO, a somewhat divergent sentiment was expressed: "When you form your ideas on the basis of words, you build from concepts, which to be meaningful depend on relation to other concepts. When you form your ideas on the basis of pictures, you form your views on the basis of impressions and of moods, that cannot even be recreated very easily, so you cannot look back and check what it was that impressed you so much."[4]

This tension is embodied in the system of real numbers, the soil in which the functions of Analysis grow, resembling Janus's head facing in two directions: on the one hand, it is the field closed under the operations of addition and multiplication; on the other hand, it is a continuous manifold the parts of which are so connected as to defy exact isolation from each other. The one is algebraic; the other is the geometric face of real numbers. Continued fractions are much more intrinsic and geometric forms of discretizing the continuum; the lack of a practical algorithm for their addition and multiplication leads to the regnancy of the discretization based on the ordinary (digital or decimal, i.e., base 10) fractions.

Whereas Newton, in his development of Calculus, was primarily motivated by "dynamics" (force, acceleration), as exemplified by the falling of the apple on his head, Leibniz, it appears, was more intrigued by what would now be described by the appellation "fractal geometry of nature." "Imagine a circle; inscribe within it three other circles congruent to each other and of maximum radius; proceed similarly within each of these circles and within each interval between them, and imagine that the process continues *ad infinitum*,"wrote Leibniz referencing configuration akin to the four mutually tangent circles appearing on Baron Bourgain's coat of arms. Leibniz's definition of the straight line as a 'curve, any part of which is similar to the whole, and it alone has this property, not only among curves but among sets' is a reflection of the fractal nature of the continuum: the Cantor set would satisfy Leibniz's definition.[5]

Dynamics, broadly conceived, is perceived as a study of change, which in its primordial (physical) context takes place within time. The Cantor set (and $\mathbb{R}$) is, so to speak, timeless, i.e., static in time, but there is "a condition of possibility" of (almost) "equiprimordial" change "in the eye of the beholder," taking form in changing the degree of magnification scale and "zooming in." This is reflected in the "multi-scale" nature of Bourgain's proof(s) of (1).

[3] In *Invariants*, Duke Mathematics Journal **5** 1939, anticipating by 4 years an even more sweeping assertion, due to Jean-Paul Sartre: "L'enfer, c'est les autres."

[4] Henry Kissinger.

[5] Leibniz also wrote the first textbook on combinatorics *Dissertatio de arte combinatoria* and invented the binary notation, which made possible modern computers and will play an important role in navigating the labyrinth of Bourgain's argument.

The first collection of Leibniz's works was published in 1735 by Rudolf Erich Raspe, better known today for his authorship of *Singular Adventures of Baron Munchausen.*

To bring this opening section to a close, let us in passing note that both results chosen by Jean are not equalities (inequalities, rather), commenting thus:

If Algebra is generally perceived as the study of equations, what perhaps lies at the heart of Analysis are inequalities, or estimates, which compare the size of two quantities or expressions. Einstein's discovery that nothing travels faster than light is an example of an inequality. The inequality 2^X *is considerably larger than* X arguably neatly encapsulates both the P vs NP problem (properly stated for finite X) and Cantor's continuum problem (when X is the first infinite ordinal). An elementary inequality, taught in the middle school, asserts that the arithmetic mean of two positive numbers is never less than their geometric mean. In between these two extremes there is a vast range of estimates of great variety and importance. Such estimates, reflecting and quantifying some subtle aspect of the underlying problem, are often exceedingly difficult to prove. It will be seen that for the inequality (1), with which we are about to get intimate, the underlying issue lies at the heart of the tension between the algebraic and (fractal)-geometric nature of the continuum. Fractal derives from Latin *fractus*, meaning broken apart; algebra derives from the Arabic *al-jabr*, meaning the reunion of broken parts.

2 Origins: Kakeya-Besicovitch Problem+

> It is difficult and often impossible to judge the value of a problem correctly in advance; for the final award depends upon the gain which science obtains from the problem. Nevertheless we can ask whether there are general criteria which mark a good mathematical problem. An old French mathematician said: 'A mathematical theory is not to be considered complete until you have made it so clear that you can *explain it to the first man whom you meet on the street.*' This clearness and ease of comprehension, here insisted on for a mathematical theory, I should still more demand for a mathematical problem if it is to be perfect; for what is clear and easily comprehended attracts, the complicated repels us.
>
> David Hilbert, *Problems of Mathematics*, 1900

In Hilbert's[6] democratic dictum, if followed by Sōichi Kakeya (writing the paper on an island nation in 1917, at the height of the Great War), the explanation of the problem now bearing his name to almost every person at just about any street in Eastern Eurasia might have run as follows: Entrusted with defending an island, possessing a huge hill, cragged and steep, your task is to purchase at the least cost to the nation's treasury, a plot of land on the flat hilltop with the following property—a cannon of length one must be capable of pointing in any direction.

Kakeya improved by a factor of one-half the obvious solution (a circle of diameter one, having area $\frac{\pi}{4}$); his proposed shape (three-cusped hypocycloid

[6] Hilbert's paper on Dirichlet's Principle is one of the two referenced by Kakeya [48]; the second one (also on Dirichlet's principle) is by Caratheodory, a student of Hilbert. In his magnificent book *Geometry and Imagination*, Hilbert refers to Besicovitch's result (described below) as "showing that this [Kakeya] problem has no solution."

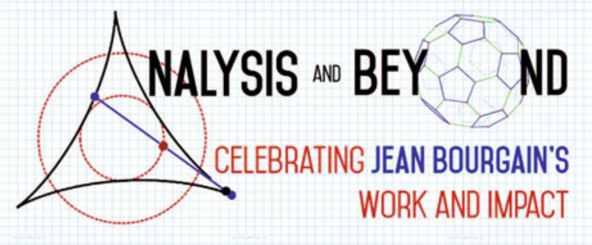


Fig. 2 Analysis and beyond

inscribed in the circle of radius 1) is alluded to in the rendering of A in the conference poster (Fig. 2). In the same year, working in Perm,[7] while the October/November Russian/Soviet Revolution was unfolding, A. S. Besicovitch reduced the minimal necessary sum to virtually[8] nothing.

In fact, Besicovitch was working on the following question: if f is a Riemann integrable function defined on the plane, is it always possible to find a pair of orthogonal coordinate axis with respect to which $\int f(x, y)dx$ exists as a Riemann integral for all y, and with resulting function of y also Riemann integrable? Besicovitch noticed that if he could construct a compact set F of plane Lebesgue measure zero containing a line segment in every direction, this would lead to a counterexample as follows. Assume (by translating F if necessary) that F contains no segment parallel to and of rational distance from either of a fixed pair of axes. Let f be the characteristic function of the set F_r consisting of those points of F with at least one rational coordinate. As F contains a segment in every direction on which both F_r and its complement are dense, there is a segment in each direction in which f is not Riemann integrable. On the other hand, the set of points of discontinuity of F is of plane measure zero, so f is Riemann integrable over the plane by the well-known criterion of Lebesgue.

The basic idea underlying the original construction of Besicovitch [5] is to form a figure obtained by splitting an equilateral triangle of unit height into many smaller triangles of the same height by dividing up the base and then sliding these elementary triangles varying distances along the base line. In 1964 Besicovitch developed a completely different approach [6], using the projection theorem due to Marstrand.

[7] Subsequently Molotov (1940–1957); currently Perm.

[8] The virtual collapse of the Russian currency appears to have had nothing to do with it. In 1924, together with Tamarkin, Besicovitch crossed the Soviet border with Norway on foot and made his way to Copenhagen to work with H. Bohr, eventually settling in Cambridge in 1927, where, in due course, he became the Rouse Ball Chair. Besicovitch's command of English remained stationary from his early days in Cambridge ("It's a story..."); for him, for example, the definite article was superfluous. A story is told that during one of his lectures, an undergraduate tittered at some distortion of English idiom. "Gentlemen,' said Besicovitch, "there are 50 million Englishmen speak English you speak; there are 500 million Russians speak English I speak." [22]

2.1 Some Fundamental Properties of Plane Sets of Fractional Dimension

In this 1954 paper [67], which was essentially the work for his doctoral thesis at Oxford and was heavily influenced by Besicovitch, John Marstrand proved the following fundamental result.

Theorem 1 (Marstrand's Projection Theorem) *Denote the projection in the direction θ by π_θ. If $X \in \mathbb{R}^2$ is a Borel subset of Hausdorff dimension s, then* $\dim_H(\pi_\theta X) = \min(s, 1)$ *for almost every θ.*

Concerning the finer information about the set of exceptional θ in Theorem 1, Kaufman proved [51] that if $\dim X \geq t$, $B \subset S^1$ with $\dim B > t$, then there exists $\theta \in B$ such that $\dim(\pi_\theta(X) \geq t$. Using crucially (1), in *The Discretized Sum-product and Projection Theorems* [14], Bourgain established the following, sharper result:

Theorem 2 *Given* $0 < \alpha < 2$ *and* $\kappa > 0$*, there is* $\eta > \frac{\alpha}{2}$ *such that if* $X \subset \mathbb{R}^2$ *is of Hausdorff dimension greater than α, then* $\dim_H(\pi_\theta(X)) \geq \eta$ *for all* $\theta \in S^1$ *except in an exceptional set E satisfying* $\dim_H(E) \leq \kappa$.

2.2 Besicovitch Type Maximal Operators and Applications to Fourier Analysis

> We must admit with humility that, while number is purely a product of our mind, space has a reality outside of our mind, so that we cannot prescribe its laws a priori.
>
> Gauss, Letter to Bessel, 1830

The Kakeya problem in $\mathbb{R}^n$ is to estimate the fractal dimension of the Besicovitch set $E \subset \mathbb{R}^n$, i.e., a set containing line segments of length one in all directions.

Conjecture 1 *Let E be a Besicovitch set in* $\mathbb{R}^n$*. Then* $\beta(n) = \dim(E) = n$.

There are several relevant notions of "fractal dimension," the simplest being the Minkowski dimension, defined as follows. Let A be a closed subset of a metric space X. Fix some radius δ. Let $\mathcal{N}(A, \delta)$ be the least number of balls of radius δ needed to cover A. If A is a rectifiable curve in $\mathbb{R}^n$, it is easy to see that $\mathcal{N}(A, \delta)$ is of order δ^{-1}. If A is a surface, $\mathcal{N}(A, \delta)$ is approximately δ^{-2}. This suggests the idea of defining the dimension of an arbitrary set as the number d for which $\mathcal{N}(A, \delta) \sim \delta^{-d}$. The limit

$$\lim_{\delta \to 0} \frac{\log \mathcal{N}(A, \delta)}{\log(\delta^{-1})},$$

if it exists, is called Minkowski dimension, $\dim_M(A)$.

The basic result proved by Davies [27] in 1971 is that $\beta(2) = 2$. The same year C. Fefferman [37] discovered the intimate connection between the Kakeya problem and the multiplier problem for the ball, proving that for $d \geq 2$ the map $f \to \int_{|\xi|\leq 1} \hat{f}(\xi)e^{ix\xi}d\xi$ defines only for $p = 2$ a bounded operator on $L^p(\mathbb{R}^d)$. This seminal result made apparent the fundamental connection between Kakeya-type questions and the higher-dimensional Fourier analysis, in particular in the theory of oscillatory integral operators.[9]

In the 1980s, Drury [31] showed that

$$\beta(n) \geq \frac{n+1}{2} \tag{2}$$

(see also Christ et al. [25]). The argument consists of intersecting the line segment $L_\xi \subset E$, L_ξ parallel to ξ in S^{d-1} by a pair of parallel hyperplanes H_1, H_2 in $\mathbb{R}^d$ and observing that for all $\delta > 0$

$$\left(\frac{1}{\delta}\right)^{d-1} \lesssim \mathcal{N}(H_1 \cap E, \delta)\mathcal{N}(H_2 \cap E, \delta). \tag{3}$$

The estimate (2) was first improved by Bourgain in 1991, in the paper eponymous with the title of this subsection [8], to $\frac{n+1}{2} + \varepsilon_n$ with ε_n given by a recursive argument (for $n = 3$, this yields bound $\frac{7}{3}$) by using a "bush" argument. A more efficient geometric argument, using "hairbrushes," was given several years later by T. Wolff, leading to

$$\dim_H(E) \geq \frac{n}{2} + 1. \tag{4}$$

The space constraints prevent me from going into the details of these arguments; referring the reader to beautiful surveys by Izabella Łaba [55], Terence Tao [89], and Thomas Wolff [95], I will restrict myself to two remarks.

The first remark is that these developments made apparent the connection between Kakeya-type problems and results in combinatorial geometry, such as the *Szemerédi-Trotter Theorem* [88], which will be briefly discussed in Sect. 3.2.

[9] A recent triumph in this area is the resolution of the Vinogradov's conjecture by Bourgain, Demeter, and Guth [15], establishing near-optimal bounds on the mean values of exponential sums such as

$$\sum_{n=1}^{N} e^{2\pi i(\alpha_1 n + \alpha_2 n^2 + \ldots \alpha_k n^k)}$$

as one varies the frequencies $\alpha_1, \ldots \alpha_k$; these are of fundamental importance in analytic number theory.

The second remark is that Bourgain's interest in Kakeya problem was stimulated by his discovery [9] of it being implied by the following version of Montgomery's conjecture[10] for Dirichlet polynomials:

Conjecture 2 *Let* $S(s) = \sum_{n=1}^{N} a_n n^s$ *with* $|a_n| \leq 1$ *and* $\mathcal{F}$ *be a set of* 1*-separated reals in the interval* $[0, T]$*,* $T > N$*. Then*

$$\sum_{t \in \mathcal{F}} |S(it)|^2 \ll T^{\varepsilon}(N + |\mathcal{F}|)N(\max_{1 \leq n \leq N} |a_n|^2). \tag{6}$$

Regrettably skipping thus over many important and pertinent developments that took place in the last decade of the past century, let us note, looking forward, that in its closing year (1999), Bourgain unveiled the connection between Kakeya problem and one of the most consequential and far-reaching results in arithmetic combinatorics, obtained by Gowers in his groundbreaking *A New Proof of Szemerédi's Theorem for Arithmetic Progressions of Length Four* [40]. This result, Balog-Szemerédi Gowers Lemma, will play a crucial role in many a subsequent development, of which some are discussed in this essay.

2.3 Balog-Szemerédi-Gowers Lemma

> Either this universe is a mere confused mass, and an intricate context of things, which shall in time be scattered and dispersed again; or it is a union consisting of order and administered by Providence.
>
> Marcus Aurelius "Meditations" 6, VIII

> Complete disorder is impossible.
>
> T.S. Motzkin

The Balog-Szemerédi-Gowers lemma is ostensibly a statement about group structure, but the main tool in its proof is a remarkable (and remarkably useful) graph-theoretic result best viewed in the context of Ramsey theory. Ramsey theory is a systematic study of the following general phenomenon. Surprisingly often, a large structure of a certain kind has to contain a fairly large highly organized substructure, even if the structure itself is completely arbitrary and apparently chaotic. It can be viewed as a vast generalization of the pigeonhole principle, which states that if a set X of n objects is colored with S colors, then there must be a subset of X of

[10] One of the consequences of Montgomery's conjecture is the density hypothesis for the Riemann zeta function

$$N(\sigma, T) \ll T^{2(1-\sigma)+\varepsilon}. \tag{5}$$

Here $\frac{1}{2} < \sigma < 1$, $T > 0$ and $N(\sigma, T)$ is the number of zeros $\rho = \beta + i\gamma$ of $\zeta(s)$ satisfying $\beta > \sigma$, $|\gamma| < T$.

size at least $\frac{n}{s}$ that uses just one color. Such a subset is called monochromatic. The situation becomes more interesting if the set X has some additional structure. It then becomes natural to ask for a monochromatic subset that keeps some of the structure X. However it also becomes much less obvious if such a subset exists. Frank Plumpton Ramsey in 1930 [76] took as his set X the set of all the edges in a complete graph and the monochromatic subset he obtained consisted of all the edges of some complete graph. One version of his theorem is as follows. For every positive integer k, there is a positive integer N such that if the edges of the complete graph are all colored either red or blue, then there must be k vertices such that all edges joining them have the same color. That is, a *sufficiently large* complete graph colored with two colors contains a complete subgraph of size k which is monochromatic. The least integer N that works is known as $R(k)$ and is known that

$$2^{\frac{k}{2}} \leq R(k) \leq 2^{2k}. \tag{7}$$

There were several results in Ramsey theory predating Ramsey's theorem; in particular, van der Waerden [93] proved that if you color the integers with some finite number r of colors, there must be some color that contains arithmetic progressions of every length. In 1935 Erdös and Turán conjectured that this holds for "the most popular" color class. More precisely, they conjectured that for any positive integer k and any real number $\varepsilon > 0$, there is a positive integer n_0 such that if $n > n_0$, any set of at least εn positive integers between 1 and n contains k-term arithmetic progression. This conjecture was proved by Szemerédi in 1975 using, among other things, his celebrated regularity lemma [87], which can be very roughly described as a statement that even the most "chaotic" systems can be decomposed into a "relatively" small number of "approximately regular" subsystems.

Using the Szemerédi regularity lemma, the following result was established by Balog and Szemerédi in 1994 [2], resulting in tower-like exponential-type dependence (cf. (7)). Gowers achievement of the polynomial bounds $K^{O(1)}$ in the statement below is crucial in the ensuing applications.

Theorem 3 (Balog-Szemerédi-Gowers Lemma) *Let $\mathcal{G}(A, B, E)$ be a finite bipartite graph, that is, a graph whose vertices can be partitioned into two disjoint sets, with $|E| \geq \frac{|A||B|}{K}$. Then there exist subsets $A' \subset A$ and $B' \subset B$ with $|A'| \gg K^{-O(1)}|A|$ and $|B'| \gg K^{-O(1)}|B|$ such that for every $a \in A$ and $b \in B$, a and b are joined by $\gg K^{-O(1)}|A||B|$ paths of length three.*

The fact that the following corollary is valid for non-commutative groups was established by Tao [90].

Corollary 4 *Let A, B be finite nonempty subsets of a group G and suppose*

$$\|1_A \star 1_B\|_{l^2(G)} \geq \frac{|A|^{\frac{3}{4}}|B|^{\frac{3}{4}}}{K} \tag{8}$$

for some[11] $K \geq 1$. *Then there exist subsets* $A' \subset A$ *and* $B' \subset B$ *with* $|A'| \gg K^{-O(1)}|A|$ *and* $|B'| \gg K^{-O(1)}|B|$ *with* $|A' \cdot B'| \ll K^{O(1)}|A||B|$ *and* $|A' \cdot (A')^1| \ll K^{O(1)}|A|$.

The quantity $\|1_A \star 1_B\|_{l^2(G)}$ counts the number of solutions to the equation $a_1 \cdot b_1 = a_2 \cdot b_2$ with $a_1, a_2 \in A$, and $b_1, b_2 \in B$ (multiplicative or additive quadruples) and is also known as the multiplicative energy of A and B.

2.4 On the Dimension of Kakeya Sets and Related Maximal Inequalities

The main result in this 1999 paper of Bourgain [10] is the following improvement of (4) for large n

$$\dim_H(E) \geq \frac{1}{25}(13n + 12). \tag{9}$$

The heart of the argument consists in applying Balog-Szemerédi-Gowers lemma to show that Kakeya set E satisfies $\mathcal{N}_\delta \geq \delta^{-\alpha(n-1)}$ with $\alpha > \frac{1}{2}$ as follows. Let L be the lattice $\delta\mathbb{Z}^n \subset \mathbb{R}^n$, and for each of the segments $\{x + te \ : \ |t| \leq \frac{1}{2}\}$ with $e \in S^{n-1}$ in the definition of Kakeya set. Let x^+ and x^- be the elements of L closest to $x + \frac{1}{2}e$ and $x - \frac{1}{2}e$, respectively. Let A be the set whose elements are the various x^+ and x^- and define $\mathcal{G} \subset A \times A$ to be the set of pairs (x^+, x^-); then let S be the set of sums $x^+ + x^-$. Clearly $|A| \lesssim \mathcal{N}_\delta(E)$, and in addition $|S| \lesssim \mathcal{N}_\delta(E)$, since the midpoint $\frac{1}{2}(x^+ + x^-)$ is within $C\delta$ of $x \in E$. But it is equally clear that point of $\mathbb{P}^{n-1}$ is within $C\delta$ of some difference $x^+ - x^-$. Thus $\delta^{-(n-1)} \lesssim \mathcal{N}_\delta(E)^{2-\varepsilon}$, as claimed.

This paper marked the first application in Harmonic Analysis of Additive Combinatorics.[12]

[11] Here $\star$ denotes the convolution operation: $f \star g = \int_G f(y)g(y^{-1}x)d\mu(y)$. Note that by Young's inequality, $\|1_A \star 1_B\|_{l^2(G)} \leq |A|^{\frac{3}{4}}|B|^{\frac{3}{4}}$.

[12] "Bourgain's argument was, to this author's knowledge, the first application of additive number theory to Euclidean harmonic analysis. It was significant, not only because it improved Kakeya bounds, but perhaps even more so because it introduced many harmonic analysts to additive number theory, including Tao, who contributed so much to the subject later on, and jump-started interaction and communication between the two communities. The Green-Tao (Fig. 3) theorem and many other developments might never have happened were it not for Bourgain's brilliant leap of thought in 1998." Izabella Łaba, BAMS, 2008 [55].

Fig. 3 Jean Bourgain and Ben Green

3 Sum-Product Phenomena and the Labyrinth of the Continuum

Additive combinatorics grew out of the classical additive number theory. Though few isolated results existed before, the turning point was Schnirelmann's approach [80] to Goldbach's conjecture asserting that any integer greater than three can be expressed as a sum of two or three primes, depending on parity. Schnirelmann proved the weaker result that there is a bound k so that every integer is a sum of at most k primes, or, in other words, the primes form an additive basis. Schnirelmann's approach, notwithstanding it being soon superseded for the Goldbach's problem by Vinogradov's method of exponential sums, kindled the interest in addition of general sets; a result of fundamental and lasting importance in this subject is due to Gregory Abelevich Freiman [38], a student of Gelfond, who was a close friend and collaborator of Schnirelmann.[13]

3.1 Freiman's Theorem and Ruzsa's Calculus

Freiman's Theorem gives characterization of sets with small doubling in terms of generalized arithmetic progression. A d-dimensional generalized arithmetic progression (GAP) is a set P of the form

$$\{a + x_1 q_q + \cdots + x_d q_d \; : \; 0 \leq x_i \leq l_i\}, \tag{10}$$

[13] Schnirelmann committed suicide on 24 September 1938, fearing imminent persecution by NKVD (subsequently KGB; currently FSB).

where $l_1, \dots l_d$ are positive integers. We call d the dimension of P; by the *size* of P, we mean $\|P\| = \prod_{i=1}^{d}(l_i + 1)$, which is the same as the number of elements if all sums in (10) are distinct (in which case we say that P is *proper*). Note that

$$|P + P| < 2^d |P| \leq 2^d \|P\|. \tag{11}$$

Theorem 5 (Freiman's Theorem) *If $A \subset \mathbb{Z}$, $|A| = n$, $|A + A| \leq \alpha n$, then A is contained in a generalized arithmetic progression of dimension at most $d(\alpha)$ and size at most $s(\alpha)n$.*

The quantitative bound in Freiman's theorem, used by Bourgain in his first proof of (1), is due to Mei-Chu Chang (Fig. 4) [24]: $d < \alpha$ (the best possible[14]) and $s \leq e^{\alpha^c}$.

AUG 21 2001

Fig. 4 Jean Bourgain and Mei-Chu Chang

[14] It is known that a bound for s must be $\gg 2^\alpha$; very likely the proper order is $e^{c\alpha}$. A beautiful survey by T. Sanders in BAMS [82] covers the recent developments.

Freiman's proof was considerably simplified by Ruzsa [77] (building on the earlier work of Plünnecke [74]). One of the fundamental notions introduced by Ruzsa is that of *Ruzsa distance* between two sets X and Y in a group, $\rho(X, Y) = \log \frac{|X-Y|}{\sqrt{|X||Y|}}$, allowing us to rewrite an elementary inequality for A, Y, Z finite sets in a group (which, as observed by Tao, is not necessarily commutative) $|A||Y - Z| \leq |A - Y||A - Z|$ as

$$\rho(Y, Z) \leq \rho(Y, A) + \rho(A, Z), \tag{12}$$

a triangle inequality-like property; ρ is also symmetric (but $\rho(X, X)$ is typically positive). The following result of Plünnecke and Ruzsa was used in Bourgain's 2+ proof in place of Freiman's theorem.

Theorem 6 *Let A, B be finite sets in a group and write $|A| = m$, $|A + B| = \alpha m$. For arbitrary nonnegative integers k, l we have*

$$|kB - lB| \leq \alpha^{k+l} m.$$

3.2 Sum-Product Phenomena and Incidence Geometry

Freiman's theorem is an example of an "inverse" result: knowing that the set has small doubling, we can characterize its structure in terms of GAPs. One of the basic "direct" results, applicable to arbitrary sets, is the "sum-product phenomenon," whose elementary and elemental nature might be described as follows. When studying addition and multiplication tables for numbers from one to nine, one might notice that there are many more numbers in the multiplication table. This basically has to do with the fact that the numbers from one to nine form an arithmetic progression. If you take a set forming an arithmetic progression (or a subset of it) and add it to itself, it will not grow much; if you take a set forming a geometric progression (or a subset of it) and multiply it by itself, it will also not grow much. However a subset of integers cannot be both an arithmetic and a geometric progression, and so it will grow either when multiplied or added with itself.

In 1983 Erdös and Szemerédi proved [35] that for any finite set of integers A

$$|A + A| + |A \cdot A| \geq C|A|^{1+\varepsilon} \tag{13}$$

for absolute constants C, ε and conjectured that in fact for any $\varepsilon > 0$ there is C_ε such that

$$|A + A| + |A \cdot A| \geq C_\varepsilon |A|^{2-\varepsilon}. \tag{14}$$

We will give a beautiful proof (due to Elekes [33] and Székeley [86]) of (14) with $\varepsilon = \frac{3}{4}$ using Szemerédi-Trotter theorem in incidence geometry, mentioned in Sect. 2.2, which in turn will follow from crossing number inequality obtained, ultimately, from a purely topological result: Euler's formula.

Crossing Number Inequality

During World War II, Turán worked as forced labor, moving wagons filled with bricks from kilns to storage places. According to his recollections, it was not a very tough job, except that they had to push much harder at the crossings. This led him to consider the following problem: for a non-planar graph $\mathcal{G}$, find a drawing for which the number of crossings is minimal. The minimal number of crossings in a drawing is called crossing number of a graph $\mathrm{Cr}(\mathcal{G})$. Another practical application of this problem appeared in the early 1980s, when it turned out that the chip area required for the realization of an electrical circuit (VLSI layout) is closely related to crossing number of underlying graph. The basic result, due to Leighton [57], is as follows:

$$\mathrm{Cr}(\mathcal{G}) \geq \frac{1}{64}\frac{|E|^3}{|V|^2} - |V|. \tag{15}$$

Here $|V|$ and E denote, respectively, the number of vertices and edges in the graph. The proof starts by observing that Euler's formula implies that if $\mathrm{Cr}(\mathcal{G}) = 0$, then $|V| - |E| + |F| = 2$. This readily implies that crossing number of any graph satisfies

$$\mathrm{Cr}(\mathcal{G}) \geq |E| - 3|V| + 6.$$

The proof is concluded by considering a planar embedding of $\mathcal{G}$ with least crossing number and choosing each vertex of $\mathcal{G}$ at random with probability p. Taking the expectations of the relevant quantities gives

$$p^4\mathrm{Cr}(\mathcal{G}) \geq p^2|E| - 3p|V| + 6;$$

letting $p = \frac{4|E|}{|V|}$ yields the desired inequality (15).

Szemerédi-Trotter Theorem

This is an assertion that given n points and m lines in the plane the number of incidences

$$I(m,n) \ll m^{\frac{2}{3}}n^{\frac{2}{3}} + m + n, \tag{16}$$

(and this is sharp). Consider a set P of m points and a set L of n lines in the plane, realizing the maximal number of incidences $I(m,n)$. Define a drawing of a graph $\mathcal{G}(V,E)$ in the plane: each point $p \in P$ becomes a vertex of $\mathcal{G}$, and two points

$p, q \in P$ are connected by an edge if they lie on a common line $l \in L$ next to one another. If a line $l \in L$ contains $k \geq 1$ points of P, then it contributes $k-1$ edges to P and hence $I(m, n) = |E| + n$. Since the edges are parts of the lines, at most $\binom{n}{2}$ pairs may cross: $\mathrm{Cr}(\mathcal{G}) \leq \binom{n}{2}$. By the crossing number theorem, $\mathrm{Cr}(\mathcal{G}) \geq \frac{1}{64}\frac{|E|^3|}{m^2} - n$, so $\frac{1}{64}\frac{|E|^3|}{m^2} - n \leq \mathrm{Cr}(\mathcal{G}) \leq \binom{n}{2}$, and a calculation gives $|E| = O(m^{\frac{2}{3}}n^{\frac{2}{3}} + m)$, proving (16).

Proof of Sum-Product Inequality

We are ready to prove (13) with $\varepsilon = \frac{1}{4}$. Let $P = \{(a, b) | a \in A + A, b \in A \cdot A\}$; P is a subset of the plane and has cardinality $|A + A||A \cdot A|$. Consider the set of lines of the form $\{(x, y) : y = a(x - b)\}$ where a, b are elements of A. Clearly L has $|A|^2$ elements. Moreover, each such line contains at least $|A|$ points in P, namely, the points $(b + c, ac)$ with $c \in P$. Thus $I(P, L) \geq |A|^3$. Applying the Szemerédi-Trotter theorem and elementary linear algebra, we conclude

$$|A + A| + |A \cdot A| = \Omega(|A|^{\frac{5}{4}}). \tag{17}$$

Before turning to the discussion of Erdös-Volkmann and Katz-Tao discretized ring conjectures, let us note that if the set A is δ-separated, by carefully adapting the preceding proofs, we obtain an inequality of the form

$$\mathcal{N}(A + A, \delta^2) + \mathcal{N}(A \cdot A, \delta^2) > \mathcal{N}(A, \delta)^{1+\tau}, \tag{18}$$

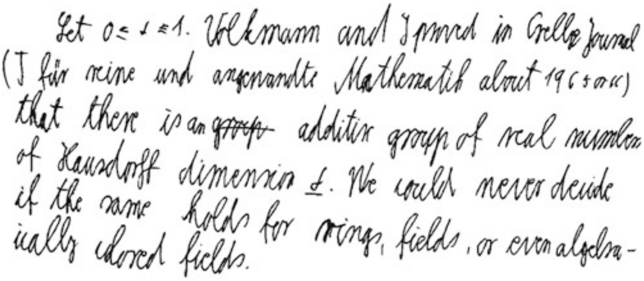
to be contrasted with Bourgain's result (1).

3.3 On the Erdös-Volkmann and Katz-Tao Discretized Ring Conjectures

Erdös-Volkmann Problem

> With Volkmann we proved that for every $0 \leq \alpha \leq 1$ there is a group of real numbers of $\dim_H = \alpha$. All our efforts so far failed in proving the existence of ring or field of Hausdorff dimension α.
>
> P. Erdös,[15] 1979

[15] Erdös expressed a similar sentiment in a letter to K. Falconer (reproduced with his kind permission (Fig. 5)). We remark that in 2016 P. Mauldin showed [69] that assuming continuum hypothesis, there exists subrings (and even subfields) of $\mathbb{R}$ of arbitrary Hausdorff dimension, which are not however Borel subsets.



Let $0 \leq \alpha \leq 1$. Volkmann and I proved in Crelle Journal (J für reine und angewandte Mathematik about 1965 or 66) that there is an ~~group~~ additive group of real numbers of Hausdorff dimension α. We could never decide if the same holds for rings, fields, or even algebraically closed fields.

Fig. 5 From a letter from P. Erdös to K. Falconer dated 18 June 1983

In 1966 Erdös and Volkmann proved [34] that for each α in (0, 1), there is an additive Borel subgroup of the reals with Hausdorff dimension α. Several proofs of this fact have now been given, all involving some sets of numbers which are well approximated by rationals. It is a well-known result that there exist infinitely many rational approximations $\frac{m}{n}$ to any real number r with an error less than n^{-2}. If $\alpha > 2$, let E be the set of real numbers r that can be "well approximated" by rational numbers in the sense that there are infinitely many rational numbers $\frac{m}{n}$ with $|r - \frac{m}{n}| < \frac{1}{n^{\alpha}}$. Jarník proved[16] in 1931 that $\dim_H(E) = \frac{2}{\alpha}$. Falconer's construction[17] of an additive Borel subset with Hausdorff dimension α builds on Jarník's Theorem: take n_k a sequence of positive integers which increases sufficiently rapidly, for example, $n_{k+1} > n_k^k$. Define the set G_α to consist of those real numbers for which there exists M such for any k there is an integer p such that $|x - \frac{p}{n_k}| < M n_k^{-\frac{1}{\alpha}}$. Clearly G_α is an additive subgroup, and it is not difficult to show, using Jarník's theorem, that its Hausdorff dimension is equal to α.

Katz-Tao Discretized Ring Conjecture

It was shown by Falconer [36] that a Borel subring R of $\mathbb{R}$ cannot have Hausdorff dimension exceeding $\frac{1}{2}$ (by considerations of the distance set $\{|a - b|; a, b, \in R \times R\} \subset \sqrt{R}$).

[16] In fact, Jarník also proved [47] a two-dimensional version of this theorem, yielding a set in $\mathbb{R}^2$ which, as was shown by Kaufmann [52], has the maximal possible set of exceptional projections discussed in Sect. 2.1.

[17] Erdös and Volkmann based their construction of G_α on the following beautiful characterization of (ir)rationals given by Cantor(1869): let $x = [x] + \sum_{k=2}^{\infty} \frac{a_k(x)}{k!}$ with the integers $a_k(x)$ satisfying $0 \leq a_k(x) \leq k - 1$. Then x is irrational iff $a_k(x) > 0$ for infinitely many k and $a_k(x) \leq k - 1$ for infinitely many k. For fixed α their G_α consists of those x which satisfy $a_k(x) \leq \kappa(x) k^{\alpha}$ or $a_k(x) \geq k - \kappa(x) k^{\alpha}$ for all $k \geq k_0(x)$ and $\kappa(x)$ positive constant.

In the 2001 paper "Some connections between Falconer's distance set conjecture and sets of Furstenberg type" [50], motivated, in part, by connections with the Kakeya problem, Nets Katz and Terence Tao formulated a quantitative version of Erdös-Volkmann problem (discretized ring conjecture). A bounded subset A of $\mathbb{R}$ is called a $(\delta,\sigma)_1$ set provided A is a union of δ-intervals and satisfies

$$|A \cap I| < (\frac{r}{\delta})^{1-\sigma}\delta^{1-\varepsilon} \tag{19}$$

whenever $I \subset \mathbb{R}$ is an arbitrary interval of size $\delta \leq r \leq 1$ ($0 < \varepsilon \ll 1$ in (19) is a small parameter).

Katz and Tao conjectured that if A is a $(\delta, \frac{1}{2})_1$ set satisfying $|A| > \delta^{\frac{1}{2}+\varepsilon}$, then necessarily $|A + A| + |A \cdot A| > \delta^{\frac{1}{2}-c}$, with $c > 0$ an absolute constant. This was proved by Bourgain in the paper eponymous with the title of this section. More generally, he proved the following result (which is the precise formulation of (1)).

Theorem 7 *If A is a $(\delta,\sigma)_1$ set, $0 < \sigma < 1$, satisfying $|A| > \delta^{\sigma+\varepsilon}$, then necessarily $|A + A| + |A \cdot A| > \delta^{\sigma-c}$, with an absolute constant $c = c(\sigma) > 0$.*

Labyrinth of the Continuum

The title of this subsection is described by Bourgain in the introduction to his paper [11] in the sentence underlined below.

> The statement in Theorem 7 is thus a purely combinatorial fact. We proceed by contradiction, assuming
>
> $$|A + A| + |A \cdot A| < \delta^{\sigma-c}. \tag{20}$$
>
> The initial stages of the argument use only the additive information, thus $|A+A| < \delta^{\sigma-c}$. It is processed through multi-scale construction, based on Ruzsa's sumset estimates, and, most importantly, quantitative versions of Freiman's famous theorem on finite sets of reals with small doubling set. ... **The final product is a subset C of A with a tree structure which exhibits a "multi-scale porosity property."** At this point, we start using multiplicative structure and prove the existence of elements $x_1, x_2 \in A - A$ such that $|x_1C + x_2C| > \delta^{\sigma-\kappa}$.

The key difficulty comes from the fact that Freiman's theorem describes the structure of sets of small doubling $|A + A| < C|A|$ with a fixed constant C, whereas the assumption (20) deals with the situation where the constant C grows with A, as A itself increases in size: the heart of Bourgain's argument is the structure theorem characterizing sets satisfying (20). The additive subgroups G_α described in Sect. 3.3 satisfy this assumption; let us look at their structure more closely, concentrating for concreteness on the case $\alpha = \frac{1}{2}$ and giving an alternative description of it as a subset of the binary tree representing the continuum (Fig. 6).

Let $P_n = \{0, \ldots, n-1\}$, and let

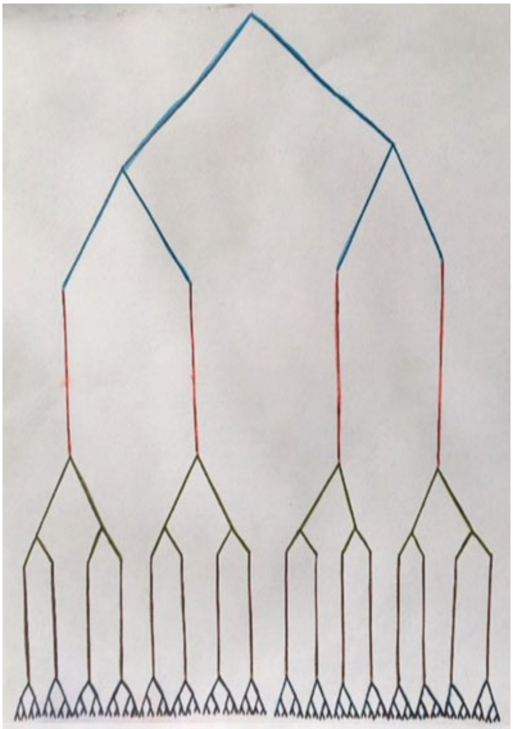


Fig. 6 Labyrinth of the continuum

$$A_n = \sum_{i=1}^{n} \frac{1}{2^{i^2}} P_{2^i} = \left\{ \sum_{i=1}^{n} a_i 2^{-i^2} \ : \ 1 \leq a_i \leq 2^i \right\}.$$

It is easy to see that the distance between distinct points $x, x' \subset A_n$ is at least $\frac{1}{4^{n^2}}$, such that x has a unique representation as a sum $\sum_{i=1}^{n} a_i 4^{-i^2}$ with $1 \leq a_i \leq 2^i$. Each term of the sum $\sum_{i=1}^{n} a_i 2^{-i^2}$ determines a distinct block of binary digits; it is seen to be GAP (defined in Sect. 3.1) as the image of $P_2 \times P_4 \cdots \times P_{2^n} \rightarrow A_n$ given by $(x_1, \ldots x_n) \rightarrow \sum_{i=1}^{n} x_i 2^{-i^2}$. The rank of this GAP is n so $|A_n + A_n| \leq 2^n |A_n|$ and $|A_n| = \prod_{i=1}^{n} |P_{2^i}| = 2^{\frac{n(n+1)}{2}}$. So we have $|A_n + A_n| = |A_n|^{1+o(1)}$.

Now we pass to the limit, akin to the way used in constructing the Cantor set: at stage n, we have a collection of $2^{\frac{n(n+1)}{2}}$ intervals of length 2^{-n^2}; from each of these intervals, we keep 2^{n+1} subintervals of length $2^{-(n+1)^2}$ separated by gaps of length $2^{-n^2-(n+1)}$. It is easy to see that the resulting fractal set coincides with $G_{\frac{1}{2}}$.

A full binary tree of height h can be identified with a set of 0, 1 valued sequences of length $\leq h$. Let us say that the tree T has *full branching for* m *generations* at the vertex σ if σ has all 2^m possible descendants m generations below it, that is, $\sigma\eta \in T$ for all $\eta \in \{0, 1\}^m$. The tree is *fully concentrated* for m generations at σ if σ has a single descendant m generations down, that is, there is a unique $\eta \in \{0, 1\}^m$ with $\sigma\eta \in T$. The sets A_n are represented by trees T_n of height n^2. For every $i < n$, every node at level i^2 has full branching for i generations and every node at level $i^2 + i$ is fully concentrated for $i + 1$ generations. Consequently, for every $j \in [i^2, i^2 + 1)$, every node at level j has full branching for one generation; for $j \in [i^2 + i, (i + 1)^2)$, every node at level j is fully concentrated for one generation. Moreover, it is not difficult to see that for every m, we can partition the levels $0, 1, \dots, n^2$ into three sets U, V, W such that:

a. For every $i \in U$, every level i node has full branching for m generations.
b. For every $j \in V$, every level j node is fully concentrated for m generations.
c. The set W constitutes a negligible fraction of the levels: $\frac{|W|}{n^2} = o(1)$ as $n \to \infty$ (with m fixed).

In the above description, $U = \bigcup_{i>m}[i^2, i^2 + i - m)$, $V = \bigcup_{i>m}[i^2 + 1, (i + 1)^2 - m)$, and W is the set of remaining levels.

Bourgain's structure theorem for sets satisfying (20) can now be informally stated as follows. Suppose $|A+A| \sim |A|^{1+\tau}$. If $b \geq 2$ is a base (say $b = 2$), we can identify A with a subset of the full b-ary tree of height m: the vertices at distance j from the root are the intervals $[kb^{m-j}, (k + 1)b^{m-j})$ which intersect A. Given ε there are $\tau > 0$ and $b \geq 2$ (which can be taken arbitrarily large) such that the following holds if m is large enough. Suppose $A \subset \{0, 1, \dots, b^{m-1}\}$ and $|A + A| \leq b^{\tau m}|A|$ (which is the case if $|A + A| \sim |A|^{1+\tau}$). Then there is a subset A' of A satisfying the following properties:

1. $|A'| \geq b^{\varepsilon m}|A|$, that is to say A' is a fairly dense subset of A.
2. The b-ary tree associated with A' is regularized in the sense that any vertex at level j has the same number N_j of children
3. Either $N_j = 1$ or $N_j \geq b^{1-\varepsilon}$, so at each level the tree has either no branching or close to full branching uniformly over all the vertices at that level.

From Theorem 7 Bourgain deduced that the answer to Erdös-Volkmann problem was negative, which was proved independently at about the same time by Edgar and Miller [32] who gave a simple and elegant proof using crucially Marstrand's projection theorem 1. The essential idea of their argument served as the starting point and inspiration for the celebrated paper by Bourgain, Katz, and Tao establishing the sum-product theorem in $\mathbb{F}_p$.

3.4 A Sum-Product Estimate in Finite Fields and Applications

The main result of this paper [19] is the following:

Theorem 8 *Let A be a subset of $\mathbb{F}_p$ such that for some $\delta > 0$*

$$p^{\delta} < |A| < p^{1-\delta}. \tag{21}$$

Fig. 7 Jean Bourgain and Terence Tao

Then

$$|A + A| + |A \cdot A| \geq c(\delta)|A|^{1+\varepsilon} \tag{22}$$

for some $\varepsilon = \varepsilon(\delta) > 0$.

Here is Terence Tao's (Fig. 7) recollection:

Regarding the prehistory of my paper with Jean Bourgain and Nets Katz, it all started with a question of Tom Wolff back in 2000, shortly before his unfortunate death. Tom had formulated the finite field version of the Kakeya conjecture (now solved by Dvir), and had observed that there appeared to be a connection between that conjecture (at least in the 3D case) and what is now the sum-product theorem. (Roughly speaking, if the sum-product phenomenon failed, then one could construct 'Heisenberg group-like' examples that almost behaved like Kakeya sets.) So he posed the question to me (as a private communication) as to whether the sum-product phenomenon was true. Nets and I chewed on this problem for a while, and found connections to some other problems (the Falconer distance problem, and the Szemeredi-Trotter theorem, over finite fields), but couldn't settle things one way or another. We then turned to Euclidean analogues, and formulated the discretized ring conjecture and showed that this was equivalent to a non-trivial improvement on the Falconer distance conjecture and on a conjecture of Wolff relating to some sets studied by Furstenberg.

After chasing some dead ends on both the finite field sum-product problem and the discretized ring problem, we gave both problems to Jean, noting that the sum-product problem would likely have applications to various finite field incidence geometry questions, including Kakeya in ${\mathbb{F}_p}^3$. Jean managed to solve the discretized ring problem using some multi-scale methods, as well as some advanced Freiman theorem type technology based on

> earlier work of Jean and Mei-Chu Chang. About the same time, Edgar and Miller solved the qualitative version of the discretized ring problem (i.e. the Erdos ring conjecture).
>
> This left the finite field sum-product problem. All the methods in our collective toolboxes were insensitive to the presence of subfields (except perhaps for Freiman's theorem, but the bounds were (and still are) too weak to get the polynomial expansion; the multi-scale amplification trick that worked in the discretized ring conjecture was unavailable here) and so were insufficient to solve the problem. We knew that it would suffice to show that some polynomial combination of A with itself exhibited expansion, but we were all stuck on how to do this for about a year, until Jean realized that the Edgar-Miller argument (based on the linear algebra dichotomy between having a maximally large span, and having a collision between generators) could be adapted for this purpose. (I still remember vividly the two-page fax from Jean conveying this point. After this breakthrough the paper got finished up quite rapidly. Of course nowadays there are many simple proofs and strengthenings of this theorem, but it was certainly a very psychologically imposing problem for us before we found the solution.

In 2006 Bourgain, Glibichuk, and Konyagin [18] proved (22) under the weaker assumption that $|A| < p^{1-\delta}$ and, combining this result with Balog-Szemerédi-Gowers lemma, made remarkable progress towards the Montgomery-Vaughan-Wooley conjecture. This asserts that multiplicative subgroups of $\mathbb{F}_p{}^*$ have "negligible additive structure" as soon as $\frac{|H|}{\log p} \to \infty$. This was established for H satisfying $|H| \geq p^{\frac{1}{4}+\delta}$ by Konyagin in 2002; Bourgain, Glibichuk, and Konyagin proved that the result holds as soon as $|H| > p^{\varepsilon}$ for any ε. Subsequently, Bourgain refined and extended this approach [12] to obtain hitherto untouchable estimates for exponential sums pertaining to Diffie-Hellman key exchange [13], a result of fundamental significance in cryptographic applications.

4 Discrete and Continuous Variations on the Expanding Theme

4.1 Bemerkung über den Inhalt von Punktmengen

> The types of creatures on the earth are countless, and on an individual level their self-preservation instinct as well as longing for procreation is always unlimited; however the space on which this entire life process plays itself out is limited. It is the *surface area of a precisely measured sphere*.
>
> Hitlers Zweites Buch, 1928

It is a pity the demented housepainter was not briefed about the Hausdorff-Banach-Tarski constructive solution of *Lebensraum* problem.[18] Building on Haus-

[18] When, as part of the "Final Solution," Hausdorff, his wife Charlotte, and a sister of hers were ordered to leave their house for local internment camp in January 1942, they opted for suicide. During the night of July 3, 1941, 40 distinguished representatives of Lvov intelligentsia, including S. Ruziewicz, perished at the hands of the S. S. "Nachtigall" battalion. Banach was saved by Rudolf Weigel, the inventor of typhus vaccine, who employed him as a feeder of lice.

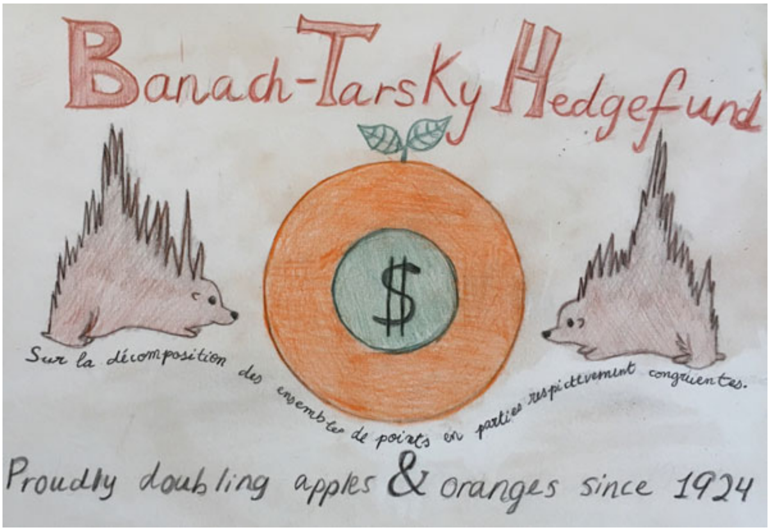

Banach-Tarsky Hedgefund
$
Sur la décomposition des ensembles de points en parties respectivement congruentes.
Proudly doubling apples & oranges since 1924

Fig. 8 Banach-Tarski hedgefund

dorff's 1914 construction [44], detailed below, Banach and Tarski, in 1924, proved [4] that there is a way of decomposing a three-dimensional ball ("precisely measured sphere") into a finite number of disjoint pieces and then reassembling the pieces to form two balls of the same radius, where "reassembling" means that the pieces are translated and rotated and that they end up still disjoint.

The construction, perhaps one of the most strikingly paradoxical in Mathematics (Fig. 8), has its origins in the question posed by Lebesgue in 1904, in the first textbook on integration bearing his name [56]. One of the properties of his integral is the monotone convergence theorem (MCT); is this property really fundamental or follows from more familiar integral axioms? Now MCT is essentially equivalent to countable additivity so the question is concerned with the existence of a positive, finitely (but not countably) additive measure on the reals assigning measure one to the unit interval.

In more detail, the problem is to assign a non-negative real number $f(A)$ to each bounded subset $A \in \mathbb{R}^n$ in such a way that:

(1) $f(E) = 1$ if E is the closed unit cube in $\mathbb{R}^n$
(2) $f(A) = f(B)$ if A and B are congruent
(3) $f(A \cup B) = f(A) + f(B)$ if A and B are disjoint
(4) $f(A_1 \cup A_2 \cup \ldots) = f(A_1) + f(A_2) + \ldots$ if $A_1, A_2, \ldots$ is any denumerable sequence of mutually disjoint sets whose union is bounded

The congruence condition in 4.1 is as follows: A and B are congruent if there exists an element g in the Euclidean group of distance preserving transformations in $\mathbb{R}^n$ such that $g(A) = g(B)$. The problem of existence of such an f is the σ-

additive measure problem; the problem of existence of f verifying only the first three properties is the finitely additive measure problem.

Lebesgue had left the countably additive measure problem in $\mathbb{R}^n$ unresolved; his construction had proved the existence of $f(A)$ for Lebesgue-measurable bounded subsets and had left the existence of non-measurable subsets as an open question. This was settled by Vitali on 1905 [92], whose construction is a forerunner of the Hausdorff-Banach-Tarski. Let l_θ be a line segment in $\mathbb{R}^2$ given by $l_\theta = \{(r, \theta) : 0 \leq r \leq 1\}$ in polar coordinates. Consider $\bigcup_\theta l_\theta = D'$ a unit disc with the origin removed. The line segments l_θ and l_ϕ belong to the same equivalence class if $\theta - \phi$ is a rational multiple of π. Consider a set E that is a union of a set of l_θ containing exactly one representative from each equivalence class. Rationals are countable: $Q \cap [0, 1] = x_1, x_2, \ldots$. Write $E_n = \{l_{\theta+2\pi x_n} : l_\theta \in E\}$. Then each E_n is obtained from E by rotation around the origin (by angle $2\pi x_n$); the sets E_n are disjoint (since E contains representative from each equivalence class), $\bigcup_n E_n = D'$. Now take D' and split it into the set F consisting of the union of the sets E_{2n} and the set G consisting of the sets E_{2n+1}. Each E_{2n} can be rotated to E_n, and the union of the E_n gives us D'. Similarly, each E_{2n+1} can be rotated to E_n, and the union of the E_n gives us D' again. Thus the punctured unit disc can be split into a countable set of disjoint pieces (all obtained by rotation of one particular set) and translated to form disjoint sets whose union is two copies of D'.[19]

Hausdorff begins his 1914 paper *Bemerkung über den Inhalt von Punktmengen* [44]by using the subgroup $G_\delta = \{n\delta\,,\ n \in \mathbb{Z}\}$ (where δ is a fixed irrational number) to show that the σ-additive problem in $\mathbb{R}^n$ has no solution for any $n \geq 1$. Both Vitali and Hausdorff use a denumerably dense subgroup of the additive group (in Hausdorff's case the dense group is $G = G_\delta + \mathbb{Z}$).

He then proceeds to show that the finitely additive measure problem in $\mathbb{R}^n$ has no solution if $n \geq 3$ by reducing the problem to the unit sphere $K = S^2$ in $\mathbb{R}^3$ and then producing the so-called Hausdorff paradoxical decomposition

$$K = A \cup B \cup C \cup Q \tag{23}$$

where A, B, C, Q are four disjoint subsets of K, Q being denumerable and $A \sim B \sim C \sim B \cup C$, the congruence here being under the group of rotations SO(3).

A decomposition (23) excludes the possibility of having an SO(3) invariant finitely additive positive measure set function defined for all subsets of K with $f(K) > 0$: indeed for such an f, $f(Q)$ must be zero and $f(A) = f(B) = f(C) = f(B \cup C) = f(B) + f(C)$, whence all of these numbers are zero, which is impossible since $0 < f(K) = f(A) + f(B) + f(C)$.

The decomposition (23) is obtained by the consideration of a denumerable subgroup $G = G(\theta, \phi)$ of SO(3) generated by two rotations θ, ϕ such that $\theta^2 = 1$, $\phi^3 = 1$, 1 being identity map, and such that θ, ϕ satisfy no other nontrivial

[19] Vitali's construction makes use of the axiom of choice (because we chose one representative from each equivalence class), and the same is true of the Banach-Tarski construction.

relations. As observed by von Neumann,[20] the group $G(\theta, \phi)$ is isomorphic to the free product of $\mathbb{Z}_2$ and $\mathbb{Z}_3$ and must necessarily contain F_2, the free group on two generators

This left open the finitely additive problem in $\mathbb{R}^1$ and $\mathbb{R}^2$; Banach begins his 1923 paper[21] (giving the title to the next subsection) by showing that in these spaces the finitely additive measure problem does have infinitely many solutions.

4.2 Sur le problème de la mesure

> Banach was not a mathematician of finesse, he was a mathematician of power. Inside he combined a spark of genius with that amazing inner imperative, which incessantly whispered to him, as in Verlaine's verse, 'Il n'y a que la glorie ardente du mètier' [There is only one thing: that intense glory of the craft] – and mathematicians know well that their craft depends on the same mystery as the craft of poets.
>
> Hugo Steinhaus[22]

In this seminal paper [3], Banach considers three questions pertaining to the invariance of finitely additive measures. First, he constructs a finitely additive, positive, translation-invariant measure μ on the family of bounded subsets of $\mathbb{R}$ such that:

(1) $\mu(A) < \infty$ for every bounded subset of $\mathbb{R}$ (so that μ gives rise in an obvious way to an element μ_A of $l^\infty(A)$).

(2) $\mu_{[a,b]}(f) = \int_a^b f(x)dx$ for every Riemann integrable function f on an interval $[a, b]$.

(3) There exists a Lebesgue integrable function g on an interval $[c, d]$ s.t. $\mu_{[c,d]}(g) \neq \int_c^d g(x)dx$.

[20] In his seminal paper *Zur allgemeinen Theorie des Masses*, which introduced the notion of amenability [72].

[21] The first equality in this paper appears just below its title: Stefan Banach (Léopol = Lwów).

"If I cared to define the single most prominent characteristic feature of Lvov school, I would mention its interest in the foundation of various theories. What I mean by this is that if one imagines mathematics as a tree, then the Lvov group was devoted to studying roots and trunks, perhaps even the main boughs, with less interest in the side branches, leaves and flowers." S. Ulam (a student of Banach and co-holder of the patent for hydrogen bomb).

[22] H. Steinhaus "discovered" Banach on the park bench of Krakow Planty promenade, discussing Lebesgue Measure with Otto Marcin Nikodym. (He viewed this as his "greatest discovery.") Lebesgue visited Lvov in 1938 to receive an honorary doctorate from Jan Kazimierz University (where Steinhaus was at that time Dean of the Faculty). Upon being given a menu in Polish at the celebratory dinner in the famous Scottish Café, Lebesgue looked at the menu for about 30 s with utmost seriousness and said, *Merci, je ne mange que des choses bien définies*[Thank you, I eat only well-defined things] [83].

The second result, which Banach calls "le probleme large de la mesure," is to show that unlike the case of $n \geq 3$, studied by Hausdorff, the finitely additive measure problem in $\mathbb{R}^n$ for $n = 1, 2$ does have infinitely many solutions.

The third question, posed by Ruziewicz in 1921, is whether Lebesgue measure on the n-sphere is the unique finitely additive rotation invariant measure defined on Lebesgue subsets. Using Hahn-Banach theorem, Banach showed that that for $n = 1$, the answer is negative, using essentially the commutativity of SO(2). He left the case of $n > 2$ open.

For $n > 3$, the affirmative answer was obtained in 1980/1981 by Margulis [65] and Sullivan [85] who used Kazhdan's property T [53].

In 1984 Drinfeld established [30] the affirmative answer in the most difficult case of $n = 2$ by proving *existence* of an element in the group ring of SU(2) which has a spectral gap. As proved by Sarnak (Fig. 9) [78], the affirmative answer for $n = 2$ implies, via inductive construction, an affirmative answer for $n \geq 2$.

Drinfeld method used some sophisticated machinery from the theory of automorphic representations, in particular Deligne's solution of Ramanujan conjecture [29]. In 1986 the explicit and optimal construction, appealing to the abovementioned tools, was obtained by Lubotzky, Phillips, and Sarnak [59, 60], in tandem with their celebrated construction (independently given by Margulis [66]) of Ramanujan graphs [61].

Fig. 9 Jean Bourgain and Peter Sarnak

4.3 Ramanujan-Selberg Conjecture

In 1916 Ramanujan [75] made two deep conjectures about the coefficients of

$$q\prod_{n=1}^{\infty}(1-q^n)^{24}=\sum_{n=1}^{\infty}\tau(n)q^n. \tag{24}$$

The first was the multiplicativity of the coefficients: if $(m,n)=1$

$$\tau(mn)=\tau(m)\tau(n); \tag{25}$$

the second was an estimate

$$|\tau(n)|\le d(n)n^{\frac{11}{2}} \tag{26}$$

where $d(n)$ is the number of divisors of n. In particular,

$$|\tau(p)|\le p^{\frac{11}{2}} \tag{27}$$

for primes p.

The first was proved by Mordell in 1917 [70] and marked the beginning of Hecke's theory of Hecke operators. The second was proved by Deligne in 1974 [29] and is one of the crowning achievements of twentieth-century mathematics.[23]

In his seminal 1965 paper *On the estimation of Fourier coefficients of modular forms*, Selberg [81] formulated an analogue of Ramanujan conjecture for non-holomorphic or Maaß forms and showed that it is equivalent to the following statement about the first positive eigenvalue of the Laplacian (*Selberg's eigenvalue conjecture*[24])

$$\lambda_1(X(p))\ge\frac{1}{4}, \tag{28}$$

where $X(p)=\mathbb{H}\backslash\Gamma(p)$, the quotient of the hyperbolic plane by the congruence subgroup

$$\Gamma(p)=\{\gamma\in\mathrm{SL}_2(\mathbb{Z})\,:\,\gamma\equiv\begin{pmatrix}1&0\\0&1\end{pmatrix}\mod p\}.$$

[23] "According to the author of the proof, Pierre Deligne, in order to present this proof, presupposing everything known by a beginning graduate student in mathematics one would need about two thousand pages of printed text. This theorem probably holds the record in modern mathematics for the ratio of the length of its proof to the length of its statement" Y. Manin [63].

[24] See the article with eponymous title by P. Sarnak in the *Notices* [79] for a tantalizing discussion.

By the variational characterization of the first eigenvalue, we have

$$\lambda_1(X(p)) = \inf_{\int_{X(p)} f d\mu=0} \frac{\int_{X(p)} |\nabla f|^2 d\mu}{\int_{X(p)} f^2 d\mu}. \tag{29}$$

Using Weil's bound for Kloosterman sums (obtained as a consequence of his proof of the Riemann hypothesis for curves), Selberg proved the following celebrated result:

$$\lambda_1(X(p)) \geq \frac{3}{16}. \tag{30}$$

This result can be viewed as (implicitly) giving rise to the first family of expander graphs.

4.4 Expanders

Expanders are highly connected sparse graphs widely used in computer science. Clearly high connectivity is desirable in any communication network. The necessity of sparsity is perhaps best seen in the case of the network of neurons in the brain: since the axons have finite thickness, their total length cannot exceed the quotient of the average volume of one's head and the area of axon's cross section. In fact, this is the context in which expander graphs first implicitly appeared in the work of Barzdin and Kolmogorov in 1967 [54].

There are several ways of making the intuitive notions of connectivity and sparsity precise; the simplest and most widely used is the following.

Given a subset of vertices, its boundary is the set of edges connecting the set to its complement. The expansion of a subset is a ratio of the size of a boundary to the size of a set. The expansion of a graph is a minimum over all expansion coefficients of its subsets. Note that the expansion coefficient is strictly positive if and only if the graph is connected.

The expansion coefficient captures the notion of being highly connected; the bigger the expansion coefficient, the more highly connected is the graph. Of course one can simply connect all the vertices, but in this case, the number of edges grows as a square of the number of vertices. The problem of constructing expanders is nontrivial because we put the second constraint: the graphs are to be sparse, i.e., the number of edges should grow linearly with the number of vertices. The simplest way to accomplish this is to demand that the graphs be regular, that is, each vertex has the same number of neighbors (say 3).

A family of k-regular graphs $\mathcal{G}_{n,k}$ forms a family of expanders if there is a fixed positive constant c, such that

$$\liminf_{n\to\infty} c(\mathcal{G}_{n,d}) \geq c > 0. \tag{31}$$

The expansion coefficient is a notion which is very easy to grasp, but it is difficult to compute numerically or to estimate analytically, as the number of subsets grows exponentially with the number of vertices. The starting point of most current work on expanders is that the expansion coefficient has a spectral interpretation:[25] to put it sonorously, if you hit a graph with a hammer, you can determine how highly connected it is by listening to the bass note. In more technical terms, high connectivity is equivalent to establishing a spectral gap for an averaging (or Laplace) operator on the graph so that condition (31) has the following alternative expression:

$$\liminf_{n\to\infty} \lambda_1(\Delta(\mathcal{G}_{n,k})) \geq \mu > 0, \tag{32}$$

making apparent the connection with Selberg's celebrated $\frac{3}{16}$ Theorem (30).

In 1973 Pinsker [73] observed that random regular graphs are expanders. In the same year, Margulis [64] gave the first explicit construction of expanders as Cayley graphs[26] of $\mathrm{SL}_3(\mathbb{F}_p)$ using Kazhdan's property T [53].

4.5 Superstrong Approximation

The strong approximation for $\mathrm{SL}_n(\mathbb{Z})$, asserting that the reduction π_q modulo q is onto, is a consequence of the Chinese remainder theorem; its extension to arithmetic groups is far less elementary but well understood. If S is a finite symmetric generating set of $\mathrm{SL}_n(\mathbb{Z})$, strong approximation is equivalent to the assertion that the Cayley graphs $\mathcal{G}(\mathrm{SL}_n(\mathbb{Z}/q\mathbb{Z}), \pi_q(S))$ are connected. The quantification of this statement, asserting that they are in fact highly connected, that is to say form a family of expanders, is what we mean by superstrong approximation. The proof of the expansion property for $\mathrm{SL}_2(\mathbb{Z})$ has its roots in Selberg's celebrated lower bound (30). The generalization of the expansion property to $G(\mathbb{Z})$ where G is a semisimple matrix group defined over $\mathbb{Q}$ is also known thanks to developments towards the general Ramanujan conjectures that have been established; this expansion property is also referred to as *property* τ for congruence subgroups.

Let Γ be a finitely generated subgroup of $\mathrm{GL}_n(\mathbb{Z})$ and let $G = \mathrm{Zcl}(\Gamma)$. The discussion of the previous paragraph applies if Γ is of finite index in G. However, if Γ is *thin*, that is to say, of infinite index in $G(\mathbb{Z})$, then $\mathrm{vol}(G(\mathbb{R})\backslash\Gamma) = \infty$, and the

[25] The connection stems from the variational characterization of the first eigenvalue, expressed in (29).

[26] Given a finite group G with a symmetric set of generators S, the Cayley graph $\mathcal{G}(G, S)$, is a graph which has elements of G as vertices and which has an edge from x to y if and only if $x = \sigma y$ for some $\sigma \in S$. The Cayley graph of $\mathrm{PSL}_2(\mathbb{F}_5)$ with respect to standard generators is a buckyball, alluded to in the rendering of O on the conference poster (Fig. 2).

techniques used to prove both of these properties do not apply. It is remarkable that under suitable natural hypothesis, strong approximation continues to hold in this thin context, as proved by Matthews, Vasserstein, and Weisfeller in 1984 [68, 94]. That the expansion property might continue to hold for thin groups was first suggested by Lubotzky and Weiss in 1993 [62]; for $\mathrm{SL}_2(\mathbb{Z})$, the issue is neatly encapsulated in the following 1-2-3 question of Lubotzky [58]. For a prime $p \geq 5$ and $i = 1, 2, 3$, let us define $S_p^i = \left\{ \begin{pmatrix} 1 & i \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ i & 1 \end{pmatrix} \right\}$. Let $\mathcal{G}_p^i = \mathcal{G}\left(\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}), S_p^i\right)$, a Cayley graph of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ with respect to S_p^i. By Selberg's theorem, $\mathcal{G}_p^1$ and $\mathcal{G}_p^2$ are families of expander graphs. However, the group $\langle \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix} \rangle$ has infinite index and thus does not come under the purview of Selberg's theorem.

Following the groundbreaking work of Helfgott [45] (which builds crucially on sum-product estimate in $\mathbb{F}_p$ discussed in Sect. 3.4), Bourgain and Gamburd [16] gave a complete answer to Lubotzky's question. The method introduced in *uniform expansion bounds for Cayley graphs of* $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ and developed in a series of papers became known as "Bourgain-Gamburd expansion machine"; thanks to a number of major developments by many people, the general superstrong approximation for thin groups is now known. The state of the art is summarized in *thin groups and superstrong approximation* [21] which contains an expanded version of most of the invited lectures from the eponymous MSRI 'Hot Topics' workshop, in the surveys by Breuillard [20] and Helfgott [46], and in the book by Tao *Expansion in Finite Simple Groups of Lie Type* [91].

4.6 On the Spectral Gap for Finitely Generated Subgroups of SU(d)

There is an Archimedean analogue of the expansion property, intimately related to the Banach-Ruziewicz problem discussed in Sect. 4.2, defined as follows.

For $k \geq 2$, let $g_1, \ldots, g_k$ be a finite set of elements in $G = SU(d)$ $(d \geq 2)$. We associate with them an averaging (or Hecke) operator $z_{g_1,\ldots,g_k}$, taking $L^2(SU(d))$ into $L^2(SU(d))$:

$$z_{g_1,\ldots,g_k} f(x) = \sum_{j=1}^{k} (f(g_j x) + f(g_j^{-1}(x)).$$

We denote by $\mathrm{supp}(z)$ the set $\{g_1, \ldots, g_k, g_1^{-1}, \ldots, g_k^{-1}\}$ and by Γ_z the group generated by $\mathrm{supp}(z)$. It is clear that $z_{g_1,\ldots,g_k}$ is self-adjoint and that the constant function is an eigenfunction of z with eigenvalue $\lambda_0(z) = 2k$. Let $\lambda_1(z_{g_1,\ldots,g_k})$ denote the supremum of the eigenvalues of z on the orthogonal complement of the constant functions in $L^2(SU(d))$. We say that z has a spectral gap if $\lambda_1(z_{g_1,\ldots,g_k}) < 2k$. It is common to, alternatively, refer to the situation described above, by asserting that the spectral gap property holds for Γ_z.

It is easy to see that affirmative solution of Banach-Ruziewicz follows from existence of z in $\mathrm{SU}(2)$ having a spectral gap. In their 1986 paper, referenced at the end of Sect. 4.2, Lubotzky, Philips and Sarnak posed a question of whether generic in measure z in $\mathrm{SU}(2)$ has a spectral gap.

In 2008 Bourgain and Gamburd [17] proved (Theorem 9 below) the spectral gap property for z in $\mathrm{SU}(2)$ satisfying the non-commutative diophantine property (NDP)—in particular for free subgroups generated by elements with algebraic entries.

The definition of non-commutative diophantine property[27] introduced in the paper "Spectra of elements in the group ring of SU(2)" by Gamburd, Jakobson, and Sarnak [41] is as follows. We say that $z_{g_1,\ldots,g_k}$ satisfies NDP if there is $D = D(g_1, \ldots, g_k) > 0$ (the diophantine constant of z) such that for any $m \geq 1$ and a word W_m in $g_1, \ldots, g_k$ of length m with $W_m \neq \pm e$ (where e denotes the identity in $\mathrm{SU}(2)$) $\|W_m \pm e\| \geq D^{-m}$.

Theorem 9 *Let $g_1, \ldots, g_k$ be a set of elements in* $\mathrm{SU}(2)$ *generating a free group and satisfying NDP (in particular, elements with algebraic entries*[28]*). Then $z_{g_1,\ldots,g_k}$ has a spectral gap.*

Regarding the proof, let me just note that in the adaption of the "expansion machine" to this Archimedean setting, the crucial role is played by the following strengthening of Theorem 7.

Theorem 10 *Given $0 < \delta < 1$ and $\kappa > 0$, there exists $\varepsilon_0 > 0$ and $\varepsilon_1 > 0$ such that if $\delta > 0$ is sufficiently small and $A \subset [1, 2]$ is a discrete set consisting of δ-separated points, satisfying $|A| = \delta^{-\sigma}$ and*

$$|A \cap I| < \rho^{\kappa}|A| \tag{33}$$

whenever I is a size ρ interval with $\delta < \rho < \delta^{\varepsilon_0}$, then

$$N(A + A, \delta) + N(A \cdot A, \delta) > \delta^{-\varepsilon_1}|A|. \tag{34}$$

[27] Recall that $\theta \in \mathbb{R}$ is called diophantine if there are positive constants C_1, C_2 s.t. for all $(k, l) \in \mathbb{Z}^2$ with $k \neq 0$ we have $|k\theta - l| \geq c_1 k^{-c_2}$. Equivalently, letting $g = e^{2\pi\theta} \in \mathrm{SO}(2)$, we may re-express this condition as follows: $|g^k - 1| \geq c_1' k^{-c_2'}$. A classical result asserts that diophantine numbers are generic in measure in $\mathbb{R}$. Given diophantine $\theta_1, \ldots, \theta_k$ and $g_1 = e^{2\pi\theta_1}, \ldots, g_k = e^{2\pi\theta_k}$ in SO(2), for any word W in $g_1, \ldots, g_k$ of length m, we have $|W_m - 1| \geq c_1 m^{-c_2}$ for some c_1, c_2. In the case of SO(3), given $g_1, \ldots, g_k$ generating a free subgroup, a pigeonhole argument shows that for any $m \geq 1$ there is always a word W in $g_1, g_1^{-1}, \ldots, g_k, g_k^{-1}$ of length at most m such that $\|W_m - e\| \leq 10(2k-1)^{-\frac{m}{6}}$, so the exponential behavior in the definition below is the appropriate one.

[28] It was established in [41] that elements with algebraic entries satisfy NDP. A major open question is whether z generic in measure in SU(2) satisfies NDP. The best known result in this direction is due to Kaloshin and Rodnianski [49]: for almost every pair (A, B) in $\mathrm{SU}(2) \times \mathrm{SU}(2)$, there is a constant $D > 0$ s.t. for any n and any word W_m $\|W_m(A, B) \pm e\| \geq D^{-m^2}$.

Theorem 9 is of importance in quantum computing [28, 43]. In the context of quantum computation, elements of a three-dimensional rotation group are viewed as "quantum gates," and a set of elements generating a dense subgroup is called "computationally universal" (since any element of rotation group can be approximated by some word in the generating set to an arbitrary precision). A set of elements is called "efficiently universal" if any element can be approximated by a word of length which is logarithmic with respect to the inverse of the chosen precision (this is the best possible). A consequence of Theorem 9 is that computationally universal sets with algebraic entries are efficiently universal.

Another application is related to the theory of quasicrystals. Generalizing Penrose's two-dimensional aperiodic tiling, John Conway and Charles Radin [26] constructed a self-similar (hierarchical) tiling of a three-dimensional space with a single prototile, such that the tiles occur in an infinite number of different orientations in the tiling. The tile is a prism, which when scaled up by two is subdivided into eight copies of itself ("daughter tiles"). If one iterates this same subdivision procedure over and over, one creates in the limit the desired tiling of three-dimensional space by prisms. Conway and Radin showed that the orientations of tiles in the tiling are uniformly distributed and posed the question of how fast this convergence to uniform distribution takes place. This question reduces to the study of the spectral gap for the averaging operator associated with eight rotations giving orientations of daughter tiles. A consequence of Theorem 9 is that this convergence takes place exponentially fast.

5 Coda

> The essence of mathematics lies precisely in its freedom.
>
> Georg Cantor

> Already history has in a sense ceased to exist, i.e. there is no such thing as a history of our own times which could be universally accepted, and the exact sciences are endangered as soon as military necessity ceases to keep people up to the mark. Hitler can say that the Jews started the war, and if he survives, that will become official history. He can't say that two and two are five, because for the purposes of, say, ballistics they have to make four.
>
> George Orwell, letter to N. Wilmett, 18 May 1944

> Freedom is the freedom to say that two plus two make four. If that is granted, all else follows.
>
> George Orwell, *Nineteen Eighty-Four*, 1949

The difficulties of explaining Bourgain's work to a broad mathematical audience turned out to be quite substantial;[29] omitting "mathematical" from the appellation renders them nearly insurmountable.

[29] "There is a continuing need to lead new generations along the thorny path which has no shortcuts. The Ancients said there is no royal road in mathematics. But the vanguard is leaving the great mass of pilgrims further and further behind, the procession is ever more strung out, and the leaders are finding themselves alone far out ahead" H. Steinhaus [84].

Ian Stewart begins his admirable book *The Problems of Mathematics* (Oxford University Press, 1987) with an interview with a mathematician conducted by Seamus Android on behalf of the proverbial man in the street[30] invoked in Hilbert's celebrated 1900 address *Problems of Mathematics*, referenced at the beginning of Sect. 2.

> Mathematician: It's one of the most important discoveries of the last decade!
>
> Android: Can you explain it in words ordinary mortals can understand?
>
> Mathematician: Look, buster, if ordinary mortals could understand it, you would not need mathematicians to do the job for you, right? You can't get a feeling for what's going on without understanding the technical details. How can I talk about manifolds without mentioning that the theorem only works if the manifolds are finite dimensional paracompact Hausdorff with empty boundary?
>
> Android: Lie a bit.
>
> Mathematician: Oh, but I could not do that!
>
> Android: Why not? *Everybody else does.*

Perhaps the most troubling omen of our times is an assault on the very basic notions of logic and truth, in their most elemental Aristotelian sense, including, in particular, the law of the excluded middle. Our discipline stands as a mighty fortress against this assault, and I, for one, believe we should not be overly defensive about our reluctance to *lie a bit* just because *everybody else does*.

***

> Of all escapes from reality, mathematics is the most successful ever. It is a fantasy that becomes all the more addictive because it works back to improve the same reality we are trying to evade. All other escapes – sex, drugs, hobbies, whatever –are ephemeral by comparison. The mathematician's feeling of triumph, as he forces the world to obey the laws his imagination has freely created, feeds on its own success. The world is permanently changed by the workings of his mind, and the certainty that his creations will endure renews his confidence as no other pursuit.
>
> Gian-Carlo Rota, 'The Lost Cafe', 1987

> The one who writes a poem writes it above all because verse writing is an extraordinary accelerator of conscience, of thinking, of comprehending the universe. Having experienced this acceleration once, one is no longer capable of abandoning the chance to repeat this experience; one falls into dependency on this process, the way others fall into dependency on drugs or on alcohol. One who finds himself in this sort of dependency on language is, I guess, what they call a poet.
>
> Joseph Brodsky, 'Nobel Lecture', 1987

To paraphrase W. H. Auden (writing *In Memory of W. B. Yeats*),

> [Mathematics][31] makes nothing happen: it survives

[30] The proverbial meaning is a function of street's location in space-time cultural continuum: it is exceedingly unlikely, I reckon, that Aristotle's remark in *Nicomachean Ethics, IX* that "without friends none would care to live, though having all other things besides" should necessarily be construed as endorsement of Facebook.

[31] Save e.g. hydrogen bomb and computer.

> In the valley of its making, where executives
> Would never want to tamper.

In attempting to explain the significance of Bourgain's remarkable and remarkably useful results to a proverbial human-on-line, one may invoke their applications in mathematical physics, computer science, and cryptography, which are of immense practical importance in contemporary life, making, in particular, the online communication possible. Their subtlety, beauty, and depth appear to be much harder to convey in "plain English." Here and now, perhaps, we must remind ourselves that the human-on-line, while attached to a digital device (built by von Neumann), is still *human* and sound bite/tweet thus: while dealing with entities seemingly fake/unreal (e.g., the real line), Bourgain's singular adventures in the labyrinth of the continuum represent a magnificent and transcendent achievement of the human spirit.

***

I met Jean in September 2005, 6 months after my daughter (who drew the pictures for this essay) was born, while visiting IAS for the program "Lie Groups, Representations and Discrete Mathematics" led by Alex Lubotzky. I do not remember the precise date but do remember the hour: it was between 2 and 3 am. After changing my daughter's diapers, I could not sleep, went to Simonyi Hall, and ran into Jean walking to the Library. It was in this discombobulated state that I was free of fear to speak to him. By dawn, the problem which had been resisting my protracted attack for a decade was vanquished in Jean's office.[32]

During this happiest year of my life, in 2005–2006, I stayed on the Lane named after Hermann Weyl who was of the view that "Mathematics is not the rigid and uninspiring schematism which the laymen is so apt to see in it; on the contrary, we stand in mathematics precisely at that point of limitation and freedom which is the essence of man himself."

During my second visit to IAS, in 2007–2008, as von Neumann Fellow participating in the "Arithmetic Combinatorics" Program led by Jean Bourgain and Van Vu, I stayed on the Lane named after Erwin Panofsky. His magnificent essay *The History of Art as a Humanistic Discipline*, based on The Spencer Trask Princeton University Lectures for 1937–38, commences thus:

> Nine days before his death Immanuel Kant was visited by his physician. Old, ill, and nearly blind, he rose from his chair and stood trembling with weakness and muttering unintelligible

[32] Jean had the following daily routine. He would arrive at the dining hall for lunch within 5 minutes of its closing and, while descending the stairs, would look for whom to join for the meal (the relevance of the person was determined primarily by their expertise in the problem Jean was currently working on). After lunch and before the sunset, the door of his office would be half-open. After getting a bottle or red wine (typically Medoc), Jean would have dinner around 9 pm, followed by a double espresso (typically in *Small World Coffee*), return to the office, call his wife and son, and then go for a brisk walk, encircling the Einstein Drive about 5 times. Between midnight and the sunrise, the office door would typically be closed. His handwritten notes (like that of Mozart's and unlike Beethoven's) are virtually free of corrections, in part, because during the dinner and the walk he would think about what would be set to paper upon his return to the office.

> words. Finally his faithful companion realized that he would not sit down until the visitor has taken a seat. This he did, and Kant then permitted himself to be helped to his chair, and, after regaining some of his strength, said, 'Das Gefühl für Humanität hat mich noch nicht verlassen' – 'The sense of humanity has not yet left me'. The two men were moved almost to tears. For, though the word *Humanität* had come, in the eighteenth century, to mean little more than politeness or civility, it had, for Kant, a much deeper significance, which the circumstances of the moment served to emphasize: man's proud and tragic consciousness of self-approved and self-imposed principles, contrasting with his utter subjection to illness, decay and all that is implied in the word 'mortality'.

Towards the end of the essay, Panofsky thus (pre-)echoes Orwell: "If the anthropocratic civilization of the Renaissance is headed, as it seems to be, for a *Middle Ages in reverse –a satanocracy* as opposed to the mediaeval theocracy – not only the humanities but also natural sciences, as we know them, will disappear, and nothing will be left but what serves the dictates of the *sub-human*."

During my third, short visit (Fig. 10), I stayed on von Neumann Drive (the only other "Drive" at IAS is named after Einstein). The similarities between von Neumann and Baron Bourgain are subtle and striking.[33] In his article *The Legend of John von Neumann* [42], Paul Halmos has the following to say: "The heroes of humanity are of two kinds: the ones who are just like all of us, but very much more so, and the ones who, apparently, have and extra-human spark. We can all run, and some of us can run the mile in less than 4 minutes; but there is nothing that most of us can do that compares with the creation of the Great G-minor Fugue. Von Neumann's greatness was the human kind. We can all think clearly, more or less, some of the time, but von Neumann's clarity of thought was orders of magnitude greater than that of most of us, all the time. Both Norbert Wiener and John von Neumann were great men, and their names will live after them, but for different reasons. Wiener saw things deeply but intuitively; von Neumann saw things clearly and logically." One may agree or disagree with Halmos's assessment; it is my belief that Bourgain's greatness combined these two kinds.

***

The IAS (where Jean did most of the work described in this essay) official seal (Fig. 11) is imprinted on the *Analysis and Beyond* conference poster. In a circular format, the quiet elegant and classical Art Deco composition depicts two graceful young ladies, one clothed and one otherwise, standing on opposite sides of a leafy tree that appears to bear abundant fruit. Their poses are complementary, one looking out towards the spectator and the other looking down, avoiding eye contact. The

[33] The following remarks about Johnny are equally applicable to Jean. "It is usually difficult to sharpen von Neumann's results. With small concern for expository simplifications or intuitive motivations, he characteristically went straight to the heart of problems, and had an uncanny ability to check all the essentially different possibilities, individually and in combination. This ability gives most of his work an objective finality, and makes later workers begin by trying to simplify von Neumann's arguments, or to apply similar techniques to related problems" [7].

"The story used to be told about him in Princeton, that while he was indeed a demi-god, he had made a detailed study of humans and could imitate them perfectly" [39].

Fig. 10 Jean Bourgain, Peter Sarnak, Alex Gamburd

figures are named in large *sans serif* letters, TRUTH to the left and BEAUTY on the right. Truth holds a mirror that overlaps the circular frame to reflect reality.

Underlying the design of the seal is the evident allusion to the famous final couplet of "Ode on a Grecian Urn": *"Beauty is truth, truth beauty," – that is all Ye know on earth, and all ye need to know* by John Keats, who was of the view that "the excellence of every art is its intensity, capable of making all disagreebles evaporate from their being in close relationship with Beauty and Truth."

Having attempted in this essay a snapshot of the excellence of Bourgain's art, let me conclude by giving a glimpse of his intensity by quoting from the interview upon receiving the *2017 Breakthrough Prize in Mathematical Sciences* (Fig. 12):

Fig. 11 The IAS Seal

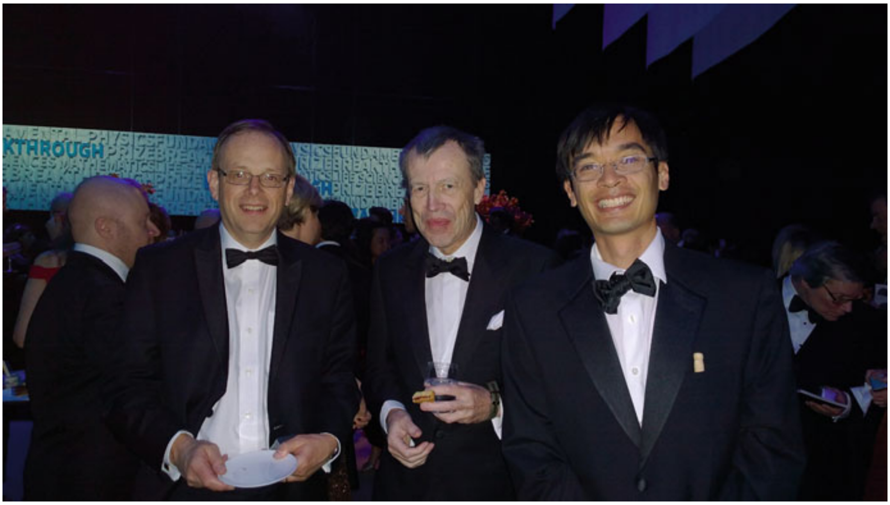


Fig. 12 Richard Taylor, Jean Bourgain, Terence Tao

> If you have a question which is generally perceived as unapproachable, it is often that you do not even quite know where you have to look to get a solution. From that point of view, we are rather like Fourier,[34] stranded in the desert, hopelessly lost. At the moment you get this insight, all of a sudden you escape the desert and things open up for you. Then we feel very

[34] Jean-Baptiste Joseph Fourier was a member of General Bonaparte's expedition to Egypt(1798–1801), important enough for the First Consul to make him, in 1802, the Prefect of the département at Grenoble, a position which he held until Emperor Napoleon's fall.

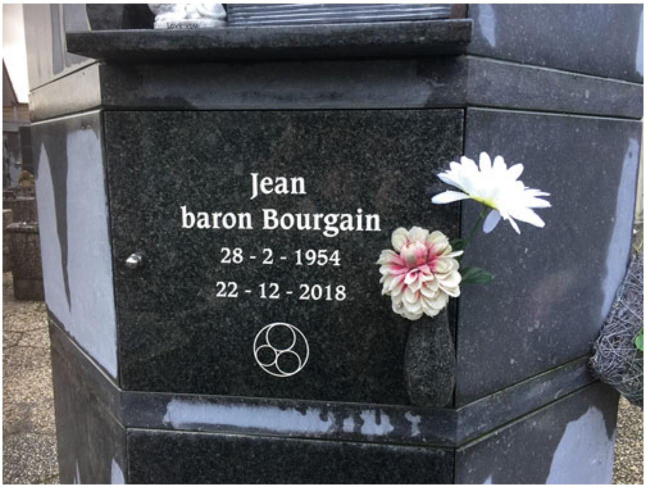


Fig. 13 Jean baron Bourgain 1954–2018

excited. These are the best moments. They make up for all the suffering with absolutely no progress worth it.

Acknowledgments The author was supported, in part, by NSF award DMS-1603715.
Opening image and Figs. 1, 2, and 11 are courtesy of the Institute for Advanced Study.
Figures 3 and 9 are courtesy of C. J. Mozzochi.
Figure 4 is courtesy of Svetlana Jitomirskaya.
Figure 5 is courtesy of Kenneth Falconer.
Figures 6 and 8 are courtesy of Anna Gamburd.
Figure 7 is courtesy of Erik Huss, Royal Swedish Academy of Sciences.
Figure 10 is courtesy of Enrico Bombieri.
Figure 12 is courtesy of Terence Tao.
Figure 13 is courtesy of Claire Bourgain.

References

1. Atiyah, M.: Geometry and Physics of the 20th Century, Géométrie au XXe Siècle, sous la direction de Joseph Kouneiher, Dominique Flament, Philippe Nabonnand, Jean-Jacques Szczeciniarz, 4–9, Hermann, 2005
2. Balog, A., Szemerédi, E.: A statistical theorem of set addition. Combinatorica **14**(3), 263–268 (1994)
3. Banach, S.: Sur le probleme de mesure. Fund. Math. 4, 7–33 (1923)
4. Banach, S, Tarski, A.: Sur la décomposition des ensembles de points en parties respectivement congruentes. Fund. Math. **6**, 244–277 (1924)
5. Besicovitch, A.S.: Sur deux questions d'intégrabilité des fonctions. J. Soc Phys.-Math. (Perm) **2**, 105–123 (1919)

6. Besicovitch A.S.: On fundamental geometric properties of plane line sets, J. Lond. Math. Soc. **39**, 441–448 (1964)
7. Birkhoff, G.: Von Neumann and lattice theory. Bull. Am. Math. Soc. **64**, 50–56 (1958)
8. Bourgain, J.: Besicovitch type maximal operators and applications to Fourier analysis. Geom. Funct. Anal. **1**(2), 147–187 (1991)
9. Bourgain, J.: Remarks on Montgomery's conjectures on Dirichlet sums. Geometric aspects of functional analysis (1989–1990), 153–165, Lecture Notes in Math., vol. 1469. Springer, Berlin (1991)
10. Bourgain, J.: On the dimension of Kakeya sets and related maximal inequalities. Geom. Func. Anal. **9**, 256–282 (1999)
11. Bourgain, J.: On the Erdös-Volkmann and Katz-Tao ring conjectures. Geom. Funct. Anal. **13**(2), 334–365 (2003)
12. Bourgain, J.: Mordell's exponential sum estimate revisited. J. Am. Math. Soc. **18**(2), 477–499 (2005)
13. Bourgain, J.: Estimates on exponential sums related to the Diffie-Hellman distributions. Geom. Funct. Anal. **15**(1), 1–34 (2005)
14. Bourgain, J.: The discretized sum-product and projection theorems. J. Anal. Math. **112**, 193–236 (2010)
15. Bourgain, J., Demeter, C., Guth, L.: Proof of the main conjecture in Vinogradov's mean value theorem for degrees higher than three. Ann. Math. (2) **184**(2), 633–682 (2016)
16. Bourgain, J., Gamburd, A.: Uniform expansion bounds for Cayley graphs of $\mathrm{SL}_2(\mathbb{F}_p)$. Ann. Math. **167**, 625–642 (2008)
17. Bourgain, J., Gamburd, A.: On the spectral gap for finitely-generated subgroups of SU(2). Invent. Math. **171**(1), 83–121 (2008)
18. Bourgain, J., Glibichuk, A.A., Konyagin, S.V.: Estimates for the number of sums and products and for exponential sums in fields of prime order. J. Lond. Math. Soc. (2) **73**(2), 380–398 (2006)
19. Bourgain, J., Katz, N., Tao, T.: A sum-product estimate in finite fields, and applications. Geom. Funct. Anal. **14**(1), 27–57 (2004)
20. Breuillard, E.: Approximate subgroups and super-strong approximation. Groups St Andrews 2013, 1–50, London Math. Soc. Lecture Note Ser., 422. Cambridge Univ. Press, Cambridge (2015)
21. Breuillard, E., Oh, H.: Thin groups and superstrong approximation, Math. Sci. Res. Inst. Publ., 61. Cambridge Univ. Press, Cambridge (2014)
22. Burkill, J.C., Abram Samoilovitch Besicovitch. 1891–1970, Biographical Memoirs of Fellows of the Royal Society, vol. 17, pp. 1–16 (1971)
23. Caffarelli, L.: The work of Jean Bourgain, Proceedings of the International Congress of Mathematicians, vol. 1, 2 (Zürich, 1994), pp. 3–5. Birkhäuser, Basel (1995)
24. Chang, M.: A polynomial bound in Freiman's theorem. Duke Math. J. **113**(3), 399–419 (2002)
25. Christ, M., Duoandikoetxea, J., Rubio de Francia, J.L.: Maximal operators associated to the Radon transform and the Calderon-Zygmund method of rotations. Duke Math. J. **53**, 189–209 (1986)
26. Conway, J., Radin, C.: Quaquaversal tilings and rotations. Invent. Math. **132**, 179–188 (1998)
27. Davies, R.: Some remarks on the Kakeya problem. Proc. Camb. Philos. Soc. **69**, 417–421 (1971)
28. Dawson, C.M., Nielsen, M.A.: The Solovay-Kitaev algorithm. Quantum Inf. Comput. **6**, 81–95 (2006)
29. Deligne, P.: La conjecture de Weil. I. (French) Inst. Hautes Études Sci. Publ. Math. No. **43**, 273–307 (1974)
30. Drinfeld, V.: Finitely-additive measures on S2 and S3, invariant with respect to rotations. Funct. Anal. Appl. **18**, 245–246 (1984)
31. Drury, S.: Lp estimates for the x-ray transform. Illinois J. Math. **27**, 125–129 (1983)
32. Edgar, G.A., Miller, C.: Borel subrings of the reals. Proc. Am. Math. Soc. **131**, 1121–1129 (2003)

33. Elekes, G.: On the number of sums and products. Acta Arith. **81**, 365–367 (1997)
34. Erdös, P., Volkmann, B.: Additive Gruppen mit vorgegebener Hausdorffscher dimension. J. Reine Angew. Math **221**, 203–208 (1966)
35. Erdös, P., Szemerédi, E.: On sums and products of integers. In: Studies in Pure Mathematics, pp. 213–218. Birkhäuser, Basel (1983)
36. Falconer, K.J.: The Hausdorff dimension of distance sets. Mathematika **32**, 206–212 (1985)
37. Fefferman, C.: The multiplier problem for the ball. Ann. Math. **94**, 330–336 (1971)
38. Freiman, G.: Foundations of a Structural Theory of Set Addition, Trans. Math. Monogr., vol. 37. Amer. Math. Soc., Providence (1973)
39. Goldstine, H.: Computer from Pascal to von Neumann. Princeton University Press, Princeton (1972)
40. Gowers, T.: A new proof of Szemeredi's theorem for arithmetic progressions of length four. GAFA **8**(3), 529–551 (1998)
41. Gamburd, A., Jakobson, D., Sarnak, P.: Spectra of elements in the group ring of SU(2). J. Eur. Math. Soc. **1**, 51–85 (1999)
42. Halmos, P.R.: The legend of John von Neumann. Am. Math. Monthly **80**, 382–394 (1973)
43. Harrow, A., Recht, B., Chuang, I.: Efficient discrete approximation of quantum gates. J. Math. Phys. **43**, 4445–4452 (2002)
44. Hausdorff, F.: Bemerkung über den Inhalt von Punktmengen. Math. Ann. **75**, 428–434 (1914)
45. Helfgott, H.A.: Growth and generation in $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. Ann. Math. (2) **167**(2), 601–623 (2008)
46. Helfgott, H.A.: Growth in groups: ideas and perspectives. Bull. Am. Math. Soc. (N.S.) **52**(3), 357–413 (2015)
47. Jarník, V.: Über die simultanen diophantischen Approximationen. Math. Zeit. **33**, 505–543 (1931)
48. Kakeya, S.: Some problems on maxima and minima regarding ovals. Tohoku Sci. Rep. **6**, 71–88 (1917)
49. Kaloshin, V., Rodnianski, I.: Diophantine properties of elements of SO(3). Geom. Funct. Anal. **11**, 953–970 (2001)
50. Katz, N., Tao, T.: Some connections between Falconer's distance set conjecture, and sets of Furstenberg type. N. Y. J. Math. **7**, 149–187 (2001)
51. Kaufman, R.: On the Hausdorff dimension of projections. Mathematika **15**, 153–155 (1968)
52. Kaufman R.: On the theorem of Jarník and Besicovitch. Acta Arithmetica **39**, 265–267 (1981)
53. Kazhdan, D.A.: On the connection of the dual space of a group with the structure of its closed subgroups. Funkcional. Anal. i Prilozhen. **1**, 71–74 (1967)
54. Kolmogorov, A.N., Barzdin, Ya.M.: On the realization of networks in three-dimensional space. Problemy Kibernet. **19**, 261–268 (1967)
55. Łaba, I.: From harmonic analysis to arithmetic combinatorics. Bull. Am. Math. Soc. (N.S.) **45**(1), 77–115 (2008)
56. Lebesgue, H.: Lecons sur l'Integration et la Recherche des Fonctions Primitives. Gauthier-Villars, Paris (1904)
57. Leighton, T.: Complexity Issues in VLSI, Foundations of Computing Series. MIT Press, Cambridge, (1983)
58. Lubotzky, A.: Cayley graphs: eigenvalues, expanders and random walks. In: Rowbinson, P. (ed.), Surveys in Combinatorics. London Math. Soc. Lecture Note Ser. 18, pp. 155–189. Cambridge Univ. Press, Cambridge (1995)
59. Lubotzky, A., Phillips, R., Sarnak, P.: Hecke operators and distributing points on S^2 I. Commun. Pure Appl. Math. **39**(S), S149–S186 (1986)
60. Lubotzky, A., Phillips, R., Sarnak, P.: Hecke operators and distributing points on S^2 II. Commun. Pure Appl. Math. **40**(4), 401–420 (1987)
61. Lubotzky, A., Phillips, R., Sarnak, P.: Ramanujan graphs. Combinatorica **8**, 261–277 (1988)
62. Lubotzky, A., Weiss, B.: Groups and expanders. In: Friedman, J. (ed.), DIMACS Series in Disc. Math. and Theor. Comp. Sci. vol. 10, pp. 95–109 (1993)
63. Manin, Yu. I.: Mathematics and physics. Translated from the Russian by Ann Koblitz and Neal Koblitz. Progress in Physics, 3. Birkhäuser, Boston (1981)

64. Margulis, G.A.: Explicit constructions of expanders. Problemy Peredaci Informacii **9**(4), 71–80 (1973)
65. Margulis, G.A.: Some remarks on invariant means. Monatsh. Math. **90**, 233–235 (1980)
66. Margulis, G.A.: Explicit group-theoretic constructions of combinatorial schemes and their applications in the construction of expanders and concentrators. Problemy Peredachi Informatsii **24**(1), 51–60 (1988)
67. Marstrand, J.M.: Some fundamental geometrical properties of plane sets of fractional dimensions. Proc. Lond. Math. Soc. (3) **4**, 257–302 (1954)
68. Matthews, C.R., Vaserstein, L.N., Weisfeiler, B.: Congruence properties of Zariski dense subgroups. I. Proc. Lond. Math. Soc. (3) **48**(3), 514–532 (1984)
69. Mauldin, R.D.: Subfields of R with arbitrary Hausdorff dimension. Math. Proc. Camb. Philos. Soc. **161**(1), 157–165 (2016)
70. Mordell, L.J.: On Mr. Ramanujan's empirical expansions of modular functions. Proc. Camb. Philos. Soc. **19**, 117–124 (1917)
71. Nahmod, A.R.: The nonlinear Schrödinger equation on tori: integrating harmonic analysis, geometry, and probability. Bull. Am. Math. Soc. (N.S.) **53**(1), 57–91 (2016)
72. von Neumann, J.: Zur allgemeinen Theorie des Maßes. Fund. Math. **13**, 73–111 (1929)
73. Pinsker, M.S.: On the complexity of a concentrator. In: 7th International Teletrafic Conference, pp. 318/1–318/4 (1973)
74. Plünnecke, H.: Eigenschaften und Abschätzungen von Wirkingsfunktionen, BMwFGMD-22 Gesellschaft für Mathematik und Datenverarbeitung, Bonn (1969)
75. Ramanujan, S.: On certain arithmetical functions. Trans. Camb. Philos. Soc. **XXII**, 159–184 (1916)
76. Ramsey, F.P.: On a problem of formal logic. Proc. Lond. Math. Soc. **30**, 264–285 (1930)
77. Ruzsa, I.: Sums of Finite Sets. Number Theory (New York, 1991–1995), pp. 281–293. Springer, New York (1996)
78. Sarnak, P.: Some Applications of Modular Forms. Cambridge Tracts in Mathematics, vol. 99. Cambridge University Press, Cambridge (1990)
79. Sarnak, P.: Selberg's eigenvalue conjecture. Notices Am. Math. Soc. **42**, 1272–1277 (1995)
80. Schnirelmann, L.: Über additive Eigenschaften von Zahlen. Ann. Inst. Polyt. Novocherkassk **14**, 3–28 (1930)
81. Selberg, A.: On the estimation of Fourier coefficients of modular forms. Proc. Sympos. Pure Math. **VII**, 1–15 (1965)
82. Sanders, T.: The structure theory of set addition revisited. Bull. Am. Math. Soc. (N.S.) **50**(1), 93–127 (2013)
83. Steinhaus, H.: Mathematician for all seasons—recollections and notes, vol. 1 (1887–1945). With a foreword by Kazimierz Dziewanowski. Translated from the Polish by Abe Shenitzer. Edited and with an introduction by Robert G. Burns, Irena Szymaniec and Aleksander Weron. Vita Mathematica, 18. Birkhäuser/Springer, Cham (2015)
84. Steinhaus, H.: Mathematician for all seasons—recollections and notes, vol. 2 (1945–1968). Translated from the Polish by Abe Shenitzer. Edited by Robert G. Burns, Irena Szymaniec and Aleksander Weron. Vita Mathematica, 19. Birkhäuser/Springer, Cham (2016)
85. Sullivan, D.: For $n > 3$ there is only one finitely additive rotationally invariant measure on the n-sphere on all Lebesgue measurable sets. Bull. Am. Math. Soc. **4**, 121–123 (1981)
86. L. Székely, Crossing numbers and hard Erdös problems in discrete geometry. Combin. Probab. Comput. **6**, 353–358 (1997)
87. Szemerédi, E.: Regular partitions of graphs, Problemes combinatoires et theorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976), pp. 399–401, Colloq. Internat. CNRS, 260, CNRS, Paris (1978)
88. Szemerédi, E., Trotter, W.: Extremal problems in discrete geometry. Combinatorica **3**, 381–392 (1983)
89. Tao, T.: From rotating needles to stability of waves: emerging connections between combinatorics, analysis and PDE. Notices AMS **48**(3), 294–303 (2001)

90. Tao, T.: Product set estimates for non-commutative groups. Combinatorica **28**(5), 547–594 (2008)
91. Tao, T.: Expansion in Finite Simple Groups of Lie Type. Graduate Studies in Mathematics, vol. 164. American Mathematical Society, Providence (2015)
92. Vitali, G.: Sul problema della misura dei gruppi di punti di una retta. Bologna, Tip. Gamberini e Parmeggiani (1905)
93. van der Waerden, B.L.: Beweis einer Baudetschen Vermutung, Nieuw. Arch. Wisk. **15**, 212–216 (1927)
94. Weisfeiler, B.: Strong approximation for Zariski-dense subgroups of semi-simple algebraic groups. Ann. Math. (2) **120**(2), 271–315 (1984)
95. Wolff, T.: Recent Work Connected with the Kakeya Problem. Prospects in Mathematics (Princeton, NJ, 1996), pp. 129–162. Amer. Math. Soc., Providence (1999)

Cartan Covers and Doubling Bernstein-Type Inequalities on Analytic Subsets of $\mathbb{C}^2$

Michael Goldstein, Wilhelm Schlag, and Mircea Voda

Dedicated to the memory of Jean Bourgain

Abstract We prove a version of the doubling Bernstein inequalities for the trace of an analytic function of two variables on an analytic subset of $\mathbb{C}^2$. The estimate applies to the whole analytic set in question including its singular points. The proof relies on a version of the Cartan estimate for maps in $\mathbb{C}^2$ which we establish in this work.

1 Introduction

In a series of papers [9–11], Fefferman and Narasimhan investigated the local behavior of a polynomial f of N real or complex variables, restricted to a given

The second author was partially supported by the NSF, DMS-1500696. The first author was partially supported by NSERC grant. The first author thanks the University of Chicago and Yale University for hospitality. The authors are grateful to János Kollár and Mihnea Popa for helpful discussions on Bézout's theorem.

M. Goldstein (✉)
Department of Mathematics, University of Toronto, Toronto, ON, Canada
e-mail: gold@math.toronto.edu

W. Schlag
Department of Mathematics, Yale University, New Haven, CT, USA

Department of Mathematics, The University of Chicago, Chicago, IL, USA
e-mail: wilhelm.schlag@yale.edu

M. Voda
Department of Mathematics, The University of Chicago, Chicago, IL, USA

Department of Mathematics, University of Toronto, Toronto, ON, Canada
e-mail: mvoda@uchicago.edu

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_5

n-dimensional algebraic variety $\mathcal{X}$. Conceptually, the problem is to quantify to what extent the local behavior of the trace of f on $\mathcal{X}$ deteriorates relative to an N-dimensional ball. Of particular interest here is to determine the dependence of quantitative estimates on the degree of the polynomials. Fefferman and Narasimhan chose the classical Bernstein inequalities for polynomials of several variables to measure the distortion of a polynomial restricted to an algebraic variety.

The authors' interest in this particular problem arose as part of their work on the Chulaevsky-Sinai conjecture. In their pioneering paper [6], Chulaevsky and Sinai analyze the spectrum of a discrete Schrödinger operator on $\mathbb{Z}$ with a quasi-periodic potential given by evaluating a generic smooth function on $\mathbb{T}^2$ along the orbit of an ergodic shift. In [16] (building on work from [15]), the authors found that some "generic versions" of these restricted Bernstein estimates play a crucial role in addressing this conjecture.

There are two major differences between the current paper and [11]: (i) we obtained estimates at singular points, and the estimates at regular points don't depend on the distance to the singular points, and (ii) we allow analytic functions and analytic sets in place of polynomials and algebraic varieties.

Fefferman and Narasimhan had considered compact subsets of algebraic varieties away from the singular points. For polynomials and algebraic varieties, Roytwarf and Yomdin [20] extended their Bernstein estimates to be independent of the distance to the singular points. However, the aforementioned spectral analysis forces us to consider analytic functions and sets rather than algebraic ones. Our estimates for analytic functions are not as sharp as for polynomials. This however is something which is absolutely natural due to elementary examples. For the same reason, the estimates for analytic functions require some "transversality conditions" since there is no way to prove the result in this setting upon the count of zeros alone.

The main result for us is Theorem A which addresses the mentioned applications to the spectral problem. In Theorem B we obtain a sharper version of Theorem A for the polynomial case, similar to [20]. The work of Roytwarf and Yomdin relies on a classical inequality for the Taylor coefficients of p-valent functions due to Biernacki [1]. In turn [1] relies on a deeper growth bound for p-valent functions obtained by Cartwright [4] (see [17] for a more detailed account of these issues). In Theorem B we show that in the context of algebraic curves, the Bernstein estimates by Roytwarf and Yomdin follow from a more elementary geometric approach in the spirit of the argument principle, without any reference to properties of p-valent functions. We also employ basic properties of the harmonic conjugate and of course Bezout's theorem (which is also needed in Roytwarf and Yomdin in order to estimate the valency). It seems that this approach can be developed for a general algebraic variety.

Regarding previous work in the analytic setting, we note that Coman-Poletsky [7] (for $n = 1$) and Brudnyi [3] (for all $n \geq 1$) studied Bernstein estimates (among other local properties) for the restriction of analytic functions of $n + 1$ variables to the graph of an analytic function of n variables. Both these papers naturally require a certain transversality condition of the zeros sets of the functions in question. We would like also to mention the Friedland–Yomdin paper [12] on doubling covering and Comte–Yomdin patper [8] on zeros of analytic functions which are close in spirit to the current paper.

We proceed to discuss the main results of the paper. First we need to introduce some notation related to Cartan sets and to Bernstein exponents. The Cartan sets will appear in our transversality condition to allow the application of the Cartan-type estimate established in Theorem C.

Definition 1.1

(1) Let $H \geq 0$, $K \geq 1$. For $\mathcal{B} \subset \mathbb{C}^2$, we say that $\mathcal{B} \in \mathrm{Car}_{2,0}(H, K)$ if

$$\mathcal{B} \subset \bigcup_{j=1}^{j_0} B(\underline{v}_j, r)$$

with $\underline{v}_j \in \mathbb{C}^2$, $r = e^{-H}$, and some $j_0 \leq K$.

(2) Let f be analytic on the ball $B(\underline{v}_0, R) \subset \mathbb{C}^2$, $\mathcal{S} \subset \mathbb{C}^2$, and $\mu \in (0, 1)$. Define

$$M_f(\underline{v}_0, R) = \sup_{B(\underline{v}_0, R)} \log |f|, \quad M_f(\mathcal{S}, \underline{v}_0, R) = \sup_{B(\underline{v}_0, R) \cap \mathcal{S}} \log |f|,$$

$$B_f(\mu; \underline{v}_0, R) = M_f(\underline{v}_0, R) - M_f(\underline{v}_0, \mu R),$$

$$B_f(\mu; \mathcal{S}, \underline{v}_0, R) = M_f(\mathcal{S}, \underline{v}_0, R) - M_f(\mathcal{S}, \underline{v}_0, \mu R).$$

We call $B_f(\mu; \underline{v}_0, R)$, $B_f(\mu; \mathcal{S}, \underline{v}_0, R)$ *Bernstein exponents*. We make the natural convention that if the function f vanishes identically, its Bernstein exponents are zero.

(3) Let f be analytic on $B(0, 1)$, $\mu \in (0, 1)$. We define

$$B_f(\mu) = \sup_{\underline{v}_0 \in B(0,1/4), 0 < R \leq 1/4} B_f(\mu; \underline{v}_0, R).$$

(4) Given an analytic function f on a disk $\mathcal{D}(z_0, R) \subset \mathbb{C}$, the quantities $M_f(z_0, R)$ and $B_f(\mu; z_0, R)$ are defined analogously to the above.

The classical Bernstein doubling inequality for a univariate polynomial f can be expressed using the above notation as

$$B_f(\mu; z_0, R) \leq (\log \mu^{-1}) \times \deg f,$$

where $\mu \in (0, 1)$, $z_0 \in \mathbb{C}$, $R > 0$.

We will use some standard conventions. Unless stated otherwise, the constants denoted by c, C might have different values each time they are used. We let $a \lesssim b$ denote $a \leq Cb$ with some positive C, and $a \simeq b$ stand for $a \lesssim b$ and $b \lesssim a$. By writing $a \ll b$, we mean that $a \leq Cb$ with a positive C, and furthermore the constant C is large enough for all related statements that we make based on $a \ll b$ to hold. All these constants are **absolute**. In particular if $a \ll b$ is part of the assumptions of a result, we mean that there exists a large enough absolute constant $C > 0$ such that if $a \leq Cb$, then the conclusion holds. We adopt such notation because we are not interested in the optimality of the implicit absolute constants. However, we note that throughout the paper, the implicit constants are absolute and can be determined explicitly.

Throughout we will impose the following *transversality condition*. Suppose the functions f_1, f_2 are analytic in the ball $B(0,1) \subset \mathbb{C}^2$ and are normalized so that $M_{f_i}(0,1) \leq 0$, $i = 1, 2$. We let $F = (f_1, f_2)$, and we define

$$\mathcal{N}_F(\varepsilon) := \{\underline{v} \in B(0,1) : |F(\underline{v})| < \varepsilon\}.$$

We require that

$$\mathcal{N}_F(\exp(-H_0)) \cap B(0,1/2) \in \mathrm{Car}_{2,0}(H_1, K_1), \quad \log K_1 \ll H_1, \tag{1}$$

for some $H_0 \gg H_1 \gg B_0 := \max_i B_{f_i}(1/4)$.

Remark 1.2 A priori it might appear that K_1 can be exponentially large, i.e., $\exp(cH_0)$ for some small $c > 0$. However, a simple argument, presented in Lemma 6.1, shows that we always have the polynomial bound $K_1 \leq H_0^C$, where C is some absolute constant.

Let $\mathcal{Z} = \{\underline{v} \in B(0,1) : f_2(\underline{v}) = 0\}$. It is well known that there exists a discrete set of singular points $\operatorname{sng} \mathcal{Z}$ (relative to $B(0,1)$) such that the set of regular points $\operatorname{reg} \mathcal{Z} := \mathcal{Z} \setminus \operatorname{sng} \mathcal{Z}$ is a one-dimensional complex manifold (see, e.g., [5]).

Theorem A *Assume the transversality condition holds and let $\mathcal{Z}$ be as above. Let $C_0 = \log(K_1 B_0^2 H_0^2)$. Then the following statements hold.*

(1) For any $\underline{v}_0 \in B(0,1/8) \cap \mathcal{Z}$ and $0 < R \leq 1/4$,

$$B_{f_1}(1/4; \mathcal{Z}, \underline{v}_0, R) \lesssim \max(\log R^{-1}, C_0) B_0^2 H_0.$$

(2) There exists an atlas of $\operatorname{reg} \mathcal{Z}$ with charts defined on $\mathcal{D}(0,1)$ such that for any chart ϕ satisfying $\phi(\mathcal{D}(0,1)) \cap B(0,1/8) \neq \emptyset$ and any $\mathcal{D}(z_0, R) \subset \mathcal{D}(0,1)$, we have

$$B_{f_1 \circ \phi}(1/4; z_0, R) \leq C(f_2) C_0 B_0^2 H_0.$$

Remark 1.3 The $\log R^{-1}$ factor from part (1) of Theorem A is needed because the estimate covers singular points. See Example 7.2.

Theorem B *Assume that f_1, f_2 are polynomials. Let $\mathcal{Z}$ be as above. Then there exists an atlas of $\operatorname{reg} \mathcal{Z}$ with charts defined on $\mathcal{D}(0,1)$ such that for any chart ϕ and any $\mathcal{D}(z_0, R) \subset \mathcal{D}(0,1)$, we have*

$$B_{f_1 \circ \phi}(1/4; z_0, R) \leq C(f_2) \times \deg f_1.$$

For our application in [16], we use the Cartan estimate for maps in $\mathbb{C}^2$ which is Theorem C we state below. The proof of Theorem A relies on Theorem C. The Cartan estimate for an analytic function $f(\underline{v})$, $\underline{v} \in \mathbb{C}^2$ (see Lemma 2.2), basically says that if the set $\{|f| < \varepsilon_0\}$ is "not two dimensional," then $\{|f| < \varepsilon\}$ is "one dimensional" for any $\varepsilon \ll \varepsilon_0$. We prove an analogue statement for mappings. Let $F : B(0,1) \subset \mathbb{C}^2 \to \mathbb{C}^2$ be analytic. We show that if the set $\{|F| < \varepsilon_0\}$ is "zero dimensional," then $\{|F| < \varepsilon\}$ is "zero dimensional" for any $\varepsilon \ll \varepsilon_0$. Of course, the quantitative details of the statement here are as important as the topological ones.

Theorem C *Assume the transversality condition holds. Then for any* $H \gg 1$, *we have*

$$\mathcal{N}_F(\exp(-HB_0^2H_0)) \cap B(0,1/4) \in \mathrm{Car}_{2,0}(H,K), \quad K \lesssim K_1 B_0^2 H_0^2.$$

The proof of Theorem C proceeds in four steps: (a) apply the Weierstrass preparation theorem to the given analytic functions in one of the two coordinates; (b) determine the resultant of the two polynomials obtained in the previous step; (c) apply Cartan's theorem in one variable so as to guarantee that this resultant is not too small off of a union of small disks in $\mathbb{C}$, which in turn gives that at least one of the two analytic functions is not too small outside of thin cylinders in $\mathbb{C}^2$; and (d) repeat the previous steps with respect to the other variable. The intersection of the two families of thin cylinders gives a $\mathrm{Car}_{2,0}$ set.

It would be interesting to extend this method to higher dimensions, i.e., to the construction of $\mathrm{Car}_{d,0}(H,K)$ sets with $d \geq 3$—at least for polynomials in d variables. In principle, this appears possible, but it seems to require the use of multivariate resultants, which are more delicate than the univariate ones. If Theorem C extends to $d \geq 3$, then one would obtain a Bernstein estimate as in Theorem A. As our applications do not require this extension, we do not pursue these matters here.

We conclude this introduction by providing some details of the aforementioned spectral theory applications. Consider a trigonometric polynomial of two variables

$$V(z,w) = \sum_{|m|,|n| \leq k} c_{m,n} e(mz + nw), \tag{2}$$

$e(\zeta) := e^{2\pi i \zeta}$. To normalize the setting, we consider the unit sphere in the space of the coefficients

$$\mathcal{C}_1 = \{(c_{m,n}) \in \mathbb{R}^{4k+2} : \sum_{m,n} |c_{m,n}|^2 = 1\}.$$

We use mes for the Lebesgue measure on the sphere. Take arbitrary $\omega \in \mathbb{T}^2$, $\lambda \in \mathbb{R}$. Consider the determinant

$$f_N(\underline{v}) = \begin{vmatrix} \lambda V(\underline{v}) & -1 & 0 & \cdots\cdots & & 0 \\ -1 & \lambda V(\underline{v}+\omega) & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ & & & & & -1 \\ 0 & \ldots\ldots\ldots & 0 & -1 & & \lambda V(\underline{v}+(N-1)\omega) \end{vmatrix} \tag{3}$$

For $\underline{v} \in \mathbb{R}^2$, $f_N(\underline{v})$ is the characteristic determinant of the Schrödinger operator with potential $V(\underline{v}+n\omega)$, $n \in \mathbb{Z}$ on the interval $[0, N-1]$ subject to Dirichlet boundary conditions. In [16], we establish the following results: *Given arbitrary* $\varepsilon > 0$, *there exists a set* $\mathcal{C} \subset \mathbb{R}^{4k+2}$ *with* $\operatorname{mes}(\mathcal{C}_1 \setminus \mathcal{C}) < \varepsilon$ *and* $\lambda_0 = \lambda_0(\varepsilon)$ *depending only on* ε *such that for any* V *with* $(c_{m,n}) \in \mathcal{C}_1$ *and any* $|\lambda| \geq \lambda_0$ *there exists a set* $\Omega(V) \subset \mathbb{T}^2$ *with* $\operatorname{mes}(\Omega(V)) < \varepsilon$ *such that for any* $\omega \in \mathbb{T}^2 \setminus \Omega(V)$, *any* N, *and any* $\underline{v}_0 \in \mathbb{T}^2$ *the functions* $f_N(\underline{v}_0 + r_0\underline{v})$ *and* $f_N(\underline{v}_n + r_0\underline{v})$, $\underline{v}_n = \underline{v}_0 + n\omega$, $|n| > N$, $\underline{v} \in B(0, 1)$, $r_0 = \exp(-(\log N)^A)$, and $A \gg 1$ *being an absolute constant, obey all conditions of Theorem A and Theorem C with* B_0, $H_0 \leq (\log N)^c$, $c \ll 1$.

The exceptional sets in this result are not artificial. In fact, the theorem fails for some $(c_{m,n}) \in \mathcal{C}$. A similar fact is true for the exceptional frequencies.

2 Cartan's Estimate

Recall the following definition from [14].

Definition 2.1 Let $H \geq 0$, $K \geq 1$. For an arbitrary set $\mathcal{B} \subset \mathbb{C}$, we say that $\mathcal{B} \in \mathrm{Car}_1(H, K)$ if $\mathcal{B} \subset \bigcup\limits_{j=1}^{j_0} \mathcal{D}(z_j, r_j)$ with $z_j \in \mathbb{C}$, $j_0 \leq K$, and $\sum_j r_j < e^{-H}$.

If $d \geq 1$ is an integer and $\mathcal{B} \subset \mathbb{C}^d$, then we define inductively that $\mathcal{B} \in \mathrm{Car}_d(H, K)$ if for any $1 \leq j \leq d$ there exists $\mathcal{B}_j \subset \mathbb{C}$, $\mathcal{B}_j \in \mathrm{Car}_1(H, K)$, so that $\mathcal{B}_z^{(j)} \in \mathrm{Car}_{d-1}(H, K)$ for any $z \in \mathbb{C} \setminus \mathcal{B}_j$, here $\mathcal{B}_z^{(j)} = \left\{(z_1, \ldots, z_d) \in \mathcal{B} : z_j = z\right\}$.

The above definition of Cartan sets is motivated by the following statement, known as Cartan estimate on the lower bound of an analytic function of several variables.

Lemma 2.2 ([14, Lem. 2.15]) *Let* $\varphi(z_1, \ldots, z_d)$ *be an analytic function defined in a polydisk* $\mathcal{P} = \prod\limits_{j=1}^{d} \mathcal{D}(z_{j,0}, 1)$, $z_{j,0} \in \mathbb{C}$. *Let* $M \geq \sup\limits_{\underline{z} \in \mathcal{P}} \log|\varphi(\underline{z})|$, $m \leq \log\left|\varphi(\underline{z}_0)\right|$, $\underline{z}_0 = (z_{1,0}, \ldots, z_{d,0})$. *There exists a constant* C_d *(depending only on the dimension* d*) such that for any given* $H \gg 1$ *there exists a set* $\mathcal{B} \subset \mathcal{P}$, $\mathcal{B} \in \mathrm{Car}_d\left(H^{1/d}, K\right)$, *and* $K = C_d H(M - m)$, *such that*

$$\log\bigl|\varphi(z)\bigr| > M - C_d H(M-m) \tag{4}$$

for any $z \in \prod_{j=1}^{d} \mathcal{D}(z_{j,0}, 1/6) \setminus \mathcal{B}$. *Furthermore, when* $d = 1$, *we can take* $K = C(M-m)$ *and keep only the disks of* $\mathcal{B}$ *containing a zero of* ϕ *in them.*

Remark 2.3

(1) The choice of the constant $1/6$ in [14, Lem. 2.15] was so that one could invoke the one-dimensional Cartan estimate as stated in Theorem 4 of [19, Lecture 11]. However, it is straightforward to adjust the result from [19] and the proof from [14] to replace $1/6$ by any $r < 1$. Of course, the constant C_d would depend (explicitly) on the particular choice of r.

(2) The definition of Cartan sets gives implicit information about their measure. For example, using Fubini and the definition of Car_d, one gets by induction that the set exceptional set $\mathcal{B}$ in the previous lemma satisfies $\mathrm{mes}_{\mathbb{C}^d}(\mathcal{B}) \le C(d)\exp(-H)$.

The following notion will be needed for our discussion of Weierstrass' preparation theorem.

Definition 2.4 Let f be analytic on the ball $B(\underline{v}_0, R_0) \subset \mathbb{C}^2$. Let $\mathfrak{e} \in \mathbb{C}^2$ be an arbitrary unit vector. We say that $\mathfrak{e}$ is m–regular for f at $\underline{v}_0$ (or just m–regular if it is clear from the context what $\underline{v}_0$ is) if

$$\sup_{z \in \mathcal{D}(0, R_0/4)} \log|f(\underline{v}_0 + z\mathfrak{e})| \ge m.$$

We show that Cartan's estimate implies that most directions are regular. We use σ to denote the standard spherical measure.

Lemma 2.5 *Let* f *be as in Definition 2.4 and let*

$$M \ge \sup_{B(\underline{v}_0, R_0)} \log|f|, \qquad \sup_{B(\underline{v}_0, R_0/4)} \log|f| \ge m.$$

Take arbitrary $H \gg 1$ *and set* $\underline{m} = M - C_2 H(M-m)$, *with* C_2 *as in Lemma 2.2. Denote by* $\mathcal{B}$ *the set of* $\mathfrak{e}$ *which are not* $\underline{m}$*–regular. Then*

$$\sigma(\mathcal{B}) \lesssim \exp(-H^{1/2}).$$

Proof Apply the Cartan estimate to find a set $\hat{\mathcal{B}}$, $\mathrm{mes}(\hat{\mathcal{B}}) \lesssim R_0^4 \exp(-H^{1/2})$, such that $\log\bigl|f(\underline{v})\bigr| > \underline{m}$ for any $\underline{v} \in B(\underline{v}_0, R_0/4) \setminus \hat{\mathcal{B}}$. Using spherical coordinates, write

$$\mathrm{mes}(\hat{\mathcal{B}}) \ge \int_{\mathcal{B}} d\sigma(\mathfrak{e}) \int_0^{R_0/4} r^3 dr \gtrsim R_0^4 \sigma(\mathcal{B})$$

and the statement follows. □

3 Bernstein Exponent and Number of Zeros

In this section we provide a relation between Bernstein exponents for one variable analytic functions and the number of their zeros.

Lemma 3.1 *Let ϕ be a non-vanishing analytic function on $\mathcal{D}(z_0, R)$. Then for any z, $|z - z_0| = r < R$, we have*

$$-\frac{2r}{R-r}(M - \log|\phi(z_0)|) \le \log|\phi(z)| - \log|\phi(z_0)| \le \frac{2r}{R+r}(M - \log|\phi(z_0)|),$$

where $M = M_\phi(z_0, R)$.

Proof The estimates follows immediately from Harnack's inequality applied to $u(z) = M - \log|\phi(z)|$. □

Proposition 3.2 *Let ϕ be an analytic function on $\mathcal{D}(0, 1)$ such that*

$$M_\phi(0, 1) \le 0, \quad M_\phi(0, 1/4) \ge m.$$

Let n be the total number of zeros of ϕ in $\mathcal{D}(0, 3/4)$. Then for any $|z_0| < 1/8$, $r < 1/8$, $\mu \in (0, 1)$, we have

$$B_\phi(\mu; z_0, r) \le Cr(n - m) - n\log\mu \lesssim -(r - \log\mu)m, \tag{5}$$

Proof Take $\zeta_0 \in \mathcal{D}(0, 1/4)$ wit $\log|f(\zeta_0)| = m$. Using Jensen's formula applied to

$$f\left(\frac{z + \zeta_0}{1 + \overline{\zeta_0}z}\right),$$

we get $n \lesssim -m$. So, we just have to prove the first estimate in (5).

Let $a_1, \dots, a_n$, be the zeros of ϕ in $\mathcal{D}(0, 7/8)$, repeated according to their multiplicities. Let $P(z) = \prod_{k=1}^{n}(z - a_k)$, $h = \phi/P$, and z_1, $|z_1 - z_0| = \mu r$, be such that $\log|h(z_1)| = M_h(z_0, \mu r)$. Note that h is non-vanishing and analytic on $\mathcal{D}(0, 3/4)$. Using Lemma 3.1 we have that for any $z \in \mathcal{D}(z_0, \mu r)$

$$\begin{aligned}\log|h(z)| &\ge \log|h(z_1)| - \frac{2|z - z_1|}{1/2 - |z - z_1|}(M_h(z_0, 1/2) - \log|h(z_1)|)\\ &\ge M_h(z_0, \mu r) - C\mu r(M_h(0, 3/4) - M_h(z_0, \mu r)).\end{aligned}$$

Therefore

$$M_\phi(z_0, \mu r) \ge M_h(z_0, \mu r) - C\mu r(M_h(0, 3/4) - M_h(z_0, \mu r)) + M_P(z_0, \mu r)$$

and

$$\begin{aligned} B_\phi(\mu; z_0, r) \le M_h(z_0, r) - M_h(z_0, \mu r) + C\mu r(M_h(0, 3/4) - M_h(z_0, \mu r)) \\ + M_P(z_0, r) - M_P(z_0, \mu r). \end{aligned} \tag{6}$$

Let z_2, $|z_2 - z_0| = r$, such that $\log|h(z_2)| = M_h(z_0, r)$, and z_3, $|z_3| = 1/4$, such that $\log|h(z_3)| = M_h(0, 1/4)$. Using Lemma 3.1, we get

$$\begin{aligned} &M_h(z_0, r) - M_h(z_0, \mu r) = \log|h(z_2)| - \log|h(z_1)| \\ &\quad \le \frac{2|z_2 - z_1|}{1/2 + |z_2 - z_1|}(M_h(z_1, 1/2) - \log|h(z_1)|) \le Cr(M_h(0, 3/4) - M_h(z_0, \mu r)), \end{aligned}$$

$$\begin{aligned} M_h(z_0, \mu r) - M_h(0, 1/4) &= \log|h(z_1)| - \log|h(z_3)| \\ &\ge -\frac{2|z_3 - z_1|}{1/2 - |z_3 - z_1|}(M_h(z_3, 1/2) - \log|h(z_3)|) \\ &\ge -C(M_h(0, 3/4) - M_h(0, 1/4)). \end{aligned}$$

Plugging these estimates in (6), we get

$$B_\phi(\mu; z_0, r) \le Cr(M_h(0, 3/4) - M_h(0, 1/4)) + B_P(\mu; z_0, r). \tag{7}$$

Recall that we know $B_P(\mu; z_0, r) \le -n\log\mu$, so to get the conclusion, we just have to estimate $M_h(0, 3/4) - M_h(0, 1/4)$. Given $z \in \mathcal{D}(0, 3/4)$, apply the submean value property to get

$$\log|h(z)| \le \frac{1}{2\pi}\int_0^{2\pi} \log|\phi(z+e^{i\theta}/4)|\, d\theta - \frac{1}{2\pi}\int_0^{2\pi} \log|P(z+e^{i\theta}/4)|\, d\theta \le n\log 4$$

and conclude $M_h(0, 3/4) \le n\log 4$. We used the assumption that $M_\phi(0, 1) \le 0$ and the fact that

$$\frac{1}{2\pi}\int_0^{2\pi} \log|z - a_k + e^{i\theta}/4|\, d\theta = \begin{cases} \log\frac{1}{4} & , |z - a_k| \le \frac{1}{4} \\ \log|z - a_k| & , |z - a_k| > \frac{1}{4} \end{cases} \ge \log\frac{1}{4}. \tag{8}$$

Since clearly $M_P(0, 1/4) \le 0$, we have $M_h(0, 1/4) \ge M_\phi(1/4)$. So,

$$M_h(0, 3/4) - M_h(0, 1/4) \le C(n - m)$$

and the conclusion follows. □

Remark 3.3

(1) It is not true that conclusion of Proposition 3.2 can be made just in terms of the number n of zeros of ϕ. Some estimate for $M_\phi(0, 1/4)$ is really needed. Here,

an elementary example $\phi(z) = \exp(-N + Nz)$, $z \in \mathcal{D}(0, 1)$, and $N > 0$ is arbitrary. Clearly, $M_\phi(0, 1) = 0, n = 0$. On the other hand, $M_\phi(0, 1/4) \simeq -N$, $B_\phi(1/4; 0, 1/8) \simeq N$.

(2) It is known from [20] that if we have control on the valency of the function ϕ, instead of just the number of zeros, then the estimate for $M_\phi(0, 1/4)$ is not needed anymore.

4 Weierstrass' Preparation Theorem and Bernstein Exponents

We start with a statement of the classical Weierstrass' preparation theorem attuned to our purposes.

Lemma 4.1 *Let $f(z, w)$ be analytic function on a polydisk*

$$\mathcal{P} := \mathcal{D}(z_0, R_0) \times \mathcal{D}(w_0, R_0) \subset \mathbb{C}^2, \qquad R_0 > 0.$$

Assume that $f(\cdot, w)$ has no zeros on some circle $\Gamma_{\rho_0} = \{z : |z - z_0| = \rho_0\}$, $0 < \rho_0 < R_0/2$, for any $w \in \mathcal{D}(w_0, r_1)$, $0 < r_1 < R_0$. Then there exists a Weierstrass polynomial $P(z, w) = z^k + a_{k-1}(w)z^{k-1} + \cdots + a_0(w)$ with $a_j(w)$ analytic in $\mathcal{D}(w_0, r_1)$ and an analytic function $g(z, w)$, $(z, w) \in \mathcal{P}' := \mathcal{D}(z_0, \rho_0) \times \mathcal{D}(w_0, r_1)$ so that the following properties hold:

(a) $f(z, w) = P(z, w)g(z, w)$ for any $(z, w) \in \mathcal{P}'$.
(b) $g(z, w) \neq 0$ for any $(z, w) \in \mathcal{P}'$.
(c) For any $w \in \mathcal{D}(w_0, r_1)$, $P(\cdot, w)$ has no zeros in $\mathbb{C} \setminus \mathcal{D}(z_0, \rho_0)$.
(d) We have

$$\left(\inf_{\Gamma_{\rho_0} \times \mathcal{D}(w_0, r_1)} \log |f|\right) - k \log(2\rho_0) \le \inf_{\mathcal{P}'} \log |g|, \tag{9}$$

$$\sup_{\mathcal{P}'} \log |g| \le \left(\sup_{\mathcal{P}} \log |f|\right) + k \log \frac{2}{R_0}. \tag{10}$$

Proof By the usual Weierstrass argument, one notes that

$$b_p(w) := \sum_{j=1}^{k} \zeta_j^p(w) = \frac{1}{2\pi i} \oint_\Gamma z^p \frac{\partial_z f(z, w)}{f(z, w)} dz$$

are analytic in $\mathcal{D}(w_0, r_1)$. Here $\zeta_j(w)$ are the zeros of $f(\cdot, w)$ in $\mathcal{D}(z_0, \rho_0)$. Since the coefficients $a_j(w)$ are linear combinations of the b_p, they are analytic in w. Analyticity of g follows by standard arguments. We just have to prove (d). Since all the roots of $P(\cdot, w)$ are in $\mathcal{D}(z_0, \rho_0)$, we have $\sup_{\mathcal{P}'} |P| \leq (2\rho_0)^k$ and (9) follows using the minimum modulus principle. Note that actually the function g can be defined on $\mathcal{P}$ as $g = f/P$ and it is analytic there. Given $(z, w) \in \mathcal{P}'$, apply the sub-mean value property for subharmonic functions to get

$$\begin{aligned}\log|g(z,w)| &\leq \frac{1}{2\pi}\int_0^{2\pi} \log|f(z+R_0e^{i\theta}/2, w)|\,d\theta \\ &\quad - \frac{1}{2\pi}\int_0^{2\pi} \log|P(z+R_0e^{i\theta}/2, w)|\,d\theta \\ &\leq \left(\sup_{\mathcal{P}} \log|f|\right) + k\log\frac{2}{R_0}.\end{aligned}$$

The estimate on the mean value of the polynomial follows by considerations analogous to (8). □

Next we describe how Bernstein exponents rule the application of Lemma 4.1.

Lemma 4.2 *Let f be analytic on $B(0,1)$, $M \geq \sup_{B(0,1)} \log|f|$, $\underline{m} = M - B$, $B \gg 1$, $\mathfrak{e}_1$ a $\underline{m}$-regular direction for f at 0 (recall Definition 2.4), and $\mathfrak{e}_2$ another non-collinear direction. With a slight abuse of notation, we denote by $f(z, w)$ the function in the new coordinates with respect to the basis $\mathfrak{e}_1, \mathfrak{e}_2$. Then there exists a circle $\Gamma_{\rho_0} = \{|z| = \rho_0\}$, $1/8 < \rho_0 < 1/4$, and $r_1 = \exp(-CB)$, with $C > 1$ an absolute constant, such that*

$$\inf_{\Gamma_{\rho_0}\times\mathcal{D}(0,r_1)} \log|f| \geq \exp(M - CB). \tag{11}$$

In particular, Lemma 4.1 applies for $f(z, w)$ with this choice of ρ_0 and r_1, as well as with $k \lesssim B$ and $\delta \geq M - CB$.

Proof Since $\mathfrak{e}_1$ is a $\underline{m}$-regular direction, there exists z_1, $|z_1| = 1/4$, such that $\log|f(z_1, 0)| \geq \underline{m}$. Due to Cartan's estimate, one has

$$\log|f(z,0)| \geq M - C(M - \underline{m}) = M - CB \tag{12}$$

for any $z \in \mathcal{D}(0, 1/4) \setminus \mathcal{B}$, where $\mathcal{B} \in \mathrm{Car}_1\left(C', C'B\right)$, $C' \gg 1$. As a consequence of the definition of Car_1 sets, we can choose $1/8 < \rho_0 < 1/4$ such that $\mathcal{B} \cap \Gamma_{\rho_0} = \emptyset$. Then

$$|f(z,0)| \geq \exp(M - CB) \tag{13}$$

for any $z \in \Gamma_{\rho_0}$. Note that due to Cauchy's estimates

$$|f(z,w) - f(z,0)| \lesssim e^M |w|$$

for any $z \in \mathcal{D}(0, 1/2)$, $w \in \mathcal{D}(0, 1/2)$. Taking into account (13), one obtains

$$|f(z, w)| > \exp(M - CB)$$

for any $z \in \Gamma_{\rho_0}$, provided $w \in \mathcal{D}(0, r_1)$, $r_1 = \exp(-CB)$, with C large enough (of course, C is larger than in (13)). This proves (11) and allows us to apply Lemma 4.1 as stated. For the bound on the degree of the Weierstrass polynomial note that by Jensen's formula applied to $f(z, 0)$, $z \in \mathcal{D}(z_1, 1/2)$,

$$k \le \#\{z \in \mathcal{D}(0, 1/4) : f(z, 0) = 0\} \le \#\{z \in \mathcal{D}(z_1, 1/2) : f(z, 0) = 0\} \lesssim B.$$

□

Remark 4.3

(1) Due to Lemma 2.5, we will always apply the previous lemma with $B \simeq B_f(1/4; 0, 1)$. This is how the Bernstein exponent determines the size of the polydisk on which we have the Weierstrass factorization.

(2) If we are given two functions f_1, f_2 satisfying the assumptions of Lemma 4.2 with the same M and B, then it is clear from the proof of the lemma that we can arrange for the conclusion to hold for both functions with the same choice of ρ_0 and r_1. Indeed, one only needs to choose ρ_0 such that $\Gamma_{\rho_0} \cap (\mathcal{B}_1 \cup \mathcal{B}_2) = \emptyset$, where $\mathcal{B}_i$ are the Cartan sets needed to guarantee (12) for f_i.

5 Resultants

We briefly recall the definition of the resultant of two univariate polynomials and some of the basic properties that we'll use. Let $f(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_0$, $g(z) = b_m z^m + b_{m-1} z^{m-1} + \cdots + b_0$ be polynomials, $a_i, b_j \in \mathbb{C}$, $a_n \neq 0$, $b_m \neq 0$. Let ζ_i, $1 \le i \le n$ and η_j, $1 \le j \le m$ be the zeros of $f(z)$ and $g(z)$, respectively. The resultant of f and g is defined as follows:

$$\operatorname{Res}(f, g) = a_n^m b_m^n \prod_{i,j} (\zeta_i - \eta_j) = (-1)^{mn} b_m^n \prod_j f(\eta_j) = (-1)^{mn} a_n^m \prod_i g(\zeta_i). \tag{14}$$

The resultant $\operatorname{Res}(f, g)$ can be expressed explicitly in terms of the coefficients (see [18]):

$$\mathrm{Res}(f, g) = \left|\begin{array}{cccccc} \multicolumn{3}{c}{\overbrace{\hphantom{a_{n-2}\ a_{n-1}\ \cdots}}^{m}} & \multicolumn{3}{c}{\overbrace{\hphantom{b_{m-2}\ b_{m-1}\ \cdots\cdots}}^{n}} \\ a_n & 0 & \cdots & b_m & 0 & \cdots 0 \\ a_{n-1} & a_n & \cdots & b_{m-1} & b_m & \cdots\cdots \\ a_{n-2} & a_{n-1} & \cdots & b_{m-2} & b_{m-1} & \cdots\cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\cdots \\ a_0 & a_1 & & & & \\ 0 & a_0 & & & & \end{array}\right| \tag{15}$$

Lemma 5.1 *Let f, g, ζ_i, η_j as above. Set*

$$t_f = \min(|a_n|, 1), \quad t_g = \min(|b_m|, 1), \quad T_f = \max_i(\max|a_i|, 1),$$

$$T_g = \max_j(\max|b_j|, 1),$$

$$R_f = t_f^{-1} T_f m, \quad R_g = t_g^{-1} T_g n.$$

The following statements hold.

(0) $\max|\zeta_i| \le R_f, \quad \max|\eta_j| \le R_g$.

(1) *If*

$$\left|\mathrm{Res}(f, g)\right| < \delta^m t_g^n, \quad 0 < \delta < 1,$$

then there exists j such that

$$\left|f(\eta_j)\right| < \delta.$$

In particular, there exists $|z| \le R_g$ such that $\max(\left|f(z)\right|, \left|g(z)\right|) < \delta$.

(2) *If there exists z such that with $s = \max(m, n)$, $t = \min(t_f, t_g)$ holds*

$$\max[\left|f(z)\right|, \left|g(z)\right|] < t\delta^s, \quad 0 < \delta < 1,$$

then

$$\left|\mathrm{Res}(f, g)\right| < t^{2s}(2R)^{s^2}\delta,$$

$R = \max(R_f, R_g)$.

Proof (0) follows by noting that, for example,

$$|a_n||\zeta_i|^n \le (\max|a_i|)(|\zeta_i|^{n-1} + \cdots + |\zeta_i| + 1) \le (\max|a_i|)n\max(|\zeta_i|^{n-1}, 1).$$

(1) follows by contradiction from (14). For (2) note that there must exist ζ_{i_0}, η_{j_0} such that $|z - \zeta_{i_0}| < \delta$, $|z - \eta_{j_0}| < \delta$ and therefore, using (0) and (14),

$$|\mathrm{Res}(f, g)| \leq t^{2s}(2R)^{s^2-1}|\zeta_{i_0} - \eta_{j_0}| < t^{2s}(2R)^{s^2}\delta.$$

□

6 Refinement of the Assumption (1)

We give a simple argument showing that by making some small adjustments, we actually have $K_1 \leq H_0^C$ in (1).

Lemma 6.1 *Using the notation and assumptions of Theorem C, we have that*

$$\mathcal{N}(F, \varepsilon_0/2) \cap B(0, 1/2) \in \mathrm{Car}_{2,0}(H_1/2, H_0^C)$$

where C is some large absolute constant.

Proof Let $f_{i,N}$ be the degree N Taylor polynomials (at the origin) associated with f_i, $i = 1, 2$ (recall that $F = (f_1, f_2)$). Since $M_{f_i}(0, 1) \leq 0$, a standard application of the Cauchy estimates yields that

$$|f_i - f_{i,N}| < \varepsilon_0/100$$

for $N = C \log \varepsilon_0^{-1} = CH_0$, $C \gg 1$. Let $F_N = (f_{1,N}, f_{2,N})$. We have

$$\begin{aligned} \mathcal{N}(F, \varepsilon_0/2) \cap B(0, 1/2) &\subset \mathcal{N}(F_N, 3\varepsilon_0/4) \cap B(0, 1/2) \\ &\subset \mathcal{N}(F, \varepsilon_0) \cap B(0, 1/2). \end{aligned} \tag{16}$$

The set $\mathcal{N}(F_N, 3\varepsilon_0/4) \cap B(0, 1/2)$ is semialgebraic of degree less than CN, and therefore it has at most N^C connected components. We refer to [2, Ch. 9] for a brief review of semialgebraic sets and their properties. It follows from our assumptions that $\mathcal{N}(F_N, 3\varepsilon_0/4) \cap B(0, 1/2)$ is covered by less than K_1 balls of radius $\exp(-H_1)$. Therefore, each connected component of $\mathcal{N}(F_N, 3\varepsilon_0/4) \cap B(0, 1/2)$ can be covered by just one ball of radius smaller than

$$CK_1 \exp(-H_1) \leq \exp(-H_1/2)$$

(recall that $\log K_1 \ll H_1$) and so

$$\mathcal{N}(F_N, 3\varepsilon_0/4) \cap B(0, 1/2) \in \mathrm{Car}_{2,0}(H_1/2, N^C).$$

The conclusion now follows from (6). □

7 Proofs of Theorems A, B, and C

We start with the proof of Theorem C.

Proof of Theorem C Take $\underline{v}_0 = (z_0, w_0) \in B(0, 1/4)$. By our assumptions

$$B_{f_i}(1/4; \underline{v}_0, 1/4), \quad M_{f_i}(\underline{v}_0, 1/4) \leq 0.$$

Due to Lemma 2.5, we can find unit vectors $\mathfrak{e}_1, \mathfrak{e}_2$, $|\langle \mathfrak{e}_1, \mathfrak{e}_2 \rangle| \ll 1$, that are $\underline{m}$-regular at $\underline{v}_0$ for both f_1, f_2 restricted to $B(\underline{v}_0, 1/4)$, with $\underline{m} = -CB_0$, $C \gg 1$. Then Lemma 4.2 applies to both f_1, f_2 and to both directions $\mathfrak{e}_1, \mathfrak{e}_2$. As in Lemma 4.2, with a slight abuse of notation, we denote by $f_i(z, w)$ the functions in the coordinates with respect to the basis $\mathfrak{e}_1, \mathfrak{e}_2$ centered at $\underline{v}_0$ and with the obvious rescaling needed to apply the lemma. Applying Lemma 4.2 (see also Remark 4.3) in the direction of $\mathfrak{e}_1$ (and $\mathfrak{e}_2$ as the choice of non-collinear direction), we can write

$$\begin{aligned} f_i(z, w) &= P_i(z, w) g_i(z, w), \\ P_i(z, w) &= z^{k_i} + a_{i,k_i-1}(w) z^{k_i-1} + \cdots + a_0(w) \end{aligned}$$

on $\mathcal{P} := \mathcal{D}(0, \rho_0) \times \mathcal{D}(0, r_1)$, $1/8 < \rho_0 < 1/4$, $r_1 = \exp(-CB_0)$, where the coefficients $a_{i,j}(w)$ are analytic on $\mathcal{D}(0, r_1)$, g_i are analytic and non-vanishing on $\mathcal{P}$, the polynomials $P_i(\cdot, w)$ and $w \in \mathcal{D}(0, r_1)$ have no zeroes in $\mathbb{C} \setminus \mathcal{D}(0, \rho_0)$, and $k_i \lesssim B_0$. Furthermore, using part (d) of Lemma 4.1,

$$-B_0 \lesssim \inf_{\mathcal{P}} \log |g_i| \leq \sup_{\mathcal{P}} \log |g_i| \lesssim B_0. \tag{17}$$

Let

$$R(w) = \operatorname{Res}\left(P_1(\cdot, w), P_2(\cdot, w)\right).$$

Note that by (15), R is analytic on $\mathcal{D}(0, r_1)$. Since we chose $\underline{v}_0 \in B(0, 1/4)$, the polydisk $\mathcal{P}$ is a subset of $B(0, 1/2)$, as a set in the standard coordinates. This allows to use the hypothesis to guarantee that there exist points $\underline{v}_j = (z_j, w_j)$ (expressed in the $\mathfrak{e}_1, \mathfrak{e}_2$ coordinates), $1 \leq j \leq J \leq K_1$ such that for

$$(z, w) \in \mathcal{P} \setminus \left(\bigcup_{j=1}^{J} B(\underline{v}_j, C \exp(-H_1)) \right)$$

we have

$$\max(|f_1(z, w)|, |f_2(z, w)|) \geq \exp(-H_0)/\sqrt{2}$$

and by (17)

$$\max(|P_1(z,w)|, |P_2(z,w)|) \gtrsim \exp(-H_0 - CB_0). \tag{18}$$

Note that we used the radius $C\exp(-H_1)$ instead of $\exp(-H_1)$ to account for the distortion under the change of coordinates. Since we are assuming that $H_1 \gg B_0$ and $\log K_1 \ll H_1$, we can find

$$w \in \mathcal{D}(0, r_1/4) \setminus \bigcup_{j=1}^{J} \mathcal{D}(w_j, C\exp(-H_1)).$$

For any such w (18) holds for any $z \in \mathcal{D}(0,\rho_0)$ and by part (1) of Lemma 5.1

$$\log|R(w)| \gtrsim -B_0H_0 - B_0^2 \gtrsim -B_0H_0.$$

Note that by the definition of the resultant (14), we have $\sup_{\mathcal{D}(0,r_1)}|R(w)| \le 1$. Take $H \gg 1$. Applying Cartan's estimate, we get

$$\log|R(w)| \gtrsim -HB_0H_0$$

for any

$$w \in \mathcal{D}(0, r_1/4) \setminus \mathcal{B}, \quad \mathcal{B} = \bigcup_{1\le k\le K} \mathcal{D}(w_k', r_1\exp(-H)), \quad K \lesssim B_0H_0.$$

By part (2) of Lemma 5.1,

$$\max(|P_1(z,w)|, |P_2(z,w)|) \ge \exp(-CHB_0^2H_0)$$

for any $w \in \mathcal{D}(0, r_1/4) \setminus \mathcal{B}$ and $z \in \mathbb{C}$. Using (17), we get

$$|F(z,w)| \gtrsim \exp(-CHB_0^2H_0 - CB_0) \ge \exp(-HB_0^2H_0) \tag{19}$$

for any $w \in \mathcal{D}(0, r_1/4) \setminus \mathcal{B}$ and $z \in \mathcal{D}(0,\rho_0)$. Applying Lemma 4.2 again in the direction of $\mathfrak{e}_2$ (and with $\mathfrak{e}_1$ as the choice of non-collinear direction) and repeating the above argument, we get that there exist $1/8 < \tilde{\rho}_0 < 1/4$, $\tilde{r}_1 = \exp(-CB_0)$, such that (19) also holds for any $z \in \mathcal{D}(0, \tilde{r}_1/4) \setminus \tilde{\mathcal{B}}$ and

$$w \in \mathcal{D}(0, \tilde{\rho}_0), \quad \tilde{\mathcal{B}} = \bigcup_{1\le \ell\le L} \mathcal{D}(z_\ell', \tilde{r}_1\exp(-H)), \quad L \lesssim B_0H_0,$$

In particular, (19) holds for any

$$(z,w) \in \mathcal{D}(0,\tilde{r}_1/4) \times \mathcal{D}(0, r_1/4) \setminus \left(\bigcup_{k,\ell} \mathcal{D}(z_\ell', \tilde{r}_1\exp(-H)) \times \mathcal{D}(w_k', r_1\exp(-H))\right).$$

Going back to standard coordinates, we obtained that there exist less than $CB_0^2H_0^2$ points $\underline{v}'_j$ such that (19) holds for any

$$(z, w) \in B(\underline{v}_0, \exp(-CB_0)) \setminus \left(\bigcup_j B(\underline{v}'_j, \exp(-H)) \right).$$

Since (19) holds outside the initial $\mathrm{Car}_{2,0}$ set, we only need to apply the above argument on K_1 balls covering the initial set to get the conclusion. □

We will need the following lemma for the proof of Theorem A.

Lemma 7.1 *Let f be analytic on the ball $B(0, 1)$, $\mathcal{Z} = \{\underline{v} \in B(0, 1) : f(\underline{v}) = 0\}$. Let $\mathcal{B} \in \mathrm{Car}_{2,0}(H_1, K_1)$, $H \gg 1$, $\log K \ll H$. If $0 \in \mathcal{Z}$, then*

$$B(0, 1/4) \cap \mathcal{Z} \setminus \mathcal{B} \neq \emptyset.$$

Proof We argue by contradiction. Assume $B(0, 1/4) \cap \mathcal{Z} \subset \mathcal{B}$. By the assumptions on $\mathcal{B}$, we can find $1/8 < r < 1/4$ such that $\mathcal{B} \cap B(0, r)$ is compactly contained in $\mathcal{B}$. Therefore the zero set of f restricted to $B(0, r)$ is compactly contained in $B(0, r)$ and $\mathcal{Z} \cap B(0, r)$ is a compact analytic variety in $\mathbb{C}^2$. This cannot be, because compact analytic varieties in $\mathbb{C}^2$ are necessarily finite sets (see, e.g., [5]) and analytic functions of several variables cannot have isolated zeros (recall that $0 \in \mathcal{Z}$). □

Proof of Theorem A

(1) Take $\underline{v}_0 \in B(0, 1/8) \cap \mathcal{Z}$, $\mathcal{Z} = \{f_2 = 0\}$, $0 < R \leq 1/4$. Let $H = C \max(\log R^{-1}, C_0)$ with C large enough (recall that $C_0 = \log(K_1 B_0^2 H_0^2)$). By Theorem C, we have

$$|F(\underline{v})| \geq \exp(-HB_0^2H_0)$$

for all $\underline{v} \in B(0, 1/4) \setminus \mathcal{B}$, $\mathcal{B} \in \mathrm{Car}_{2,0}(H, K)$, $K \lesssim K_1B_0^2H_0^2$. Note that $B(\underline{v}_0, R) \subset B(0, 1/4)$ and our choice of H is such that we can apply Lemma 7.1 to f_2 restricted to $B(\underline{v}_0, R)$ and the above $\mathcal{B}$ (after an obvious rescaling). So, there exists $\underline{v}_1 \in \mathcal{B}(\underline{v}_0, R/4) \cap \mathcal{Z} \setminus \mathcal{B}$. Note that we have

$$|f_1(\underline{v}_1)| = |F(\underline{v}_1)| \geq \exp(-HB_0^2H_0)$$

and therefore

$$M_{f_1}(\mathcal{Z}, \underline{v}_0, R/4) \geq -HB_0^2H_0.$$

The first statement now follows by recalling that

$$M_{f_1}(\mathcal{Z}, \underline{v}_0, R) \leq M_{f_1}(0, 1) \leq 0.$$

(2) Take $\underline{v}_0 \in B(0, 1/8)$. By our assumptions

$$B_{f_2}(1/4; \underline{v}_0, 1/4) \leq B_0, \quad M_{f_2}(\underline{v}_0, 1/4) \leq 0.$$

Due to Lemma 2.5, we can find a unit vector $\mathfrak{e}_1$, that is $\underline{m}$-regular at $\underline{v}_0$ for f_2 restricted to $B(\underline{v}_0, 1/4)$, with $\underline{m} = -CB_0$, $C \gg 1$. Let $\mathfrak{e}_2$ be another unit vector orthogonal to $\mathfrak{e}_1$. As in Lemma 4.2, with a slight abuse of notation, we denote by $f_i(z, w)$ the functions in the coordinates with respect to the basis $\mathfrak{e}_1, \mathfrak{e}_2$ centered at $\underline{v}_0$ and with the obvious rescaling needed to apply the lemma. Applying Lemma 4.2 in the direction of $\mathfrak{e}_1$ (with $\mathfrak{e}_2$ as the choice of non-collinear direction), we can write

$$\begin{aligned} f_2(z, w) &= P(z, w)g(z, w), \\ P(z, w) &= z^k + a_{k-1}(w)z^{k-1} + \cdots + a_0(w) \end{aligned}$$

on $\mathcal{P} := \mathcal{D}(0, \rho_0) \times \mathcal{D}(0, r_1)$, $1/8 < \rho_0 < 1/4$, $r_1 = \exp(-CB_0)$, where the coefficients $a_j(w)$ are analytic on $\mathcal{D}(0, r_1)$ and $k \lesssim B_0$. Since we also have that g is analytic and non-vanishing on $\mathcal{P}$,

$$\mathcal{Z} \cap \mathcal{P} = \mathcal{Z}_P \cap \mathcal{P}, \quad \mathcal{Z}_P := \{(z, w) \in \mathbb{C} \times \mathcal{D}(0, r_1) : P(z, w) = 0\}.$$

It is well known (see [5]) that for any point (z, w) of the variety $\mathcal{Z}_P$, there exist $\varepsilon > 0$, $\delta > 0$ such that the following statements hold.

(i) If (z, w) is a regular point, then there exists an analytic function $\zeta : \mathcal{D}(0, \varepsilon) \to \mathcal{D}(0, \delta)$ such that

$$\mathcal{Z}_P \cap (\mathcal{D}(z, \delta) \times \mathcal{D}(w, \varepsilon)) = \{(z + \zeta(w' - w), w') : w' \in \mathcal{D}(w, \varepsilon)\}.$$

(ii) If (z, w) is a singular point, then there exist integers $p_i \geq 1$ and analytic functions $\zeta_i : \mathcal{D}(0, \varepsilon) \to \mathcal{D}(0, \delta)$, $1 \leq i \leq i_0(z, w) \leq k$, such that $\sum_i p_i \leq k$ and

$$\mathcal{Z}_P \cap (\mathcal{D}(z, \delta) \times \mathcal{D}(w, \varepsilon)) = \bigcup_i \{(z + \zeta_i((w' - w)^{\frac{1}{p_i}}), w') : w' \in \mathcal{D}(w, \varepsilon)\}.$$

By compactness we can cover $B(0, 1/8) \cap \mathcal{Z}$ by finitely many polydisks $\frac{1}{2}\mathcal{P}$ (more precisely, by their preimages under the change of variables we assumed above), and in turn $\mathcal{Z} \cap \frac{1}{2}\mathcal{P}$ can be covered by finitely many polydisks $\mathcal{D}(z_j, \delta_j) \times \mathcal{D}(w_j, \varepsilon_j/8)$ with $(z_j, w_j) \in \mathcal{Z}_P$ and ε_j, δ_j as above. We will also use ζ_j and $\zeta_{i,j}$ the functions associated with (z_j, w_j). Let $r_0 > 0$ be the minimum over all the ε_j needed to cover $B(0, 1/8) \cap \mathcal{Z}$. Near each (z_j, w_j), we will define local charts and show we can control the Bernstein exponent of f_1 in the local charts. The control over the Bernstein exponent will follow from Theorem C and Proposition 3.2. To this end, we take $H = C(\log r_0^{-1})C_0$, with $C \gg 1$ large enough, and we note that, with this choice of H, Theorem C guarantees that

$$|F(z, w)| \geq \exp(-HB_0^2 H_0),\ \forall (z, w) \in \mathcal{P} \setminus (\mathbb{C} \times \mathcal{B}) \tag{20}$$

where $\mathcal{B}$ is a union of disks with the sum of the radii much smaller than r_0^k (recall that $k \lesssim B_0 \ll H_0$). To define the charts, we distinguish two cases.
(i) (z_j, w_j) is regular. Let

$$\psi_j(w) = (z_j + \zeta_j(w), w_j + w),\ w \in \mathcal{D}(0, \varepsilon_j).$$

It follows from (20) that

$$M_{f_1 \circ \psi_j}(0, 1/4) \geq -HB_0^2 H_0.$$

By Proposition 3.2 (recall that $M_{f_1}(0, 1) \leq 0$), it is clear that

$$B_{f_1 \circ \psi_j}(1/4; z, r) \lesssim H \leq C(f_2)C_0 B_0^2 H_0$$

when $\mathcal{D}(z, r) \subset \mathcal{D}(0, \varepsilon_j/8)$. This shows the conclusion of part (2) holds if we define the local chart by rescaling $\psi_j|_{\mathcal{D}(0,\varepsilon_j/8)}$.
(ii) (z_j, w_j) is singular. Let

$$\psi_{i,j}(w) = (z_j + \zeta_{i,j}(w), w_j + w^{p_i}),\ w \in \mathcal{D}(0, \varepsilon_j).$$

It follows from (20) that

$$M_{f_1 \circ \psi_{i,j}}(0, 1/4) \geq -HB_0^2 H_0$$

(recall that $p_i \leq k$), and therefore Proposition 3.2 guarantees that

$$B_{f_1 \circ \psi_{i,j}}(1/4; z, r) \lesssim HB_0^2 H_0 \leq C(f_2)C_0 B_0^2 H_0$$

when $\mathcal{D}(z, r) \subset \mathcal{D}(0, \varepsilon_j/8)$. This shows that the conclusion holds if corresponding to each $w \in \mathcal{D}(0, \varepsilon_j/8) \setminus \{0\}$ we define a local chart by rescaling $\psi_{i,j}|_{\mathcal{D}(w,r)}$, where $\mathcal{D}(w, r)$ is the largest disk about w in $\mathcal{D}(0, \varepsilon_j/8)$ on which w^{p_i} is one to one.

Clearly the above charts cover $\operatorname{reg} \mathcal{Z} \cap B(0, 1/8)$, and we can complete an atlas of $\operatorname{reg} \mathcal{Z}$ by adding charts whose ranges don't intersect $B(0, 1/8)$. This concludes the proof. □

Next we give an example showing that the $\log R^{-1}$ is actually necessary in part (1) of Theorem A.

Example 7.2 Let

$$f_1(z,w) = z^2 + w, \quad f_2(z,w) = zw, \quad \mathcal{Z} = \{f_2 = 0\}.$$

Let $R \ll 1$, $\underline{v}_0 = (R/4, 0)$. Then straightforward computations show that

$$\sup_{\mathcal{B}(\underline{v}_0, R/4)\cap\mathcal{Z}} \log|f_1(z,w)| = \sup_{|z-R/4|<R/4} \log|z^2| = \log\left(\frac{R}{2}\right)^2,$$

and

$$\sup_{\mathcal{B}(\underline{v}_0, R)\cap\mathcal{Z}} \log|f_1(z,w)| = \max\left(\sup_{|z-R/4|<R} \log|z^2|, \sup_{|w|^2+(R/4)^2<R^2} \log|w|\right)$$

$$= \max\left(\log\left(\frac{5R}{4}\right)^2, \log\frac{\sqrt{15}R}{4}\right) = \log\frac{\sqrt{15}R}{4},$$

provided R is small enough ($R < 1/2$ is enough). Therefore,

$$B_{f_1}(1/4; \mathcal{Z}, \underline{v}_0, R) = C + \log R^{-1}.$$

Finally, we will prove Theorem B, but we first establish an auxiliary result. To this end we will need the following extension of the classical Bézout theorem. Suppose we have a system of n complex polynomial equations $f_i(z_1, ..., z_n) = 0$, $i = 1, .., n$. Let $\mathcal{Z}_1, \dots, \mathcal{Z}_s$ be the irreducible components of the variety defined by the system. Then

$$\deg(\mathcal{Z}_1) + \cdots + \deg(\mathcal{Z}_s) \leq \deg f_1 \times \cdots \times \deg f_n. \tag{21}$$

The authors are grateful to János Kollár and Mihnea Popa for pointing out this version of the Bézout bound (for a more general result, see [13, Thm. 12.3]).

Lemma 7.3 *Let $f(z,w)$, $g(z,w)$ be non-constant polynomials with no common factors. Let $\zeta(w)$ be an analytic function on $\mathcal{D}(0, r_0)$ such that*

$$\{(\zeta(w), w) : w \in \mathcal{D}(0, r_0)\} \subset \operatorname{reg}\{f(z,w) = 0\}. \tag{22}$$

Then there exists at most one straight line $\mathcal{L} \subset \mathbb{C}$ through the origin such that

$$\#\{\xi \in (-r_0, r_0) : g(\zeta(\xi), \xi) \in \mathcal{L}\} > (\deg f)^2 \deg g. \tag{23}$$

Proof Let $\mathcal{L}$ be a line through the origin. We first argue that if (23) holds, then we must have $\{g(\zeta(\xi), \xi) : \xi \in (-r_0, r_0)\} \subset \mathcal{L}$. Write

$$f(z,\xi) := P(x+iy,\xi) + iQ(x+iy,\xi) = \hat{P}(x,y,\xi) + i\hat{Q}(x,y,\xi)$$

where P, Q are the real and imaginary parts of f, and $\hat{P}$, $\hat{Q}$ are real polynomials of three real variables x, y, ξ. Clearly $\deg\hat{P} = \deg\hat{Q} = \deg f$. Similarly write

$$g(z,\xi) := U(x+iy,\xi) + iV(x+iy,\xi) = \hat{U}(x,y,\xi) + i\hat{V}(x,y,\xi).$$

Without loss of generality, we may assume that the line $\mathcal{L}$ is horizontal. If (23) holds, then the system

$$\hat{P} = 0, \quad \hat{Q} = 0, \quad \hat{V} = 0 \tag{24}$$

has more than $\deg\hat{P} \times \deg\hat{Q} \times \deg\hat{V}$ solutions $\underline{v}_j = (x_j, y_j, \xi_j)$ with

$$\xi_j \in (-r_0, r_0), \quad x_j + iy_j = \zeta(\xi_j), \quad \xi_{j_1} \neq \xi_{j_2}.$$

Complexify the variables x, y, ξ, and let $\mathcal{Z}_1, \dots, \mathcal{Z}_s$ be the irreducible components of the complex variety defined by the system (24). By the Bézout bound (21), there exists a component $\mathcal{Z}_k$ that contains at least two of the solutions $\underline{v}_j$ and therefore has dimension at least one. Let $\underline{v}_0$ be one of the solutions contained in $\mathcal{Z}_k$. We will argue that there exists an analytic mapping

$$t \to \underline{v}(t) = (x(t), y(t), \xi(t)) \in \mathcal{Z}_k,\ t \in \mathcal{D}(0,\delta) \tag{25}$$

such that $\underline{v}(0) = \underline{v}_0$ and $\xi(t)$ is non-constant. By [21] we know that there exists a neighborhood N of $\underline{v}_0$ in $\mathcal{Z}_k$, such that for any $\underline{v} \in N \setminus \{\underline{v}_0\}$, there exists a one-dimensional irreducible variety $\mathcal{V}$ through both $\underline{v}$ and $\underline{v}_0$. Since $\mathcal{V}$ can be parametrized by a Riemann surface (see [5, Prop. 6.2]), we get the existence of a mapping of the form (25). If $\xi(t)$ is constant, then by the uniqueness theorem (see [5, Cor. 5.3.2]), we must have $\mathcal{V} \subset \{\xi = \xi_0\}$. If this happens for all such mappings obtained by choosing different $\underline{v} \in N \setminus \{\underline{v}_0\}$, then $N \subset \{\xi = \xi_0\}$, and by the uniqueness theorem, $\mathcal{Z}_k \subset \{\xi = \xi_0\}$. This would contradict the fact that $\mathcal{Z}_k$ contains two of the solutions $\underline{v}_j$ (recall that $\xi_{j_1} \neq \xi_{j_2}$). So we proved the existence of the mapping (25) with the desired properties. We have

$$f(x(t)+iy(t),\xi(t)) = 0, \quad V(x(t)+iy(t),\xi(t)) = 0,\ t \in \mathcal{D}(0,\delta).$$

By the assumption (22), we get

$$x(t) + iy(t) = \zeta(\xi(t)),$$

provided we choose δ small enough. Therefore

$$V(\zeta(\xi(t)),\xi(t)) = 0,\ t \in \mathcal{D}(0,\delta)$$

and since $\xi(t)$ is non-constant, there exists $\varepsilon > 0$ so that

$$V(\zeta(\xi), \xi) = 0,\ \xi \in (\xi_0 - \varepsilon, \xi_0 + \varepsilon).$$

So $V(\zeta(\xi), \xi) = 0$ for all $\xi \in (-r_0, r_0)$, that is, $\{g(\zeta(\xi), \xi) : \xi \in (-r_0, r_0)\} \subset \mathcal{L}$.

Now we can finish the proof by arguing by contradiction. If the conclusion doesn't hold, it follows that we have $\{g(\zeta(\xi), \xi) : \xi \in (-r_0, r_0)\} \subset \mathcal{L}_1 \cap \mathcal{L}_2 = \{0\}$ and therefore the system $f = g = 0$ has infinitely many solutions. By the classical Bézout theorem, this would contradict the assumption that f and g don't have common factors. □

Proof of Theorem B Let $\mathcal{Z}_1, \dots, \mathcal{Z}_s$ be the irreducible components of $\mathcal{Z}$. Each of them is the zero set of an irreducible factor of f_2. Let $f_{2,1}, \dots, f_{2,s}$ be such irreducible factors. Fix $k \in \{1, \dots, s\}$ and $(z_0, w_0) \in \operatorname{reg} \mathcal{Z} \cap \mathcal{Z}_k$. We can make a change of variables (as in the proof of part (2) of Theorem A) such that (z_0, w_0) is mapped to the origin, and we can find an analytic function $\zeta : \mathcal{D}(0, \varepsilon_0) \to \mathcal{D}(0, \delta_0)$ so that

$$\phi(w) = (\zeta(w), w), w \in \mathcal{D}(0, \varepsilon_0)$$

is a chart for $\operatorname{reg} \mathcal{Z} \cap \mathcal{Z}_k$ around the origin.

If $f_{2,k}$ divides f_1, then f_1 vanishes identically on $\mathcal{Z}_k$, and its Bernstein exponent is 0 by convention (in any chart). So, we just need to treat the case when $f_{2,k}$ and f_1 have no common factors. Let $\psi(w) = f_1(\phi(w))$. We claim that

$$B_\psi(1/4; w_0, R) \le C(f_2)\deg f_1 \tag{26}$$

provided $\mathcal{D}(w_0, R) \subset \mathcal{D}(0, \varepsilon_0/8)$. We will check this claim by using the previous lemma and Proposition 3.2. Let $a_1, \dots, a_n$ be the zeros of ψ. Since f_1 and $f_{2,k}$ are co-prime, using the classical Bézout theorem, we have

$$n \le \deg f_{2,k} \times \deg f_1.$$

Factorize

$$\psi(w) = h(w)P(w), \quad P(w) = \prod_{k=1}^{n}(w - a_k).$$

From the proof of Proposition 3.2 (see (7)), we have

$$B_\psi(1/4; w_0, R) \le CR(M_h(0, 3\varepsilon_0/4) - M_h(0, \varepsilon_0/4)) + B_P(1/4; w_0, R).$$

Recall that $B_P(1/4; w_0, R) \le n \log 4 \le C(f_2)\deg f_1$. So, to check the claim (26), we just need to estimate $M_h(0, 3\varepsilon_0/4) - M_h(0, \varepsilon_0/4)$. Without loss of generality, we can assume $h(0) = 1$, and therefore $M_h(0, \varepsilon_0/4) \geq 0$. Since h does not vanish, we have

$$h(w) = e^{u(w)+iv(w)},$$

where $u + iv$ is analytic and u, v are real valued. Then

$$M := M_h(0, 3\varepsilon_0/4) = \sup_{w \in \mathcal{D}(0,3\varepsilon_0/4)} |u(w)|.$$

Due to the Borel-Carathéodory estimate (see [19, Thm. 11.1.1]),

$$N := \sup_{w \in \mathcal{D}(0,7\varepsilon_0/8)} |v(w)| \gtrsim M.$$

Choose $|\hat{w}| = 7\varepsilon_0/8$ such that $|v(\hat{w})| \geq N/2$ and at the same time no root a_k falls on the straight line through $\hat{w}$ and the origin. This allows us to define the continuous functions $\theta_k(\xi) := \arg(\xi\hat{w} - a_k) \in [0, 2\pi]$, $\xi \in (-\infty, +\infty)$. Set

$$\theta(\xi) = \sum_{1 \leq k \leq n} \theta_k(\xi).$$

Take $\theta \in (0, 2\pi)$ arbitrary. We have

$$\operatorname{Im} e^{-i\theta}\psi(\xi\hat{w}) = e^{u(\xi\hat{w})}|P(w)| \sin(v(\xi\hat{w}) + \theta(\xi) - \theta).$$

It is clear form this formula that if $N \gg n$, then for any θ,

$$\#\{\xi \in (-7\varepsilon_0/8, 7\varepsilon_0/8) : f_1(\zeta(\xi\hat{w}), \xi\hat{w}) \in \mathcal{L}_\theta\} \geq N/4,$$

where $\mathcal{L}_\theta$ is the line of angle θ through the origin. This and Lemma 7.3 imply that we must have

$$N \lesssim (\deg f_{2,k})^2 \deg f_1.$$

Putting the above together, we have

$$M_h(0, 3\varepsilon_0/4) - M_h(0, \varepsilon_0/4) \lesssim C(f_2) \deg f_1,$$

which completes the proof of claim (26).

Finally, it is clear that the conclusion holds by choosing the charts to be rescaled versions of $\phi|_{\mathcal{D}(0,\varepsilon_0/8)}$, for each $(z_0, w_0) \in \operatorname{reg} \mathcal{Z}$. □

References

1. Biernacki, M.: Sur les fonctions multivalentes de ordre p. CR. Acad. Sci. (Paris) **203**, 449–451 (1936)
2. Bourgain, J.: Green's Function Estimates for Lattice Schrödinger Operators and Applications, volume 158 of Annals of Mathematics Studies (Princeton University Press, Princeton, 2005)
3. Brudnyi, A.: On local behavior of holomorphic functions along complex submanifolds of $\mathbb{C}^N$. Invent. Math. **173**(2), 315–363 (2008)
4. Cartwright, M.L.: Some inequalities in the theory of functions. Math. Ann. **111**, 98–118 (1935)
5. Chirka, E.M.: Complex Analytic Sets, volume 46 of Mathematics and Its Applications (Soviet Series) (Kluwer Academic Publishers Group, Dordrecht, 1989)
6. Chulaevsky, V.A., Sinaĭ, Ya. G.: Anderson localization for the 1-D discrete Schrödinger operator with two-frequency potential. Commun. Math. Phys. **125**(1), 91–112 (1989)
7. Coman, D., Poletsky, E.A.: Transcendence measures and algebraic growth of entire functions. Invent. Math. **170**(1), 103–145 (2007)
8. Comte, G., Yomdin, Y.: Zeroes and rational points of analytic functions. Preprint arXiv:1608.02455
9. Fefferman, C., Narasimhan, R.: Bernstein's inequality on algebraic curves. Ann. Inst. Fourier (Grenoble) **43**(5), 1319–1348 (1993)
10. Fefferman, C., Narasimhan, R.: On the polynomial-like behaviour of certain algebraic functions. Ann. Inst. Fourier (Grenoble) **44**(4), 1091–1179 (1994)
11. Fefferman, C., Narasimhan, R.: A local Bernstein inequality on real algebraic varieties. Math. Z. **223**(4), 673–692 (1996)
12. Friedland, O., Yomdin, Y.: Doubling coverings of algebraic hypersurfaces. Pure Appl. Funct. Anal. **2**(2), 221–241 (2017)
13. Fulton, W.: Intersection Theory, volume 2 of Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics], 2nd edn. (Springer, Berlin, 1998)
14. Goldstein, M., Schlag, W.: Fine properties of the integrated density of states and a quantitative separation property of the Dirichlet eigenvalues. Geom. Funct. Anal. **18**(3), 755–869 (2008)
15. Goldstein, M., Schlag, W., Voda, M.: On localization and spectrum of multi-frequency quasi-periodic operators (2016). Preprint
16. Goldstein, M., Schlag, W., Voda, M.: On the spectrum of multi-frequency quasi-periodic Schrödinger operators at large coupling. Invent. Math. **217**(2), 603–701 (2019)
17. Hayman, W.K.: Multivalent Functions, volume 110 of Cambridge Tracts in Mathematics, 2nd edn. (Cambridge University Press, Cambridge, 1994)
18. Lang, S.: Algebra, volume 211 of Graduate Texts in Mathematics, 3rd edn. (Springer, New York, 2002)
19. Levin, B. Ya.: Lectures on Entire Functions, volume 150 of Translations of Mathematical Monographs (American Mathematical Society, Providence, 1996)
20. Roytwarf, N., Yomdin, Y.: Bernstein classes. Ann. Inst. Fourier (Grenoble) **47**(3), 825–858 (1997)
21. Shiffman, B.: Local complex analytic curves in an analytic variety. Proc. Am. Math. Soc. **24**, 432–437 (1970)

A Weighted Prékopa–Leindler Inequality and Sumsets with Quasicubes

Ben Green, Dávid Matolcsi, Imre Ruzsa, George Shakan, and Dmitrii Zhelezov

Dedicated to the memory of Jean Bourgain

Abstract We give a short, self-contained proof of two key results from a paper of four of the authors. The first is a kind of weighted discrete Prékopa–Leindler inequality. This is then applied to show that if $A, B \subseteq \mathbb{Z}^d$ are finite sets and U is a subset of a "quasicube", then $|A + B + U| \geqslant |A|^{1/2}|B|^{1/2}|U|$. This result is a key ingredient in forthcoming work of the fifth author and Pälvölgyi on the sum-product phenomenon.

2000 Mathematics Subject Classification Primary 11B30

BG and GS are supported by BG's Simons Investigator Grant number 376201.

Dávid Matolcsi is supported by the ÚNKP-21-1 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund.

B. Green (✉)
University of Oxford, Oxford, UK
e-mail: ben.green@maths.ox.ac.uk

D. Matolcsi
Eötvös Loránd University, Budapest, Hungary

I. Ruzsa
Rényi Institute of Mathematics, Budapest, Hungary

G. Shakan
Centre de Recherches Mathématiques, Montréal, QC, Canada

D. Zhelezov
Radon Institute for Computational and Applied Mathematics RICAM, Linz, Austria

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_6

1 Introduction

Quasicubes. The notion of a quasicube $\Sigma \subseteq \mathbb{Z}^d$ is defined inductively. When $d = 1$, a quasicube is simply a set of size two. For larger d, Σ is a quasicube if

1. $\pi(\Sigma) = \{x_0, x_1\}$ is a set of size two, where $\pi : \mathbb{Z}^d \to \mathbb{Z}$ is the coordinate projection onto the final coordinate, and
2. The fibre $\Sigma_i := \Sigma \cap \pi^{-1}(x_i)$ (considered as a subset of $\mathbb{Z}^{d-1}$) is a quasicube.

Thus, for instance, the usual cube $\{0, 1\}^d$ is a quasicube. Another example of a quasicube with $d = 2$ is the set $\Sigma = \{(0, 0), (1, 0), (0, 1), (1, 2)\}$.

The following result is established in [3].

Theorem 1.1 *Let $A, B \subseteq \mathbb{Z}^d$ be finite sets and suppose that $U \subseteq \mathbb{Z}^d$ is contained in a quasicube. Then $|A + B + U| \geqslant |A|^{1/2}|B|^{1/2}|U|$.*

Our aim in this note is to give a short, self-contained proof of this result.

2 A Weighted Discrete Prékopa–Leindler Inequality

As in [3], we deduce Theorem 1.1 from a weighted discrete Prékopa–Leindler inequality. Let $a, b : \mathbb{Z} \to [0, \infty)$ be compactly supported functions. We define the *max-convolution*

$$a \overline{*} b(n) := \sup_{m \in \mathbb{Z}} a(n - m)b(m),$$

and we write

$$\|a\|_2 := \big(\sum_n a(n)^2\big)^{1/2}, \quad \|b\|_2 := \big(\sum_n b(n)^2\big)^{1/2}.$$

The following result is equivalent to [3, Theorem 11.1].

Proposition 2.1 *Let $a, b : \mathbb{Z} \to [0, \infty)$ be compactly supported functions and let $p \in [0, 1]$. Then we have*

$$\sum_n \max(pa\overline{*}b(n), (1 - p)a\overline{*}b(n - 1)) \geqslant \|a\|_2\|b\|_2.$$

In the case $p = \frac{1}{2}$, this is (2.4) in the paper of Prékopa [4], where it is used to establish the one-dimensional case of what is now known as the Prékopa–Leindler inequality (we will recall the statement of this below). We will proceed in the opposite direction, deducing Proposition 2.1 from Prékopa–Leindler.

Suppose that $f, g : \mathbb{R} \to [0, \infty)$ are compactly supported, piecewise continuous functions. Then the (one-dimensional) Prékopa–Leindler inequality states that

$$\int f \overline{*} g \geqslant 2\|f\|_2 \|g\|_2, \tag{1}$$

where the max-convolution is defined by

$$f \overline{*} g(x) := \sup_{y \in \mathbb{R}} f(x-y)g(y),$$

and the norms are the usual Lebesgue norms

$$\|f\|_2 := \left(\int f^2\right)^{1/2}, \quad \|g\|_2 := \left(\int g^2\right)^{1/2}.$$

(It should always be clear from context whether we are applying $\overline{*}$ or $\|\cdot\|_2$ with functions on $\mathbb{Z}$ or functions on $\mathbb{R}$.) We note that Brascamp and Lieb [1] found a much shorter proof of (1) than the original (see also this survey of Gardner [2]).

Proof of Proposition 2.1 By continuity we may assume that $p \in (0, 1)$. Set $\lambda := \log(\frac{1}{p} - 1)$. Apply (1) with functions f, g defined by

$$f(x) := e^{\lambda\{x\}} a(\lfloor x \rfloor), \quad g(y) := e^{\lambda\{y\}} b(\lfloor y \rfloor).$$

Let $n \in \mathbb{Z}$ and $0 \leqslant t < 1$. Suppose that $x + y = n + t$. Then, since $x - 1 < \lfloor x \rfloor \leqslant x$, we have $n - 2 < \lfloor x \rfloor + \lfloor y \rfloor < n + 1$, or in other words $\lfloor x \rfloor + \lfloor y \rfloor = n - 1$ or n. If $\lfloor x \rfloor + \lfloor y \rfloor = n - 1$, then

$$f(x)g(y) \leqslant e^{\lambda(t+1)} a \overline{*} b(n-1),$$

while if $\lfloor x \rfloor + \lfloor y \rfloor = n$, then

$$f(x)g(y) \leqslant e^{\lambda t} a \overline{*} b(n).$$

Therefore

$$f \overline{*} g(n+t) \leqslant e^{\lambda t} \max(a \overline{*} b(n), e^{\lambda} a \overline{*} b(n-1)).$$

Integrating over $t \in [0, 1)$ and then summing over $n \in \mathbb{Z}$ yield

$$\int f \overline{*} g \leqslant \frac{e^{\lambda} - 1}{\lambda} \sum_n \max(a \overline{*} b(n), e^{\lambda} a \overline{*} b(n-1)). \tag{2}$$

On the other hand,

$$\|f\|_2^2 = \frac{e^{2\lambda}-1}{2\lambda}\|a\|_2^2, \quad \|g\|_2^2 = \frac{e^{2\lambda}-1}{2\lambda}\|b\|_2^2.$$

Substituting into (1) gives

$$\sum_n \max(a\overline{*}b(n), e^{\lambda}a\overline{*}b(n-1)) \geqslant (e^{\lambda}+1)\|a\|_2\|b\|_2.$$

Recalling the choice of λ (thus $p = \frac{1}{e^{\lambda}+1}$), the proposition follows. □

3 Proof of the Main Theorem

The arguments of this section are all in [3], but there they form part of a more general framework. Here we provide a self-contained account tailored to the specific purpose of proving Theorem 1.1.

Proof of Theorem 1.1 We proceed by induction on d. The proof of the inductive step also proves the base case $d = 1$.

Suppose that U is contained in a quasicube $\Sigma \subset \mathbb{Z}^d$. Suppose that $\pi(\Sigma) = \{x_0, x_1\}$, where $\pi : \mathbb{Z}^d \to \mathbb{Z}$ is projection onto the last coordinate. Since the inequality is translation-invariant, we may assume that $x_0 = 0$ and $x_1 = q > 0$. Suppose first that $q = 1$.

Let $A_i := A \cap \pi^{-1}(n)$ be the fibre of A above n, and similarly for B. The set U has just two fibres U_0, U_1, and, by the definition of quasicubes, they are both contained in quasicubes of dimension $d - 1$.

Observe that the fibre of $A + B + U$ above n contains $A_x + B_y + U_0$ whenever $x + y = n$ and $A_x + B_y + U_1$ whenever $x + y = n - 1$. By induction,

$$|A_x + B_y + U_0| \geqslant |A_x|^{1/2}|B_y|^{1/2}|U_0|,$$

$$|A_x + B_y + U_1| \geqslant |A_x|^{1/2}|B_y|^{1/2}|U_1|,$$

and so the fibre $(A + B + U)_n$ of $A + B + U$ above n has size at least

$$\max\big(|U_0| \max_{x+y=n} |A_x|^{1/2}|B_y|^{1/2}, |U_1| \max_{x+y=n-1} |A_x|^{1/2}|B_y|^{1/2}\big).$$

This is equal to

$$|U| \max\big(pa\overline{*}b(n) + (1-p)a\overline{*}b(n-1)\big),$$

where $p := |U_0|/|U|$, $a(x) := |A_x|^{1/2}$ and $b(y) := |B_y|^{1/2}$. Summing over n and applying Proposition 2.1, we obtain

$$\begin{aligned} |A+B+U| &= \sum_n |(A+B+U)_n| \\ &\geqslant |U| \sum_n \max\big(pa\overline{*}b(n) + (1-p)a\overline{*}b(n-1)\big) \\ &\geqslant |U|\|a\|_2\|b\|_2 = |U||A|^{1/2}|B|^{1/2}. \end{aligned}$$

This proves the result when $q = 1$. Suppose now that q is arbitrary, and foliate $A = \bigcup_{r\in\mathbb{Z}/q\mathbb{Z}} A_r$, $B = \bigcup_{s\in\mathbb{Z}/q\mathbb{Z}} B_s$, where $A_r := \{a \in A : \pi(a) \equiv r (\mathrm{mod}\ q)\}$ and similarly for B_s. Let r_* be such that $|A_r| \leqslant |A_{r_*}|$ for all r, and s_* be such that $|B_s| \leqslant |B_{s_*}|$ for all s.

The sets $A_{r_*} + B_s + U$ are disjoint as s varies, and so by the case $q = 1$ (rescaled), we have

$$|A+B+U| \geqslant \sum_s |A_{r_*} + B_s + U| \geqslant |U||A_{r_*}|^{1/2} \sum_s |B_s|^{1/2}. \tag{3}$$

Similarly,

$$|A+B+U| \geqslant |U||B_{s_*}|^{1/2} \sum_r |A_r|^{1/2}. \tag{4}$$

Taking products of (3), (4) and using

$$|A_{r_*}|^{1/2} \sum_r |A_r|^{1/2} \geqslant \sum_r |A_r| = |A|,$$

$$|B_{s_*}|^{1/2} \sum_s |B_s|^{1/2} \geqslant \sum_s |B_s| = |B|,$$

the result follows. □

References

1. Brascamp, H.J., Leib, E.H.: On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. J. Funct. Anal. **22**(4), 366–389 (1976)
2. Gardner, R.: The Brunn-Minkowski inequality. Bull. Am. Math. Soc. **39**(3), 355–405 (2002)
3. Matolcsi, D., Ruzsa, I.Z., Shakan, G., Zhelezov, D.: An analytic approach to cardinalities of sumsets. Preprint
4. Prékopa, A.: Logarithmic concave measures with application to stochastic programming. Acta Sci. Math. (Szeged) **32**, 301–316 (1971)

Equidistribution of Affine Random Walks on Some Nilmanifolds

Weikun He, Tsviqa Lakrec, and Elon Lindenstrauss

Dedicated to the memory of Jean Bourgain

Abstract We study quantitative equidistribution in law of affine random walks on nilmanifolds, motivated by a result of Bourgain, Furman, Mozes, and the third named author on the torus. Under certain assumptions, we show that a failure to having fast equidistribution is due to a failure on a factor nilmanifold. Combined with equidistribution results on the torus, this leads to an equidistribution statement on some nilmanifolds such as Heisenberg nilmanifolds. In an appendix we strengthen results of de Saxce and the first named author regarding random walks on the torus by eliminating an assumption on Zariski connectedness of the acting group.

1 Introduction

In this paper we consider random walks on compact nilmanifolds by automorphisms of the nilmanifolds as well as by affine maps. Recall that a nilmanifold is a space of the form $X = N/\Lambda$, where N is a connected simply connected nilpotent Lie group and $\Lambda < N$ is a lattice (which in a nilpotent Lie group is necessarily cocompact; cf. [18]). An *automorphism* of X is defined to be the homeomorphism of X induced by a Lie group automorphism of N that preserves Λ; we denote the group of all

The authors are supported by ERC 2020 grant HomDyn (grant no. 833423). W.H. is also supported by KIAS Individual Grant (no. MG080401).

W. He
Korea Institute for Advanced Study, Seoul, Republic of Korea
e-mail: heweikun@kias.re.kr

T. Lakrec · E. Lindenstrauss (✉)
Einstein Institute of Mathematics, The Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: elon@math.huji.ac.il

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_7

such automorphisms by $\mathrm{Aut}(X)$. An *affine transformation* on X is the composition of an automorphism of X by left translation by an element of N; the group of affine transformations of X, denoted by $\mathrm{Aff}(X)$, is the semidirect product $\mathrm{Aut}(X) \ltimes N$. The projection $\mathrm{Aff}(X) \to \mathrm{Aut}(X)$ will be denoted by θ.

Given a Borel probability measure μ on $\mathrm{Aut}(X)$ (or more generally $\mathrm{Aff}(X)$) and a starting point $x \in X$, we can define a random walk by successively applying to x a sequence of elements $g_1, g_2, \ldots$ each g_i chosen i.i.d according to μ. Thus the distribution of the random walk after n steps, i.e., of the random element $g_n \ldots g_1 x$ in X, is given by $\mu^{*n} * \delta_x$.

In this situation, Bekka and Guivarc'h give a sufficient and necessary condition for the random walk defined by μ to have a spectral gap on $L^2(X)$:

Theorem A (Bekka-Guivarc'h [2, Theorem 1]) *Let $X = N/\Lambda$ be a nilmanifold and let H be a countable subgroup of* $\mathrm{Aff}(X)$. *The following are equivalent*

(i) The action of H on $X = N/\Lambda$ has a spectral gap.
(ii) The action of H on $T = N/[N, N]\Lambda$ has a spectral gap.
(iii) There is no non-trivial H-invariant factor torus T' of T such that the projection of $H \subset$ $\mathrm{Aff}(X)$ to $\mathrm{Aut}(T')$ is virtually abelian.

Recall for a torus $T = V/\Delta$ where V is an Euclidean space and Δ is a lattice in V; a factor torus is some $T' = T/S$ where S is a subtorus of T (i.e., $S = W/(W \cap \Delta)$ for W a rational linear subspace of V relative to the rational structure defined by Δ).[1] If H is some subgroup of $\mathrm{Aff}(T)$, then the factor torus T' is said to be H-invariant if W is invariant under $\theta(H)$, equivalently if the action of H on T induces an action of H on T' so that the projection map from T to T' is H-equivariant. For a nilmanifold $X = N/\Lambda$, the quotient $N/[N, N]\Lambda$ is a torus, called the *maximal torus factor* of X.

When these equivalent conditions in Theorem A are satisfied, for all but a set of x of exponentially small measure, the random walk $g_n \ldots g_1 x$ emanating from x is exponentially close to being equidistributed. The purpose of this paper is to understand the random walk $g_n \ldots g_1 x$ starting from *any* $x \in X$.

Under certain assumptions on μ, that are substantially stronger than those in Theorem A, we show that either the random walk equidistributes, namely, $\mu^{*n} * \delta_x$ converges to the Haar measure m_X on X in the weak-$*$ topology or the random walk is trapped in a proper closed set invariant under the group generated by $\mathrm{Supp}(\mu)$. Furthermore, the equidistribution result is quantitative. Informally, we show that if the equidistribution is not fast, it is because the random walk is close to a "small" orbit.

[1] Note that finite index quotients are not considered factor tori under this definition.

1.1 Quantitative Equidistribution

Similar to the spirit of Theorem A, we aim to prove that if a random walk on a nilmanifold does not equidistribute, it is because the projected random walk on a factor torus does not equidistribute. This leads to the following definitions.

Consider a nilmanifold $X = N/\Lambda$. We fix a Riemannian distance on X. For $\alpha \in (0, 1)$, let $\mathcal{C}^{0,\alpha}(X)$ denote the space of α-Hölder continuous functions on X, equipped with the norm

$$\|f\|_{0,\alpha} = \|f\|_\infty + \sup_{x \neq y \in X} \frac{|f(x) - f(y)|}{d(x, y)^\alpha}.$$

For ν and η Borel measures on X, recall that the *α-Wasserstein distance* between them is defined by

$$\mathcal{W}_\alpha(\nu, \eta) = \sup_{f \in \mathcal{C}^{0,\alpha}(X) : \|f\|_{0,\alpha} \leq 1} \left| \int_X f \, \mathrm{d}\nu - \int_X f \, \mathrm{d}\eta \right|.$$

Let $T = V/\Delta$ be a torus of dimension d. We choose an identification $\mathbb{Z}^d$ with its group of unitary characters via some isomorphism $a \mapsto \chi_a$. Each closed subgroup L of T is uniquely determined by its dual

$$L^* = \{ a \in \mathbb{Z}^d \mid L \subset \ker \chi_a \}.$$

Definition 1 A closed subgroup L of a torus T is said to have *height* $\leq h$ if its dual $L^* \subset \mathbb{Z}^d$ can be generated by integer vectors of norm $\leq h$.

Remark This notion depends on the choice of the isomorphism from $\mathbb{Z}^d$ to the group of unitary characters. For any torus we encounter in this paper, we assume such choice is implicitly fixed in advance.

For the next two definitions, we will denote by $T = N/[N, N]\Lambda$ the maximal torus factor of X and by $\pi : X \to T$ the canonical projection.

Definition 2 Let $\lambda > 0$, $C > 1$, and $\alpha \in (0, 1]$ be parameters. Let μ be a Borel probability measure on a $\mathrm{Aut}(X)$, and let $\Gamma = \langle \mathrm{Supp}(\mu) \rangle$. We say that the μ-induced random walk on X satisfies (C, λ, α)-*quantitative equidistribution* if the following holds for any integer $m \geq 1$ and any $t \in (0, \frac{1}{2})$. Assume

$$m \geq C \log \frac{1}{t} \quad \text{and} \quad \mathcal{W}_\alpha(\mu^{*m} * \delta_x, \mathrm{m}_X) > t.$$

Then there exists a point $x' \in X$ such that

(i) $d(x, x') \leq e^{-\lambda m}$
(ii) $\pi(\Gamma x')$ lies in a proper closed Γ-invariant subgroup of T of height $\leq t^{-C}$

In the case where Γ acts irreducibly on T (i.e., Γ acts irreducibly on $N/[N, N]$ over $\mathbb{Q}$), the condition (ii) can be replaced by

(ii') $\pi(\Gamma x')$ consists of rational points of denominator $\leq t^{-C}$

In the situation of an affine random walk, the definition needs to be adjusted. We fix a left-invariant Riemannian distance on the Lie group $\mathrm{Aff}(X)$.

Definition 3 Let $\lambda > 0$, $C > 1$, and $\alpha \in (0, 1]$ be parameters. Let μ be a finitely supported Borel probability measure on $\mathrm{Aff}(X)$. We say that the μ-induced random walk on X satisfies (C, λ, α)-*quantitative equidistribution* if the following holds for any integer $m \geq 1$ and any $t \in (0, \frac{1}{2})$. Assume

$$m \geq C \log \frac{1}{t} \quad \text{and} \quad \mathcal{W}_\alpha(\mu^{*m} * \delta_x, \mathrm{m}_X) > t.$$

Then there exist a point $x' \in X$ and a closed subgroup $H' \subset \mathrm{Aff}(X)$ such that

(i) $d(x, x') \leq e^{-\lambda m}$
(ii) $d(g, H') \leq e^{-\lambda m}$ for every $g \in \mathrm{Supp}(\mu)$
(iii) $\pi(H'x') - \pi(x')$ lies in a proper closed $\theta(H')$-invariant subgroup of T of height $\leq t^{-C}$

Here, we are thinking of H' as generated by $e^{-\lambda m}$-perturbations of elements of $\mathrm{Supp}(\mu)$. Since $\mathrm{Aut}(X)$ is discrete, the perturbation only happens on the translation part. In particular, we will have $\theta(H') = \theta(H)$ where $H = \langle \mathrm{Supp}(\mu) \rangle$. Again, if $\theta(H)$ acts irreducibly on T, then the condition (iii) can be replaced by

(iii') $\pi(H'x') - \pi(x')$ consists of rational points of denominator $\leq t^{-C}$

In both the linear and affine case, we say that the μ-induced random walk on X satisfies (λ, α)-*quantitative equidistribution* if it satisfies (C, λ, α)-quantitative equidistribution for some constant C.

Remark For $0 < \alpha < \alpha' \leq 1$, we have $\mathcal{W}_{\alpha'}(\nu, \eta) \leq \mathcal{W}_\alpha(\nu, \eta)$ for any measures ν and η. Hence, (C, λ, α)-quantitative equidistribution implies (C, λ, α')-quantitative equidistribution for any $\alpha' \in (\alpha, 1]$.

1.2 Statement of the Main Result

Let μ be a Borel probability measure on a Lie group H. If H acts on an Euclidean space Z via $\theta_Z \colon H \to \mathrm{GL}(Z)$, we define the *essential exponential growth rate* of the action on Z to be the quantity

$$\tau_Z(\mu) = \inf_{\kappa > 0} \limsup_{m \to +\infty} \frac{1}{m} \min\big\{ \log \# A \mid A \subset \mathrm{Aut}(Z) \text{ and } (\theta_Z)_* \mu^{*m}(A) \geq 1 - e^{-\kappa m} \big\}.$$

Clearly, if $(\theta_Z)_*\mu$ is finitely supported,

$$\tau_Z(\mu) \le \lim_{m\to+\infty} \frac{1}{m} \log\left(\#\,\mathrm{Supp}\left((\theta_Z)_*\mu^{*m}\right)\right).$$

Let $H \curvearrowright (X, \mathrm{m}_X)$ be a probability measure preserving action of H on a compact space X. Let $(U_X, L^2(X, \mathrm{m}_X))$ denote the corresponding Koopman representation. Let $\pi : (X, \mathrm{m}_X) \to (Y, \mathrm{m}_Y)$ be a factor, i.e., $\pi_*\,\mathrm{m}_X = \mathrm{m}_Y$, H acts on (Y, m_Y), and π is H-equivariant. By composing with π, we can embed $L^2(Y, \mathrm{m}_Y)$ in $L^2(X, \mathrm{m}_X)$ as an H-invariant subspace. Let $U_{X,Y}$ be the restriction of U_X to the orthogonal complement of $L^2(Y, \mathrm{m}_Y)$ in $L^2(X, \mathrm{m}_X)$. For a Borel probability measure μ on H, define

$$\sigma_{X,Y}(\mu) = -\lim_{m\to+\infty} \frac{1}{m} \log\|U_{X,Y}(\mu)^m\|.$$

We say a measure μ on $\mathrm{Aff}(X)$ has a *finite exponential moment* if there exists $\beta > 0$ such that

$$\int_{\mathrm{Aff}(X)} \mathrm{Lip}_X(g)^\beta \,\mathrm{d}\mu(g) < +\infty, \tag{1}$$

where for $g \in \mathrm{Aff}(X)$,

$$\mathrm{Lip}_X(g) = \sup_{x,x'\in X,\, x\neq x'} \frac{d(gx, gx')}{d(x, x')}.$$

To keep track of the parameter β, we say more precisely that μ has a finite β-exponential moment.

Theorem 1.1 *Let μ be a probability measure on* $\mathrm{Aff}(X)$ *having a finite β-exponential moment for some $\beta > 0$. Let Γ denote the subgroup generated by the support of $\theta_*\mu$. Assume that there exists a rational Γ-invariant connected central subgroup $Z \subset N$ such that*

$$\tau_Z(\mu) < 2\sigma_{X,Y}(\mu) \tag{2}$$

where $Y = N/(\Lambda Z)$ is the corresponding factor nilmanifold.

If the μ-induced random walk on Y satisfies (λ, α)-quantitative equidistribution for some $\lambda > 0$ and $0 < \alpha \le \min\{1, \beta\}$, then the μ-induced random walk on X satisfies (λ', α)-quantitative equidistribution for any $\lambda' \in (0, \lambda)$.

Note that when the equivalent conditions in Theorem A hold for X and $H = \langle \mathrm{Supp}(\mu)\rangle$, we have $\sigma_{X,Y}(\mu) > 0$. In some special situations, for instance, if $\theta_Z(H)$ is a virtually nilpotent group, we have easily $\tau_Z(\mu) = 0$. Thus, in these situations, Theorem 1.1 applies and reduces the problem of quantitative equidistribution on X

to whether there is one on Y, a nilmanifold of smaller dimension. The idea is that this will eventually reduce to the case of random walks on a torus, where much more is known.

We believe that a result akin to Theorem 1.1 should hold more generally with the assumption (2) relaxed to $\sigma_{X,Y}(\mu) > 0$, with an appropriate (relative) irreducibility assumption, e.g.,

Conjecture 1.2 *Let μ be a probability measure on* $\mathrm{Aff}(X)$ *having a finite β-exponential moment for some $\beta > 0$. Let H denote the subgroup generated by* $\mathrm{Supp}(\mu)$ *and $\Gamma = \theta(H)$. Assume that there exists a rational Γ-invariant connected central subgroup $Z \subset N$ with corresponding factor nilmanifold $Y = N/(\Lambda Z)$ so that*

(i) $\sigma_{X,Y}(\mu) > 0$.
(ii) For any finite index subgroup $H' < H$, and any proper H'-invariant affine subnilmanifold $X' \subset X$, the projection of X' to Y is a proper affine subnilmanifold of Y.
(iii) The μ-induced random walk on Y satisfies (λ, α)-quantitative equidistribution for some $\lambda > 0$ and $0 < \alpha \leq \min\{1, \beta\}$.

Then the μ-induced random walk on X satisfies (λ', α)-quantitative equidistribution for any $\lambda' \in (0, \lambda)$.

1.3 The Case of a Torus

Previous works [6, 7, 14–16] on the case of a torus have been conducted by Bourgain, Furman, Mozes, Boyer, Saxcé, and the authors of the present paper. The most general result known when this paper is written can be summarized as follows. Recall that if μ is a Borel probability measure on $\mathrm{GL}_d(\mathbb{R})$, the top Lyapunov exponent of μ is

$$\lambda_{1,\mathbb{R}^d}(\mu) = \lim_{n\to+\infty} \frac{1}{n} \int \log\|g\| \, \mathrm{d}\mu^{*n}(g).$$

Theorem B ([6, 14, 15]; cf. Appendix B) *Let $X = \mathbb{R}^d/\mathbb{Z}^d$ for some $d \geq 2$. Let μ be a probability measure on* $\mathrm{Aut}(X)$ *having finite exponential moment. Denote by $\Gamma \subset \mathrm{GL}_d(\mathbb{Z})$ the subgroup generated by* $\mathrm{Supp}(\mu)$.

Assume that the action of Γ on $\mathbb{R}^d$ is strongly irreducible. Then the μ-induced random walk on X satisfies (λ, α)-quantitative equidistribution for any λ in the range $(0, \lambda_{1,\mathbb{R}^d}(\mu))$ and any $\alpha \in (0, 1]$.

Note that the assumption that $\Gamma \curvearrowright \mathbb{R}^d$ is strongly irreducible implies that X itself is the only non-trivial Γ-invariant factor torus. Thus, a proper Γ-invariant closed subgroup of X is a finite set of rational points. Its height controls the size of denominators.

Theorem C ([7, 16]; cf. Appendix B) *Let $X = \mathbb{R}^d/\mathbb{Z}^d$ for some $d \geq 2$. Let μ be a finitely supported probability measure on* $\mathrm{Aff}(X)$*. Denote by $\Gamma \subset \mathrm{GL}_d(\mathbb{Z})$ the subgroup generated by* $\mathrm{Supp}(\theta_*\mu)$*.*

Assume that the action of Γ on $\mathbb{R}^d$ is strongly irreducible. Then given $\lambda \in (0, \lambda_{1,\mathbb{R}^d}(\theta_*\mu))$ and $\alpha \in (0, 1]$, there exists $C = C(\theta_*\mu, \lambda, \alpha)$ such that the μ-induced random walk on X satisfies (C, λ, α)-quantitative equidistribution.

Note that the constant C depends only on $\theta_*\mu$ and not on the translation part of the elements in $\mathrm{Supp}(\mu)$.

Note that the statements of Theorem B and Theorem C are somewhat stronger than those in [15] and [16], in that there is no assumption that the Zariski closure of Γ is connected. In Appendix B we explain how this assumption can be eliminated.

1.4 Consequences of the Main Theorem

Assume that there is a filtration

$$1 = Z_0 \subset Z_1 \subset \cdots \subset Z_{l-1} \subset Z_l = N$$

of rational closed connected subgroups such that Z_k/Z_{k-1} is central in N/Z_{k-1} for all $k = 1, \ldots, l$. Denote $X_k = N/(Z_k\Lambda)$ for $k = 0, \ldots, l-1$. Thus, we have a tower of nilmanifolds,

$$X = X_0 \to X_1 \to \cdots \to X_{l-1} = Z_l/(Z_{l-1}\Lambda)$$

where the last nilmanifold is a torus. Theorem 1.1 combined with Theorem B immediately leads to the following statement.

Theorem 1.3 *Let μ be a probability measure on* $\mathrm{Aut}(X)$ *having a finite exponential moment or a finitely supported probability measure on* $\mathrm{Aff}(X)$*. Let Γ denote the subgroup of* $\mathrm{Aut}(X)$ *generated by the support of $\theta_*\mu$. Assume*

(i) For all $k = 1, \ldots, l-1$, Z_k is Γ-invariant.
(ii) For all $k = 1, \ldots, l-1$, $\tau_{Z_k/Z_{k-1}}(\mu) < 2\sigma_{X_{k-1},X_k}(\mu)$.
(iii) The action of Γ on Z_l/Z_{l-1} is strongly irreducible.

Then the μ-random walk on X is (λ, α)-quantitatively equidistributed for any $\lambda \in (0, \lambda_{1,Z_l/Z_{l-1}}(\mu))$ and any $\alpha \in (0, 1]$.

Note that unlike in [16, Theorem 1.3] (cf. Theorem C), here the implicit constant C of the (λ, α)-quantitative equidistribution does depend on the translation part, though mildly. For more details about the dependence of the implicit constants on the translation part, see Lemma 4.5.[2]

[2] Lemma 4.5 contains the key inductive step, combined with [16, Theorem 1.3] one can easily see how the constants in Theorem 1.3 depend on the translation parts. As this is not particularly illuminating, we do not give an explicit discussion here.

From this quantitative statement, i.e., Theorem 1.3, we can deduce easily the following qualitative statement.

Corollary 1.4 *Let μ be either probability measure on* $\mathrm{Aut}(X)$ *with finite exponential moment or a Borel probability measure on* $\mathrm{Aff}(X)$ *with finite support. Let* Γ *denote the subgroup generated by* $\mathrm{Supp}(\theta_*\mu)$*. Let H denote the subgroup generated by* $\mathrm{Supp}(\mu)$.

Assume the same assumptions as in Theorem 1.3. Then for any $x \in X$, either $\mu^{*m} * \delta_x$ converge to m_X *in the weak-$*$ topology or the projection of the orbit Hx to the maximal torus factor is contained in a proper closed H-invariant subset.*

Clearly, the two options in Corollary 1.4 are mutually exclusive.

From Corollary 1.4, follow easily the following classification theorem about orbit closures and stationary measures: if μ and Γ are as in in Theorem 1.3, then a Γ-orbit closure is either X or projects to a proper closed Γ-invariant subset on the maximal torus factor. Similarly, an ergodic μ-stationary measure on X is either m_X or supported on a proper closed invariant subset. However, these classification theorems can be deduced from the work of Benoist and Quint [4] and the work of Eskin and the third-named author [11], works that deal with the much more general context of random walks on homogeneous spaces. For instance, [4, Corollary 1.10] states as follows. In the case of automorphism action, if Γ is a finitely generated subgroup of $\mathrm{Aut}(X)$, whose Zariski closure in $\mathrm{Aut}(N)$ is a Zariski connected semisimple subgroup with no compact factor, then every Γ-orbit closure $\overline{\Gamma x}$ is a finite homogeneous union of affine submanifolds. If moreover μ is a probability measure Γ whose support generates Γ, then the Cesàro mean $\frac{1}{n}\sum_{m=1}^{n}\mu^{*m} * \delta_x$ converges in the weak-$*$ topology to the homogeneous measure probability measure on $\overline{\Gamma x}$ (the measure induced by the Haar measure on the stabilizer of $\overline{\Gamma x}$). In [11], the requirement on semisimplicity is relaxed. However, there does not seem to be at present a purely ergodic theoretic approach to Corollary 1.4 without the additional Cesàro mean.

1.5 Idea of the Proof

To conclude this introduction, let us explain the conceptual ideas behind the proof of the main theorem. Let f be a test function on X which witnesses $\mathcal{W}_\alpha(\mu^{*m} * \delta_x, \mathrm{m}_X) > t$. The goal is to construct another witness which has the additional property that it is constant on each fiber of $\pi : X \to Y$. Using Fourier analysis on the fibers (i.e., the Z-direction), we may assume without loss of generality that f behaves as a character on the fibers. Now sample random elements g_1 and g_2 in $\mathrm{Aff}(X)$ according to $\mu^{*m'}$ with some $m' \leq m$ and set $f_1 = f \circ g_1$ and $f_2 = f \circ g_2$. Because $\tau_Z(\mu)$ is small, with large probability, $\theta_Z(g_1) = \theta_Z(g_2)$ so that $f_1\overline{f_2}$ is constant on each fiber. And because $\sigma_{X,Y}(\mu)$ is large, $f_1\overline{f_2}$ is a witness to $\mathcal{W}_\alpha(\mu^{*(m-m')} * \delta_x, \mathrm{m}_X) > t^{O(1)}$ also with large probability.

2 Examples

This section is devoted to a few concrete examples where our main theorem applies and an example where it does not apply.

2.1 Heisenberg Nilmanifold

Let N be the $(2d+1)$-dimensional Heisenberg group. Recall that a Heisenberg group is a two-step simply connected nilpotent Lie group of one-dimensional center. Let Z denote the center of N. Note that $[N, N] = Z$ and it is isomorphic to $\mathbb{R}$.

Let Λ be a lattice in N and set $X = N/\Lambda$. The maximal factor torus of X is $T = N/[N, N]\Lambda = N/Z\Lambda$. Let μ be a Borel probability measure on $\mathrm{Aff}(X)$ with finite support. Let H denote the subgroup generated by $\mathrm{Supp}(\mu)$ and $\Gamma = \theta(H)$. Assume that

$$\text{the action of } \Gamma \text{ on } N/Z \text{ is strongly irreducible.} \tag{3}$$

We claim that the assumptions of Theorem 1.3 are satisfied for the filtration $\{0\} \subset Z \subset N$. Indeed, $\{0\} \subset Z \subset N$ is the ascending central series of N. Hence the assumptions on the filtration are satisfied. It remains to see that $\tau_Z(\mu) < 2\sigma_{X,T}(\mu)$. On the one hand, (3) implies the condition A of Theorem A. Hence the action of Γ on X has a spectral gap, which implies $\sigma_{X,T}(\mu) > 0$ (we remark that the special case of Theorem A we use here for Heisenberg nilmanifolds was established by Bekka and Heu in [3]). On the other hand, any $\gamma \in \mathrm{Aut}(X)$ preserves both Z and the lattice $Z \cap \Lambda$ in Z. Hence the action of $\mathrm{Aut}(X)$ on Z consists only of $\{\pm 1\}$. It follows that $\tau_Z(\mu) = 0$, establishing condition (ii) of Theorem 1.3, and hence Theorem 1.3 applies to Heisenberg nilmanifolds.

Qualitatively, we can say a little bit more than Corollary 1.4.

Theorem 2.1 *Let X be a Heisenberg nilmanifold and μ a probability measure on* $\mathrm{Aut}(X)$ *having a finite exponential moment, and let Γ denote the subgroup generated by* $\mathrm{Supp}(\mu)$. *Assume the irreducibility condition* (3) *holds. Then for every $x \in X$, either $\mu^{*n} * \delta_x$ converges to m_X in the weak-$*$ topology or the Γ-orbit of x is finite.*

Proof By the discussion above, Corollary 1.4 applies. Thus, it is enough to see that if the image of x in $T = N/Z\Lambda$ is rational, then the $\mathrm{Aut}(X)$-orbit of x is finite.

By [10, Theorem 5.1.8], the $\mathbb{Q}$-span of $\log(\Gamma)$ is a $\mathbb{Q}$-structure of the Lie algebra $\mathrm{Lie}(N)$ of N. We can choose a basis of this $\mathbb{Q}$-structure and identify both $\mathrm{Lie}(N)$ and N with $\mathbb{R}^{2d+1} = \mathbb{R}^{2d} \oplus Z$ so that the projection of Λ to $\mathbb{R}^{2d}$ is exactly $\mathbb{Z}^{2d}$. Then every automorphism $\gamma \in \mathrm{Aut}(X)$ is of the form

$$(y, t) \in \mathbb{R}^{2d} \oplus Z \mapsto (A_\gamma y, \epsilon_\gamma t + L_\gamma y,) \in \mathbb{R}^{2d} \oplus Z$$

where $A_\gamma \in \mathrm{GL}_{2d}(\mathbb{Z})$, $\epsilon_\gamma \in \{\pm 1\}$, and $L_\gamma \colon \mathbb{R}^{2d} \to \mathbb{R}$ is a linear form. From [10, Theorem 5.4.2], we know that there is an integer $q \in \mathbb{N}$ such that $\Lambda \subset \mathbb{Z}^{2d} \oplus \frac{1}{q}\mathbb{Z} \subset \mathbb{R}^{2d} \oplus Z$. This implies that the linear form L_γ must be rational of denominator q.

It follows that if $(y, t) \in \mathbb{R}^{2d} \oplus Z$ with y rational of denominator q', then

$$\mathrm{Aut}(X)(y, t) \subset \frac{1}{q'}\mathbb{Z}^{2d} \times \left(\{\pm t\} + \frac{1}{qq'}\mathbb{Z}\right) \subset \mathbb{R}^{2d} \oplus Z.$$

Since the group law in these coordinates is bilinear with rational structural constants, this allows to conclude that $\Gamma.(y, t)$ is finite. □

2.2 Heisenberg Nilmanifold over Number Fields

In the example above, the growth rate $\tau_Z(\mu)$ for the action on the center Z is equal to 0 because this action is virtually trivial. In the next example, we have again a two-step nilpotent group N, but the group $\Gamma \subset \mathrm{Aut}(X)$ will have a non-trivial action on the center.

Let $B \colon \mathbb{C}^{2d} \times \mathbb{C}^{2d} \to \mathbb{C}$ be a bilinear form with integral coefficients in the standard basis. For a commutative ring with unity R, define $\mathrm{Heis}_B(R)$ to be the group with underlying set $R^{2d+1} = R^{2d} \times R$ and with the group law $\forall (y, t), (y', t') \in R^{2d} \times R$,

$$(y, t)(y', t') = \big(y + y', t + t' + B(y, y')\big).$$

Let K be a number field. Denote by $\mathcal{O}_K$ its ring of integers and by $\mathcal{O}_K^\times$ the group of units. Let r_1 be the number of embeddings of K in $\mathbb{R}$ and r_2 the number of conjugate pairs of embeddings of K in $\mathbb{C}$. Let $\iota \colon K \to \mathbb{R}^{r_1} \times \mathbb{C}^{r_2}$ be the corresponding ring embedding so that $\iota(\mathcal{O}_K)$ is discrete, which in fact is a lattice in $\mathbb{R}^{r_1} \times \mathbb{C}^{r_2}$. This embedding of rings induces an embedding of groups $\mathrm{Heis}_B(\mathcal{O}_K) \to \mathrm{Heis}_B(\mathbb{R}^{r_1} \times \mathbb{C}^{r_2})$, which we denote again by ι. Let $N = \mathrm{Heis}_B(\mathbb{R}^{r_1} \times \mathbb{C}^{r_2})$ and $\Lambda = \iota(\mathrm{Heis}_B(\mathcal{O}_K))$. It is easy to check that Λ is a lattice in N. Hence $X = N/\Lambda$ is a nilmanifold.

Inside $\mathrm{Aut}(\Lambda)$, we have automorphisms of the form

$$(y, t) \in \mathrm{Heis}(\mathcal{O}_K) \mapsto (Ay, \epsilon t + Ly) \in \mathrm{Heis}(\mathcal{O}_K)$$

where $A \subset \mathrm{GL}_{2d}(\mathcal{O}_K)$, $\epsilon \in \mathcal{O}_K^\times$, and $L \in (\mathcal{O}_K^{2d})^*$ such that

$$\forall y, y' \in K^{2d}, B(Ay, Ay') = \epsilon B(y, y').$$

They extend to $\mathrm{Aut}(N)$ via ι. Denote by $\Gamma_0 \subset \mathrm{Aut}(X)$ the group consisting of such automorphisms. For example, for $d = 1$ and $B = \det$, A can be any matrix in $\mathrm{GL}_2(\mathcal{O}_K)$ and $\epsilon = \det(A)$, and Γ_0 is isomorphic to a semi-direct product $\mathrm{GL}_2(\mathcal{O}_K) \ltimes \mathcal{O}_K^2$.

Consider the central subgroup

$$Z = \{0\} \oplus \mathbb{R}^{r_1} \times \mathbb{C}^{r_2} \subset (\mathbb{R}^{r_1} \times \mathbb{C}^{r_2})^{2d} \times (\mathbb{R}^{r_1} \times \mathbb{C}^{r_2}) = N.$$

Then Γ_0 preserves Z. Let μ be a probability measure on Γ_0. Necessarily, $\tau_Z(\mu) = 0$ because Γ_0 acts on Z via the abelian group $\mathcal{O}_K^\times$, which grows at polynomial rate. Let Γ be the group generated by $\mathrm{Supp}(\mu)$. The action of Γ on $N/Z = (\mathbb{R}^{r_1} \times \mathbb{C}^{r_2})^{2d}$ can be identified with $\Gamma \to \mathrm{GL}_{2d}(\mathcal{O}_K) \xrightarrow{\iota} \mathrm{GL}_{2d}(\mathbb{R})^{r_1} \times \mathrm{GL}_{2d}(\mathbb{C})^{r_2}$. If the action of Γ on K^{2d} is strongly irreducible over K, then the action of Γ on N/Z is strongly irreducible over $\mathbb{Q}$. If moreover Γ is not virtually abelian, then the condition (iii) of Theorem A is satisfied. We conclude that

$$\tau_Z(\mu) = 0 < 2\sigma_{X,T}(\mu)$$

where $T = N/(\Lambda Z)$. Hence Theorem 1.1 applies. However, Theorem B does **not** apply to the induced random walk on T, as the action of Γ on N/Z is not irreducible over $\mathbb{R}$ unless $r_1 + r_2 = 1$ (it is strongly irreducible over $\mathbb{Q}$ unless K is a totally complex extension of a totally real field, c.f. e.g., [19, §2]). However, it is conjectured that a quantitative equidistribution holds for such random walks on T, at least under the assumption that the projection of Γ to $\mathrm{GL}_{2d}(\mathbb{R})^{r_1} \times \mathrm{GL}_{2d}(\mathbb{C})^{r_2}$ has semisimple Zariski closure with no compact factor.

2.3 A Non-semisimple Group of Toral Automorphisms

In both examples above, the growth rates of the action on the fibers are all zero. Now we give an example where we have a positive growth rate while our result still applies.

Consider $X = \mathbb{T}^{2d} = \mathbb{R}^{2d}/\mathbb{Z}^{2d}$ with $d \geq 2$. Let A and D be independent random elements in $\mathrm{SL}_d(\mathbb{Z})$. Denote by η the law of A and ν that of D. Let I_d denote the $d \times d$ identity matrix. Let μ be the law of the random block triangular matrix

$$\left(\begin{array}{c|c} A & I_d \\ \hline 0 & D \end{array}\right).$$

Let $Z = \mathbb{R}^d \oplus \{0\} \subset \mathbb{R}^{2d}$ and $Y = \mathbb{R}^{2d}/\mathbb{R}^d \oplus \mathbb{Z}^d$. The filtration $\{0\} \subset Z \subset \mathbb{R}^{2d}$ is preserved by Γ, the group generated by the support of μ.

Proposition 2.2 *In the above setting, given the measure η, there is some ν such that Theorem 1.3 can be applied to μ and the filtration $\{0\} \subset Z \subset \mathbb{R}^{2d}$.*

As a consequence, we can say the following about orbit closures under the action of Γ, the group generated by the support of the constructed μ. For every $x \in \mathbb{T}^{2d}$, either Γx is dense or Γx is contained in a finite union of affine subtori parallel to $\mathbb{R}^d/\mathbb{Z}^d \oplus \{0\}$. For properly chosen η, the group Γ will **not** have semisimple Zariski closure. Thus, the work of Benoist-Quint [4] does not apply to such group. Neither does the work of Guivarc'h-Starkov [13] nor that of Muchnik [17] (though stationary measures even in this case are analyzed by Eskin and the third named author in [11]).

To show the proposition, we need the following lemma to control $\sigma_{X,Y}(\mu)$.

Lemma 2.3 *In the setting above, denote by* $(U_Y, L^2(Y, \mathrm{m}_Y))$ *the Koopman representation associated to the action of* $\mathrm{Aut}(Y)$ *on* Y *and by* $U_{Y,0}$ *the restriction of* U_Y *to the subspace of mean zero functions. Then we have*

$$\|U_{X,Y}(\mu)^2\| \leq \sqrt{3}\|U_{Y,0}(\nu)\|.$$

Proof Let $F\colon L^2(X, m_X) \to \ell^2(\mathbb{Z}^{2d})$ denote the isometry given by the Fourier transform. Under this isometry, $U_{X,Y}$ is conjugated to a unitary representation T of Γ on $\ell^2((\mathbb{Z}^d \setminus \{0\}) \times \mathbb{Z}^d)$. Explicitly, let $\varphi \in \ell^2((\mathbb{Z}^d \setminus \{0\}) \times \mathbb{Z}^d)$. Then for all $(a, b) \in (\mathbb{Z}^d \setminus \{0\}) \times \mathbb{Z}^d$,

$$(T(\mu)\varphi)(a, b) = \int_\Gamma \varphi({}^t g(a, b))\,\mathrm{d}\mu(g) = \mathbb{E}\big[\varphi({}^t Aa, a + {}^t Db)\big].$$

Let P_0 be the orthogonal projection $\ell^2((\mathbb{Z}^d \setminus \{0\}) \times \mathbb{Z}^d) \to \ell^2((\mathbb{Z}^d \setminus \{0\}) \times \{0\})$. Concretely, for $\varphi \in \ell^2((\mathbb{Z}^d \setminus \{0\}) \times \mathbb{Z}^d)$ and $(a, b) \in (\mathbb{Z}^d \setminus \{0\}) \times \mathbb{Z}^d$,

$$(P_0\varphi)(a, b) = \delta_0(b)\varphi(a, 0).$$

Then $P_0 T(\mu) P_0 = 0$ because

$$(P_0 T(\mu) P_0 \varphi)(a, b) = \delta_0(b)\mathbb{E}\big[\delta_0(a)\varphi({}^t Aa, 0)\big] = 0.$$

Hence, taking the square of the equality $T(\mu) = P_0 T(\mu) + (1 - P_0)T(\mu)$, we see,

$$\|T(\mu)^2\| \leq 3\|(1 - P_0)T(\mu)\|.$$

To conclude, it suffices to show

$$\|(1 - P_0)T(\mu)\| \leq \|U_{Y,0}(\nu)\|. \tag{4}$$

We first show the inequality in the case where A is almost surely some fixed matrix $g \in \mathrm{SL}_d(\mathbb{Z})$. Consider, for $a \in \mathbb{Z}^d \setminus \{0\}$, the subspace

$$\mathcal{H}_a = \ell^2(\{a\} \times \mathbb{Z}^d) \subset \ell^2((\mathbb{Z}^d \setminus \{0\}) \times \mathbb{Z}^d).$$

Let Q_a denote the orthogonal projection onto $\mathcal{H}_a$. Observe that

$$\forall a \in \mathbb{Z}^d \setminus \{0\}, T(\mu)\mathcal{H}_a \subset \mathcal{H}_{{}^t g^{-1} a}.$$

Moreover P_0 preserves the subspaces $\mathcal{H}_a$. Hence, for any $\varphi \in \ell^2((\mathbb{Z}^d \setminus \{0\}) \times \mathbb{Z}^d)$, the vectors $(1 - P_0)T(\mu)Q_a\varphi$, $a \in \mathbb{Z}^d \setminus \{0\}$, are all orthogonal to each other. Thus

$$\begin{aligned}\|(1 - P_0)T(\mu)\varphi\|^2 &= \sum_{a\in\mathbb{Z}^d\setminus\{0\}} \|(1 - P_0)T(\mu)Q_a\varphi\|^2 \\ &\leq \sum_{a\in\mathbb{Z}^d\setminus\{0\}} \|(1 - P_0)T(\mu)Q_a\|^2 \|Q_a\varphi\|^2 \\ &\leq \Big(\sup_{a\in\mathbb{Z}^d\setminus\{0\}} \|(1 - P_0)T(\mu)Q_a\|^2\Big)\|\varphi\|^2.\end{aligned}$$

By identifying $\mathcal{H}_a$ with $\ell^2(\mathbb{Z}^d)$ in the obvious way, we see that $\|(1-P_0)T(\mu)Q_a\| = \|V_a\|$ where $V_a \colon \ell^2(\mathbb{Z}^d) \to \ell^2(\mathbb{Z}^d \setminus \{0\})$ is the operator defined by

$$\forall\psi \in \ell^2(\mathbb{Z}^d),\ \forall b \in \mathbb{Z}^d \setminus \{0\}, \quad (V_a\psi)(b) = \mathbb{E}\big[\psi({}^t g^{-1}a + {}^t Db)\big].$$

Let $W_a \colon \ell^2(\mathbb{Z}^d) \to \ell^2(\mathbb{Z}^d)$ be the isometry induced by translating the index by ${}^t g^{-1}a$, so that $V_a = V_0 W_a$. But V_0 is conjugated to $U_{Y,0}(\nu)$ via the Fourier transform. Hence

$$\sup_{a\in\mathbb{Z}^d\setminus\{0\}} \|(1 - P_0)T(\mu)Q_a\| \leq \sup_{a\in\mathbb{Z}^d\setminus\{0\}} \|V_a\| \leq \|V_0\| = \|U_{Y,0}(\nu)\|.$$

This shows (4) for the special case where A is almost surely constant.

Using the independence between A and D, we can write

$$\mu = \int_{\mathrm{SL}_d(\mathbb{Z})} \mu_g \,\mathrm{d}\eta(g),$$

with μ_g being the law of the random matrix

$$\left(\begin{array}{c|c} g & I_d \\ \hline 0 & D \end{array}\right).$$

Then $\|T(\mu)\| \leq \int_{\mathrm{SL}_d(\mathbb{Z})} \|T(\mu_g)\| \,\mathrm{d}\eta(g)$ proves (4). □

Proof of Proposition 2.2 Once η is chosen, the action of Γ on Z is determined. Hence $\tau_Z(\mu)$ is determined.

Let ν_0 be a symmetric probability measure on $\mathrm{SL}_d(\mathbb{Z})$ whose support generates a Zariski-dense subgroup. Then by a result of Furman and Shalom [12, Theorem 6.5] (which is a special case of Theorem A),

$$\|U_{Y,0}(\nu_0)\| < 1.$$

Let $\nu = \nu_0^{*k}$ where k is an integer. By choosing k large enough, we can make $\|U_{Y,0}(\nu_0)\|$ arbitrarily small and hence $\sigma_{X,Y}(\mu)$ arbitrarily large by Lemma 2.3. This ensures that

$$\tau_Z(\mu) < 2\sigma_{X,Y}(\mu).$$

At the same time, the support of ν generates a Zariski dense subgroup Γ in SL_d. In particular the action of Γ on $\mathbb{R}^{2d}/Z$ is strongly irreducible. This is why Theorem 1.3 can be applied. □

2.4 A Non-example

Let N be the connected and simply connected nilpotent Lie group whose Lie algebra is the free *two*-step nilpotent Lie algebra on three generators. It can be realized as $N = \mathbb{R}^3 \oplus \mathbb{R}^3$ with the group law being

$$(x, y)(x', y') = (x + x', y + y' + x \wedge x'), \quad \text{for all } x, x', y,' y \in \mathbb{R}^3,$$

where $\wedge$ denotes the usual cross product on $\mathbb{R}^3$. As explained in [2, Example 35], the automorphism group $\mathrm{Aut}(N)$ of N is isomorphic to the subgroup of $\mathrm{GL}_6(\mathbb{R})$ of matrices $g_{A,B}$ of the form

$$g_{A,B} = \left(\begin{array}{c|c} A & 0 \\ \hline B & \det(A)(A^{\mathrm{tr}})^{-1} \end{array}\right),$$

with $A \in \mathrm{GL}_3(\mathbb{R})$ and B any 3×3 matrix with real coefficients. Here $\mathrm{Aut}(N)$ acts on the center Z of N via $\theta_Z \colon g_{A,B} \mapsto \det(A)(A^{\mathrm{tr}})^{-1}$ and acts on N/Z via $\theta_{N/Z} \colon g_{A,B} \mapsto A$.

Let Λ be any lattice in N and set $X = N/\Lambda$. Let μ be a probability measure on $\mathrm{Aut}(X)$ and Γ the group generated by its support. Denote moreover $Y = N/(\Lambda Z)$. In order to apply Theorem 1.1 to the factor map $X \to Y$, we need

(i) $\tau_Z(\mu)$ to be small; informally, that is $\theta_Z(\Gamma)$ is a small group.
(ii) $\sigma_{X,Y}(\mu)$ to be large; in view of Theorem A, this requires $\theta_{N/Z}(\Gamma)$ to be a large group (not virtually amenable by [2, Theorem 1 and Theorem 5]).

But $\theta_Z(\Gamma)$, isomorphic to $\theta_{N/Z}(\Gamma)$, cannot be small and large at the same time. This is why, very likely, Theorem 1.1 does not apply to such random walks. However, we still expect the conclusion of Theorem 1.1 to hold, provided that $\theta_{N/Z}(\Gamma)$ is a large group (e.g., Zariski dense in $\mathrm{SL}_3(\mathbb{R})$).

3 The Setup

Throughout this paper, $X = N/\Lambda$ denotes a nilmanifold. As recalled in the introduction, this means that N is a connected simply connected nilpotent Lie group and $\Lambda \subset N$ is a lattice, which is necessarily cocompact ([18, Theorem 2.1]). Recall that the $\mathbb{Q}$-span of $\log(\Lambda)$ defines a $\mathbb{Q}$-structure on $\mathrm{Lie}(N)$, the Lie algebra of N. A connected closed subgroup of N is said to be rational if its Lie algebra is rational in $\mathrm{Lie}(N)$ with respect to this $\mathbb{Q}$-structure. For a connected closed subgroup $M \subset N$ to be rational, it is necessary and sufficient that $M \cap \Lambda$ is a lattice in M. For these, see [10, §5.1].

Denote by $\mathrm{Aut}(X) = \mathrm{Aut}(N/\Lambda)$ denote the group of continuous automorphisms of N preserving Λ. Let $\mathrm{Aff}(X) = \mathrm{Aff}(N/\Lambda) = \mathrm{Aut}(X) \ltimes N$ denote the group of (invertible) affine transformations of X. More precisely for $\gamma \in \mathrm{Aut}(X)$ and $n \in N$, let $(\gamma, n) \in \mathrm{Aff}(X)$ denote the map $X \to X$, $x\Lambda \mapsto n\gamma(x)\Lambda$. Denote by $\theta\colon \mathrm{Aff}(X) \to \mathrm{Aut}(X)$ the projection to the automorphism part, that is, $\theta(\gamma, n) = \gamma$ for all $(\gamma, n) \in \mathrm{Aff}(X)$.

Moreover, we will identify an automorphism $\gamma \in \mathrm{Aut}(X)$ with $(\gamma, 1_N) \in \mathrm{Aff}(X)$ and an element $n \in N$ with the left translation $(1, n) \in \mathrm{Aff}(X)$. With this notation we have, for all $\gamma \in \mathrm{Aut}(X)$ and all $n \in N$, $\gamma n \gamma^{-1} = \gamma(n)$. If $g \in \mathrm{Aff}(X)$ and $n \in N$ is central, then $gng^{-1} = \theta(g)(n)$.

Let m_X denote the normalized N-invariant measure on X induced by the Haar measure of N. The action $\mathrm{Aff}(X) \curvearrowright X$ preserves m_X. Let $(U, L^2(X, \mathrm{m}_X))$ denote the associated Koopman representation. That is, for $g \in \mathrm{Aff}(X)$, $U(g)$ is the unitary operator on $L^2(X, \mathrm{m}_X)$ defined by

$$\text{for all } f \in L^2(X, \mathrm{m}_X) \text{ and almost all } x \in X,\ U(g)f(x) = f(g^{-1}(x)).$$

Let also $U^*(g) = U(g)^* = U(g^{-1})$. By an abuse of notation, we let $U(g)$ and $U^*(g)$ denote also the operators from $\mathcal{C}^0(X)$, the space of continuous functions, to itself defined in the obvious way.

Let μ be a Borel measure on $\mathrm{Aff}(X)$. We set $U(\mu) = \int U(g)\,\mathrm{d}\mu(g)$ and $U^*(\mu) = \int U^*(g)\,\mathrm{d}\mu(g)$. For any integer $m \geq 0$, any Borel measure η on X, and any continuous function $f \in \mathcal{C}^0(X)$, we have

$$\int_X f\,\mathrm{d}\mu^{*m} * \eta = \int_X U^*(\mu)^m f\,\mathrm{d}\eta.$$

3.1 Hölder Functions

We fix a Riemannian metric on X and let $d\colon X \times X \to [0, +\infty)$ denote the associated distance function. Let $\alpha \in (0, 1]$. Denote by $\mathcal{C}^{0,\alpha}(X)$ the set of α-Hölder continuous functions from X to $\mathbb{C}$. Endow it with the norm

$$\|f\|_{0,\alpha} = \|f\|_\infty + \omega_\alpha(f)$$

where

$$\omega_\alpha(f) = \sup_{x \neq y \in X} \frac{|f(x) - f(y)|}{d(x, y)^\alpha}.$$

For $g \in \mathrm{Aff}(X)$, define

$$\mathrm{Lip}_X(g) = \sup_{x,x' \in X,\, x \neq x'} \frac{d(gx, gx')}{d(x, x')}.$$

This quantity is finite since g is of class $\mathcal{C}^\infty$ and d is a Riemannian distance. It is greater or equal to 1 since X is compact. Moreover, $\mathrm{Lip}_X\colon \mathrm{Aff}(X) \to [1, +\infty)$ is continuous and submultiplicative, i.e., for all $g, h \in G$,

$$\mathrm{Lip}_X(gh) \le \mathrm{Lip}_X(g)\,\mathrm{Lip}_X(h). \tag{5}$$

It is straightforward to check that if $g \in \mathrm{Aff}(X)$ and $f \in \mathcal{C}^{0,\alpha}(X)$, then $U^*(g)f$ is still α-Hölder continuous and

$$\|U^*(g)f\|_{0,\alpha} \le \mathrm{Lip}_X(g)^\alpha \|f\|_{0,\alpha}.$$

Remark also that for $f_1, f_2 \in \mathcal{C}^{0,\alpha}(X)$, then $f_1 f_2 \in \mathcal{C}^{0,\alpha}(X)$ and

$$\|f_1 f_2\|_{0,\alpha} \le \|f_1\|_{0,\alpha} \|f_2\|_{0,\alpha}. \tag{6}$$

4 The Main Argument

As in the statement of Theorem 1.1, let μ be Borel measure on $\mathrm{Aff}(X)$ having a finite exponential moment. Let $\Gamma \subset \mathrm{Aut}(X)$ denote the subgroup generated by the support of $\theta_*\mu$. Let $Z \subset N$ be a Γ-invariant rational connected closed central subgroup. Then $Y = N/(\Lambda Z)$ is a nilmanifold, and we have a $\Gamma \ltimes N$-equivariant factor map $\pi : X \to Y$. Let m_Y denote the N-invariant probability measure on Y induced by the Haar measure of N. We defined two quantities $\tau_Z(\mu)$ and $\sigma_{X,Y}(\mu)$ in the introduction. This section is dedicated to the proof of the following proposition.

Proposition 4.1 *Assume that μ has a finite β-exponential moment. Assume*

$$\tau_Z(\mu) < 2\sigma_{X,Y}(\mu).$$

Then given $0 < \alpha \leq \min\{1, \beta\}$, there exists a constant $C \geq 2$ such that the following holds.

For any Borel probability measure η on X, any $t \in (0, 1/2)$, and any $m \geq C \log \frac{1}{t}$, if

$$\mathcal{W}_\alpha(\mu^{*m} * \eta, \mathrm{m}_X) \geq t,$$

then

$$\mathcal{W}_\alpha(\pi_*\eta, \mathrm{m}_Y) \geq e^{-Cm}.$$

In other words, if there is $f \in \mathcal{C}^{0,\alpha}(X)$ satisfying

$$\left| \int_X f \,\mathrm{d}\mu^{*m} * \eta - \int_X f \,\mathrm{d}\,\mathrm{m}_X \right| > t \|f\|_{0,\alpha},$$

then there exists $\varphi \in \mathcal{C}^{0,\alpha}(Y)$ such that

$$\left| \int_X \varphi \,\mathrm{d}\pi_*\eta - \int_X \varphi \,\mathrm{d}\,\mathrm{m}_Y \right| > e^{-Cm} \|\varphi\|_{0,\alpha}.$$

4.1 Principal Torus Bundle

Let $S = Z/(Z \cap \Lambda)$. Let $d = \dim Z$. Then S is a torus of dimension d. Note that π is a fiber bundle of fiber S. Moreover, it is a principal bundle: since Z is contained in the center of N, the action of Z by left translation on X factors through S.

By choosing a basis in $Z \cap \Lambda$, we fix an isomorphism between $\mathbb{Z}^d$ and the group $\mathrm{Hom}(S, S^1)$ of unitary characters of S. Denote the isomorphism as $a \mapsto \chi_a$, $a \in \mathbb{Z}^d$. The Koopman representation U restricted to Z factors through S. Hence, we can decompose $L^2(X, \mathrm{m}_X)$ into a Hilbert sum of characteristic subspaces

$$L^2(X, \mathrm{m}_X) = \sum_{a \in \mathbb{Z}^d} \mathcal{H}_a \tag{7}$$

where for $a \in \mathbb{Z}^d$,

$$\mathcal{H}_a = \{ f \in L^2(X, \mathrm{m}_X) \mid \forall z \in Z,\ U(z)f = \chi_a(z)f \}.$$

Here we identified χ_a with its lift as character of Z. For $a = 0$, $\mathcal{H}_0$ is the subspace of functions that are constant on each fiber of π. Since $\pi_* \mathrm{m}_X = \mathrm{m}_Y$, we have the isometry

$$\mathcal{H}_0 = L^2(Y, \mathrm{m}_Y).$$

Thus, the Hilbert space of the representation $U_{X,Y}$ is precisely $\sum_{a\in\mathbb{Z}^d\setminus\{0\}} \mathcal{H}_a$.

Since for all $g \in \mathrm{Aff}(X)$ and $z \in Z$, $zg = g\theta(g)^{-1}(z)$, we have

$$\forall g \in \mathrm{Aff}(X),\ \forall a \in \mathbb{Z}^d, \quad U(g)\mathcal{H}_a = \mathcal{H}_{\theta(g)\cdot a},$$

where $\gamma \cdot a \in \mathbb{Z}^d$ is such that $\chi_{\gamma\cdot a} = \chi_a \circ \gamma^{-1}$ for $\gamma \in \Gamma$. This defines an action of Γ on $\mathbb{Z}^d$. Note that Γ acts via some homomorphism $\Gamma \to \mathrm{GL}_d(\mathbb{Z})$.

4.2 Fourier Transform

For continuous functions, the decomposition (7) can be made more explicit using Fourier transforms. The aim here is to prove the following lemma using Fourier transforms.

Lemma 4.2 *Given $\alpha \in (0, 1]$, there is a constant C depending on α such that the following holds. If a measure η on X, a function $f \in \mathcal{C}^{0,\alpha}(X)$, and $t \in (0, 1/2)$ satisfy*

$$\left| \int_X f \,\mathrm{d}\eta - \int_X f \,\mathrm{d}\,\mathrm{m}_X \right| \geq t \|f\|_{0,\alpha},$$

then there exist $a_0 \in \mathbb{Z}^d$ with $\|a_0\| \leq t^{-C}$ and $f_0 \in \mathcal{C}^{0,\alpha}(X) \cap \mathcal{H}_{a_0}$ such that

$$\left| \int_X f_0 \,\mathrm{d}\eta - \int_X f_0 \,\mathrm{d}\,\mathrm{m}_X \right| \geq t^C \|f_0\|_{0,\alpha}.$$

Specializing this lemma to the case where X is a torus and Y is a point, we can recover Lemma 4.5 in Boyer [8]. Our proof is slightly shorter.

Let m_S denote the normalized Haar measure on S. For $a \in \mathbb{Z}^d$, define for any $f \in \mathcal{C}^0(X)$,

$$F_a f(x) = \int_S \chi_a(z)(U(z)f)(x) \,\mathrm{d}\,\mathrm{m}_S(z).$$

It is readily check that $F_a f \in \mathcal{H}_a$. It preserves $\mathcal{C}^{0,\alpha}(X)$ for any $\alpha \in (0, 1]$. Moreover, since Lip_X is continuous and S is compact, we have, uniformly in a,

$$\forall f \in \mathcal{C}^{0,\alpha}(X), \quad \|F_a f\|_{0,\alpha} \ll \|f\|_{0,\alpha}. \tag{8}$$

Define also the Féjer kernel: for $N \in \mathbb{N}$,

$$\mathcal{F}_N = \sum_{(a_i)\in[-N,N]^d} \left(\prod_{i=1}^{d}\left(1 - \frac{|a_i|}{N}\right)\right) F_{(a_i)}.$$

Lemma 4.3 *Let $\alpha \in (0, 1)$ and $N \in \mathbb{N}$. For any $f \in \mathcal{C}^{0,\alpha}(X)$,*

$$\|\mathcal{F}_N f - f\|_\infty \ll N^{-\alpha}\|f\|_{0,\alpha}. \tag{9}$$

Here the implied constant depend on the choice of the basis on $Z \cap \Lambda$.

Proof For $t \in \mathbb{T} = \mathbb{R}/\mathbb{Z}$, write $e(t) = e^{2\pi i t}$. While defining χ_a, we had chosen a basis of the lattice $Z \cap \Lambda$. This choice induces an isomorphism $\varphi\colon \mathbb{T}^d \to S$ so that for all $t = (t_1, \dots, t_d) \in \mathbb{T}^d$ and all $a = (a_1, \dots, a_d) \in \mathbb{Z}^d$,

$$\chi_a(\varphi(t)) = \prod_{i=1}^{d} e(a_i t_i).$$

For $N \geq 1$, denote by $K_N : \mathbb{T} \to \mathbb{R}$ the N-th Féjer kernel on the circle, i.e.,

$$\forall t \in \mathbb{T},\ K_N(s) = \sum_{a=-N+1}^{N-1} \left(1 - \frac{|a|}{N}\right) e(at) = \frac{1}{N}\left(\frac{\sin(N\pi t)}{\sin(\pi t)}\right)^2.$$

Let $f \in \mathcal{C}^{0,\alpha}(X)$. It follows from the definition that for all $x \in X$,

$$\mathcal{F}_N f(x) = \int_{\mathbb{T}^d} \left(\prod_{i=1}^{d} K_N(t_i)\right) f(\varphi(t_1, \dots, t_d)^{-1}x)\, \mathrm{d}t_1 \cdots \mathrm{d}t_d.$$

We fix a Riemannian distance $d_{\mathbb{T}^d}$ on $\mathbb{T}^d$. Since φ is smooth and both $\mathbb{T}^d$ and X are compact,

$$\forall t \in \mathbb{T}^d,\ \forall x \in X, \quad d(\varphi(t)^{-1}x, x) \ll d_{\mathbb{T}^d}(t, 0)$$

where the implied constant depends on the choice of $d_{\mathbb{T}^d}$. It follows that

$$\begin{aligned}|f(x) - \mathcal{F}_N f(x)| &= \int_{\mathbb{T}^d} \left(\prod_{i=1}^{d} K_N(t_i)\right) |f(\varphi(t_1, \ldots, t_d)^{-1}x) - f(x)| \, \mathrm{d}t_1 \cdots \mathrm{d}t_d \\ &\leq \int_{\mathbb{T}^d} \prod_{i=1}^{d} K_N(t_i) \|f\|_{0,\alpha} d(\varphi(t_1, \ldots, t_d)^{-1}x, x)^{\alpha} \, \mathrm{d}t_1 \cdots \mathrm{d}t_d \\ &\ll \|f\|_{0,\alpha} \int_{\mathbb{T}^d} \left(\prod_{i=1}^{d} K_N(t_i)\right) d_{\mathbb{T}^d}((t_1, \ldots, t_d), 0)^{\alpha} \, \mathrm{d}t_1 \cdots \mathrm{d}t_d\end{aligned}$$

Note that for $(t_1, \ldots, t_d) \in [-\frac{1}{2}, \frac{1}{2}]^d$, we have $d_{\mathbb{T}^d}((t_1, \ldots, t_d), 0)^{\alpha} \ll t_1^{\alpha} + \cdots + t_d^{\alpha}$. Hence

$$\int_{\mathbb{T}^d} \left(\prod_{i=1}^{d} K_N(t_i)\right) d_{\mathbb{T}^d}((t_1, \ldots, t_d), 0)^{\alpha} \, \mathrm{d}t_1 \cdots \mathrm{d}t_d \ll \int_{-\frac{1}{2}}^{\frac{1}{2}} K_N(t)|t|^{\alpha} \, \mathrm{d}t.$$

The last quantity is bounded by $N^{-\alpha}$, by [9, Lemma 1.6.4]. □

Proof of Lemma 4.2 We first prove the lemma for $\alpha \in (0, 1)$. Let C denote the implied constant in (9). Pick an integer N such that

$$\frac{t}{8C} \leq N^{-\alpha} \leq \frac{t}{4C}.$$

By Lemma 4.3, we get

$$\|\mathcal{F}_N f - f\|_{\infty} \leq \frac{t}{4} \|f\|_{0,\alpha}.$$

Combined with the assumption, this gives

$$\left| \int_X \mathcal{F}_N f \, \mathrm{d}(\eta - \mathrm{m}_X) \right| \geq \frac{t}{2} \|f\|_{0,\alpha}.$$

Then by the definition of the Féjer kernel,

$$\frac{t}{2} \|f\|_{0,\alpha} \leq \sum_{a \in [-N,N]^d} \left| \int_X F_a f \, \mathrm{d}(\eta - \mathrm{m}_X) \right|$$

Hence there exists $a \in [-N, N]^d$ such that

$$\left|\int_X F_a f \,\mathrm{d}(\eta - \mathrm{m}_X)\right| \geq \frac{t}{2(2N+1)^d}\|f\|_{0,\alpha} \gg t^{1+\frac{d}{\alpha}}\|f\|_{0,\alpha}.$$

Thus, on account of (8), $f_0 = F_a f$ satisfies the required properties.

If $\alpha = 1$, then (9) in Lemma 4.3 becomes (cf. [9, Lemma 1.6.4])

$$\|\mathcal{F}_N f - f\|_\infty \ll \frac{\log N}{N}\|f\|_{0,1}.$$

The rest of the proof is similar. □

4.3 Essential Growth Rate

Recall the definition of the quantity $\tau_Z(\mu)$ from the introduction. Consider a Borel probability measure μ on $\mathrm{Aut}(Z)$ where Z is a connected simply connected abelian Lie group. For $\kappa > 0$, let

$$\tau_Z(\mu,\kappa) = \limsup_{m\to+\infty} \frac{1}{m}\log\min\{\#A \mid A \subset \mathrm{Aut}(Z) \text{ with } \mu^{*m}(A) \geq 1 - e^{-\kappa m}\}.$$

This quantity is non-decreasing in κ. Let

$$\tau_Z(\mu) = \lim_{\kappa\to 0}\tau_Z(\mu,\kappa).$$

We define similarly $\tau_Z(\mu)$ if, more generally, μ is a measure on a group which acts measurably on Z by automorphisms.

Under an exponential moment assumption, this quantity is finite. Moreover it can be bounded in terms of the top Lyapunov exponent of μ.

Lemma 4.4 *Assume that the support of μ preserves a lattice of Z. Assume that μ has a finite exponential moment. Then*

$$\tau_Z(\mu) \leq (d^2-1)\lambda_{1,Z}(\mu)$$

where $d = \dim Z$ and $\lambda_{1,Z}(\mu)$ denote the top Lyapunov exponent of the linear random walk defined by μ on Z.

Proof Without loss of generality, we assume $Z = \mathbb{R}^d$ and that $\mathrm{Supp}(\mu)$ preserves the lattice $\mathbb{Z}^d$. By the large deviation estimate (Theorem A.1 proved in the Appendix), for any $\omega > 0$, there is some $\kappa > 0$ such that for all m sufficiently large

$$\mu^{*m}(B(0, e^{(\lambda_{1,Z}(\mu)+\omega)m})) \geq 1 - e^{-\kappa m}.$$

By taking $A = \mathrm{GL}_d(\mathbb{Z}) \cap B(0, e^{(\lambda_{1,Z}(\mu)+\omega)m})$, we get

$$\tau_Z(\mu) \leq \tau_Z(\mu, \kappa) \leq (d^2 - 1)(\lambda_{1,Z}(\mu) + \omega).$$

We obtain the desired inequality by letting $\omega \to 0$. □

4.4 The Cauchy-Schwarz Argument

The heart of the proof of Proposition 4.1 is a use of the Cauchy-Schwarz inequality. Let

$$C_\beta = \int_{\mathrm{Aff}(X)} \mathrm{Lip}_X(g)^\beta \, \mathrm{d}\mu(g).$$

Lemma 4.5 *Assume that μ has a finite β-exponential moment (i.e., that $C_\beta < \infty$) and that*

$$\tau_Z(\mu) < 2\sigma_{X,Y}(\mu).$$

Then for every $0 < \alpha \leq \min\{1, \beta\}$, there exists a constant m_0 depending on μ and C depending on $\theta_*\mu$ and $2\sigma_{X,Y}(\mu) - \tau_Z(\mu)$, such that the following holds. Let $t \in (0, 1/2)$ and $f \in \mathcal{C}^{0,\alpha}(X) \cap \mathcal{H}_{a_0}$ with $a_0 \in \mathbb{Z}^d \setminus \{0\}$. Let η be a Borel probability measure on X. If

$$\left| \int_X f \, \mathrm{d}\mu^{*m} * \eta \right| \geq t \|f\|_{0,\alpha}, \tag{10}$$

for some $m \geq \max(C \log \frac{1}{t}, m_0)$, then there exists $f_1 \in \mathcal{C}^{0,\alpha}(X) \cap \mathcal{H}_0$ such that

$$\left| \int_X f_1 \, \mathrm{d}\eta - \int_X f_1 \, \mathrm{d}\,\mathrm{m}_X \right| \geq (2C_\beta)^{-2m} \, t^2 \, \|f_1\|_{0,\alpha}.$$

Proof Without loss of generality, we may assume $\|f\|_{0,\alpha} = 1$.

We are going the partition $\Gamma \ltimes N$ according to the action on $\mathbb{Z}^d$. For $a \in \mathbb{Z}^d$, define

$$P_a = \{ g \in \Gamma \ltimes N \mid \theta(g)^{-1} \cdot a_0 = a \}.$$

For $m \geq 1$ and $a \in \mathbb{Z}^d$, define $\mu_a^{(m)}$ to be renormalized restriction of μ^{*m} to P_a so that we have

$$\mu^{*m} = \sum_{a \in \mathbb{Z}^d} \mu^{*m}(P_a) \mu_a^{(m)}.$$

Define also

$$f_a^{(m)} = U^*(\mu_a^{(m)})f$$

so that

$$U^*(\mu)^m f = \sum_{a\in\mathbb{Z}^d} \mu^{*m}(P_a) f_a^{(m)}. \tag{11}$$

From the definition of P_a, we know that $f_a^{(m)} \in \mathcal{H}_a$. Hence the sum in (11) is an orthogonal one. In particular,

$$\|U^*(\mu)^m f\|_2^2 = \sum_{a\in\mathbb{Z}^d} \mu^{*m}(P_a)^2 \| f_a^{(m)}\|_2^2. \tag{12}$$

From $f_a^{(m)} \in \mathcal{H}_a$, follow that $\left|f_a^{(m)}\right|^2 \in \mathcal{H}_0$. The functions $\left|f_a^{(m)}\right|^2$ are going to be our candidates for f_1.

The core of the argument is the following applications of the Cauchy-Schwarz inequality. From (10) and (11), we get

$$t \le \left|\int_X U^*(\mu)^m f \,\mathrm{d}\eta\right| \le \sum_{a\in\mathbb{Z}^d} \mu^{*m}(P_a) \left|\int_X f_a^{(m)} \,\mathrm{d}\eta\right|.$$

By the Cauchy-Schwarz inequality for the sum,

$$t^2 \le \sum_{a\in\mathbb{Z}^d} \mu^{*m}(P_a) \left|\int_X f_a^{(m)} \,\mathrm{d}\eta\right|^2.$$

By the Cauchy-Schwarz inequality for the integral,

$$t^2 \le \sum_{a\in\mathbb{Z}^d} \mu^{*m}(P_a) \int_X \left|f_a^{(m)}\right|^2 \mathrm{d}\eta. \tag{13}$$

We want to compare the right-hand side of the above equation (where the integration is over the unknown measure η) to the following analogous expression involving $\|\cdot\|_2^2$, i.e., when the integration is with respect to the Haar measure m_X:

$$\sum_{a\in\mathbb{Z}^d} \mu^{*m}(P_a) \| f_a^{(m)}\|_2^2 = \sum_{a\in\mathbb{Z}^d} \mu^{*m}(P_a) \int_X \left|f_a^{(m)}\right|^2 \mathrm{d}\,\mathrm{m}_X\,.$$

But first, we need throw away the a's for which the $\| f_a^{(m)}\|_{0,\alpha}$ is too large. From the exponential moment assumption and (5),

$$\sum_{a\in\mathbb{Z}^d}\mu^{*m}(P_a)\int_{P_a}\mathrm{Lip}_X(g)^\beta\,\mathrm{d}\mu_a^{(m)}(g)=\int_{\mathrm{Aff}(X)}\mathrm{Lip}_X(g)^\beta\,\mathrm{d}\mu^{*m}(g)\le C_\beta^m.$$

By the Markov inequality, we have for any $\kappa>0$,

$$\mu^{*m}*\delta_{a_0}(B_\kappa)\le e^{-\kappa m}$$

where

$$B_\kappa=\Big\{a\in\mathbb{Z}^d\mid\int_{P_a}\mathrm{Lip}_X(g)^\beta\,\mathrm{d}\mu_a^{(m)}(g)>C_\beta^m e^{\kappa m}\Big\}.$$

For $a\in\mathbb{Z}^d\setminus B_\kappa$, we have, since $\alpha\le\beta$,

$$\begin{aligned}\big\|f_a^{(m)}\big\|_{0,\alpha}&\le\int_{P_a}\|U^*(g)f\|_{0,\alpha}\,\mathrm{d}\mu_a^{(m)}(g)\\&\le\int_{P_a}\mathrm{Lip}_X(g)^\alpha\|f\|_{0,\alpha}\,\mathrm{d}\mu_a^{(m)}(g)\\&\le C_\beta^m e^{\kappa m}.\end{aligned}\tag{14}$$

Next, we need to exploit the assumption $\tau_Z(\mu)<2\sigma_{X,Y}(\mu)$. Choose $\tau>\tau_Z(\mu)$ and $\sigma<\sigma_{X,Y}(\tau)$ such that $\sigma-\tau/2=\frac{2\sigma_{X,Y}(\mu)-\tau_Z(\mu)}{4}>0$ and moreover,

(i) For $m\ge m_1=m_1(\mu)$, $\|U_{X,Y}(\mu)^m\|\le e^{-\sigma m}$.
(ii) There exists $\kappa=\kappa(\theta_*\mu)\in(0,\frac{1}{2})$ such that for $m\ge m_2=m_2(\theta_*\mu)$, there exists $A\subset\mathbb{Z}^d$ satisfying $\#A\le e^{\tau m}$ and $\mu^{*m}*\delta_{a_0}(A)\ge1-e^{-\kappa m}$.

Note that κ only depends on $\theta_*\mu$ because we are letting $\langle\mathrm{Supp}(\mu)\rangle$ act on Z via $\langle\mathrm{Supp}(\mu)\rangle\overset{\theta}{\to}\Gamma\to\mathrm{Aut}(Z)$. By replacing A by $A\setminus B_\kappa$, we may assume without loss of generality that $A\subset\mathbb{Z}^d\setminus B_\kappa$.

Using the fact that $\|U^*(\mu_a^{(m)})\|\le1$ (hence $\big\|f_a^{(m)}\big\|_2\le\|f\|_2\le1$), the Cauchy-Schwarz inequality, and (12) we obtain

$$\begin{aligned}\sum_{a\in A}\mu^{*m}(P_a)\big\|f_a^{(m)}\big\|_2^2&\le\sum_{a\in A}\mu^{*m}(P_a)\big\|f_a^{(m)}\big\|_2\\&\le\sqrt{\#A}\sqrt{\sum_{a\in A}\mu^{*m}(P_a)^2\big\|f_a^{(m)}\big\|_2^2}\\&=\sqrt{\#A}\big\|U^*(\mu)^m f\big\|_2\\&\le\sqrt{\#A}\big\|U_{X,Y}(\mu)^m\big\|\\&\le e^{-(\sigma-\tau/2)m}.\end{aligned}$$

Now remember (13). Bounding $\|f_a^{(m)}\|_\infty \leq \|f\|_\infty \leq 1$ for $a \in \mathbb{Z}^d \setminus A$, we obtain

$$t^2 \leq e^{-\kappa m} + \sum_{a \in A} \mu^{*m}(P_a) \int_X \left|f_a^{(m)}\right|^2 \mathrm{d}\eta,$$

which we rewrite as

$$t^2 \leq e^{-\kappa m} + \sum_{a \in A} \mu^{*m}(P_a) \| f_a^{(m)} \|_2^2 + \sum_{a \in A} \mu^{*m}(P_a) \int_X \left|f_a^{(m)}\right|^2 \mathrm{d}(\eta - \mathrm{m}_X).$$

Then it follows from the above that

$$t^2 \leq e^{-\kappa m} + e^{-(\sigma - \tau/2)m} + \sum_{a \in A} \mu^{*m}(P_a) \int_X \left|f_a^{(m)}\right|^2 \mathrm{d}(\eta - \mathrm{m}_X),$$

Now if

$$m \geq 2 \max\left\{\frac{1}{\kappa}, \frac{1}{\sigma - \tau/2}\right\} \log \frac{2}{t},$$

then

$$\frac{t^2}{2} \leq \sum_{a \in A} \mu^{*m}(P_a) \int_X \left|f_a^{(m)}\right|^2 \mathrm{d}(\eta - \mathrm{m}_X).$$

Hence there exists $a \in A$ such that

$$\int_X \left|f_a^{(m)}\right|^2 \mathrm{d}(\eta - \mathrm{m}_X) \geq \frac{t^2}{2}.$$

Moreover, since $A \subset \mathbb{Z}^d \setminus B_\kappa$, we have by (6) and (14),

$$\left\| \left|f_a^{(m)}\right|^2 \right\|_{0,\alpha} \leq \left\| f_a^{(m)} \right\|_{0,\alpha}^2 \leq C_\beta^{2m} e^{2\kappa m} \leq (2C_\beta)^{2m}.$$

Thus, $f_1 = \left|f_a^{(m)}\right|^2$ satisfies the required properties, proving the lemma with $m_0 = \max\{m_1, m_2\}$ and $C = 4 \max\left\{\frac{1}{\kappa}, \frac{4}{2\sigma_{X,Y}(\mu) - \tau_Z(\mu)}\right\}$. □

4.5 Proof of the Key Proposition

We need one more lemma before we prove Proposition 4.1.

Lemma 4.6 *Assume that μ has a finite β-exponential moment. For every $0 < \alpha \leq \min\{1, \beta\}$, there exists a constant $C \geq 1$ such that the following holds for any parameter $t \in (0, 1/2)$ and any $m \in \mathbb{N}$ sufficiently large. If there exists $f_0 \in C^{0,\alpha} \cap \mathcal{H}_0$ satisfying*

$$\left| \int_X f_0 \, \mathrm{d}\mu^{*m} * \eta - \int_X f_0 \, \mathrm{d}\, \mathrm{m}_X \right| \geq t \| f_0 \|_{0,\alpha},$$

then there exists $f_1 \in C^{0,\alpha} \cap \mathcal{H}_0$ such that

$$\left| \int_X f_1 \, \mathrm{d}\eta - \int_X f_1 \, \mathrm{d}\, \mathrm{m}_X \right| \geq e^{-Cm} t^C \| f_1 \|_{0,\alpha}.$$

Proof Without loss of generality, assume $\| f_0 \|_{0,\alpha} = 1$. By the moment assumption, there is $C_\beta \geq 1$ such that for any $m \in \mathbb{N}$, $\int \mathrm{Lip}_X(g)^\beta \, \mathrm{d}\mu^{*m}(g) \leq C_\beta^m$. Set

$$E = \{ g \in \mathrm{Aff}(X) \mid \mathrm{Lip}_X(g)^\beta > 4C_\beta^m t^{-1} \}$$

so that we have

$$\mu^{*m}(E) \leq \frac{t}{4}$$

by the Markov inequality . Thus for any $\alpha \in (0, \beta]$,

$$\forall g \in \mathrm{Aff}(X) \setminus E, \quad \| U^*(g) f_0 \|_{0,\alpha} \leq \mathrm{Lip}_X(g)^\alpha \| f_0 \|_{0,\alpha} \leq 4C_\beta^m t^{-1}.$$

By the assumption on f_0,

$$\begin{aligned} t &\leq \int_{\mathrm{Aff}(X)} \left| \int_X U^*(g) f_0 \, \mathrm{d}(\eta - \mathrm{m}_X) \right| \, \mathrm{d}\mu^{*m}(g) \\ &\leq 2\mu^{*m}(E) + \int_{\mathrm{Aff}(X) \setminus E} \left| \int_X U^*(g) f_0 \, \mathrm{d}(\eta - \mathrm{m}_X) \right| \, \mathrm{d}\mu^{*m}(g). \end{aligned}$$

Hence

$$\int_{\mathrm{Aff}(X) \setminus E} \left| \int_X U^*(g) f_0 \, \mathrm{d}(\eta - \mathrm{m}_X) \right| \, \mathrm{d}\mu^{*m}(g) \geq \frac{t}{2}.$$

Hence there exists $g \in \mathrm{Aff}(X) \setminus E$ such that $f_1 = U^*(g) f_0$ satisfies

$$\left| \int_X f_1 \, \mathrm{d}(\eta - \mathrm{m}_X) \right| \geq \frac{t}{2}.$$

Moreover, since $g \notin E$,

$$\|f_1\|_{0,\alpha} \leq 4C_\beta^m t^{-1},$$

showing the required property for f_1. □

Proof of Proposition 4.1 Let $t \in (0, 1)$ be such that there exists $f \in \mathcal{C}^{0,\alpha}(X)$ such that

$$\left|\int_X f \, \mathrm{d}\mu^{*m} * \eta - \int_X f \, \mathrm{d}\,\mathrm{m}_X\right| \geq t\|f\|_{0,\alpha}.$$

By Lemma 4.2 there is $a_0 \in \mathbb{Z}^d$ and $f_0 \in \mathcal{C}^{0,\alpha}(X) \cap \mathcal{H}_{a_0}$ such that

$$\left|\int_X f_0 \, \mathrm{d}\mu^{*m} * \eta - \int_X f_0 \, \mathrm{d}\,\mathrm{m}_X\right| \geq t^{O(1)}\|f_0\|_{0,\alpha}.$$

Using either Lemma 4.6 in the case where $a = 0$ or Lemma 4.5 otherwise (note that $a_0 \neq 0$ implies $\int_X f_0 \, \mathrm{d}\,\mathrm{m}_X = 0$), we obtain some $f_1 \in \mathcal{C}^{0,\alpha}(X) \cap \mathcal{H}_0$ such that

$$\left|\int_X f_1 \, \mathrm{d}\eta - \int_X f_1 \, \mathrm{d}\,\mathrm{m}_X\right| \geq e^{-O(m)} t^{O(1)}\|f_1\|_{0,\alpha}. \tag{15}$$

Letting $\varphi \in \mathcal{C}^{0,\alpha}(Y)$ be such that $f_1 = \varphi \circ \pi$, we have $\int_X f_1 \, \mathrm{d}\eta = \int_Y \varphi \, \mathrm{d}\pi_*\eta$, $\int_X f_1 \, \mathrm{d}\,\mathrm{m}_X = \int_Y \varphi \, \mathrm{d}\,\mathrm{m}_Y$, and $\|\varphi\|_{0,\alpha} \ll \|f_1\|_{0,\alpha}$. The last implied constant depends only on the choice of Riemannian metrics on X and on Y. Therefore,

$$\left|\int_Y \varphi \, \mathrm{d}\pi_*\eta - \int_Y \varphi \, \mathrm{d}\,\mathrm{m}_Y\right| \gg e^{-O(m)} t^{O(1)}\|\varphi\|_{0,\alpha},$$

finishing the proof of the proposition. □

5 Proof of the Main Theorems

We are ready to prove the main theorem of this paper.

Proof of Proposition 4.1 We use the same notation μ, β, X, N, Λ, Z, Y, Γ, θ in Proposition 4.1 as in Theorem 1.1.

Assume that the μ-induced walk on Y satisfies (C_Y, λ, α)-quantitative equidistribution for parameters $C_Y > 0$, $\lambda > 0$ and $0 < \alpha \leq \min(1, \beta)$. Let $\lambda' \in (0, \lambda)$. We want to show that the μ-induced walk on X satisfies (C_X, λ', α)-quantitative equidistribution for a large constant C_X. Assume that for some $t \in (0, \frac{1}{2})$, $m \geq C_X \log \frac{1}{t}$ holds that $\mathcal{W}_\alpha(\mu^{*m} * \delta_x, \mathrm{m}_X) > t$.

Denote by $\pi : X \to Y$ the natural projection. Now we can apply Proposition 4.1, whose constant we denote by C_π, on $\eta = \mu^{*(m-m')} * \delta_x$ with m' random walk steps. Choose m' to be such that

$$C_\pi \log \frac{1}{t} < m' < 2C_\pi \log \frac{1}{t}.$$

By the proposition,

$$\mathcal{W}_\alpha(\mu^{*(m-m')} * \delta_{\pi(x)}, \mathrm{m}_Y) = \mathcal{W}_\alpha(\pi_* \eta, \mathrm{m}_Y) \geq e^{-C_\pi m'} > t^{2C_\pi^2},$$

If C_X is large enough, then we can guarantee that $m - m' \geq C_Y \log(t^{-2C_\pi^2})$, so that the premise of the (C_Y, λ, α)-quantitative equidistribution of the random walk induced on Y applies.

For simplicity, we will assume for the remainder of the proof that μ is supported on $\mathrm{Aut}(X)$ and leave the case that it is supported on $\mathrm{Aff}(X)$ to the reader. The two proofs are almost identical.

The quantitative equidistribution on Y tells us that there exists $y' \in Y$ with $d(\pi(x), y') < e^{-\lambda(m-m')}$ such that the projection of $\Gamma y'$ to the maximal torus factor T_Y of Y is contained in a proper closed Γ-invariant subgroup L of T_Y of height $\leq t^{-2C_Y C_\pi^2}$. Note that if $C_X > \frac{2C_\pi \lambda}{\lambda - \lambda'}$, then $e^{-\lambda(m-m')} < e^{-\lambda' m}$. By choosing $x' \in X$ to be the point closest to x in $\pi^{-1}(y')$, we get $d(x, x') = d(\pi(x), y') < e^{-\lambda' m}$.

Let T_X denote the maximal torus factor of X. Then T_Y is a factor of T_X. Moreover the following the diagram of Γ-equivariant maps commutes.

$$\begin{tikzcd} X \arrow[r, "\pi"] \arrow[d] & Y \arrow[d] \\ T_X \arrow[r, "\pi'"] & T_Y \end{tikzcd}$$

Thus, the projection of $\Gamma x'$ to T_X is contained in $\pi'^{-1}(L)$, which is a proper closed Γ-invariant subgroup of T_X of height $\leq O(t^{-2C_Y C_\pi^2})$ by the following observation.

Lemma 5.1 *Let T' be a factor torus of a torus T and let $\pi' : T \to T'$ be the factor map. There exists $C' > 1$ such that if L is a proper closed subgroup of T' of height $\leq h$, then $\pi'^{-1}(L)$ is a proper closed subgroup of T of height $\leq C'h$.*

Proof A generating set of the dual of L can be mapped by the dual of π' to a generating set of the dual of $\pi'^{-1}(L)$. The dual of π' changes the norm of the vectors by at most a finite factor C', the operator norm of this linear transformation.

By taking a new C_X that is large enough, this give us (C_X, λ', α)-quantitative equidistribution of the random walk on X. □

Theorem 1.3 follows immediately.

Proof of Theorem 1.3 Remark that for each $k = 1, \dots, l-1$, because $X \to X_k$ is a smooth map between compact Riemannian manifolds, the condition that μ has a finite exponential moment implies that the image measure in $\mathrm{Aff}(X_k)$ also has finite exponential moment. It suffices then to use Theorem B for the random walk on the torus X_{l-1} and apply Theorem 1.1 repeatedly $l-1$ times. □

Corollary 1.4 follows from the following lemma.

Lemma 5.2 *Let X be a nilmanifold and μ a probability measure on* $\mathrm{Aff}(X)$. *Let T denote the maximal torus factor of X. Let H denote the subgroup generated by* $\mathrm{Supp}(\mu)$. *If the μ-induced random walk on X satisfies a (C, λ, α)-quantitative equidistribution for some $C > 0$, $\lambda > 0$, and $\alpha \in (0, 1]$, then for any $x \in X$*

(i) Either $\mu^{*m} * \delta_x$ converges to m_X in the weak-$*$ topology
(ii) Or the projection of Hx to T is contained in a proper closed H-invariant subset

Proof Let $\pi : X \to T$ be the projection from X to its maximal torus factor. Assume that $\mu^{*m} * \delta_x$ does not converge to m_X in the weak-$*$ topology. The space of α-Hölder functions $C^{0,\alpha}(X)$ is dense in the space of continuous functions. It follows that there is $t > 0$ such that

$$\mathcal{W}_\alpha(\mu^{*m} * \delta_x, \mathrm{m}_X) > t$$

for an unbounded sequence of m.

From the quantitative equidistribution, we get

(i) A sequence (x_k) of points in X
(ii) A sequence (H_k) of subgroups of $\mathrm{Aff}(X)$
(iii) A sequence (L_k) of proper closed subgroup of T of height $\leq t^{-C}$ and invariant under $\theta(H_k) = \theta(H)$

such that

(i) $\lim_{k\to+\infty} x_k = x$
(ii) $\lim_{k\to+\infty} \sup_{g\in\mathrm{Supp}(\mu)} d(g, H_k) = 0$
(iii) $\pi(H_k x_k) - \pi(x_k) \subset L_k$ for all k

In T, there are only finitely many closed subgroup of height $\leq t^{-C}$. Therefore, after extracting a subsequence, we may assume that $L_k =: L$ are all equal. Letting k go to $+\infty$, we find

$$\forall g \in \mathrm{Supp}(\mu), \quad \pi(gx) - \pi(x) \in L.$$

This is enough to conclude that $\pi(x) + L \subset T$ is H-invariant and $\pi(Hx) \subset \pi(x) + L$. □

Acknowledgments The proof in Appendix B grew out of a discussion together with Nicolas de Saxcé. We are grateful to him for sharing his ideas.

This paper is dedicated to the memory of Jean Bourgain, a great man and a profound mathematician, whose deep work laid the framework for much that is done in this paper. In particular, much of what the third named author knows about arithmetic combinatorics he learned from Jean Bourgain. While working on this paper, we have been acutely aware of him being no longer with us—no doubt if we could have discussed these questions with him we could have gone much further.

Appendix A: A Large Deviation Estimate

Let μ be Borel probability measure on $\mathrm{GL}_d(\mathbb{R})$, $d \geq 2$. Consider the random walk in the linear group defined by μ. Recall that μ is said to have a finite exponential moment if there is $\beta > 0$ such that

$$\int_{\mathrm{GL}_d(\mathbb{R})} \max\bigl\{\|g\|, \|g^{-1}\|\bigr\}^{\beta} \,\mathrm{d}\mu(g) < +\infty. \tag{16}$$

Recall also that the top Lyapunov exponent of μ is defined by

$$\lambda_1(\mu) = \lim_{m\to+\infty} \frac{1}{m} \int_{\mathrm{GL}_d(\mathbb{R})} \log\|g\| \,\mathrm{d}\mu^{*m}(g)$$

Theorem A.1 *Let μ be a Borel probability measure on $\mathrm{GL}_d(\mathbb{R})$. Assume μ has a finite exponential moment. For any $\omega > 0$ there is $\kappa > 0$ such that for all m large enough.*

$$\mu^{*m}\{\, g \in \mathrm{GL}_d(\mathbb{R}) \mid \log\|g\| > m(\lambda_1(\mu) + \omega)\,\} \leq e^{-\kappa m}.$$

Proof Let $g_1, g_2, \ldots$ be independent random variables distributed according to μ.

Given $\omega > 0$, let $l \geq 1$ be such that

$$\mathbb{E}\bigl[\log\|g_1 \cdots g_l\|\bigr] < l(\lambda_1(\mu) + \omega/3).$$

Observe that $\bigl(\log\|g_{kl-l+1} \cdots g_{kl}\|\bigr)_{k\geq 1}$ is a sequence of i.i.d. real-valued random variables having a finite exponential moment. Thus, by Crámer's theorem, there exists $\tau > 0$ such that for k large enough,

$$\mathbb{P}\bigl[\log\|g_1 \cdots g_l\| + \cdots + \log\|g_{kl-l+1} \cdots g_{kl}\| > kl(\lambda_1(\mu) + 2\omega/3)\bigr] \leq e^{-\tau k}.$$

The norm is submultiplicative, hence for k large enough,

$$\mathbb{P}\bigl[\log\|g_1 \cdots g_{kl}\| > kl(\lambda_1(\mu) + 2\omega/3)\bigr] \leq e^{-\tau k}.$$

For any m, write $m = kl + j$ with $0 \leq j < k$. Using submultiplicativity again, we see that if $\log\|g_1 \cdots g_m\| > m(\lambda_1(\mu)+\omega)$, then either $\log\|g_1 \cdots g_{kl}\| > kl(\lambda_1(\mu)+2\omega/3)$ or there is $1 \leq i \leq j$ such that $\log\|g_{kl+i}\| > \frac{\omega}{3l}m$. Thus,

$$\mathbb{P}\big[\log\|g_1 \cdots g_m\| > m(\lambda_1(\mu) + \omega)\big] \leq e^{-\tau k} + l\mathbb{P}\big[\log\|g_1\| > \frac{\omega}{3l}m\big].$$

Finally, since μ has a finite exponential moment, there is some $\beta > 0$ such that $\mathbb{E}\big[\|g_1\|^{\beta}\big]$ is finite. Hence by Markov's inequality,

$$\mathbb{P}\big[\log\|g_1\| > \frac{\omega}{3l}m\big] = \mathbb{P}\big[\|g_1\|^{\beta} > e^{\frac{\beta\omega}{3l}m}\big] \leq e^{-\frac{\beta\omega}{3l}m}\mathbb{E}\big[\|g_1\|^{\beta}\big].$$

Put together, we find

$$\mathbb{P}\big[\log\|g_1 \cdots g_m\| > m(\lambda_1(\mu) + \omega)\big] \leq e^{-\kappa m}$$

for $\kappa = \frac{1}{2}\min\{\frac{\tau}{l}, \frac{\beta\omega}{3l}\}$ and m large enough. □

Appendix B: The Case of a Torus

Here we explain how to remove the Zariski connectedness assumption in the main theorem of [15]. Namely, the goal is the following.

Theorem B.1 *Let $X = \mathbb{R}^d/\mathbb{Z}^d$. Let μ be a probability measure on $\mathrm{Aut}(X) = \mathrm{GL}_d(\mathbb{Z})$ having a finite exponential moment. Let Γ denote the subgroup generated by the support of μ. Assume that the action of Γ on $\mathbb{R}^d$ is strongly irreducible. Then given any $\lambda \in (0, \lambda_{1,\mathbb{R}^d}(\mu))$, there exists a constant $C = C(\mu, \lambda) \geq 1$ such that the following holds. If $x \in X$ satisfies*

$$|\widehat{\mu^{*n} * \delta_x}(a)| > t \quad \textit{and} \quad n \geq C \log \frac{\|a\|}{t}$$

for some $a \in \mathbb{Z}^d \setminus \{0\}$ and $t \in (0, \frac{1}{2})$, then there exists a rational point $x' \in X$ of denominator at most $(\frac{\|a\|}{t})^C$ such that $d(x, x') \leq e^{-\lambda m}$.

The corresponding statement for affine random walks is the following. Recall that $\theta\colon \mathrm{Aff}(X) \to \mathrm{Aut}(X)$ denote the linear part.

Theorem B.2 *Let $X = \mathbb{R}^d/\mathbb{Z}^d$. Let μ be a finitely supported probability measure on $\mathrm{Aff}(X) = \mathrm{GL}_d(\mathbb{Z}) \ltimes \mathbb{R}^d$. Let Γ denote the subgroup generated by the support of $\theta_*\mu$. Assume that the action of Γ on $\mathbb{R}^d$ is strongly irreducible. Then given any $\lambda \in (0, \lambda_{1,\mathbb{R}^d}(\mu))$, there exists a constant $C = C(\theta_*\mu, \lambda) \geq 1$ such that the following holds. If $x \in X$ satisfies*

$$|\widehat{\mu^{*n} * \delta_x}(a)| > t \quad \textit{and} \quad n \geq C \log \frac{\|a\|}{t}$$

for some $a \in \mathbb{Z}^d \setminus \{0\}$ *and* $t \in (0, \frac{1}{2})$*, then there exists a point* $x' \in X$ *and a finite set* $F \subset \mathrm{Aff}(X)$ *such that* $d(x, x') \leq e^{-\lambda m}$*,* $d_{\mathrm{H}}(\mathrm{Supp}(\mu), F) \leq e^{-\lambda m}$*, and moreover, denoting by* H *the subgroup generated by* F*, the orbit* Hx' *is finite of cardinality at most* $(\frac{\|a\|}{t})^C$*.*

In view of [8, Lemma 4.5] or alternatively Lemma 4.2, Theorem B follows.

The key point is a Fourier decay estimate for $(\theta_*\mu)^{*n}$, stated as Theorem B.7 below, replacing [15, Theorem 3.20]. To establish this Fourier decay property, we first need a Fourier decay estimate for multiplicative convolutions of measures having nice non-concentration properties, Theorem B.3. Then in Sect. B.2, using return times and the special case of Zariski-connected groups, we obtain a decomposition of $(\theta_*\mu)^{*n}$ as a sum of multiplicative convolutions of measures having the required non-concentration properties. Once Theorem B.7 is established, the rest of the proof of Theorem B.1 is identical to the corresponding part in [15] and that of Theorem B.2 to the corresponding part in [16].

B.1 Multiplicative Convolutions in Simple Algebras

First, we need a slight improvement of [15, Theorem 2.1] by allowing the measures we convolve to be different.

Let E be a normed simple algebra over $\mathbb{R}$ of finite dimension. For $x \in E$, denote by $\det(x)$ the determinant of the linear endomorphism $E \to E$, $y \mapsto xy$. For $\rho > 0$, write

$$S(\rho) = \{x \in E \mid |\det(x)| \leq \rho\}.$$

Definition 4 Let $\epsilon > 0$, $\kappa > 0$, and $\tau > 0$ be parameters. We say a Borel measure η on E satisfies $\mathrm{NC}_0(\epsilon, \kappa, \tau)$ at scale $\delta > 0$ if

(i) $\eta\big(E \setminus B(0, \delta^{-\epsilon})\big) \leq \delta^\tau$
(ii) For every $x \in E$, $\eta(x + S(\delta^\epsilon)) \leq \delta^\tau$
(iii) For every $\rho \geq \delta$ and every proper affine subspace $W \subset E$, $\eta(W^{(\rho)}) \leq \delta^{-\epsilon}\rho^\kappa$, where $W^{(\rho)}$ denotes the ρ-neighborhood of W

Throughout this appendix, each occurrence of $((t))$ with $t > 0$ denotes an unspecified Borel measure of total mass at most t.

Definition 5 We say a Borel probability measure η on E satisfies $\mathrm{NC}(\epsilon, \kappa, \tau)$ at scale δ if it can be written as $\eta = \eta_0 + ((\delta^\tau))$ with η_0 satisfying $\mathrm{NC}_0(\epsilon, \kappa, \tau)$ at scale δ.

Here, NC stands for non-concentration.

Let E^* denote the linear dual of E over $\mathbb{R}$. Recall that the Fourier transform of a finite Borel measure ν on E is defined as

$$\forall \xi \in E^*, \quad \hat{\nu}(\xi) = \int_E e(\xi(x))\,\mathrm{d}\nu(x)$$

where $e(t) = e^{2\pi i t}$, for $t \in \mathbb{R}$.

Theorem B.3 (Fourier Decay of Multiplicative Convolutions in Simple Algebra) *Let E be a normed simple algebra over $\mathbb{R}$ of finite dimension. Given $\kappa > 0$, there exists $s = s(E,\kappa) \in \mathbb{N}$ and $\epsilon = \epsilon(E,\kappa) > 0$ such that for any parameter $\tau \in (0, \epsilon\kappa)$, the following holds for any scale $\delta > 0$ sufficiently small.*

If $\eta_1, \dots, \eta_s$ are Borel probability measures on E satisfying $\mathrm{NC}(\epsilon,\kappa,\tau)$ at scale δ, then for all $\xi \in E^*$ with $\delta^{-1+\epsilon} \le \|\xi\| \le \delta^{-1-\epsilon}$,

$$|(\eta_1 * \cdots * \eta_s)^{\wedge}(\xi)| \le \delta^{\epsilon\tau}.$$

The special case where $\eta_1 = \cdots = \eta_s$ are the same measure is precisely [15, Theorem 2.1]. We will deduce the general case from the special case using a trick from an article of Bourgain and Dyatlov [5].

For measures η and η' on E, we write $\eta \boxplus \eta'$ for the additive convolution between η and η'. Similarly, $\eta \boxminus \eta'$ denotes the image measure of $\eta \otimes \eta'$ under the map $(x,y) \mapsto x - y$. Finally, for integer $k \ge 1$, we write

$$\eta^{\boxplus k} = \underbrace{\eta \boxplus \cdots \boxplus \eta}_{k \text{ times}}.$$

The following two observations on the NC property are immediate.

Lemma B.4 *Let $\epsilon, \kappa, \tau, \sigma > 0$ be parameters and let $\delta > 0$.*

(i) If η is a Borel probability measures on E satisfying $\mathrm{NC}(\epsilon,\kappa,\tau)$ at scale δ, then $\eta \boxminus \eta$ satisfies $\mathrm{NC}(O(\epsilon),\kappa,\tau/2)$ at scale δ.

(ii) Convex combinations of probability measures satisfying $\mathrm{NC}(\epsilon,\kappa,\tau)$ also satisfy $\mathrm{NC}(\epsilon,\kappa,\tau)$ at scale δ.

Lemma B.5 *Let $\epsilon, \kappa, \tau, \sigma > 0$ be parameters and let $\delta > 0$. Let η and η' be Borel probability measures on E such that $\eta = \delta^{\sigma}\eta' + ((1))$. If η satisfies $\mathrm{NC}(\epsilon,\kappa,\tau)$ at scale δ, then η' satisfies $\mathrm{NC}(\epsilon+\sigma,\kappa,\tau-\sigma)$ at scale δ.*

Finally, we will need to compare Fourier transform of multiplicative convolutions with that of multiplicative convolutions of additive convolutions.

Lemma B.6 *Let ν, ν', ν'' be Borel probability measures on E, and then for any integer $k \ge 1$, the Fourier transform of $\nu * (\nu'^{\boxplus k} \boxminus \nu'^{\boxplus k}) * \nu''$ takes non-negative real values and, moreover,*

$$\forall \xi \in E^*, \quad |(\nu * \nu' * \nu'')^{\wedge}(\xi)|^{2k} \le \big(\nu * (\nu'^{\boxplus k} \boxminus \nu'^{\boxplus k}) * \nu''\big)^{\wedge}(\xi).$$

Proof By definition,

$$(\nu * \nu' * \nu'')^\wedge(\xi) = \iiint e(\xi(xyz))\,\mathrm{d}\nu(x)\,\mathrm{d}\nu'(y)\,\mathrm{d}\nu''(z).$$

By Hölder's inequality applied to the function $(x, z) \mapsto \int e(\xi(xyz))\,\mathrm{d}\nu'(y)$,

$$\begin{aligned} &|(\nu * \nu' * \nu'')^\wedge(\xi)|^{2k} \\ &\quad \leq \iint \left|\int e(\xi(xyz))\,\mathrm{d}\nu'(y)\right|^{2k} \mathrm{d}\nu(x)\,\mathrm{d}\nu''(z) \\ &\quad = \iiint e\big(\xi\big(x(y_1 + \cdots + y_k - y_{k+1} - \cdots - y_{2k})z\big)\big) \\ &\qquad \times \mathrm{d}\nu'^{\otimes 2k}(y_1, \ldots, y_{2k})\,\mathrm{d}\nu(x)\,\mathrm{d}\nu''(z) \\ &\quad = \big(\nu * (\nu'^{\boxplus k} \boxminus \nu'^{\boxplus k}) * \nu''\big)^\wedge(\xi). \end{aligned}$$

□

Proof of Theorem B.3 For $\lambda = (\lambda_1, \ldots, \lambda_s) \in \mathbb{C}^s$, define

$$\eta_\lambda = \lambda_1 \eta_1 \boxminus \eta_1 + \cdots + \lambda_s \eta_s \boxminus \eta_s.$$

Consider the function $F\colon \mathbb{C}^s \to \mathbb{C}$ defined by

$$F(\lambda) = \widehat{\eta_\lambda^{*s}}(\xi) = (\eta_\lambda * \cdots * \eta_\lambda)^\wedge(\xi).$$

For all $\lambda = (\lambda_1, \ldots, \lambda_s) \in \mathbb{R}^s$ with $\lambda_1 + \cdots + \lambda_s = 1$, by Lemma B.4, η_λ satisfy $\mathrm{NC}(\epsilon, \kappa, \tau/2)$ at scale δ. Hence by [15, Theorem 2.1], we can bound

$$|F(\lambda)| \leq \delta^{\epsilon_0 \tau}$$

for some $\epsilon_0 = \epsilon_0(E, \kappa)$.

Observe that $F(\lambda)$ is a homogeneous polynomial function of degree s. Then above implies

$$|\partial_1 \cdots \partial_s F(0, \ldots, 0)| \ll \delta^{\epsilon_0 \tau}.$$

The left-hand side is the coefficient of the monomial term $\lambda_1 \cdots \lambda_s$, which is

$$\partial_1 \cdots \partial_s F(0, \ldots, 0) = \sum_{\sigma \in \mathfrak{S}_s} \big((\eta_{\sigma(1)} \boxminus \eta_{\sigma(1)}) * \cdots * (\eta_{\sigma(s)} \boxminus \eta_{\sigma(s)})\big)^\wedge(\xi).$$

By Lemma B.6, each term of the right-hand side is non-negative real. It follows that

$$\left|\left((\eta_1 \boxminus \eta_1) * \cdots * (\eta_s \boxminus \eta_s)\right)^{\wedge}(\xi)\right| \ll \delta^{\epsilon_0 \tau}.$$

In view of Lemma B.6, this concludes the proof of the theorem. □

B.2 Fourier Decay for Linear Random Walks

From now on let μ be a probability measure on $\mathrm{Aut}(\mathbb{T}^d) = \mathrm{GL}_d(\mathbb{Z})$. Let λ_1 denote the top Lyapunov exponent of μ and let Γ denote the subgroup generated by $\mathrm{Supp}(\mu)$. We assume

(i) The measure μ has a finite exponential moment
(ii) The action of Γ on $\mathbb{R}^d$ is strongly irreducible

Let G denote the Zariski closure of Γ in GL_d and G° the identity component of G; then $\Gamma_0 = \Gamma \cap G^\circ$ is a finite index subgroup of Γ. Let E denote the subalgebra generated by $G^\circ(\mathbb{R})$. If $\gamma_1, \ldots, \gamma_J$ are a complete set of representatives for the cosets in Γ/Γ_0, then for any $\gamma \in \Gamma$, we have that $\gamma E = \gamma_j E$ for some $1 \le j \le J$.

Theorem B.7 (Fourier Decay for Random Walks in $\mathrm{GL}_d(\mathbb{Z})$**)** *Let Γ, μ, and $\gamma_1, \ldots, \gamma_J$ be as above. Then there exists $\alpha_0 = \alpha_0(\mu) > 0$ such that for every $\alpha \in (0, \alpha_0)$, there exists $c = c(\mu, \alpha) > 0$ such that for all n sufficiently large, all $1 \le j \le J$ and $\xi \in E^*$ with*

$$e^{\alpha n} \le e^{\lambda_1 n} \|\xi\| \le e^{\alpha_0 n}$$

the following estimate on Fourier coefficients of μ^{*n} holds:

$$\left|\int_{\gamma_j E} e\big(\xi(\gamma_j^{-1} g)\big)\, \mathrm{d}\mu^{*n}(g)\right| \le e^{-c_0 n}.$$

Let $(g_n)_{n \ge 1}$ be a sequence of independent random variables distributed according to μ. Consider the return times to G°,

$$\tau(1) = \inf\{n \ge 1 \mid g_n \cdots g_1 \in G^\circ\}$$

and recursively for $m \ge 2$,

$$\tau(m) = \inf\{n > \tau(m) \mid g_n \cdots g_1 \in G^\circ\}.$$

They are the return times of a Markov chain on the finite space G/G°. Thus for every $m \ge 1$, $\tau(m)$ is almost surely finite.

Let μ° denote the law of $g_{\tau(1)} \cdots g_1$, which is a probability measure on G°. It has the following properties.

Lemma B.8 ([1, Lemma 4.40]) *If μ has a finite exponential moment, then so does μ°.*

Denote $T = \mathbb{E}[\tau(1)]$. Let $\lambda_1 = \lambda_1(\mu)$ denote the top Lyapunov exponent of μ.

Lemma B.9 ([1, Lemma 4.42]) *The top Lyapunov exponent of μ° is*

$$\lambda_1(\mu^\circ) = T\lambda_1.$$

Lemma B.10 ([1, Lemma 4.42]) *Given $\omega > 0$, there is $c = c(\mu, \omega) > 0$ such that for all m sufficiently large,*

$$\mathbb{P}\big[|\tau(m) - Tm| \geq \omega m\big] \leq e^{-cm}.$$

Note that the support of μ° generates $\Gamma \cap G^\circ$, whose Zariski closure is G°. For $m \geq 1$, in view of Lemma B.9, it is appropriate to rescale $(\mu^\circ)^{*m}$ by a factor of $e^{-T\lambda_1 m}$. Put

$$\tilde{\mu}^\circ_m = (e^{-T\lambda_1 m})_*(\mu^\circ)^{*m}.$$

Under the assumptions recalled at the beginning of the paragraph, G° acts irreducibly on $\mathbb{R}^d$ and is Zariski-connected. Thus, we can apply the results in [15, Section 3] to the random walk defined by μ°. As explained in [15, Proof of Theorem 3.20], Proposition 3.1 and Proposition 3.2 of [15] imply the following.

Lemma B.11 *Write $D = \dim E$. There exists $\kappa = \kappa(\mu^\circ) > 0$ such that given any $\alpha > 0$ and $\epsilon > 0$ there exists $\tau > 0$ such that the additive convolution $(\tilde{\mu}^\circ_m)^{\boxplus D} \boxminus (\tilde{\mu}^\circ_m)^{\boxplus D}$ satisfies $\mathrm{NC}(\epsilon, \kappa, \tau)$ in E at all scales $\delta \in [e^{-m}, e^{-\alpha m}]$ for all $m \geq 1$ sufficiently large.*

For $m \geq 1$ and $l \geq 1$, we define ν_l to be the law of the variable

$$g_{\tau(m)} \cdots g_1 \quad \text{conditional to the event} \quad \tau(m) = l.$$

By this definition,

$$(\mu^\circ)^{*m} = \sum_{l \in \mathbb{N}} p_l \nu_l. \tag{17}$$

where $p_l = \mathbb{P}[\tau(m) = l]$. Here, we are hiding the dependency of ν_l and p_l on m in order to make the notations less cumbersome.

Let n, s and $l_1, \ldots, l_s$ be integers. Consider the events $\tau(jm) = l_1 + \cdots + l_j$, $j = 1, \ldots, s$. By the Markov property, we have

$$\mathbb{P}\big[\forall j = 1, \ldots, s,\ \tau(jm) = l_1 + \cdots + l_j\big] = p_{l_1} \cdots p_{l_s}$$

Now assume that $l_1 + \cdots + l_s + k = n$ with $k \geq 0$ and condition the variable $g_n \cdots g_1$ according to the events above. We obtain a decomposition

$$\mu^{*n} = \sum_{l_1+\cdots+l_s+k=n} p_{l_1} \cdots p_{l_s} \mu^{*k} * \nu_{l_s} * \cdots * \nu_{l_1} + ((\mathbb{P}[\tau(sm) > n])). \tag{18}$$

With these preparations, the proof of Theorem B.7 is not difficult.

Proof of Theorem B.7 Let $\alpha > 0$ be given. In this proof, each occurrence of c denotes a small positive constant depending on μ and α but independent of n.

Let $\kappa = \kappa(\mu^\circ) > 0$ be the constant given by Lemma B.11. Let $s = s(E, \kappa) \geq 1$ and $\epsilon = \epsilon(E, \kappa) > 0$ be the constants given by Theorem B.3. By Lemma B.11, there exists $\tau > 0$ such that $(\tilde{\mu}_m^\circ)^{\boxplus D} \boxminus (\tilde{\mu}_m^\circ)^{\boxplus D}$ satisfies $\mathrm{NC}(\epsilon/2, \kappa, 2\tau)$ in E at all scales $\delta \in [e^{-m}, e^{-\alpha m/2}]$, provided that $m \geq 1$ is large enough. Without loss of generality, we may assume $\tau < \kappa\epsilon$, $\tau < \epsilon/2$.

Let $\omega = \omega(\mu, \alpha)$ to be a constant whose value is to be determined later. For $n \geq 1$, and choose $m = \left\lfloor \frac{n}{(T+\omega)s} \right\rfloor$. Everything below is true for n sufficiently large (larger than some n_0 depending on μ and α). By Lemma B.10, we have

$$\mathbb{P}[\tau(sm) > n] \leq e^{-cn}$$

and

$$\mathbb{P}[\tau(sm) < n - 3\omega n] \leq e^{-cn}.$$

Put

$$\mathcal{L} = \{\, l \in \mathbb{N} \mid p_l \geq e^{-\frac{\alpha\tau}{4D} m} \,\}.$$

We can bound

$$\sum_{(l_1,\ldots,l_s) \notin \mathcal{L}^s} p_{l_1} \cdots p_{l_s} \leq sne^{-\frac{\alpha\tau}{4D} m} \leq e^{-cn}.$$

Thus, (18) becomes

$$\mu^{*n} = \sum_{\substack{l_1,\ldots,l_s \in \mathcal{L}, k \leq 3\omega n \\ l_1+\cdots+l_s+k=n}} p_{l_1} \cdots p_{l_s} \mu^{*k} * \nu_{l_s} * \cdots * \nu_{l_1} + ((e^{-cn})).$$

Let γ be one of $\gamma_1, \ldots, \gamma_J$. To finish the proof of the theorem, it suffices to establish an upper bound for the quantity

$$I_{l_1,\ldots,l_s,k}(\xi) := \int_{\gamma E} e\big(\xi(\gamma^{-1} g)\big) \, \mathrm{d}\big(\mu^{*k} * \nu_{l_s} * \cdots * \nu_{l_1}\big)(g)$$

uniformly for all $l_1, \dots, l_s \in \mathcal{L}$, $k \le 3\omega n$ with $l_1 + \cdots + l_s + k = n$.

Indeed, developing $(\tilde{\mu}_m^\circ)^{\boxplus D} \boxminus (\tilde{\mu}_m^\circ)^{\boxplus D}$ using (17), we see

$$(\tilde{\mu}_m^\circ)^{\boxplus D} \boxminus (\tilde{\mu}_m^\circ)^{\boxplus D} = p_l^{2D}(e^{-T\lambda_1 m})_*\big(\nu_l^{\boxplus D} \boxminus \nu_l^{\boxplus D}\big) + ((1)).$$

Since $p_l^{2D} \ge e^{-\alpha\tau m/2} \ge \delta^\tau$ for $l \in \mathcal{L}$, it follows from Lemma B.5 that $(e^{-T\lambda_1 m})_*\big(\nu_l^{\boxplus D} \boxminus \nu_l^{\boxplus D}\big)$ satisfies $\mathrm{NC}(\epsilon, \kappa, \tau)$ at all scales $\delta \in [e^{-m}, e^{-\alpha m/2}]$, provided that $m \ge 1$ is large enough.

Theorem B.3 tells us that for $(l_1, \dots, l_s) \in \mathcal{L}^s$, for all $\xi \in E^*$ with $e^{\alpha m/2} \le e^{T\lambda_1 sm}\|\xi\| \le e^m$,

$$\Big|\Big(\big(\nu_{l_s}^{\boxplus D} \boxminus \nu_{l_s}^{\boxplus D}\big) * \cdots * \big(\nu_{l_1}^{\boxplus D} \boxminus \nu_{l_1}^{\boxplus D}\big)\Big)^\wedge(\xi)\Big| \le e^{-\alpha\epsilon\tau m/2}.$$

Using Lemma B.6 repeatedly s times, we obtain, for all $\xi \in E^*$ in the same range,

$$\big|\big(\nu_{l_s} * \cdots * \nu_{l_1}\big)^\wedge(\xi)\big| \le e^{-\frac{\alpha\epsilon\tau}{2(2D)^s} m} \le e^{-cn},$$

Let E acts on E^* on the right by

$$\forall x, y \in E,\ \forall \xi \in E^*, \quad (\xi \cdot x)(y) = \xi(xy).$$

For every $\gamma \in \Gamma$ and every $\xi \in E^*$, we have

$$I_{l_1,\dots,l_s,k}(\xi) = \int_{\gamma E} (\nu_{l_s} * \cdots * \nu_{l_1})^\wedge(\xi \cdot \gamma^{-1} g)\, \mathrm{d}\mu^{*k}(g).$$

Note that for any $g \in \gamma E \cap \Gamma$,

$$\|\xi\|\|g^{-1}\|^{-1} \ll_\gamma \|\xi \cdot \gamma^{-1} g\| \ll_\gamma \|\xi\|\|g\|. \tag{19}$$

Using the assumption that μ has a finite exponential moment and Markov's inequality, we can find a constant $C = C(\mu) \ge 1$ such that for any $k \ge 1$, the μ^{*k}-measure of the set of $g \in \Gamma$ such that

$$\|g\| \le e^{Ck} \quad \text{and} \quad \|g^{-1}\| \le e^{Ck} \tag{20}$$

is at least $1 - e^{-k}$.

Set $\alpha_0 = \frac{1}{4Ts}$ and let $\xi \in E^*$ be such that $e^{\alpha n} \le e^{\lambda_1 n}\|\xi\| \le e^{\alpha_0 n}$. Using $(1 - 2\omega)n \le Tsm \le n$ and $k \le 3\omega n$, we have, for any $g \in \mathrm{Supp}(\mu^{*k})$ satisfying (20),

$$e^{(\alpha - (2\lambda_1 + 4C)\omega)n} \le e^{T\lambda_1 sm}\|\xi \cdot \gamma^{-1} g\| \le e^{(\alpha_0 + 4C\omega)n}.$$

Here we assumed n to be larger than a constant depending on γ to beat the implied constant in the $\ll_\gamma$ notation in (19). With the choice $\omega = \min\{\frac{\alpha}{4\lambda_1+8C}, \frac{1}{16CTs}\}$, we can guarantee that this implies

$$e^{\alpha m/2} \leq e^{T\lambda_1 sm}\|\xi \cdot \gamma^{-1} g\| \leq e^m.$$

Putting everything together, we obtain

$$|I_{l_1,\dots,l_s,k}(\xi)| \leq e^{-cn}$$

for all $l_1, \dots, l_s \in \mathcal{L}$, $k \leq 3\omega n$ with $l_1 + \cdots + l_s + k = n$. This concludes the proof of the theorem. □

B.3 Proof of Theorems B.1 and B.2

Let μ be a Borel probability measure on $\mathrm{GL}_d(\mathbb{Z}) \ltimes \mathbb{R}^d$ having a finite exponential moment. Let $x \in X$ be a point. We shall use the shorthand $\nu_n = \mu^{*n} * \delta_x$. Assume that for some $a \in \mathbb{Z}^d \setminus \{0\}$ and $t \in (0, \frac{1}{2})$ and for some large n, we have

$$|\widehat{\nu_n}(a)| > t. \tag{21}$$

Let Γ denote the group generated by $\theta_*\mu$. Let G denote the Zariski closure of Γ and G° the identity component of G. Let E be the subalgebra generated by $G^\circ(\mathbb{R})$. Let $\gamma_1, \dots, \gamma_J$ be a complete set of representatives for the cosets in $\Gamma/(\Gamma \cap E)$. For any integer m, the we can decompose

$$(\theta_*\mu)^{*m} = \sum_{j=1}^{J} (\gamma_j)_*\mu_{m,j}$$

where $\mu_{m,j}$ is a measure on $\Gamma \cap E$. By Theorem B.7, for m large enough, we have the Fourier decay property for each $\mu_{m,j}$,

$$\forall \xi \in E^* \text{ with } e^{\alpha m} \leq e^{\lambda_1 m}\|\xi\| \leq e^{\alpha_0 n}, \quad |\widehat{\mu_{m,j}}(\xi)| \leq e^{-c_0 n}.$$

Writing $\nu_n = \mu^{*m} * \nu_{n-m}$, we have

$$\begin{aligned}\widehat{\nu_n}(a) &= \iint e(\langle a, gy\rangle)\,\mathrm{d}\mu^{*m}(g)\,\mathrm{d}\nu_{n-m}(y)\\ &= \sum_{j=1}^{J} \iint e(\langle a, gy\rangle)\mathbb{1}_{\gamma_j E}(\theta(g))\,\mathrm{d}\mu^{*m}(g)\,\mathrm{d}\nu_{n-m}(y)\end{aligned}$$

Thus, (21) implies that there exists $j \in \{1, \dots, J\}$ such that

$$t \ll \left| \iint e(\langle a, gy \rangle) \mathbb{1}_{\gamma_j E}(\theta(g)) \, \mathrm{d}\mu^{*m}(g) \, \mathrm{d}\nu_{n-m}(y) \right|$$

By Hölder's inequality,

$$t^{2k} \ll \int \left| \int e(\langle a, gy \rangle) \mathbb{1}_{\gamma_j E}(\theta(g)) \, \mathrm{d}\mu^{*m}(g) \right|^{2k} \mathrm{d}\nu_{n-m}(y)$$

After developing the $2k$-power and separating the linear part with the translation part, we obtain

$$\begin{aligned} t^{2k} \ll \int_{(\gamma_j E)^{2k}} & \left| \widehat{\nu_{n-m}}((g_1 + \dots + g_k - g_{k+1} - \dots - g_{2k})^{\mathrm{tr}} a) \right| \\ & \times \, \mathrm{d}\big((\theta_* \mu)^{*m}\big)^{\otimes 2k}(g_1, \dots, g_{2k}). \end{aligned}$$

That is,

$$t^{2k} \ll \int \left| \widehat{\nu_{n-m}}(g^{\mathrm{tr}} \gamma_j^{\mathrm{tr}} a) \right| \mathrm{d}\big(\mu_{m,j}^{\boxplus k} \boxminus \mu_{m,j}^{\boxplus k}\big)(g).$$

Then, the same argument in the proof of [15, Proposition 4.1] leads to

Proposition B.12 *There are constants $C \geq 1$ and $\sigma > \tau > 0$ depending only on $\theta_* \mu$ such that for $m \geq C |\log t|$, the above implies that there exists a r_0-seperated subset $Q \subset \mathbb{R}^d / \mathbb{Z}^d$ such that*

$$\nu_{n-m}\Big(\bigcup_{x \in Q} B(x, \rho_0) \Big) \geq t^C$$

where $\rho_0 = e^{-\sigma m} \|a\|$ and $r_0 = e^{\tau m} \rho_0$.

From here on, the proof of Theorem B.1 is identical to that of [15, Theorem 1.2] and that of Theorem B.2 is identical to that of [16, Theorem 1.3]. That is, the Zariski-connectedness condition is not used in the relevant parts in [15] and [16].

References

1. Aoun, R.: Random subgroups of linear groups are free. Duke Math. J. **160**(1), 117–173 (2011). MR 2838353
2. Bekka, B., Guivarc'h, Y.: On the spectral theory of groups of affine transformations of compact nilmanifolds. Ann. Sci. Éc. Norm. Supér. (4) **48**(3), 607–645 (2015). MR 3377054

3. Bekka, B., Heu, J.-R.:Random products of automorphisms of Heisenberg nilmanifolds and Weil's representation. Ergodic Theory Dyn. Syst. **31**(5), 1277–1286 (2011). MR 2832245
4. Benoist, Y., Quint, J.: Stationary measures and invariant subsets of homogeneous spaces (iii). Ann. Math. **178**, 1017–1059 (2013)
5. Bourgain, J., Dyatlov, S.: Fourier dimension and spectral gaps for hyperbolic surfaces. Geom. Funct. Anal. **27**(4), 744–771 (2017). MR 3678500
6. Bourgain, J., Furman, A., Lindenstrauss, E., Mozes, S.: Stationary measures and equidistribution for orbits of nonabelian semigroups on the torus. J. Am. Math. Soc. **24**(1), 231–280 (2011). MR 2726604
7. Boyer, J.-B.: On the affine random walk on the torus (2017). e- prints: arXiv:1702.08387
8. Boyer, J.-B.: Almost sure functional central limit theorem for the linear random walk on the torus. Probab. Theory Related Fields **173**(1–2), 651–696 (2019). MR 3916116
9. Butzer, P.L., Nessel, R.J.: Fourier Analysis and Approximation (Academic Press, New York-London, 1971). Volume 1: One-Dimensional Theory, Pure and Applied Mathematics, vol. 40. MR 0510857
10. Corwin, L.J., Greenleaf, F.P.: Representations of Nilpotent Lie Groups and Their Applications. Part I. Cambridge Studies in Advanced Mathematics, vol. 18 (Cambridge University Press, Cambridge, 1990). Basic theory and examples. MR 1070979
11. Eskin, A., Lindenstrauss, E.: Random walks on locally homogeneous spaces. Preprint
12. Furman, A., Shalom, Y.: Sharp ergodic theorems for group actions and strong ergodicity. Ergodic Theory Dyn. Syst. **19**(4), 1037–1061 (1999). MR 1709429
13. Guivarc'h, Y., Starkov, A.N.: Orbits of linear group actions, random walks on homogeneous spaces and toral automorphisms. Ergodic Theory Dyn. Syst. **24**(3), 767–802 (2004). MR 2060998
14. He, W.: Random walks on linear groups satisfying a Schubert condition. Israel J. Math. **238**(2), 593–627 (2020). MR 4145811
15. He, W., de Saxcé, N.: Linear random walks on the torus. Duke Math. J. **171**(5), 1061–1133, (1 April 2022). MR 4402559
16. He, W., Lakrec, T., Lindenstrauss, E.: Affine Random Walks on the Torus. International Mathematics Research Notices (2021), rnaa322
17. Muchnik, R.: Semigroup actions on $\mathbb{T}^n$. Geom. Dedicata **110**, 1–47 (2005). MR 2136018
18. Raghunathan, M.S.: Discrete Subgroups of Lie Groups (Springer, New York-Heidelberg, 1972). Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 68. MR 0507234
19. Wang, Z.: Quantitative density under higher rank abelian algebraic toral actions. Int. Math. Res. Not. IMRN (16), 3744–3821 (2011). MR 2824843

Logarithmic Quantum Dynamical Bounds for Arithmetically Defined Ergodic Schrödinger Operators with Smooth Potentials

Svetlana Jitomirskaya and Matthew Powell

Dedicated to the memory of Jean Bourgain

Abstract We present a method for obtaining power-logarithmic bounds on the growth of the moments of the position operator for one-dimensional ergodic Schrödinger operators. We use Bourgain's semialgebraic method to obtain such bounds for operators with multifrequency shift or skew-shift underlying dynamics with arithmetic conditions on the parameters.

1 Introduction

It is well known that Anderson localization (pure point spectrum with exponentially decaying eigenfunctions) is highly unstable with respect to various perturbations. For quasiperiodic operators, it very sensitively depends on the arithmetics of the phase (a seemingly irrelevant parameter from the point of view of the physics of the problem) and doesn't hold generically [15]. It can also be destroyed by generic rank one perturbations [7, 10]. This instability is therefore also present for the—very physically relevant—notion of dynamical localization, defined as non-spread of the initially localized wave packet or boundedness in time of the moments of the position operator (see (3)).

Thus moments of the position operator for generic rank one perturbations of many operators with a.e. dynamical localization are unbounded in time. This bizarre situation is partially rescued by a result of [5, 6]: when eigenfunctions have an additional SULE (semi-uniform localization) property, the moments of the position operators of *all* rank-one perturbations grow at most power-logarithmically. Indeed SULE has since been proved for all operators with localization that come from physically realizable models. From this point of view, power-logarithmic bounds of

S. Jitomirskaya (✉) · M. Powell
Department of Mathematics, University of California, Irvine, CA, USA
e-mail: szhitomi@uci.edu; szhitomi@math.uci.edu; mtpowell@uci.edu

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_8

the moments are the stable—and therefore physically relevant—property, making it worthwhile to prove directly for operator families with (expected) a.e. localization, bypassing the localization proof. This, in particular, includes one-dimensional ergodic operator families $H_{\omega,x} : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ given by

$$(H_{\omega,x}\psi)(n) = \psi(n-1) + \psi(n+1) + V(T_\omega^n(x))\psi(n), \tag{1}$$

where T_ω is an ergodic transformation and V is a real-valued function, in the regime of positive Lyapunov exponents.

Direct proofs of upper quantum dynamical bounds for quasiperiodic and other ergodic operators with positive Lyapunov exponents have been done, in increasing generality in [8, 11, 14]. In all these cases, the results featured the desired stability in phase and often were also arithmetic in frequency (in contrast with many localization proofs). All the papers mentioned above obtain vanishing of the transport exponents $\beta(p)$ (see (4)), which implies *sub-polynomial* growth of the moments. Here we present a method that allows to improve this to the desired *power-logarithmic* bounds. We note that our results are also phase-stable and our frequency conditions are arithmetic. The only previous direct proof of power-logarithmic bounds was done for the Anderson model in [16] based on different considerations, but we note that for the Anderson model, localization always holds ([4] or sees a very simple recent argument in [18]). Thus, to the best of our knowledge, we present the first proof of power-logarithmic quantum dynamical bounds for models without localization.

To get such bounds we, inspired by the theory of logarithmic dimension developed in [23], introduce the notion of logarithmic transport exponents (see (5)) and obtain estimates for them.

Technically, our method goes back to [12] where the existence of transfer matrices growing appropriately along a subsequence was first used to prove zero Hausdorff dimension of spectral measures for one-frequency quasiperiodic operators, including in situations where localization cannot hold. The ideas of [12] were first applied in [8] to obtain vanishing transport exponents for those models, and then this was further modified and developed in [14] to allow very rough functions. These methods however required continued fraction techniques and did not extend naturally even to the case of higher-dimensional tori. This was tackled in [11] which developed a method allowing to handle general dynamics of zero topological entropy. Here, for our one-frequency result, we go back to the approach of [8, 12, 14]. The method of [11] however is too rough for the logarithmic scale. It turns out that for higher-dimensional shifts and skew-shifts already the basics of the Bourgain's semialgebraic/large deviations method [3] are ideally suited to obtain the desired power-logarithmic bounds on the moments.

The key estimate from Bourgain's method used here is the sublinear bound (23) on the number of hits of a semialgebraic set by a shift [3] or skew-shift [22] trajectory. In fact, all we need is a much weaker statement: the existence of at least one miss in sublinear time, which of course follows from the sublinear bound. We make some explicit estimates on the power used in the sublinear bound (23) in Sect. 4. The sublinear bound was also fruitfully used in a recent work [13] to establish vanishing of transport exponents $\beta(p)$ (thus subpolynomial bounds on the moment growth) for long-range quasiperiodic operators, for which the authors of [13] developed a non-transfer matrix-based approach. It is an interesting question whether power-logarithmic bounds can be also obtained in that case.

We cover all scenarios where a.e. Anderson localization has been proved for one-dimensional operators with analytic quasiperiodic and skew-shift potentials as described in Bourgain's book [3] and with Gevrey extensions in [19, 20]. For all these models, the a.e. dynamical localization was also shown to hold [2]. Essentially, what we demonstrate by this work is that power-logarithmic bounds on transport can be viewed as *dynamical localization-light*, since the proof is considerably simpler than that of localization and in fact can be obtained in many known scenarios as a part of the latter proof. Yet the results are phase-stable and presumably optimal as far as phase-stable results go. Just as with Anderson localization, our theorems are non-perturbative (obtained as a corollary of positive Lyapunov exponents) for analytic potentials over toral shifts and Gevrey potentials for one-frequency shifts, while they require large coupling constants dependent on the frequency for the multifrequency Gevrey and skew-shift cases. We note, however, that all such dependence comes from the large deviation estimates that we use as a black box; we don't add any further "perturbative" components through our technique.

We proceed to formulate our main results. Consider the time-averaged quantity:

$$a(n, T) = \frac{2}{T}\int_0^\infty e^{2t/T}\frac{1}{2}\left(\left|\left\langle e^{itH_{\omega,x}}\delta_0, \delta_n\right\rangle\right|^2 + \left|\left\langle e^{itH_{\omega,x}}\delta_1, \delta_n\right\rangle\right|^2\right)dt, \tag{2}$$

where $\delta_n(m) = 1$ when $m = n$ and 0 otherwise.

Dynamical localization is characterized by boundedness in time of the moments of the position operator:

$$\left\langle |X|^p(T)\right\rangle = \sum_{n\in\mathbb{Z}}(1+|n|)^p a(n, T). \tag{3}$$

For simplicity, we are restricting our attention to time-averaged quantities rather than considering $a(n,t) = \frac{1}{2}\left(\left|\left\langle e^{itH_{\omega,x}}\delta_0, \delta_n\right\rangle\right|^2 + \left|\left\langle e^{itH_{\omega,x}}\delta_1, \delta_n\right\rangle\right|^2\right)$, but our analysis can be carried through for non-time-averaged quantities as well, following the ideas in [8]. We only consider time-averaging for a small simplification.

Dynamical localization always implies Anderson localization but is strictly stronger [6, 17] . When dynamical localization does not hold, the moments of the position are unbounded in time, and a natural quantity of interest is how fast this growth is. Classically, this is captured by the upper and lower transport exponents:

$$\beta^+(p) = \limsup_{t\to\infty}\frac{\ln\langle |X|^p(t)\rangle}{p\ln t}; \quad \beta^-(p) = \liminf_{t\to\infty}\frac{\ln\langle |X|^p(t)\rangle}{p\ln t}, \tag{4}$$

which describe power-law bounds on the growth of the moments. It is known that, under very relaxed conditions (c.f. [11]), the transport exponents vanish when the Lyapunov exponent is positive. Let us refine the notion of transport exponents by defining the logarithmic transport exponents as

$$\beta_{\ln}^{+}(p) = \limsup_{t\to\infty} \frac{\ln \langle |X|^{p}(t)\rangle}{p \ln\ln t}; \quad \beta_{\ln}^{-}(p) = \liminf_{t\to\infty} \frac{\ln \langle |X|^{p}(t)\rangle}{p \ln\ln t}. \tag{5}$$

Our first result is that positivity of the Lyapunov exponent will imply that this exponent is finite for every p.

Let T_ω represent either the shift or the skew-shift on the torus, $\mathbb{T}^\nu$, $G^\sigma(\mathbb{T}^\nu)$ denote the Gevrey class, $L(E)$ denote the Lyapunov exponent, and $DC(A, c)$ and $SDC(A, c)$ denote Diophantine conditions (see Sect. 2 for the relevant definitions). In this regime, we have the following.

Theorem 1.1 *Let $H_{\omega,x}$ be an operator of the form* (1) *with T_ω given by the shift on $\mathbb{T}$, and either f is analytic or $f \in G^\sigma(\mathbb{T}), \sigma > 1$, and obeys the transversality condition* (12)*. Suppose that $L(E) > 0$ for every $E \in \mathbb{R}$. Then for any $x \in \mathbb{T}, \epsilon > 0$ and $m > 0$,*

(1) If $\omega \in \mathbb{R}\backslash\mathbb{Q}$, then $\liminf_{T\to\infty} \frac{\langle |X|^m(T)\rangle}{(\ln T)^{m(\sigma+1+\epsilon)}} < \infty$.
(2) If $\omega \in DC(A, c)$, then $\limsup_{T\to\infty} \frac{\langle |X|^m(T)\rangle}{(\ln T)^{m(\sigma+1+\epsilon)}} < \infty$.

Remark 1 We can rewrite the conclusions of Theorem 1.1 as follows:

(1) If $\omega \in \mathbb{R}\backslash\mathbb{Q}$, then $\beta_{\ln}^{-}(p) \le 1 + \sigma$ for every $p > 0$ and $x \in \mathbb{T}$.
(2) If $\omega \in DC(A, c)$, then $\beta_{\ln}^{+}(p) \le 1 + \sigma$ for every $p > 0$ and $x \in \mathbb{T}$.

Remark 2 For analytic f the conclusion holds with $\sigma = 1$.

We have similar logarithmic quantum-dynamical bounds for non-constant analytic potentials on higher-dimensional tori.

Theorem 1.2 *Let $H_{\omega,x}$ be an operator of the form* (1) *with T_ω given by the shift on $\mathbb{T}^\nu$ with $\nu > 1$. Suppose also that f is a non-constant analytic function on $\mathbb{T}^\nu$, $\omega \in DC(A, c)$ and that $L(E) > 0$ for every $E \in \mathbb{R}$. Then there exists $\gamma = \gamma(\nu, A)$ such that, for every $m > 0$,*

$$\beta_{\ln}^{\pm}(m) \le \gamma. \tag{6}$$

for all $x \in \mathbb{T}^\nu$.

Remark 3 For analytic f, the condition $L(E) > 0$ for every $E \in \mathbb{R}$ is satisfied for λf, where $\lambda > \lambda_0(f)$. Also we have as an immediate corollary that there exists $\gamma(\nu)$ such that for a.e. $\omega \in \mathbb{T}^\nu$, $\beta_{\ln}^{\pm}(m) \le \gamma(\nu)$ for every $m > 0$.

Things become a bit more technical when we consider the multifrequency shift with potentials in the Gevrey class, or when considering the skew-shift instead of the shift.

Theorem 1.3 *Let $x \in \mathbb{T}^\nu$. Let $H_{\omega,x}$ be an operator of the form* (1) *with T_ω given by the shift on $\mathbb{T}^\nu$ with $\nu > 1$. Suppose also that $f = \lambda f_0 \in G^\sigma(\mathbb{T}^\nu)$ such that f_0 obeys the transversality condition* (12)*, $\omega \in DC(A, c)$ and that $L(E) > 0$ for every $E \in \mathbb{R}$. Then there exists $\lambda_0 = \lambda_0(f_0, \omega) > 0$ and $\gamma = \gamma(\sigma, \nu, A)$ such that, for every $\lambda > \lambda_0$ and $m > 0$,*

$$\beta_{\ln}^{\pm}(m) \leq \gamma. \tag{7}$$

Remark 4 The condition on λ_0 comes from [19] and is necessary to obtain and use a large deviation estimate which is critical to our proof. See Theorem 2.4.

Theorem 1.4 *Let $H_{\omega,x}$ be an operator of the form* (1) *with T_ω given by the skew-shift on $\mathbb{T}^\nu$, suppose $f = \lambda f_0 \in G^\sigma(\mathbb{T}^\nu)$ such that f_0 obeys* (12)*, and $\omega \in SDC(A, c)$, for some $A \leq 2$. Suppose that $L(E) > 0$ for every $E \in \mathbb{R}$. Then there exists $\lambda_0 = \lambda_0(f_0, \omega) > 0$ and $\gamma = \gamma(\sigma, \nu, A)$ such that for every $\lambda > \lambda_0$ and $m > 0$,*

$$\beta_{\ln}^{\pm}(m) \leq \gamma. \tag{8}$$

for all $x \in \mathbb{T}^\nu$.

Remark 5 As mentioned earlier, the perturbative nature of Theorems 1.3 and 1.4 is fully captured in the ω-dependence of λ_0 that comes from [19, 20], while the bound γ that we prove to exist is constant for a.e. Diophantine ω.

Remark 6 We will see in our proof that the γ that appears in Theorems 1.3 and 1.4 has ω-dependence which appears precisely as the constant δ from (23). It is possible to explicitly compute $\gamma = C(\sigma\nu + 1)\left(\frac{1}{\delta}\right)$. Here C is a universal constant $C = C(\nu)$. The constant δ is different for the shift and skew-shift and will be obtained by semialgebraic methods in Sect. 4, where we obtain the explicit estimates $\delta \leq \frac{1}{A+\nu}$ for the shift and $\delta < \frac{1}{A\nu 2^{\nu-1}}$ for the skew-shift.

Remark 7 One of the only places where there is still room for improvement in this approach is the estimate on δ in Theorem 2.2. The closer δ is to 1, the smaller γ will be and thus the better the localization result. Our estimate for the shift follows from a harmonic analysis approach given by Bourgain. For $\omega \in DC(A, c)$, other estimates have been obtained by other authors using alternative methods (c.f. [11] and [22]), but when $A \gg 1$, our localization result is stronger.

We note that the method in [11] while applicable to all our models and a lot more is insufficient to obtain ln-type estimates which we are after here, largely because it allows to find the required exponential growth of the transfer matrix only on polynomially large length scales, whereas the growth needs to be on logarithmic length scales to obtain ln-type estimates.

Related to dynamical bounds are dimensional bounds on spectral measures. It is known that positive Lyapunov exponent implies that the spectral measures have

Hausdorff dimension zero for every phase. A finer notion, introduced in [21] and explored in more generality in [23], is the logarithmic dimension. In short, we say that the upper logarithmic dimension of a measure, μ, is less than α if the measure is supported on a set of logarithmic dimension less than α. A result due to Simon [24] says that spectral measures for 1D quasiperiodic operators with positive Lyapunov exponent are supported on a set of logarithmic capacity 0 for a.e. phase. This implies that the upper logarithmic dimension of the spectral measures is at most 1 for a.e. phase. It leaves unclear what happens on this null set of phases. Moreover, while upper bounds on quantum dynamics imply suitable upper bounds on upper dimension of spectral measures, the reverse is not, in general, true. Indeed, examples are known where the spectral measure is pure point but quantum dynamics is even quasi-ballistic (see [6]). Since we prove power-logarithmic quantum dynamics bounds for all phase, a consequence is a (weaker) bound on the upper logarithmic dimension for every phase. Thus, while we obtain weaker dimensional estimates this way, we are able to handle every phase, not just a.e. phase.

By Theorem 2.6 from [23], we have the following corollary.

Corollary 1.1 *Under the assumptions of Theorem 1.1, with $\omega \in DC(A, c)$, we have $\dim_{\ln}^{+}(\mu) \leq 1 + \sigma$, where μ is the spectral measure related to δ_0 and $H_{\omega,x}$. Under the assumptions of Theorem 1.3, we have $\dim_{\ln}^{+}(\mu) \leq \gamma$.*

Other quantities have been proposed for studying dynamical localization-type estimates, see [1, 8], but one of the major advantages of $\beta_{\ln}^{\pm}(p)$ is that, similar to $\beta^{\pm}(p)$, it is stable under perturbations in certain circumstances. See Theorem 1.5 part (b) for a precise statement.

One transfer matrix-based way to approach upper dynamical bounds goes back to a scheme by Damanik and Tcheremchantsev [8] wherein the quantity $\beta^{\pm}(p)$ was related to suitable growth of the transfer matrices along suitable length scales (see also [16]). In this paper, we refine this scheme to allow us to obtain finer dynamical estimates. Our contribution is the following theorem, which required us to address certain technical limitations in the original argument (see Sect. 2.2 for the relevant definitions and Sect. 3 for full details).

Theorem 1.5 *Suppose H_1 is of the form (1) with bounded potential v_1 and $\sigma(H_1) \subset [-K + 1, K - 1]$.*

(a) *Suppose for all $\delta < \infty$ and $T > T_0$, we have*

$$\int_{-K}^{K} \left(\min_{l=\pm 1} \max_{1 \leq lj \leq (\ln T)^{\gamma}} \left\| A_j^{v_1, E+i/T}(x) \right\|^2 \right)^{-1} dE = O(T^{-\delta}) \tag{9}$$

for some $\gamma > 1$. Then $\beta_{\ln,1}^{+}(p) \leq \gamma$, where $\beta_{\ln,1}^{+}(p)$ is the transport exponent associated to H_1. If the above condition holds for a sequence $T_n \to \infty$, then $\beta_{\ln,1}^{-}(p) \leq \gamma$.

(b) *In addition to the above, suppose also that H_2 is an operator of the form (1) with bounded potential v_2 such that $\sigma(H_2) \subset [-K + 1, K - 1]$ and suppose* *that there exists* $B > 0$ *such that for all* $E \in [-K+1, K-1]$, $0 < \epsilon \leq 1$, *and* $|n| \leq \ln(\epsilon^{-1})$,

$$\epsilon^B \left\| A_n^{v_1, E+i\epsilon} \right\| \lesssim \left\| A_n^{v_2, E+i\epsilon} \right\| \lesssim \epsilon^{-B} \left\| A_n^{v_1, E+i\epsilon} \right\|. \tag{10}$$

Then $\beta^{\pm}_{\ln,2}(p) \leq \gamma$ *for every* $p > 0$, *where* $\beta^{\pm}_{\ln,2}(p)$ *is the transport exponent associated to* H_2.

Remark 8 It is worth noting that Theorem 1.5 is a purely deterministic result and thus holds for general operators of the form

$$(Hu)(n) = u(n-1) + u(n+1) + V(n)u(n),$$

where V is a bounded sequence of real numbers.

Theorem 1.5 is similar to Theorem 1 in [8], but there is a major issue with just repeating the proof of Theorem 1 in [8] using $(\ln T)^{\gamma}$ in place of T^{γ}. The problem is that the result in [8] a priori assume that $\beta^{\pm}(p) < \infty$ for every $p > 0$. This is the well-known ballistic upper bound. We do not, unfortunately, have a similar a priori estimate on $\beta^{\pm}_{\ln}(p)$, even when $\beta^{\pm}(p) = 0$, which means the original argument is insufficient. Our main technical achievement on the way to a proof of Theorem 1.5 is a sufficient condition (Theorem 3.2) under which we can say $\beta^{\pm}_{\ln}(p) < C < \infty$ for every $p > 0$. Once we have this, we can use the ideas from [8] to obtain Theorem 1.5.

This essentially reduces the problem of bounding log-transport exponents to obtaining lower bounds on the growth of the transfer matrix along particular length scales. This will be done in a two-step process. First, we will demonstrate that, for a fixed energy and frequency, transfer matrix growth can be suboptimal only for a set of phases of small measure. This will be captured by so-called large deviation estimates. Then we will show that every phase will correspond to a transfer matrix with good growth after at most power-log many iterates of the transformation.

The rest of our paper is organized in the following way. In Sect. 2 we introduce the relevant definitions needed for our paper. Section 2.2 is devoted to those definitions needed for the proof of Theorem 1.5. Section 2.3 recalls facts about semialgebraic sets which will be necessary for the proof of Theorem 1.3. Section 2.4 recalls the large deviation theorems needed for measure estimates. We prove Theorem 1.5 in Sect. 3. We explicitly compute discrepancy bounds in Sect. 4. We prove two technical lemmas regarding the set of "good" phases in Sect. 5. Finally, we prove Theorem 1.1 in Sect. 6 and Theorem 1.3 in Sect. 7. Proofs of Theorems 1.2 and 1.4 are essentially identical to that of Theorem 1.3. However, we describe the small changes needed in, correspondingly, Sects. 8 and 9.

2 Preliminaries

2.1 Schrödinger Operators and Transfer Matrices

We consider the two particular types of Schrödinger operator, $H_{\omega,x} : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ given by

$$(H_{\omega,x}\psi)(n) = \psi(n-1) + \psi(n+1) + f(T_\omega^n(x))\psi(n), n \in \mathbb{Z}. \tag{11}$$

The first case we consider is where $x \in \mathbb{T}^\nu$, T_ω is the shift: $T_\omega x = x + \omega$ and $\omega = (\omega_1, ..., \omega_\nu)$ and $(\omega_1, ..., \omega_\nu, 1)$ are rationally independent. The second case we consider is where $x \in \mathbb{T}^\nu$, T_ω is the skew-shift: $T_\omega(x_1, ..., x_\nu) = (x_1 + \omega, x_2 + x_1, x_3 + x_2, ..., x_\nu + x_{\nu+1})$ and $\omega \in \mathbb{R}\backslash\mathbb{Q}$.

Additionally, we recall that $G^\sigma(\mathbb{T}^\nu)$ denotes the Gevrey class:

$$G^\sigma(\mathbb{T}^\nu) = \left\{ f : \mathbb{T}^\nu \to \mathbb{R} : \left|\left|D^\alpha f\right|\right|_\infty < C^{|\alpha|+1}(\alpha!)^\sigma \right\}.$$

An equivalent definition of G^σ which we will take advantage of is:

$$G^\sigma(\mathbb{T}^\nu) = \left\{ f : \mathbb{T}^\nu \to \mathbb{R} : |\hat{f}(n)| \le e^{-|n|^{1/\sigma}} \right\}.$$

In both of the cases, we will consider $f \in G^\sigma(\mathbb{T}^\nu)$ in (11).

For technical reasons, we will further restrict our attention to those Gevrey class functions that obey a transversality condition:

$$D^\alpha f(x) \neq 0 \quad \text{for any } x \in \mathbb{T}^\nu, \alpha \in \mathbb{N}^\nu. \tag{12}$$

From this point forward, when discussing $f \in G^\sigma(\mathbb{T}^\nu)$, we will mean those $f \in G^\sigma(\mathbb{T}^\nu)$ that satisfy (12). Recall that, for any $E \in \mathbb{C}$, any solution to the eigenequation $H_{\omega,x}\psi = E\psi$ can be reconstructed from the n-step transfer matrix:

$$A_n^{f,E}(x) = \prod_{k=n}^{1} \begin{pmatrix} f(T_\omega^k(x)) - E & -1 \\ 1 & 0 \end{pmatrix} \tag{13}$$

by

$$\begin{pmatrix} \psi(n+1) \\ \psi(n) \end{pmatrix} = A_n^{f,E}(x) \begin{pmatrix} \psi(1) \\ \psi(0) \end{pmatrix}. \tag{14}$$

We can then define

$$L_n(E) = \frac{1}{n} \int \ln \left|\left| A_n^{f,E}(x) \right|\right| dx$$

and the Lyapunov exponent is given by

$$L(E) = \lim L_n(E) = \inf L_n(E).$$

We will also need a Diophantine condition. We say that $\omega \in DC(A, c)$ if $||k \cdot \omega|| > c|k|^{-A}$ for every $k \in \mathbb{Z}^\nu \backslash \{0\}$. We say that $\omega \in SDC(A, c)$ if $||k \cdot \omega|| > c\frac{1}{|k|(\ln |k|)^A}$. We will only consider $\omega \in SDC(A, c)$ for $A \leq 2$, which is a restriction imposed by Theorem 2.4. See [19] for details.

In what follows, C and c will denote finite constants and ϵ will denote a small constant, all of which can only depend on f, ν, ω, or E. Moreover, these constants may change throughout a proof, but ϵ will always denote a small constant, and boundedness of C and c will be unchanged.

2.2 Transport Exponents

Recall that we have defined

$$\beta_{\ln}^{+}(p) = \limsup \frac{\ln \langle |X|^p(t) \rangle}{p \ln \ln t}; \quad \beta_{\ln}^{-}(p) = \liminf \frac{\ln \left\langle |X|_t^p \right\rangle}{p \ln \ln t}.$$

It is simple to verify via Hölder's inequality that $\beta_{\ln}^{\pm}(p)$ is non-decreasing in p, so obtaining a bound on $\beta_{\ln}^{\pm}(+\infty)$ is sufficient for bounding $\beta_{\ln}^{\pm}(p)$ for any $p > 0$.

To bound $\beta_{\ln}^{\pm}(+\infty)$, for general operators, we will need to define the so-called outside probabilities:

$$P_l(N, T) = \sum_{n<-N} a(n, T) \tag{15}$$

$$P_r(N, T) = \sum_{n>N} a(n, T) \tag{16}$$

$$P(N, T) = P_l(N, T) + P_r(N, T) \tag{17}$$

$$= \sum_{|n|>N} a(n, T) \tag{18}$$

along with associated log-transport quantities:

$$S_{\ln}^{+}(\alpha) = -\limsup \frac{\ln(P((\ln T)^\alpha - 2, T))}{\ln \ln T} \tag{19}$$

$$S_{\ln}^{-}(\alpha) = -\liminf \frac{\ln(P((\ln T)^\alpha - 2, T))}{\ln \ln T} \tag{20}$$

$$\alpha_{\ln}^{\pm} = \sup \left\{ \alpha \geq 0 : S_{\ln}^{\pm}(\alpha) < \infty \right\}. \tag{21}$$

A quick note on our convention here: we use $(\ln T)^{\alpha} - 2$ so that $S_{\ln}^{\pm}(0) = 0$ as in [8].

Our goal in Sect. 3 will be to show that, under suitable conditions, $\beta_{\ln}^{\pm}(p) \leq \alpha_{\ln}^{\pm}$ for every $p > 0$, which will be used to establish Theorem 1.5.

2.3 Semialgebraic Sets

Definition 2.1 We say that a set $\mathcal{S} \subset \mathbb{R}^n$ is semialgebraic if it can be written as a finite union of polynomial inequalities. More precisely, suppose $P = \{p_1, \ldots, p_s\} \subset \mathbb{R}[X_1, \ldots, X_n]$ is a finite collection of real polynomials in n variables, whose degrees are bounded by d. A closed semialgebraic set, $\mathcal{S} \subset \mathbb{R}^n$, is given by an expression of the form

$$\mathcal{S} = \bigcup_{j=1}^{k} \bigcap_{m \in Q_j} \left\{x \in \mathbb{R}^n : p_m s_{jm} 0\right\}, \tag{22}$$

where $Q_j \subset \{1, ..., s\}$ and $s_{jm} \in \{\leq, =, \geq\}$ are arbitrary. Moreover, we say that $\mathcal{S}$ has degree at most sd and its degree is the infimum of sd over all representations as in (22).

Theorem 2.1 ([3] Corollary 9.6) *Let $\mathcal{S} \subset [0, 1]^n$ be semialgebraic of degree B. Let $\epsilon > 0$ be a small number and $|\mathcal{S}| < \epsilon^n$, where $|\cdot|$ represents Lebesgue measure. Then there exists $C = C(n)$ such that $\mathcal{S}$ may be covered by at most $B^C \epsilon^{1-n}$ ϵ-balls.*

Using these results for general semialgebraic sets, we can obtain sublinear bounds for the shift and skew-shift.

Theorem 2.2 *Let T_ω represent either the shift or the skew-shift. Let $\mathcal{S} \subset [0, 1]^n$ be semialgebraic of degree B and $|\mathcal{S}| < \eta$. Let $\omega \in DC(A, c)$ (when considering the shift) or $\omega \in SDC(A, c)$ (when considering the skew-shift), and let N be an integer such that*

$$B \leq N < \frac{1}{\eta}.$$

Then there is $C = C(n)$ and $\delta = \delta(\omega)$ such that for any $x_0 \in \mathbb{T}^n$,

$$\#\left\{k = 1, ..., N : T_{\omega}^{k}(x_0) \in \mathcal{S}\right\} < N^{1-\delta} B^C. \tag{23}$$

Remark 9 While the above result holds for any $N \geq B$, the resulting bound, $N^{1-\delta} B^C$, will only be smaller than N when $\ln(N) > C \ln(B)$, where $C = C(n, \delta)$.

The case where T_ω is the shift is due to Bourgain [[3] Corollary 9.7], and the case for the skew-shift follows from Lemma 8.4 in [22]. The particular δ obtained differs between the shift and skew-shift, as we will show in Sect. 4.

Remark 10 Different authors obtain different values of δ for the shift (c.f. [22] and [11]) depending on what method they use. In Sect. 4 we explicitly estimate δ for the shift using the approach from [3], which turns out to be better than the values from [22] and [11] when $\omega \in DC(A, c)$, $A \gg 1$.

2.4 Large Deviation Theorems

Throughout the section, we will assume that the energy, E, is such that $L(E) > 0$.

The estimate we will obtain in Sect. 4 will rely on estimates on the measure of semialgebraic sets. The particular semialgebraic sets we are interested in are the set of phases, x, for which $\frac{1}{n}\left|\left|A_n^{f,E}(x)\right|\right|$ converges to $L(E)$ slowly. To this end, we recall the following large deviation theorems, the first of which is due to Bourgain, Goldstein, and Schlag, and the second is due to S. Klein, which quantitatively measure the rate of convergence.

For the shift model with non-constant analytic potential, there is a well-known large deviation estimate.

Theorem 2.3 ([3] Theorem 5.5) *Assume $\omega \in \mathbb{T}^\nu$ satisfies $\omega \in DC(A, c)$. Let f be a non-constant real analytic function on $\mathbb{T}^\nu$. Then there is $\alpha = \alpha(A) > 0$ such that*

$$\left|\left\{x \in \mathbb{T}^\nu : \left|\frac{1}{N}\ln\left|\left|A_N^{f,E}(x)\right|\right| - L_N(E)\right| < N^{-\alpha}\right\}\right| < e^{-N^\alpha}. \tag{24}$$

For the shift model with Gevrey class potential and skew-shift with analytic or Gevrey class potential satisfying a transversality condition, we have:

Theorem 2.4 ([19] Theorem 6.1) *Assume $f \in G^\sigma(T^\nu)$ satisfies a transversality condition, and suppose $f = \lambda f_0$, for some $\lambda \in \mathbb{R}$ and $f_0 \in G^\sigma$ fixed. Let $\omega \in DC(c, A)$ (for the shift) or $\omega \in SDC(A, c)$, $A \leq 2$ (for the skew-shift). Then there exists $\lambda_0 = \lambda_0(f_0, A)$ such that for every fixed $|\lambda| > \lambda_0$ and for every energy E, we have*

$$\left|\left\{x \in \mathbb{T}^\nu : \left|\frac{1}{N}\ln\left|\left|A_N^{f,E}(x)\right|\right| - L_N(E)\right| < N^{-\tau}\right\}\right| < e^{-N^\alpha}, \tag{25}$$

for some constants $\tau, \alpha > 0$ depending only on ν, and every $N > N_0(\lambda, c, f_0, \sigma, \nu)$.

3 Transport Exponents

Our first goal in this section is to relate $\beta_{\ln}^{\pm}(p)$ to $S_{\ln}^{\pm}$. Observe that if $S_{\ln}^{-}(\alpha) < +\infty$, we have:

$$P((\ln T)^{\alpha} - 2, T) > (\ln T)^{-S_{\ln}^{-}(\alpha)-} \tag{26}$$

and so

$$\begin{aligned}\left\langle |X|^p(T)\right\rangle &= \sum_{n=-\infty}^{+\infty} (|n|+1)^p a(n,T) &(27)\\ &\geq \sum_{|n|>(\ln T)^{\alpha}-2} (|n|+1)^p a(n,T) &(28)\\ &\geq C(\ln T)^{\alpha p} P((\ln T)^{\alpha} - 2, T) &(29)\\ &\geq C(\ln T)^{\alpha p}(\ln T)^{-S_{\ln}^{-}(\alpha)-} &(30)\\ &= C(\ln T)^{\alpha p - S_{\ln}^{-}(\alpha)-} &(31)\end{aligned}$$

and thus

$$\beta_{\ln}^{-}(p) \geq \alpha - \frac{S_{\ln}^{-}(\alpha)}{p}. \tag{32}$$

A similar analysis for $S_{\ln}^{+}(\alpha) < +\infty$ shows

$$\beta_{\ln}^{+}(p) \geq \alpha - \frac{S_{\ln}^{+}(\alpha)}{p}. \tag{33}$$

Together, this shows that

$$\beta_{\ln}^{\pm}(+\infty) \geq \alpha_{\ln}^{\pm}. \tag{34}$$

On the other hand, it is possible to use $\alpha_{\ln}^{\pm}$ to bound $\beta_{\ln}^{\pm}(+\infty)$ from above:

Theorem 3.1 *Let H be an operator of the form (1) with bounded potential and suppose that for some $\eta > 0$, and for all $p > 0$, we have*

$$\left\langle |X|^p(T)\right\rangle < C_p(\ln T)^{\eta p}. \tag{35}$$

Then $0 \leq \alpha_{\ln}^{\pm} \leq \eta$ and

$$\beta_{\ln}^{\pm}(+\infty) \leq \alpha_{\ln}^{\pm}. \tag{36}$$

Remark 11 We can replace (35) with the condition $\beta^+_{\ln}(p) < \eta$ for every $p > 0$.

Remark 12 The following proof uses the same ideas as the proof of Theorem 4.1 in [9].

Proof The bound $0 \le \alpha^{\pm}_{\ln} \le \eta$ follows from the computation performed above, so we will focus on proving (36).

Fix $0 \le \alpha \le \alpha^+_{\ln}$, $\epsilon > 0$ and consider the following:

$$\begin{aligned}\left\langle |X|^p(T)\right\rangle &= \sum_{n=-\infty}^{+\infty} (|n|+1)^p a(n,T) &\quad(37)\\ &= \sum_{|n|\le(\ln T)^\alpha - 2} + \sum_{(\ln T)^\alpha-2<|n|\le(\ln T)^{\alpha^+_{\ln}+\epsilon/2}} &\quad(38)\\ &\quad + \sum_{(\ln T)^{\alpha^+_{\ln}+\epsilon/2}<|n|\le(\ln T)^{\eta+\epsilon}} + \sum_{(\ln T)^{\eta+\epsilon}<|n|} . &\quad(39)\end{aligned}$$

Let us label these sums 1–4. A few notes before we start bounding these sums. First, we will assume $\alpha > 0$. If $\alpha = 0$, then we may proceed by removing the second sum and replacing α with $\alpha^+_{\ln}$ in the first sum. Second, if $\alpha^+_{\ln} = \eta$, then the third sum is unnecessary.

We can bound sum 1 by

$$\sum_{|n|\le(\ln T)^\alpha-2} < C(\ln T)^{\alpha p}.$$

We can bound sum 2:

$$\sum_{(\ln T)^\alpha-2<|n|\le(\ln T)^{\alpha^+_{\ln}+\epsilon/2}} \le C(\ln T)^{p\alpha^+_{\ln}+p\epsilon/2} P((\ln T)^\alpha - 2, T).$$

If $\alpha^+_{\ln} = \eta$, then sum 3 is unnecessary. If $\alpha^+_{\ln} < \eta$, then we can bound sum 3 by

$$\sum_{(\ln T)^{\alpha^+_{\ln}+\epsilon/2}<|n|\le(\ln T)^{\eta+\epsilon}} \le (\ln T)^{\eta p+p\epsilon} P((\ln T)^{\alpha^+_{\ln}+\epsilon/2}, T),$$

and by definition of $\alpha^+_{\ln}$, the right-hand side goes to 0, so it can be further bounded by some constant C.

Finally, we have the bound for sum 4. For any m,

$$\sum_{(\ln T)^{\eta+\epsilon}<|n|} \le (\ln T)^{-(\eta+\epsilon)m} \left\langle |X|^{p+m}(T)\right\rangle$$

$$\leq C_{p+m}(\ln T)^{-(\eta+\epsilon)m}(\ln T)^{\eta(p+m)}.$$

By taking $m > \eta p/\epsilon$, we have

$$\sum_{(\ln T)^{\eta+\epsilon}<|n|} < C.$$

Putting everything together, we have

$$\left\langle |X|^p(T)\right\rangle < C + C(\ln T)^{p\alpha} + C(\ln T)^{p\alpha_{\ln}^{+}+p\epsilon/2}P((\ln T)^{\alpha}-2,T). \tag{40}$$

Taking ln throughout, and letting

$$f(T,p,\alpha,\epsilon) = \max\left\{\alpha p \ln\ln(T), (p\alpha_{\ln}^{+} + \frac{p\epsilon}{2})\ln\ln(T) + \ln(P((\ln T)^{\alpha}-2,T))\right\},$$

we have

$$\ln\left(\left\langle |X|^p(T)\right\rangle\right) < C + f(T,p,\alpha,\epsilon) \tag{41}$$

so

$$\beta_{\ln}^{+}(p) \leq \max\left\{\alpha, \alpha_{\ln}^{+} + \frac{\epsilon}{2} - \frac{S_{\ln}^{+}(\alpha)}{p}\right\}. \tag{42}$$

Taking $p \to \infty$ yields our result for $\beta_{\ln}^{+}(p)$. The proof for $\beta_{\ln}^{-}(p)$ is similar.

□

The major roadblock to using this result to obtain bounds on $\beta_{\ln}^{\pm}(p)$ is that it requires an a priori finite estimate on $\beta_{\ln}^{\pm}(p)$ for every $p > 0$, which we do not have in general. This differs from the situation arising when we merely want to bound $\beta^{\pm}(p)$, since in that case we usually have a trivial ballistic upper bound: $\beta^{\pm}(p) \leq 1$. To remedy this, we have the following, which provides a sufficient condition for $\beta^{\pm}(p) < C < \infty$ for every $p > 0$.

Theorem 3.2 *Let H be an operator of the form* (1) *with bounded potential and suppose that $\alpha_{\ln}^{\pm} < +\infty$. Moreover, suppose that, for some $\xi > 0$,*

$$P((\ln T)^{\xi}, T) = O(T^{-a}) \tag{43}$$

for every $a > 1$, and for some $\gamma < \infty$, we have

$$\left\langle |X|^p(T)\right\rangle < C_p T^{\gamma p}. \tag{44}$$

Then for some $\eta < \infty$ (35) *holds.*

Remark 13 As noted above, (44) always holds with $\gamma = 1$ when the potential is bounded.

Proof The proof proceeds the same as before, expressing $\langle |X|^p(T)\rangle$ as a sum and decomposing that sum into four further sums, except we take η to be ξ. With this modification, the bounds for sums 1–3 still hold, but we need to be more careful with the fourth sum.

We have:

$$\sum_{(\ln T)^{\xi+\epsilon}<|n|} = \sum_{(\ln T)^{\xi+\epsilon}<|n|\leq T^{\gamma+\epsilon}} + \sum_{T^{\gamma+\epsilon}<|n|} . \tag{45}$$

Let us denote the first sum by I and the second sum by II. We can bound sum I by

$$\begin{aligned}\sum_{(\ln T)^{\xi+\epsilon}<|n|\leq T^{\gamma+\epsilon}} &\leq T^{(\gamma+\epsilon)p} P((\ln T)^{\xi+\epsilon}, T) &(46)\\ &\leq T^{p(\gamma+\epsilon)-a} &(47)\end{aligned}$$

for large T, where we can take any $a > 1$. Taking $a > p(\gamma+\epsilon)$, we have $\sum_{(\ln T)^{\xi+\epsilon}<|n|\leq T^{\gamma+\epsilon}} < C$. For sum II, we have

$$\begin{aligned}\sum_{T^{\gamma+\epsilon}<|n|} &= T^{-m(\gamma+\epsilon)} \sum_{T^{\gamma+\epsilon}<|n|} (|n|+1)^{p+m} a(n,T) &(48)\\ &\leq T^{-m(\gamma+\epsilon)} \left\langle |X|^{p+m}(T)\right\rangle &(49)\\ &\leq C_{m+p} T^{(p+m)\gamma - m(\gamma+\epsilon)} < C. &(50)\end{aligned}$$

for $m > \gamma p/\epsilon$. With these two bounds, we may proceed as before to conclude that $\beta^+_{\ln}(p) < C < +\infty$. □

We will now turn our attention to the proof of Theorem 1.5. We start with a lemma due to Damanik and Tcheremchantsev:

Lemma 3.1 ([8] Theorem 7) *Suppose H is of the form* (1)*, where V is a bounded real-valued function, and $K \geq 4$ is such that $\sigma(H) \subset [-K+1, K-1]$. Then*

$$P_r(N,T) \lesssim e^{-cN} + T^3 \int_{-K}^{K} \left(\max_{1\leq n\leq N} \left\| A_n^{f,E+i/T} \right\|^2 \right)^{-1} dE \tag{51}$$

$$P_l(N,T) \lesssim e^{-cN} + T^3 \int_{-K}^{K} \left(\max_{1\leq n\leq N} \left\| A_{-n}^{f,E+i/T} \right\|^2 \right)^{-1} dE \tag{52}$$

With this lemma, and the preceding theorems, we will prove Theorem 1.5.

Proof of Theorem 1.5(a) In light of Theorem 3.1, it suffices to show that $\alpha_{\ln}^{\pm} \leq \gamma$. We will do this for $\alpha_{\ln}^{+}$ and observe that the proof for $\alpha_{\ln}^{-}$ is the same.

Using (9) and Lemma 3.1, since $\gamma > 1$, we have

$$P((\ln T)^{\gamma}, T) = O(T^{-\delta}) \tag{53}$$

for every $\delta < \infty$. Thus

$$\frac{\ln\left(P((\ln T)^{\gamma}, T)\right)}{\ln\ln(T)} \leq \frac{-\delta \ln(T)}{\ln\ln(T)}. \tag{54}$$

We are left with

$$S_{\ln}^{+}(\gamma) = +\infty, \tag{55}$$

so $\alpha_{\ln}^{+} \leq \gamma$. □

We will now prove the second part.

Proof of Theorem 1.5(b) Fix H_1 and H_2 of the form (1) with bounded potentials, v_1 and v_2, and let $K \geq 4$ be such that $\sigma(H_i) \subset [-K+1, K-1]$ for $i = 1, 2$. Denote the corresponding transfer matrices by A^{v_1} and A^{v_2} and the corresponding transport exponents by $\beta_{\ln,1}^{\pm}(p)$, $\beta_{\ln,2}^{\pm}(p)$. Suppose that there is $\gamma < \infty$ such that for every $M > 0$ and $T > T_0(M)$,

$$\int_{-K}^{K} \left(\max_{0 \leq |n| \leq (\ln T)^{\gamma}} \left|\left|A_n^{v_1}(x, E + i/T)\right|\right|^2 \right)^{-1} dE \leq CT^{-M}.$$

Moreover, suppose that there exists $A > 0$ such that for all $E \in [-K+1, K-1]$, $0 < \epsilon \leq 1$, and $|n| \leq \ln(\epsilon^{-1})$,

$$\epsilon^{A} \left|\left|A_n^{v_1, E+i\epsilon}\right|\right| \lesssim \left|\left|A_n^{v_2, E+i\epsilon}\right|\right| \lesssim \epsilon^{-A} \left|\left|A_n^{v_1, E+i\epsilon}\right|\right|. \tag{56}$$

Let $P_1(N, T)$ and $P_2(N, T)$ be the corresponding outside probabilities.

Observe, by Lemma 3.1 and our assumptions above, that for any $M > 0$, and $T > T_0(M)$,

$$P_2((\ln T)^{\gamma}, T) \leq e^{-C(\ln T)^{\gamma}} + T^3 \int \int_{-K}^{K} \left(\max_{0 \leq |n| \leq (\ln T)^{\gamma}} \left|\left|A_n^{v_2}(x, E + i/T)\right|\right|^2 \right)^{-1} dE \tag{57}$$

$$\leq e^{-C(\ln T)^{\gamma}} + T^{3+A} \int \int_{-K}^{K} \left(\max_{0 \leq |n| \leq (\ln T)^{\gamma}} \left|\left|A_n^{v_1}(x, E + i/T)\right|\right|^2 \right)^{-1} dE \tag{58}$$

$$\leq CT^{-M}, \tag{59}$$

and thus

$$\frac{\ln(P_2((\ln T)^\gamma, T))}{\ln\ln(T)} \leq \frac{-M\ln(T)+\ln(C)}{\ln\ln(T)}. \tag{60}$$

We conclude as before.

□

4 Semialgebraic Sets

Here we obtain an explicit estimate on the δ from Theorem 2.2.

Theorem 4.1 *When T_ω is the shift on $\mathbb{T}^n$, and $\omega \in DC(A,c)$, we can take $\delta \leq \frac{1}{A+n}$ in Theorem 2.2. When T_ω is the skew-shift on $\mathbb{T}^n$, and $\omega \in SDC(A,c)$, we can take $\delta < \frac{1}{n2^{n-1}(1+\epsilon)}$ for any $\epsilon > 0$.*

Remark 14 The general idea of the proof is the same in both cases. We first prove a bound of the form $\#\{k=1,...,N : T_\omega(x_0) \in B_\epsilon\} \leq N^{-\zeta}$, where B_ϵ is a ball of radius ϵ. Then we use the covering lemma for semialgebraic sets (Theorem 2.1) to cover the desired semialgebraic set by ϵ-balls. Because of this similarity, we will only give a proof for the shift. The details for the skew-shift can be found in [22] (Lemma 8.4 and Theorem 8.7).

Proof Fix $\epsilon = N^{-\delta}$ and let $\chi(x) = \chi_{B(0,\epsilon)}(x)$ be the characteristic function of the ball of radius ϵ centered at 0. Let $R = \frac{1}{10\epsilon}$ and let

$$F_R(x_j) = \frac{1}{R}\left(\frac{\sin(Rx/2)}{\sin(x/2)}\right)^2 = \sum_{|m|<R}\left(1-\frac{|m|}{R}\right)e^{imx_j} = \sum_{|m|<R}\widehat{F_R}(m)e^{imx_j}$$

be the usual Fejer kernel on $\mathbb{R}$.

If $\chi(x)=0$, then $\chi(x) \leq CR^{-n}\prod_{j=1}^n F_R(x_j)$ holds trivially. On the other hand, by our choice of ϵ and R, if $\chi(x)=1$, then $F_R(x_j) \sim R$, since, for small x_j,

$$F_R(x_j) = \frac{1}{R}\left(\frac{\sin(Rx_j/2)}{\sin(x_j/2)}\right)^2 \sim \frac{1}{R}R^2 = R,$$

and we also have $\chi(x) \leq CR^{-n}\prod_{j=1}^n F_R(x_j)$. Thus we have

$$\begin{aligned}\prod_{j=1}^n F_R(x_j) &= \prod_{j=1}^n \sum_{|m|<R}\widehat{F_R}(m)e^{imx_j} \\ &= \sum_{|m|<R}\widehat{F_R}(m_1)\cdots\widehat{F_R}(m_n)e^{im\cdot x}.\end{aligned} \tag{61}$$

Hence, if we set $m = (m_1, ..., m_n)$, we have

$$\sum_{j=1}^{N} \chi(x_0 + j\omega) \leq \quad CR^{-n} \sum_{j=1}^{N} \sum_{|m_k|<R; 1\leq k\leq n} \widehat{F_R}(m_1) \cdots \widehat{F_R}(m_n) e^{im\cdot(x_0+j\omega)} \tag{62}$$

$$\leq \quad CR^{-n} \sum_{|m_k|<R; 1\leq k\leq n} \left(\widehat{F_R}(m_1) \cdots \widehat{F_R}(m_n) e^{im\cdot x} \left(\sum_{j=1}^{N} e^{ijm\cdot\omega} \right)\right) \tag{63}$$

$$\leq \quad CR^{-n} \sum_{|m_k|<R; 1\leq k\leq n} \left(\widehat{F_R}(m_1) \cdots \widehat{F_R}(m_n) \left| \sum_{j=1}^{N} e^{ijm\cdot\omega} \right| \right). \tag{64}$$

At this point, we can split the sum into two parts: either $m_k = 0$ for all $1 \leq k \leq n$ or at least one $m_k \neq 0$. Thus we can write (64) = (65) + (66), where (65) and (66) are given by

$$CR^{-n} \widehat{F_R}(0)^n \left| \sum_{j=1}^{N} e^{ij0\cdot\omega} \right| \tag{65}$$

and

$$CR^{-n} \sum_{0\leq|m_k|<R; 1\leq k\leq n; \text{ some } m_k\neq 0} \left(\widehat{F_R}(m_1) \cdots \widehat{F_R}(m_n) \left| \sum_{j=1}^{N} e^{ijm\cdot\omega} \right| \right). \tag{66}$$

Since $0 < \widehat{F_R}(m) \leq 1$ and $\left| \sum_{j=1}^{N} e^{ijm\cdot\omega} \right| \leq N$, we have for any x_0

$$\begin{aligned}
\sum_{j=1}^{N} \chi(x_0 + j\omega) &\leq CR^{-n} N + CR^{-n} \sum_{0<|m|<R} \left| \sum_{j=1}^{N} e^{ijm\cdot\omega} \right| \\
&= CR^{-n} N + CR^{-n} \sum_{0<|m|<R} \left| \frac{1 - e^{iNm\cdot\omega}}{1 - e^{im\cdot\omega}} \right| \\
&\leq CR^{-n} N + CR^{-n} \sum_{0<|m|<R} 2|1 - e^{im\cdot\omega}|^{-1} \\
&\leq CR^{-n} N + C \max_{0<|m|<R} 2|1 - e^{im\cdot\omega}|^{-1}.
\end{aligned}$$

Since $\omega \in DC(c, A)$, we know $||m \cdot \omega|| > c|m|^{-A}$, for every $m \neq 0$, so $|1 - e^{im\cdot\omega}|^{-1} \lesssim R^A$, and we conclude

$$\begin{aligned}\sum_{j=1}^{N} \chi(x_0 + j\omega) &\leq CR^{-n}N + CR^{A}\\ &\leq CN(R^{-n} + R^{A}N^{-1})\\ &\leq CN(\epsilon^{n} + \epsilon^{-A}N^{-1}).\end{aligned}$$

Now, if we take $\delta = \frac{1}{n+A}$, then by our choice of ϵ, we have

$$\begin{aligned}\epsilon^{-A}N^{-1} &= \epsilon^{-A}\epsilon^{A+n}\\ &= \epsilon^{n},\end{aligned}$$

so

$$\sum_{j=1}^{N} \chi(x_0 + j\omega) \leq CN\epsilon^{n}.$$

We conclude the proof by observing that, by Theorem 2.1, it is possible to cover $\mathcal{S}$ using no more than $B^{C}\epsilon^{1-n}$ ϵ-balls, where $C = C(n)$. Thus the above computation shows that

$$\begin{aligned}\#\{k = 1, ..., N : x_0 + k\omega \in \mathcal{S}\} &\leq CN\epsilon^{n}B^{C}\epsilon^{1-n}\\ &= CNB^{C}\epsilon\\ &\leq N^{1-\delta}B^{C}.\end{aligned}$$

For the skew-shift, we have, by Lemma 8.3 and Theorem 8.7 from [22], that for any $\epsilon' > 0$,

$$\#\left\{k = 1, ..., N : T_{\omega}^{k}(x_0) \in B_{\epsilon}\right\} \leq CN^{-\frac{1}{2^{n-1}(1+\epsilon)}+\epsilon'}.$$

Applying Theorem 2.1, we have

$$\#\left\{k = 1, ..., N : T_{\omega}^{k}(x_0) \in \mathcal{S}\right\} \leq CB^{C}\epsilon^{1-n}N^{-\frac{1}{2^{n-1}(1+\epsilon)}+\epsilon'}$$

□

5 Technical Lemmas

We will prove our results for right cocycles and observe that the exact same arguments establish the same results for left cocycles.

Let us define

$$V_k^f(E,a) := \left\{ x \in \mathbb{T}^\nu : \frac{1}{k} \ln \left\| A_k^{f,E}(x) \right\| \geq a \right\}.$$

We will begin with the following lemma, which reduces everything to the study of semialgebraic sets. Fix $\tau < 1$ and $1 - \tau/16 > a > c > d > 1 - \tau/8 > 1 - \tau$.

Lemma 5.1 *Let $f \in G^\sigma(\mathbb{T}^\nu)$. There is some $k_\tau(E) < \infty$ so that for $k > k_\tau(E)$ and $|E - z| < e^{-\frac{k\tau L(E)}{\|f\|_\infty}}$, we can find $N_1 < \infty$ so that we have the following sequence of inclusions:*

$$V_k^f(E, aL(E)) \subset V_k^{\tilde{f}_{N_1}}(E, cL(E)) \subset V_k^f(z, dL(E)) \tag{67}$$

where $\tilde{f}_{N_1}(x)$ is a certain polynomial of degree N_1, so $V_k^{\tilde{f}_{N_1}}(E, cL(E))$ is semialgebraic of degree at most kN_1.

Remark 15 We may take $N_1(k) \sim k^{\sigma\nu+}$ in the above lemma.

Proof Let us fix $k \in \mathbb{N}$ large and $\epsilon > 0$ small. First, since $f \in G^\sigma(\mathbb{T}^\nu)$, we know that

$$|\hat{f}(n)| \leq C_1 e^{-|n|^{1/(\sigma+)}}. \tag{68}$$

Let $f_{N_0}(x) = \sum_{|n| \leq N_0} \hat{f}(n) e^{in \cdot x}$. For $N_0 \geq k^{\sigma+\epsilon}$, we have

$$|f(x) - f_{N_0}(x)| \leq e^{-k^{1+\epsilon}} \leq e^{-k(1-c)L(E)}.$$

Now for such N_0, there exists a polynomial $\tilde{f}_{N_1}(x)$ of degree N_1 with $N_1 = k^{\sigma\nu+\epsilon}$ so that

$$|f_{N_0}(x) - \tilde{f}_{N_1}(x)| \leq e^{-k(1-d)L(E)}.$$

This can be seen by approximating $e^{in_j x_j}$ by a Taylor polynomial of degree $k^{\sigma+}$ and then bounding the error as usual. Note that these two inequalities hold for k sufficiently large (dependent only on the dimension ν and ϵ).

By upper semicontinuity, compactness considerations, and a standard telescoping argument, we have

$$\left\| A_k^{f,E}(x) - A_k^{f_{N_0},E}(x) \right\| < e^{-k^{1+\epsilon}} \tag{69}$$

$$\left|\left|A_k^{f,E}(x) - A_k^{\tilde{f}_{N_1}(x),z}\right|\right| < e^{-k(1-d+\tau)L(E)}e^{k(L(E)+\epsilon)} < e^{k(L(E)/2+\epsilon)} \tag{70}$$

for k sufficiently large and $|E - z| < e^{-\frac{k\tau(L(E)+\epsilon)}{||f||_\infty}}$. The first inclusion can now be established by observing that, for $x \in V_k^f(E, aL(E))$, we have

$$\begin{aligned}\left|\left|A_k^{f_{N_0},E}(x)\right|\right| &\geq \left|\left|A_k^{f,E}(x)\right|\right| - \left|\left|A_k^{f,E}(x) - A_k^{f_{N_0},E}(x)\right|\right| \\ &\geq e^{ckL(E)}.\end{aligned}$$

The other inclusion is proved in the same way.

The semialgebraic bound on $V_k^{\tilde{f}_{N_1}}(E, cL(E))$ follows from the fact that $V_k^{\tilde{f}_{N_1}}(E, cL(E))$ is given by a single inequality involving a polynomial of degree kN_1. □

Now we have

Lemma 5.2 *Let* $k, E, z, d,$ *and* $V_k^f(z, dL(E))$ *be as in Lemma 5.1. Then* $|V_k^f(z, dL(E))| > 1/2$, *where* $|\cdot|$ *represents Lebesgue measure.*

Proof By definition of $L(E)$, we have

$$\begin{aligned}L(E) &\leq \frac{1}{k}\int \ln\left|\left|A_k^{f,E}(x)\right|\right| dx \\ &\leq |V_k^f(E, aL(E))|(L(E)+\epsilon) + (1 - |V_k^f(E, aL(E))|)(aL(E)) \\ &\leq |V_k^f(E, aL(E))|((1-a)L(E)+\epsilon) + aL(E).\end{aligned}$$

Thus, by choosing ϵ appropriately (which can be done by upper semicontinuity and taking $k > k_0(\epsilon)$ sufficiently large), and the fact that $a < 1$, we have

$$|V_k^f(E, aL(E))| \geq \frac{1}{2}. \tag{71}$$

The set inclusion proved above now yields the result. □

Our next goal is to show that for T_ω either the shift or skew-shift, there is some $N_k < \infty$ such that for every $x \in \mathbb{T}^\nu$, $T_\omega(x) \in V_k^f(z, dL(E))$ for some $1 \leq j \leq N_k$ and then obtain the required transfer matrix bounds. We will split the remaining argument up into three cases: the shift with $\nu = 1$, the shift with $\nu > 1$,, and the skew-shift with $\nu > 1$.

6 The Case $\nu = 1$

Our goal is to first establish the following estimates. Let d be as in Lemma 5.1.

Theorem 6.1 *Let $f \in G^{\sigma}(\mathbb{T})$, $\omega \in \mathbb{R}\backslash\mathbb{Q}$, and $E \in \mathbb{C}$ such that $L(E) > 0$. For any $0 < \tau < 1$, there exist $k_\tau = k_\tau(E) < \infty$ such that for any $\epsilon > 0, k > k_\tau$, and $x \in \mathbb{T}$, there is $1 \leq j \leq Ck^{1+\sigma+\epsilon}$ so that for any $z \in \mathbb{C}$ with $|z - E| < e^{-\frac{\tau k L(E)}{\|f\|_\infty}}$, we have*

$$\left|\left|A_k^{f,z}(x + j\omega)\right|\right|^2 > e^{dkL(E)}. \tag{72}$$

Theorem 6.2 *Fix $\epsilon > 0$. Let $f \in G^{\sigma}(\mathbb{T})$, $\omega \in DC(A, c)$, and $L(E) > 0$. Then for any $\xi, \zeta > 1$, there is $C, c > 0$ and $T_E < \infty$ such that for $T > T_E$,*

$$\inf\left\{\min_{\iota=\pm 1} \max_{1\leq \iota m \leq C(\ln T)^{\zeta(1+\sigma+\epsilon)}} \left|\left|A_m^{f,z}(x)\right|\right|^2 T^{-\xi}\right\} > c \tag{73}$$

where the infimum is over all $x \in \mathbb{T}$ and $z \in \mathbb{C}$ with $|z - E| < T^{-\zeta}$. Moreover, T_E is uniformly bounded below for E in compact sets with positive $L(E)$.

In particular, for $E \in [-K, K]$, we have $\max_{1\leq n\leq C(\ln T)^{\zeta(1+\sigma)}} \left|\left|A_n^{f,E+i/T}\right|\right|^2 \geq cT^{\xi}$ for every $\xi > 1$ and large T.

If $\omega \in \mathbb{R}\backslash\mathbb{Q}$, then the above holds for a sequence, T_n for $n > n_E$ for all E, and for $n > n_0$ for $E \in [-K, K]$.

When $\nu = 1$, we can write ω as a continued fraction. Let $\frac{p_n}{q_n}$ be the continued fraction approximation of ω. We then have the following lemma.

Lemma 6.1 (Lemma 9 from [12]) *Suppose $\Delta \subset \mathbb{T}$ is an interval with $|\Delta| > 1/q_n$. Then for every $x \in \mathbb{T}$, there exists $1 \leq j \leq q_n + q_{n-1} - 1$ such that $x + j\omega \in \Delta$.*

Lemmas 5.1 and 5.2, along with Remark 15, imply $V_k^f(z, dL(E))$ contains an open set, Δ, of measure

$$\frac{1}{2k^{1+\sigma+\epsilon}} \lesssim |\Delta|.$$

Now if we take $k > Cq_n^{1/(1+\sigma+\epsilon)}$, we have $|\Delta| > 1/q_n$, and so, by Lemma 6.1,

Lemma 6.2 *Let f, E, z, and d be as in Lemma 5.1. For $k \sim q_n^{1/(1+\sigma+\epsilon)}$, there exists $1 \leq j \lesssim k^{1+\sigma+\epsilon}$ such that $x + j\omega \in V_k^f(z, dL(E))$.*

Theorem 6.1 now follows by the set inclusion we proved in the previous section.

Since the proof of Theorem 6.2 is identical to the proof of Theorem 7.2 in the next section, we omit it and refer readers to the next section for the details.

With Theorem 6.2, we can prove Theorem 1.1.

Proof of Theorem 1.1 Let us begin by fixing $x \in \mathbb{T}$ and $f \in G^{\sigma}(\mathbb{T})$. Moreover suppose that $L(E) > 0$ for every $E \in \mathbb{R}$. First, we will consider the case $\omega \in DC(A, c)$. Fix $\epsilon > 0$ and set $\gamma = 1 + \sigma$. The hypotheses of Theorem 6.2 are satisfied, and we can combine the conclusion of Theorem 6.2 with the conclusion of Lemma 3.1 to obtain

$$P((\ln T)^{\gamma+\epsilon} - 2, T) \le e^{-C(\ln T)^{\zeta(\gamma+\epsilon)}} + CT^{-\delta}$$

for every $\zeta, \delta > 1$. Since $\gamma > 1$, we can further bound this by

$$P((\ln T)^{\gamma+\epsilon} - 2, T) \le CT^{-\delta},$$

using a different constant C. As before, we obtain $\alpha_{\ln}^{+} \le 1 + \sigma < +\infty$.

We can now appeal to Theorem 3.2 to establish the hypotheses of Theorem 3.1, so $\beta_{\ln}^{+}(p) \le \alpha_{\ln}^{+} \le 1 + \sigma$.

Now we turn to the case $\omega \in \mathbb{R}\backslash\mathbb{Q}$. We can appeal to Theorem 6.2 to obtain the above for a sequence $T_n \to \infty$. With a sequence, we have analogous statements as above, but for S^{-} and α^{-}. Thus we obtain $\beta_{\ln}^{-}(p) \le 1 + \sigma$. □

7 The Case $\nu > 1$

As in the case $\nu = 1$, our goal is to first establish the following estimates:

Theorem 7.1 *Let $f = \lambda f_0 \in G^{\sigma}(T^{\nu})$, $\nu > 1$, $\omega \in DC(A, c)$, $\lambda > \lambda_0(f_0, \omega)$, and $E \in \mathbb{R}$ such that $L(E) > 0$. For any $0 < \tau < 1$, there exist $k_\tau = k_\tau(E) < \infty$, $\delta = \delta(\omega, \nu)$, and $\gamma = \gamma(\sigma, \nu, \delta)$ such that for any $\epsilon > 0$, $k > k_\tau$, and $x \in \mathbb{T}^{\nu}$, there is $1 \le j \le k^{\gamma+\epsilon}$ so that for any $z \in \mathbb{C}$ with $|z - E| < e^{-\frac{\tau k L(E)}{\|f\|_\infty}}$, we have*

$$\left\|A_k^{f,z}(x + j\omega)\right\| > e^{k(1-\tau)L(E)}. \tag{74}$$

Theorem 7.2 *Fix $\epsilon > 0$. Let $f = \lambda f_0 \in G^{\sigma}(\mathbb{T}^{\nu})$, $\nu > 1$, $\omega \in DC(c, A)$, $\lambda > \lambda_0(f_0, \omega)$, and $L(E) > 0$. Then for any $\xi, \zeta > 1$, there is $c > 0$ and $T_E < \infty$ such that for $T > T_E$,*

$$\inf\left\{\min_{\iota=\pm 1} \max_{1 \le \iota m \le (\ln T)^{\zeta(\gamma+\epsilon)}} \left\|A_m^{f,z}(x)\right\|^2 T^{-\xi}\right\} > c \tag{75}$$

where γ and δ are as above, and the infimum is over all $x \in \mathbb{T}^{\nu}$ and $z \in \mathbb{C}$ with $|z - E| < T^{-\zeta}$. Moreover, the dependence of T_E on E is through $L(E)$, as in Theorem 6.2. Thus, as before, T_E is uniformly bounded below for E in compact sets with positive $L(E)$.

Remark 16 If we consider just $E \in [-K, K]$ in the above theorem, then continuity of $L(E)$, which was established for our situation in [19], and compactness of $[-K, K]$ yield the desired uniform lower bound on T.

When $\nu > 1$, we need to do a bit more work to obtain an analogue of Lemma 6.1. We may appeal to Theorems 2.4 and 2.2 to obtain:

Lemma 7.1 *Let $\omega \in DC(A, c)$. For $f = \lambda f_0 \in G^\sigma(\mathbb{T}^\nu)$, there exists $\lambda_0(f_0, \omega)$ such that for $\lambda > \lambda_0$ and every $x \in \mathbb{T}^\nu$, there exists $1 \le j \le k^{C(\nu+A)(\sigma\nu+1)+}$ such that $x + j\omega \in V_k^{\tilde{f}_{N_1}}(E, cL(E))$.*

Proof Recall that by Theorem 2.4, combined with (69), with N_1 as in Lemma 5.1, there exists a λ_0 so that for all $\lambda > \lambda_0$ and $f = \lambda f_0$, we have

$$\left|\left\{x \in \mathbb{T}^\nu : \left|\frac{1}{k}\ln\left\|A_k^{\tilde{f}_{N_1},E}(x)\right\| - L_k(E)\right| > 2k^{-\tau}\right\}\right| < e^{-k^\alpha}. \tag{76}$$

This implies

$$\left|\left\{x \in \mathbb{T}^\nu : \frac{1}{k}\ln\left\|A_k^{\tilde{f}_{N_1},E}(x)\right\| - L(E) < -2k^{-\tau}\right\}\right| < e^{-k^\alpha}, \tag{77}$$

since $L_k(E) \ge L(E)$. Thus, for k sufficiently large, and $N_1(k) \sim k^{\sigma\nu+}$, by Remark 15,

$$\left|\mathbb{T}^\nu \backslash V_k^{\tilde{f}_{N_1}}(E, cL(E))\right| < e^{-k^\alpha}. \tag{78}$$

Since the left-hand side is the complement of a semialgebraic set of degree at most kN_1, it is itself semialgebraic of degree at most kN_1. By Theorem 4.1, for fixed $0 < \epsilon < \delta = \frac{1}{\nu+A}$, we can thus set $\mathcal{S} = \left(\mathbb{T}^\nu \backslash V_k^{\tilde{f}_{N_1}}(E, cL(E))\right)$, $\eta = e^{-k^\alpha}$, $B = kN_1$, and $N = B^{C/(\delta-\epsilon)}$ and then appeal to Theorem 2.2 to obtain, for any $0 < \epsilon < \delta$,

$$\#\{1 \le j \le N : x + j\omega \in \mathcal{S}\} < B^{C\frac{1-\delta}{\delta-\epsilon}} B^C = B^{C\frac{1-\epsilon}{\delta-\epsilon}}. \tag{79}$$

Thus, for every $x \in \mathbb{T}^\nu$, there is a $1 \le j \le (kN_1)^{C\frac{1-\epsilon}{\delta-\epsilon}} < N^{1-\epsilon}$ so that $x + j\omega \in V_k^{\tilde{f}_{N_1}}(E, cL(E))$. The result now follows from our choice of $N_1 \sim k^{\sigma\nu+}$ in Lemma 5.1.

□

Theorem 7.1 now follows from the fact that $V_k^{\tilde{f}_{N_1}}(E, cL(E)) \subset V_k^f(z, dL(E))$, and observing that $d > 1 - \tau$, just as in the case $\nu = 1$.

Theorem 7.2 can now be proved using Theorem 7.1.

Proof of Theorem 7.2 Fix $\xi, \zeta > 1$ and $0 < \tau < \frac{\zeta ||f||_\infty}{\zeta ||f||_\infty + \xi} < 1$. Consider any $M_k = M_k(\xi, \zeta)$ such that the following holds:

$$e^{k\tau L(E)/(\zeta ||f||_\infty)} < M_k < e^{k(1-\tau)L(E)/\xi} \tag{80}$$

and

$$(\ln M_k)^{(\gamma+\epsilon)\zeta} > k^{\gamma+} + k. \tag{81}$$

Both conditions can be satisfied by taking k sufficiently large due to our choice of τ and $\zeta > 1$. Appealing to Theorem 7.1, for every $x \in \mathbb{T}^\nu$, there is $1 \le j \le (\ln M_k)^{(\gamma+\epsilon)\zeta} - k$ so that for $|z - E| < M_k^{-\zeta}$, we have

$$\left|\left|A_k^{f,z}(x + j\omega)\right|\right| \ge M_k^{\xi}. \tag{82}$$

Now recall that, by definition,

$$A_{k+j}^{f,z}(x) = A_k^{f,z}(x + j\omega) A_j^{f,z}(x). \tag{83}$$

Moreover, A is an $SL_2(\mathbb{R})$ cocycle, so $||A_k|| = \left|\left|A_k^{-1}\right|\right|$, and thus

$$\left|\left|A_k^{f,z}(x + j\omega)\right|\right| \le \left|\left|A_{k+j}^{f,z}(x)\right|\right| \left|\left|A_j^{f,z}(x)\right|\right|. \tag{84}$$

This together with (82) implies

$$\max_{1 \le j \le (\ln M_k)^{(\gamma+\epsilon)\zeta} - k} \left\{\left|\left|A_{k+j}^{f,z}(x)\right|\right|, \left|\left|A_j^{f,z}(x)\right|\right|\right\} \ge M_k^{\xi}. \tag{85}$$

Thus we must have

$$\max_{1 \le j \le (\ln M_k)^{(\gamma+\epsilon)\zeta}} \left|\left|A_j^{f,z}(x)\right|\right|^2 \ge M_k^{\xi}. \tag{86}$$

It is not difficult to show that for some $T_0 = T_0(E) < \infty$ and any $T > T_0$, we can find $k < \infty$ and $M_k = T$ satisfying (80) and (81). Thus, we have, for any $\xi, \zeta > 1$,

$$\inf_{|z-E|<T^{-\zeta}; x \in \mathbb{T}^\nu} \left\{ \max_{1 \le \iota j \le (\ln T)^{(\gamma+\epsilon)\zeta}} \left|\left|A_j^{f,z}(x)\right|\right|^2 T^{-\xi} \right\} > c > 0. \tag{87}$$

It remains to show that we can also use the same M_k to obtain an analogous bound for the left transfer matrix. Note that for an ergodic invertible cocycle, the

Lyapunov exponent of the forward cocycles and the Lyapunov exponent of the backward cocycles agree. Moreover, if $A_k(\omega, x)$ is the cocycle over rotations by ω, then $A_{-k}(\omega, x) = A_k(-\omega, x + \omega)$. Since ω and $-\omega$ obey the same Diophantine condition, Lemma 7.1 also holds for $A_{-k}^{f,z}(x)$, which means we can use the exact same M_k to obtain a bound as above. □

Now we can turn to the proof of Theorem 1.3.

Proof of Theorem 1.3 We can follow the same idea as in the proof of Theorem 1.1, using Theorem 7.2 in place of Theorem 6.2. Let us fix $x \in \mathbb{T}^\nu$, $\omega \in DC(A, c) \subset \mathbb{T}^\nu$, and $f = \lambda f_0 \in G^\sigma(\mathbb{T}^\nu)$, where $\lambda > \lambda_0(f_0, \omega)$ so that we satisfy the conclusions of Theorem 2.4. Moreover, suppose that $L(E) > 0$ so that we may appeal to Theorem 7.2.

By Theorem 7.2, along with Theorem 3.1, we have

$$P((\ln T)^{\gamma+\epsilon} - 2, T) \leq CT^{-\beta}$$

for some $\gamma = \gamma(A, c, \sigma, \nu) < +\infty$ and every $\beta > 1$. Moreover, it is clear that

$$\frac{\ln(P((\ln T)^{\gamma+\epsilon} - 2, T))}{\ln\ln(T)} \leq -\delta \frac{\ln(T)}{\ln\ln(T)}, \tag{88}$$

so by Theorems 3.2 and 3.1, $\beta_{\ln}^{\pm}(p) \leq \alpha_{\ln}^{\pm} \leq \gamma$. □

8 The Analytic Case

The proofs of our main results in the case of an analytic potential are morally the same as those for Gevrey potentials. Indeed, we can quickly obtain the following using the same proofs as the analogous results above.

Theorem 8.1 *Let f be a non-constant analytic function on $\mathbb{T}^\nu, \nu \geq 1, \omega \in DC(A, c)$, and $E \in \mathbb{R}$ such that $L(E) > 0$. For any $0 < \tau < 1$, there exist $k_\tau = k_\tau(E) < \infty, \delta = \delta(\omega, \nu)$, and $\gamma = \gamma(\nu, \delta)$ such that for any $\epsilon > 0, k > k_\tau$, and $x \in \mathbb{T}^\nu$, there is $1 \leq j \leq k^{\gamma+\epsilon}$ so that for any $z \in \mathbb{C}$ with $|z - E| < e^{-\frac{\tau k L(E)}{\|f\|_\infty}}$, we have*

$$\left|\left|A_k^{f,z}(x + j\omega)\right|\right| > e^{k(1-\tau)L(E)}. \tag{89}$$

Theorem 8.2 *Fix $\epsilon > 0$. Let f be a non-constant analytic function on $\mathbb{T}^\nu, \nu \geq 1, \omega \in DC(c, A)$, and $L(E) > 0$. Then for any $\xi, \zeta > 1$, there is $c > 0$ and $T_E < \infty$ such that for $T > T_E$,*

$$\inf\left\{\min_{\iota=\pm 1}\max_{1\le \iota m\le (\ln T)^{\zeta(\gamma+\epsilon)}}\left\|A_m^{f,z}(x)\right\|^2 T^{-\xi}\right\} > c \tag{90}$$

where γ and δ are as before, and the infimum is over all $x \in \mathbb{T}^\nu$ and $z \in \mathbb{C}$ with $|z - E| < T^{-\zeta}$.

Moreover, the dependence of T_E on E is through $L(E)$, as in Theorem 6.2. Thus, as before, T_E is uniformly bounded below for E in compact sets with positive $L(E)$.

The main difference between these two results and the variants from Sects. 6 and 7 is the assumption on f. Here, we do not need to assume $f = \lambda f_0$ for $\lambda > \lambda_0(f_0, \omega)$. Indeed, this condition is needed for the Gevrey case in order to use the large deviation estimate Theorem 2.4, but the analogous estimate for analytic potentials, Theorem 2.3, does not require such a condition. Once we have a large deviation estimate, the proofs proceed exactly as in the proof of Theorem 7.1, with (68) replaced by $|\hat{f}(n)| \le CE^{c|n|}$. Note that continuity of $L(E)$, which is required in the uniform minoration of T_E, was established in [3].

9 The Skew-Shift Case, $\nu > 1$

Let T_ω denote the skew-shift on $\mathbb{T}^\nu$. As in the shift case, our goal is to first establish the following estimates:

Theorem 9.1 *Let $f = \lambda f_0 \in G^\sigma(T^\nu)$, $\nu > 1$, $\omega \in SDC(A, c)$, $\lambda > \lambda_0(f_0, \omega)$ and $E \in \mathbb{R}$ such that $L(E) > 0$. For any $0 < \tau < 1$, there exist $k_\tau = k_\tau(E) < \infty$, $\delta = \delta(\omega, \nu)$, and $\gamma = \gamma(\sigma, \nu, \omega)$ such that for any $\epsilon > 0$, $k > k_\tau$, and $x \in \mathbb{T}^\nu$, there is $1 \le j \le k^{\gamma+\epsilon}$ so that for any $z \in \mathbb{C}$ with $|z - E| < e^{-\frac{\tau k L(E)}{\|f\|_\infty}}$, we have*

$$\left\|A_k^{f,z}(x + j\omega)\right\| > e^{k(1-\tau)L(E)}. \tag{91}$$

Theorem 9.2 *Fix $\epsilon > 0$. Let $f = \lambda f_0 \in G^\sigma(\mathbb{T}^\nu)$, $\nu > 1$, $\omega \in SDC(c, A)$, $\lambda > \lambda_0(f_0, \omega)$, and $L(E) > 0$. Then for any $\xi, \zeta > 1$, there is $c > 0$ and $T_E < \infty$ such that for $T > T_E$,*

$$\inf\left\{\min_{\iota=\pm 1}\max_{1\le \iota m\le (\ln T)^{\zeta(\gamma+\epsilon)}}\left\|A_m^{f,z}(x)\right\|^2 T^{-\xi}\right\} > c \tag{92}$$

where γ and δ are as above, and the infimum is over all $x \in \mathbb{T}^\nu$ and $z \in \mathbb{C}$ with $|z - E| < T^{-\zeta}$. Moreover, if we restrict our attention to E in some compact interval $[-K, K]$, we can take T_E uniformly bounded below.

In particular, for $E \in [-K, K]$, we have $\max_{1\le n\le (\ln T)^{\zeta(\gamma+\epsilon)}}\left\|A_n^{f,E+i/T}\right\|^2 \ge CT^\xi$ for every $\xi > 1$ and T large.

An analogue of Lemma 6.1 follows using the same argument as in the multifrequency shift case. The proof is identical to the proof of Lemma 9.1, but we use the skew-shift bound from Theorem 2.2 instead of the shift bound.

Lemma 9.1 *Let δ be defined as above. For $f = \lambda f_0 \in G^{\sigma}(\mathbb{T}^{\nu})$, there exists $\lambda_0(f_0, \omega)$ such that for $\lambda > \lambda_0$, every $\epsilon > 0$ and $x \in \mathbb{T}^{\nu}$ there exists $1 \leq j \leq k^{C(1/\delta)(\sigma\nu+1)+\epsilon}$ such that $T_{\omega}(x) \in V_k^{\tilde{f}_{N_1}}(E, cL(E))$.*

Theorem 9.1 now follows from the fact that $V_k^{\tilde{f}_{N_1}}(E, cL(E)) \subset V_k^{f}(z, dL(E))$, and observing that $d > 1 - \tau$, just as in the case $\nu = 1$.

Theorem 9.2 can now be proved using Theorem 9.1 in the same way that Theorem 7.2 was proved using Theorem 7.1.

Proof of Theorem 1.4 We can use the same argument as the proof of Theorem 1.3, using the analogous results from this section rather than those from Sect. 7. □

Acknowledgments We thank Wencai Liu for useful comments on the earlier version of the manuscript. This work was partially supported by NSF DMS-2155211, DMS-2052899, and Simons 681675. S.J. was a 2020-2021 Simons fellow.

References

1. Barbaroux, J.-M., Germinet, F., Tcheremchantsev, S.: Fractal dimensions and the phenomenon of intermittency in quantum dynamics. Duke Math. J. **1**, 161–193 (2001)
2. Bourgain, J., Jitomirskaya, S.: Anderson localization for the band model. In: Geometric Aspects of Functional Analysis. Lecture Notes in Mathematics, vol. 1745, pp. 67–79 (2000)
3. Bourgain, J.: Green's Function Estimates for Lattice Schrödinger Operators and Applications. Princeton University Press, Princeton (2005)
4. Carmona, R., Klein, A., Martinelli, F.: Anderson localization for Bernoulli and other singular potentials. Commun. Math. Phys. **188**, 41–66 (1987)
5. del Rio, R., Jitomirskaya, S., Last, Y., Simon, B.: What is localization. Phys. Rev. Lett **75**, 117 (1995)
6. del Rio, R., Jitomirskaya, S., Last, Y., Simon, B.: Operators with singular continuous spectrum, IV. Hausdorff dimensions, rank one perturbations, and localization. J. Anal. Math. **69**(1), 153–200 (1996)
7. del Rio, R., Makarov, M., Simon, B.: Operators with singular continuous spectrum. II. Rank one operators. Commun. Math. Phys. **165**(1), 59–67 (1994)
8. Damanik, D., Tcheremchantsev, S.: Quantum dynamics via complex analysis methods: general upper bounds without time-averaging and tight lower bounds for the strongly coupled Fibonacci Hamiltonian. J. AMS. **20**(3), 799–827 (2007)
9. Germinet, F., Kiselev, A., Tcheremchantsev, S.: Transfer matrices and transport for Schrödinger operators. Annales de L'Institut Fourier **54**(3), 787–830 (2004)
10. Gordon, A.: The point spectrum of the one-dimensional Schrödinger operator. Uspehi Mat. Nauk **31**, 257–258 (1976)
11. Han, R., Jitomirskaya, S.: Quantum dynamical bounds for ergodic potentials with underlying dynamics of zero topological entropy. Anal. PDE **12**(4), 867–902 (2019)

12. Jitomirskaya, S., Last, Y.: Power-law subordinacy and singular spectra. II. Line operators. Commun. Math. Phys. **211**, 643–658 (2000)
13. Jitomirskaya, S., Liu, W.: Upper bounds on transport exponents for long range operators. J. Math. Phys. **62**(7) Paper No. 073506, 9 (2021)
14. Jitomirskaya, S., Mavi, R.: Dynamical bounds for quasiperiodic schrödinger operators with rough potentials. Int. Math. Res. Notices **2017**(1), 96–120 (2016)
15. Jitomirskaya, S., Simon, B.: Operators with singular continuous spectrum, III. Almost periodic Schrödinger operators. Commun. Math. Phys. **165**, 201–205 (1994)
16. Jitomirskaya, S., Schulz-Baldes, H.: Upper bounds on wavepacket spreading for random Jacobi matrices. Commun. Math. Phys. **273**, 601–618 (2007)
17. Jitomirskaya, S., Schulz-Baldes, H., Stolz, G.: Delocalization in random polymer models. Commun. Math. Phys. **233**(1), 27–48 (2003)
18. Jitomirskaya, S., Zhu, X.: Large deviations of the Lyapunov exponent and localization for the 1D Anderson model. Commun. Math. Phys. **370**(3), 311–324 (2019)
19. Klein, S.: Anderson localization for the discrete one-dimensional quasi-periodic Schrödinger operator with potential defined by a Gevrey-class function. J. Funct. Anal. **218**, 255–292 (2005)
20. Klein, S.: Localization for quasiperiodic Schrödinger operators with multivariable Gevrey potential functions. J. Spectr. Theory **4**(3), 431–484 (2014)
21. Landrigan, M.: Log-dimensional properties of spectral measures. Ph.D. Thesis, UC Irvine, 2001
22. Liu, W.: Quantitative inductive estimates for Green's functions of non-self-adjoin matrices. Analysis and PDE, to appear
23. Landrigan, M., Powell, M.: Fine dimensional properties of spectral measures (2021). arxiv:2107.10883, to appear J. Spectr. Theory 2022
24. Simon, B.: Equilibrium measures and capacities in spectral theory. Inverse Probl. Imaging **1**, 713–772 (2007)

The Slicing Problem by Bourgain

B. Klartag and V. Milman

Dedicated to the memory of Jean Bourgain

Abstract In the context of his work on maximal functions in the 1980s, Jean Bourgain came across the following geometric question: Is there $c > 0$ such that for any dimension n and any convex body $K \subseteq \mathbb{R}^n$ of volume one, there exists a hyperplane H such that the $(n-1)$-dimensional volume of $K \cap H$ is at least c? This innocent and seemingly obvious question (which remains unanswered!) has established a new direction in high-dimensional geometry. It has emerged as an "engine" that inspired the discovery of many deep results and unexpected connections. Here we provide a survey of these developments, including many of Bourgain's results.

Foreword by V. Milman: Some Historical Reminiscences

In August 1984, I visited Jean Bourgain for a couple of days in Brussels where he worked at the time. We intended to spend a year together at IHES, Paris (during the 1984–1985 academic year). Jean was preparing his trip to Leningrad (now St. Petersburg) in September, and I wanted to see him before he left (I had many colleagues and friends there). When he brought me to the train station on my way back to Paris, he proposed the following question: "Let K be a centrally symmetric convex body in $\mathbb{R}^n$; let $Vol(K) = 1$. Does there exist $u \in SL_n$ such that all hyperplane central sections of $u(K)$ will have around the same $(n-1)$-dimensional volume?"

B. Klartag (✉)
Department of Mathematics, Weizmann Institute of Science, Rehovot, Israel
e-mail: boaz.klartag@weizmann.ac.il

V. Milman
School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel
e-mail: milman@tauex.tau.ac.il

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_9

To say more precisely for the non-experts: Is there a universal number C (independent of anything, including the dimension n) such that for every $(n-1)$-dimensional subspace H the following holds:

$$\frac{1}{C}a \leq Vol_{n-1}(u(K) \cap H) \leq Ca? \tag{1}$$

Here C means a universal constant, as usual. Jean added that the question had arisen in his work on maximal functions.

During the train trip back to Paris, I suddenly realized that some of our recent joint observations with Gromov (see Lemma 1 in [21]) on some consequences of the Brunn-Minkowski inequality may lead to the answer (see Lemma 2 in [21]).

I informed Jean about this upon arriving to Paris. This result of Jean on maximal functions was published in 1986 in Amer. J. Math [21]. I should note here that only a few years later did we learn about Hensley's paper [45] where isotropic position and Lemma 2 were already considered for problems of analytic number theory.

However, going back to 1984–1985, Jean asked me a few months later whether the number "a" in (1), which already depended on the body K, is actually uniformly independent of the dimension n, bounded from below. He knew how to prove that it is bounded from above (by its value for the euclidean ball of volume 1). "I don't need this information for my paper" – Jean said – "this number is cancelled in computations; but, I feel I should know it if I need to use it." Jean thought that I may see some geometric point from which it will easily follow, and indeed I thought at first it would be easy: If "a" for some K is extremely small, this means that all central hyperplane sections have a very small volume for a body of volume 1. It looks very counterintuitive; however, it is not yet proved, 35 years later.

The question appears in the "Remark" after Lemma 2 in [21]. Shortly after, Alain Pajor and I produced some advanced study of the isotropic position and this problem [83] and demonstrated some equivalent problems. In the meantime, Jean proved a lower bound of $\frac{1}{n^{0.25}\log n}$ (see [23]), and 20 years later, Boaz Klartag improved upon it to obtain the better lower bound, of $\frac{1}{n^{0.25}}$ by a different approach.

It is surprising and striking how far-reaching and how consequential this problem has become. We will demonstrate this in this survey.

Jean had revisited the aforementioned problem many times, from his 1986 congress talk in Berkeley [22] to his works [24, 26, 27] in later years.

I was told once by Jean that he had spent more time on this problem and had dedicated more efforts to it than to any other problem he had ever worked on. A few months before his passing, Jean wrote to me again, inquiring about any recent progress. He wanted to know the answer before he would leave.

1 Introduction

The classical Busemann-Petty problem, which is closely related to the slicing problem, reads as follows: Let K and T be centrally symmetric convex bodies in $\mathbb{R}^n$ and $Vol_{n-1}(K \cap \theta^\perp) \leq Vol_{n-1}(T \cap \theta^\perp)$ for all $\theta \in S^{n-1} = \{x \in \mathbb{R}^n \, ; \, |x| = 1\}$. Does it follow that $Vol_n K \leq Vol_n T$? Here $\theta^\perp = \{x \in \mathbb{R}^n \, ; \, \langle x, \theta \rangle = 0\}$ is the hyperplane orthogonal to θ. This is Problem 1 in [30], where it is shown that the answer is affirmative when K is an ellipsoid.

For general K and T, the answer to the Busemann-Petty question turned out to be "yes" for dimensions $n \leq 4$. However, surprisingly, the intuition breaks, and for dimensions $n \geq 5$, it does not hold (see the book by Gardner [39] and Koldobsky [64] for history and references). In fact, the intuition in high dimension fails so miserably, and the computations are so difficult that the counterexample in a sufficiently high dimension is simple to describe: Just take K for the cube and T for a Euclidean ball, as shown by K. Ball [5, 7]. Indeed, for $n \geq 10$, $K = [-1/2, 1/2]^n$ and for T, a Euclidean ball of volume $9/10$ centered at the origin in $\mathbb{R}^n$,

$$Vol_{n-1}(K \cap \theta^\perp) \leq \sqrt{2} < 0.9\sqrt{e} \approx \frac{\Gamma\left(\frac{n}{2}+1\right)^{(n-1)/n}}{\Gamma\left(\frac{n+1}{2}\right)} \cdot 0.9^{(n-1)/n} = Vol_{n-1}(T \cap \theta^\perp).$$

In order to overcome this obstruction, a question that looks more sensible to us today is the following:

Question 1.1 *Let $K, T \subseteq \mathbb{R}^n$ be centrally symmetric convex bodies such that $Vol_{n-1}(K \cap \theta^\perp) \leq Vol_{n-1}(T \cap \theta^\perp)$ for all $\theta \in S^{n-1}$. Does it follow that $Vol_n K \leq C \cdot Vol_n T$ for some universal constant C?*

In particular, is it true that there exists a constant $c > 0$ (perhaps, very small) such that for every dimension n and for any convex centrally symmetric body $K \subset \mathbb{R}^n$, if $Vol_{n-1}(K \cap \theta^\perp) < c$ for every $\theta \in S^{n-1}$ then $Vol_n K \leq 1$? This is the essence of the slicing problem, sometimes referred to as the hyperplane conjecture. The assumption of central symmetry is not very essential (see, e.g., [50]), and Question 1.1 is in fact equivalent to the following:

Question 1.2 *Let $K \subset \mathbb{R}^n$ be a convex set of volume one. Does there exist a hyperplane $H \subset \mathbb{R}^n$, such that*

$$Vol_{n-1}(K \cap H) > 1/C$$

for some universal constant $C > 0$, independent of the dimension n?

This is known as *Bourgain's slicing problem.* It is not just a nice riddle; a positive answer would have several consequences in convex geometry. In fact, in some sense, the hyperplane conjecture is the "opening gate" to a better understanding of uniform measures in high dimensions. It is simpler and it is implied by the *thin-shell problem* of Anttila, Ball, and Perisinnaki [2] and by the conjecture of Kannan, Lovász, and Simonovits (KLS) on the *isoperimetric inequality* in convex sets [48], which we discuss below. In fact, the slicing problem appears virtually in any study of the uniform measure on convex sets in high dimension. Here is a sample of entirely equivalent formulations of Question 1.2 mostly taken from [83]. We write $A \simeq B$ if $cA \leq B \leq CA$ for some universal constants $c, C > 0$.

1. Let $K \subset \mathbb{R}^n$ be a convex body (i.e., a non-empty, bounded, open convex set). Does there exist an ellipsoid $\mathcal{E} \subset \mathbb{R}^n$, with $Vol_n \mathcal{E} = Vol_n K$, such that $Vol_n(K \cap C\mathcal{E})/Vol_n(K) \geq 1/2$, where $C > 0$ is a universal constant?
2. Let $K \subset \mathbb{R}^n$ be a convex body. Select $n+2$ independent, random points according to the uniform measure on K. Let $p(K)$ be the probability that these $n+2$ points are the vertices of a convex polytope. Is it true that $(1-p(K))^{1/n} \simeq 1/\sqrt{n}$? This question is known as the *Sylvester problem*.
3. Let $K \subset \mathbb{R}^n$ be a convex body of volume one. Is it true that there exists a volume-preserving, affine map $T : \mathbb{R}^n \to \mathbb{R}^n$, such that

$$Vol_{n-1}(T(K) \cap H) \simeq 1$$

   for any hyperplane through the origin $H \subset \mathbb{R}^n$?
4. Let $K \subset \mathbb{R}^n$ be a convex body. Denote by $\mathrm{Cov}(K)$ the covariance matrix of a random vector that is distributed uniformly in K. In Bourgain's notation, the *isotropic constant* of K is defined as

$$L_K = \frac{\det(\mathrm{Cov}(K))^{\frac{1}{2n}}}{Vol_n(K)^{\frac{1}{n}}}. \tag{2}$$

   The isotropic constant is invariant under invertible, affine transformations. It is known that $(2\pi e)^{-1/2} + o(1) \leq L_K$ for any convex body in $\mathbb{R}^n$ (the minimizer is the Euclidean ball or an ellipsoid). Is it true that $L_K < C$, for some universal constant $C > 0$, independent of the dimension?

However, let us now take a step back. The slicing problem is part of the study of measures in a high-dimensional space. One of the earliest results on probability distributions in high-dimensional spaces is of course the classical *central limit theorem*: The sum of independent random variables is approximately Gaussian when the number of variables approaches infinity, under quite general assumptions. In other words, for large n, suppose that $f_1, \dots, f_n$ are probability densities on the real line of mean zero and variance one, satisfying certain mild regularity conditions. Then, the integral

$$\int_{H_t} \prod_{i=1}^{n} f_i(x_i)$$

with $H_t = \{x \in \mathbb{R}^n ; \sum_{i=1}^n x_i = t\sqrt{n}\}$ is approximately the Gaussian density $e^{-t^2/2}/\sqrt{2\pi}$. Therefore, the value of this integral does not depend too much on the specific form of the densities we started with, and the behavior is asymptotically *universal*. This is a marvelous effect of universality in high dimensions, indicating that when viewed correctly, high-dimensional measures exhibit regularity and order rather than incomprehensible complications.

Another example for regularity in high dimensions is Dvoretzky's theorem, which asserts that any high-dimensional convex body has nearly Euclidean sections of a large dimension; see [81] and references therein for background. Thus, the symmetries of the Euclidean ball appear, even though we made only minimal assumptions: only convexity and the high dimension. The central limit theorem and Dvoretzky's theorem are high-dimensional effects that lack clear analogs in low dimensions.

As it turns out, there are motifs in high-dimensional geometry which seem to compensate for the difficulties that arise from high dimensionality. One of these motifs is the *concentration of measure phenomenon*. Quite unexpectedly, a scalar Lipschitz function on a high-dimensional space behaves in many cases as if it were a constant function. For example, if we sample five random points from the n-dimensional unit sphere, for large n, and substitute them into a 1-Lipschitz function, then we will almost certainly obtain five numbers that are very close to one another. This phenomenon is reminiscent of the well-known geometric property that in the high-dimensional Euclidean sphere, "most of the mass is close to the equator, for any equator." This geometric property, which follows from the isoperimetric inequality, is unthinkable in, say, three dimensions. Since the second-named author's proof of Dvoretkzy's theorem in the 1970s, the concentration of measure has become a major tool in high-dimensional analysis.

It was a big surprise in the 1970s and 1980s that the asymptotic behavior (i.e., when the dimension increases) of high-dimensional normed spaces is "well organized" and not chaotic, as one could expect from the intuition which was based, perhaps, on exponential growth of entropy (=covering) for n-dimensional spaces. However, the concentration of measure balances the exponentially high entropy of n-dimensional spaces and leads to a "regularity" in high dimension, limiting the "geometric diversity" in high dimensions. The absolute constants involved in the analysis may balance the rate of exponential decay (coming from concentration) with the rate of exponential expansion (coming from covering/entropy). Surprisingly, both exponents have "roughly" the same order of decay via expansion by dimension, and only a constant factor is needed in order to compensate and obtain a regularity result in high dimensions. The constant factors in the exponent are needed for compensation, and this explains the "isomorphic" nature of the results, the fact that absolute constants appear in their formulation.

So, a new intuition had to be created roughly four decades ago, and it was built upon results which showed very regular patterns. Today, we may state that these results were observed roughly in two different forms:

(a) Geometric and structural results (e.g., Dvoretzky's theorem, quotient of subspace theorem, Ramsey's theorem in combinatorics).
(b) The uniform measure distribution (volume behavior) in high-dimensional convex bodies.

In both of these forms, there is striking regularity and almost no pathology when the dimension increases. Bourgain's slicing problem had a major influence on (b), and the entire direction actually stemmed from his conjecture and his work. The Bourgain-Milman inequality [25] is one of the results from that period of time that is closest to a bridge between (a) and (b).

The spatial arrangement of volume due to the geometry of $\mathbb{R}^n$, for large n, imposes rigidity on convex sets and convexity-related measures. Convexity is one of the ways in which one may harness the concentration of measure phenomenon in order to formulate clean, non-trivial theorems. The Brunn-Minkowski inequality from the end of the nineteenth century states that for any non-empty Borel sets $A, B \subseteq \mathbb{R}^n$,

$$|A + B|^{1/n} \geq |A|^{1/n} + |B|^{1/n}, \tag{3}$$

where $A + B = \{x + y\,;\, x \in A, y \in B\}$ and $|A|$ is the volume of the set A. The Brunn-Minkowski inequality is a close relative of the isoperimetric inequality, and equality holds in (3) essentially only when A and B are congruent convex bodies. In addition to (1) above, let us mention another consequence of the Brunn-Minkowski theory: the reverse Hölder inequalities, proven by Berwald [13] and Borell (see [83] or Borell's papers [17, 18, 20]). For any convex body $K \subseteq \mathbb{R}^n$, a linear functional $f : \mathbb{R}^n \to \mathbb{R}$ and $p, q > 0$,

$$\left(\int_K |f(x)|^p \frac{dx}{|K|} \right)^{1/p} \leq C \left(\int_K |f(x)|^q \frac{dx}{|K|} \right)^{1/q} \tag{4}$$

where $C = C_{p,q} > 0$ depends solely on p, q and neither on K nor on the dimension n. This amusing property of convex domains goes beyond linear functionals. Suppose now that $f : \mathbb{R}^n \to \mathbb{R}$ is an arbitrary polynomial of degree at most d. Bourgain proved in his paper [23] that (4) holds true in this case, with the constant C depending only on p, q and d, and not on the convex body or the dimension. These results serve as evidence for the general hypothesis that in many respects the uniform measure on a high-dimensional convex body resembles a Gaussian measure.

In the same paper, [23] Bourgain proved that the constant C from Question 1.2 (or Question 1.1) may be replaced by $Cn^{0.25} \log n$. The logarithmic factor was later removed by the first-named author in [51]. We proceed with a more detailed account of the development of the study of the regularity of high-dimensional convexity-related measures and the major influence that Jean Bourgain had on this development.

2 The Isotropic Position

The covariance matrix $\mathrm{Cov}(K) = (\mathrm{Cov}_{ij}(K))_{i,j=1,\dots,n}$ of a convex body $K \subset \mathbb{R}^n$ is given by

$$\mathrm{Cov}_{ij}(K) = \int_K x_i x_j \frac{dx}{Vol_n(K)} - \int_K x_i \frac{dx}{Vol_n(K)} \int_K x_j \frac{dx}{Vol_n(K)}. \tag{5}$$

When $Vol_n(K) = 1$, its isotropic constant satisfies $L_K^{2n} = \det \mathrm{Cov}(K)$, according to (2).

A convex body $K \subset \mathbb{R}^n$ of volume one is *isotropic* (or in *isotropic position*) if its barycenter lies at the origin and its covariance matrix is a scalar matrix. Any convex body $K \subset \mathbb{R}^n$ can be transformed into an isotropic convex body by applying an affine transformation of the form $Tx = \alpha \mathrm{Cov}(K)^{-1/2} x + v$ for appropriate $\alpha > 0$ and $v \in \mathbb{R}^n$. It follows from (2) that when K is isotropic,

$$\mathrm{Cov}(K) = L_K^2 \cdot \mathrm{Id}. \tag{6}$$

In other words, when K is isotropic, for any $\theta \in S^{n-1}$,

$$\int_K \langle x, \theta \rangle^2 dx = L_K^2. \tag{7}$$

It follows from (6) that for any convex body $K \subseteq \mathbb{R}^n$ with $Vol_n(K) = 1$ and for any invertible linear map $T : \mathbb{R}^n \to \mathbb{R}^n$,

$$L_K^2 \leq \frac{1}{n |\det T|^{2/n}} \int_K |Tx|^2 dx. \tag{8}$$

Indeed, it suffices to prove (8) in the case where K is isotropic. In this case, the right-hand side of (8) equals $L_K^2 \cdot \mathrm{Trace}[T^* T]/(n |\det T|^{2/n}) \geq L_K^2$, by the arithmetic-geometric means inequality.

When $K \subseteq \mathbb{R}^n$ is a convex body of volume one with barycenter at the origin, an alternative definition of L_K^2 is that it is the minimum of the right-hand side of (8) over all linear transformations of determinant one. Consequently, for such $K \subseteq \mathbb{R}^n$ there exists a linear map T with $\det(T) = 1$ and

$$L_K^2 = \frac{1}{n} \int_K |Tx|^2 dx = \frac{1}{n} \int_0^\infty Vol_n(K \setminus \sqrt{s} T^{-1}(B^n)) ds \geq \frac{1}{n} \int_0^{\kappa_n^{-2/n}} \left[1 - \kappa_n s^{n/2} \right] ds = L_{B^n}^2,$$

where $B^n = \{x \in \mathbb{R}^n \, ; \, |x| = 1\}$ is the Euclidean unit ball and $\kappa_n = Vol_n(B^n)$. Since $L_{B^n}^2 = 1/(2\pi e) + o(1)$, we conclude that $L_K > c$ for some universal constant $c > 0$.

If $K \subset \mathbb{R}^n$ is an isotropic convex body, then for any two hyperplanes $H_1, H_2 \subseteq \mathbb{R}^n$ through the origin,

$$\frac{Vol_{n-1}(K \cap H_1)}{Vol_{n-1}(K \cap H_2)} \leq C \tag{9}$$

for a universal constant $C > 0$. This was proven by Hensley [45] in the case where K is centrally symmetric and rediscovered by the second-named author in Lemma 2 in [21]. Fradelizi [38] eliminated the assumption that K is centrally symmetric and obtained the sharp bound $C \leq \sqrt{6}$ in (9). In order to prove (9), one fixes a unit vector $\theta \in S^{n-1}$ and denotes

$$\rho(t) = Vol_{n-1}(K \cap (t\theta + \theta^{\perp})).$$

A crucial property that follows from the Brunn-Minkowski inequality is that ρ is *log-concave*, that is, the function $\log \rho$ is a concave function (which is allowed to attain the value $-\infty$). Therefore, the proof of (9) boils down to the proof of the following one-dimensional inequality: For any log-concave probability density $\rho : \mathbb{R} \to [0, \infty)$ with $\int t\rho(t)dt = 0$,

$$\frac{1}{\sqrt{12}} \leq \rho(0) \cdot \sqrt{\int_{-\infty}^{\infty} t^2 \rho(t)dt} \leq \frac{1}{\sqrt{2}}. \tag{10}$$

The space of one-dimensional, log-concave probability densities of mean zero and variance one is compact in the L^1-topology. A compactness argument shows that an inequality such as (10) holds true with *some* numerical constants. The sharp values of the constants in (10) are due to Fradelizi [38]. Consequently, whenever $K \subseteq \mathbb{R}^n$ is an isotropic convex body, for any hyperplane $H \subseteq \mathbb{R}^n$ through the origin,

$$\frac{1}{\sqrt{12} \cdot L_K} \leq Vol_{n-1}(K \cap H) \leq \frac{1}{\sqrt{2} \cdot L_K}. \tag{11}$$

The assumption that H passes through the origin is not entirely necessary for the right-hand side inequality in (11), if one is willing to increase the constant. This follows from a version of inequality (10) where $\rho(0)$ is replaced by $\sup \rho$; see Fradelizi [38]. It follows that when $K \subseteq \mathbb{R}^n$ is convex and isotropic, for any hyperplane $H \subseteq \mathbb{R}^n$,

$$Vol_n(K \cap H) \leq \frac{1}{L_K}. \tag{12}$$

From (12), we obtain the relatively trivial bound $L_K \leq C\sqrt{n}$ for the isotropic constant. Indeed, since K is convex and of volume one, it cannot have a width larger than $5\sqrt{n}$ in all directions, as otherwise $K - K$ would contain a Euclidean ball of volume larger than 4^n, in contradiction to the Rogers-Shephard inequality

[86]. We recall that this inequality states that $Vol_n(K - K) \leq 4^n Vol_n(K)$ for any convex body $K \subseteq \mathbb{R}^n$. Pick a direction in which the width of K is at most $5\sqrt{n}$, and use Fubini's theorem to find a hyperplane H orthogonal to this direction with $Vol_n(K \cap H) \geq 1/(5\sqrt{n})$. Now (12) shows that $L_K \leq 5\sqrt{n}$.

The idea demonstrated above, of reducing statements on convex bodies to one-dimensional inequalities pertaining to log-concave functions, is a common theme in convex geometry. For example, the reverse Hölder inequality (4) may be proven by reducing matters to a one-dimensional inequality with log-concave probability densities. A log-concave function in one dimension of a finite integral decays exponentially at infinity (see, e.g. [52, Lemma 2.1]). It follows (see [83] or [82]) that for any convex body $K \subseteq \mathbb{R}^n$ of volume one, the "ψ_1-norm" of a linear functional $f : \mathbb{R}^n \to \mathbb{R}$ satisfies

$$\|f\|_{\psi_1(K)} \leq C\|f\|_{L^2(K)} \tag{13}$$

where $C > 0$ is a universal constant, where $\|f\|_{L^p(K)}^p = \int_K |f|^p$ and where for $\alpha \geq 1$,

$$\|f\|_{\psi_\alpha(K)} = \inf\left\{\lambda > 0;\ \int_K \exp(|f/\lambda|^\alpha) \leq 2\right\}. \tag{14}$$

The ψ_1-norm of a function f is finite if its value distribution is subexponential, and the ψ_2-norm is finite if the distribution is sub-Gaussian. The contrast between ψ_1-norm and ψ_2-norm, or between subexponential tail and sub-Gaussian tail, lies at the heart of Bourgain's bound $L_K \leq Cn^{1/4}\log n$. Before proceeding with Bourgain's proof, let us provide a bit of background on ψ_2-processes and on certain results from the local theory of Banach spaces that are related to Bourgain's proof. Suppose that μ is a probability measure on $\mathbb{R}^n$, and denote

$$A := \sup_{\theta \in S^{n-1}} \|f_\theta\|_{\psi_2(\mu)} \tag{15}$$

where $f_\theta(x) = \langle x, \theta \rangle$ and where the ψ_2-norm of a function f with respect to the measure μ is defined analogously to (14) above. A key result by Talagrand [96, 97] continuing the work of Fernique [35] states that for any norm $\|\cdot\|$ on $\mathbb{R}^n$,

$$\int_{\mathbb{R}^n} \|x\| d\mu(x) \leq CA \int_{\mathbb{R}^n} \|x\| d\gamma_n(x) \tag{16}$$

where $C > 0$ is a universal constant and where γ_n is the standard Gaussian measure on $\mathbb{R}^n$. The proof of inequality (16) involves concepts such as majorizing measures and generic chaining.

Bounds for the Gaussian integral of a norm, as on the right-hand side of (16), are of great importance in the local theory of Banach spaces. One of the most important and useful technical statements in this direction is the following theorem, which is a combination of three results, by Lewis [69], by Figiel and Tomczak-Jaegermann [36], and, the most non-trivial, by Pisier [90, 91] (see also the appendix of [25] for a complete proof):

Theorem 2.1 *For any norm $\|\cdot\|$ on $\mathbb{R}^n$, there exists an invertible linear transformation T such that*

$$\int_{\mathbb{R}^n} \|Tx\| d\gamma_n(x) \cdot \int_{\mathbb{R}^n} \|(T^{-1})^* y\|_* d\gamma_n(y) \leq Cn \log d_{BM}$$

where $\|\cdot\|_*$ is the dual norm and where d_{BM} is the Banach-Mazur distance of the norm $\|\cdot\|$ from a Euclidean norm. The linear map T determines the so-called ℓ-position.

When a norm $\|\cdot\|$ on $\mathbb{R}^n$ has K as its unit ball, its Banach-Mazur distance from a Euclidean norm is

$$d_{BM} = d_{BM}(K) = \inf\{rs > 0\,;\, \exists T : \mathbb{R}^n \to \mathbb{R}^n \text{ linear, with } r^{-1}B^n \subseteq T(K) \subseteq sB^n\}. \tag{17}$$

It is well-known that $d_{BM} \leq \sqrt{n}$; see e.g. [82]. We remark in passing that **up to logarithmic factors, the slicing problem is equivalent to the question of whether the isotropic position is an ℓ-position**, as one may show; see [27] for related results. Theorem 2.1 is a central ingredient of the original proof of the Bourgain-Milman inequality [25], which states that for any convex body $K \subseteq \mathbb{R}^n$ with barycenter at the origin,

$$Vol_n(K)Vol_n(K^\circ) \geq c^n Vol_n(B^n)^2 \geq (c'/n)^n, \tag{18}$$

where $K^\circ = \{x \in \mathbb{R}^n\,;\, \forall y \in K,\ |\langle x, y\rangle| \leq 1\}$ is the polar body, i.e., the unit ball of the dual norm. There are by now several proofs of (18) using methods and ideas from very different parts of mathematics. Kuperberg's proof relies on topology [66], Nazarov's proof on complex analysis [85], and the proof by Giannopoulos, Paouris, and Vritsiou on transportation of measure via the logarithmic Laplace transform [42] as in Section 4 below. Inequality (18) is a converse to the Santaló inequality, which states that

$$Vol_n(K)Vol_n(K^\circ) \leq Vol_n(B^n)^2,$$

and may be proven via Steiner symmetrizations [76, 77]. A clever application of Hölder's inequality shows that $\int_{\mathbb{R}^n} \|x\| d\gamma_n(x) \geq cn \cdot v^{-1/n}$ where $v > 0$ is the volume of the unit ball of the norm $\|\cdot\|$ in $\mathbb{R}^n$. It thus follows from Theorem 2.1 and from the above that for any norm $\|\cdot\|$ on $\mathbb{R}^n$, there exists a linear map $T : \mathbb{R}^n \to \mathbb{R}^n$ of determinant one such that

$$\int_{\mathbb{R}^n} \|Tx\| d\gamma_n(x) \leq Cn \log n \cdot V^{1/n} \tag{19}$$

where now $V > 0$ is the volume of the unit ball of the dual norm $\|\cdot\|_*$ in $\mathbb{R}^n$.

Let us now return to the proof of Bourgain's bound for the isotropic constant. It follows from (7) and from the Markov-Chebyshev inequality that for any isotropic convex body $K \subseteq \mathbb{R}^n$,

$$Vol_n(K \cap (\sqrt{2n} L_K B^n)) = 1 - Vol_n(K \setminus (\sqrt{2n} L_K B^n)) \geq 1 - \frac{\int_K |x|^2 dx}{2n L_K^2} = \frac{1}{2}. \tag{20}$$

The first step of the proof is to use (20) in order to show the following: When replacing K with $K \cap C\sqrt{n} L_K B^n$, the isotropic constant changes by a factor of at most C, and the new convex body is still roughly in isotropic position (up to a constant). Thus, it suffices to bound the isotropic constant of an isotropic convex body $K \subseteq \mathbb{R}^n$ which satisfies the additional assumption that

$$K \subseteq 10\sqrt{n} L_K B^n. \tag{21}$$

One corollary of (21) is that for any $\theta \in S^{n-1}$, the linear functional $f_\theta(x) = \langle x, \theta \rangle$ satisfies

$$\|f_\theta\|_{L^\infty(K)} \leq 10\sqrt{n} L_K.$$

The $\psi_1(K)$-norm of f_θ is at most CL_K, according to (7) and (13) above. There is a simple interpolation inequality between the ψ_1-norm and the L^∞-norm that yields a bound for the ψ_2-norm. Namely, for any $\theta \in S^{n-1}$,

$$\|f_\theta\|_{\psi_2(K)} \leq \sqrt{\|f_\theta\|_{\psi_1(K)} \cdot \|f_\theta\|_{L^\infty(K)}} \leq \sqrt{CL_K \cdot 10\sqrt{n} L_K} = C' n^{1/4} L_K. \tag{22}$$

The proof of the interpolation inequality on the left-hand side of (22) is simple; note that when $\sup |f| \leq M$,

$$\int_K e^{|f/\sqrt{\lambda M}|^2} \leq \int_K e^{|f/\lambda|} \leq 2$$

if $\lambda \geq \|f\|_{\psi_1(K)}$. The next step in Bourgain's proof is to apply (8) and conclude that for any symmetric, positive-definite linear map $T : \mathbb{R}^n \to \mathbb{R}^n$ with $\det T = 1$,

$$n L_K^2 \leq \int_K \langle Tx, x \rangle \leq \int_K \sup_{y \in TK} |\langle x, y \rangle| = \int_K \|Tx\| dx \tag{23}$$

where $\|x\| = \sup_{y \in K} |\langle x, y \rangle|$ is a norm on $\mathbb{R}^n$ whose unit ball is polar to $K \cap (-K)$. An interesting feature of the maneuver (23) is the comparison between an integral with quadratic dependence on x, which is reflected in the square of L_K on the left-hand side, and an integral whose dependence on x is not quadratic but only linear. Next, thanks to (22), we may apply the Talagrand bound (16) and conclude that

$$nL_K^2 \le \int_K \|Tx\|dx \le Cn^{1/4}L_K \int_{\mathbb{R}^n} \|Tx\|d\gamma_n(x). \tag{24}$$

Inequality (24) is valid for any linear map $T : \mathbb{R}^n \to \mathbb{R}^n$ of determinant one (the assumption that T is symmetric and positive-definite is immaterial due to the symmetries of the Gaussian measure). We may now choose T to be a map leading to ℓ-position and apply Pisier's bound in the form of inequality (19) above. This shows that

$$nL_K^2 \le Cn^{1/4}L_K \cdot n\log n \cdot Vol_n(K \cap (-K))^{1/n} \le C'n^{1/4}L_K \cdot n\log n.$$

This completes the proof of Bourgain's bound $L_K \le Cn^{1/4}\log n$.

In his paper [21], Bourgain claimed a positive answer to the slicing problem in the case where $K \subseteq \mathbb{R}^n$ is *unconditional*, i.e., when for any $x = (x_1, \dots, x_n) \in \mathbb{R}^n$,

$$(x_1, \dots, x_n) \in K \qquad \Longleftrightarrow \qquad (|x_1|, \dots, |x_n|) \in K.$$

In this case, one may use the Loomis-Whitney inequality [71], which is valid for any compact set in $\mathbb{R}^n$:

$$Vol_n(K) \le \prod_{i=1}^n Vol_{n-1}(Proj_{e_i^\perp} K)^{1/(n-1)}, \tag{25}$$

where $Proj_{e_i^\perp}$ is the orthogonal projection onto the hyperplane $e_i^\perp$ and e_i is the ith-standard unit vector in $\mathbb{R}^n$. When K is convex and unconditional, $Proj_{e_i^\perp} K = K \cap e_i^\perp$. Hence, (11) and (25) imply that $L_K \le 1/\sqrt{2}$ when K is convex and unconditional (this numerical constant may be improved). Moreover, if $K \subseteq T$ and T is an unconditional convex body such that $Vol_n(T)/Vol_n(K) \le A^n$, it is known that $L_T \le CA$; see [83] or the recent book by Brazitikos, Giannopoulos, Valettas, and Vritsiou [12], a large part of which is concerned with the slicing problem.

In addition to unconditional bodies, there are other classes of convex bodies for which an affirmative answer to the slicing problem is known. These include zonoids [87], their duals, more generally subspaces and quotients of L_p spaces [8, 46, 47, 78], unit balls of Schatten class norms [65], random convex bodies [59], 2-convex bodies, and other examples described in [12].

In [24], Bourgain proved the boundness of the isotropic constant for "ψ_2-bodies" which are convex bodies for which the ψ_1-estimate (13) can be upgraded to a ψ_2-estimate. That is, for a convex body $K \subseteq \mathbb{R}^n$ of volume one with barycenter at the origin and for $1 \le \alpha \le 2$, we write $b_\alpha(K)$ to be the minimum $b > 0$ such that for any linear functional $f : \mathbb{R}^n \to \mathbb{R}$,

$$\|f\|_{\psi_\alpha(K)} \le b\|f\|_{L^2(K)}.$$

Thus, $b_1(K) \leq C$ according to (13), while for ellipsoids $\mathcal{E} \subseteq \mathbb{R}^n$ we have $b_2(\mathcal{E}) \leq C$. Bourgain proved that $L_K \leq Cb_2(K) \log b_2(K)$, and the logarithmic factor was later removed by the first-named author and E. Milman in [63] using methods related to those described in Section 4. The current state of the art is the bound $L_K \leq C\sqrt{b_\alpha(K)^\alpha n^{1-\alpha/2}}$ for any $1 \leq \alpha \leq 2$, from [63].

A class of convex bodies in high dimensions with favorable properties is the class of convex bodies of a *finite volume ratio*, a notion introduced by Szarek and Tomczak-Jaegermann [94, 95]. These are centrally symmetric convex bodies $K \subseteq \mathbb{R}^n$ that contain an ellipsoid $\mathcal{E}$ such that $Vol_n(K)/Vol_n(\mathcal{E}) \leq C^n$ for a universal constant C. Dvoretzky's theorem asserts that an arbitrary convex body in $\mathbb{R}^n$ has a k-dimensional section that is approximately Euclidean for k of the order of magnitude of $\log n$. This estimate is dramatically improved in the class of finite-volume ratio bodies, and it was proven by Kashin [49] and then using this terminology by Szarek and Tomczak-Jaegermann [94, 95] that such bodies contain an approximately Euclidean section of dimension $k \geq cn$. In our joint work with Bourgain [26, 27], we proved that the validity of the hyperplane conjecture in the class of finite-volume ratio bodies would imply its validity in the class of all convex bodies. This is proven via a method based on Steiner symmetrization.

We move on to describe yet another equivalent formulation of the slicing problem, which is also related to Steiner symmetrizations. Let $K \subseteq \mathbb{R}^n$ be a convex body and let $H = h^\perp \subseteq \mathbb{R}^n$ be a hyperplane, where $h \in S^{n-1}$ is a unit vector. Define the *Steiner symmetral of K with respect to H* as the set

$$S_H(K) = \left\{ x + th \,;\, x \in H, t \in \mathbb{R},\ K \cap (x + \mathbb{R}h) \neq \emptyset\,,\ |t| \leq \frac{1}{2} Meas\{K \cap (x + \mathbb{R}h)\} \right\}$$

where $Meas$ is the one-dimensional Lebesgue measure in the line $x + \mathbb{R}h$. Steiner symmetrization preserves the volume of the set K, and it transforms convex sets to convex sets. Applying consecutive Steiner symmetrizations with respect to a sequence of hyperplanes makes K "more symmetric," or "closer to a Euclidean ball." It was proven in [60] that for any convex body $K \subseteq \mathbb{R}^n$ with $Vol_n(K) = Vol_n(B^n)$ there exist $3n$ Steiner symmetrizations that transform K into a convex body $\tilde{K}$ with

$$\frac{1}{C} B^n \subseteq \tilde{K} \subseteq CB^n. \tag{26}$$

When $\tilde{K}$ satisfies (26), we say that it is an "isomorphic Euclidean ball." If one applies only $n - \ell$ symmetrizations for some positive ℓ, then there exists an ℓ-dimensional projection of K that remains unchanged in the symmetrization process. Hence, at least $n - O(1)$ symmetrizations are required to arrive at an isomorphic Euclidean ball as in (26), or even at an isomorphic ellipsoid. However, there exist certain special convex bodies, such as the unit cube, that can be transformed into an isomorphic ellipsoid using fewer symmetrizations. Given a convex body $K \subset \mathbb{R}^n$ and a function $c(\varepsilon)$ $(0 < \varepsilon < 1)$, we say that "K is $c(\varepsilon)$-symmetrizable" if for any $\varepsilon > 0$ there exist $\lfloor \varepsilon n \rfloor$ Steiner symmetrizations that transform K into a convex body

$\tilde{K}$ with

$$d_{BM}(K) < c(\varepsilon)$$

where $d_{BM}(K)$ is the Banach-Mazur distance between K and a Euclidean ball, defined analogously to (17). For example, the cube $[-1, 1]^n$ is $c(\varepsilon)$-symmetrizable for $c(\varepsilon) = C\sqrt{|\log \varepsilon|/\varepsilon}$. Suppose that we are allowed to remove 10% of the mass of a convex body K. Can we now apply only εn Steiner symmetrizations and obtain a body that resembles an ellipsoid, up to a universal constant?

Question 2.1 *Does there exist $C, d > 0$, such that for any dimension n and for any convex body $K \subseteq \mathbb{R}^n$, there exists a convex body $T \subseteq K$ with $Vol_n(T) > 0.9 \cdot Vol_n(K)$ such that T is $(C/\varepsilon)^d$-symmetrizable?*

In [61], it is proven that **Question 2.1 has an affirmative answer if and only if Bourgain's hyperplane conjecture holds true**.

3 Distribution of Volume in Convex Bodies

The assumption that K is *convex* was used in Bourgain's proof through the ψ_1-bound (13), the fact that the distribution of values of a linear functional on a convex set has a uniformly subexponential tail. In fact, instead of dealing with the uniform measure on a given convex body $K \subseteq \mathbb{R}^n$ of volume one, we may consider a more general situation: Suppose that μ is a probability measure supported on K whose continuous density is denoted by f. Assume that μ satisfies a ψ_1-condition: For any linear functional $f : \mathbb{R}^n \to \mathbb{R}$,

$$\|f\|_{\psi_1(\mu)} \leq A\|f\|_{L^2(\mu)},$$

for some parameter $A > 0$, where the definition of the $\|\cdot\|_{\psi_1(\mu)}$ norm is analogous to (14). A straightforward adaptation of Bourgain's argument (see the Appendix of [57]) shows that under these assumptions there exists a hyperplane $H \subseteq \mathbb{R}^n$ with

$$\int_H f \geq \frac{c(A)}{n^{1/4}\log n}, \tag{27}$$

where $c(A) > 0$ depends solely on A. Up to the logarithmic factor, the estimate "$n^{1/4}\log n$" in (27) is sharp, as shown by the first-named author and Koldobsky [57]. Thus, Bourgain's bound for the slicing problem is sharp up to logarithmic factors, if all that one takes from convexity is the uniform ψ_1-estimate for linear functionals.

Nevertheless, there is more to say about the distribution of volume in convex bodies beyond the subexponential tail of linear functionals. We begin by discussing the point of view emphasized by K. Ball [6] that connects between volume distribution of convex bodies and that of log-concave measures. Recall that a function $\rho : \mathbb{R}^n \to [0, \infty)$ is log-concave if $-\log \rho$ is a convex function, which is allowed to attain the value $+\infty$. For example, the characteristic function of a convex body, which equals one on the body and vanishes elsewhere, is a log-concave function.

A Borel measure on $\mathbb{R}^n$ is log-concave if it is supported in an affine subspace with a log-concave density in this subspace. It was proven by Borell [19] that a finite, Borel measure μ on $\mathbb{R}^n$ is a log-concave measure if and only if the following Brunn-Minkowski type inequality holds true: For any compacts $A, B \subseteq \mathbb{R}^n$ and for any $0 < \lambda < 1$,

$$\mu(\lambda A + (1-\lambda)B) \geq \mu(A)^{\lambda}\mu(B)^{1-\lambda}. \tag{28}$$

It follows from (28) that the push-forward of a log-concave measure under a linear map is again log-concave. In particular, by projecting the uniform measure on a convex body to a lower-dimensional subspace, we obtain a log-concave measure on this subspace. Given an integrable log-concave function $\rho : \mathbb{R}^n \to \mathbb{R}$ with $\rho(0) > 0$, define

$$K(\rho) = \left\{ x \in \mathbb{R}^n \, ; \, \int_0^{\infty} \rho(tx)t^n dt \geq \frac{\rho(0)}{n+1} \right\}. \tag{29}$$

As shown by K. Ball [6], the set in (29) is always convex. The convexity of $K(\rho)$ is closely related to the Busemann inequality [29]; see [83].

The convex body $K(\rho)$ seems to represent rather well the volume distribution of the measure μ whose density is ρ. For example, if the barycenter of μ lies at the origin, so does the barycenter of $K(\rho)$. As in [51], we define the isotropic constant of a log-concave function $\rho : \mathbb{R} \to [0, \infty)$ with $0 < \int \rho < \infty$ as

$$L_{\rho} = \left(\frac{\sup \rho}{\int \rho} \right)^{1/n} \cdot \det \mathrm{Cov}(\rho)^{1/(2n)}$$

where the covariance matrix $\mathrm{Cov}(\rho)$ is defined analogously to (5). It was proven by Ball [6] (see also [51]) that we always have

$$L_{\rho} \simeq L_{K(\rho)}. \tag{30}$$

Thus, **the slicing problem is equivalent to the problem of bounding the isotropic constant of an arbitrary log-concave measure** μ on $\mathbb{R}^n$.

A new era in the study of volume distribution in high-dimensional convex sets began in 2005 when Paouris [89] found applications of the following property: For any absolutely continuous, log-concave probability measure μ on $\mathbb{R}^n$, there exists $\theta \in S^{n-1}$ with

$$\left(\int_{\mathbb{R}^n} |\langle x,\theta\rangle|^n d\mu(x)\right)^{1/n} \simeq \sqrt{\int_{\mathbb{R}^n} |x|^2 d\mu(x)}. \tag{31}$$

This property is proven by associating a certain convex body with the measure μ similarly to (29); see [62] or (better) the short argument in [1]. Note that this property of log-concave measures does not follow from the ψ_1-bound for linear functionals, used in Bourgain's proof (e.g., look at the random variable $X = R\Gamma$ where R is a standard one-dimensional Gaussian and Γ is an n-dimensional standard Gaussian independent of R). Recalling that the orthogonal projection of the uniform measure of a convex body is always log-concave, we conclude the following from (31): For any isotropic convex body $K \subseteq \mathbb{R}^n$ and for any ℓ-dimensional subspace $E \subseteq \mathbb{R}^n$, there exists a unit vector $\theta \in E$ with

$$\left(\int_K |\langle x,\theta\rangle|^\ell dx\right)^{1/\ell} \simeq \sqrt{\ell} L_K. \tag{32}$$

Note that an isotropic convex body K is a ψ_2-body if and only if (32) holds true for any $\theta \in \mathbb{R}^n$ and for any ℓ. Thus, property (32) may be viewed as a weak form of a ψ_2-estimate, which is valid for any convex body. Paouris used (32) in order to prove the following large deviation estimate:

Theorem 3.1 (Paouris [89]) *Let $K \subseteq \mathbb{R}^n$ be an isotropic convex body of volume one. Then, for any $t > 1$,*

$$Vol_n(\{x \in K\,;\ |x| \geq CtL_K\sqrt{n}\}) \leq e^{-t\sqrt{n}} \tag{33}$$

where $C > 0$ is a universal constant.

In order to appreciate Theorem 3.1, recall from (20) that

$$Vol_n(\{x \in K\,;\ |x| \leq 2L_K\sqrt{n}\}) \geq 1/2,$$

i.e., at least half of the mass of K is located in a ball of radius $2L_K\sqrt{n}$ centered at the origin. Theorem 3.1 tells us that only a tiny fraction, just an $e^{-\sqrt{n}}$-fraction of the mass of K, is located outside a ball of radius $CL_K\sqrt{n}$. This effect is a precursor to the *thin-shell estimate* for isotropic convex bodies that we will discuss shortly. The Paouris proof of Theorem 3.1 applies the second-named author's estimates for the Dvoretzky theorem (see [81]) in the context of the norm

$$\|y\|_{L^p(K)} = \|\langle\cdot,y\rangle\|_{L^p(K)} = \left(\int_K |\langle x,y\rangle|^p dx\right)^{1/p} \qquad (y \in \mathbb{R}^n).$$

The unit ball of the dual norm is denoted by $Z_p(K) \subseteq \mathbb{R}^n$, and it is referred to as the *L_p-centroid body of K*; see also Lutwak and Zhang [72]. If K is isotropic, then the set $Z_2(K)$ is a Euclidean ball of radius L_K. In the case where K is centrally symmetric, we have $Z_\infty(K) = K$. The ψ_1-estimate for linear functionals on convex bodies (13) is equivalent to the assertion that

$$Z_p(K) \subseteq CpZ_2(K) \qquad \text{for all } p \geq 1. \tag{34}$$

The idea of Paouris was to apply the quantitative theory of Dvoretzky's theorem, due to the second-named author [81], to the L_p-centroid bodies. Together with the estimate (34), this quantitative theory yields the following: Suppose that $K \subseteq \mathbb{R}^n$ is an isotropic convex body and $1 \leq \ell \leq c\sqrt{n}$. Then, for a random ℓ-dimensional subspace $E \subseteq \mathbb{R}^n$, with high probability, the orthogonal projection of $Z_\ell(K)$ onto E denoted by

$$\mathrm{Proj}_E(Z_\ell(K)) \tag{35}$$

is an isomorphic Euclidean ball. In other words, the convex body in (35) contains a Euclidean ball of radius r centered at the origin, and it is contained in a Euclidean ball of radius Cr. We may now invoke (32) and conclude that r has the order of magnitude of $\sqrt{\ell}L_K$. Thus, by the quantitative estimates revolving around Dvoretzky's theorem, due to Litvak, Milman, and Schechtman [70],

$$\sqrt{\ell}L_K \simeq r \simeq \left(\int_{S^{n-1}} \|y\|^\ell_{L^\ell(K)} d\sigma(y)\right)^{1/\ell} \simeq \sqrt{\frac{\ell}{n}}\left(\int_K |x|^\ell dx\right)^{1/\ell}, \tag{36}$$

where σ is the rotationally invariant probability measure on S^{n-1}. Thus, (36) yields estimates for L_p-moments of the Euclidean norm for all $p \leq c\sqrt{n}$, which imply that only a fraction of at most $e^{-\sqrt{n}}$ of the volume of K is located outside a ball of radius $CL_K\sqrt{n}$.

The tension between ψ_1-estimates and ψ_2-estimates for convex bodies, going back to Bourgain's work in the 1980s, is a central issue in the analysis of the slicing problem. Recall that the inclusion (34) follows from the ψ_1-bound, while a ψ_2-estimate with constant A would yield that $Z_p(\mu) \subseteq CA\sqrt{p}Z_2(\mu)$. In this respect, it is worthwhile to mention yet another equivalent formulation of the hyperplane conjecture, which may be extracted from [63, Remark 3.3]: **Question 1.2 has an affirmative answer if and only if for any isotropic convex body $K \subseteq \mathbb{R}^n$ and any $1 \leq p \leq n$,**

$$\boldsymbol{Vol^{1/n}(Z_p(\mu)) \simeq \sqrt{p} \cdot Vol^{1/n}(Z_2(\mu)).}$$

A question by the second-named author (see [10, 87, 88]) asks whether for any convex body $K \subseteq \mathbb{R}^n$ there exists a non-zero linear functional $\varphi : \mathbb{R}^n \to \mathbb{R}$ for which

$$\|\varphi\|_{\psi_2(K)} \leq C\|\varphi\|_{L^2(K)},$$

with a universal constant C. In other words, does any convex body have at least one direction with a uniformly sub-Gaussian tail? In some sense, a direction where the tail is approximately exponential resembles a "cone-like behavior" of the convex body (see [52]), and the question is whether there always exists a direction in which better, sub-Gaussian behavior is observed. It was proven by the first-named author in [52] that the answer is affirmative up to logarithmic factors. The logarithmic factor that the proof in [52] yielded is $\log^5(t+1)$ (in formula (37) below), and it was subsequently improved to $\log^2(t+1)$ in Giannopoulos, Pajor, and Paouris [40] and then to $\log(t+1)$ in Giannopoulos, Paouris, and Valettas [41]:

Theorem 3.2 *Let $n \geq 1$ be an integer, and let $K \subset \mathbb{R}^n$ be a convex body of volume one. Then, there exists a non-zero linear functional $\varphi : \mathbb{R}^n \to \mathbb{R}$ such that for any $t \geq 1$,*

$$Vol_n\left(\{x \in K; |\varphi(x)| \geq t\|\varphi\|_{L_1(K)}\}\right) \leq e^{-c\frac{t^2}{\log(t+1)}} \tag{37}$$

where $c > 0$ is a universal constant.

In the case of unconditional convex bodies, the logarithmic factor in (37) is not needed at all, as proven in Bobkov and Nazarov [15]. In general, it is not known whether the logarithmic factor is necessary, or equivalently, whether any convex body admits at least one uniformly sub-Gaussian direction. We move on to discuss the question of the existence of *approximately Gaussian directions*. A conjecture that appears in the works of Anttila, Ball, and Perissinaki [2] and Brehm and Voigt [28] suggests that any high-dimensional convex body admits at least one approximately Gaussian direction. That is, it was conjectured that whenever X is a random vector in $\mathbb{R}^n$, uniformly distributed in some convex body, then there exists $0 \neq \theta \in \mathbb{R}^n$ such that the random variable

$$\langle X, \theta \rangle$$

is approximately a Gaussian random variable. The degree of the approximation is expected to improve when the dimension n increases. The conjecture clearly holds true in the case where X is uniform in an n-dimensional cube, by the classical central limit theorem, and in the case where X is uniform in a Euclidean ball or an ellipsoid, by the so-called Maxwell observation. The conjecture has turned out to be true in general, as proven by the first-named author [53]:

Theorem 3.3 ("Central Limit Theorem for Convex Bodies") *There exists a sequence $\varepsilon_n \searrow 0$ for which the following holds: let $K \subset \mathbb{R}^n$ be a convex body, and let X be a random vector that is distributed uniformly in K. Then, there exist a unit vector $\theta \in S^{n-1}$, $t_0 \in \mathbb{R}$ and $V > 0$ such that*

$$\sup_{A \subseteq \mathbb{R}} \left| \mathbb{P}\{\langle X, \theta \rangle \in A\} - \frac{1}{\sqrt{2\pi V}} \int_A e^{-\frac{(t-t_0)^2}{2V}} dt \right| \leq \varepsilon_n,$$

where the supremum runs over all measurable sets $A \subseteq \mathbb{R}$.

Moreover, if the convex body $K \subseteq \mathbb{R}^n$ is an isotropic body of volume one, then there exists a subset $\Theta \subseteq S^{n-1}$ with $\sigma(\Theta) \geq 1 - \varepsilon_n$ such that for all $\theta \in \Theta$,

$$\sup_{A \subseteq \mathbb{R}} | \mathbb{P}\{ \langle X, \theta \rangle \in A \} - \mathbb{P}\{ Z \in A \} | \leq \varepsilon_n,$$

where Z is a Gaussian random variable of mean zero and variance L_K^2.

The bound obtained in [54] for ε_n is $\varepsilon_n \leq C/n^{\alpha}$ where $C, \alpha > 0$ are universal constants and $\alpha \geq 1/15$.

Theorem 3.3 exposes a universal property of high-dimensional convex bodies: they all have approximately Gaussian one-dimensional marginals. Moreover, most of these marginals of a high-dimensional convex body, with the isotropic normalization, are approximately Gaussian. In fact, this phenomena is not restricted to one-dimensional marginals. As was proven by Eldan and the first-named author [33], when one projects the uniform measure of an istoropic convex body $K \subset \mathbb{R}^n$ to a random k-dimensional subspace E with $k \leq cn^{\alpha}$, the probability measure obtained in E has a density that is approximately Gaussian, both in total variation sense and in the sense that the ratio between this density and a Gaussian density in E is very close to 1 in large parts of the subspace E. Here, $c, \alpha > 0$ are universal constants.

Interestingly, the Gaussian approximation property of convex bodies may be reformulated in terms of a *thin-shell condition*, according to a beautiful general principle that goes back to Sudakov [93] and to Diaconis and Freedman [31] (see also Anttila, Ball, and Perissinaki [2], Bobkov [14], and von Weizsäcker [98]). This principle reads as follows: suppose that X is any random vector in $\mathbb{R}^n$ with finite second moments, normalized to have mean zero and identity covariance. Then, most of the one-dimensional marginals of X are approximately Gaussian *if and only if* the random variable $|X|/\sqrt{n}$ is concentrated around the value one, i.e.,

$$\mathbb{E}\left(\frac{|X|}{\sqrt{n}} - 1\right)^2 \leq \varepsilon,$$

for a small number $\varepsilon > 0$. These assumptions imply that the Kolmogorov distance between a typical marginal of X and a Gaussian distribution is bounded by $C(\varepsilon + n^{-\alpha})$ for universal constants $C, \alpha > 0$; see the formulation in [55]. In other words, typical marginals are approximately Gaussian if and only if most of the mass of X is concentrated in a "thin spherical shell" whose radius is $\sqrt{n}$ and whose width is much smaller than $\sqrt{n}$. Theorem 3.3 is therefore parallel to the estimate

$$\frac{\sigma_K^2}{n} := \mathbb{E}\left(\frac{|X|}{\sqrt{n}L_K} - 1\right)^2 \leq \frac{C}{n^{\alpha}}, \tag{38}$$

valid for any random vector X that is distributed uniformly in an isotropic convex body $K \subseteq \mathbb{R}^n$, where $C, \alpha > 0$ are universal constants. Thus, most of the volume of

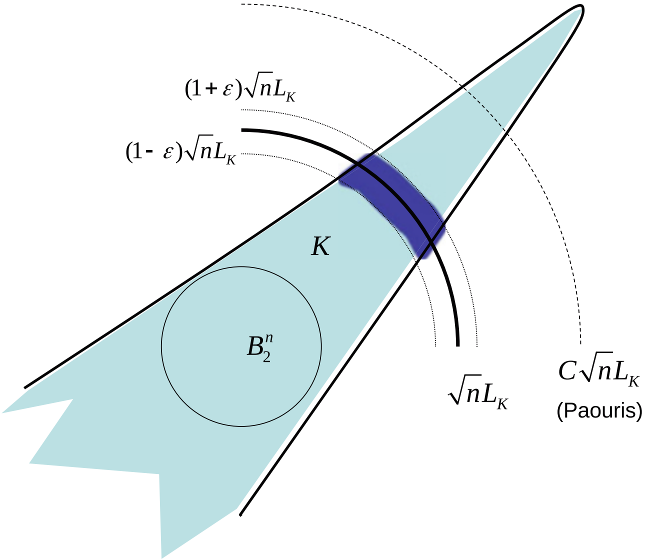


Fig. 1 The thin-shell and large deviation regimes (illustration by S. Artstein-Avidan)

a convex body in high dimensions, with the isotropic normalization, is contained in a thin spherical shell, whose width is much smaller than its radius. This complements the Paouris large deviation bound, Theorem 3.1. See Fig. 1 for an illustration.

The parameter α from (38) is related to the width of the thin spherical shell that contains most of the mass of an isotropic convex body $K \subseteq \mathbb{R}^n$. The argument in [54] leads to the estimate $\alpha \geq 1/6$, which was improved to $\alpha \geq 1/4$ by Fleury [37], to $\alpha \geq 1/3$ by Guédon and Milman [44], and then to $\alpha \geq 1/2$ by Lee and Vempala [67] who built upon a stochastic localization technique of Eldan [32]. In terms of the *thin-shell parameter* σ_K defined in (38), the current best bound due to Lee and Vempala [67] is that for any isotropic convex body $K \subseteq \mathbb{R}^n$,

$$\sigma_K \leq Cn^{1/4} \tag{39}$$

where $C > 0$ is a universal constant. The "1/4" in (39) perhaps reminds us of the best known result for the slicing problem $L_K \leq Cn^{1/4}$ from [51], which is up to logarithmic factors due to Bourgain. This is not fully a coincidence. It was proven by Eldan and the first-named author in [34] that

$$\sup_{K \subseteq \mathbb{R}^n} L_K \leq C \sup_{K \subseteq \mathbb{R}^n} \sigma_K. \tag{40}$$

Thus, any progress on the thin-shell parameter σ_K beyond the bound (39) would automatically lead to progress in the slicing problem. It was conjectured in Anttila, Ball, and Perissinaki [2] and in Bobkov and Koldobsky [16] that σ_K is bounded by a universal constant, perhaps up to a logarithmic factor. In view of (40), **the thin-shell conjecture would imply the hyperplane conjecture**.

We move on to a brief discussion of further developments related to the isoperimetric problem in convex bodies; see, e.g.. the recent survey by Lee and Vempala [67] for a thorough treatment. In addition to the isotropic constant L_K and the thin-shell parameter σ_K, an important quantity related to an isotropic convex body $K \subseteq \mathbb{R}^n$ is its *isoperimetric constant* or *Cheeger constant*, defined as follows:

$$\frac{1}{\psi_K} := L_K \cdot \inf_{A \subseteq \overline{K}} \frac{Vol_{n-1}(K \cap \partial A)}{\min\{Vol_n(A), Vol_n(K \setminus A)\}} \tag{41}$$

where the infimum runs over all subsets $A \subseteq \overline{K}$ with smooth boundary and where we recall that in this paper a convex body K is an open set and $\overline{K}$ is its closure. In the infimum in (41), we partition K into two parts so as to minimize the surface area of the interface between them; we do not include in the surface area the part of ∂A that lies on the boundary of K, only the interface between the two parts inside the convex, open set K. The reason for the normalization in (41) is the following chain of inequalities:

$$\sup_{K \subseteq \mathbb{R}^n} L_K \leq C \sup_{K \subseteq \mathbb{R}^n} \sigma_K \leq \tilde{C} \sup_{K \subseteq \mathbb{R}^n} \psi_K \leq \bar{C} n^{1/4}, \tag{42}$$

where the suprema run over all isotropic convex bodies in $\mathbb{R}^n$ and where the last inequality was proven by Lee and Vempala [67]. It was conjectured in Kannan, Lovász, and Simonovits (KLS) [48] that $\psi_K \leq C$ for any isotropic convex body $K \subseteq \mathbb{R}^n$, where $C > 0$ is a universal constant. This is a stronger conjecture than slicing, in view of (42). Ball and Nguyen [9] proved the bound $L_K \leq \exp(C\psi_K^2)$ for any isotropic convex body in $\mathbb{R}^n$. Eldan reduced the study of the isoperimetric KLS conjecture to the thin-shell conjecture, up to logarithmic factors. Denoting $\psi_n = \sup_{K \subseteq \mathbb{R}^n} \psi_K$ and $\sigma_n = \sup_{K \subseteq \mathbb{R}^n} \psi_K$, it was proven in Eldan's breakthrough paper [32] that

$$\psi_n \leq C \sqrt{\log n \cdot \sum_{\ell=1}^{n} \frac{\sigma_\ell^2}{\ell}},$$

where $C > 0$ is a universal constant. It follows that **up to factors logarithmic in the dimension, the thin-shell conjecture is equivalent to the isoperimetric KLS conjecture**. We summarize this section by noting that current progress in the thin-shell and KLS conjectures stops at $n^{1/4}$, which is the best known bound for the seemingly innocent slicing problem.

4 Bound for the Isotropic Constant

In this section, we provide some details regarding the logarithmic improvement of Bourgain's bound for the isotropic constant. This improvement is related to the following theorem due to the first-named author [51], the so-called *isomorphic version of the slicing problem*:

Theorem 4.1 *Let $K \subset \mathbb{R}^n$ be a convex body and $0 < \varepsilon < 1$. Then, there exists a convex body $T \subset \mathbb{R}^n$ such that*

(i) $(1-\varepsilon)T \subseteq K \subseteq (1+\varepsilon)T$.
(ii) $L_T < C/\sqrt{\varepsilon}$*, where* $C > 0$ *is a universal constant.*

In [56], it is proven that T from Theorem 4.1 can be additionally assumed to be a projective image of K. Recall that the projective image of a polytope is itself a polytope with the same number of vertices and faces.

The Paouris large deviation estimate, which is Theorem 3.1 above, implies the following: For any convex bodies $K, T \subseteq \mathbb{R}^n$, if

$$\left(1 - \frac{1}{\sqrt{n}}\right) T \subseteq K \subseteq \left(1 + \frac{1}{\sqrt{n}}\right) T \tag{43}$$

then

$$L_K \simeq L_T. \tag{44}$$

Indeed, since isotropic constants are invariant under affine transformations, we may assume that K is an isotropic convex body. Theorem 3.1 shows that at most an $e^{-10\sqrt{n}}$-fraction of the volume of K is located outside the ball $C\sqrt{n}L_K B^n$. It thus follows from (43) that at most an $e^{-\sqrt{n}}$-fraction of the volume of T is located outside this ball. The variational characterization (8) of the isotropic constant of T now implies that $L_T \leq CL_K$, and (44) follows by symmetry. Consequently, by substituting $\varepsilon = 1/\sqrt{n}$ in Theorem 4.1, we conclude that for any convex body $K \subseteq \mathbb{R}^n$,

$$L_K \leq Cn^{1/4},$$

as advertised. Let us now elaborate on the ideas behind the proof of Theorem 4.1. We are given a convex body $K \subseteq \mathbb{R}^n$. After translation, we may assume that the barycenter of K lies at the origin. Consider the *logarithmic Laplace transform*

$$F(x) = F_K(x) = \log \int_{\mathbb{R}^n} e^{\langle x,y \rangle} dy.$$

The function F is smooth and convex in $\mathbb{R}^n$, by the Cauchy-Schwartz inequality. In fact, for any $x \in \mathbb{R}^n$, denoting by μ_x the probability measure on K with density $y \mapsto e^{\langle x,y\rangle - F(x)} 1_K(y)$, we have

$$\nabla F(x) = \mathrm{bar}(\mu_x) \qquad \text{and} \qquad \nabla^2 F(x) = \mathrm{Cov}(\mu_x), \tag{45}$$

where $\mathrm{bar}(\mu_x) = \int y d\mu_x(y)$ is the barycenter of μ_x and where $\nabla^2 F(x)$ is the Hessian matrix of F. Note that $\nabla F(x) \in K$ for any $x \in \mathbb{R}^n$, since μ_x is a measure supported on K, and hence its barycenter is in K by convexity. The Hessian of F is positive-definite everywhere, according to (45). This convexity property of F implies that the map $x \mapsto \nabla F(x)$ is a diffeomorphism from $\mathbb{R}^n$ onto an open subset of K (which is in fact K itself).

Fix $0 < \varepsilon < 1$ as in the formulation of Theorem 4.1. We may use the point of view of "transportation of measure" and change variables as follows:

$$\int_{\varepsilon n K^\circ} \det \mathrm{Cov}(\mu_x) dx = \int_{\varepsilon n K^\circ} \det \nabla^2 F(x) dx \overset{\text{“}y=\nabla F(x)''}{=} \int_{\nabla F(\varepsilon n K^\circ)} 1 dy \leq Vol_n(K),$$

as $\nabla F(\varepsilon n K^\circ) \subseteq \nabla F(\mathbb{R}^n) \subseteq K$. In particular, there exists $x \in \varepsilon n K^\circ$ such that

$$\det \mathrm{Cov}(\mu_x) \leq \frac{Vol_n(K)}{Vol_n(\varepsilon n K^\circ)} = \varepsilon^{-n} \frac{Vol_n(K)^2}{n^n Vol_n(K) Vol_n(K^\circ)} \leq \left(\frac{C}{\varepsilon}\right)^n Vol_n(K)^2, \tag{46}$$

where the last passage follows from the Bourgain-Milman inequality (18). Let us take a closer look at the probability measure μ_x. It is a log-concave measure with density

$$\rho(y) = e^{\langle x,y\rangle - F(x)} 1_K(y).$$

Since $x \in \varepsilon n K^\circ$, we know that $|\langle x, y\rangle| \leq \varepsilon n$ for all $y \in K$. Hence,

$$\frac{\sup_K \rho}{\inf_K \rho} = \frac{\sup_{y \in K} e^{\langle x,y\rangle}}{\inf_{y\in K} e^{\langle x,y\rangle}} \leq \frac{e^{\varepsilon n}}{e^{-\varepsilon n}} = e^{2\varepsilon n}. \tag{47}$$

Recall the convex body $K(\rho)$ associated with the log-concave density ρ via formula (29). It follows from (29) and (47) that

$$(1 - C\varepsilon)K \subseteq K(\rho) \subseteq (1 + C\varepsilon)K$$

for some universal constant $C > 0$. We still need to show that $L_{K(\rho)} \leq \tilde{C}/\sqrt{\varepsilon}$, so that Theorem 4.1 would follow with $T = K(\rho)$. In view of (30), it suffices to show that

$$L_\rho \leq C/\sqrt{\varepsilon}. \tag{48}$$

However, since the barycenter of K lies at the origin, we know that $\nabla F(0) = 0$ by (45). Since F is a convex function, its critical points are global minimum points, and hence $F(x) \geq F(0) = \log Vol_n(K)$ for any $x \in \mathbb{R}^n$. Consequently, by (46),

$$L_\rho = (\sup \rho)^{1/n} \cdot \det \mathrm{Cov}(\rho)^{1/(2n)} = \exp\left(\frac{\sup_{y \in K} \langle x, y \rangle - F(x)}{n} \right) \cdot \det \mathrm{Cov}(\rho)^{1/(2n)}$$
$$\leq \exp\left(\frac{\varepsilon n - \log Vol_n(K)}{n} \right) \cdot \left(\left(\frac{C}{\varepsilon} \right)^n Vol_n(K)^2 \right)^{1/(2n)} \leq \frac{\tilde{C}}{\sqrt{\varepsilon}}.$$

This completes the proof of (48), as well as our sketch of the proof of Theorem 4.1.

It is possible to view the hyperplane conjecture as a strong conjectural version of the Bourgain-Milman inequality and of the *M-ellipsoid theory* due to the second-named author. We present two interpretations of this point of view. The first interpretation is related to the strong slicing conjecture, which suggests that for any convex body $K \subseteq \mathbb{R}^n$,

$$L_K \leq L_{\Delta^n} = \frac{(n!)^{\frac{1}{n}}}{(n+1)^{\frac{n+1}{2n}} \cdot \sqrt{n+2}}, \tag{49}$$

where $\Delta^n \subseteq \mathbb{R}^n$ is any simplex whose vertices span $\mathbb{R}^n$ and add up to zero. This conjecture holds true in two dimensions. See also Rademacher [92] for supporting evidence. On the other hand, the Mahler conjecture suggests that for any convex body $K \subseteq \mathbb{R}^n$ containing the origin in its interior,

$$Vol_n(K) Vol_n(K^\circ) \geq Vol_n(\Delta^n) \cdot Vol_n((\Delta^n)^\circ) = \frac{(n+1)^{n+1}}{(n!)^2}. \tag{50}$$

In two dimensions, the conjecture was proven by Mahler [74]; see also Meyer [75], and see Barthe and Fradelizi [11] for the case of convex bodies with symmetries. The Bourgain-Milman inequality established (50) up to a factor of c^n, for a universal constant $c > 0$. In [56], it is shown that **the strong version (49) of Bourgain's slicing conjecture implies Mahler's conjecture (50)**. Let us also mention in passing that in the centrally symmetric case, the strong version of Bourgain's slicing conjecture is that the isotropic constant is maximized for the cube. If this is true, then an old conjecture by Minkowski would follow; see Magazinov [73] and also Autissier [3]. The Minkowski conjecture suggests that for any lattice $L \subseteq \mathbb{R}^n$ of unit covolume and for any $x \in \mathbb{R}^n$ there exists $y \in L$ with $\prod_{i=1}^n |x_i - y_i| \leq 2^{-n}$.

There is also a second interpretation of the relationship between the slicing problem and the notion of the M-ellipsoid and the Bourgain-Milman inequality. For a convex body $K \subseteq \mathbb{R}^n$, an ellipsoid $\mathcal{E} \subseteq \mathbb{R}^n$ is called an *M-ellipsoid for K with constant A* if

$$Vol_n(\mathcal{E}) = Vol_n(K) \qquad \text{and} \qquad N(K, \mathcal{E}) \cdot N(\mathcal{E}, K) \leq e^{An}.$$

Here, $N(K, T) = \inf\{N\,;\ \exists x_1, \dots, x_N \in \mathbb{R}^n,\ K \subseteq \bigcup_{i=1}^{N}(x_i + T)\}$ is the covering number of K by T, the minimal number of translates of T that may cover K. The second-named author proved [79, 80] that there exists a universal constant $C > 0$, such that any convex body $K \subseteq \mathbb{R}^n$ has an M-ellipsoid with constant C. This fact plays an important role in asymptotic geometric analysis. For example, assume the normalization $Vol_n(K) = Vol_n(B^n)$, and let $u \in SL_n$ be such that $u(\mathcal{E}) = B^n$. Set $K_1 = u(K)$. Then, for any $0 < \lambda < 1$, with high probability of choosing a random $\lfloor \lambda n \rfloor$-dimensional subspace $E \subseteq \mathbb{R}^n$, the convex body

$$\mathrm{Proj}_E(K_1)$$

is a convex body of finite volume ratio, with a volume ratio constant depending solely on λ. It was observed by K. Ball [4] that for an isotropic convex body $K \subseteq \mathbb{R}^n$, the Euclidean ball of volume one is an M-ellipsoid for K with a constant depending solely on L_K. Hence, a positive solution to the slicing problem, i.e., a universal bound on L_K, would imply the theorem on the existence of the M-ellipsoid.

Note that the M-ellipsoid is an isomorphic notion: if $\mathcal{E}$ is an M-ellipsoid for K with constant A and $K/2 \subseteq T \subseteq 2K$, then a homothetic copy of $\mathcal{E}$ of volume $Vol_n(T)$ is also an M-ellipsoid of T with constant 4α. Therefore, Theorem 4.1 implies the existence of an M-ellipsoid for any K, with some universal constant $C > 0$.

Note Added in Proofs In November 2020, Yuansi Chen posted a breakthrough paper that significantly improves upon the above bounds for the isotropic constant of a general convex body in $\mathbb{R}^n$. In the notation of (42), it is shown that

$$\sup_{K \subseteq \mathbb{R}^n} L_K \le C \sup_{K \subseteq \mathbb{R}^n} \psi_K \le C_1 \exp(C_2 \sqrt{\log n \cdot \log\log n}),$$

where the suprema run over all isotropic convex bodies in $\mathbb{R}^n$. In particular, $L_K \le C_\varepsilon n^\varepsilon$ for any convex body $K \subseteq \mathbb{R}^n$ and $\varepsilon > 0$ with a coefficient $C_\varepsilon > 0$ depending solely on ε. The proof utilizes the stochastic localization technique of Eldan [32], as refined by Lee and Vempala [67]. See Chen, Y., *An Almost Constant Lower Bound of the Isoperimetric Coefficient in the KLS Conjecture.* Geom. Funct. Anal. (GAFA), Vol. 31, Issue 1, (2021). In March/April 2022, the first named author and Joseph Lehec posted a paper showing that in fact

$$\sup_{K \subseteq \mathbb{R}^n} L_K \le C \log^4 n$$

and

$$\sup_{K \subseteq \mathbb{R}^n} \psi_K \le C \log^5 n$$

where again the suprema run over all isotropic convex bodies in $\mathbb{R}^n$. The new ingredients in the proof include a certain functional analytic interpretation of the heat flow of a log-concave distribution, as well as an inequality involving dual Sobolev norms (see [58]).

References

1. Adamczak, R., Latała, R., Litvak, A. E., Oleszkiewicz, K., Pajor, A., Tomczak-Jaegermann, N.: A short proof of Paouris' inequality. Canad. Math. Bull. **57**(1), 3–8 (2014)
2. Anttila, M., Ball, K., Perissinaki, I.: The central limit problem for convex bodies. Trans. Am. Math. Soc. **355**(12), 4723–4735 (2003)
3. Autissier, P.: Un lemme matriciel effectif. Math. Z. **273**(1–2), 355–361 (2013)
4. Ball, K.: Isometric problems in ℓ_p and sections of convex sets. Ph.D. Thesis, Cambridge University, 1986
5. Ball, K.: Cube slicing in $\mathbb{R}^n$. Proc. Am. Math. Soc. **97**(3), 465–473 (1986)
6. Ball, K.: Logarithmically concave functions and sections of convex sets in $\mathbb{R}^n$. Studia Math. **88**(1), 69–84 (1988)
7. Ball, K.: Some remarks on the geometry of convex sets. Geometric aspects of functional analysis, Israel seminar (1986/87). Springer Lecture Notes in Math., vol. 1317, pp. 224–231 (1988)
8. Ball, K.: Normed spaces with a weak Gordon-Lewis property. Lecture Notes in Mathematics, vol. 1470, pp. 36–47. Springer (1991)
9. Ball, K., Nguyen, V.H.: Entropy jumps for isotropic log-concave random vectors and spectral gap. Studia Math. **213**(1), 81–96 (2012)
10. Barthe, F., Guédon, O., Mendelson, S., Naor, A.: A probabilistic approach to the geometry of the l_p^n-ball. Ann. Probab. **33**, 480–513 (2005)
11. Barthe, F., Fradelizi, M.: The volume product of convex bodies with many hyperplane symmetries. Am. J. Math. **135**(2), 311–347 (2013)
12. Brazitikos, S., Giannopoulos, A., Valettas, P., Vritsiou, B.-H.: Geometry of isotropic convex bodies. American Mathematical Society, Providence, RI (2014)
13. Berwald, L.: Verallgemeinerung eines Mittelwertsatzes von J. Favard für positive konkave Funktionen. Acta Math. **79**, 17–37 (1947)
14. Bobkov, S.G.: On concentration of distributions of random weighted sums. Ann. Prob. **31**(1), 195–215 (2003)
15. Bobkov, S.G., Nazarov, F.L.: On convex bodies and log-concave probability measures with unconditional basis. Geometric Aspects of Functional Analysis. Lecture Notes in Math., vol. 1807, pp. 53–69. Springer (2003)
16. Bobkov, S.G., Koldobsky, A.: On the central limit property of convex bodies. Geometric Aspects of Functional Analysis. Lecture Notes in Math., vol. 1807, pp. 44–52. Springer (2003)
17. Borell, C.: Inverse Hölder inequalities in one and several dimensions. J. Math. Anal. Appl. **41**, 300–312 (1973)
18. Borell, C.: Complements of Lyapunov's inequality. Math. Ann. **205**, 323–331 (1973)
19. Borell, C.: Convex measures on locally convex spaces. Ark. Mat. **12**, 239–252 (1974)
20. Borell, C.: The Brunn-Minkowski inequality in Gauss space. Invent. Math. **30**(2), 207–216 (1975)
21. Bourgain, J.: On high-dimensional maximal functions associated to convex bodies. Am. J. Math. **108**(6), 1467–1476 (1986)
22. Bourgain, J.: Geometry of Banach spaces and harmonic analysis. In: Proceedings of the International Congress of Mathematicians (ICM Berkeley 1986), pp. 871–878. Amer. Math. Soc. (1987)

23. Bourgain, J.: On the distribution of polynomials on high-dimensional convex sets. Geometric Aspects of Functional Analysis. Israel Seminar (1989–90). Lecture Notes in Math., vol. 1469, pp. 127–137. Springer (1991)
24. Bourgain, J.: On the isotropy-constant problem for "Psi-2" bodies. Geom. Aspects of Funct. Anal. Israel Seminar (2001–02). Lecture Notes in Math., vol. 1807, pp. 114–121. Springer (2002)
25. Bourgain, J., Milman, V.D.: New volume ratio properties for convex symmetric bodies in $\mathbb{R}^n$. Invent. Math. **88**(2), 319–340 (1987)
26. Bourgain, J., Klartag, B., Milman V.: A reduction of the slicing problem to finite volume ratio bodies. C. R. Acad. Sci. Paris Ser. I **336**, 331–334 (2003)
27. Bourgain, J., Klartag, B., Milman V.: Symmetrization and isotropic constants of convex bodies. Geom. aspects of Funct. Anal. Israel Seminar. Springer Lecture Notes in Math., vol. 1850, pp. 101–116. Springer (2004)
28. Brehm, U., Voigt, J.: Asymptotics of cross sections for convex bodies. Beiträge Algebra Geom. **41**(2), 437–454 (2000)
29. Busemann, H.: A theorem on convex bodies of the Brunn-Minkowski type. Proc. Natl. Acad. Sci. USA **35**, 27–31 (1949)
30. Busemann, H., Petty, C.M.: Problems on convex bodies. Mathematica Scandinavica **4**(1), 88–94 (1956)
31. Diaconis, P., Freedman, D.: Asymptotics of graphical projection pursuit. Ann. Stat. **12**(3), 793–815 (1984)
32. Eldan, R.: Thin shell implies spectral gap up to polylog via a stochastic localization scheme. Geom. Funct. Anal. (GAFA) **23**(2), 532–569 (2013)
33. Eldan, R., Klartag, B.: Pointwise estimates for marginals of convex bodies. J. Funct. Anal. **254**(8), 2275–2293 (2008)
34. Eldan, R., Klartag, B.: Approximately Gaussian marginals and the hyperplane conjecture. Concentration, Functional Inequalities and Isoperimetry. Contemp. Math., vol. 545, pp. 55–68. Amer. Math. Soc. (2011)
35. Fernique, X: Régularité de processus gaussiens. Invent. Math. **12**, 304–320 (1971)
36. Figiel, T., Tomczak-Jaegermann, N.: Projections onto Hilbertian subspaces of Banach spaces. Isr. J. Math. **53**, 155–171 (1979)
37. Fleury, B.: Concentration in a thin euclidean shell for log-concave measures. J. Funct. Anal. **259**, 832–841 (2010)
38. Fradelizi, M.: Hyperplane sections of convex bodies in isotropic position. Beiträge Algebra Geom. **40**(1), 163–183 (1999)
39. Gardner, R.J.: Geometric tomography. Second edition. Encyclopedia of Mathematics and its Applications, vol. 58. Cambridge University Press, New York (2006)
40. Giannopoulos, A., Pajor, A., Paouris, G.: A note on subgaussian estimates for linear functionals on convex bodies. Proc. Am. Math. Soc. **135**(8), 2599–2606 (2007)
41. Giannopoulos, A., Paouris, G., Valettas, P.: On the existence of subgaussian directions for log-concave measures. Concentration, Functional Inequalities and Isoperimetry. Contemp. Math., vol. 545, pp. 103–122. Amer. Math. Soc. (2011)
42. Giannopoulos, A., Paouris, G., Vritsiou, B.-H.: The isotropic position and the reverse Santaló inequality. Isr. J. Math. **203**(1), 1–22 (2014)
43. Gromov, M., Milman, V.D.: Brunn theorem and a concentration of volume phenomena for symmetric convex bodies. Israel Seminar on Geometrical Aspects of Functional Analysis (1983/84), 12 p. Tel Aviv Univ., Tel Aviv (1984)
44. Guédon, O., Milman, E.: Interpolating thin-shell and sharp large-deviation estimates for isotropic log-concave measures. Geom. Funct. Anal. (GAFA) **21**(5), 1043–1068 (2011)
45. Hensley, D.: Slicing convex bodies—bounds for slice area in terms of the body's covariance. Proc. Am. Math. Soc. **79**(4), 619–625 (1980)
46. Junge, M.: Hyperplane conjecture for quotient spaces of L_p. Forum Math. **6**(5), 617–635 (1994)

47. Junge, M.: Proportional subspaces of spaces with unconditional basis have good volume properties. Geometric Aspects of Functional Analysis (Israel, 1992–1994). Oper. Theory Adv. Appl., vol. 77, pp. 121–129. Birkhäuser (1995)
48. Kannan, R., Lovász, L., Simonovits M.: Isoperimetric problems for convex bodies and a localization lemma. J. Discr. Comput. Geom. **13**, 541–559 (1995)
49. Kashin, B.S.: The widths of certain finite-dimensional sets and classes of smooth functions. Izv. Akad. Nauk SSSR Ser. Mat. **41**(2), 334–351 (1977). English translation: Math. USSR-Izv. **11**(2), 317–333 (1978)
50. Klartag, B.: An isomorphic version of the slicing problem. J. Funct. Anal. **218**(2), 372–394 (2005)
51. Klartag, B.: On convex perturbations with a bounded isotropic constant. Geom. and Funct. Anal. (GAFA) **16**(6), 1274–1290 (2006)
52. Klartag, B.: Uniform almost sub-gaussian estimates for linear functionals on convex sets. Algebra Anal. (St. Petersburg Math. J.) **19**(1), 109–148 (2007)
53. Klartag, B.: A central limit theorem for convex sets. Invent. Math. **168**, 91–131 (2007)
54. Klartag, B.: Power-law estimates for the central limit theorem for convex sets. J. Funct. Anal. **245**, 284–310 (2007)
55. Klartag, B.: High-dimensional distributions with convexity properties. Proc. of the Fifth Euro. Congress of Math., Amsterdam, July 2008, pp. 401–417. Eur. Math. Soc. Publishing House (2010)
56. Klartag, B.: Isotropic constants and Mahler volumes. Adv. Math. **330**, 74–108 (2018)
57. Klartag, B., Koldobsky, A.: An example related to the slicing inequality for general measures. J. Funct. Anal. **274**(7), 2089–2112 (2018)
58. Klartag, B., Lehec, J.: Bourgain's slicing problem and KLS isoperimetry up to polylog. Preprint, arXiv:2203.15551
59. Klartag, B., Kozma, G.: On the hyperplane conjecture for random convex sets. Isr. J. Math. **170**, 253–268 (2009)
60. Klartag, B., Milman, V.D.: Isomorphic Steiner symmetrization. Invent. Math. **153**(3), 463–485 (2003)
61. Klartag, B., Milman, V.: Rapid Steiner symmetrization of most of a convex body and the slicing problem. Combin. Probab. Comput. **14**(5–6), 829–843 (2005)
62. Klartag, B., Milman, V.: Geometry of log-concave functions and measures. Geom. Dedicata **112**, 169–182 (2005)
63. Klartag, B., Milman, E.: Centroid bodies and the logarithmic Laplace transform – a unified approach. J. Funct. Anal. **262**(1), 10–34 (2012)
64. Koldobsky, A.: Fourier analysis in convex geometry. Mathematical Surveys and Monographs, vol. 116. American Mathematical Society, Providence, RI (2005)
65. König, H., Meyer, M., Pajor, A.: The isotropy constants of the Schatten classes are bounded. Math. Ann. **312**, 773–783 (1998)
66. Kuperberg, G.: From the Mahler conjecture to Gauss linking integrals. Geom. Funct. Anal. (GAFA) **18**(3), 870–892 (2008)
67. Lee, Y.T., Vempala, S.: Eldan's stochastic localization and the KLS hyperplane conjecture: an improved lower bound for expansion. Symposium on Foundations of Computer Science (FOCS), pp. 998–1007. IEEE Computer Soc. (2017)
68. Lee, Y.T., Vempala, S.: The Kannan-Lovász-Simonovits conjecture. Current Developments in Mathematics 2017, pp. 1–36. Int. Press (2019)
69. Lewis, D.R.: Ellipsoids defined by Banach ideal norms. Mathematica **26**, 18–29 (1979)
70. Litvak, A.E., Milman, V.D., Schechtman, G.: Averages of norms and quasi-norms. Math. Ann. **312**(1), 95–124 (1998)
71. Loomis, L.H., Whitney, H.: An inequality related to the isoperimetric inequality. Bull. Am. Math. Soc. **55**, 961–962 (1949)
72. Lutwak, E., Zhang, G.: Blaschke-Santaló inequalities. J. Differ. Geom. **47**(1), 1–16 (1997)
73. Magazinov, A.: A proof of a conjecture by Haviv, Lyubashevsky and Regev on the second moment of a lattice Voronoi cell. Adv. Geom. **20**(1), 117–120 (2020)

74. Mahler, K.: Ein Übertragungsprinzip für konvexe körper. Časopis Pest Mat. Fys. **68**, 93–102 (1939)
75. Meyer, M.: Convex bodies with minimal volume product in $\mathbb{R}^2$. Monatsh. Math. **112**(4), 297–301 (1991)
76. Meyer, M., Pajor, A.: On Santaló's inequality. Geometric Aspects of Functional Analysis (1987–88). Lecture Notes in Math., vol. 1376, pp. 261–263. Springer (1989)
77. Meyer, M., Pajor, A.: On the Blaschke-Santaló inequality. Arch. Math. **55**(1), 82–93 (1990)
78. Milman, E.: Dual mixed volumes and the slicing problem. Adv. Math. **207**, 566–598 (2006)
79. Milman, V.D.: Inégalité de Brunn-Minkowski inverse et applications á la théorie locale des espaces normés. [An inverse form of the Brunn-Minkowski inequality, with applications to the local theory of normed spaces] C. R. Acad. Sci. Paris Sér. I Math. **302**(1), 25–28 (1986)
80. Milman, V.D.: Isomorphic symmetrization and geometric inequalities. Geometric Aspects of Functional Analysis (1986/87). Lecture Notes in Math., vol. 1317, pp. 107–131. Springer (1988)
81. Milman, V.D.: Dvoretzky's theorem—thirty years later. Geom. Funct. Anal. **2**(4), 455–479 (1992)
82. Milman, V.D., Schechtman, G.: Asymptotic theory of finite-dimensional normed spaces. With an appendix by M. Gromov. Lecture Notes in Mathematics, vol. 1200. Springer (1986)
83. Milman, V.D., Pajor, A.: Isotropic position and inertia ellipsoids and zonoids of the unit ball of a normed n-dimensional space. Geometric Aspects of Functional Analysis (1987–88). Lecture Notes in Math., vol. 1376, pp. 64–104. Springer (1989)
84. Milman, V.D., Pajor, A.: Entropy and asymptotic geometry of non-symmetric convex bodies. Adv. Math. **152**(2), 314–335 (2000)
85. Nazarov, F.: The Hörmander proof of the Bourgain-Milman theorem. Geometric Aspects of Functional Analysis. Lecture Notes in Math., vol. 2050, pp. 335–343. Springer (2012)
86. Rogers, C.A., Shephard, G.C.: The difference body of a convex body. Arch. Math. **8**, 220–233 (1957)
87. Paouris, G.: ψ_2-estimates for linear functionals on zonoids. Geometric Aspects of Functional Analysis. Lecture Notes in Math., vol. 1807, 211–222. Springer, Berlin (2003)
88. Paouris, G.: On the Ψ_2 behavior of linear functionals on isotropic convex bodies. Studia Math. **168**(3), 285–299 (2005)
89. Paouris, G.: Concentration of mass on convex bodies. Geom. Funct. Anal. (GAFA) **16**(5), 1021–1049 (2006)
90. Pisier, G.: Holomorphic semigroups and the geometry of Banach spaces. Ann. Math. **115**(2), 375–392 (1982)
91. Pisier, G.: The volume of convex bodies and Banach space geometry. Cambridge Tracts in Mathematics, Vol. 94. Cambridge University Press, Cambridge (1989)
92. Rademacher, L.: A simplicial polytope that maximizes the isotropic constant must be a simplex. Mathematika **62**(1), 307–320 (2016)
93. Sudakov, V.N.: Typical distributions of linear functionals in finite-dimensional spaces of high-dimension. (Russian) Dokl. Akad. Nauk. SSSR **243**(6), 1402–1405 (1978). English translation in Soviet Math. Dokl. **19**, 1578–1582 (1978)
94. Szarek, S.J.: On Kashin's almost Euclidean orthogonal decomposition of ℓ_n^1. Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys. **26**(8), 691–694 (1978)
95. Szarek, S., Tomczak-Jaegermann, N.: On nearly Euclidean decomposition for some classes of Banach spaces. Compositio Math. **40**(3), 367–385 (1980)
96. Talagrand, M.: The generic chaining. Upper and Lower Bounds of Stochastic Processes. Springer (2005)
97. Talagrand, M.: Upper and lower bounds for stochastic processes. Modern Methods and Classical Problems. Springer (2014)
98. von Weizsäcker, H.: Sudakov's typical marginals, random linear functionals and a conditional central limit theorem. Probab. Theory Relat. Fields **107**(3), 313–324 (1997)

On the Work of Jean Bourgain in Nonlinear Dispersive Equations

Carlos E. Kenig

Dedicated to the memory of Jean Bourgain

Abstract In this brief note, we survey a sample of the deep and influential contributions of Jean Bourgain to the field of nonlinear dispersive equations. Bourgain also made many fundamental contributions to other areas of partial differential equations and mathematical physics (as well as to a myriad of other areas in analysis, number theory, combinatorics, theoretical computer science, and more). Quoting the citation of the American Mathematical Society L. P. Steele Prize for Lifetime Achievement awarded to Bourgain in 2018, “Jean Bourgain is a giant in the field of mathematical analysis, which he has applied broadly and to great effect.”

Jean Bourgain’s contributions to mathematics will be remembered forever. Those of us who knew him will also remember his warmth, generosity, and graciousness.

1 Introduction

In this brief note, we survey a sample of the deep and influential contributions of Jean Bourgain to the field of nonlinear dispersive equations. Bourgain also made many fundamental contributions to other areas of partial differential equations and mathematical physics (as well as to a myriad other areas in analysis, number theory, combinatorics, theoretical computer science, and more). Quoting the citation of the AMS L. P. Steele Prize for Lifetime Achievement, awarded to Bourgain in 2018, “Jean Bourgain is a giant in the field of mathematical analysis, which he has applied broadly and to great effect.”

Partially supported by NSF grant DMS-1800082. Reprinted with permission from the AMS.

C. E. Kenig (✉)
University of Chicago, Chicago, IL, USA
e-mail: cek@math.uchicago.math.edu

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_10

Jean Bourgain's contributions to mathematics will be remembered forever. Those of us who knew him will also remember his warmth, generosity, and graciousness.

2 Nonlinear Dispersive Equations: The Well-Posedness Theory Before Bourgain

The theory of nonlinear dispersive equation goes back to the nineteenth century, in connection with water waves in shallow water. The Korteweg-de Vries equation, which governs this phenomenon, was proposed by Boussinesq and by Korteweg-de Vries, in the late nineteenth century, as a way of explaining the discovery by Scott Russell (1835) of traveling waves. The generalized KdV equations (gKdV)$_k$ ($k = 1$ being the Korteweg-de Vries equation) are

$$(gKdV)_k \begin{cases} \partial_t u + \partial_x^3 u + u^k \partial_x u = 0, \ x \in \mathbb{R}, \text{ or } x \in \mathbb{T}, t \in \mathbb{R} \\ u|_{t=0} = u_0(x) \end{cases}$$

(here, $\mathbb{T}$ and $\mathbb{T}^d$ are the 1-dimensional (d-dimensional) torus). Another example of nonlinear dispersive equations is the nonlinear Schrödinger equations (NLS),

$$(NLS) \begin{cases} i\partial_t u + \Delta u \pm |u|^{p-1} u = 0, \ x \in \mathbb{R}^d, \text{ or } x \in \mathbb{T}^d \\ u|_{t=0} = u_0(x) \end{cases}$$

When $d = 1$, $p = 3$, these equations model the propagation of wave packets in the theory of water waves. The equations also appear in non-linear optics and in quantum field theory. These equations have a Hamiltonian structure and preserve mass and energy (although the energy maybe negative). For both equations, the conserved mass is $\int |u_0|^2$, where the integral is over $\mathbb{R}^d$ or $\mathbb{T}^d$. For (gKdV)$_k$, the conserved energy is $E(u_0) = \int [(\partial_x u_0)^2 - c_k u_0^{k+1}] dx$, and for (NLS), it is $E(u_0) = \int [(\nabla_x u_0)^2 \mp c_p |u_0|^{p+1}] dx$, where the integrals are over $\mathbb{R}^d$ or $\mathbb{T}^d$.

These equations are called dispersive because their linear parts are dispersive. Heuristically, the linear equations, when defined for $x \in \mathbb{R}^d$, are called dispersive, because the initial data gets "spread out" or "dispersed" by the evolution. (The linear equations can be solved by using Fourier's method). Since the mass of the solution is constant (the L^2 norm is conserved), this requires the size of the linear solution to become small for large t, the so called "dispersive effect." Note that this is a feature of linear dispersive equations, the traveling wave solutions discovered by Russell do not have this property, and they are purely nonlinear objects. Moreover, when $x \in \mathbb{T}^d$, there is no room for the solution to "spread out," and the "dispersive effect" disappears.

Even though these equations were introduced in the nineteenth century/early twentieth century, their systematic study started much later. One of the first things to understand for such equations is the "well-posedness." An equation like $(\text{gKdV})_k$ or (NLS) is said to be locally well-posed (LWP) in a space B (with $u_0 \in B$), if the equation has a unique solution u (in a suitable sense) for $u_0 \in B$, for some $T = T(u_0)$, $0 \le t \le T$, $u \in C([0,T]; B)$, and the mapping $u_0 \in B \to u \in C([0,T]; B)$ is continuous. (That is to say, in analogy with ODE, we have existence, uniqueness, and continuous dependence on the initial data). If we can take $T = +\infty$, we say that the problem is globally well-posed (GWP). Since dispersive equations are (essentially) time reversible, we can replace $[0,T]$ by $[-T,T]$. Usually in this subject, the space B is taken to be an L^2-based Sobolev space, (or sometimes a weighted L^2-based Sobolev space, with power weights, in case we are working in $\mathbb{R}^d$). The reason for using L^2-based spaces as opposed to L^p-based spaces is the failure of estimates for $u_0 \in L^p$, $p \ne 2$, in the associated linear problems. The first (LWP) results used the analogy of these problems to classical hyperbolic ones, which led (by the classical energy method and its refinements and compactness arguments [5, 6])to the (LWP) of $(\text{gKdV})_k$ in $H^s(\mathbb{R})$, for $s > \frac{3}{2}$, for $k = 1, 2, \dots$, with the same result holding in $H^s(\mathbb{T})$, and to the (LWP) of (NLS) in $H^s(\mathbb{R}^d)$, for $s > \frac{d}{2}$, with the same result holding in $H^s(\mathbb{T}^d)$. (In the case of (NLS), some restrictions on p arise also, coming from the possible lack of "smoothness" of $\alpha \to |\alpha|^{p-1}\alpha$). Here, for f defined on $\mathbb{R}^d$, we set $\widehat{f}(\xi) = \int_{\mathbb{R}^d} e^{2\pi i x\cdot\xi} f(x)dx$, $H^s(\mathbb{R}^d) = \{f : \int (1+|\xi|^2)^s |\widehat{f}(\xi)|^2 d\xi < \infty\}$ and for f defined on $\mathbb{T}^d$, we set $\widehat{f}(n) = \int_{\mathbb{T}^d} e^{2\pi i x\cdot n} f(x)dx$, $n \in \mathbb{Z}^d$, and $H^s(\mathbb{T}^d) = \{f : \sum_{n\in\mathbb{Z}^d} |\widehat{f}(n)|^2 (1+|n|^2)^s < \infty\}$. An inspection of these proofs shows that "dispersive properties" of $(\partial_t + \partial_x^3)$ or of $(i\partial_t + \Delta)$ are not used at all in the case of $\mathbb{R}^d$, and hence they remain valid for the case of $\mathbb{T}^d$. Particular cases of $(\text{gKdV})_k$ and (NLS) are closely connected to complete integrability, a theory which was first developed largely in this regard [1]. These are the cases $k = 1, 2$ in $(\text{gKdV})_k$ and $p = 3, d = 1$ in (NLS). The applicability of this method initially required high order of differentiability of the data u_0, and, in the case $x \in \mathbb{R}$, fast decay of u_0. More recently, this has been greatly improved (see [41, 42, 53]) but still only applies to a few specific cases.

In the late 1970s and early 1980s, the pioneering works of Ginibre-Velo [32–34], and Kato [45], through the use of important new advances in harmonic analysis [83, 86], led to "low regularity" (LWP) and (GWP) results for (NLS) in $\mathbb{R}^d$, culminating with the definitive results of Tsutsumi [85]and Cazenave-Weissler [22]. This approach exploited the "dispersive properties" of $(i\partial_t + \Delta)$ and the connection with the "restriction problem" for the Fourier transform (discovered and formulated in the visionary work of E.M. Stein (see [81]) uncovered by Segal [78] and Strichartz [83]).

More precisely, the solution of the initial value problem

$$(LS) \begin{cases} i\partial_t u + \Delta u = 0, \ x \in \mathbb{R}^d, t \in \mathbb{R} \\ u|_{t=0} = u_0(x) \end{cases}$$

is given by $\widehat{u}(\xi, t) = e^{it|\xi|^2}\widehat{u}_0(\xi) = (e^{it\Delta}u_0)\hat{(}\xi)$ or, $u(x,t) = \frac{c_d}{|t|^{\frac{d}{2}}} \int_{\mathbb{R}^n} e^{i|x-y|^2/4t}$ $u_0(y)dy$.

The second formula gives that, for u solving (LS),

$$|u(x,t)| \leq \frac{c_d}{|t|^{\frac{d}{2}}} \|u_0\|_{L^1}, \tag{1}$$

which clearly shows the "dispersive effect" mentioned earlier. The relevant "restriction problem" here is the one to the paraboloid $= \{(\xi, |\xi|^2) : \xi \in \mathbb{R}^d\} \subset \mathbb{R}^{d+1}$. In this case, we have the "restriction" inequality (for $f \in \mathscr{S}(\mathbb{R}^{d+1})$)

$$\left(\int |\widehat{f}(\xi, |\xi|^2)|^2 d\xi \right)^{\frac{1}{2}} \lesssim \|f\|_{L^{\frac{2(d+2)}{d+4}}(\mathbb{R}^{d+1})} \tag{2}$$

(see [83, 86]). The connection with (LS) is that the dual inequality to (2) is the "extension inequality," which gives, from the first formula for the solution u of (LS), the estimate

$$\|u\|_{L^{\frac{2(d+2)}{d}}(\mathbb{R}^{d+1})} \lesssim \|u_0\|_{L^2(\mathbb{R}^d)}. \tag{3}$$

Now, to solve (NLS), one needs to solve (by Duhamel's principle) the equation (with the notation $e^{it\Delta}u_0 = S(t)u_0$)

$$u(t) = S(t)u_0 \pm \int_0^t S(t-t')|u|^{p-1}u(t')dt'. \tag{4}$$

This is solved by using the contraction mapping principle on spaces constructed exploiting the estimate (2) and related ones [32–34, 45].

The result of Cazenave-Weissler [22] is

Theorem 2.1 *Assume that $u_0 \in H^s(\mathbb{R}^d)$, $s \geq 0$, $s \geq s_0$, where $p - 1 = \frac{4}{d-2s_0}$. Assume also that $p - 1 > [s] + 1$ if $p - 1 \notin 2\mathbb{Z}^\star$, where $[s]$ is the greatest integer smaller than s. Then (NLS) is locally well-posed for $t \in [-T, T]$. In the subcritical case $s > s_0$, we can take $T = T(\|u_0\|_{H^s})$, in the critical case $s = s_0$, $T = T(u_0)$.*

This approach, relying on the estimates (1) and (3), uses crucially the "dispersive properties"' of $(i\partial_t + \Delta)$ in $\mathbb{R}^d$, and hence it does not apply to $\mathbb{T}^d$. On the other hand, on $\mathbb{R}^d$ it yields essentially optimal results in terms of the values of s when $B = H^s(\mathbb{R}^d)$, which greatly improve the results obtained by the energy method described earlier.

There are several motivations for hoping to have "low regularity" well-posedness results for $(\text{gKdV})_k$ and (NLS). The first one is that, if one can obtain (LWP) at the regularity level given by the conserved mass, or the conserved energy, with time of existence $T = T(\|u_0\|)_{L^2}$, or $T = T(\|u_0\|_{H^1})$, one can use the a priori control given by the conserved quantity, to obtain global well-posedness, simply iterating the local result. Another one is the belief that, since for the associated linear problem we have well-posedness in H^s, for any s, the threshold $\bar{s}$ for the nonlinear problem gives information on the nonlinear effects present in the problem. We will see later another motivation, at very low regularity levels, stemming from the connection with quantum field theory and giving global well-posedness for "generic" data. Turning to the "low regularity" local well-posedness theory for $(\mathrm{gKdV})_k$, the new difficulty is the fact that the nonlinear term contains a derivative, which needs to be "recovered." One might think that the fact that $(\partial_t + \partial_x^3)$ has a "stronger dispersive effect" (we have for instance the bound $|u(x,t)| \lesssim \frac{1}{t^{1/3}} \|u_0\|_{L^1}$ for the linear solution, which is stronger for small t than the $\frac{1}{t^{1/2}}$ we get for (LS), $d = 1$) would compensate for the derivative in the nonlinearity, but this is not obviously the case. Kato [43, 44] found a "local smoothing" effect for solutions of $(\mathrm{gKdV})_k$ which allowed, when $x \in \mathbb{R}$, to control "a priori," with $u_0 \in L^2(\mathbb{R})$ quantities like $\int_j^{j+1} \int_0^1 \Big(\partial_x u(x,t)\Big)^2 dxdt,\ j \in \mathbb{Z}$, uniformly in j, but this only gave rise to "weak solutions" with L^2 data, but did not give uniqueness or continuous dependence on the data. This was also restricted to $x \in \mathbb{R}$, since such an estimate in $\mathbb{T}$ would contradict time reversibility and conservation of mass. In the 1980s and early 1990s, in a joint project with G. Ponce and L. Vega, we developed a new approach to the "low regularity" local and global well-posedness theory (for $x \in \mathbb{R}$) for $(\mathrm{gKdV})_k$, which in the case $k \geq 4$ gave essentially optimal (in some sense) results [4, 51]. This was also based on the contraction mapping theorem and used tools from harmonic analysis. In addition to the analogs of the "extension inequality" (3), (with $(\xi, |\xi|^2)$ being replaced by (ξ, ξ^3)), we used a sharp form (for linear equations) of the Kato"local smoothing" estimate, introduced in [30, 82, 87], as well as an analog of the "maximal function" estimate introduced in [21] and motivated by statistical mechanics (see also [31, 87]). The combination of these two estimates allowed us to control well the nonlinear term $u^k \partial_x u$. In addition, we also applied the multilinear harmonic analysis tools developed by Coifman-Meyer [23, 24]. This was all completely tied to dispersion and was totally dependent or the fact that $x \in \mathbb{R}$. A sample result obtained, for KdV ($k = 1$), was

Theorem 2.2 ([49]) *Let* $s > \frac{3}{4}$, $u_0 \in H^s(\mathbb{R})$. *Then,* $\exists T = T(\|u_0\|_{H^s})$, *and a space* $X_T^s \subset C([-T,T]; H^s)$, *such that KdV has a unique solution* $u \in X_T^s$, *which depends continuously on* u_0.

The space X_T^s is constructed by using the estimates mentioned earlier, namely the sharp "local smoothing" estimate, the "maximal function" estimate, and the variants of the "extension estimate." One then proves the result by the contraction mapping principle in the space X_T^s, $T = T(\|u_0\|_{H^s})$, showing that the mapping $\Phi_{u_0}(u) = W(t)u_0 + \int_0^t W(t-t')(u\partial_x u)(t')dt'$ has a fixed point in X_T^s, where $\widehat{W(t)f}(\xi) = e^{it\xi^3} \widehat{f}(\xi)$.

Remark 1 The approach was, in a certain sense, sharp: if we have a space X_T^s such that $\forall u_0 \in H^s(\mathbb{R})$, the linear solution $W(t)u_0$ belongs to X_T^s and such that, for all $v, w \in X_T^s$, we have $v\partial_x w \in L^1_{loc}(\mathbb{R})$ and then $s \geq \frac{3}{4}$.

At this point, we had no idea on how to improve the results for $k = 1, 3$ (the $k = 2$ result in [49] was also "optimal," as was shown in [51]), or how to do anything other than the $s > \frac{3}{2}$ result given by the energy method in the case $x \in \mathbb{T}$.

3 Bourgain's Transformative Work on the Well-Posedness Theory of Dispersive Equations

In the spring of 1990, I gave a lecture on the work (then in progress) in [49], and E. Speer was in the audience. He asked me the following question: consider the quintic (NLS) on $\mathbb{T}$:

$$\begin{cases} i\partial_t u + \Delta u \pm |u|^4 u = 0, \; x \in \mathbb{T}, t \in \mathbb{R} \\ u|_{t=0} = u_0(x) \in H^s(\mathbb{T}). \end{cases} \tag{5}$$

Is this problem well-posed for $s < \frac{1}{2}$?

I knew that the energy method gave $s > \frac{1}{2}$, that complete integrability did not apply, and that the methods we developed with Ponce and Vega, which relied on dispersion, did not apply. Speer explained the reason for the question, which was in connection with the work [56] of Lebowitz, Rose, and Speer, in which they had constructed a Gibbs measure associated to the problem (5). The points that the authors of [56] were concerned with were that the measure they constructed used the periodic setting crucially and that the support of the measure was contained in very low regularity spaces. So, they wanted to have a flow for (5), in the support of the Gibbs measure, which kept the Gibbs measure invariant. If so, a by-product of all this would be that, for data in the support of the measure, local in time existence could be globalized in time, similarly to the arguments in the presence of conserved quantities that we saw before. I told Speer that I felt that the question was very hard and that I thought that the person who could make progress in it, and would probably be interested in the problem, was Jean Bourgain! Bourgain did get interested and resolved completely the Lebowitz-Rose-Speer questions [7, 8, 10]. In doing so, he transformed the theory of nonlinear dispersive equations, starting with his papers [7–9]. Moreover, he continued making fundamental contributions to all aspects of this theory and transformed not only the well-posedness theory and created the probabilistic theory suggested by [10, 11], and [56] but also many other central areas in the field. Let me now turn to Bourgain's papers [7, 8], in which he made his first groundbreaking contributions to the well-posedness theory. These works address the following two fundamental questions:

1. How to prove low regularity well-posedness results for (NLS) and $(\text{gKdV})_k$, for $x \in \mathbb{T}^d$?
2. How to improve the well-posedness results on (KdV) on $\mathbb{R}$?

It turns out that, in solving the first question, Bourgain also found the path to solving the second one. Also, once the first question was solved, Bourgain turned to the Gibbs measure questions from [56], in [10, 11], settling them and extending their scope, as we shall see below. We thus turn to (NLS) on $\mathbb{T}^d$, and we will concentrate on Bourgain's results for $d = 1, 2$, which are the most relevant to our exposition.

Theorem 3.1 **([7])**

(i) (NLS) is locally well-posed in $H^s(\mathbb{T})$, *for* $s \geq 0$, $p - 1 < \frac{4}{1-2s}$. *Thus, for* $p - 1 = 4$, *(NLS) is (LWP) in* $H^s(\mathbb{T})$ *for all* $s > 0$.
(ii) (NLS) is locally well-posed in $H^s(\mathbb{T}^2)$, *for* $p - 1 = 2$, $s > 0$.

Compared with corresponding results in $\mathbb{R}, \mathbb{R}^2$, that we discussed earlier, one key difficulty is the lack of a "dispersive effect." Another difficulty is that, in the periodic case, the Fourier transform, in the solution of the associated linear problem, is replaced by Fourier series, leading to "exponential sums" that are much more difficult to estimate than integrals. For instance, the operator $e^{it\Delta}u_0 = S(t)u_0$, now takes the form

$$S(t)u_0(x) = \sum_{n \in \mathbb{Z}^d} e^{i(xn + t|n|^2)} \widehat{u}_0(n).$$

The proof the Theorem 3.1 proceeds by using the contraction mapping principle. The first step is to find estimates that replace the inequality (3), crucial in the case of $\mathbb{R}^d$, which is proved using oscillatory integral estimates. Bourgain achieved this by using analytic number theory, and the results that he obtained in doing this have independent interest in analytic number theory. As a sample, let me mention two such estimates:

(a)

$$\| \sum_{n \in \mathbb{Z}, |n| \leq N} a_n e^{i(nx + n^2 t)} \|_{L^6(\mathbb{T}^2)} \lesssim N^{\varepsilon} \Big(\sum |a_n|^2 \big)^{\frac{1}{2}}, \forall \varepsilon > 0,$$

which is used in Theorem 3.1(i).

and

(b)

$$\| \sum_{n \in \mathbb{Z}^2, |n_1| \leq N, |n_2| \leq N} a_n e^{i(nx + |n|^2 t)} \|_{L^4(\mathbb{T}^3)} \lesssim N^{\varepsilon} \Big(\sum_{n \in \mathbb{Z}^2} |a_n|^2 \Big)^{\frac{1}{2}}, \forall \epsilon > 0,$$

which is used in Theorem 3.1(ii).

Their proof uses the argument of Tomas [86] in the proof of the "restriction inequality," combined with the "major arc" description of exponential sums (due to Vinogradov) and number theoretic arguments inspired by Weyl type lemmas [88]. The second main contribution of Bourgain here is the introduction of new function spaces in which to apply the contraction mapping principle.

For K, N positive integers, consider $\Lambda_{K,N} = \{\zeta = (\xi, \lambda) \in \mathbb{Z}^d \times \mathbb{R} : N \le |\xi| \le 2N$ and $K \le |\lambda - |\xi|^2| \le 2K\}$. For a function u in $L^2(\mathbb{T}^d \times \mathbb{R})$, let

$$u(x,t) = \sum_{\xi \in \mathbb{Z}^d} \int \widehat{u}(\zeta) e^{2\pi i(\xi x + t\lambda)} d\lambda,$$

and define $|||u|||_s = \sup_{K,N} (K+1)^{\frac{1}{2}} (N+1)^s \left(\int_{\Lambda_{K,N}} |\widehat{u}(\zeta)|^2 d\zeta \right)^{\frac{1}{2}}$.

Fixing an interval of t in $[-\delta, \delta]$, one considers the restriction norm

$$|||u|||_{X^s} = \inf |||\tilde{u}|||_s, \tag{6}$$

where the infimum is taken over all $\tilde{u}$ coinciding with u in $[-\delta, \delta]$ and shows that the integral equation has a solution in X^s, for small δ, by (4), now on $\mathbb{T}^d$, using the contraction mapping theorem. This applies to (i) and (ii) and uses crucially the bounds (a) and (b).

It is difficult to overestimate the impact of this work in the well-posedness theory. It was simply a complete game changer. While versions of the spaces just described were in the literature before, in earlier works of Rauch and Reed [76] and M. Beals [3] dealing with propagation of singularities for solutions of semilinear wave equations, and also implicit in the contemporary work of Klainerman-Machedon [55] on the local well-posedness of semilinear wave equations, the flexibility and universality of Bourgain's formulation of these spaces contributed decisively to their wide applicability in solving a large number of previously intractable problems, in the work of many researchers.

We now turn to the work in [8], on $(gKdV)_k$, on $\mathbb{T}$. We will restrict ourselves to commenting on the results for $k = 1$.

Theorem 3.2 ([8]) *(KdV) is locally well-posed on $L^2(\mathbb{T})$, with time of existence depending on $\|u_0\|_{L^2}$, and hence by conservation of the L^2 norm, it is globally well-posed in $L^2(\mathbb{T})$.*

The proof also proceeds by a contraction mapping argument, in spaces related to the ones given by (6) but adapted to the linear operator $\partial_t + \partial_x^3$. A first reduction is to the case of data of integral 0, that is, whose zero Fourier coefficient vanishes. The space X_s now has norm

$$|||u|||_s = \left\{ \sum_{n \in \mathbb{Z}, n \neq 0} |n|^{2s} \int_{-\infty}^{+\infty} (1 + |\lambda - n^3|) |\widehat{u}(n,\lambda)|^2 \, d\lambda \right\}^{1/2}$$

for u defined for $(x,t) \in \mathbb{T}^2$, with mean in x equal to 0. The relevant version of (a), when $s = 0$, is now

(a')

$$\|f\|_{L^4(\mathbb{T}^2)} \lesssim \left(\sum_{m,n\in\mathbb{Z}} (1 + |n - m^3|)^{2/3} |\hat{f}(m,n)|^2 \right)^{1/2} .$$

A very important difference with (NLS) is the fact that there is a derivative in the non-linearity, and no linear local smoothing effect, as we mentioned earlier. Bourgain's crucial insight here was that there is a nonlinear smoothing effect, best captured by the function spaces introduced above. This is given in the following estimates: let $w(x,t) = \partial_x(u^2)(x,t)$, where we assume that $\int_{\mathbb{T}} u(x,t)\,dx = 0$. Then, for $s \geq 0$,

$$\left(\sum_{n\neq 0} |n|^{2s} \int \frac{|\widehat{w}(n,\lambda)|^2}{(1 + |\lambda - n^3|)}\, d\lambda \right)^{1/2} \lesssim |||u|||_{X_s},$$

$$\left(\sum_{n\neq 0} |n|^{2s} \left(\int \frac{|\widehat{w}(n,\lambda)|}{(1 + |\lambda - n^3|)}\, d\lambda \right)^2 \right)^{1/2} \lesssim |||u|||_{X_s},$$

It is through these estimates, controlling $\partial_x(u^2)$ by u, that we see this nonlinear smoothing effect, which is a consequence of the "curvature" of (n, n^3).

Finally, also in [8], Bourgain observed that this nonlinear smoothing effect also carries over to the case $x \in \mathbb{R}$, using the function spaces

$$X_b^s = \left\{ u(x,t) : \iint (1 + |\lambda - \xi^3|)^{2b} \cdot |1 + |\xi||^{2s} |\widehat{u}(\xi,\lambda)|^2\, d\xi\, d\lambda < \infty,\ \text{where } (\xi,\lambda) \in \mathbb{R}^2 \right\}.$$

He proved:

Theorem 3.3 ([8]) *(KdV) is globally well-posed in* $L^2(\mathbb{R})$.

Remark 2 By using a nonlinear smoothing effect, and thus replacing $v\partial_x w$ in Remark 1 by $\partial_x(u^2)$, Bourgain bypassed the objection for improving $s > \frac{3}{4}$, given in Remark 1. To Ponce, Vega, and myself, this was a shocking observation. Of course, this was just one of the many shocking observations made by Bourgain over the years! These works of Bourgain have been and continue to be remarkably influential.

Remark 3 Theorem 3.2 and Theorem 3.3 generated substantial interest in the question of finding the optimal s for (LWP) in each theorem. In [50], it was shown that (LWP) for $\mathbb{T}$ holds for $s > -\frac{1}{2}$ and for $\mathbb{R}$ for $s > -\frac{3}{4}$, both by the contraction mapping principle. In [12], Bourgain observed that (LWP) cannot be proved by the contraction mapping principle, for $s < -\frac{1}{2}$ on $\mathbb{T}$ and for $s < -\frac{3}{4}$ on $\mathbb{R}$. In [40] and [54], it was shown (independently) that (LWP) holds in $H^{-\frac{1}{2}}(\mathbb{T})$ and $H^{-\frac{3}{4}}(\mathbb{R})$, by the contraction mapping principle, using a modification of the spaces X_b^s introduced by Bourgain. That a modification of the spaces was needed was shown by Nakanishi, Takaoka, and Tsutsumi [72]. Finally, (LWP) was shown in $H^{-1}(\mathbb{T})$ by Kappeler-Topalov [42] and by Killip-Visan in $H^{-1}(\mathbb{R})$ [53], using inverse scattering. These are the optimal spaces for (LWP) in the scale of Sobolev spaces, as was shown by Molinet [69, 70].

4 A Quick Sampling of Some of the Other Groundbreaking Contributions of Bourgain to Nonlinear Dispersive Equations

4.1 Gibbs Measure Associated to Periodic (NLS)

We again consider the (NLS) equation

$$\begin{cases} i\partial_t u + \Delta u \pm |u|^{p-1}u = 0, \, p > 1, \, u : \mathbb{T}^d \times \mathbb{R} \to \mathbb{C} \\ u|_{t=0} = u_0 \end{cases}$$

and recall the two conserved quantities: the mass

$$M(u) = \int_{\mathbb{T}^d} |u|^2 \, dx = M(u_0)$$

and the Hamiltonian (the energy)

$$H(u) = \frac{1}{2} \int_{\mathbb{T}^d} |\nabla u|^2 \, dx \pm \frac{1}{p+1} \int_{\mathbb{T}^d} |u|^{p+1} \, dx = H(u_0).$$

If we set $\hat{u}(n,t) = a_n(t) + ib_n(t)$, we see that u solves (NLS) if and only if $\dot{a}_n(t) = \frac{\partial H}{\partial b_n}$ and $\dot{b}_n(t) = -\frac{\partial H}{\partial a_n}$, $n \in \mathbb{Z}^d$. Thus, (NLS) can be viewed as an infinite-dimensional Hamiltonian system. If the Hamiltonian system is finite-dimensional, say we consider $|n| \leq N$, then the Gibbs measure $d\mu$, given by

$$d\mu = \frac{1}{Z_N} e^{-H(a_n,b_n)} \prod_{|n|\leq N} da_n \, db_n,$$

where Z_N is a normalization constant, is well-defined and invariant with respect to the flow. In the paper [56], Lebowitz-Rose-Speer were able to make sense of the

Gibbs measure associated to (NLS) in $\mathbb{T}$, with $p = 5$. They considered the formal expression

$$\text{“}d\mu = \frac{1}{Z} e^{-H(a_n, b_n)} \prod_{n\in\mathbb{Z}} da_n\, db_n\text{"},$$

by introducing first the Gaussian measure

$$d\rho = \frac{1}{\tilde{Z}} e^{-\sum_n (1+n^2)(|a_n|^2+|b_n|^2)} \prod_n da_n\, db_n,$$

with support in $H^s(\mathbb{T})$, $s < \frac{1}{2}$, and then proved that $d\mu$ is absolutely continuous with respect to $d\rho$. The questions they formulated were as follows:

1. Is (NLS) on $\mathbb{T}$, with $p = 5$, on $H^s(\mathbb{T})$, $0 < s < \frac{1}{2}$, well-defined for all times, at least for data in the support of the measure?
2. Is $d\mu$ invariant with respect to the (NLS) flow?

In the paper [10], Bourgain answered both questions in the positive. To treat both issues, he used the (LWP) result in H^s, $0 < s < \frac{1}{2}$, given in Theorem 3.1, and then used the invariance of the measure under the flow to establish global well-posedness almost surely $d\mu$.

Bourgain then treated in [11] a very challenging question along these lines: Can one do this for the cubic (NLS) on $\mathbb{T}^2$, at least in the defocusing case, that is, for the equation

$$i\partial_t u + \Delta u - |u|^2 u = 0, x \in \mathbb{T}^2?$$

The existence of $d\mu$ in this case was due to Glimm-Jaffe [36], but supp$\mu \subset H^s(\mathbb{T}^2)$, $s < 0$, while Theorem 3.1 gives (LWP) in $H^s(\mathbb{T}^2)$, $s > 0$.

Bourgain overcame this difficulty through another shocking breakthrough. He considered the following random data:

$$u_0^\omega = \sum_{n\in\mathbb{Z}^2} \frac{g_n(\omega)}{(1+|n|^2)^{\frac{1}{2}}} e^{inx},$$

where the $\{g_n\}$ are identically distributed complex Gaussian random variables. Since $u_0^\omega \in H^s(\mathbb{T}^2)$, $s < 0$, u_0^ω belongs to the support of the Gibbs measure μ. (We are going to ignore here the need for “Wick-ordering” the (NLS) equation here; see [11]). The key observation is that if u is the (NLS) solution, $w(t) = u(t) - S(t)u_0^\omega$ is (almost surely in ω) well-defined in $H^{\bar{s}}(\mathbb{T}^2)$, where $\bar{s} > 0$, and one can then solve for w, to obtain a local in time solution. Finally, the local in time solution is extended globally in time, using the invariance of the Gibbs measure. This very influential paper led to the notion of “probabilistic well-posedness” in dispersive equations in works of Burq-Tzvelkov [20], T. Oh [73], and many others, including Bourgain-Bulut [17, 18].

4.2 Bourgain's "High-Low Decomposition"

In Theorem 2.1, the local in time result can be extended to a global in time one, in case the H^s norm of the data is small, $s \geq s_0$. In the mass (L^2) subcritical case, when $p - 1 < 4/d$, that is when $s_0 < 0$, the problem is locally well-posed in L^2 and hence globally well-posed in L^2. When $p - 1 \geq 4/d$, in the focusing case, that is when the sign in front of the nonlinearity in (NLS) is negative, and hence the Hamiltonian does not have a definite sign, sufficiently large smooth solution may blow-up in finite time (see Glassey [35], Merle [58, 59], Bourgain-Wang [19], Merle-Raphaël [60–64], Raphaël [74, 75], Merle-Raphaël-Rodnianski [65], etc.). Also, if the nonlinearity is "defocusing," that is, the sign in front of the nonlinear term in (NLS) is negative so that the conserved Hamiltonian

$$H(u) = \frac{1}{2}\int |\nabla u|^2 + \frac{1}{p+1}\int |u|^{p+1},$$

controls $\int |\nabla u|^2$, and if $p - 1 < \frac{4}{d-2}$ (that is $s_0 < 1$) and hence the problem is energy subcritical, (NLS) is globally well-posed in the energy sphere $H^1(\mathbb{R}^d)$, by iterating the result in Theorem 2.1.

Bourgain [13] developed a very general method to, in such circumstances, obtain global well-posedness below the energy norm. A sample result is

Theorem 4.1 ([13]) *The problem*

$$\begin{cases} i\partial_t u + \Delta u - u|u|^2 = 0 \\ u|_{t=0} = u_0 \in H^s(\mathbb{R}^2) \end{cases}$$

is globally well-posed for $s > \frac{3}{5}$. Moreover, the solution u satisfies $u(t) - S(t)u_0 \in H^1(\mathbb{R}^2)$ for all t (with a polynomial control in $|t|$ of the H^1 norm).

The general scheme of the method is as follows: first, one has to have a conserved quantity (say $I(u_0)$), such that $I(u_0)$ controls a certain H^{s_0} norm. Next, one needs a local well-posedness result (LWP) in H^{s_1}, for $s_1 < s_0$, with the flow map satisfying $I(u(t)-S(t)u_0) \leq F(\|u_0\|_{H^{s_1}})$, where $S(t)$ is the associated linear evolution, acting unitarily on all H^s spaces. One then expects a global well-posedness result in H^{s_2}, for some $s_1 < s_2 < s_0$. In the theorem stated, I is the Hamiltonian. One then splits, for some T large and fixed, $u_0 = u_{0,1}^{(N_0)} + u_{0,2}^{(N_0)}$, with $u_{0,1}^{(N)} = \int_{|\xi|\leq N_0} \hat{u}_0(\xi)e^{ix\cdot\xi}\, d\xi$, where $N_0 = N_0(T)$ is to be chosen.

It is simple to see that $H(u_{0,1}^{(N_0)}) \lesssim N_0^{2(1-s)}$. One then solves the nonlinear problem with initial data $u_{0,1}^{(N_0)}$, for all times. If we choose the time interval $I = [0, \delta]$, where $\delta = N_0^{-2(1-s)-\epsilon}$,

$$\|u_{0,1}^{(N_0)}\|_{L^4(\mathbb{R}^d \times I)} = o(1).$$

If we let $u = u_1^{(N_0)} + v$, where $u_1^{(N_0)}$ is the global solution just mentioned, v satisfies the difference equation

$$\begin{cases} i\partial_t v + \Delta v - 2|u_1^{(N_0)}|^2 v - (u_1^{(N_0)})^2 \bar{v} - \overline{(u_1^{(N_0)})} v^2 - 2u_1^{(N_0)}|v|^2 - |v|^2 v = 0 \\ v|_{t=0} = u_{0,2}^{(N_0)}, \end{cases}$$

with $\|u_{0,2}^{(N_0)}\|_{L^2} \lesssim N_0^{-s}$; $\|u_{0,2}^{(N_0)}\|_{H^s} \leq C$. One then gets, after calculations, $v = S(t)(u_{0,2}^{(N_0)}) + w$, where $w(t) \in H^1$, $\|w(t)\|_{L^2} \lesssim N_0^{-s}$ and $\|w(t)\|_{H^1} \lesssim N_0^{1-2s+\epsilon}$.

Then, fixing $t_1 = \delta$, we obtain $u(t_1) = u_1 + v_1$, where $u_1 = u_1^{(N_0)}(t_1) + w(t_1)$, $v_1 = S(t_1)(u_{0,2}^{(N_0)})$. Using the conservation of H, and the bounds for w, this yields

$$H(u_1) \leq H(u_0) + CN_0^{2-3s+\epsilon},$$

while v_1 has the same properties as $u_{0,2}^{(N_0)}$. Iterating the procedure, to reach time T, we need a number of steps:

$$\frac{T}{\delta} \simeq T \cdot N_0^{2(1-s)+\epsilon}.$$

Thus, we need to ensure that

$$T \cdot N_0^{2(1-s)+\epsilon} \cdot N_0^{2-3s+\epsilon} < H(u_{0,1}^{(N_0)}) \approx N_0^{2(1-s)}.$$

This can be achieved for $s > \frac{2}{3}$. A more elaborate argument gives $s > \frac{3}{5}$.

This method, as mentioned before, is very general and has led to many global well-posedness results, due to many researchers, for instance, in energy subcritical, defocusing problems. The method also stimulated the "I-team" (Colliander, Keel, Staffilani, Takaoka and Tao) to develop the "I-method" to treat similar types of situations. The "I-method" has been extraordinarily successful (see, for instance, [25–28], etc.).

Besides his interest in global well-posedness for defocusing, energy subcritical (NLS), Bourgain was very interested in corresponding global in time results for energy critical and supercritical (NLS). In the next section, we will discuss Bourgain's work in the energy critical case. Understanding the global in time, energy supercritical case was a problem that Bourgain considered very natural and intriguing. In [16], Bourgain conjectured the global existence of classical solutions, with smooth, well-localized data, for defocusing energy supercritical (NLS). For years, this problem was considered out of reach. Recently, this conjecture was disproved for $d \geq 5$ in the spectacular series of papers by Merle, Raphaël, Rodnianski, and Szeftel [66, 67], who also were able to obtain corresponding results for the compressible Euler and Navier-Stokes flows [68].

4.3 Bourgain's Work on the Defocusing Energy Critical (NLS)

In the remarkable paper [14], Bourgain considered the defocusing, energy critical (NLS)

$$\begin{cases} i\partial_t u + \Delta u - |u|^{\frac{4}{d-2}} u = 0, d \geq 3 \\ u|_{t=0} = u_0 \in H^1(\mathbb{R}^d) \end{cases} \tag{7}$$

Theorem 4.2 *(7) is globally well-posed for u_0 radial, when $d = 3, 4$. Moreover, higher regularity of u_0 is preserved for all times.*

Remark 4 The result was proved independently by Grillakis [39], when $d = 3$. It was extended to all $d \geq 3$, still under u_0 radial, by Tao in 2005.

Remark 5 In addition to global well-posedness, Bourgain established scattering, that is, to say, there exist $u_0^{\pm} \in H^1(\mathbb{R}^d)$, radial such that

$$\lim_{t\to\pm\infty} \left\| u(t) - S(t)(u_0^{\pm}) \right\|_{H^1(\mathbb{R}^d)} = 0.$$

Remark 6 The corresponding result for the defocusing energy critical nonlinear wave equation

$$\begin{cases} \partial_t^2 u - \Delta u + |u|^{\frac{4}{d-2}} u = 0 \\ u|_{t=0} = u_0 \in H^1(\mathbb{R}^d) \\ \partial_t u|_{t=0} = u_1 \in L^2(\mathbb{R}^d) \end{cases}$$

was established by Struwe [84] in the radial case and by Grillakis [37, 38] in the non-radial case (see also [79, 80]), with scattering being obtained in [2]. The key idea was to use the Morawetz identity [71], which for the wave equation has energy critical scaling, combined with finite speed of propagation (another important feature of the wave equation) to prevent "energy concentration."

For the proof of Theorem 4.2, when $d = 3$, the starting point is to show that if

$$\int_0^{T_\star}\int_{\mathbb{R}^3}|u(x,t)|^{10}\,dx\,dt<\infty, \tag{8}$$

where $T_\star$ is the "final time of existence" of u, then $T_\star=\infty$ and u scatters. This fact is now referred to as "the standard finite time blow-up" criterion. In order to achieve (8), Bourgain's idea was to do so by induction on the size of the Hamiltonian of u_0 and show that

$$\|u\|_{L^{10}_x L^{10}_{[0,T_\star]}}\le M(H(u_0)),$$

for some function M. It is easy to show, from the proof of the local well-posedness result (since $\|u_0\|_{H^1}\lesssim H(u_0)$), that this is the case if $H(u_0)$ is small. Arguing by contradiction, one assumes that

$$\|u\|_{L^{10}_x L^{10}_{[0,T_\star]}}>M,$$

for some M large and that $\|v\|_{L^{10}_x L^{10}_t}<M_1$, whenever

$$\begin{cases} i\partial_t v+\Delta v-|v|^4v=0\\ v|_{t=0}=v_0,\end{cases}$$

provided $H(v_0)<H(u_0)-\eta^4$, for some small η (depending only on $H(u_0)$), and then one reaches a contradiction for large M.

In order to reach this contradiction, Bourgain introduced a modification of the Morawetz estimate for the Schrödinger equation, due to Lin-Strauss [57]. Comparing Theorem 4.2 with the earlier work on the wave equation, by Grillakis, mentioned in Remark 6, key difficulties are the infinite speed of propagation and the unfavorable scaling of the estimate in [57]. This is addressed in

Proposition 1 *Let u be a solution of (7) in the energy space on a time interval I on which (7) is well-posed in the energy space. Then,*

$$\int_I\int_{|x|<|I|^{1/2}}\frac{|u(x,t)|^6}{|x|}\,dx\,dt\le CH(u_0)|I|^{1/2}.$$

It is in the application of this Proposition (which allows one to handle energy concentration) that the radial hypothesis is used. The details of the proof are intricate. The "induction on energy" used in the proof is an audacious idea, which has been extremely influential. In [29], the "I-team" (Colliander-Keel-Staffilani-Takaoka-Tao) in a major breakthrough extended the $d=3$ result in Theorem 4.2 to the non-radial case. An important ingredient of their proof is the introduction of an "interaction Morawetz" inequality, a version of Proposition 1, in which the origin is not a privileged point. This was extended to $d=4$ by Ryckman-Visan [77] and to $d \geq 5$ by Visan [89]. Later on, a new method, dubbed the "concentration-compactness/rigidity theorem method," was introduced in [46–48], which is very flexible and which could also treat focusing problems, under sharp size conditions. This method also led to many more developments in this type of problems, in the works of many researchers. For a proof of Theorem 4.2, and its non-radial version in [29], using this new method, see the work of Killip-Visan [52].

5 Conclusion

The work of Jean Bourgain transformed the field of nonlinear dispersive equations by settling old conjectures, introducing new methods and ideas, and posing important problems. The works briefly described in this note are just a small (hopefully representative) sample of Bourgain's influential contributions to this field. They will continue to inspire researchers for generations to come.

Acknowledgments I am very grateful to Gigliola Staffilani, for sharing with me the slides of her lecture at the meeting "Honoring the Life and Work of Jean Bourgain" held at IAS, Princeton on May 31-June 01, 2019. In particular, the presentation in Section 4.1 follows very closely her exposition.

References

1. Ablowitz, M.J., Clarkson, P.A.: Solitons, nonlinear evolution equations and inverse scattering. London Mathematical Society Lecture Note Series, vol. 149. Cambridge University Press, Cambridge (1991)
2. Bahouri, H., Shatah, J.: Decay estimates for the critical semilinear wave equation. Ann. inst. H. Poincaré Anal. Non-Linéaire **15**(6), 783–789 (1998)
3. Beals, M.: Self-spreading and strength of singularities for solution to semilinear wave equation. Ann. Math. **118**, 187–214 (1983)
4. Birnir, B., Kenig, C., Ponce, G., Svanstedt, N., Vega, L.: On the ill-posedness of the IVP for the generalized Korteweg-de Vries and nonlinear Schrödinger equations. J. Lond. Math. Soc. **53**, 551–559 (1996)
5. Bona, J.L., Scott, R.: Solutions of the Korteweg-de Vries equation in fractional order Sobolev spaces. Duke Math. J. **43**, 87–99 (1976)
6. Bona, J.L., Smith, R.: The initial value problem for the Korteweg-de Vries equation. Roy. Soc. Lond. Ser. A **278**, 555–601 (1978)
7. Bourgain, J.: Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations. Part I: Schrödinger equations. Geom. Funct. Anal. (GAFA) **3**, 107–156 (1993)
8. Bourgain, J.: Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations. Part II. The KdV equation. Geom. Funct. Anal. (GAFA) **3**, 209–262 (1993)
9. Bourgain, J.: Exponential sums and nonlinear Schrödinger equations. Geom. Funct. Anal. (GAFA), 157–178 (1993)

10. Bourgain, J.: Periodic nonlinear Schródinger equations and invariant measures. Comm. Math. Phys. **160**, 1–26 (1994)
11. Bourgain, J.: Invariant measures for the 2D-defocusing nonlinear Schródinger equation. Comm. Math. Phys. **176**, 421–445 (1996)
12. Bourgain, J.: Periodic Korteweg-de Vries equation with measures as initial data. Selecta Math **3**, 115–159 (1997)
13. Bourgain, J.: Refinements of Strichartz inequality and applications to 2D-NLS with critical nonlinearity. IMRN NS, 253–283 (1998)
14. Bourgain, J.: Global well-posedness of defocusing critical nonlinear Schrödinger equation in the radial case. J. Am. Math. Soc. **12**, 145–171 (1999)
15. Bourgain, J.: Global solutions of nonlinear Schrödinger equations. Am. Math. Soc. Colloq. Publ., vol. 46. American Mathematical Society, Providence, RI (1999)
16. Bourgain, J.: Problems in Hamiltonian PDE's, GAFA 2000 (Tel Aviv, 1999). Geom. Funct. Anal. 2000, Special Volume, Part I, pp. 32–56
17. Bourgain, J., Bulut, A.: Almost sure global well-posedness for the radial nonlinear Schrödinger equation on the unit ball I: the 2D case. Ann. Inst. H. Poincaré Anal. Non Linéaire **31**, 1267–1288 (2014)
18. Bourgain, J., Bulut, A.: Almost sure well-posedness for the radial nonlinear Schrödinger equation on the unit ball II: the 3D case. J. Eur. Math. Soc. (JEMS) **16**, 1289–1325 (2014)
19. Bourgain, J., Wang, W.: Construction of blow-up solutions for the nonlinear Schrödinger equation with critical nonlinearity. Dedicated to Ennio de Giorgi. Ann. Scuole Norm. Sup. Pisa Cl. Sc. **25**, 197–215 (1997)
20. Burq, N., Tzvetkov, N.: Random data Cauchy theory for supercritical wave equations II. A global existence result. Invent. Math. **173**, 477–496 (2008)
21. Carleson, L.: Some analytical problems related to statistical mechanics. LNM **779**, 5–45 (1979)
22. Cazenave, T., Weissler, F.: The Cauchy problem for the critical nonlinear Schrödinger equation in H^s. Nonlinear Anal. TMA **14**, 807–836 (1990)
23. Coifman, R., Meyer, Y.: Au delà des operateurs pséudo-differentiels. Astérisque, vol. 57. Société Mathématique de France, Paris (1978)
24. Coifman, R., Meyer, Y.: Nonlinear harmonic analysis, operator theory and PDE, Beijing lectures in harmonic analysis. Ann. Math. Stud., vol. 112. Princeton Univ. Press, Princeton, NJ (1986)
25. Colliander, J., Keel, M., Staffilani, G., Takaoka, H., Tao, T.: A refined global well-posedness result for Schrödinger equations with derivative. SIAM J. Math. Anal. **34**, 64–86 (2002)
26. Colliander, J., Keel, M., Staffilani, G., Takaoka, H., Tao, T.: Sharp global well-posedness for KdV and modified KdV on $\mathbb{R}$ and $\mathbb{T}$. J. Am. Math. Soc. **16**, 705–749 (2003)
27. Colliander, J., Keel, M., Staffilani, G., Takaoka, H., Tao, T.: Global existence and scattering for rough solutions of a nonlinear Schrödinger equation on $\mathbb{R}^3$. Comm. Pure Appl. Math. **57**, 987–1014 (2004)
28. Colliander, J., Keel, M., Staffilani, G., Takaoka, H., Tao, T.: Multilinear estimates for periodic KdV equations, and applications. J. Funct. Anal. **211**, 173–218 (2004)
29. Colliander, J., Keel, M., Staffilani, G., Takaoka, H., Tao, T.: Global well-posedness and scattering for the energy-critical nonlinear Schrödinger equation in $\mathbb{R}^3$. Ann. Math. **167**, 767–865 (2008)
30. Constantin, P., Saut, J.-C.: Local smoothing properties of dispersive equations. J. Am. Math. Soc. **1**, 413–439 (1998)
31. Dahlberg, B., Kenig, C.: A note on the almost everywhere behavior of solutions to the Schrödinger equation. LNM 908, 205–209 (1982)
32. Ginibre, J., Velo, G.: On a class of nonlinear Schrödinger equations, special theories in dimensions 1, 2 and 3. Ann. Inst. H. Poincaré **28**, 287–316 (1978)
33. Ginibre, J., Velo, G.: On the class of nonlinear Schrödinger equations. J. Funct. Anal. **32**, 1–71 (1979)
34. Ginibre, J., Velo, G.: The global Cauchy problem for the nonlinear Schrödinger equation revisited. Ann. Inst. H. Poincaré, Analyse NonLinëaire **2**, 309–327 (1985)

35. Glassey, R.T.: On the blowing up of solutions to the Cauchy problem for nonlinear Schrödinger equations. J. Math. Phys. **18**, 1794–1795 (1977)
36. Glimm, J., Jaffe, A.: Quantum Physics. Springer (1987)
37. Grillakis, M.: Regularity and asymptotic behavior of the wave equation with a critical nonlinearity. Ann. Math. **132**, 485–509 (1990)
38. Grillakis, M.: Regularity for the wave equation, with a critical nonlinearity. Comm. Pure Appl. Math. **45**, 749–774 (1992)
39. Grillakis, M.: On nonlinear Schrödinger equations. Comm. PDE **25**, 1827–1844 (2000)
40. Guo, Z.: Global well-posedness of Korteweg-de Vries equation in $H^{-\frac{3}{4}}$. J. Math. Pures Appl. **91**, 583–597 (2009)
41. Kappeler, T., Topalov, P.: Global well-posedness of mKdV in $L^2(\mathbb{T}, \mathbb{R})$. Comm. PDE **30**, 435–449 (2005)
42. Kappeler, T., Topalov, P.: Global well-posedness of KdV in $H^{-1}(\mathbb{T}, \mathbb{R})$. Duke Math. J. **135**, 327–360 (2006)
43. Kato, T.: On the Korteweg-de Vries equation. Manuscripta Math **29**, 89–99 (1979)
44. Kato, T.: On the Cauchy problem for the (generalized) Korteweg-de Vries equation. Adv. Math. Supp. Stud. Stud. Appl. Math. **8**, 93–126 (1983)
45. Kato, T.: Nonlinear Schrödinger equations. Ann. Inst. H. Poincaré, Phys-Théor. **46**, 113–129 (1987)
46. Kenig, C., Merle, F.: Global well-posedness, scattering and blow-up for the energy critical, focusing, non-linear Schrödinger equation in the radial case. Invent Math. **166**, 646–675 (2006)
47. Kenig, C., Merle, F.: Global well-posedness, scattering and blow-up for the energy-critical focusing non-linear wave equation. Acta Math. **201**, 147–212 (2008)
48. Kenig, C., Merle, F.: Scattering for $H^{\frac{1}{2}}$ bounded solutions to the cubic, defocusing (NLS) in 3 dimensions. Trans. Am. Math. Soc. **362**, 1937–1962 (2010)
49. Kenig, C., Ponce, G., Vega, L.: Well-posedness and scattering results for the generalized Korteweg-de Vries equation via the contraction principle. Comm. Pure Appl. Math. **46**, 527–620 (1993)
50. Kenig, C., Ponce, G., Vega, L.: A bilinear estimate with applications to the KdV equation. J. Am. Math. Soc. **9**, 573–603 (1996)
51. Kenig, C., Ponce, G., Vega, L.: On the ill-posedness of some canonical dispersive equations. Duke Math. J. **106**, 617–633 (2006)
52. Killip, R., Visan, M.: Global well-posedness and scattering for the defocusing quintic NLS in three dimensions. Anal. PDE **5**, 855–885 (2012)
53. Killip, R., Visan, M.: KdV is well-posed in H^{-1}. Ann. Math. **190**, 249–305 (2019)
54. Kishimoto, N.: Well-posedness of the Cauchy problem for the Korteweg-de Vries equation at the critical regularity. Differ. Integr. Equ. **22**, 447–464 (2009)
55. Klainerman, S., Machedon, M.: Space-time estimates for the null forms and the local existence theorem. Comm. Pure Appl. Math. **46**, 1221–1268 (1993)
56. Lebowitz, J., Rose, H., Speer, E.: Statistical mechanics of the nonlinear Schrödinger equation. J. Stat. Phys. **50**, 657–687 (1988)
57. Lin, J., Strauss, W.: Decay and scattering of solutions of a nonlinear Schrödinger equation. J. Funct. Anal. **30**, 245–263 (1978)
58. Merle, F.: Construction of solutions with exact k blow-up points for the Schrödinger equation with critical power nonlinearity. Comm. Math. Phys. **149**, 205–214 (1992)
59. Merle, F.: Determination of blow-up solutions with minimal mass for nonlinear Schrödinger equation with critical power. Duke Math. J. **69**, 427–453 (1993)
60. Merle, F., Raphaël, P.: Sharp upper bound on the blow-up rate for the critical nonlinear Schrödinger equation. Geom. Funct. Anal. **13**, 591–642 (2003)
61. Merle, F., Raphaël, P.: On universality of blow-up profile for L^2 critical nonlinear Schrödinger equation. Invent. Math. **156**, 565–672 (2004)
62. Merle, F., Raphaël, P.: Profiles and quantization of the blow-up mass for critical nonlinear Schrödinger equation. Comm. Math. Phys. **253**, 675–704 (2005)

63. Merle, F., Raphaël, P.: The blow-up dynamic and upper bound on the blow-up rate for critical nonlinear Schrödinger equation. Ann. Math. **161**, 157–222 (2005)
64. Merle, F., Raphaël, P.: On a sharp lower bound on the blow-up rate for the L^2 critical nonlinear Schrödinger equation. J. Am. Math. Soc. **19**, 37–90 (2006)
65. Merle, F., Raphaël, P., Rodnianski, I.: Type II blow-up for the energy supercritical NLS. Camb. J. Math. **3**, 439–617 (2015)
66. Merle, F., Raphaël, P., Rodnianski, I., Szeftel, J.: On smooth self similar solutions to the compressible Euler equations. arXiv:1912.10998
67. Merle, F., Raphaël, P., Rodnianski, I., Szeftel, J.: On blow-up for the energy super critical defocusing non linear Schrödinger equations. arXiv: 1912.11009
68. Merle, F., Raphaël, P., Rodnianski, I., Szeftel, J.: On the implosion of a three dimensional compressible fluid. arXiv:1912.11009
69. Molinet, L.: A note on ill-posedness for the KdV equation. Differ. Integr. Equ. **24**, 759–765 (2011)
70. Molinet, L.: Sharp ill-posedness and well-posedness for the KdV and mKdV equations on the torus. Adv. Math. **230**, 1895–1930 (2012)
71. Morawetz, C.S.: Notes on time decay and scattering for some hyperbolic problems. SIAM (1975)
72. Nashanishi, K., Takaoka, H., Tsutsumi, Y.: Counterexamples to bilinear estimates related with the KdV equation and the nonlinear Schrödinger equation. Methods Appl. Anal. **8**, 569–578 (2001)
73. Oh, T.: Invariant Gibbs measures and a.s. global well-posedness for coupled KdV systems. Differ. Integr. Equ. **22**, 637–668 (2009)
74. Raphaël, P.: Stability of the log log bound for blow-up solutions to the critical non-linear Schrödinger equation. Math. Ann. **331**, 577–609 (2005)
75. Raphaël, P.: Existence and stability of a solution blowing up on a sphere for the L^2 supercritical nonlinear Schrödinger equation. Duke Math. J. **134**, 199–258 (2006)
76. Rauch, J., Reed, M.: Nonlinear microlocal analysis of semilinear hyperbolic systems in one variable. Duke Math. J. **49**, 397–475 (1982)
77. Ryckman, E., Visan, M.: Global well-posedness and scattering for the defocusing energy-critical nonlinear Schrödinger equation in $\mathbb{R}^{1+4}$. Am. J. Math. **129**, 1–60 (2007)
78. Segal, I.E.: Space-time decay for solutions of wave equations. Adv. Math. **22**, 305–311 (1976)
79. Shatah, J., Struwe, M.: Regularity results for nonlinear wave equations. Ann. Math. **138**, 503–518 (1993)
80. Shatah, J., Struwe, M.: Well-posedness in the energy space for semilinear wave equations with critical growth. Int. Math. Res. Notices **1994**(7), 303–309 (1994)
81. Stein, E.M.: Harmonic Analysis: real variable methods, orthogonality and oscillatory integrals. Princeton Mathematical Series, vol. 43. Princeton University Press, Princeton, NJ (1993)
82. Sjölin, P.: Regularity of solutions to the Schrödinger equation. Duke Math. J. **55**, 699–715 (1987)
83. Strichartz, R.S.: Restriction of Fourier Transforms to quadratic surfaces and decay of solutions to wave equations. Duke Math. J. **44**, 705–714 (1977)
84. Struwe, M.: Globally regular solutions to the u^5 Klein-Gordon equation. Ann. Scuola Norm. Sup. Pisa Cl. Sci. **15**, 495–513 (1989)
85. Tsutsumi, Y.: L^2 solutions for nonlinear Schrödinger equations and nonlinear groups. Funkcial. Ekvac. **30**, 115–125 (1987)
86. Tomas, P.: A restriction theorem for the Fourier transform. Bull. AMS **81**, 477–478 (1979)
87. Vega, L.: Schrödinger equations: pointwise convergent to the initial data. Proc. AMS **102**, 874–878 (1988)
88. Vinogradov, I.M.: The Method of Trigonometric Sums in the Theory of Numbers. Intersciences, NY (1954)
89. Visan, M.: The defocusing energy-critical nonlinear Schrödinger equation in higher dimensions. Duke Math. J. **138**, 281–374 (2007)

On Trace Sets of Restricted Continued Fraction Semigroups

Alex Kontorovich

Dedicated to the memory of Jean Bourgain

Abstract We record an argument due to Jean Bourgain which gives lower bounds on the size of the trace sets of certain semigroups related to continued fractions on finite alphabets. These bounds are motivated by the "Classical Arithmetic Chaos" Conjecture of McMullen (Dynamics of units and packing constants of ideals, 2012). Specifically, a power is gained in the asymptotic size of the trace set over a "trivial" exponent. The proof involves a new application of the Balog-Szemerédi-Gowers Lemma from additive combinatorics.

1 Introduction

We begin with some personal remarks and reminiscences on the occasion of this Dedicated Volume; we allow ourselves to be descriptive, returning to precision and science in Sect 1.1. My collaboration with Jean Bourgain began in the fall of 2008, when I applied for the 2009-2010 IAS Special Year in Analytic Number Theory. To explain properly what we were trying to accomplish, I have to back up to my 2007 thesis. There I was interested in a kind of mixture between the theorems of Friedlander-Iwaniec [14] and Piatetskii-Shapiro [35], the former being that the polynomial

The author is partially supported by an NSF grant DMS-1802119, and the Simons Foundation through MoMath's Distinguished Visiting Professorship for the Public Dissemination of Mathematics.

A. Kontorovich (✉)
Department of Mathematics, Rutgers University, Piscataway, NJ, USA

National Museum of Mathematics, New York, NY, USA
e-mail: alex.kontorovich@rutgers.edu

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_11

$$FI(x, y) := x^2 + (y^2)^2 \tag{1}$$

represents infinitely many primes and the latter that the sequence

$$PS(n) := \lfloor n^\alpha \rfloor$$

does too, for sufficiently small values of the fixed constant $\alpha > 1$. Both sequences are "thin": the number of integers up to X represented by FI is about $X^{3/4}$, whereas for PS, it is about $X^{1/\alpha}$, so it is rather difficult to produce primes in such sparse sequences[1] (the latter being still much easier than the former![2]). The nice thing about PS is that there is a parameter, α, to play with and thus a potential range of thinness where one can succeed. The main idea of my thesis (suggested to me by Peter Sarnak, motivated by his work with Jean and Alex Gamburd on the Affine Sieve [11]), was to see whether an amalgam of the two was possible in the group setting; by this we mean the following:

Let $\Gamma < SL_2(\mathbb{Z})$ be some Zariski-dense subgroup of the modular group; if it is of infinite index (or even just non-congruence!), then we have no idea exactly which pairs (c, d) arise as bottom rows, say, of elements in the group $\left(\begin{smallmatrix} * & * \\ c & d \end{smallmatrix}\right) \overset{?}{\in} \Gamma$. We would, in principle, first need to try to write any such matrices as words in the generators of Γ. Regardless, consider the sequence

$$\mathcal{S} := \{c^2 + d^2 : \left(\begin{smallmatrix} * & * \\ c & d \end{smallmatrix}\right) \in \Gamma\}. \tag{2}$$

The total number of such values $c^2+d^2 < X$ *can* be counted effectively, that is, with power savings, as was done in my thesis [24] (under a technical assumption that was removed in [30]). The answer is roughly X^δ, where δ, assumed to exceed one-half, is the critical exponent of Γ (equivalently [34], the Hausdorff dimension of the limit set of Γ; the condition $\delta > 1/2$ is needed to relate δ to the base eigenvalue $\lambda_0 = \delta(1-\delta)$ of the hyperbolic Laplacian acting on square-integrable functions on the upper half plane $\mathbb{H}$ invariant under Γ). Since one can exhibit Γ with δ arbitrarily close to 1, one can play with this "thinness" parameter, similarly to Piatetskii-Shapiro, where $1/\alpha < 1$ plays the role of δ. If instead we returned to all integer pairs (c, d) but forced $d = y^2$ to be a perfect square, then we would exactly be in the situation of Friedlander-Iwaniec (1). So this set S has both features, studying $c^2 + d^2$ for restricted (by the group) values of (c, d), with the flexibility of a parameter δ. Since this phenomenon of δ being thin but not "too" thin will appear again and again, let me refer to it as being **slightly thin**, that is, allowing $\delta < 1$ but also requiring that $\delta > 1 - \varepsilon_0$ for some (usually small) $\varepsilon_0 > 0$.

[1] Heath-Brown [19] was later able to do the same for the even thinner polynomial $x^3 + 2y^3$, which takes about $X^{2/3}$ values up to X.

[2] See also [25] for a simpler instance of this "parity breaking."

The problem of producing primes in $\mathcal{S}$ for *any* value of $\delta < 1$ is still wide open. The tools in my thesis managed to produce R-almost primes (i.e., numbers with at most R prime factors) for $R = 13$ in slightly thin groups,[3] and in my application to IAS, I had proposed not just to use the linear sieve but to introduce bilinear form techniques into the affine sieve to attack this problem, with the hope of producing actual primes. Admittedly, this is perhaps a rather niche question, but one I enjoyed thinking about for its mixture of geometric, combinatorial, spectral, dynamical, algebraic, and number theoretic techniques.

Jean must have read my application, because the next time I visited Peter at IAS, Jean requested to speak with me. At our meeting, he outlined how to execute such bilinear form ideas to produce primes, not in $\mathcal{S}$ but in certain algebraic traces of entries of slightly thin subgroups of the Picard group $\mathrm{SL}_2(\mathbb{Z}[i])$, the added dimension allowing for more variables.[4] Together, we whittled away at the problem until we could produce, for slightly thin subgroups of the modular group $\mathrm{SL}_2(\mathbb{Z})$, primes in the values of the *linear* map $f : \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \mapsto d$ (note that, for $\mathcal{S}$ in (2), we would instead apply a quadratic map $f : \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \mapsto c^2 + d^2$). At some point during a conversation with Jean and Peter, we realized that in fact we already had almost all the tools needed to prove something much stronger: the reason we were able to produce prime values of f is because we were actually producing *almost all* numbers! (This is in contradistinction to $f = c^2 + d^2$ which is genuinely a thin subset of $\mathbb{Z}$.[5])

Thus, producing an almost-all statement in the $f = d$ values of slightly thin subgroups of $\mathrm{SL}_2(\mathbb{Z})$ became our first joint paper [4]. It required an additional stubborn technical ingredient (of independent interest) to count effectively in bisectors in thin groups, which we proved in a companion paper jointly also with Peter [12]; such ideas have since been generalized many times by many authors. We had also noticed some similarities between this problem and the local-global problem for Apollonian packings (see [4, Remark 1.12]), but there did not seem to be an obvious way to transfer our technology, given that the Apollonian group was **not** slightly thin, but had a fixed dimension, $\delta \approx 1.30$; there was no parameter to adjust!

At this point, I thought our collaboration was basically done and I could return to civilian life. But as luck would have it, Curt McMullen pointed out to us the similarity between the problem we had just attacked and Zaremba's conjecture [40] on bounded continued fractions of rationals. (See, e.g., [26] for a detailed discussion of this problem.) We *should* have already been aware of the connection, since years before, Jeff Lagarias had pointed out to me the similarity between Zaremba and Apollonius (see [16, p. 37]), but somehow it took Curt's urging for us to begin working on it. The Zaremba problem was nearly identical, except that we were missing a number of technical ingredients, including the

[3] It turns out that I should have been able to product R-almost primes with $R = 7$, see [21].

[4] Much later, I would exploit a similar feature in [29].

[5] For a formal definition of thinness in a general context, see [27, p. 954].

main consequences of [12]. The reason is that, in Zaremba, one must deal with a sub-**semi-**group of $SL_2(\mathbb{Z})$, not a sub*group*, thus rendering our spectral and representation theoretic counting developments useless. Nevertheless, one could substitute the thermodynamic formalism to count [13, 20], and we were able to show density one for Zaremba [5]. Here the analog of being "slightly thin" is having a sufficiently large allowed alphabet for the restricted partial quotients.

Again I thought that would basically be the end of things, but about a year later, we realized that using some rather different techniques (relying not on "slight thinness," but instead exploiting the existence of values of shifted binary quadratic forms inside the bend set, as observed by Sarnak [36], and taking inspiration from Jean's paper [3] on prime values in Apollonian packings), we could actually extend the local-global technology to prove density one in the Apollonian problem; see [6].

It quickly became clear that the bilinear form technology developed on our "Orbital Circle Method" phase could be applied much more widely, and we turned our attention back to the original sequence (2) from my thesis. There we were able to implement these ideas, along with some others (e.g., the "dispersion method" in the group context), to push past the sieve level of distribution (see [27]) which follows "for free" from counting arguments and "expansion" (i.e., certain families of Cayley graphs being expanders), to a level "Beyond Expansion." In the end, we could produce, for any slightly thin group, R-almost primes in $\mathcal{S}$, with $R = 4$ [7]. This became Part I in our Beyond Expansion program.

For Part II [8], we turned our attention to a problem of Einsiedler-Lindenstrauss-Michel-Venkatesh, which itself actually served as the original motivation for the Affine Sieve (see [37]). This problem, involving the same semigroup as in our Zaremba work, required "only" a square-free sieve, but as it turned out, expansion alone was just barely insufficient to solve the problem. An added difficulty was that, unlike Zaremba where the linear function on the semigroup was $f : \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \mapsto d$, here one needed to deal with the trace, $f : \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \mapsto a + d$. The nice thing about $f = d$ is that it is *already* itself bilinear, being expressible as $f(\gamma) = \langle e_2, \gamma e_2 \rangle$ (with $e_2 = (0, 1)^t$), but trace is not. Nevertheless, with some new ideas, we were able to solve the problem, producing an infinitude of "low-lying" but "fundamental" closed geodesics on the modular surface. In Part III of the series [10], we added the adjective "reciprocal" to the closed geodesics, inching such a tad closer to the Markoff geodesics they are meant to imitate.

And the final Part IV of the series, which is related to the theorem we wish to explain in this note, was motivated by the "Classical Arithmetic Chaos Conjecture" posed by Curt McMullen; see Sect 1.1. This problem was basically too difficult for us to say very much about at all, except that we could improve the "expansion" exponent of distribution in the trace set all the way to what it would have been, had some analog of the Ramanujan conjecture (on average) existed in this setting; see [9].

Thus, it ended my collaboration with Jean Bourgain. Echoing what others have said, with his passing, the mathematical world has lost an Archimedes, an Euler, and a Gauss. It was an incredible privilege and honor to work with Jean, and I am forever grateful.

The last few of our papers were being finalized as Jean was undergoing various surgeries and chemotherapies, and one only appeared posthumously (one of ours, that is, I'm sure Jean will continue co-authoring posthumously for a few more years). In our conversations over the last few years, he never once showed any signs of fear or despair at his condition, treating "mundane" things very matter of factly and wanting to steer discussions back to theorems and (scientific) battles still to be waged.

Of Jean's many hand-written and scanned notes to me (as in the example below), all have been converted to publications save one, which is the one we aim to record now. To be perfectly honest, Jean thought it should be possible to do more here and wanted to return to the problem later, not publish things as they stand. But now there is no "later," so I would like to record his theorem as is. At some point, Michael Magee and I worked on this note as an appendix to our paper with Jean (which itself later became an appendix); I would like to thank Michael for his work on it and his permission to reuse some of it here. On to the science.

1.1 McMullen's Arithmetic Chaos Conjecture

For $x \in \mathbb{R}$, we write its continued fraction expansion as

$$x = [a_0; a_1, \dots, a_\ell, \dots] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots \cfrac{1}{a_\ell + \ddots}}},$$

which may be either finite or infinite; here $a_0 \in \mathbb{Z}$, and for $j \geq 1$, the "partial quotients" a_j are positive integers. The bar in

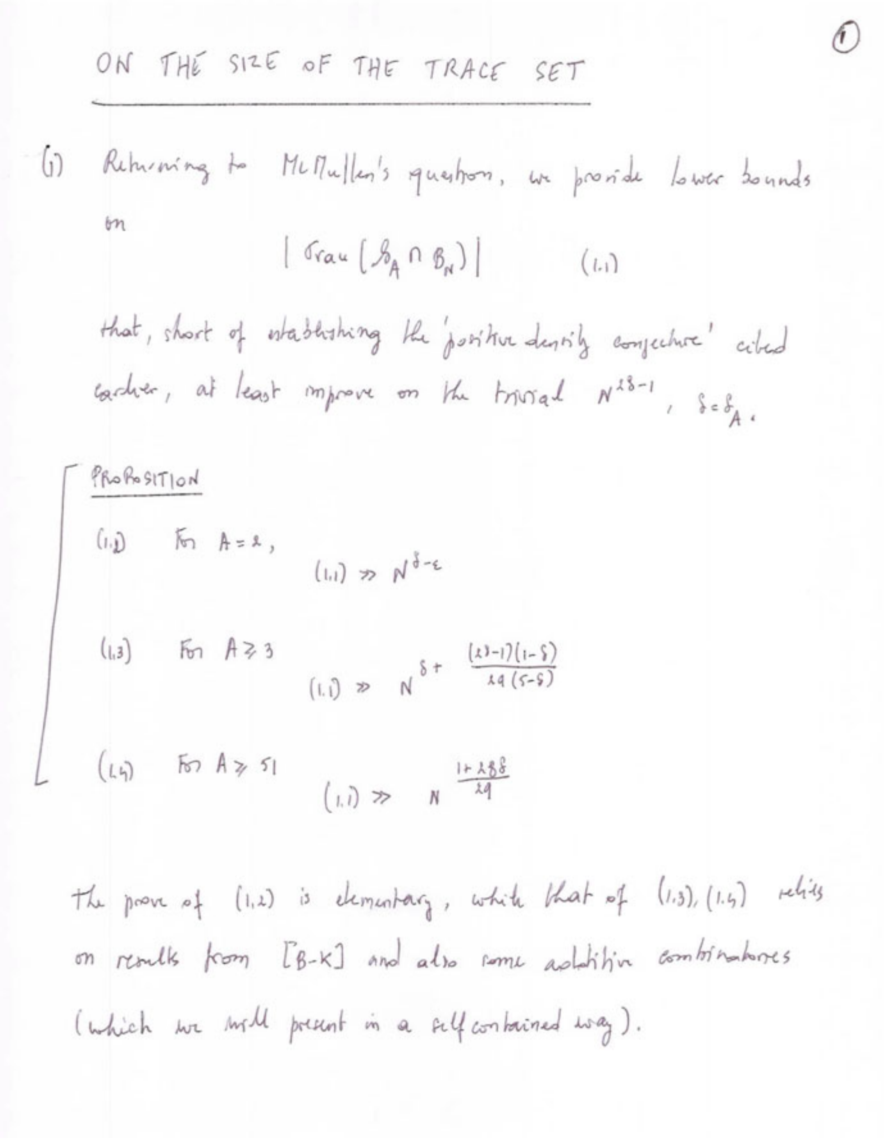
$$[\overline{a_0, a_1, \dots, a_\ell}]$$

denotes periodically repeating partial quotients; it is very well-known that such numbers are quadratic surds. For a finite "alphabet" $\mathcal{A} \subset \mathbb{N}$, let $\mathfrak{C}_{\mathcal{A}}$ denote the Cantor-like set of numbers in the unit interval whose partial quotients lie in $\mathcal{A}$,

$$\mathfrak{C}_{\mathcal{A}} := \{[0; a_1, \dots, a_\ell, \dots] \, : \, a_1, a_2, \dots \in \mathcal{A}\},$$

and let $\delta_{\mathcal{A}}$ be its Hausdorff dimension,

$$\delta_{\mathcal{A}} := H.dim(\mathfrak{C}_{\mathcal{A}}) \in [0, 1).$$



1

ON THE SIZE OF THE TRACE SET

(i) Returning to McMullen's question, we provide lower bounds on

$$|\mathrm{Trace}(\mathcal{S}_A \cap B_N)| \qquad (1.1)$$

that, short of establishing the 'positive density conjecture' cited earlier, at least improve on the trivial $N^{2\delta-1}$, $\delta=\delta_A$.

PROPOSITION

(1.2) For $A=2$,

$$(1.1) \gg N^{\delta-\varepsilon}$$

(1.3) For $A \geqslant 3$

$$(1.1) \gg N^{\delta+\frac{(2\delta-1)(1-\delta)}{29(5-\delta)}}$$

(1.4) For $A \geqslant 51$

$$(1.1) \gg N^{\frac{1+28\delta}{29}}$$

The prove of (1,2) is elementary, while that of (1,3), (1.4) relies on results from [B-K] and also some additive combinatorics (which we will present in a self contained way).

Motivated by conjectures on the rigidity of higher-rank diagonal flows, McMullen [31, 32] formulated the following rank-one problem:

Conjecture 1.1 (McMullen's Classical Arithmetic Chaos Conjecture) *Let $\mathcal{A}$ be any alphabet with dimension $\delta_{\mathcal{A}}$ exceeding $1/2$. Then for any real quadratic field K, the set*

$$\{[\overline{a_0, a_1, \dots, a_\ell}] \in K \ : \ all\ a_j \in \mathcal{A}\} \tag{3}$$

grows exponentially as the length $\ell \to \infty$.

Exponential growth is not known for a single choice of $\mathcal{A}$ and K. Worse yet, it is unknown unconditionally whether there is an alphabet $\mathcal{A}$ such that every K has at least *one* surd with all partial quotients in $\mathcal{A}$, that is, whether the union over all ℓ of (3) is non-empty! On the other hand, Mercat [33] has proven this last statement assuming the validity of Zaremba's conjecture [5, 40]. Unconditionally, Wilson [38] has shown that for any K, there is some $\mathcal{A} = \mathcal{A}(K)$ so that (3) is non-empty infinitely often; see also [39].

1.2 Thin Semigroups

To connect this problem to "thin semigroups," let $\mathcal{G}_\mathcal{A} \subset \mathrm{GL}_2(\mathbb{Z})$ denote the semigroup generated by matrices of the form $\left(\begin{smallmatrix} 0 & 1 \\ 1 & a \end{smallmatrix}\right)$ with $a \in \mathcal{A}$,

$$\mathcal{G}_\mathcal{A} := \left\langle \begin{pmatrix} 0 & 1 \\ 1 & a \end{pmatrix} \ : \ a \in \mathcal{A} \right\rangle^+,$$

where the superscript "+" indicates generation without inverses. (See [28, Lecture 3] for why $\mathcal{G}_\mathcal{A}$ is "thin".) This matrix semigroup was introduced in [5] to study Zaremba's conjecture, but is equally germane to McMullen's problem, due to the following elementary observation: if

$$\gamma = \begin{pmatrix} 0 & 1 \\ 1 & a_0 \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & a_\ell \end{pmatrix} \in \mathcal{G}_\mathcal{A},$$

then

$$\mathbb{Q}([\overline{a_0, \dots, a_\ell}]) = K,$$

where

$$K = \mathbb{Q}(\sqrt{\mathrm{tr}^2\,\gamma - 4\det\gamma}).$$

(Recall that $\det\gamma = \pm 1$.) That is, one can read off the discriminant of the real quadratic field corresponding to adjoining $[\overline{a_0, \dots, a_\ell}]$ in terms of the trace of γ.

1.3 The Local-Global and Positive Density Conjectures

The above simple observation motivates one to study the set $\mathcal{T}_{\mathcal{A}}$ of traces of $\mathcal{G}_{\mathcal{A}}$,

$$\mathcal{T}_{\mathcal{A}} := \{\operatorname{tr}\gamma : \gamma \in \mathcal{G}_{\mathcal{A}}\}.$$

Indeed, Bourgain and the author have formulated a certain "Local-Global Conjecture" for linear forms on $\mathcal{G}_{\mathcal{A}}$ (see [28, Conjecture 6.3.1]) which implies both Zaremba's conjecture and McMullen's Conjecture 1.1, in particular, predicting which traces should arise and with what multiplicity. A weaker problem, formulated already by McMullen [32], is the following:

Conjecture 1.2 (McMullen's Positive Density Conjecture for Traces) *Let $\mathcal{A}$ be an alphabet with $\delta_{\mathcal{A}} > 1/2$. Then the trace set $\mathcal{T}_{\mathcal{A}}$ comprises a positive proportion of integers, that is,*

$$\#\mathcal{T}_{\mathcal{A}} \cap [1, N] \gg N, \tag{4}$$

as $N \to \infty$.

The restriction to alphabets having $\delta_{\mathcal{A}}$ exceed $1/2$ is necessary, in light of the following result of Hensley [20].

Theorem 1.3 (Hensley) *As $N \to \infty$,*

$$|\mathcal{G}_{\mathcal{A}} \cap B_N| \asymp N^{2\delta_{\mathcal{A}}}. \tag{5}$$

Indeed, if $\delta_{\mathcal{A}} < 1/2$, then $\mathcal{T}_{\mathcal{A}}$ is automatically a thin subset of the integers.

This positive density Conjecture 1.2, despite being much weaker than a full local-global statement, is also wide open, even for any choice of (finite) alphabet $\mathcal{A}$. If, instead of traces, one considers the set $\mathfrak{D}_{\mathcal{A}}$ of "bottom-right" entries,

$$\mathfrak{D}_{\mathcal{A}} := \{d \in \mathbb{N} : \exists \begin{pmatrix} * & * \\ * & d \end{pmatrix} \in \mathcal{G}_{\mathcal{A}}\},$$

then one can show not just positive density but density one for "slightly thin" alphabets (ones with $\delta_{\mathcal{A}} > 1 - \varepsilon_0$); see [5]. Zaremba's conjecture is equivalent to a local-global statement for $\mathfrak{D}_{\mathcal{A}}$. The proof technique there shows the following:

Theorem 1.4 ([5]) *Let $\mathcal{A}$ be an alphabet with $\delta = \delta_{\mathcal{A}}$ sufficiently near 1, $\delta > 1 - \varepsilon_0$. Then, there exist subsets $S_N \subset \mathcal{G}_{\mathcal{A}} \cap B_N$ of nearly full cardinality,*

$$\#S_N \gg N^{2\delta}$$

such that, for every $d \ll N$, the multiplicity of the map $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \mapsto d$ is bounded by

$$\#\{\gamma \in S_N, \gamma = \left(\begin{smallmatrix} * & * \\ * & d \end{smallmatrix}\right)\} \ll N^{2\delta-1}. \tag{6}$$

This estimate will be the only "black box" used; besides this, the paper is self-contained.

1.4 Statements of the Main Theorems

Returning to the trace set $\mathcal{T}_{\mathcal{A}}$, the "trivial" bound towards (4) is

$$\#\mathcal{T}_{\mathcal{A}} \cap [1, N] \gg N^{2\delta_{\mathcal{A}}-1-o(1)}. \tag{7}$$

Indeed, a simple argument shows that each trace $t < N$ occurs with multiplicity $\ll N^{1+o(1)}$, whence (7) follows from (5).

Our goal here is to give Jean Bourgain's proofs of the following two results, which improve over this.

Theorem 1.5 *When $\delta_{\mathcal{A}} > 1/2$,*

$$\#\mathcal{T}_{\mathcal{A}} \cap [1, N] \gg N^{\delta-o(1)}. \tag{8}$$

Theorem 1.6 *Suppose $\{1, 2, 3\} \subset \mathcal{A}$ and $\delta > 1 - \varepsilon_0$ so that (6) holds. Then as $N \to \infty$,*

$$\#\mathcal{T}_{\mathcal{A}} \cap [1, N] \gg N^{\delta+\frac{1-\delta}{29}-o(1)}. \tag{9}$$

Remark 1.7 The proof of Theorem 1.5 is elementary, and yet it already improves upon (7), sometimes dramatically so. Indeed, when $\mathcal{A} = \{1, 2\}$, we have $\delta_{\{1,2\}} \approx 0.531$ [17]; the trivial bound (7) gives only

$$\#\mathcal{T}_{\{1,2\}} \cap [1, N] \gg N^{0.062},$$

while (8) gives

$$\#\mathcal{T}_{\{1,2\}} \cap [1, N] \gg N^{0.531}.$$

Remark 1.8 The original work [5] showed (6) as long as $\delta > 0.984$, and this bound was relaxed in [15] and [23] to $\delta > 0.781$; the latter holds already for $\mathcal{A} = \{1, 2, 3, 4\}$ which has dimension $\delta_{\{1,2,3,4\}} \approx 0.789$; see [22]. Thus, for this alphabet, Theorem 1.6 improves from (8) that

$$\#\mathcal{T}_{\{1,2,3,4\}} \cap [1, N] \gg N^{0.789},$$

to

$$\#\mathcal{T}_{\{1,2,3,4\}} \cap [1, N] \gg N^{0.796}.$$

The proof of Theorem 1.6 applies more generally to give an improvement in the exponent whenever $\delta > 1/2$ and $\mathcal{A}$ contains a three-term progression; but for ease of exposition, we state this simpler version.

The core of Theorem 1.6 is the following version of the Balog-Szemerédi-Gowers Lemma with polynomial dependencies of constants on one another. The original version of the Balog-Szemerédi-Gowers Lemma with polynomial dependencies of constants appeared in Gowers' work on arithmetic progressions [18].

For subsets A, B of an ambient additive group and $G \subset A \times B$ an arbitrary subset, we use the notation

$$A \overset{G}{+} B := \{a + b : (a, b) \in G\}.$$

Lemma 1.9 *Let $A \subset \mathbb{Z}$ be a finite set and $G \subset A \times A$ satisfy*

$$|G| > \frac{1}{K}|A|^2$$

and

$$|A \overset{G}{+} A| \le |A|. \tag{10}$$

Then, there is a subset $A' \subset A$ such that

$$|(A' \times A') \cap G| \gg K^{-2}|A|^2$$

and

$$|A' - A'| \ll K^{13}|A|,$$

where the implied constants are absolute.

This version is a refinement of Bourgain's work [1, Lemma 2.1] on the dimension of Kakeya sets. To make the argument almost self-contained (modulo Theorem 1.4), we give a quick proof of Lemma 1.9 in Sect. 5.

1.5 Notation

Whenever we write B_N, we mean the ball in the space of 2×2 matrices with respect to the ℓ^1 norm on their entries, and when we write $\|g\|$, we mean the ℓ^1 norm. We write sqf for the squarefree part of a number and ω for the number of distinct prime factors of a number. We use Vinogradov notation $\ll$,$\gg$, O,o in the standard way and indicate dependence of implied constants on other parameters by subscripts, e.g., $\ll_\epsilon$. We use $f \asymp g$ to mean $f \ll g$ and $g \ll f$. Normally, we view $\mathcal{A}$ as fixed so any implied constant may depend on $\mathcal{A}$. For a subset A of an ambient additive group, we write $A + A$ and $A - A$ for setwise sums and differences, e.g., $A - A = \{a_1 - a_2 \ : \ a_1, a_2 \in A\}$, etc.

2 Preliminary Remarks

Write

$$\gamma_a := \left(\begin{smallmatrix} 0 & 1 \\ 1 & a \end{smallmatrix}\right),$$

for a generator of the semigroup $\mathcal{G}_\mathcal{A}$. A ping-pong argument using the action of γ_a on [0, 1] by Möbius transformations shows that $\mathcal{G}_\mathcal{A}$ is freely generated by the γ_a, for $a \in \mathcal{A}$. Let

$$\Gamma_\mathcal{A} := \mathcal{G}_\mathcal{A} \cap \mathrm{SL}_2$$

be the sub-semigroup of orientation-preserving elements; equivalently, these are even words in the generators (each of the latter has determinant $\det \gamma_a = -1$).

The key (trivial) observation used throughout is the following:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & \alpha \end{pmatrix} = \begin{pmatrix} b & a + \alpha b \\ d & c + \alpha d \end{pmatrix}, \tag{11}$$

whence $b+c+\alpha d$ is a trace in $\mathcal{G}_\mathcal{A}$. Taking $\alpha = 1, 2, 3$ (if we assume that $\{1, 2, 3\} \subset \mathcal{A}$), we see that all three of

$$b + c + d,\ b + c + 2d,\ b + c + 3d \in \mathcal{T}_\mathcal{A}, \tag{12}$$

whenever $\left(\begin{smallmatrix} * & b \\ c & d \end{smallmatrix}\right) \in \mathcal{G}_\mathcal{A}$.

3 Proof of Theorem 1.5

For simplicity, assume that $\{1, 2\} \subset \mathcal{A}$; in general, we know by $\delta > 1/2$ that $|\mathcal{A}| \geq 2$, and trivial modifications are needed in what follows. In light of (12), we would like to know the multiplicity of the map

$$\varphi : \Gamma_{\mathcal{A}} \to \mathbb{N}^2 : \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \mapsto (b + c, d).$$

Lemma 3.1 *Let $(n, m) \in \mathbb{N}^2$ with $n, m \ll N$. Then, the preimage of (n, m) has cardinality at most*

$$|\varphi^{-1}(n, m)| \leq \gcd(n^2 + 4, m)^{1/2} N^{o(1)}.$$

Proof Suppose that $(b + c, d) = (n, m)$. Clearly, $d = m$ is determined. Since $ad - bc = 1$ and $c = n - b$, we have

$$1 + b(n - b) \equiv 0 (\operatorname{mod} d).$$

The discriminant of this quadratic in b is $\Delta := n^2 + 4$, and it is elementary that the number of solutions to $b (\operatorname{mod} d)$ is $\ll_{\varepsilon} (\Delta, d)^{1/2} N^{\varepsilon}$. Since $b \leq d$, it is determined once it is known mod d; then so is $c = n - b$, and then $a = (1 + bc)/d$. □

The issue becomes to discard $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_{\mathcal{A}}$ having large $\gcd(\Delta, d)$.

Lemma 3.2 *For any $\epsilon > 0$, there is a subset $B'_N \subset \Gamma_{\mathcal{A}} \cap B_N$ satisfying*

$$|B'_N| > \frac{1}{2} |\Gamma_{\mathcal{A}} \cap B(N)|, \tag{13}$$

and if $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in B'_N$, then $\gcd((b + c)^2 + 4, d) \ll_{\epsilon} N^{\epsilon}$.

Lemma 3.2 follows immediately from the following: Lemma that we will also use later.

Lemma 3.3 *Suppose that $\delta > \frac{1}{2}$. For all $\epsilon > 0$, there is $\eta = \eta(\epsilon) > 0$ such that*

$$\left| \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_{\mathcal{A}} \cap B_N \ : \ \gcd((b + c)^2 + 4, d) > N^{\epsilon} \right\} \right| \ll_{\epsilon} N^{2\delta - \eta}.$$

In particular, in comparison to Theorem 1.3, these elements form a negligible subset.

Proof Given $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_{\mathcal{A}} \cap B_N$, suppose that there is "large" $q > N^{\epsilon}$ dividing both $(b + c)^2 + 4$ and d. Then, $bc + 1 \equiv 0 \bmod d$ implies

$$(b - c)^2 \equiv b^2 + c^2 + 2 \equiv (b + c)^2 + 4 \equiv 0 \bmod q.$$

Therefore, $b \equiv c \bmod q_1$ for some $q_1 | q$ with $q_1 > N^{\epsilon/2}$. Then,

$$d \equiv b^2 + 1 \equiv c^2 + 1 \equiv 0 \bmod q_1. \tag{14}$$

We write each $g \in \Gamma_{\mathcal{A}}$ with $\|g\| \asymp N$ in the form

$$g = g_1 g_2$$

with

$$\|g_2\| \asymp N' := N^{\frac{\epsilon}{10}}.$$

(Note that for $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathcal{G}_{\mathcal{A}}$, the entries a, b, c, d are all commensurate and that word length in the generators is log-commensurate to the Archimedean norm.) Accordingly, we write

$$g_1 = \begin{pmatrix} \alpha & \beta \\ \gamma & \zeta \end{pmatrix}, \quad g_2 = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$$

so that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \alpha x + \beta z & \alpha y + \beta w \\ \gamma x + \zeta z & \gamma y + \zeta w \end{pmatrix}. \tag{15}$$

For each choice of g_1 and y, we will show there are few possibilities for w if (14) is to hold for some $q_1 > N^{\frac{\epsilon}{2}}$. On the other hand, y and w determine g_2, and hence g_1, y, w determine g. Combining (14) and (15), we get

$$\gamma y + \zeta w \equiv 0 \bmod q_1 \tag{16}$$

and

$$(\gamma x + \zeta z)^2 \equiv -1 \bmod q_1.$$

Thus, using $\det g_2 = 1$ gives

$$-y^2 \equiv y^2(\gamma x + \zeta z)^2 \equiv (\gamma y x + \zeta y z)^2 \equiv (\gamma y x + \zeta(xw - 1))^2 \equiv \zeta^2 \bmod q_1$$

where the last equality uses (16). In other words,

$$y^2 + \zeta^2 \equiv 0 \bmod q_1.$$

For fixed g_1, y the number of w so that (14) holds for some q_1 is bounded by

$$\sum_{q_1 | y^2+\zeta^2} |\{w \ : \ \gamma y + \zeta w = 0 \bmod q_1\}|. \tag{17}$$

For each q_1 in the sum, let $q_2 = \gcd(q_1, \zeta^2)$. We have $y^2 \equiv 0 \bmod q_2$, and since $0 < y < N'$, this implies $q_2 < N'$. This means $\gcd(q_1, \zeta) < N'$, and then $\gamma y + \zeta w = 0 \bmod q_1$ specifies $w \bmod q_3 := q_1 / \gcd(q_1, \zeta)$, with $q_3 > N^{\frac{\epsilon}{2}} N^{-\frac{\epsilon}{10}}$. But since $0 < w \ll N^{\frac{\epsilon}{10}}$, this specifies w.

Then, each term in (17) is bounded by 1, and we can bound (17) by the number of divisors of $y^2 + \zeta^2 \le N^2$, which is $N^{o(1)}$. It remains to sum over g_1 and y, and this gives that the number of $g \in \Gamma_{\mathcal{A}} \cap B_N$ such that (14) holds for some $q_1 > N^{\frac{\epsilon}{2}}$ is

$$\begin{aligned} &\le |\{g_1 : \|g_1\| \ll N^{1-\frac{\epsilon}{10}}\}| \cdot |\{y \ll N^{\frac{\epsilon}{10}}\}| \cdot N^{o(1)} \\ &\ll N^{2\delta(1-\frac{\epsilon}{10})} N^{\frac{\epsilon}{10}} N^{o(1)} \ll N^{2\delta - \eta} \end{aligned}$$

for some $\eta = \eta(\epsilon) > 0$. □

Proof of Theorem 1.5 For small $\epsilon > 0$, let B'_N be the family of subsets from Lemma 3.2. Combining Lemmas 3.1 and 3.2, the map

$$B'_N \to \mathcal{T}_{\mathcal{A}}(3N) \times \mathcal{T}_{\mathcal{A}}(4N) : \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto (b + c + d, b + c + 2d)$$

has multiplicity at most $N^{o(1)}$ and so

$$|\mathcal{T}_{\mathcal{A}}(N)|^2 > \frac{1}{2} |\Gamma_{\mathcal{A}} \cap B_N| N^{-o(1)} \gg N^{2\delta - o(1)}.$$

Taking square roots completes the proof. □

4 Proof of Theorem 1.6

Assume now that $\{1, 2, 3\} \subset \mathcal{A}$, so we can exploit the full force of (12). By Theorem 1.4, there is a family of subsets

$$S(N) \subset \Gamma_{\mathcal{A}} \cap B_N$$

with $|S(N)| \gg N^{2\delta}$ and such that the map $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \mapsto d$ has multiplicity $M \ll N^{2\delta - 1}$. Then, using Lemma 3.3 and echoing the previous argument, we can find a subset $S'(N) \subset S(N)$ such that the map

$$\psi : S'(N) \to \mathcal{T}_{\mathcal{A}}(3N) \times \mathcal{T}_{\mathcal{A}}(5N), \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto (b+c+d, b+c+3d)$$

has multiplicity $< N^{o(1)}$. Let

$$T_0 = \{b+c+jd \ : \ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in S'(N), \ 1 \le j \le 3\} \subset \mathcal{T}_{\mathcal{A}}(5N).$$

We apply Lemma 1.9 with $A = T_0$ and

$$G = \{(b+c+d, b+c+3d) \ : \ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in S'(N)\} = \psi(S'(N)).$$

By our previous bound on the multiplicity of ψ,

$$|G| > N^{2\delta - o(1)}.$$

Also,

$$T_0 \overset{G}{+} T_0 = \{2(b+c+d) \ : \ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in S'(N)\}$$

so

$$|T_0 \overset{G}{+} T_0| \le |T_0|.$$

We can thus apply Lemma 1.9 with

$$K = |T_0|^2 N^{-2\delta + o(1)}$$

Let A' be the subset obtained from Lemma 1.9.

The key point is that for each element of $(t_1, t_2) \in (A' \times A') \cap G$, one has $t_2 - t_1 \in 2\mathcal{D}_{\mathcal{A}}$. Moreover, if $(t_1, t_2) = \psi\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, then $t_2 - t_1 = 2d$. Since the multiplicity of the denominator mapping on $S'(N)$ is at most M, the multiplicity of $(t_1, t_2) \mapsto t_2 - t_1$ is at most M on $(A' \times A') \cap G$. Therefore,

$$K^{13}|A| \gg |A' - A'| \ge |(A' \times A') \cap G| M^{-1} \gg K^{-2}|A|^2 M^{-1}$$

where the outer two inequalities are the output of Lemma 1.9. This gives

$$|A| \ll K^{15} M,$$

or recalling the value of K,

$$|T_0| < |T_0|^{30} N^{-30\delta + o(1)} M.$$

Substituting the value of $M = N^{2\delta - 1}$ gives the result as claimed.

5 Proof of Lemma 1.9

The following argument is a modification of [2, Section 2]: By Cauchy-Schwarz and (10),

$$\begin{aligned}|G| &= \sum^{G}_{z\in A+A} |\{(x,y)\in G \,:\, x+y=z\}| \\ &\leq |A|^{\frac{1}{2}} |\{(x_1,y_1;x_2,y_2)\in G\times G \,:\, x_1+y_1=x_2+y_2\}|^{\frac{1}{2}},\end{aligned}$$

implying that

$$|\{(x_1,y_1;x_2,y_2)\in G\times G \,:\, x_1+y_1=x_2+y_2\}| > \frac{1}{K^2}|A|^3. \tag{18}$$

Denote $w(x) = |\{(x_1,x_2)\in A \,:\, x_1-x_2=x\}|$ and set

$$D = \{x \,:\, w(x) > \frac{1}{10K^2}|A|\}, \quad R = \{(x,x')\in A^2 : x_1-x_2\in D\},$$

and also write

$$R_{x_1} = \{x_2\in A \,:\, x_1-x_2\in D\}.$$

The set D is the "popular differences." Then,

$$\begin{aligned}&|\{(x_1,y_1;x_2,y_2)\in A^4 \,:\, x_1+y_1=x_2+y_2,\ \text{either } x_1-x_2\notin D \text{ or } y_1-x_2\notin D\}| \\ &\qquad \leq 2|A|^2\frac{1}{10K^2}|A| = \frac{1}{5K^2}|A|^3,\end{aligned}$$

since, for example, $x_1-x_2=y_1-y_2$ so each of at most $|A|^2$ pairs (x_1,x_2) with $x_1-x_2\notin D$ contribute at most $\frac{1}{10K^2}|A|$ possibilities for (y_1,y_2). The other contributions are estimated similarly. This estimate together with (18) gives

$$\begin{aligned}\frac{1}{2K^2}|A|^3 &< |\{(x_1,y_1;x_2,y_2)\in G\times G \,:\, x_1+y_1=x_2+y_2, x_1-x_2\in D, y_1-x_2\in D\}| \\ &\leq \sum_{(x_1,y_1)\in G} |R_{x_1}\cap R_{y_1}| \\ &= \sum_y |(R_y\times R_y)\cap G|. \end{aligned} \tag{19}$$

Let

$$Y = \{(x, x') \in A^2 \ : \ |R_x \cap R_{x'}| < \theta|A|\}$$

where θ is a parameter to be specified. Obviously from the definition of Y, we have

$$\sum_y |(R_y \times R_y) \cap Y| = \sum_{(x,x')\in Y} |R_x \cap R_{x'}| < \theta|A|^3.$$

Therefore, from (19), we see that

$$\sum_y |(R_y \times R_y) \cap G| > \frac{1}{4K^2}|A|^3 + \frac{1}{4K^2\theta}\sum_y |(R_y \times R_y) \cap Y|.$$

Thus, there is $y_0 \in A$ such that

$$|(R_{y_0} \times R_{y_0}) \cap G| > \frac{1}{4K^2}|A|^2 + \frac{1}{4K^2\theta}|(R_{y_0} \times R_{y_0}) \cap Y|. \tag{20}$$

In particular,

$$|R_{y_0}| > \frac{1}{2K}|A|. \tag{21}$$

Let

$$A' = \{x \in R_{y_0} \ : \ |(\{x\} \times R_{y_0}) \cap Y| < \frac{1}{3}|R_{y_0}|\}. \tag{22}$$

Then, clearly

$$\frac{1}{3}|R_{y_0}||R_{y_0}\backslash A'| < |(R_{y_0} \times R_{y_0}) \cap Y|,$$

and by (20), (5), we have

$$\begin{aligned} |(A' \times A') \cap G| &\geq |(R_{y_0} \times R_{y_0}) \cap G| - 2|R_{y_0}\backslash A'|.|R_{y_0}| \\ &\geq \frac{1}{4K^2}|A|^2 + \left(\frac{1}{4K^2\theta} - 6\right)|(R_{y_0} \times R_{y_0}) \cap Y| \\ &\geq \frac{1}{4K^2}|A|^2 \end{aligned} \tag{23}$$

if we take

$$\theta = \frac{1}{24K^2}.$$

Take now $(x_1, x_2) \in (A' \times A') \cap G$. By (22), there are at least $\frac{1}{3}|R_{y_0}|$ values of $x \in R_{y_0}$ such that $(x_1, x) \notin Y$ and $(x_2, x) \notin Y$. For each of these x, we have by the definition of Y

$$|R_{x_1} \cap R_x| \geq \frac{1}{24K^2}|A|, \quad |R_{x_2} \cap R_x| \geq \frac{1}{24K^2}|A|,$$

and then

$$\begin{aligned} x_1 - x_2 &= (x_1 - x) - (x_2 - x) \\ &= (x_1 - y_1) - (x - y_1) - (x_2 - y_2) + (x - y_2) \end{aligned} \tag{24}$$

for at least $\frac{|A|^2}{576K^4}$ pairs (y_1, y_2) (depending on x, x_1, x_2) with

$$(x_1, y_1), (x, y_1), (x_2, y_2), (x, y_2) \in R.$$

By definition of R and D, each of the parenthetical terms in (24) admits a representation in at least $\frac{|A|}{10K^2}$ ways as a difference of elements of A, and therefore, the number of representations

$$x_1 - x_2 = (\tau_1 - \tau_2) - (\tau_3 - \tau_4) - (\tau_5 - \tau_6) + (\tau_7 - \tau_8), \quad \tau_i \in A \tag{25}$$

is at least

$$\frac{1}{3}|R_{y_0}|.\frac{|A|^2}{576K^4}.\left(\frac{|A|}{10K^2}\right)^4 \overset{(21)}{\gg} \frac{1}{K^{13}}|A|^7.$$

Considering the map on $(\tau_i)_{i=1}^8$ given by (25), we get

$$|A' - A'| \ll \frac{|A|^8}{K^{-13}|A|^7} = K^{13}|A|.$$

This together with the previously established (23) proves Lemma 1.9.

References

1. Bourgain, J.: On the dimension of Kakeya sets and related maximal inequalities. Geom. Funct. Anal. **9**(2), 256–282 (1999)
2. Bourgain, J.: Sum-product theorems and applications. In: Additive Number Theory, pp. 9–38. Springer, New York (2010)
3. Bourgain, J.: Integral Apollonian circle packings and prime curvatures. J. Anal. Math. **118**(1), 221–249 (2012)
4. Bourgain, J., Kontorovich, A.: On representations of integers in thin subgroups of SL(2, **Z**). GAFA **20**(5), 1144–1174 (2010)

5. Bourgain, J., Kontorovich, A.: On Zaremba's conjecture. Annals Math. **180**(1), 137–196 (2014)
6. Bourgain, J., Kontorovich, A.: On the local-global conjecture for integral Apollonian gaskets. Invent. Math. **196**(3), 589–650 (2014)
7. Bourgain, J., Kontorovich, A.: The affine sieve beyond expansion I: Thin hypotenuses. Int. Math. Res. Not. IMRN (19), 9175–9205 (2015)
8. Bourgain, J., Kontorovich, A.: Beyond expansion II: low-lying fundamental geodesics. J. Eur. Math. Soc. (JEMS) **19**(5), 1331–1359 (2017)
9. Bourgain, J., Kontorovich, A.: Beyond expansion IV: Traces of thin semigroups. Discrete Anal., Paper No. 6, 27 (2018)
10. Bourgain, J., Kontorovich, A.: Beyond expansion III: Reciprocal geodesics, 2019. To appear, Duke Math. J. arXiv:1610.07260
11. Bourgain, J., Gamburd, A., Sarnak, P.: Affine linear sieve, expanders, and sum-product. Invent. Math. **179**(3), 559–644 (2010)
12. Bourgain, J., Kontorovich, A., Sarnak, P.: Sector estimates for hyperbolic isometries. GAFA **20**(5), 1175–1200 (2010)
13. Bourgain, J., Gamburd, A., Sarnak, P.: Generalization of Selberg's 3/16th theorem and affine sieve. Acta Math **207**, 255–290 (2011)
14. Friedlander, J., Iwaniec, H.: The polynomial $X^2 + Y^4$ captures its primes. Ann. Math. (2) **148**(3), 945–1040 (1998)
15. Frolenkov, D.A., Kan, I.D.: A strengthening of a theorem of Bourgain-Kontorovich II. Mosc. J. Comb. Number Theory **4**(1), 78–117 (2014)
16. Graham, R.L., Lagarias, J.C., Mallows, C.L., Wilks, A.R., Yan, C.H.: Apollonian circle packings: number theory. J. Number Theory **100**(1), 1–45 (2003)
17. Good, I.J.: The fractional dimensional theory of continued fractions. Proc. Camb. Philos. Soc. **37**, 199–228 (1941)
18. Gowers, W.T.: A new proof of Szemerédi's theorem for arithmetic progressions of length four. Geom. Funct. Anal. **8**(3), 529–551 (1998)
19. Heath-Brown, D.R.: Primes represented by $x^3 + 2y^3$. Acta Math. **186**(1), 1–84 (2001)
20. Hensley, D.: The distribution of badly approximable numbers and continuants with bounded digits. In: Théorie des nombres (Quebec, PQ, 1987), pp. 371–385. de Gruyter, Berlin (1989)
21. Hong, J., Kontorovich, A.: Almost prime coordinates for anisotropic and thin pythagorean orbits. Isr. J. Math. **209**(1), 397–420 (2015)
22. Jenkinson, O.: On the density of Hausdorff dimensions of bounded type continued fraction sets: the Texan conjecture. Stoch. Dyn. **4**(1), 63–76 (2004)
23. Kan, I.D.: A strengthening of a theorem of Bourgain and Kontorovich. V. Tr. Mat. Inst. Steklova **296**(Analiticheskaya i Kombinatornaya Teoriya Chisel), 133–139 (2017)
24. Kontorovich, A.: The hyperbolic lattice point count in infinite volume with applications to sieves. Duke J. Math. **149**(1), 1–36 (2009). arXiv:0712.1391
25. Kontorovich, A.: A pseudo-twin primes theorem (2012). In: Proceedings of the Edinburg Conference, Workshop on Multiple Dirichlet Series. arXiv:0507569. To appear
26. Kontorovich, A.: From Apollonius to Zaremba: local-global phenomena in thin orbits. Bull. Am. Math. Soc. (N.S.) **50**(2), 187–228 (2013)
27. Kontorovich, A.: Levels of distribution and the affine sieve. Ann. Fac. Sci. Toulouse Math. (6) **23**(5), 933–966 (2014)
28. Kontorovich, A.: Applications of thin orbits. In: Dynamics and Analytic Number Theory, vol. 437. London Math. Soc. Lecture Note Ser., pp. 289–317. Cambridge Univ. Press, Cambridge (2016)
29. Kontorovich, A.: The local-global principle for integral Soddy sphere packings. J. Modern Dynam. **15**, 209–236 (2019)
30. Kontorovich, A., Oh, H.: Almost prime Pythagorean triples in thin orbits. J. Reine Angew. Math. **667**, 89–131 (2012). arXiv:1001.0370
31. McMullen, C.T.: Uniformly Diophantine numbers in a fixed real quadratic field. Compos. Math. **145**(4), 827–844 (2009)

32. McMullen, C.: Dynamics of units and packing constants of ideals (2012). Online lecture notes, http://www.math.harvard.edu/~ctm/expositions/home/text/papers/cf/slides/slides.pdf
33. Mercat, P.: Construction de fractions continues périodiques uniformément bornées (2012). To appear, J. Théor. Nombres Bordeaux
34. Patterson, S.J.: The Laplacian operator on a Riemann surface. Compositio Math. **31**(1), 83–107 (1975)
35. Pjateckiĭ-Šapiro, I.I.: On the distribution of prime numbers in sequences of the form $[f(n)]$. Mat. Sb. **33**, 559–566 (1953)
36. Sarnak, P.: Letter to J. Lagarias (2007). http://web.math.princeton.edu/sarnak/AppolonianPackings.pdf
37. Sarnak, P.: Reciprocal geodesics. In: Analytic Number Theory, vol. 7. Clay Math. Proc., pp. 217–237. Amer. Math. Soc., Providence, RI (2007)
38. Wilson, S.M.J.: Limit points in the Lagrange spectrum of a quadratic field. Bull. Soc. Math. France **108**, 137–141 (1980)
39. Woods, A.C.: The Markoff spectrum of an algebraic number field. J. Aust. Math. Soc. Ser. A **25**(4), 486–488 (1978)
40. Zaremba, S.K.: La méthode des “bons treillis” pour le calcul des intégrales multiples. In: Applications of Number Theory to Numerical Analysis (Proc. Sympos., Univ. Montreal, Montreal, Que., 1971), pp. 39–119. Academic Press, New York (1972)

Polynomial Equations in Subgroups and Applications

Sergei V. Konyagin, Igor E. Shparlinski, and Ilya V. Vyugin

Dedicated to the Memory of Jean Bourgain

Abstract We obtain a new bound for the number of solutions to polynomial equations in cosets of multiplicative subgroups in finite fields, which generalizes previous results of P. Corvaja and U. Zannier (2013). We also obtain a conditional improvement of recent results of J. Bourgain, A. Gamburd, and P. Sarnak (2016) and S. V. Konyagin, S. V. Makarychev, I. E. Shparlinski, and I. V. Vyugin (2019) on the structure of solutions to the reduction of the Markoff equation $x^2 + y^2 + z^2 = 3xyz$ modulo a prime p.

2010 Mathematics Subject Classification 11D79, 11T06

S. V. Konyagin (✉)
Steklov Mathematical Institute, Moscow, Russia
e-mail: konyagin@mi-ras.ru

I. E. Shparlinski
Department of Pure Mathematics, University of New South Wales, Sydney, NSW, Australia
e-mail: igor.shparlinski@unsw.edu.au

I. V. Vyugin
Institute for Information Transmission Problems RAS, Moscow, Russia

HSE University, Moscow, Russia

Steklov Mathematical Institute, Moscow, Russia
e-mail: vyugin@gmail.com

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_12

1 Introduction

1.1 Background and Motivation

Bourgain, Gamburd, and Sarnak [2, 3] have recently initiated the study of reductions modulo p of the set $\mathcal{M}$ of *Markoff triples* $(x, y, z) \in \mathbb{N}^3$ which are positive integer solutions to the Diophantine equation

$$x^2 + y^2 + z^2 = 3xyz, \qquad (x, y, z) \in \mathbb{Z}^3. \tag{1}$$

Simple computation shows that the map

$$\mathcal{R}_1 : (x, y, z) \mapsto (3yz - x, y, z)$$

and similarly defined maps $\mathcal{R}_2$, $\mathcal{R}_3$ (which are all involutions) send one Markoff triple to another. Due to the symmetry of (1), the set $\mathcal{M}$ is also invariant under permutations. Let S_3 be the group of permutations of order 3. For $\sigma \in S_3$ we denote by Π_σ the mapping $\pi(x_1, x_2, x_3) = (x_{\sigma(1)}, x_{\sigma(2)}, x_{\sigma(3)})$. It is easy to check that the transformations $\mathcal{R}_i$, $i = 1, 2, 3$ and the mappings Π_σ generate a group of transformations acting on $\mathcal{M}$.

A celebrated result of Markoff [18, 19] asserts that all integer positive solutions to (1) can be generated from the solution (1, 1, 1) by using sequences of the above transformations.

This naturally leads to the notion of the *functional graph* on Markoff triples with the "root" (1, 1, 1) and edges $(x_1, y_1, z_1) \to (x_2, y_2, z_2)$, povided that $(x_2, y_2, z_2) = \mathcal{T}(x_1, y_1, z_1)$, where

$$\mathcal{T} \in \{\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3\} \cup \{\Pi_\sigma : \sigma \in S_3\}. \tag{2}$$

In this terminology, the result of Markoff [18, 19] asserts that this graph is *connected*.

Baragar [1, Section V.3] and, more recently, Bourgain, Gamburd, and Sarnak [2, 3] conjecture that this property is preserved modulo all sufficiently large primes p and the set of non-zero solutions $\mathcal{M}_p$ to (1) considered modulo p. In particular, this means that $\mathcal{M}_p$ can be obtained from the set of Markoff triples $\mathcal{M}$ reduced modulo p.

This conjecture, which we can also write as $\mathcal{M}_p = \mathcal{M} \pmod p$, means that the functional graph $\mathcal{X}_p$ associated with the transformation (2) remains connected.

Accordingly, if we define by $\mathcal{C}_p \subseteq \mathcal{M}_p$ the set of the triples in the largest connected component of the above graph $\mathcal{X}_p$, then we can state:

Conjecture 1.1 (Baragar [1]; Bourgain, Gamburd, and Sarnak [2, 3]) *For every prime p, we have $\mathcal{C}_p = \mathcal{M}_p$.*

Bourgain, Gamburd, and Sarnak [2, 3] have obtained several major results towards Conjecture 1.1; see also [4, 8, 9, 11]. For example, by [2, Theorem 1], we have

$$\#\left(\mathcal{M}_p \setminus \mathcal{C}_p\right) = p^{o(1)}, \qquad \text{as } p \to \infty, \tag{3}$$

and also by [2, Theorem 2], we know that Conjecture 1.1 holds for all but maybe at most $X^{o(1)}$ primes $p \le X$ as $X \to \infty$.

The bound (3) has been improved in [16, Theorem 1.2] as

$$\#\left(\mathcal{M}_p \setminus \mathcal{C}_p\right) \le \exp\left((\log p)^{2/3+o(1)}\right), \qquad \text{as } p \to \infty. \tag{4}$$

Furthermore, Bourgain, Gamburd, and Sarnak [2, 3] have also proved that the size of any connected component of the graphs $\mathcal{X}_p$ is at least

$$\#\mathcal{X}_p \ge c(\log p)^{1/3}, \tag{5}$$

for some absolute constant $c > 0$. In turn, the bound (5) has been improved in [16, Theorem 1.3] as

$$\#\mathcal{X}_p \ge c(\log p)^{7/9}. \tag{6}$$

The improvements in (4) and (6) are based on a bound of Corvaja and Zannier [7, Corollary 2], on the number of solutions to the equation

$$P(u, v) = 0, \quad (u, v) \in \mathcal{G}_1 \times \mathcal{G}_2,$$

where P is a bivariate absolutely irreducible polynomial over the finite field $\mathbb{F}_p$ of p elements and $\mathcal{G}_1, \mathcal{G}_2 \subseteq \overline{\mathbb{F}}_p$ are multiplicative groups in the algebraic closure $\overline{\mathbb{F}}_p$ of $\mathbb{F}_p$; see also [12, 14, 17, 20] for some related results.

Motivated by the above results and connections, here we

- Derive a new bound on the number of solutions in subgroups to a systems of several polynomials which covers under a unified setting the results of [7, 17, 20];
- Obtain an improvement of (4) under a very plausible conjecture on the number of solutions in subgroups of some particular equation over $\mathbb{F}_p^*$.

1.2 New Results

As before, for a prime p, we use $\overline{\mathbb{F}}_p$ to denote the algebraic closure of the finite field $\mathbb{F}_p$ of p elements.

We also say that a polynomial is irreducible if and only it is absolutely irreducible.

For a bivariate irreducible polynomial

$$P(X, Y) = \sum_{i+j \le d} a_{ij} X^i Y^j \in \overline{\mathbb{F}}_p[X, Y] \tag{7}$$

of total degree $\deg P \le d$, we define $P^\sharp(X, Y)$ as the homogeneous polynomial of degree $d^\sharp = \min\{i + j : a_{ij} \ne 0\}$ given by

$$P^\sharp(X, Y) = \sum_{i+j=d^\sharp} a_{ij} X^i Y^j. \tag{8}$$

We also consider the set of polynomials $\mathcal{P}$:

$$\mathcal{P} = \{P(\lambda X, \mu Y) \mid \lambda, \mu \in \overline{\mathbb{F}}_p^*\}.$$

Since $P(X, Y)$ is irreducible, it is not homogenous, and thus $P(X, Y) \ne P^\sharp(X, Y)$. Hence, we can define g as the greatest common divisor of the following set of differences:

$$g = \gcd\{i_1 + j_1 - i_2 - j_2 \ : \ a_{i_1,j_1} a_{i_2,j_2} \ne 0\}. \tag{9}$$

Given a multiplicative subgroup $\mathcal{G} \subseteq \overline{\mathbb{F}}_p$, we say that two polynomials $P, Q \in \overline{\mathbb{F}}_p[X, Y]$ are *$\mathcal{G}$-independent* if there is no $(u, v) \in \mathcal{G}^2$ and $\gamma \in \overline{\mathbb{F}}_p^*$ such that polynomials $P(X, Y)$ and $\gamma Q(uX, vY)$ coincide.

We now fix h polynomials

$$P_k(X, Y) = P(\lambda_k X, \mu_k Y) \in \mathcal{P}, \qquad k = 1, \dots, h, \tag{10}$$

which are $\mathcal{G}$-independent.

The following result generalizes a series of previous estimates of a similar type; see [7, 12, 14, 17, 20] and references therein.

Theorem 1.2 *Suppose that P is irreducible,*

$$\deg_X P = m \qquad \text{and} \qquad \deg_Y P = n$$

and also that $P^\sharp(X, Y)$ consists of at least two monomials. There exists a constant $c_0(m, n)$, depending only on m and n, such that for any multiplicative subgroup $\mathcal{G} \subseteq \overline{\mathbb{F}}_p$ of order $t = \#\mathcal{G}$ satisfying

$$\frac{1}{2} p^{3/4} h^{-1/4} \ge t \ge \max\{h^2, c_0(m, n)\},$$

and $\mathcal{G}$-independent polynomials (10) *we have*

$$\sum_{i=1}^{h} \# \left\{ (u, v) \in \mathcal{G}^2 \ : \ P_i(u, v) = 0 \right\} < 12mn(m+n)gh^{2/3}t^{2/3}.$$

Our next result is conditional on the following:

Conjecture 1.3 *There exist constants $\varepsilon_0 > 0$ and A such that for any prime p, any subgroup $\mathcal{G} \subseteq \overline{\mathbb{F}}_p$ with $\#\mathcal{G} \le p^{\varepsilon_0}$, and any elements $\alpha_{1,1}, \alpha_{1,2}, \alpha_{2,1}, \alpha_{2,2} \in \overline{\mathbb{F}}_p$ satisfying*

$$\alpha_{1,1} \neq 0, \quad \alpha_{1,2} \neq 0, \quad \alpha_{1,1}\alpha_{2,2} - \alpha_{1,2}\alpha_{2,1} \neq 0, \tag{11}$$

the equation

$$\frac{\alpha_{1,1}u - \alpha_{1,2}}{\alpha_{2,1}u - \alpha_{2,2}} = v \tag{12}$$

has at most A solutions in $u, v \in \mathcal{G}$.

Remark 1.4 It is likely that the constant A in Conjecture 1.3 cannot be taken less than 9, even for $\mathcal{G} \subseteq \mathbb{F}_p$ rather than for $\mathcal{G} \subseteq \overline{\mathbb{F}}_p$; see some heuristic arguments in Sect. 6. It is possible that this is optimal and Conjecture 1.3 holds with $A = 9$. Also we must have $\varepsilon_0 \le 1/2$; see Sect. 6.

Remark 1.5 It is easy to see that using the bound (4) instead of (3) in the argument of the proof of Theorem 1.6 immediately allows us to relax the condition of Conjecture 1.3 to counting solutions in subgroups $\mathcal{G} \subseteq \mathbb{F}_{p^2}$ of order $\#\mathcal{G} \le \exp\left((\log p)^{2/3+\varepsilon_0}\right)$. However, we believe Conjecture 1.3 holds as stated.

Theorem 1.6 *If Conjecture 1.3 holds for some ε_0 and A, then for sufficiently large p we have*

$$\#\left(\mathcal{M}_p \setminus \mathcal{C}_p\right) \le (\log p)^B,$$

where $B = 16 \log A + c$ for an absolute constant c.

Remark 1.7 Recently (after this work has been submitted) Chen [6] presented a striking result giving a full resolution of Conjecture 1.1 (for all sufficiently large p). However, we still believe that our present argument as well as the argument of [16] are of interest since they apply to more general equations than (1), for example, to equations of the form $x^2 + y^2 + z^2 = axyz + b$, which the method of [6] is limited to (1).

2 Solutions to Polynomial Equations in Subgroups of Finite Fields

2.1 Stepanov's Method

Consider a polynomial $\Phi \in \overline{\mathbb{F}}_p[X, Y, Z]$ such that

$$\deg_X \Phi < A, \quad \deg_Y \Phi < B, \quad \deg_Z \Phi < C,$$

that is,

$$\Phi(X, Y, Z) = \sum_{0\le a<A} \sum_{0\le b<B} \sum_{0\le c<C} \omega_{a,b,c} X^a Y^b Z^c.$$

We assume

$$A < t,$$

where $t = \#\mathcal{G}$ is the order of the subgroup $\mathcal{G} \subseteq \mathbb{F}_p^*$, and consider the polynomial

$$\Psi(X, Y) = Y^t \Phi(X/Y, X^t, Y^t).$$

Clearly,

$$\deg \Psi \le t + t(B-1) + t(C-1) = (B + C - 1)t.$$

We now fix some $\mathcal{G}$-independent polynomials (10) and define the sets

$$\mathcal{F}_k = \left(\lambda_k^{-1}\mathcal{G} \times \mu_k^{-1}\mathcal{G}\right), \quad k = 1, \ldots, h, \qquad \text{and} \qquad \mathcal{E} = \bigcup_{k=1}^{h} \mathcal{F}_k. \tag{13}$$

We also consider the locus of singularity

$$\mathcal{M}_{sing} = \Big\{(X, Y) \mid XY = P(X, Y) = 0 \text{ or } \frac{\partial}{\partial Y} P(X, Y) = P(X, Y) = 0\Big\}.$$

Lemma 2.1 *Let $P(X, Y)$ be an irreducible polynomial of bi-degree*

$$\left(\deg_X P, \deg_Y P\right) = (m, n)$$

and let $n \ge 1$. Then, for the cardinality of the set $\mathcal{M}_{sing}$, the following holds:

$$\#\mathcal{M}_{sing} \le (m + n)^2.$$

Proof If the polynomial $P(X, Y)$ is irreducible, then the polynomials $P(X, Y)$ and $\frac{\partial P}{\partial Y}(X, Y)$ are relatively prime. Thus, the Bézout theorem yields the bound $L \le (m+n)(m+n-1)$, where L is the number of roots of the system

$$\frac{\partial}{\partial Y} P(X, Y) = P(X, Y) = 0.$$

Clearly, the number of X with $P(X, 0) = 0$ is less than or equal to $\deg_X P(X, Y) = m$, the number of pairs $(0, Y)$ on the curve

$$P(X, Y) = 0, \tag{14}$$

where P is given by (7), is less than or equal to $\deg_Y P(X, Y) = n$. The total numbers of such pairs is at most $L + m + n \le (m+n)^2$. □

Assume that the polynomial Ψ and the $\mathcal{G}$-independent polynomials (10) satisfy the following conditions:

- All pairs in the set

$$\left\{(X, Y) \in \mathcal{E} \setminus \mathcal{M}_{sing} \mid P(X, Y) = 0\right\}$$

  are zeros of orders at least D of the function $\Psi(X, Y)$ on the curve (14);
- The polynomials $\Psi(X, Y)$ and $P(X, Y)$ are relatively prime.

If these conditions are satisfied, then the *Bézout theorem* gives us the upper bound $D^{-1} \deg \Psi \deg P + \#\mathcal{M}_{sing}$ for the number of roots (x, y) of the system

$$\Psi(X, Y) = P(X, Y) = 0, \qquad (X, Y) \in \mathcal{G}.$$

Since the polynomials P_k are $\mathcal{G}$-independent, the sets $\mathcal{F}_k$ given by (13) are disjoint, and also there is a one-to-one correspondence between the zeros:

$$\begin{aligned} &P_k(X, Y) = 0,\ (X, Y) \in \mathcal{G}^2, \\ &\qquad\qquad \iff P(u, v) = 0,\ (u, v) = (\lambda_k^{-1} X, \mu_k^{-1} Y) \in \mathcal{F}_k. \end{aligned}$$

Therefore, we obtain the bound

$$\begin{aligned} N_h &\le \frac{\deg \Psi \cdot \deg P}{D} + \#\mathcal{M}_{sing} \\ &\le \frac{(m+n)(B+C-1)t}{D} + \#\mathcal{M}_{sing} \end{aligned} \tag{15}$$

on the total number of zeros of P_k in $\mathcal{G}^2$, $k = 1, \ldots, h$:

$$N_h = \sum_{k=1}^{h} \#\{(u,v) \in \mathcal{G}^2 \ : \ P_k(u,v) = 0\}.$$

For completeness, we present proofs of several results from [17] which we use here as well.

2.2 Some Divisibilities and Non-divisibilities

We begin with some simple preparatory results on the divisibility of polynomials.

Lemma 2.2 *Suppose that $Q(X,Y) \in \mathbb{F}_p[X,Y]$ is an irreducible $\mathcal{G}$-independent polynomial such that*

$$Q(X,Y) \mid \Psi(X,Y)$$

and $Q^{\sharp}(X,Y)$ consists of at least two monomials. Then,

$$Q^{\sharp}(X,Y)^{\lfloor t/e \rfloor} \mid \Psi^{\sharp}(X,Y),$$

where $Q^{\sharp}(X,Y)$ and $\Psi^{\sharp}(X,Y)$ are defined as in (8) and e is defined as g in (9) with respect to $Q(X,Y)$ instead of $P(x,y)$.

Proof Consider $\rho \in \mathcal{G}$ and substitute $X = \rho\widetilde{X}$ and $Y = \rho\widetilde{Y}$ in the polynomials $Q(X,Y)$ and $\Psi(X,Y)$. Then,

$$Q(X,Y) \longmapsto Q_\rho(\widetilde{X},\widetilde{Y}) = Q(\rho\widetilde{X},\rho\widetilde{Y}),$$

and

$$\begin{aligned}\Psi(X,Y) &= \Psi(\rho\widetilde{X},\rho\widetilde{Y})\\ &= (\rho\widetilde{Y})^t\Phi((\rho\widetilde{X})/(\rho\widetilde{Y}),(\rho\widetilde{X})^t,(\rho\widetilde{Y})^t) = \Psi(\widetilde{X},\widetilde{Y}),\end{aligned}$$

because $\rho^t = 1$. Hence, for any $\rho \in \mathcal{G}$, we have

$$Q_\rho(X,Y) \mid \Psi(X,Y),$$

and we also note that $Q_\rho(X,Y)$ is irreducible.

Since $Q(X,Y)$ is irreducible, $e \geqslant 1$ is correctly defined, and there exist at least $s = \lfloor t/e \rfloor$ elements $\rho_1,\ldots,\rho_s \in \mathcal{G}$ such that all nontrivial ratios $Q_{\rho_i}(X,Y)/Q_{\rho_j}(X,Y)$ are not constants, that is,

$$Q_{\rho_i}(X,Y)/Q_{\rho_j}(X,Y) \notin \overline{\mathbb{F}}_p, \qquad 1 \le i < j \le s. \tag{16}$$

Obviously, the polynomials $Q_{\rho_1}(X, Y), \ldots, Q_{\rho_s}(X, Y)$ are pairwise relatively prime, because they are irreducible and satisfy (16). Furthermore, the polynomials $Q^{\sharp}_{\rho_i}(X, Y)$ are homogeneous of degree $d^{\sharp}$, and the following holds

$$Q^{\sharp}(X, Y) = \rho_1^{-d^{\sharp}} Q^{\sharp}_{\rho_1}(X, Y) = \ldots = \rho_s^{-d^{\sharp}} Q^{\sharp}_{\rho_s}(X, Y).$$

So, we have

$$Q_{\rho_1}(X, Y) \cdot \ldots \cdot Q_{\rho_s}(X, Y) \mid \Psi(X, Y);$$

consequently,

$$Q^{\sharp}_{\rho_1}(X, Y) \cdot \ldots \cdot Q^{\sharp}_{\rho_s}(X, Y) \mid \Psi^{\sharp}(X, Y).$$

Since

$$Q^{\sharp}_{\rho_1}(X, Y) \cdot \ldots \cdot Q^{\sharp}_{\rho_s}(X, Y) = (\rho_1 \cdot \ldots \cdot \rho_s)^{d^{\sharp}} Q^{\sharp}(X, Y)^s,$$

we obtain the desired result. □

Lemma 2.3 *Let $G(X, Y), H(X, Y) \in \mathbb{F}_p[X, Y]$ be two homogeneous polynomials. Also suppose that $G(X, Y)$ consists of at least two nonzero monomials,* $\deg H < p$, *and the number of monomials of the polynomial $H(X, Y)$ does not exceed s for some positive integer $s < p$. Then,*

$$G(X, Y)^s \nmid H(X, Y).$$

Proof Clearly, if $G(X, Y)^s \mid H(X, Y)$, then $G(X, 1)^s \mid H(X, 1)$. The polynomial $G(X, 1)$ has at least one nonzero root. It has been proved in [14, Lemma 6] that such a polynomial $H(X, 1)$ cannot have a nonzero root of order s and the result follows. □

Lemma 2.4 *If $AB < t/g$ and* $\deg \Psi < p$, *then for the polynomial $P(X, Y)$ given by* (7) *we have*

$$P(X, Y) \nmid \Psi(X, Y).$$

2.3 Derivatives on Some Curves

There we study derivatives on an algebraic curve and define some special differential operators. Throughout this section, we use

$$\frac{\partial}{\partial X}, \quad \frac{\partial}{\partial Y} \qquad \text{and} \qquad \frac{d}{dX}$$

for standard partial derivatives with respect to X and Y and for the derivative with respect to X along the curve (14), respectively. In particular,

$$\frac{d}{dX} = \frac{\partial}{\partial X} + \frac{dY}{dX}\frac{\partial}{\partial Y}, \tag{17}$$

where by the implicit function theorem from the Eq. (14), we have

$$\frac{dY}{dX} = -\frac{\frac{\partial P}{\partial X}(X, Y)}{\frac{\partial P}{\partial Y}(X, Y)}.$$

We also define inductively

$$\frac{d^k}{dX^k} = \frac{d}{dX}\frac{d^{k-1}}{dX^{k-1}}$$

the k-th derivative on the curve (14).

Consider the polynomials $q_k(X, Y)$ and $r_k(X, Y)$, $k \in \mathbb{N}$, which are defined inductively as

$$q_1(X, Y) = -\frac{\partial}{\partial X}P(X, Y), \qquad r_1(X, Y) = \frac{\partial}{\partial Y}P(X, Y),$$

and

$$\begin{aligned} q_{k+1}(X, Y) &= \frac{\partial q_k}{\partial X}\left(\frac{\partial P}{\partial Y}\right)^2 \\ &\qquad - \frac{\partial q_k}{\partial Y}\frac{\partial P}{\partial X}\frac{\partial P}{\partial Y} - (2k-1)q_k(X, Y)\frac{\partial^2 P}{\partial X \partial Y}\frac{\partial P}{\partial Y} \\ &\qquad\qquad + (2k-1)q_k(X, Y)\frac{\partial^2 P}{\partial Y^2}\frac{\partial P}{\partial X}, \\ r_{k+1}(X, Y) &= r_k(X, Y)\left(\frac{\partial P}{\partial Y}\right)^2 = \left(\frac{\partial P}{\partial Y}\right)^{2k+1}. \end{aligned} \tag{18}$$

We now show by induction that

$$\frac{d^k}{dX^k}Y = \frac{q_k(X, Y)}{r_k(X, Y)}, \qquad k \in \mathbb{N}. \tag{19}$$

The base of induction is

$$\frac{d}{dX}Y = -\frac{\frac{\partial}{\partial X}P(X,Y)}{\frac{\partial}{\partial Y}P(X,Y)} = \frac{q_1(X,Y)}{r_1(X,Y)}.$$

One can now easily verify that assuming (19) and (17) we have

$$\frac{d^{k+1}}{dX^{k+1}}Y = \frac{d}{dX}\frac{d^k}{dX^k}Y = \frac{d}{dX}\frac{q_k(X,Y)}{r_k(X,Y)} = \frac{q_{k+1}(X,Y)}{r_{k+1}(X,Y)},$$

where q_{k+1} and r_{k+1} are given by (18), which concludes the induction and proves the formula (19).

The implicit function theorem gives us the derivatives $\frac{d^{k+1}}{dX^{k+1}}Y$ at a point (X, Y) on the algebraic curve (14), if the denominator $r_k(X, Y)$ is not equal to zero. Otherwise, $r_k(X, Y) = 0$ if and only if the following system holds:

$$\frac{\partial}{\partial Y}P(X,Y) = P(X,Y) = 0.$$

Let us give the following estimates:

Lemma 2.5 *For all integers $k \geq 1$, the degrees of the polynomials $q_k(X, Y)$ and $r_k(X, Y)$ satisfy the bounds*

$$\deg_X q_k \leq (2k-1)m - k, \qquad \deg_Y q_k \leq (2k-1)n - 2k + 2,$$
$$\deg_X r_k \leq (2k-1)m, \qquad \deg_Y r_k \leq (2k-1)(n-1).$$

Proof Direct calculations show that

$$\deg_X q_1 \leq m - 1 \qquad \text{and} \qquad \deg_Y q_1 \leq n,$$

and using (18) (with $k - 1$ instead of k) and examining the degree of each term, we obtain the inequalities

$$\deg_X q_k \leq \deg_X q_{k-1} + 2m - 1 \leq (2k-1)m - k,$$
$$\deg_Y q_k \leq \deg_y q_{k-1} + 2n - 2 \leq (2k-1)n - 2k + 2.$$

We now obtain the desire bounds on $\deg_X q_k$ and $\deg_Y q_k$ by induction.

For the polynomials r_k, the statement is obvious. □

Lemma 2.6 *Let $Q(X, Y) \in \mathbb{F}_p[X, Y]$ be a polynomial such that*

$$\deg_X Q(X,Y) \leq \mu, \quad \deg_Y Q(X,Y) \leq \nu \tag{20}$$

and $P(X, Y) \in \mathbb{F}_p[X, Y]$ be a polynomial such that

$$\deg_X P(X,Y) \le m, \quad \deg_Y P(X,Y) \le n.$$

Then, the divisibility condition

$$P(X,Y) \mid Q(X,Y) \tag{21}$$

on the coefficients of the polynomial $Q(X,Y)$ is equivalent to a certain system of not more than $(\mu+\nu+1)mn$ homogeneous linear algebraic equations in coefficients of $Q(X,Y)$ as variables.

Proof The dimension of the vector space $\mathcal{L}$ of polynomials $Q(X,Y)$ that satisfy (20) is equal to $(\mu+1)(\nu+1)$. Let us call the vector subspace of polynomials $Q(X,Y)$ that satisfy (20) and (21) by $\widetilde{\mathcal{L}}$. Because $Q(X,Y) = P(X,Y)R(X,Y)$ where the polynomial $R(X,Y)$ is such that

$$\deg_X R(X,Y) \le \mu - m \qquad \text{and} \qquad \deg_Y R(X,Y) \le \nu - n, \tag{22}$$

then the vector space $\widetilde{\mathcal{L}}$ is isomorphic to the vector space of the coefficients of the polynomials $R(x,y)$ satisfying (22). The dimension of the vector space $\widetilde{\mathcal{L}}$ is equal to

$$\dim \widetilde{\mathcal{L}} = (\mu - m + 1)(\nu - n + 1).$$

It means that the subspace $\widetilde{\mathcal{L}}$ of the space $\mathcal{L}$ is given by a system of

$$\begin{aligned}(\mu+1)(\nu+1) &- (\mu-m+1)(\nu-n+1)\\ &= \mu n + \nu m - mn + m + n + 1 \le (\mu+\nu+1)mn\end{aligned}$$

homogeneous linear algebraic equations. □

As in [17], we now consider the differential operators:

$$D_k = \left(\frac{\partial P}{\partial Y}\right)^{2k-1} X^k Y^k \frac{d^k}{dX^k}, \qquad k \in \mathbb{N}, \tag{23}$$

where, as before, $\frac{d^k}{dX^k}$ denotes the k-th derivative on the algebraic curve (14) with the local parameter X. We note now that the derivative of a polynomial in two variables along a curve is a rational function. As one can see from the inductive formula for $\frac{d^k}{dX^k}$, the result of applying any operator D_k to a polynomial in two variables is again a polynomial in two variables.

Consider non-negative integers a, b, c such that $a < A$, $b < B$, $c < C$. From the formulas (19) for derivatives on the algebraic curve (14), we obtain by induction the following relations:

$$\begin{aligned} &D_k \left(\frac{X}{Y}\right)^a X^{bt} Y^{(c+1)t} = R_{k,a,b,c}(X,Y) \left(\frac{X}{Y}\right)^a X^{bt} Y^{(c+1)t}, \\ &D_k \Psi(X,Y)|_{X,Y \in \mathcal{F}_i} = R_{k,i}(X,Y)|_{X,Y \in \mathcal{F}_i}, \end{aligned} \tag{24}$$

where $\mathcal{F}_i$ are from formula (13),

$$\begin{aligned} &R_{k,i}(X,Y) \\ &\quad = \sum_{0 \le a < A} \sum_{0 \le b < B} \sum_{0 \le c < C} \omega_{a,b,c} R_{k,a,b,c}(X,Y) \left(\frac{X}{Y}\right)^a \lambda_i^{-bt} \mu_i^{-(c+1)t} \end{aligned} \tag{25}$$

for some coefficients $\omega_{a,b,c} \in \mathbb{F}_p$, $a < A$, $b < B$, $c < C$, and λ_i, μ_i from (13).

We now define

$$\widetilde{R}_{k,i}(X,Y) = Y^{A-1} R_{k,i}(X,Y). \tag{26}$$

Lemma 2.7 *The rational functions $R_{k,a,b,c}(X,Y)$ and $\widetilde{R}_{k,i}(X,Y)$, given by (24) and (26), are polynomials of degrees*

$$\deg_X R_{k,a,b,c} \le 4km, \qquad \deg_Y R_{k,a,b,c} \le 4kn,$$

and

$$\deg_X \widetilde{R}_{k,i} \le A + 4km, \qquad \deg_Y \widetilde{R}_{k,i} \le A + 4kn.$$

Proof We have

$$\begin{aligned} \frac{d^k}{dX^k} X^{a+bt} Y^{(c+1)t-a} = \sum_{(\ell_1,\ldots,\ell_s)} &C_{\ell_1,\ldots,\ell_s} X^{a+bt-k+\sum_{i=1}^s \ell_i} \\ &Y^{(c+1)t-a-s} \left(\frac{d^{\ell_1} Y}{dX^{\ell_1}}\right) \ldots \left(\frac{d^{\ell_s} Y}{dX^{\ell_s}}\right), \end{aligned} \tag{27}$$

where $(\ell_1, \ldots, \ell_s)$ runs through all s-tuples of positive integers with $\ell_1 + \ldots + \ell_s \le k$, $s = 0, \ldots, k$, and $C_{\ell_1,\ldots,\ell_s}$ are some constants.

By the formula (27) and the form of the operator (23), we obtain that $R_{k,a,b,c}(x,y)$ are polynomials and $R_{k,i}(x,y)$ are rational functions. Actually, from the formulas (27) and (19), we easily obtain that the denominator of

$$\frac{d^k}{dX^k} \left(\frac{X}{Y}\right)^a X^{bt} Y^{(c+1)t}$$

divides $\left(\frac{\partial P}{\partial Y}(X,Y)\right)^{2k-1}$. Hence, we obtain that $R_{k,a,b.c}(X,Y)$ are polynomials. From the formula (25), we obtain that $R_{k,i}$ is a rational function with denominator divided by Y^{A-1}. Consequently, $\widetilde{R}_{k,i}$ are polynomials.

The result now follows from Lemma 2.5 and the formulas (23) and (24). □

2.4 Multiplicity Points on Some Curves

We recall that D_k, $k = 1, 2, \ldots$ are the differential operators defined by (23).

Lemma 2.8 *If $P(X,Y) \mid \Psi(X,Y)$ and $P(X,Y) \mid D_j\Psi(X,Y)$, $j = 1, \ldots, k-1$, then at least one of the following alternatives holds:*

- *either (x, y) is a root of order at least k of $\Psi(X,Y)$ on the algebraic curve (14).*
- *or $(x, y) \in \mathcal{M}_{sing}$.*

Proof If $D_j\Psi(X,Y)$ vanishes on the curve $P(X,Y) = 0$, then either

$$\frac{d^j}{dX^j}\Psi(x,y) = 0, \tag{28}$$

where, as before, $\frac{d^j}{dX^j}$ is j-th derivative on the algebraic curve (14) with the local parameter X, or

$$xy = 0, \tag{29}$$

or

$$\frac{\partial P}{\partial Y}(x,y) = 0, \tag{30}$$

on the curve (14).

If we have (28) for $j = 1, \ldots, k-1$ and also $\Psi(x,y) = 0$, then the pair (x,y) satisfies the first case of conditions of Lemma 2.8.

If we have (29) or (30) on the curve (14), then the pair (x, y) satisfies the second case of conditions of Lemma 2.8. □

3 Small Divisors of Integers

3.1 Smooth Numbers

As usual, we say that a positive integer is y-smooth if it is composed of prime numbers up to y. Then, we denote by $\psi(x,y)$ the number of y-smooth positive integers $n \le x$. Among a large variety of bounds and asymptotic formulas for $\psi(x, y)$ (see [13, 15, 22]), the most convenient bound for our applications is given by [22, Theorem 5.1].

Lemma 3.1 *There is an absolute constant c_0 such that for any fixed real-positive $x \ge y \ge 2$, we have*

$$\psi(x, y) \le c_0 e^{-u/2} x,$$

where

$$u = \frac{\log x}{\log y}.$$

3.2 Number of Small Divisors of Integers

For a real z and an integer n, we use $\tau_z(n)$ to denote the number of positive integer divisors $d \mid n$ with $d \le z$. We present a bound on $\tau_z(n)$ for small values of z (which we put in a slightly more general form than we need for our applications).

Lemma 3.2 *There is an absolute constant C_0 such that for any fixed real-positive $\varepsilon < 1$, there is $n(\varepsilon)$ such that if $n \ge n(\varepsilon)$ and $z \ge (\log n)^{2\log(1/\varepsilon)}$, then*

$$\tau_z(n) \le C_0 \varepsilon z.$$

Proof Let s be the number of all distinct prime divisors of n, and let $p_1, \dots, p_s$ be the first s primes. We note that

$$\tau_z(n) \le \psi(z, p_s). \tag{31}$$

By the prime number theorem, we have $n \ge p_1 \dots p_s = \exp(p_s + o(p_s))$, and thus

$$p_s \ll \log n \le z^{1/b}, \tag{32}$$

where $b = 2\log(1/\varepsilon)$. Combining Lemma 3.1 with (31) and (32), we see that

$$\tau_z(n) \le \psi(z, z^{1/b+o(1)}) \le c_0 e^{-b/2+o(1)} z = (c_0 + o(1)) e^{-b/2} z \le C_0 \varepsilon z$$

for any $C_0 > c_0$ (where c_0 is as in Lemma 3.1), provided that n and thus z are large enough. □

4 Proof of Theorem 1.2

4.1 Preliminary Estimates

We define the following parameters:

$$A = \left\lfloor \frac{t^{2/3}}{gh^{1/3}} \right\rfloor, \quad B = C = \left\lfloor h^{1/3}t^{1/3} \right\rfloor, \qquad D = \left\lfloor \frac{t^{2/3}}{4gh^{1/3}mn} \right\rfloor.$$

The exact values of A, B, C, and D play no role until the optimization step at the very end of the proof. However, it is important to note that their choice ensures (36) and (37).

If $P_i(x, y) = 0$ for at least one $i = 1, \dots, h$, then

$$D_k\Psi(x, y) = 0, \qquad (x, y) \in \bigcup_{i=1}^{h} \mathcal{F}_i, \tag{33}$$

with the operators (23), where the sets $\mathcal{F}_i$ are as in (13). The condition (33) is given by a system of linear homogeneous algebraic equations in the variables $\omega_{a,b,c}$. The number of equations can be calculated by means of Lemmas 2.6 and 2.7. To satisfy the condition (33) for some k, we have to make sure that the polynomials $\widetilde{R}_{k,i}(X, Y)$, $i = 1, \dots, h$, given by (26), vanish identically on the curve (14). The bi-degree of $\widetilde{R}_{k,i}(X, Y)$ is given by Lemma 2.7:

$$\deg_X \widetilde{R}_{k,i} \le A + 4km, \qquad \deg_Y \widetilde{R}_{k,i} \le A + 4kn.$$

The number of equations on the coefficients that guarantee the vanishing of the polynomial $\widetilde{R}_{k,i}(X, Y)$ on the curve (14) is given by Lemma 2.6 and is equal to $(\mu + \nu + 1)mn$, where μ, ν are as in Lemma 2.6 and

$$\mu \le A + 4km, \quad \nu \le A + 4kn.$$

Finally, the condition (33) for some k is given by $h(\mu + \nu + 1)mn \le mnh(2A + 4k(m + n))$ linear algebraic homogeneous equations. Consequently, the condition (33) for all $k = 0, \dots, D - 1$ is given by the system of

$$L = hmn \sum_{k=0}^{D-1} (4k(m + n) + 2A + 1)$$

linear algebraic homogeneous equations in variables $\omega_{a,b,c}$. Now it is easy to see that

$$L = h\left((2A + 1)Dmn + 2nm(m + n)D(D - 1)\right)$$

$$\le 2hADmn + 2hmn(m+n)D^2 = 2hmn(AD + (m+n)D^2).$$

4.2 Optimization of Parameters

The system has a nonzero solution if the number of equations is less than to the number of variables, in particular, if

$$2hmn(AD + (m+n)D^2) < ABC, \tag{34}$$

as we have ABC variables. It is easy to get an upper bound for the left-hand side of (34). For sufficiently large $t > c_0(m,n)$, where $c_0(m,n)$ is some constant depending only on m and n, we have

$$\begin{aligned} &2hmn(AD + (m+n)D^2) \\ &\quad < 2hmn\left(\frac{h^{-1/3}t^{2/3}}{g}\frac{h^{-1/3}t^{2/3}}{4mng} + (m+n)\frac{h^{-2/3}t^{4/3}}{16m^2n^2g^2}\right) \\ &\quad < \frac{3}{4}\frac{h^{1/3}t^{4/3}}{g^2}. \end{aligned} \tag{35}$$

Assuming that $c_0(m,n)$ is large enough, we obtain

$$ABC = \left\lfloor \frac{h^{-1/3}t^{2/3}}{g} \right\rfloor \lfloor h^{1/3}t^{1/3} \rfloor^2 > \frac{3}{4}\frac{h^{1/3}t^{4/3}}{g^2},$$

which together with (35) implies (34).

It is clear that

$$gAB \le t. \tag{36}$$

We also require that the degree of the polynomial $\Psi(x,y)$ should be less than p,

$$\deg \Psi(x,y) \le (B-1)t + Ct < p. \tag{37}$$

Actually, the inequality $(B-1)t + Ct < 2h^{1/3}t^{4/3} < p$ is satisfied because $t < \frac{1}{2}p^{3/4}h^{-1/4}$.

Finally, recalling Lemmas 2.2, 2.3 and 2.4, and also the irreducibility of the polynomial $P(x,y)$, we see that $P_k(X,Y)$ and $\Psi(X,Y)$ are co-prime. Hence, by Lemmas 2.1 and 2.8 and the inequality (15), we obtain that N_h satisfies the inequality

$$\begin{aligned} N_h &\le \#\mathcal{M}_{sing} + (m+n)\frac{(B+C-1)t}{D} \\ &< (m+n)^2 + (m+n)\frac{2h^{1/3}t^{4/3}}{\lfloor h^{-1/3}t^{2/3}/(4mng)\rfloor} \\ &< 12mn(m+n)gh^{2/3}t^{2/3} \end{aligned}$$

for sufficiently large $t > c_0(m,n)$, which concludes the proof.

5 Proof of Theorem 1.6

5.1 Outline of the Proof

Before giving technical details, we first outline the sequence of the following steps:

- We consider the set $\mathcal{R} = \mathcal{M}_p \setminus \mathcal{C}_p$ and show that if it is large then by Lemma 3.2 there is a large set $\mathcal{L} \subseteq \mathcal{R}$ elements of large orders.
- Each element $x \in \mathcal{L}$ has an orbit of size at least $t(x)/2$, which is also in $\mathcal{R}$.
- Using Conjecture 1.3, we estimate the size of intersections of these orbits for distinct elements $x_1, x_2 \in \mathcal{L}$.
- We conclude that all intersections together are small, and so to fit all orbits in $\mathcal{R}$, the size of $\mathcal{R}$ must be even larger than we have initially assumed.

5.2 Formal Argument

We always assume that p is large enough. Define the mapping

$$\mathcal{T}_0\,(x,y,z) \mapsto (x,z,3xz-y)\,,$$

where $\mathcal{T}_0 = \Pi_{1,3,2} \circ \mathcal{R}_2$ is the composition of the permutations

$$\Pi_{1,3,2} = (x,y,z) \mapsto (x,z,y)$$

and the involution

$$\mathcal{R}_2 : (x,y,z) \mapsto (x,3xz-y,z)$$

as in the above.

Therefore, the orbit $\Gamma(x,y,z)$ of (x,y,z) under the above group of transformations Γ contains, in particular, the triples (x,u_n,u_{n+1}), $n = 1,2,\ldots$, where the sequence u_n satisfies a binary linear recurrence relation

$$u_{n+2} = 3xu_{n+1} - u_n, \qquad n = 1, 2, \dots, \tag{38}$$

with the initial values, $u_1 = y$, $u_2 = z$. This also means that $\Gamma(x, y, z)$ contains all triples obtained by the permutations of the elements in (x, u_n, u_{n+1}).

Let $\xi, \xi^{-1} \in \mathbb{F}_{p^2}^*$ be the roots of the characteristic polynomial $Z^2 - 3xZ + 1$ of the recurrence relation (38). In particular, $3x = \xi + \xi^{-1}$. Then, it is easy to see that unless $(x, y, z) = (0, 0, 0)$, which we eliminate from the consideration, the sequence u_n is periodic with period $t(x)$ which is the order of ξ in $\mathbb{F}_{p^2}^*$.

Let B be a fixed positive number to be chosen later. We denote

$$M_0 = (\log p)^B \qquad \text{and} \qquad M_1 = M_0^{1/4}/3 = (\log p)^{B/4}/3.$$

Assume that the remaining set of nodes $\mathcal{R} = \mathcal{M}_p \setminus \mathcal{C}_p$ is of size $\#\mathcal{R} > M_0$. Note that if $(x, y, z) \in \mathcal{R}$, then also $(y, x, z) \in \mathcal{R}$, and for any x, y, there are at most two values of z such that $(x, y, z) \in \mathcal{R}$. Therefore, there are more than $(M_0/2)^{1/2}$ elements $x \in \mathbb{F}_p^*$ with $(x, y, z) \in \mathcal{R}$ for some $y, z \in \mathbb{F}_p$.

Since there are obviously at most $T(T + 1)/2$ elements $\xi \in \mathbb{F}_{p^2}^*$ of order at most T, we conclude that there is a triple $(x^*, y^*, z^*) \in \mathcal{R}$ with

$$t(x^*) > \sqrt{(M_0/2)^{1/2}} > 2M_1, \tag{39}$$

where $t(x^*)$ is the period of the sequence u_n which is defined as in (38) with respect to (x^*, y^*, z^*).

Then, the orbit $\Gamma(x^*, y^*, z^*)$ of this triple has at least $2M_1$ elements. Let M be the cardinality of the set $\mathcal{X}$ of projections along the first components of all triples $(x, y, z) \in \Gamma(x^*, y^*, z^*)$. Since the orbits are closed under the permutation of coordinates and permutations of the triples

$$(x^*, u_n, u_{n+1}), \qquad n = 1, \dots, t(x^*),$$

where as above the sequence u_n is defined as in (38) with respect to (x^*, y^*, z^*) and $t(x^*)$ is its period, produce the same projection no more than twice, we obtain

$$M \geq \frac{1}{2} t(x^*). \tag{40}$$

Recalling (39), we obtain

$$M > M_1 = (\log p)^{B/4}/3. \tag{41}$$

Using that $(x, y, z) \notin \mathcal{M}_p$, we notice that by the bound (3),

$$M = p^{o(1)}. \tag{42}$$

For $t \mid p^2 - 1$, we denote $g(t)$ the number of $x \in \mathcal{X}$ for which the period of the sequence u_n defined as in (38) satisfies $t(x) = t$. Observe that

$$\sum_{t \mid p^2-1} g(t) = M.$$

The same argument as used in the bound (40) implies that

$$g(t) = 0 \quad \text{for} \quad t > 2M. \tag{43}$$

We apply Lemma 3.2 with

$$\varepsilon = \frac{1}{40AC_0}, \tag{44}$$

where A is a bound from Conjecture 1.3 and C_0 is as in Lemma 3.1. Take

$$B = 16 \log(1/\varepsilon) + 1. \tag{45}$$

Since $g(t) < t$ for any t and also since due to (41) we have

$$4\sqrt{AM} > (\log p)^{B/8} \ge (\log(p^2 - 1))^{2\log(1/\varepsilon)},$$

by Lemma 3.2,

$$\begin{aligned} \sum_{\substack{t \le 4\sqrt{AM} \\ t \mid p^2-1}} g(t) < \sum_{\substack{t \le 4\sqrt{AM} \\ t \mid p^2-1}} t &\le 4\sqrt{AM}\tau_{4\sqrt{AM}}(p^2-1) \\ &\le C_0\varepsilon(4\sqrt{AM})^2 = 0.4M. \end{aligned}$$

Hence, we conclude that

$$\sum_{\substack{t > 4\sqrt{AM} \\ t \mid p^2-1}} g(t) \ge 0.6M.$$

Let $\mathcal{L}$ be the set of $x \in \mathcal{X}$ with $t(x) > 4\sqrt{AM}$. We have shown that

$$\#\mathcal{L} \ge 0.6M. \tag{46}$$

For each $x \in \mathcal{L}$, we fix some $y, z \in \mathbb{F}_p$ such $(x, y, z) \in \Gamma(x^*, y^*, z^*)$ and again consider the sequence u_n, $n = 1, 2, \ldots$, given by (38) having the period $t(x) = t_0$, so we consider the set

$$\mathcal{Z}(x) = \{u_n \ : \ n = 1, \dots, t_0\}.$$

Let $\mathcal{H}_x$ be the subgroup of $\mathbb{F}_{p^2}^*$ of order $t(x)$ and $\xi(x)$ satisfy the equation $3x = \xi(x) + \xi(x)^{-1}$. One can easily check, using an explicit expression for binary recurrence sequences via the roots of the characteristic polynomial, that

$$\mathcal{Z}(x) = \left\{\alpha(x)u + \frac{r(x)}{\alpha(x)u} \ : \ u \in \mathcal{H}_x\right\},$$

where

$$r(x) = \frac{(\xi(x)^2 + 1)^2}{9(\xi(x)^2 - 1)^2},$$

and $\alpha(x) \in \mathbb{F}_{p^2}^*$. If for some r an element $\xi = \xi_0$ satisfies the equation

$$r = \frac{(\xi^2 + 1)^2}{9(\xi^2 - 1)^2},$$

then other solutions are $-\xi_0$, $1/\xi_0$, $-1/\xi_0$. Moreover, $3x = \xi + \xi^{-1}$ can take, for a fixed r, at most two values whose sum is 0. Since every value is taken at most twice among the elements of the sequence u_n, $n = 1, \dots, t(x)$, we have

$$\#\mathcal{Z}(x) \geq \frac{1}{2}t(x) > 2\sqrt{AM}. \tag{47}$$

Now we construct a set $\mathcal{L}^* \subseteq \mathcal{L}$. If $x, x^* \in \mathcal{L}$ and $x + x^* = 0$, then we put one of the elements x, x^* in $\mathcal{L}^*$. If $x \in \mathcal{L}$ and $-x \notin \mathcal{L}$, then we set $x \in \mathcal{L}^*$. Due to (46), we get

$$\#\mathcal{L}^* \geq 0.3M. \tag{48}$$

Moreover, for any distinct $x, x^* \in \mathcal{L}^*$, we have $x+x^* \neq 0$ and, hence, $r(x) \neq r(x^*)$.

We claim that under Conjecture 1.3 for any distinct $x, x^* \in \mathcal{L}^*$, the inequality

$$\#\left(\mathcal{Z}(x) \bigcap \mathcal{Z}(x^*)\right) \leq 2A \tag{49}$$

holds.

Indeed, take distinct elements $x, x^* \in \mathcal{L}^*$. By $\mathcal{G}$, we denote the subgroup of $\mathbb{F}_{p^2}^*$ generated by $\mathcal{H}_x$ and $\mathcal{H}_{x^*}$. Notice that due to (42) and (43), we have

$$\#\mathcal{G} = p^{o(1)}. \tag{50}$$

Next, $\#(Z(x) \cap Z(x^*)$ is the number of solutions to the equation

$$\alpha(x)u + \frac{r(x)}{\alpha(x)u} = \alpha(x^*)v + \frac{r(x^*)}{\alpha(x^*)v}, \qquad (u, v) \in \mathcal{H}_x \times \mathcal{H}_{x^*},$$

as in the above or, equivalently,

$$P_{x,x^*}(u, v) = 0, \qquad (u, v) \in \mathcal{H}_x \times \mathcal{H}_{x^*},$$

where

$$P_{x,x^*}(X, Y) = \alpha(x)^2\alpha(x^*)X^2Y - \alpha(x)\alpha(x^*)^2XY^2 \\ - \alpha(x)r(x^*)X + \alpha(x^*)r(x)Y.$$

The number of solutions to the last equation in $(u, v) \in \mathcal{H}_x \times \mathcal{H}_{x^*}$ does not exceed the number of solutions in $(u, v) \in \mathcal{G}^2$. Let $Z = X/Y$. Then, the equation is reduced to

$$\frac{\alpha(x)^2\alpha(x^*)Z - \alpha(x)\alpha(x^*)^2}{\alpha(x)r(x^*)Z - \alpha(x^*)r(x)} = U, \tag{51}$$

where $U = Y^{-2}Z^{-1}$.

Now we are in position to use Conjecture 1.3. The conditions (11) on the coefficients of linear functions in the numerator and in the denominator of the fraction in (51) are satisfied since $\alpha(x) \neq 0$, $\alpha(x^*) \neq 0$, and $r(x) \neq r(x^*)$.

Also, for large p we have $\#\mathcal{G} \leq p^{\varepsilon_0}$ due to (50). By Conjecture 1.3, Eq. (51) has at most A solutions in Z, Y. For each solution, there are at most two possible values of Y. Fixing Y, we determine X. So, the inequality (49) holds.

Denote

$$h = [\sqrt{M/A}] + 1.$$

Due to (41) and (48), we have $\#\mathcal{L}^* \geq h$ provided that p is large enough. We choose h elements $x_1, \ldots, x_h$ from $\mathcal{L}^*$. It follows from (49) that for $j = 1, \ldots, h$ we have

$$\sum_{i=1}^{j-1} \# \left(\mathcal{Z}\left(x_j\right) \bigcap \mathcal{Z}\left(x_i\right) \right) \leq 2(j-1)A,$$

which implies by (47)

$$\# \left(\mathcal{Z}\left(x_j\right) \setminus \bigcup_{i=1}^{j-1} \mathcal{Z}\left(x_i\right) \right) \geq 2\sqrt{AM} - 2(j-1)A.$$

Observe that

$$\#\left(\bigcup_{j=1}^{h}\mathcal{Z}\left(x_j\right)\right)=\sum_{j=1}^{h}\#\left(\mathcal{Z}\left(x_j\right)\setminus\bigcup_{i=1}^{j-1}\mathcal{Z}\left(x_i\right)\right).$$

Hence,

$$\begin{aligned}\#\left(\bigcup_{j=1}^{h}\mathcal{Z}\left(x_j\right)\right) &> 2\sqrt{AM}h-(h-1)hA\\ &=(2\sqrt{AM}-(h-1)A)h\\ &>(2\sqrt{AM}-\sqrt{AM})\sqrt{M/A}>M,\end{aligned}$$

but this inequality contradicts the definition of M. Together with the choice of B given by (44) and (45), this concludes the proof.

6 Comments

Let $P(n)$ be the largest primitive prime divisor of 2^n-1, that is, the largest prime which divides 2^n-1, but does not divide any of the numbers 2^d-1 for $1\le d<n$. Note that $P(n)\equiv 1 \pmod{n}$. By a striking result of Stewart [21, Theorem 1.1], we have

$$P(n)\ge n\exp\left(\frac{\log n}{104\log\log n}\right),$$

provided that n is large enough. It is also natural to assume that $\log P(n)/\log n\to\infty$ for $n\to\infty$. However, for us a weaker assumption is sufficient. Namely, assume that

$$\limsup\frac{\log P(24m)}{\log m}=\infty.$$

We then take $n=24m$, $m\in\mathbb{N}$, and $p=P(n)$ such that $n=p^{o(1)}$. Then, $p\equiv 1 \pmod{24}$. Since 2 is a quadratic residue modulo p, we can take $\xi\in\mathbb{F}_p$ such that $\xi^2=2$. We consider a group $\mathcal{G}$ generated by ξ. Note that $\#\mathcal{G}=2n=p^{o(1)}$ as $n\to\infty$. The group $\mathcal{G}$ contains an element ζ_4 of order 4 and an element ζ_6 of order 6. It is easy to check that

$$((\pm\zeta_4\pm 1)/\xi)^8=1.$$

Thus,

$$(\pm\zeta_4 \pm 1)^{2n} = \xi^{6n} = 1.$$

Hence, $\pm\zeta_4 \pm 1 \in \mathcal{G}$. Also,

$$(\pm\zeta_6 - 1)^3 = 1.$$

Hence, similarly $\pm\zeta_6 - 1 \in \mathcal{G}$. Consider a set $\mathcal{D}$ consisting of 9 elements

$$\mathcal{D} = \{(p - 1/2), 1, -2, \zeta_4, -\zeta_4, \zeta_4 - 1, -\zeta_4 - 1, \zeta_6 - 1, -\zeta_6 - 1\}.$$

Clearly, $x \in \mathcal{G}, x + 1 \in \mathcal{G}$ for any $x \in \mathcal{D}$. This shows that probably A in Conjecture 1.3 should be at least 9.

We also observe that in Conjecture 1.3 the value of ε_0 cannot be taken greater than $1/2$.

Indeed, suppose that p is a prime and $p - 1$ has a divisor $t = p^{\varepsilon_0 + o(1)}$, as $p \to \infty$ with a fixed $\varepsilon_0 > 1/2$ (the infinitude of such primes follows instantly from [10, Theorem 7]).

Let us fix any $\alpha_{1,1}, \alpha_{1,2}, \alpha_{2,1}, \alpha_{2,2} \in \mathbb{F}_p$. Clearly, the Eq. (12) has $N = p + O(1)$ of solutions $(u, v) \in \left(\mathbb{F}_p^*\right)^2$. Let $\mathcal{G} \subseteq \mathbb{F}_p^*$ be a subgroup of order t. Since $\mathbb{F}_p^*$ is the union of $(p - 1)/t$ cosets $a\mathcal{G}$ of $\mathcal{G}$, the direct product $\mathbb{F}_p^* \times \mathbb{F}_p^*$ is the union of $(p - 1)^2/t^2$ products of cosets of $\mathcal{G}$. By the Dirichlet principle is that there is at least one product $a\mathcal{G} \times b\mathcal{G}$ such that the number of solutions $(u, v) \in a\mathcal{G} \times b\mathcal{G}$ (with some $a, b \in \mathbb{F}_p^*$) is not less than

$$\frac{N}{(p-1)^2/t^2} \geq (1 + o(1))t^2/p \geq p^{2\varepsilon_0 - 1 + o(1)}$$

and hence is not bounded as $p \to \infty$. Changing the variables $\widetilde{u} = a^{-1}u$, $\widetilde{v} = b^{-1}v$ in (12) we obtain another equation of the same type

$$\frac{\alpha_{1,1}ab^{-1}\widetilde{u} - \alpha_{1,2}b^{-1}}{\alpha_{2,1}a\widetilde{u} - \alpha_{2,2}} = \widetilde{v}$$

with an unbounded number of solutions $(\widetilde{u}, \widetilde{v}) \in \mathcal{G}^2$.

Finally, we note that using [5, Theorem 1.2] one concludes that Conjecture 1.3 holds (in much stronger and general form) for a sequence of primes of relative density 1. However, this does not give any new results for the sets $\mathcal{M}_p$ because, as we mentioned, Bourgain, Gamburd, and Sarnak [2, Theorem 2] have already shown that Conjecture 1.1 holds for an overwhelming majority of primes $p \leq X$ as $X \to \infty$.

Acknowledgments The authors are very grateful to William Chen for clarifying discussion concerning his work [6] and to the referee for the careful reading of the manuscript and valuable remarks.

This work was supported by the Australian Research Council Grants DP180100201 and DP200100355 (Shparlinski) and by the Russian Science Foundation Grant 19-11-00001 (Vyugin).

References

1. Baragar, A.: The Markoff equation and equations of Hurwitz. Ph.D. Thesis, Brown University, 1991
2. Bourgain, J., Gamburd, A., Sarnak, P.: Markoff triples and strong approximation. C. R. Acad. Sci. Paris, Ser. I **354**, 131–135 (2016)
3. Bourgain, J., Gamburd, A., Sarnak, P.: Markoff surfaces and strong approximation, I. Preprint (2016). http://arxiv.org/abs/1607.01530
4. Cerbu, A., Gunther, E., Magee, M., Peilen, L.: The cycle structure of a Markoff automorphism over finite fields. J. Number Theory **211**, 1–27 (2020)
5. Chang, M.-C., Kerr, B., Shparlinski, I., Zannier, U.: Elements of large order on varieties over prime finite fields. J. Théor. Nombres Bordeaux **26**, 579–593 (2014)
6. Chen, W.: Nonabelian level structures, Nielsen equivalence, and Markoff triples. Preprint (2020). http://arxiv.org/abs/2011.12940
7. Corvaja, P., Zannier, U.: Greatest common divisors of $u - 1$, $v - 1$ in positive characteristic and rational points on curves over finite fields. J. Eur. Math. Soc. **15**, 1927–1942 (2013)
8. de Courcy-Ireland, M., Lee, S.: Experiments with the Markoff surface. Experimental Math. (2018), (to appear)
9. de Courcy-Ireland, M., Magee, M.: Kesten-McKay law for the Markoff surface mod$p$. Annales Henri Lebesgue (to appear)
10. Ford, K.: The distribution of integers with a divisor in a given interval. Annals Math. **168**, 367–433 (2008)
11. Gamburd, A., Magee, M. and Ronan, R.: An asymptotic formula for integer points on Markoff-Hurwitz varieties. Annals Math., **190**, 751–809 (2019)
12. Garcia, A., Voloch, J.F.: Fermat curves over finite fields. J. Number Theory **30**, 345–356 (1988)
13. Granville, A.: Smooth numbers: Computational number theory and beyond. Algorithmic Number Theory: Lattices, Number Fields, Curves, and Cryptography, pp. 267–322. Cambridge University Press (2008)
14. Heath-Brown, D.R., Konyagin, S.V.: New bounds for Gauss sums derived from kth powers, and for Heilbronn's exponential sum. Quart. J. Math. **51**, 221–235 (2000)
15. Hildebrand, A., Tenenbaum, G.: Integers without large prime factors. J. Théorie des Nombres de Bordeaux **5**, 411–484 (1993)
16. Konyagin, S.V., Makarychev, S.V., Shparlinski, I.E., Vyugin, I.V.: On the structure of graphs of Markoff triples. Quart. J. Math. **71**, 637–648 (2020)
17. Makarychev, S.V., Vyugin, I.V.: Solutions of polynomial equations in subgroups of $\mathbb{F}_p$. Arnold Math J. **5**, 105–121 (2019)
18. Markoff, A.: Sur les formes quadratiques binaires indéfinies. Math. Ann. **15**, 381–409 (1879)
19. Markoff, A.: Sur les formes quadratiques binaires indéfinies. Math. Ann. **17**, 379–399 (1880)
20. Shkredov, I.D., Vyugin, I.V.: On additive shifts of multiplicative subgroups. Mat. Sb. **203**, 81–100 (2012) (in Russian)
21. Stewart, C.L.: On divisors of Lucas and Lehmer numbers. Acta Math. **211**, 291–314 (2013)
22. Tenenbaum, G.: Introduction to analytic and probabilistic number theory. Grad. Studies Math., vol. 163. Amer. Math. Soc. (2015)

Exponential Sums, Twisted Multiplicativity, and Moments

E. Kowalski and K. Soundararajan

Dedicated to the memory of Jean Bourgain

Abstract We study averages over squarefree moduli of the size of exponential sums with polynomial phases. We prove upper bounds on various moments of such sums, and obtain evidence of un-correlation of exponential sums associated to different suitably unrelated and generic polynomials. The proofs combine analytic arguments with the algebraic interpretation of exponential sums and their monodromy groups.

1 Introduction

Some of Jean Bourgain's many interactions with number theory involved exponential sums in different ways. Among these, one can mention his ground-breaking use of ideas from the circle method to solve Bellow's problems concerning pointwise ergodic theorems at times $f(n)$, where $f \in \mathbf{Z}[X]$ is a polynomial (see, in particular, [5–7]) or its combination with bilinear forms in joint works with A. Kontorovich to study some aspects of the sieve in orbits beyond a simple appeal to expansion and spectral gaps (see, for instance, [8]). We respectfully dedicate this paper to his memory.

E. Kowalski (✉)
ETH Zürich—D-MATH, Zürich, Switzerland
e-mail: kowalski@math.ethz.ch

K. Soundararajan
Department of Mathematics, Stanford University, Stanford, CA, USA
e-mail: ksound@stanford.edu

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_13

1.1 Exponential Sums with Polynomials

This paper is primarily concerned with exponential sums with polynomial phases. Let $f \in \mathbf{Z}[X]$ be a non-constant polynomial with degree d. For $q \geqslant 1$ squarefree and a coprime to q, we define

$$W(a;q) = W_f(a;q) = \frac{1}{\sqrt{q}} \sum_{x \,(\mathrm{mod}\, q)} e\Big(\frac{af(x)}{q}\Big),$$

where the sum is over residue classes modulo q. For simplicity, we restrict attention to squarefree q and set $W(a;q) = 0$ if q is not squarefree or if $(a,q) > 1$.

An application of the Chinese Remainder Theorem shows that the exponential sums $W(a;q)$ satisfy the following "twisted multiplicativity": if $(q_1,q_2) = 1$, then

$$W(a;q_1q_2) = W(a\bar{q_1};q_2)W(a\bar{q_2};q_1),$$

where $q_1\bar{q_1} \equiv 1 \mod q_2$ and $q_2\bar{q_2} \equiv 1 \mod q_1$. Apart from finitely many primes, the Weil bound gives $|W(a;p)| \leqslant (d-1)$, so that $|W(a;q)| \ll (d-1)^{\omega(q)}$ where $\omega(q)$ denotes the number of (distinct) prime factors of q. It follows that

$$\sum_{q \leqslant x} |W(a;q)| \ll \sum_{q \leqslant x} (d-1)^{\omega(q)} \ll x(\log x)^{d-2},$$

and we seek an improvement over this "trivial" bound, as well as bounds for related mean values such as $\sum_{q \leqslant x} |W(a;q)|^2$. The possibility of obtaining such improvements was first recognized by Hooley and explored further in the work of Fouvry and Michel [13].

One of our main theorems gives a refinement of these earlier results. Given a field K, we say that a polynomial $f \in K[X]$ is *decomposable* if there are polynomials g and h in $K[X]$, both with degree $\geqslant 2$, such that $f = g \circ h$. If f cannot be expressed as such a composition, we call f *indecomposable*.

Theorem 1.1 *Let $f \in \mathbf{Q}[X]$ be an indecomposable polynomial with* $\deg(f) = d \geqslant 3$.

(1) *For any $a \geqslant 1$,*

$$\sum_{q \leqslant x} |W(a;q)|^2 \ll x(\log\log x)^{(d-1)^2}.$$

(2) *There exists $\gamma > 0$, depending only on d, such that for any $a \geqslant 1$,*

$$\sum_{q \leqslant x} |W(a;q)| \ll \frac{x}{(\log x)^{\gamma}}.$$

Remark 1.2 The implied constants above (and in what follows) are allowed to depend on f. Throughout, we ignore linear polynomials where $W(a;q)$ is usually 0 and quadratic polynomials where $|W(a;q)|$ is usually 1 (since these are quadratic Gauss sums).

One can compute effectively a possible value of the constant γ (see Remark 5.4).

The possibility of obtaining non-trivial bounds for

$$\sum_{q \leqslant x} |W(a;q)|$$

(with f allowed to be a rational function) was first pointed out by Hooley in [18] in the case of Kloosterman sums. Introducing ideas from algebraic geometry (notably from the work of Katz [22]), Fouvry and Michel [13] refined and extended Hooley's work to more general exponential sums. Under a hypothesis that the polynomial f is generic (in a sense to be made precise below; see [13, H.1, H.2, H.3, H.3']; note that in fact Fouvry and Michel consider rational functions and not only polynomials), Fouvry and Michel proved in [13, Th. 1.5] that

$$\sum_{q \leqslant x} |W(a;q)| \ll x(\log\log x)^{k_f - 1} \tag{1}$$

for some explicit integer $k_f \geqslant 1$. Theorem 1.1 refines this in two ways. Firstly, it applies to a larger class of polynomials f, with the much simpler criterion of being indecomposable. For instance, if the degree $d \geqslant 3$ of f is prime, then f is automatically indecomposable, so that our result applies, but any polynomial f such that f' has a multiple root, say $f = X^3 g$ for some g of degree $d-3$, fails to satisfy the condition H.1 of [13], since the zeros of f' are not simple. On the other hand, it is elementary to check that a polynomial f satisfying the conditions H.1, H.2, H.3 (or H.3') is indecomposable (see Lemma 6.1 below, combined with Remark 1.10 for the terminology). Secondly, part (2) of the theorem improves on (1) qualitatively by showing that the average of $|W(a;q)|$ over $q \leqslant x$ tends to 0, which does not follow from the method of Fouvry and Michel.

The proof of the second part of Theorem 1.1 relies on the following result, which may be of independent interest.

Theorem 1.3 *Let $f \in \mathbf{Z}[X]$ of degree $d \geqslant 3$. Then, one of the following two possibilities holds:*

(1) *The limit*

$$\lim_{p \to +\infty} \frac{1}{p} \sum_{a \in \mathbf{F}_p^\times} |W(a;p)|^4 \qquad \textit{exists and equals } 2.$$

(2) *There exists $\delta > 0$ (depending only on d) and a subset of primes with positive density $\geqslant \delta$ on which*

$$\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times}|W(a;p)|^4\geqslant 3+O(p^{-1/2}).$$

For a generic (again in a sense to be made precise later) polynomial f, the first case of the theorem holds.

Remark 1.4

(1) The work of Katz [21] contains material from which it is likely that one can deduce Theorem 1.3. However, in view of the different focus and the generality of [21], our independent and slightly more elementary proof seems worth including.

(2) Using the method of [13, § 4], one can show that

$$\sum_{q\leqslant x}|W(a;q)|\gg\frac{x}{\log x}$$

(or even a slightly better lower bound), and it is a natural question to ask whether there exists a constant $\delta>0$ such that

$$\frac{x}{(\log x)^{\delta+\varepsilon}}\ll\sum_{q\leqslant x}|W(a;q)|\ll\frac{x}{(\log x)^{\delta-\varepsilon}}\tag{2}$$

for any $\varepsilon>0$. This is an open problem; in Remark 3.6, we will mention a potential candidate value of δ, at least for the upper bound for generic polynomials.

(3) It might be possible to extend Theorem 1.1 to certain rational functions, but some additional work is required (e.g., to properly understand the analogue of indecomposability for rational functions and to extend [22, Lemma 7.7.5]).

1.2 Sums of Twisted Multiplicative Functions

A key feature of the exponential sums considered above is their twisted multiplicativity. In this section, we formulate, following Hooley [18], Fouvry and Michel [13], and our own recent paper [26], a general result on bounding averages of twisted multiplicative functions.

Suppose we are given a function V that associates to each prime p and each reduced residue class $a \pmod p$ a complex number $V(a;p)$. Extend this to a function $V(a;q)$ where q is squarefree and $a \pmod q$ is a reduced residue class by "twisted multiplicativity": that is, if $q=q_1q_2$ with $(q_1,q_2)=1$, then

$$V(a;q_1q_2)=V(a\bar{q}_1;q_2)V(a\bar{q}_2;q_1).\tag{3}$$

Set $V(a;q)=0$ if q is not squarefree, or if a is not coprime to q. For each prime p, let $G(p)\geqslant g(p)\geqslant 0$ be such that

$$\max_{(a,p)=1}|V(a;p)|\leqslant G(p),\qquad \text{and}\qquad \frac{1}{p}\sum_{(a,p)=1}|V(a;p)|\leqslant g(p).\tag{4}$$

Extend g and G to all squarefree integers using multiplicativity, so that (4) remains valid for all q.

The question then is to obtain, under suitable conditions, a bound for

$$\sum_{q\leqslant x}|V(a;q)|$$

that improves upon the trivial bound

$$\sum_{q\leqslant x}|V(a;q)|\leqslant\sum_{q\leqslant x}G(q).$$

Theorem 1.5 *Let $M>0$ be such that $G(p)\leqslant M$ for all primes p. Then, for any fixed integer $a\geqslant 1$ and for all large x, we have*

$$\sum_{q\leqslant x}|V(a;q)|\ll\frac{x}{\log x}\prod_{p\leqslant x}\Big(1+\frac{g(p)}{p}\Big)(\log\log x)^M,$$

where the implied constant may depend on M.

Remark 1.6

(1) The twisted multiplicativity (3) is naturally connected to the Chinese Remainder Theorem via the Fourier transform. Suppose that for each prime p and any residue class $a\pmod p$, we are given a complex number $v(a;p)$. We extend v to squarefree moduli q and any residue class $a\pmod q$ by means of the Chinese Remainder Theorem: that is, we set

$$v(a;q)=\prod_{p|q}v(a;p).$$

Consider now the Fourier transform of v:

$$V(a;q)=\sum_{b\,(\mathrm{mod}\,q)}v(b;q)e(ab/q).$$

Then, $V(a;q)$ satisfies the twisted multiplicative relation (3).

If $v(a;p)$ corresponds to a probability measure (thus all $v(a;p)$ are non-negative and $\sum_a v(a;p)=1$), then $|V(a;p)|\leqslant 1$ for all $a\pmod p$, so that we may use $G(p)=1$. Bounding the L^1-norm by the L^2-norm, we may take

$$g(p) = \Big(\frac{1}{p}\sum_{a=1}^{p-1} |V(a;p)|^2\Big)^{\frac{1}{2}} = \Big(\sum_{a=1}^{p} |v(a;p)|^2 - \frac{1}{p}\Big)^{\frac{1}{2}},$$

upon using Parseval.

(2) In the applications to equidistribution in [26], the functions that occur are Weyl sums of the form

$$V(a;q) = \frac{1}{\varrho(q)} \sum_{x \in A_q} e\Big(\frac{a \cdot x}{q}\Big)$$

for some $h \in \mathbf{Z}^n \setminus \{0\}$, where $A_q \subset (\mathbf{Z}/q\mathbf{Z})^n$ are non-empty sets "defined by the Chinese Remainder Theorem", and $\varrho(q) = |A_q|$.

1.3 Non-correlation of Exponential Sums for Different Polynomials

Our next results are attempts to establish that the exponential sums associated to two different polynomials f and g are uncorrelated. Here we use the notation $W_f(a;q)$ instead of $W(a;q)$ to keep track of the dependency on the polynomial. The results here will depend on polynomials being suitably generic (as in the work of Fouvry and Michel [13] mentioned earlier), and we begin by making this notion precise.

Definition 1.7 (Morse Polynomial) Let K be a field. A polynomial $f \in K[X]$ of degree $d \geqslant 1$ is called *Morse* if it has no repeated roots, its derivative f' is squarefree of degree $d-1$, and the values of f at the zeros of f' (in an algebraic closure of K) are distinct.

Remark 1.8 The values of f at the zeros of the derivative of f are known as *critical values* of f. Note that when f' is even, the critical values appear in pairs $a + f(0)$, $-a + f(0)$ where a is a critical value of $f(x) - f(0)$.

If d is smaller than the characteristic of K, then the condition that $\deg(f') = d - 1$ is automatically fulfilled.

If f is a Morse polynomial, then 0 is not a critical value of f (since there would then be a double zero).

We recall that in an abelian group A, a subset $S \subset A$ is called *Sidon* if the equation $a + b = c + d$ with $(a,b,c,d) \in S^4$ has only the obvious solutions where $a \in \{c,d\}$.

We will say that $S \subset A$ is a *symmetric Sidon set* if there exists $\alpha \in A$ such that $S = \alpha - S$, and the equation $a + b = c + d$ with $(a,b,c,d) \in S^4$ has only the obvious solutions where $a \in \{c,d\}$ or $b = \alpha - a$.

We require one last item of terminology. For any field K, two polynomials f and g in $K[X]$ are *linearly equivalent over* K if there exist a, b, c, d in K, with a

and c non-zero, such that

$$g(X) = af(cX + d) + b.$$

Note that the sets V_f and V_g of critical values of f and g are then related by

$$V_g = aV_f + d.$$

In particular, if V_f is a Sidon set (resp. a symmetric Sidon set), then so is V_g.

Definition 1.9 (Sidon–Morse Polynomial) Let K be a field. A polynomial $f \in K[X]$ of degree $d \geqslant 2$ is called *Sidon–Morse* if it is Morse and one of the following holds:

(1) The set of critical values of f is a Sidon set in the additive group of K.
(2) The polynomial f is linearly equivalent to an odd polynomial g, and the set of critical values of g is a symmetric Sidon set in K.

For a polynomial $f \in A[X]$, with A an integral domain, we say that f is Morse (or Sidon–Morse) if the definition is satisfied for the field of fractions of A.

Remark 1.10 (1) To distinguish between the two alternatives above, we will say that f is a *symmetric Sidon–Morse polynomial* in the second case.

(2) It would seem to be more natural to define a symmetric Sidon polynomial to be one where the set of critical values of f is a symmetric Sidon set. This condition is implied by our definition, and it may in fact be that this is an equivalent definition (at least over $\mathbf{Q}$), but we do not know if this is the case. We will see how, at some crucial point in the proof of Theorem 6.3 below, this alternative definition is not sufficient to proceed.

(3) Any polynomial f of degree $d \geqslant 3$ in $\mathbf{Z}[X]$ whose derivative has Galois group $\mathfrak{S}_{d-1}$ is a (non-symmetric) Sidon–Morse polynomial over $\mathbf{Q}$ (see [22, proof of Th. 7.10.6]). It is then a Sidon–Morse polynomial over $\mathbf{F}_p$ for all but finitely many p. In particular, a "generic" polynomial in $\mathbf{Z}[X]$, in a natural sense, is Sidon–Morse over $\mathbf{Q}$.

(4) The genericity conditions H.1, H.2, H.3 for $f \in \mathbf{Z}[X]$ used by Fouvry and Michel are equivalent to asking that f is a Sidon–Morse polynomial; if f satisfies H.1, H.2, H.3', then it is a symmetric Sidon–Morse polynomial (but the converse is not always true, since H.3' requires f to be odd, not merely linearly equivalent to an odd polynomial).

Theorem 1.11

(1) *Let f and g be polynomials in $\mathbf{Z}[X]$ with degree $d_f \geqslant 3$ and d_g, respectively. Assume that f is Sidon–Morse over $\mathbf{Q}$ and that $d_f > d_g$. Then,*

$$\sum_{q \leqslant x} |W_f(a; q)\overline{W_g(a; q)}|^2 \ll x(\log\log x)^A$$

for some A depending only on d_f and d_g, where the implied constant depends on f and g.

(2) *Let $m \geqslant 1$ be an integer and let $f_1, \ldots, f_m$ be polynomials of degrees $d_i = \deg(f_i) \geqslant 3$. Assume that all f_i are Sidon–Morse polynomials over $\mathbf{Q}$ and moreover that, for any $i \neq j$, the polynomials f_i and f_j are not linearly equivalent over $\bar{\mathbf{Q}}$.*

Let s be the number of polynomials f_i such that f_i is a symmetric Sidon–Morse polynomial of odd degree $\geqslant 5$. Then, for $x \geqslant 2$, we have

$$\sum_{q \leqslant x} |W_1(a;q) \cdots W_m(a;q)| \ll \frac{x}{(\log x)^{\gamma}}$$

$$\sum_{q \leqslant x} |W_1(a;q) \cdots W_m(a;q)|^2 \ll x(\log\log x)^A$$

$$\sum_{q \leqslant x} |W_1(a;q) \cdots W_m(a;q)|^4 \ll x(\log x)^{2^{m-s}3^s-1}(\log\log x)^A$$

for some $\gamma > 0$ and some $A \geqslant 0$ depending only on m and $(d_1, \ldots, d_m)$, where $W_i(a;q) = W_{f_i}(a;q)$. The implied constants depend on the polynomials.

Remark 1.12

(1) Since the upper bounds for two polynomials essentially match those in Theorem 1.1, this result suggests that the exponential sums are uncorrelated. However, we cannot prove it rigorously, since we would need to prove some matching lower bound, such as

$$\sum_{q \leqslant x} |W_f(a;q)|^4 \gg x(\log\log x)^B$$

for any $B \geqslant 1$, for instance. The best current lower bound that we can achieve in general (by adapting the method of Fouvry and Michel [13, §4]) is

$$\sum_{q \leqslant x} |W_f(a;q)|^4 \gg \frac{x}{\log x}(\log\log x)^B$$

for any $B \geqslant 1$ (and the best upper bound that we can give for the last sum is

$$\sum_{q \leqslant x} |W_f(a;q)|^4 \ll x(\log x)(\log\log x)^A$$

for some A).

(2) In another paper, Fouvry and Michel [12, Th. 1.2, 1.3] proved that if f is a Sidon–Morse polynomial, then there are infinitely many squarefree integers q with two prime factors such that

$$|W_f(a;q)| \leqslant q^{-\beta}$$

where $\beta > 0$ depends only on the degree of f. It would be interesting to extend this property to all indecomposable polynomials.

1.4 Previous Work

Fouvry and Michel also consider rational functions and lower bounds. In the case of the Kloosterman sums

$$\mathrm{Kl}_2(a;q) = \frac{1}{\sqrt{q}} \sum_{(x,q)=1} e\Big(\frac{ax+\bar{x}}{q}\Big)$$

(i.e., $f(x) = x + 1/x$), they obtain

$$\frac{x}{\log x} \exp((\log\log x)^{5/12}) \ll \sum_{q \leqslant x} |\mathrm{Kl}_2(a;q)| \ll \frac{x}{(\log x)^{\delta}} \tag{5}$$

for any $\delta < 1 - \frac{8}{3\pi}$ (see [13, Th. 1.2, 1.3]).

In this particular case, it is known that if we sum the Kloosterman sums without taking absolute values, one can prove much stronger estimates using the spectral theory of automorphic forms, like

$$\sum_{q \leqslant x} \mathrm{Kl}_2(1;q) \ll x^{2/3+\varepsilon}$$

for any $\varepsilon > 0$ (see, e.g., [19, §16.6]). Patterson [28] has also proved a strong result for certain cubic sums, namely for any non-zero integer a, the asymptotic formula

$$\sum_{q \leqslant x} \sum_{0 \leqslant n < q} e\Big(\frac{an^3}{q}\Big) \sim c(a) X^{4/3}$$

holds for some explicit constant $c(a) > 0$, and Patterson [29, Conj. 2.2] has conjectured similar asymptotic formulas for all cubic polynomials.

It would be of considerable interest to obtain general conditions on a twisted multiplicative function $V(a;q)$, bounded at primes, that ensure a power saving in the sums

$$\sum_{q \leqslant x} V(a;q).$$

Outline of the Paper We prove Theorem 1.5 in the next section. Section 3 gathers a number of properties of exponential sums with polynomials, and Sect. 4 uses these results to prove Theorem 1.1, assuming Theorem 1.3. The latter is proved in Sect. 5, and Sect. 6 discusses generic polynomials. In both of these, we rely heavily on the foundational studies of Katz. Section 7 concludes with the proof of Theorem 1.11, and Sect. 8 contains some hopefully enlightening comments concerning parts of the results of Katz that we use.

2 Sums of Twisted Multiplicative Functions

Since the proof of Theorem 1.5 follows the broad plan of our earlier work (and is not far from that of Fouvry and Michel [13, §3]), we shall be brief.

Put $z = x^{1/(\alpha \log\log x)}$ with $\alpha = 3(M^2+1)$. We factor any integer $q \leqslant x$ as $q = rs$ where all prime factors of s are $\leqslant z$ and all prime factors of r are $> z$. We then have

$$V(a;q) = V(a;rs) = V(\bar{r}a;s)V(\bar{s}a;r)$$

by twisted multiplicativity; hence

$$|V(a;q)| \leqslant G(r)|V(\bar{r}a;s)|.$$

We handle first the terms where $s \leqslant x^{1/3}$. We split the sum over $q \leqslant x$ according to the residue class of r modulo s, getting

$$\sum_{\substack{q\leqslant x\\ s\leqslant x^{1/3}}} |V(a;q)| \leqslant \sum_{s\leqslant x^{1/3}} \sum_{r\leqslant x/s} G(r)|V(\bar{r}a;s)| \leqslant \sum_{s\leqslant x^{1/3}} \sum_{t \,(\mathrm{mod}\, s)} |V(\bar{t}a;s)| \sum_{\substack{r\leqslant x/s\\ r\equiv t \,(\mathrm{mod}\, s)}} G(r).$$

By Shiu's work on the Brun–Titchmarsh theorem for multiplicative functions (see [32, Th. 1]), we may bound the sum over r above by

$$\ll \frac{x/s}{\varphi(s)\log(x/s)} \exp\Big(\sum_{z<p\leqslant x} \frac{G(p)}{p}\Big) \ll \frac{x}{s\varphi(s)\log x}\Big(\frac{\log x}{\log z}\Big)^{\mathrm{M}} \ll \frac{x}{s\varphi(s)\log x}(\log\log x)^{\mathrm{M}}.$$

Therefore,

$$\sum_{\substack{q\leqslant x\\ s\leqslant x^{1/3}}} |V(a;q)| \ll \frac{x}{\log x}(\log\log x)^M \sum_{s\leqslant x^{1/3}} \frac{1}{s\varphi(s)} \sum_{t \,(\mathrm{mod}\, s)} |V(\bar{t}a;s)|$$

$$\ll \frac{x}{\log x}(\log\log x)^M \sum_{s\leqslant x^{1/3}} \frac{g(s)}{\varphi(s)} \ll \frac{x}{\log x}(\log\log x)^M \prod_{p\leqslant x}\Big(1+\frac{g(p)}{p}\Big).$$

We now consider the contribution of the terms with $s > x^{1/3}$. Since $G(p) \leqslant M$ for all p,

$$\sum_{\substack{q \leqslant x \\ s > x^{1/3}}} |V(a;q)| \leqslant \sum_{r \leqslant x^{2/3}} M^{\omega(r)} \sum_{x^{1/3} < s \leqslant x/r} M^{\omega(s)}.$$

Applying the Cauchy–Schwarz inequality and [26, Lemma 3.2] to the inner sum, we find that

$$\begin{aligned}\sum_{x^{1/3} < s \leqslant x/r} M^{\omega(s)} &\ll \Big(\sum_{s \leqslant x/r} M^{2\omega(s)}\Big)^{1/2} \Big(\sum_{x^{1/3} < s \leqslant x/r} 1\Big)^{1/2} \\ &\ll \frac{x}{r}(\log x)^{(M^2-1)/2} \exp\Big(-\frac{\log(x/r)}{2\log z}\Big) \ll \frac{x}{r}(\log x)^{(M^2-1)/2-\alpha/6} \ll \frac{x}{r\log x}.\end{aligned}$$

Therefore,

$$\sum_{\substack{q \leqslant x \\ s > x^{1/3}}} |V(a;q)| \ll \frac{x}{\log x} \sum_{r \leqslant x^{2/3}} \frac{M^{\omega(r)}}{r} \ll \frac{x}{\log x} \exp\Big(\sum_{z \leqslant p \leqslant x} \frac{M}{p}\Big) \ll \frac{x}{\log x}(\log\log x)^M.$$

The proof of Theorem 1.5 is now complete.

3 Exponential Sums of Polynomials: Preliminary Results

In this section, we collect together some results on the exponential sums $W_f(a;p)$. We shall use and expand on some of these results in later sections. First we recall the Weil bound: if $f \in \mathbf{Z}[X]$ has degree $d \geqslant 1$ and $(a,p)=1$, then

$$|W_f(a;p)| \leqslant (d-1). \tag{6}$$

Next we quote a result from Shao [31, Th. 2.1].

Lemma 3.1 *Let $f \in \mathbf{Z}[X]$ be a polynomial of degree d. Let κ denote the number of irreducible factors of $f(X) - f(Y) \in \mathbf{Q}[X,Y]$. Then, $\kappa \leqslant \tau(d)$ (the number of divisors of d), and for large x, we have*

$$\sum_{p \leqslant x} \frac{1}{p}\Big(\frac{1}{p}\sum_{(a,p)=1} |W(a;p)|^2\Big) = (\kappa-1)\log\log x + O(1).$$

Proof The asymptotic for the sum over primes is given in Theorem 2.1 of Shao [31], and the bound on κ is described in the remark after Theorem 2.1. □

While Lemma 3.1 involves the factorization of $F(X,Y)=(f(X)-f(Y))/(X-Y)$ in $\mathbf{Q}[X,Y]$, it is of greater significance to understand the factorization of $F(X,Y)$ over $\bar{\mathbf{Q}}[X,Y]$ (or equivalently over $\mathbf{C}[X,Y]$).

Lemma 3.2 *Let $f\in\mathbf{Z}[X]$ be a polynomial of degree d, and suppose that the polynomial $F(X,Y)=(f(X)-f(Y))/(X-Y)$ factors into m irreducible factors over $\bar{\mathbf{Q}}[X,Y]$. If $m=1$, then for all p we have*

$$\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times}|W(a;p)|^2=1+O(p^{-1/2}).$$

If $m>1$, then there is a set of primes $\mathscr{P}$ of density $\geqslant\delta>0$ (with δ depending only on the degree d) such that for $p\in\mathscr{P}$

$$\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times}|W(a;p)|^2=m+O(p^{-1/2}).$$

Proof If $m=1$, then the affine curve with equation $F(X,Y)=0$ is geometrically irreducible over $\mathbf{Q}$, so that for all large p it is geometrically irreducible over $\mathbf{F}_p$. Orthogonality of characters and the Riemann hypothesis for curves over finite fields then show that

$$\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times}|W(a;p)|^2=\frac{1}{p}\Big|\Big\{(x,y)\in\mathbf{F}_p^2\,:\,F(x,y)=0\Big\}\Big|=1+O(p^{-1/2}).$$

Now suppose $m>1$, and let K be a finite Galois extension of $\mathbf{Q}$ such that $F(X,Y)$ factors in $K[X,Y]$ into m different factors, each of which is irreducible in $\bar{\mathbf{Q}}[X,Y]$. Thus, the affine curve defined by $F(X,Y)$ is the union of m geometrically irreducible curves over K. Note that the degree of the field K may be bounded in terms of d. We take $\mathscr{P}$ to be the set of primes splitting completely in K. By the Chebotarev density theorem, $\mathscr{P}$ has density $1/[K:\mathbf{Q}]$, which is bounded away from 0 by an amount depending only on d. For $p\in\mathscr{P}$, the m geometrically irreducible components of the curve $F(X,Y)=0$ are defined over $\mathbf{F}_p$, and the Riemann hypothesis gives here

$$\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times}|W(a;p)|^2=m+O(p^{-1/2}).$$

□

Our next result is due to Fried [16, Th. 1] (see also the more elementary account by Turnwald in [33, Th. 1]). It describes when the polynomial $F(X,Y)=(f(X)-$

$f(Y))/(X-Y)$ is absolutely irreducible, i.e., when $m=1$ in the notation of the previous lemma, and therefore, $\kappa=2$ in the notation of Lemma 3.1.

We recall that for any integer $d\geqslant 0$, the Dickson polynomial $D_d\in\mathbf{Z}[X,a]$ is defined to be the unique polynomial such that

$$D_d(X+aX^{-1},a)=X^d+(a/X)^d$$

(see, e.g., [33, §1]); in particular, $D_d(X,0)=X^d$.

Proposition 3.3 (Fried) *Let $f\in\mathbf{Z}[X]$ with degree $d\geqslant 1$, and let*

$$F=(f(X)-f(Y))/(X-Y)\in\mathbf{Q}[X,Y].$$

(1) *If $\deg(f)$ is not an odd prime, then F is absolutely irreducible if and only if f is indecomposable in $\mathbf{Q}[X]$.*
(2) *If d is an odd prime $\geqslant 5$, then F is absolutely irreducible if it is not linearly equivalent in $\mathbf{Q}[X]$ to a Dickson polynomial $D_d(X,a)$.*
(3) *If $d=3$, then F is absolutely irreducible if and only if f is not linearly equivalent in $\mathbf{Q}[X]$ to a Dickson polynomial $D_3(X,0)$.*

Putting Lemmas 3.1, 3.2, and Proposition 3.3 together, we arrive at the following corollary:

Corollary 3.4 *Let $f\in\mathbf{Z}[X]$ be a polynomial of degree $d\geqslant 1$. If f is indecomposable, then for large x we have*

$$\sum_{p\leqslant x}\frac{1}{p}\Bigl(\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times}|W(a;p)|^2\Bigr)=\log\log x+O(1),$$

whereas if f is decomposable, then for large x we have

$$\sum_{p\leqslant x}\frac{1}{p}\Bigl(\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times}|W(a;p)|^2\Bigr)\geqslant 2\log\log x+O(1).$$

Proof If d is prime, then κ must be $2=\tau(d)$ in Lemma 3.1. Moreover, f is automatically indecomposable, and so the stated result holds in this case. If $f=g\circ h$ is decomposable, then $f(X)-f(Y)$ has $(X-Y)$, $(h(X)-h(Y))/(X-Y)$ and $(g(h(X))-g(h(Y)))/(h(X)-h(Y))$ as factors, so that $\kappa\geqslant 3$ in Lemma 3.1 and the stated result holds. Finally, if the degree d is composite and f is indecomposable, then the first part of Proposition 3.3 shows that $(f(X)-f(Y))/(X-Y)$ is irreducible in $\bar{\mathbf{Q}}[X,Y]$ and, therefore, in $\mathbf{Q}[X,Y]$. Either Lemma 3.1 or Lemma 3.2 now gives the stated result. □

Lastly, we consider the behavior of $W(a;p)$ when f is assumed to be Sidon–Morse over $\mathbf{Q}$. Here the work of Katz permits a very precise understanding of such exponential sums.

Proposition 3.5 *Let $f \in \mathbf{Z}[X]$ be a polynomial of degree d, and suppose that f is Sidon–Morse over $\mathbf{Q}$. Let K_d denote the compact group $\mathrm{USp}_{d-1}(\mathbf{C})$ if f is symmetric Sidon–Morse and the compact group $\mathrm{SU}_{d-1}(\mathbf{C})$ if f is Sidon–Morse but not symmetric. For any integer $k \geqslant 0$, we have*

$$\lim_{p\to+\infty} \frac{1}{p} \sum_{a\in\mathbf{F}_p^\times} |W(a;p)|^{2k} = \int_{K_d} |\mathrm{tr}(g)|^{2k} d\mu(g),$$

where μ is the Haar measure on K_d normalized to have total volume 1. Furthermore,

$$\int_{\mathrm{USp}_{d-1}(\mathbf{C})} |\mathrm{tr}(g)|^{2k} d\mu(g) \begin{cases} = (2k-1)!! & \textit{for } 1 \leqslant k \leqslant (d-1)/2 \\ \leqslant (2k-1)!! & \textit{for all } k \geqslant 1, \end{cases}$$

and

$$\int_{\mathrm{SU}_{d-1}(\mathbf{C})} |\mathrm{tr}(g)|^{2k} d\mu(g) \begin{cases} = k! & \textit{for } 0 \leqslant k \leqslant (d-1) \\ \leqslant k! & \textit{for all } k \geqslant 0, \end{cases}$$

Proof This is largely a consequence of the work of Katz [22]. We recall the relevant result of Katz in Theorem 6.3 and explain the link to the moments over K_d in Remark 6.10. Further discussion of Katz's theorem may be found in Sect. 8.

The moments over K_d for small k (which match the moments of a standard complex Gaussian for $K_d = \mathrm{SU}_{d-1}(\mathbf{C})$ and the moments of a standard real Gaussian for $K_d = \mathrm{USp}_{d-1}(\mathbf{C})$) were computed by Diaconis and Shahshahani, and the upper bounds for all k may be found in the work of Perret-Gentil [30, Prop. 2.2]. □

Remark 3.6 Katz's Theorem also leads to a possible guess for the optimal value of the upper bound in (2), in the case of Sidon–Morse polynomials, namely

$$\delta = 1 - \int_{K_d} |\mathrm{tr}(g)| d\mu(g)$$

(for instance, if $K_d = \mathrm{SU}_2(\mathbf{C})$, this leads to $\delta = 1 - 8/(3\pi)$, as in (5)).

Asymptotically, for large d, we have the Gaussian approximations

$$\int_{K_d} |\mathrm{tr}(g)| d\mu(g) \approx \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} |x| e^{-x^2/2} dx = \sqrt{\frac{2}{\pi}} = 0.79788456\ldots,$$

$$\int_{K_d} |\mathrm{tr}(g)| d\mu(g) \approx \frac{1}{\pi}\int_{\mathbf{C}} |z| e^{-|z|^2} dz = \frac{\sqrt{\pi}}{2} = 0.8662269\ldots$$

in the $\mathrm{USp}_{d-1}(\mathbf{C})$ and $\mathrm{SU}_{d-1}(\mathbf{C})$ cases, respectively.

4 Proof of Theorem 1.1

We begin with the first part of the theorem, which seeks a bound for $\sum_{q\leqslant x} |W(a;q)|^2$. We apply Theorem 1.5 to the function $q \mapsto W(a;q)^2$, which is twisted multiplicative. The Weil bound (6) allows us to take $G(p) = (d-1)^2$ for all but finitely many primes. Writing

$$g(p) = \frac{1}{p}\sum_{(a,p)=1} |W(a;p)|^2,$$

and recalling that f is indecomposable, Corollary 3.4 gives

$$\sum_{p\leqslant x} \frac{g(p)}{p} = \log\log x + O(1).$$

Theorem 1.5 yields

$$\sum_{q\leqslant x} |W(a;q)|^2 \ll \frac{x}{\log x}\exp\Big(\sum_{p\leqslant x}\frac{g(p)}{p}\Big)(\log\log x)^{(d-1)^2} \ll x(\log\log x)^{(d-1)^2}.$$

Now we turn to the proof of the second part of the theorem, which we will deduce from Theorem 1.5 and Theorem 1.3 (to be proved in Sect. 5). Applying Theorem 1.5 to the twisted multiplicative function $|W(a;q)|$ and using the Weil bound (which permits $M = d-1$ here), we obtain

$$\sum_{q\leqslant x} |W(a;q)| \ll \frac{x}{\log x}(\log\log x)^{d-1}\exp\Big(\sum_{p\leqslant x}\frac{1}{p}\Big(\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times} |W(a;p)|\Big)\Big). \tag{7}$$

Let ϵ be a small positive number, and let $\mathscr{P}$ denote the set of primes p for which

$$\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times} |W(a;p)|^4 \geqslant 2-\epsilon.$$

By Theorem 1.3, we know that the set $\mathscr{P}$ has density $\geqslant \delta = \delta(d) > 0$ with δ depending only on d. For any real number y with $|y| \leqslant d-1$, we claim that

$$|y| \leqslant \frac{1+y^2}{2}, \qquad \text{and} \qquad |y| \leqslant \frac{1+y^2}{2} + \frac{3/2-y^4}{200(d-1)^4}.$$

The first inequality is clear and so is the second inequality in the range $y^4 \leqslant 3/2$. In the range $3/2 < y^4 \leqslant (d-1)^4$, note that $(1+y^2)/2 - |y| \geqslant (1+\sqrt{3/2})/2 - (3/2)^{1/4} > 1/200$, so that the desired inequality holds in this case also.

Applying the first inequality above for primes $p \notin \mathscr{P}$, we find

$$\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times} |W(a;p)| \leqslant \frac{1}{2} + \frac{1}{2p}\sum_{a\in\mathbf{F}_p^\times} |W(a;p)|^2,$$

while applying the second inequality above for primes $p \in \mathscr{P}$, we find

$$\begin{aligned}\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times} |W(a;p)| &\leqslant \frac{1}{2} + \frac{1}{2p}\sum_{a\in\mathbf{F}_p^\times} |W(a;p)|^2 + \frac{1}{200(d-1)^4}\Big(\frac{3}{2} - \frac{1}{p}\sum_{a\in\mathbf{F}_p^\times} |W(a;p)|^4\Big)\\ &\leqslant \frac{1}{2} + \frac{1}{2p}\sum_{a\in\mathbf{F}_p^\times} |W(a;p)|^2 - \frac{1}{400(d-1)^4}.\end{aligned}$$

Combining both inequalities, and using the first part of Corollary 3.4, we conclude that

$$\sum_{p\leqslant x}\frac{1}{p^2}\sum_{a\in\mathbf{F}_p^\times} |W(a;p)| \leqslant \Big(\frac{1}{2} + \frac{1}{2} - \frac{\delta}{400(d-1)^4} + o(1)\Big)\log\log x.$$

Inserting this bound in (7), the second part of the theorem follows:

5 The Fourth Moment: Proof of Theorem 1.3

As we shall see, for Sidon–Morse polynomials, the work of Katz [22] can be used to show that Case (1) of Theorem 1.3 holds. The main challenge is to handle all polynomials of degree $\geqslant 3$ and not just the generic ones.

Let $f \in \mathbf{Z}[X]$ be a polynomial with $d = \deg(f) \geqslant 3$. If $(f(X)-f(Y))/(X-Y)$ is not absolutely irreducible, then Lemma 3.2 shows that there is a positive density of primes on which the second moment of $W(a;p)$ is at least $2+O(p^{-1/2})$, so that by Cauchy–Schwarz a stronger form of the second case of Theorem 1.3 holds (with the fourth moment being $\geqslant 4+O(p^{-1/2})$).

From now on, we will therefore assume that the polynomial

$$F(X,Y) = (f(X)-f(Y))/(X-Y)$$

is absolutely irreducible. The remaining part of the proof will use in an essential way the algebraic interpretation of the exponential sums $W(a; p)$, which goes back to Weil, and it seems difficult to prove the lower bound for the fourth moment with a direct elementary argument.

Fix a prime ℓ (for instance, $\ell = 2$); all primes p will be assumed to be different from ℓ and to be larger than d. Let ι be a fixed isomorphism $\bar{\mathbf{Q}}_\ell \to \mathbf{C}$; we use it to identify ℓ-adic numbers and complex numbers.

Let $p \neq \ell$, $p > d$, be a prime number. We denote by ψ_p the ℓ-adic additive character of $\mathbf{F}_p$ such that

$$\iota(\psi_p(a)) = e\Big(\frac{a}{p}\Big)$$

for $a \in \mathbf{F}_p$.

Let $\mathscr{G}_p$ be the ℓ-adic sheaf $f_*\bar{\mathbf{Q}}_\ell/\bar{\mathbf{Q}}_\ell$ on the affine line $\mathbf{A}^1_{\mathbf{F}_p}$; it has rank $d-1$ and is everywhere tamely ramified (since $p > d$). The sheaf $\mathscr{G}_p$ is a Fourier sheaf in the sense of Katz [22, 7.3.5], and we denote by $\mathscr{F}_p$ its (unitarily normalized) Fourier transform with respect to ψ_p (defined in [22, 7.3.3], up to the normalization). The trace function of $\mathscr{F}_p$ takes value 0 for $a = 0$ and takes value (after applying ι)

$$\frac{1}{\sqrt{p}} \sum_{x \in \mathbf{F}_p} e\Big(\frac{af(x)}{p}\Big) = W(a; p)$$

for $a \in \mathbf{F}_p^\times$ (see [22, Th. 7.3.8, (4)], where again the Fourier transform is not normalized). The rank of $\mathscr{F}_p$ is also equal to $d-1$, and $\mathscr{F}_p$ is lisse and pure of weight 0 outside 0 and ∞ (see [22, Lemma 7.3.9]).

Lemma 5.1 *If the polynomial $(f(X) - f(Y))/(X - Y)$ is absolutely irreducible over $\mathbf{Q}$, then for all p large enough, the sheaf $\mathscr{F}_p$ is geometrically irreducible.*

Proof This is a Fourier-side variant of Lemma 3.2. If the polynomial

$$F(X, Y) = (f(X) - f(Y))/(X - Y)$$

is absolutely irreducible, then the curve $C_{f,p}$ over $\mathbf{F}_p$ with equation

$$(f(x) - f(y))/(x - y) = 0$$

is geometrically irreducible, which by the Riemann hypothesis for curves implies that as $\nu \to +\infty$, we have

$$|C_{f,p}(\mathbf{F}_{p^\nu})| \sim p^\nu.$$

However, the discrete Parseval formula implies that

$$\frac{1}{p^\nu}|C_{f,p}(\mathbf{F}_{p^\nu})| = \frac{1}{p^\nu}\sum_{a\in\mathbf{F}_{p^\nu}^\times}\Bigl|\frac{1}{p^{\nu/2}}\sum_{x\in\mathbf{F}_{p^\nu}} e\Bigl(\frac{\mathrm{tr}(af(x))}{p}\Bigr)\Bigr|^2$$

(with the trace from $\mathbf{F}_{p^\nu}$ to $\mathbf{F}_p$) so we obtain

$$\lim_{\nu\to+\infty}\frac{1}{p^\nu}\sum_{a\in\mathbf{F}_{p^\nu}^\times}\Bigl|\frac{1}{p^{\nu/2}}\sum_{x\in\mathbf{F}_{p^\nu}} e\Bigl(\frac{\mathrm{tr}(af(x))}{p}\Bigr)\Bigr|^2 = 1,$$

and this implies that $\mathscr{F}_p$ is geometrically irreducible by Katz's diophantine criterion for irreducibility (see, e.g., [27, Lemma 4.14]). □

We now consider only primes p such that the sheaf $\mathscr{F}_p$ is geometrically irreducible.

Let G_p be the arithmetic monodromy group of $\mathscr{F}_p$ and G_p^g the geometric monodromy subgroup; we can view these as algebraic subgroups of $\mathrm{GL}_{d-1}(\bar{\mathbf{Q}}_\ell)$. The irreducibility property of $\mathscr{F}_p$ means that G_p^g acts irreducibly on $\bar{\mathbf{Q}}_\ell^{d-1}$.

By a deep theorem of Deligne (see [11, Th. 3.4.1 (iii) and Cor. 1.3.9]), the connected component of the identity $G_{p,0}^g$ of the group G_p^g is semisimple. It is invariant under all automorphisms of G_p^g; hence, it is a normal subgroup of G_p (since inner automorphisms of G_p induce automorphisms of its normal subgroup G_p^g). Let f_p denote a fixed element of the conjugacy class of the Frobenius automorphism at p.

Let $\mathscr{E}_p$ be the sheaf $\mathrm{End}(\mathrm{End}(\mathscr{F}_p))$. Its trace function for $a\in\mathbf{F}_p^\times$ is $|W(a;p)|^4$.

Let V_p be the subspace $\mathrm{End}(\mathrm{End}(\bar{\mathbf{Q}}_\ell^{d-1}))^{G_p^g}$ of vectors invariant under G_p^g, the action of G_p on the space $\mathrm{End}(\mathrm{End}(\bar{\mathbf{Q}}_\ell^{d-1}))$ being "the obvious one" induced by the action on $\bar{\mathbf{Q}}_\ell^{d-1}$ (if a group G acts on a vector space E, it acts on $\mathrm{End}(E)$ by $g\cdot u = g\circ u\circ g^{-1}$).

Applying the Grothendieck–Lefschetz trace formula and Deligne's version of the Riemann hypothesis, we get a formula

$$\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times}|W(a;p)|^4 = \iota(\mathrm{tr}(f_p|V_p)) + O(p^{-1/2}) \tag{8}$$

where the implied constant depends only on d (e.g., by conductor estimates, much as in [15, Th. 9.1]).

Proposition 5.2 *There exists a finite Galois extension K of $\mathbf{Q}$ of degree bounded in terms of d only such that for all but finitely many primes p that are totally split in K, the action of f_p on $V_p = \mathrm{End}(\mathrm{End}(\bar{\mathbf{Q}}_\ell^{d-1}))^{G_p^g}$ is trivial.*

Let us admit this proposition and conclude the proof of Theorem 1.3. For primes totally split in the number field K, we have $\iota(\mathrm{tr}(f_p|V_p)) = \dim(V_p)$. On the other hand, the definition of the action of G_p^g on $\mathrm{End}(\bar{\mathbf{Q}}_\ell^{d-1})$ shows that the space V_p is the space of all linear maps $\mathrm{End}(\bar{\mathbf{Q}}_\ell^{d-1}) \to \mathrm{End}(\bar{\mathbf{Q}}_\ell^{d-1})$ which commute with the G_p^g-action. The identity is an element of this space, so its dimension is $\geqslant 1$. Since the action on $\mathrm{End}(\bar{\mathbf{Q}}_\ell^{d-1})$ is semisimple (e.g., by Deligne's Theorem [11, Th. 3.4.1] because it is still pure of weight 0), Schur's Lemma in representation theory (see, e.g., [25, Prop. 2.7.15 (3)]) implies that the dimension of V_p is exactly 1 if and only if the action of G_p^g on $\mathrm{End}(\bar{\mathbf{Q}}_\ell^{d-1})$ is irreducible. However, V_p contains both the multiples of the identity and the space $\mathrm{End}^0(\bar{\mathbf{Q}}_\ell^{d-1})$ of matrices of trace zero as stable subspaces, so this irreducibility can only hold if $\mathrm{End}^0(\bar{\mathbf{Q}}_\ell^{d-1})$ is zero, i.e., if $d = 2$. So for primes totally split in K, we have $\dim(V_p) \geqslant 2$; hence

$$\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times} |W(a;p)|^4 \geqslant 2 + O(p^{-1/2})$$

by (8).

To improve on this unless the limit is equal to 2, we use very deep work of Katz [22, Th. 14.3.4] that implies that $G_{p,0}^g$ is independent of p for all p large enough. Take a prime p large enough so that $G_{p,0}^g$ has stabilized, and suppose that $\dim(V_p) = 2$ for some p split in K. Then, the group $G_{p,0}^g$ must act irreducibly on matrices of trace zero. However, the Lie algebra of G_p^g is a stable subspace, so that we must have $\mathrm{Lie}(G_{p,0}^g) = \mathrm{End}^0(\bar{\mathbf{Q}}_\ell^{d-1})$. That means that $G_{p,0}^g$ is equal to $\mathrm{SL}_{d-1}(\bar{\mathbf{Q}}_\ell)$. Then, for all primes p large enough, we have $ZG_{p,0}^g = \mathrm{GL}_{d-1}(\bar{\mathbf{Q}}_\ell)$, where Z is the group of scalar matrices in $\mathrm{GL}_{d-1}(\bar{\mathbf{Q}}_\ell)$, which implies that f_p acts trivially for all p large enough and then that the limit of the fourth moments exists and is equal to 2.

To finally show that the constant 2 is best possible, we recall that Katz has proved that if f is a Sidon–Morse polynomial (e.g., the derivative f' has Galois group S_{d-1}), then G_p^g contains $\mathrm{SL}_{d-1}(\bar{\mathbf{Q}}_\ell)$ for all p large enough (see Theorem 6.3), in which case it is well-known that the action of G_p^g on the space of matrices of trace zero is irreducible, so that the dimension of V_p is then equal to 2 for all p large enough.

Remark 5.3 The arguments above are related to the easiest part of the Larsen's Alternative [23].

Proof of Proposition 5.2 We will begin by proving the statement without the information that the degree of K can be bounded in terms of d only, since the latter requires extra ingredients.

Step 1. We first prove that, for all primes p large enough, the action of f_p on V_p is of finite order. Since we are assuming that G_p^g acts irreducibly on $\bar{\mathbf{Q}}_\ell^{d-1}$, a result of Katz shows that the connected component of the identity $G_{p,0}^g$ of G_p^g acts irreducibly on $\bar{\mathbf{Q}}_\ell^{d-1}$, provided p is large enough (see [22, 7.7.3, Lemma 7.7.5]).

Recall that the group of outer automorphisms of $G^g_{p,0}$ is the group $\mathrm{Out}(G^g_{p,0})$ of automorphisms modulo inner automorphisms. For $g \in G_p$, let $\alpha_p(g) \in \mathrm{Out}(G^g_{p,0})$ be the class modulo inner automorphisms of the automorphism $x \mapsto xgx^{-1}$ of $G^g_{p,0}$ (it is an automorphism since $G^g_{p,0}$ is normal in G_p). This defines a group homomorphism

$$G_p \xrightarrow{\alpha_p} \mathrm{Out}(G^g_{p,0}).$$

We claim that the kernel of α_p is $G^g_{p,0}Z \cap G_p$ where Z is again the group of scalar matrices in $\mathrm{GL}_{d-1}(\bar{\mathbf{Q}}_\ell)$. Indeed, the condition $\alpha_p(g) = 1$ means that there exists $h \in G^g_{p,0}$ such that $gxg^{-1} = hxh^{-1}$ for all $x \in G^g_{p,0}$, which is equivalent to $h^{-1}g$ belonging to the centralizer of $G^g_{p,0}$ in $\mathrm{GL}_{d-1}(\bar{\mathbf{Q}}_\ell)$ or, in other words, to $h^{-1}g$ commuting with the action of $G^g_{p,0}$ on $\bar{\mathbf{Q}}_\ell^{d-1}$. By Schur's Lemma (see, e.g., [25, Prop. 2.7.15 (2)]), the irreducibility of the action of $G^g_{p,0}$ implies that this centralizer is equal to Z. Thus, $g \in \ker(\alpha_p)$ is equivalent to $g \in G^g_{p,0}Z \cap G_p$. We deduce therefore that we have an injective group homomorphism

$$G_p/(G^g_{p,0}Z \cap G_p) \xrightarrow{\alpha_p} \mathrm{Out}(G^g_{p,0}).$$

Because $G^g_{p,0}$ is a connected semisimple group, its outer automorphism group is finite (see, e.g., [4, p. 42, prop. 18] in the case of compact groups). Hence, α_p injects $G_p/(G^g_{p,0}Z \cap G_p)$ in a finite group. Since Z acts trivially on $\mathrm{End}(W)$ for any representation W, and since G^g_p acts trivially on V_p, this shows that the order of the action of f_p on V_p is a divisor of the order of the outer automorphism group.

Step 2. We next prove that there exists a finite-dimensional continuous ℓ-adic Galois representation

$$\varrho\colon \mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}(E)$$

for some $\bar{\mathbf{Q}}_\ell$-vector space E, such that for all but finitely many primes, the action of Frobenius at p on E "is" the same as the action of f_p on V_p. It is enough to define a constructible ℓ-adic sheaf $\mathscr{V}$ on $\mathrm{Spec}(\mathbf{Z}[1/\ell N])$ for some integer $N \geqslant 1$ such that the stalk over all but finitely many primes p "is" the space V_p and such that the action of f_p coincides with the action of the Frobenius at p. Indeed, this sheaf $\mathscr{V}$ will be lisse outside of a finite set S of primes and hence will correspond to a Galois representation of the Galois group of the maximal extension unramified outside S, and this is a quotient of the Galois group of $\mathbf{Q}$. To construct $\mathscr{V}$, we use [27, Lemma 4.23] (see also [27, Lemma 4.27] for a more difficult application), applied to the data

$$(X, Y, f, g) = (\mathbf{A}^4, \mathrm{Spec}(\mathbf{Z}[1/\ell]), \text{the structure morphism},$$
$$g(x, y, z, w) = f(x) + f(y) - f(z) - f(w))$$

and take the second cohomology sheaf of the complex resulting from this application of [27, Lemma 4.23].

That this "works" results from the expression

$$\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times}\Bigl|\frac{1}{\sqrt{p}}\sum_{x\in\mathbf{F}_p}e\Bigl(\frac{af(x)}{p}\Bigr)\Bigr|^4=\frac{1}{p^3}\sum_{x,y,z,w\in\mathbf{F}_p}\sum_{a\in\mathbf{F}_p^\times}e\Bigl(\frac{ag(x,y,z,w)}{p}\Bigr),$$

combined with the cohomological expression

$$V_p \simeq H_c^2(\mathbf{G}_m \times \bar{\mathbf{F}}_p, \mathrm{End}(\mathrm{End}(\mathscr{F}_p)))(1). \tag{9}$$

Step 3. By the compatibility with Frobenius of the isomorphism (9) in Step 2, and by Step 1, the action of Frobenius at p under ϱ is of finite order for all but finitely many primes p. The image H of ϱ is a compact ℓ-adic Lie group (identifying $\mathrm{GL}(E)$ with $\mathrm{GL}_m(\bar{\mathbf{Q}}_\ell)$ for some $m \geqslant 1$, we first note that H is contained in $\mathrm{GL}_m(L)$ for some finite extension L of $\mathbf{Q}_\ell$, by an oft-rediscovered lemma—see, for instance, [24, Lemma 9.0.8]—and then it is a closed subgroup of an ℓ-adic Lie group, hence itself an ℓ-adic Lie group by, e.g., [1, p. 227, th. 2]). It follows from [1, Cor. 1, p. 169] that there is a neighborhood U of $1 \in H$ which contains no non-trivial finite subgroup; there is then a number field K such that the finite-index subgroup $\mathrm{Gal}(\bar{\mathbf{Q}}/K)$ maps to U. All the Frobenius elements in this subgroup (which exist outside any given finite set of primes because Frobenius elements are dense, by a form of Chebotarev's density theorem) must map to the identity, which means that $\mathrm{Gal}(\bar{\mathbf{Q}}/K)$ is in the kernel of ϱ. This implies that for a prime p that is totally split in K, the action of f_p, which "is" the action of Frobenius under ϱ, is trivial. This proves the result, up to the bound on the degree of K.

Step 4. Now we explain how to bound the degree of K in terms of d only.

The first ingredient is a fact from the theory of finite groups: for given positive integers k and m, if Γ is a finite subgroup of $\mathrm{GL}_k(\bar{\mathbf{Q}}_\ell)$ such that all elements of Γ have order dividing m, then the order of Γ is bounded in terms of k and m only. Indeed, by a well-known theorem of Jordan (see, e.g., [10, Th. 36.13]), there exists a normal abelian subgroup Γ_0 of Γ of index bounded in terms of k and m. This reduces the problem to the abelian case; but Γ_0 can be diagonalized, and the bound on the order of its elements shows that Γ_0 is isomorphic to a subgroup of $(\mathbf{Z}/m\mathbf{Z})^k$, hence the result.

We want to apply this to the image $\Gamma \subset \mathrm{GL}(E)$ of the Galois representation ϱ. We have $\dim(E) \leqslant (d-1)^4$. By the Chebotarev density theorem, it is then enough to prove that the order of the action of f_p on V_p is uniformly bounded in terms of d only. For this we use the fact that there are, up to isomorphism, only finitely many possibilities for $G^g_{p,0}$, since it is a connected and semisimple subgroup of GL_{d-1} (this follows, in the equivalent case of compact Lie groups, from the discussion in [4, §4, n^{0}9, Scholie], which shows that such subgroups are classified by their root system R, which here has rank $\leqslant d-1$, which gives only finitely many possibilities, and for each root system R by a subgroup of the quotient $Q(R)/P(R)$ discussed in loc. cit.; since this quotient is finite by [2, §1, n^{o}10], there are again only finitely many possibilities). So the order of f_p is a divisor of the order of one of finitely many finite groups (depending only on d). □

Remark 5.4 The argument in Step 4 shows that it is possible to give an effective value for the constant γ in Theorem 1.1. Indeed, the index of Γ_0 in Jordan's Theorem can be bounded effectively (for instance, one gets from [10, Th. 36.14] that

$$|\Gamma| \leqslant m^k |\Gamma_0| \leqslant m^k(\sqrt{8k}+1)^{2k^2},$$

and better bounds are known), and the order of the groups $G_{p,0}/G^g_{0,0}$ can also be bounded effectively from the classification of roots systems.

6 Generic Polynomials

In this section, we will prove the kind of non-correlation estimates modulo primes that are needed in the proof of Theorem 1.11. We also explain Proposition 3.5 at the end.

We first make some remarks concerning Sidon–Morse polynomials:

Lemma 6.1 *Let K be any field and let $f \in K[X]$ be a Morse polynomial of degree $d \geqslant 2$.*

(1) *The polynomial f is indecomposable over K.*
(2) *For any $c \in K$, the polynomials $f+c$ and $-f+c$ are Morse polynomials. If f is a Sidon–Morse polynomial, then $f+c$ and $-f+c$ are Sidon–Morse polynomials.*

Proof

(1) We show that if f is decomposable, then it is not a Morse polynomial. Let $f = g \circ h$ where $\deg(g) \geqslant 2$ and $\deg(h) \geqslant 2$ be a decomposable polynomial. Note that p does not divide either $\deg(g)$ or $\deg(f)$ since $p \nmid d$.

For any critical point α of g, the critical values of f contain, with multiplicity, the values $g(h(\beta))$ where $h(\beta) = \alpha$. This will give rise to a critical value with multiplicity at least 2 *unless* $h - \alpha = \gamma(X-\beta)^{\deg(h)}$ for some $\gamma \in K^\times$. Since p does not divide $\deg(h)$, this can only occur for a single value of α, so that g is of the form

$$g = \delta(X-\alpha)^{\deg(g)} + \eta$$

for some $\delta \in K^\times$ and $\eta \in K$. Then, we get

$$g \circ h = \eta + \delta \gamma^{\deg(g)} (X - \beta)^d,$$

which has a single critical value and is therefore not a Morse polynomial.

(2) This is straightforward from the definition, since the critical points of $g = f + c$ (resp. $g = -f + c$) are the same as those of f, so the critical values of g are those of f translated by c (resp. the negative of those of f, translated by c).

□

Let p be a prime number and $f \in \mathbf{F}_p[X]$ a Sidon–Morse polynomial. We define the ℓ-adic sheaf $\mathscr{F}_f$ associated to f as in the previous section. We will now normalize it in a specific way. We denote by $\mathscr{L}_2$ the Kummer sheaf associated to the Legendre character, with trace function $a \mapsto (a/p)$.

Definition 6.2 (Normalized Sheaf) Let p be a prime and $f \in \mathbf{F}_p[X]$ a Sidon–Morse polynomial with $p \nmid \deg(f) - 1$.

(1) If f is not symmetric Sidon, then there is a unique $c \in \mathbf{F}_p$ such that the sum of the critical values of $f + c$ is equal to 0, and the *normalized* sheaf $\widetilde{\mathscr{F}}_f$ of f is defined to be

$$\widetilde{\mathscr{F}}_f = \mathscr{F}_{f+c} \otimes \mathscr{L}_2^{d-1}.$$

We then say that $c = c_f$ is the *critical shift* of f.

(2) If f is symmetric Sidon polynomial, and

$$f = g(\beta X + \gamma) + \delta$$

where g is odd, then we put

$$\widetilde{\mathscr{F}}_f = \mathscr{F}_g.$$

We note that the sum of critical values of g is then equal to 0.

The trace function of $\widetilde{\mathscr{F}}_f$ is 0 for $a = 0$ and for $a \in \mathbf{F}_p^\times$ is given either by

$$\widetilde{W}_f(a; p) = \frac{1}{\sqrt{p}} \Big(\frac{a}{p}\Big)^{d-1} \sum_{x \in \mathbf{F}_p} e\Big(\frac{a(f(x) + c)}{p}\Big) = \Big(\frac{a}{p}\Big)^{d-1} e\Big(\frac{ac}{p}\Big) \sum_{x \in \mathbf{F}_p} e\Big(\frac{af(x)}{p}\Big) \tag{10}$$

or by

$$\widetilde{W}_f(a; p) = \frac{1}{\sqrt{p}} \sum_{x \in \mathbf{F}_p} e\Big(\frac{ag(x)}{p}\Big)$$

$$= \frac{1}{\sqrt{p}} e\Big(-\frac{a\delta}{p}\Big) \sum_{x\in\mathbf{F}_p} e\Big(\frac{af((x-\gamma)/\beta)}{p}\Big) = e\Big(-\frac{a\delta}{p}\Big) W_f(a;p). \qquad (11)$$

in the symmetric case. In particular, we see that in all cases, the formula

$$|\widetilde{W}_f(a;p)| = |W_f(a;p)|$$

is valid all a modulo p.

The point of this normalization is the following theorem of Katz:

Theorem 6.3 (Katz) *Let p be a prime number. Let $f \in \mathbf{F}_p[X]$ be a Sidon–Morse polynomial of degree $d \geqslant 3$. Assume that $p > 2d-1$ and that $p \nmid d-1$.*

(1) *If f is not a symmetric Sidon–Morse polynomial, then the geometric monodromy group of $\widetilde{\mathscr{F}}_f$ is equal to $\mathrm{SL}_{d-1}(\bar{\mathbf{Q}}_\ell)$.*
(2) *If f is a symmetric Sidon–Morse polynomial, which implies that d is odd, then the geometric monodromy group of $\widetilde{\mathscr{F}}_f$ is isomorphic to $\mathrm{Sp}_{d-1}(\bar{\mathbf{Q}}_\ell)$.*

Proof

(1) If f is not of symmetric type, then the geometric monodromy group contains SL_{d-1} under the assumption on p, by [22, Th. 7.9.6], and has trivial determinant by [22, Lemma 7.10.4, (2)], so it must be SL_{d-1}.
(2) If f is of symmetric type, then under the assumption on p, a conjugate of the geometric monodromy group of $\widetilde{\mathscr{F}}_f$ is contained in Sp_{d-1} by [22, Lemma 7.10.4, (3)] (since the associated polynomial g is odd). By [22, Th. 7.9.7], it contains either SL_{d-1} or Sp_{d-1} or SO_{d-1}; the only possibility that is compatible with both these facts is that it is Sp_{d-1}.

□

Remark 6.4 If we consider a Morse polynomial f such that the set of critical values is a symmetric Sidon set, we might hope that (2) still holds. However, although one can still deduce from the work of Katz that the geometric monodromy group of $\widetilde{\mathscr{F}}_f$ contains a symplectic group, we currently do not know if this condition is sufficient to ensure that $\widetilde{\mathscr{F}}$ has conversely a symplectic symmetry.

We will also need a result that is essentially a consequence of the ideas of Fried.

Proposition 6.5 *Let p be a prime number. Let f and g in $\mathbf{F}_p[X]$ be Sidon–Morse polynomials of respective degree $d_f \geqslant 3$ and $d_g \geqslant 3$. Assume that $d_f < p$ and $d_g < p$.*

If f and g are not linearly equivalent over $\bar{\mathbf{F}}_p$, then $f(X)-g(Y)+c$ and $f(X)+g(Y)+c$ are absolutely irreducible for any c.

Proof Since $f+c$ is a Sidon–Morse polynomial (Lemma 6.1), and linearly equivalent to g if and only if f is, we can assume that $c=0$. Since $-g$ is a Sidon–Morse polynomial, and linearly equivalent to f if and only if so is g, we need to only consider the case of $f(X)-g(Y)$.

Let G be the Galois group of the equation $f(X) - Y = 0$ over the field $\bar{\mathbf{F}}_p(Y)$ (so X is the variable). If $f(X) - g(Y)$ is not absolutely irreducible, then G is also isomorphic to the one for the equation $g(X) - Y = 0$ by [9, §2.1.1].[1] By [9, §2.1.4], if f and g are not linearly equivalent over $\bar{\mathbf{F}}_p$, then the faithful permutation representations of G on the roots of these two equations are not equivalent as permutation representations, but have the same character (i.e., are equivalent as linear representations). However, for Sidon–Morse polynomials f and g, the group G and its permutation representation are isomorphic to $\mathfrak{S}_d$ with the standard permutation representation on d letters (see [22, Proof of Lemma 7.10.2.3]). However, this is a contradiction, since this faithful permutation representation of $\mathfrak{S}_d$ is characterized by its character (the only non-obvious case is when $d = 6$, and we consider the standard permutation representation and that given by a non-trivial outer automorphism of $\mathfrak{S}_6$, but these have different characters, e.g., because a transposition is mapped to, respectively, a transposition, with 4 fixed points, or a product of three disjoint transpositions, without fixed points). □

Proposition 6.6 *Let p be a prime. Let $m \geqslant 1$ be an integer and let $f_1, \ldots, f_m$ be Sidon–Morse polynomials in $\mathbf{F}_p[X]$. Assume that $p > 2\deg(f_i) - 1$ and $p \nmid (\deg(f_i) - 1)$ for all i. Assume also that for all $i \neq j$, the polynomials f_i and f_j are not linearly equivalent over $\bar{\mathbf{F}}_p$.*

Then, the geometric monodromy group of the sheaf

$$\bigoplus_{1 \leqslant i \leqslant m} \widetilde{\mathscr{F}}_{f_i}$$

is the direct product of the geometric monodromy groups of the sheaves $\widetilde{\mathscr{F}}_{f_i}$.

Proof We write $d_i = \deg(f_i)$ and $\widetilde{\mathscr{F}}_i = \widetilde{\mathscr{F}}_{f_i}$. We also denote by $\widetilde{\mathscr{F}}_i^{\vee}$ the dual of $\widetilde{\mathscr{F}}_i$.

We will apply the Goursat–Kolchin–Ribet criterion, as developed by Katz [22, Prop. 1.8.2] and expounded by Fouvry, Kowalski, and Michel [14, Lemma 2.4]. In the language of loc. cit., it suffices to check that the family $(\widetilde{\mathscr{F}}_i)$ is $\mathbf{G}_m$-generous [14, Def. 2.1], since the individual geometric monodromy groups of $\widetilde{\mathscr{F}}_i$ are connected by Theorem 6.3.

This desired property is the combination of four conditions. Condition (1) holds because the sheaves $\widetilde{\mathscr{F}}_i$ are pure of weight 0 on $\mathbf{G}_m$ and have a geometric monodromy group (namely SL_{d_i-1} or Sp_{d_i-1} by Theorem 6.3) that acts irreducibly on $\bar{\mathbf{Q}}_\ell^{d_i-1}$. Conditions (2) and (3) are then known properties of SL_{d_i-1} and Sp_{d_i-1} (see [14, §3.1]).

To prove the most important Condition (4), it is enough to check that if $i \neq j$, there is no geometric isomorphism

[1] This is written for the base field $\mathbf{C}$, but the argument extends to any algebraically closed field when the polynomials involved have degree less than the characteristic of the field.

$$\widetilde{\mathscr{F}}_i \simeq \widetilde{\mathscr{F}}_j \otimes \mathscr{L}, \quad \text{or} \quad \widetilde{\mathscr{F}}_i^{\vee} \simeq \widetilde{\mathscr{F}}_j \otimes \mathscr{L} \tag{12}$$

where $\mathscr{L}$ is a rank one sheaf lisse on $\mathbf{G}_m$ (see [14, Remark 2.2]). This is impossible unless $d_i = d_j$ and unless either none or both of f_i and f_j are symmetric Sidon–Morse. We now assume that $d_i = d_j$, and we denote by d this common value.

Case 1. Assume first that neither f_i nor f_j is symmetric and that we have the isomorphism $\widetilde{\mathscr{F}}_i \simeq \widetilde{\mathscr{F}}_j \otimes \mathscr{L}$ in (12). We denote by c_i and c_j the critical shifts of f_i and f_j.

We recall that since $p > 2d-1$, the sheaf $\widetilde{\mathscr{F}}_i$ is, for all i, tamely ramified at 0 [22, Lemma 7.10.4, (1)], with local monodromy isomorphic to the sum of the non-trivial characters of order d [22, Lemma 7.10.4, (1)]. These must be permuted by multiplication by the monodromy character χ_0 of $\mathscr{L}$ at 0, which is only possible if $\chi_0 = 1$, i.e., if $\mathscr{L}$ is lisse at 0.

Next, by [22, Th. 7.8.4, (2)] and the construction of $\widetilde{\mathscr{F}}_i$, the wild monodromy representation of $\widetilde{\mathscr{F}}_i$ at ∞ is the direct sum

$$\bigoplus_{v \in V_i} \mathscr{L}_{\psi(vX)} \tag{13}$$

where V_i is the set of critical values of $f_i + c_i$ and $\mathscr{L}_{\psi(vX)}$ denotes the Artin–Schreier sheaf modulo p with trace function $a \mapsto e(av/p)$. Let $v \in V_i$. The putative isomorphism $\widetilde{\mathscr{F}}_i \simeq \widetilde{\mathscr{F}}_j \otimes \mathscr{L}$ implies that there exists $w \in V_j$ such that

$$\mathscr{L}_{\psi(vX)} = \mathscr{L} \otimes \mathscr{L}_{\psi(wX)},$$

as representations of the wild inertia group at ∞. In particular, $\mathscr{L}$ is an Artin–Schreier sheaf at infinity, say $\mathscr{L} \simeq \mathscr{L}_{\psi(cX)}$ for some c, as representations of the wild inertia group. The local isomorphism becomes

$$\bigoplus_{v \in V_i} \mathscr{L}_{\psi(vX)} \simeq \bigoplus_{w \in V_j} \mathscr{L}_{\psi((c+w)X)},$$

so that $V_i = V_j + c$ as subsets of $\bar{\mathbf{F}}_p$. But taking the sum of the values on both sides, and using the definition of the normalized sheaf, we deduce that $c = 0$. Thus, the sheaf $\mathscr{L}$ is trivial on the wild monodromy group and, therefore, is also tamely ramified at ∞.

Since $\mathscr{L}$ is lisse on $\mathbf{G}_m$ and tame, it is a Kummer sheaf attached to some multiplicative character χ of $\mathbf{F}_p^{\times}$ (which is its trace function). Since it is lisse at 0, this character must be trivial. Hence, we deduce that $\widetilde{\mathscr{F}}_i$ and $\widetilde{\mathscr{F}}_j$ are in fact geometrically isomorphic.

By the diophantine criterion for irreducibility (see, e.g., [27, Lemma 4.14]), this implies that

$$\limsup_{\nu\to+\infty}\frac{1}{p^\nu}\Big|\sum_{a\in\mathbf{F}_{p^\nu}^\times}\widetilde{W}_i(a;p^\nu)\overline{\widetilde{W}_j(a;p^\nu)}\Big| = \limsup_{\nu\to+\infty}\frac{1}{p^\nu}\sum_{a\in\mathbf{F}_{p^\nu}^\times}|\widetilde{W}_i(a;p^\nu)|^2 = 1, \tag{14}$$

where $\widetilde{W}_i(a;p^\nu)$ is the trace function of $\widetilde{\mathscr{F}}_i$ over the extension of degree ν of $\mathbf{F}_p$. By (10) and orthogonality of characters, the sum on the left-hand side is equal to

$$\frac{1}{p^\nu}|\{(x,y)\in\mathbf{F}_{p^\nu}^2 \mid f_i(x)+c_i = f_j(y)+c_j\}| - 1$$

(noting that if the trace function of f_i has the Legendre factor, then so does f_j, and they cancel out). If the polynomial $f_i(X) - f_j(Y) + c_i - c_j$ is absolutely irreducible, then we get

$$\frac{1}{p^\nu}|\{(x,y)\in\mathbf{F}_{p^\nu}^2 \mid f_i(x)+c_i = f_j(y)+c_j\}| - 1 \ll p^{-\nu/2}$$

by the Riemann Hypothesis for curves, which contradicts (14). Thus, the polynomial

$$f_i(X) - f_j(Y) + c_i - c_j$$

is not absolutely irreducible, which can only happen if f_i and f_j are linearly equivalent over $\bar{\mathbf{F}}_p$ (Proposition 6.5).

Case 2. We continue assuming that neither f_i nor f_j is symmetric and consider the second case of an hypothetical isomorphism (12). It is elementary that the dual $\widetilde{\mathscr{F}}_i^\vee$ is the normalized sheaf associated to $-f_i$ (because $\mathscr{F}_i$ is the Fourier transform of a sheaf that is self-dual, being the direct image of the self-dual constant sheaf; see [22, Th. 7.3.8, (2)]). Thus, we are reduced to the previous case.

Case 3. Now we assume that f_i and f_j are symmetric. Since $\widetilde{\mathscr{F}}_i$ and $\widetilde{\mathscr{F}}_j$ are then self-dual by Theorem 6.3 (2), we need to only exclude the possibility of a geometric isomorphism of the form

$$\widetilde{\mathscr{F}}_i \simeq \widetilde{\mathscr{F}}_j \otimes \mathscr{L}.$$

Assume there is such an isomorphism. We denote by g_i and g_j the odd polynomials associated to f_i and f_j so that $\widetilde{\mathscr{F}}_i = \mathscr{F}_{g_i}$ and $\widetilde{\mathscr{F}}_j = \mathscr{F}_{g_j}$. Arguing exactly as in Case 1, we see that the sheaf $\mathscr{L}$ is trivial. Then, continuing again as in Case 1 using (11), we find that there are δ_i and δ_j such that

$$f_i(X) - f_j(Y) - \delta_i + \delta_j$$

is not absolutely irreducible, and Proposition 6.5 allows us to conclude that f_i and f_j would have to be linearly dependent.

□

Lemma 6.7 *Let f and g in $\mathbf{Z}[X]$ be polynomials of common degree $d \geqslant 3$. The polynomials f and g are linearly equivalent over $\bar{\mathbf{Q}}$ if and only if $f \pmod p$ and $g \pmod p$ are linearly equivalent over an algebraic closure $\bar{\mathbf{F}}_p$ of $\mathbf{F}_p$ for infinitely many primes.*

Proof The set $X_{f,g}$ of tuples (a, b, c, d) in $\bar{\mathbf{Q}}$ such that

$$g = af(cX + d) + b$$

is defined by polynomial equations with rational coefficients. The polynomials f and g are linearly equivalent over $\bar{\mathbf{Q}}$ if and only if $X_{f,g}(\bar{\mathbf{Q}})$ is not empty. Since $X_{f,g}$ is an algebraic variety, this is true if and only if $X_{f,g}(\bar{\mathbf{F}}_p)$ is not empty for all p large enough (e.g., by the Nullstellensatz: if $X_{f,g}(\bar{Q})$ is empty, then there is a representation of 1 as belonging to the ideal generated by the equations of $X_{f,g}$, and this leads to a representation of 1 over $\bar{\mathbf{F}}_p$ for all primes large enough), which proves the assertion. □

Corollary 6.8 *Let $m \geqslant 1$ be an integer, and let $f_1, \ldots, f_m$ be Sidon–Morse polynomials in $\mathbf{Z}[X]$ that are pairwise not linearly equivalent over $\bar{\mathbf{Q}}$. Let $s \leqslant m$ be the number of f_i such that f_i is symmetric Sidon–Morse of degree $\geqslant 5$.*

We have

$$\frac{1}{p} \sum_{(a,p)=1} |W_{f_1}(a; p) \cdots W_{f_m}(a; p)|^2 = 1 + O(p^{-1/2}) \tag{15}$$

$$\frac{1}{p} \sum_{(a,p)=1} |W_{f_1}(a; p) \cdots W_{f_m}(a; p)|^4 = 2^{m-s}3^s + O(p^{-1/2}) \tag{16}$$

where the implied constant depends only on m and on the degrees of the polynomials f_i.

Proof Applying Lemma 6.7, we see that for p large enough, the assumptions of Proposition 6.6 hold modulo p. Let p be such a prime. Using the same notation as in (8), the left-hand side of (15) is equal to

$$\iota(\mathrm{tr}(f_p | \operatorname{End}(W_p)^G)) + O(p^{-1/2})$$

where W_p is the tensor product space

$$\bigotimes_i \bar{\mathbf{Q}}_\ell^{d_i - 1}$$

as a representation of the geometric monodromy group G of

$$\bigoplus_i \widetilde{\mathscr{F}}_i.$$

By Proposition 6.6, this representation can be identified with the external tensor product of the representations of the individual geometric monodromy groups; since this external tensor product is an irreducible representation (see, e.g., [25, Prop. 2.3.23]), the invariant space has dimension one, spanned by the scalar matrices, on which f_p acts trivially, and the first result follows.

For the second result, we get in the same way the main term of (16) equal to

$$\prod_{i=1}^{m} \dim(\mathrm{End}(\mathrm{End}(\bar{\mathbf{Q}}_\ell^{d_i-1}))^{G_i}$$

where G_i is the geometric monodromy group of $\widetilde{\mathscr{F}}_i$. By the simplest case of the Larsen Alternative (see [23, Th. 1.1.6]), each factor is equal to 3 if f_i is a symmetric Sidon–Morse polynomial of degree $\geqslant 5$ (with symplectic monodromy) and to 2 for the others. □

We conclude this section with the following proposition, which is used in the proof of the first part of Theorem 1.11, where only one polynomial is assumed to be a Sidon–Morse polynomial.

Proposition 6.9 *Let f and g be non-constant polynomials in $\mathbf{Z}[X]$ of degrees d_f and d_g, respectively. Suppose that f is a Sidon–Morse polynomial, that $d_f < d_g$, and that g is absolutely irreducible. Then,*

$$\frac{1}{p} \sum_{(a,p)=1} |W_f(a;p)\overline{W_g(a;p)}|^2 = 1 + O(p^{-1/2})$$

where the implied constant depends only on d_f and d_g.

Proof This is a variant of the Goursat–Kolchin–Ribet argument, but where we only fully control one of the sheaves.

Let $p > d_f - 1$ be a prime such that f is a Sidon–Morse polynomial modulo p. We denote by $\widetilde{\mathscr{F}}_f$ the normalized sheaf associated to f modulo p and by G_f (resp. G_g) the geometric monodromy group of $\widetilde{\mathscr{F}}$ (resp. of $\mathscr{F}_g$). Since f is a Sidon–Morse polynomial, we have $G_f = \mathrm{SL}_{d_f-1}$ or $G_f = \mathrm{Sp}_{d_f-1}$ (the latter when f is symmetric Sidon–Morse) by Theorem 6.3.

Let further H be the geometric monodromy group of $\widetilde{\mathscr{F}}_f \oplus \mathscr{F}_g$. We have a natural inclusion $H \to G_f \times G_g$, and the composition of this inclusion with either projection is surjective.

We denote by W_p the space

$$\mathrm{End}(\widetilde{\mathscr{F}}_f \otimes \mathscr{F}_g)^H$$

and by f_p a representative of the Frobenius automorphism in H. The analogue of (8) in this case is the formula

$$\frac{1}{p}\sum_{a\in\mathbf{F}_p^\times}|W_f(a;p)\overline{W_g(a;p)}|^2=\iota(\mathrm{tr}(f_p|W_p))+O(p^{-1/2})$$

where the implied constant depends only on d_f and d_g (and we used the fact that the trace function of $\widetilde{\mathscr{F}}_f$ has the same modulus as that of $\mathscr{F}_f$). By Schur's Lemma, it then suffices to prove that the representation of H on $\widetilde{\mathscr{F}}_f\otimes\mathscr{F}_g$ is irreducible, and in turn, it is enough to prove that $H=G_f\times G_g$ (using again the irreducibility of external tensor product of irreducible representations, see [25, Prop. 2.3.23]).

We denote by L the kernel of the composition homomorphism

$$G_f\to H\subset G_f\times G_g\to G_g.$$

This is a normal subgroup of G_f; hence, L is either finite or equal to G_f. If the latter holds, then H contains $G_f\times\{1\}$, and it follows easily that $H=G_f\times G_f$.

Thus, we need to exclude the possibility that L is finite. However, if that is the case, then G_f/L is isomorphic to a subgroup of G_g; hence, the Lie algebra of G_f has a faithful representation of dimension $\leqslant d_g-1$. Since we assumed that $d_f>d_g$, this is impossible in view of the minimal dimensions of faithful representations of the Lie algebras of SL_{d_f-1} or Sp_{d_f-1} (which are equal to d_f-1; see, e.g., [3, p. 249, Exercice 2 and p. 214, Table 2]). □

Remark 6.10 Theorem 6.3 also implies Proposition 3.5. Indeed, using the same notation as in (8), the Riemann Hypothesis and conductor estimates imply that for k fixed and p large, we have

$$\sum_{a\in\mathbf{F}_p^\times}|W(a;p)|^{2k}=\nu_k+O(p^{-1/2}),$$

where ν_k is the multiplicity of the trivial representation of the geometric monodromy group in the representation $\mathrm{End}(\bar{\mathbf{Q}}_\ell^{d-1})^{\otimes k}$. By character theory for compact groups, we have

$$\nu_k=\int_{K_d}|\mathrm{tr}(g)|^{2k}d\mu(g)$$

for a maximal compact subgroup K_d of the geometric monodromy group, where μ is the Haar measure on K_d normalized to have total volume 1. We can take $K_d=\mathrm{SU}_{d-1}(\mathbf{C})$ if the geometric monodromy group is SL_{d-1} and $K_d=\mathrm{USp}_{d-1}(\mathbf{C})$ if it is Sp_{d-1}.

7 Multiple Correlations

We now come to Theorem 1.11. For the first part, we apply Theorem 1.5 to the function $a \mapsto |W_f(a;q)\overline{W_g(a;q)}|^2$. We can take $M = (d_f - 1)^2(d_g - 1)^2$. By Proposition 6.9, we have

$$\frac{1}{p}\sum_{(a,p)=1} |W_f(a;p)\overline{W_g(a;p)}|^2 = 1 + O(p^{-1/2})$$

so we can take $g(p) = 1 + O(p^{-1/2})$. Thus, Theorem 1.5 gives, for some constant $C \geqslant 0$, the bound

$$\begin{aligned}\sum_{q\leqslant x} |W_f(a;q)W_g(a;q)|^2 &\ll \frac{x}{\log x}\prod_{p\leqslant x}\Big(1+\frac{1}{p}+\frac{C}{p^{3/2}}\Big)(\log\log x)^{(d_f-1)^2(d_g-1)^2}\\ &\ll x(\log\log x)^{(d_f-1)^2(d_g-1)^2}.\end{aligned}$$

For the second part, we apply Theorem 1.5 to the functions

$$a \mapsto |W_1(a;q)\cdots W_m(a;q)|,$$

$$a \mapsto |W_1(a;q)\cdots W_m(a;q)|^2,$$

$$a \mapsto |W_1(a;q)\cdots W_m(a;q)|^4$$

and argue as in the proof of Theorem 1.1 using Corollary 6.8.

8 Remarks on Katz's Theorem

We want to observe that Katz's Theorem (Theorem 6.3) can be explained, in the case of monodromy SL_{d-1}, as the combination of two facts:

(1) The local monodromy computation (13), which has an intuitive meaning as the algebraic analogue of the stationary phase expansion for oscillatory integrals

$$g(t) = \int e^{itf(x)}dx,$$

(2) A result of Gabber (see [22, Th. 1.0]) which (essentially) deduces the nature of the monodromy group from the Sidon property of the critical values.

Since the proof of Gabber's result, in this special case, is relatively accessible and (in our opinion) quite enlightening with respect to the relevance of the Sidon condition, we include the precise statement and its proof.

Proposition 8.1 *Let V be a finite-dimensional complex vector space of dimension $r \geqslant 1$, and let G be a connected semisimple compact subgroup of* $\mathrm{GL}(V)$ *which acts irreducibly on V. Let D be the subgroup of elements of* $\mathrm{GL}(V)$ *which are diagonal with respect to some basis, and let χ_i, for $1 \leqslant i \leqslant r$, be the characters $D \to \mathbf{C}^\times$ giving the coefficients of the elements of D.*

Let $A \subset D$ be a subgroup of the normalizer of G in $\mathrm{GL}(V)$. *Let $S \subset \widehat{A}$ be the subset of the group of characters of A given by the restrictions to A of the diagonal characters χ_i. If $|S| = r$ and S is a Sidon set in $\widehat{A}$, then $G = \mathrm{SU}(V)$.*

Proof We denote by $Z \subset D$ the subgroup of scalar matrices. We may assume that $G \subset \mathrm{U}(V)$.

The group G is a compact real Lie group. We consider the representation of the group A on $\mathrm{End}(V)$ by conjugation. It acts on the elementary matrices $E_{i,j}$ by $\chi_i \chi_j^{-1}$. The assumption that S has r elements and is a Sidon set means then that

$$\mathrm{End}(V) = \bigoplus_{i,j} \mathbf{C} E_{i,j}$$

is a decomposition of the representation as a sum of characters where, for $i \neq j$, the line $\mathbf{C}E_{i,j}$ is a non-trivial character of multiplicity one.

Since $A \subset N_{\mathrm{GL}(V)}(G)$, the complexified Lie algebra $L \subset \mathrm{End}(V)$ of G is a subrepresentation of the representation of A on $\mathrm{End}(V)$. Thus, there exists a subspace H of the diagonal matrices and a subset X of pairs (i, j) of distinct integers such that

$$L = H \oplus \bigoplus_{(i,j)\in X} \mathbf{C} E_{i,j}.$$

This implies that L is in fact stable under conjugation by all of D. We have therefore an induced morphism

$$D \to \mathrm{Aut}(L),$$

which induces an injective morphism $D/Z \to \mathrm{Aut}(L)$. Its image is contained in the neutral component of $\mathrm{Aut}(L)$. Since L is semisimple, the latter is equal to the adjoint group of G (see, e.g., [2, p. 244, Prop. 30, (ii)]). It follows that the connected semisimple group $G \subset \mathrm{SU}(V)$ has rank $r - 1$; it follows that $G = \mathrm{SU}(V)$ (e.g., by the Borel–de Siebenthal Theorem: the group G coincides with the connected component of the identity of the centralizer in $\mathrm{SU}(V)$ of the center of G, for instance by [4, p. 36, prop. 13], and the center is contained in the group of scalar matrices by Schur's Lemma, so its centralizer is $\mathrm{SU}(V)$). □

This proposition is applied to a conjugate of the *finite* subgroup A of elements of the form

$$\mathrm{diag}(e(xv_1/p), \ldots, e(xv_{d-1}/p))$$

where $(v_1, \ldots, v_{d-1})$ are the critical values of f; indeed, the local monodromy computation implies that such a subgroup is contained in a maximal compact subgroup of the monodromy group.

Acknowledgments E.K. was partially supported by a DFG-SNF lead agency program grant (grant number 200020L_175755). K.S. is partially supported through a grant from the National Science Foundation and a Simons Investigator Grant from the Simons Foundation. This work was started when K.S. was a senior Fellow at the ETH Institute for Theoretical Studies, whom he thanks for their warm and generous hospitality.

We thank W. Sawin for his comments concerning Sect. 5.

References

1. Bourbaki, N.: Groupes et algèbres de Lie, chapitre III. Springer
2. Bourbaki, N.: Groupes et algèbres de Lie, chapitre VI. Springer.
3. Bourbaki, N.: Groupes et algèbres de Lie, chapitre VIII. Springer
4. Bourbaki, N.: Groupes et algèbres de Lie, chapitre IX. Springer
5. Bourgain, J.: On the maximal ergodic theorem for certain subsets of the integers. Isr. J. Math. **61**, 39–72 (1988)
6. Bourgain, J.: An approach to pointwise ergodic theorems. In: Geometric Aspects of Functional Analysis (1986/87). Lecture Notes in Math., vol. 1317, pp. 204–223. Springer (1988)
7. Bourgain, J.: Pointwise ergodic theorems for arithmetic sets. Publications Mathématiques de l'IHéS **69**, 5–41 (1989)
8. Bourgain, J., Kontorovich, A.: On the local-global conjecture for integral Apollonian gaskets, with an appendix by Péter Varjú. Invent. math. **196**, 589–650 (2014)
9. Cassou-Noguès, P., Couveignes, J.M.: Factorisations explicites de $g(y) - h(z)$. Acta Arith. **87**, 291–317 (1999)
10. Curtis, C., Reiner, I.: Representation Theory of Finite Groups and Associative Algebras, vol. 356. AMS Chelsea Publishing (1962)
11. Deligne, P.: La conjecture de Weil, II. Publ. Math. IHÉS **52**, 137–252 (1980)
12. Fouvry, É., Michel, Ph.: À la recherche de petites sommes d'exponentielles. Annales de l'Institut Fourier **52**, 47–80 (2002)
13. Fouvry, É., Michel, Ph.: Sommes de modules de sommes d'exponentielles. Pacific J. Math. **209**, 261–288 (2003); erratum, Pacific J. Math. **225**, 199–200 (2006)
14. Fouvry, É., Kowalski, E., Michel, Ph.: A study in sums of products. Phil. Trans. R. Soc. A **373**, 20140309 (2014)
15. Fouvry, É., Kowalski, E., Michel, Ph.: Algebraic twists of modular forms and Hecke orbits. Geom. Funct. Anal. **25**, 580–657 (2015); https://doi.org/10.1007/s00039-015-0310-2
16. Fried, M.: On a conjecture of Schur. Michigan Math. J. **17**, 41–55 (1970)
17. Fried, M., Jarden, M.: Field arithmetic. Ergebnisse der Math., vol. 11. Springer (2008)
18. Hooley, C.: On the distribution of the roots of polynomial congruences. Mathematika **11**, 39–49 (1964)
19. Iwaniec, H., Kowalski, E.: Analytic Number Theory, vol. 53. AMS Colloquium Publ., (2004)

20. Katz, N.M.: Gauss sums, Kloosterman sums and monodromy groups. Annals of Math. Studies, vol. 116. Princeton Univ. Press (1988)
21. Katz, N.M.: Perversity and exponential sums. In: Algebraic Number Theory in honor of K. Iwasawa. Adv. Studies Pure Math., vol. 17, pp. 209–259 (1989)
22. Katz, N.M.: Exponential sums and differential equations. Annals of Math. Studies, vol. 124. Princeton Univ. Press (1990)
23. Katz, N.M.: Larsen's alternative, moments and the monodromy of Lefschetz pencils. In: Contributions to Automorphic Forms, Geometry, and Number Theory (Collection in Honor of J. Shalika's 60th Birthday), pp. 521–560. J. Hopkins Univ. Press (2004)
24. Katz, N.M., Sarnak, P.: Random matrices, Frobenius eigenvalues and monodromy. Colloquium Publ., vol. 45. A.M.S., (1999)
25. Kowalski, E.: An introduction to the representation theory of groups. Grad. Studies in Math., vol. 155. A.M.S., (2014)
26. Kowalski, E., Soundararajan, K.: Equidistribution from the Chinese remainder theorem. Adv. Math. **385**, 107776 (2021)
27. Kowalski, E., Michel, Ph., Sawin, W.: Bilinear forms with Kloosterman sums and applicatins. Annals Math. **186**, 413–500 (2017)
28. Patterson, S.J.: On the distribution of certain Hua sums, II. Asian J. Math. **6**, 719–730 (2002)
29. Patterson, S.J.: The asymptotic distribution of exponential sums, II. Experiment. Math. **14**, 87–98 (2005)
30. Perret-Gentil, C.: Gaussian distribution of short sums of trace functions over finite fields. Math. Proc. Camb. Phil. Soc. **163**, 385–422 (2017)
31. Shao, X.: Polynomial values modulo primes on average and sharpness of the larger sieve. Algebra Number Theory **9**, 2325–2346 (2015)
32. Shiu, P.: A Brun-Titchmarsh theorem for multiplicative functions. J. Reine Angew. Math. **313**, 161–170 (1980)
33. Turnwald, G.: On Schur's conjecture. J.. Aust. Math. Soc. **58**, 312–357 (1995)

The Ternary Goldbach Problem with a Missing Digit and Other Primes of Special Types

Helmut Maier and Michael Th. Rassias

Dedicated to the Memory of Jean Bourgain

Abstract The goal of the present paper is to prove on assumption of the Generalized Riemann Hypothesis that each sufficiently large odd integer N_0 can be expressed in the form

$$N_0 = p_1 + p_2 + p_3 \, ,$$

where p_1, p_2 are Piatetski-Shapiro primes and p_3 is a prime with a missing digit.

2010 Mathematics Subject Classification: 11P32, 11N05, 11A63

1 Introduction

The ternary Goldbach problem concerns the representation of a large odd integer N as a sum of three primes. It was first treated by Vinogradov [15] for arbitrary primes (cf. [13]). Helfgott [5] showed that this is true for all odd $N \geq 7$.
The problem was later modified by the request that the primes be of special type. Balog and Friedlander [1] considered the ternary Goldbach problem with variables restricted to Piatetski-Shapiro primes.

H. Maier
Department of Mathematics, University of Ulm, Ulm, Germany
e-mail: helmut.maier@uni-ulm.de

M. Th. Rassias (✉)
Department of Mathematics and Engineering Sciences, Hellenic Military Academy, Vari Attikis, Greece

Institute for Advanced Study, Program in Interdisciplinary Studies, Princeton, NJ, USA
e-mail: michail.rassias@math.uzh.ch

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_14

In [10], H. Maier and M. Th. Rassias showed that on assumption of the Generalized Riemann Hypothesis each sufficiently large odd integer is the sum of a prime and two isolated primes. For other combinations of the types, see [1–3, 8, 9, 14].
The goal of the present paper is the proof of the following:

Theorem 1.1 *On assumption of the Generalized Riemann Hypothesis (GRH), each sufficiently large odd integer N_0 can be represented in the form*

$$N_0 = p_1 + p_2 + p_3\ ,$$

where for $i = 1, 2$ the p_i are of the form $p_i = [n_i^{c_0}]$, $n_i \in \mathbb{N}$, with $c_0 = 1/\gamma_0$, $\gamma^* < \gamma_0 \leq 1$, where

$$\gamma^* = \frac{8}{9} + \frac{2}{3}\,\frac{\log(10/9)}{\log 10}\ (\approx 0.919\ldots)$$

and the decimal expansion of p_3 does not contain the digit a_0.

2 Outline of the Proof

In the sequel, we provide the main ideas as well as a sketch of proofs of the key statements. For a more detailed presentation of the proof, the reader is referred to [11].
One ingredient of the proof is Maynard's approach [12] using sieve decompositions based on ideas of Harman [4]. Another ingredient is the discrete circle method.
We recall the following definition from [11] which we complement by a few new definitions.

Definition 2.1 Let $a_0 \in \{0, 1, \ldots, 9\}$, $k \in \mathbb{N}$, and let

$$\mathcal{A} := \left\{ \sum_{0 \leq i \leq k} n_i 10^i\ :\ n_i \in \{0, 1, \ldots, 9\} \setminus \{a_0\} \right\}\ ,$$

$$X := 10^k\ ,\ \ \mathcal{B} := \{n \leq X,\ n \in \mathbb{N}\}\ ,$$

$\mathbb{P}$ the set of prime numbers,

$$S_{\mathcal{A}}(\theta) := \sum_{a \in \mathcal{A}} e(a\theta)\ ,\ \ S_{\mathbb{P}}(\theta) := \sum_{p \leq x} e(p\theta)\ ,\ \ S_{\mathcal{A}\cap\mathbb{P}}(\theta) := \sum_{p \in \mathcal{A}\cap\mathbb{P}} e(p\theta)\ .$$

Let $\mathcal{C}$ be a set of integers. We define the characteristic function $1_{\mathcal{C}}$ by

$$1_{\mathcal{C}}(n) := \begin{cases} 1\,, & \text{if } n \in \mathcal{C} \\ 0\,, & \text{if } n \notin \mathcal{C}\,. \end{cases}$$

For $d \in \mathbb{N}$, we set

$$\mathcal{C}_d := \{c \;:\; cd \in \mathcal{C}\}.$$

The sifted set $\mathcal{U}(\mathcal{C}, z)$ is defined by

$$\mathcal{U}(\mathcal{C}, z) := \{c \in \mathcal{C} \;:\; p \mid c \;\Rightarrow\; p > z\}\,.$$

The sieving function $S(\mathcal{C}, z)$—the counting function of $\mathcal{U}(\mathcal{C}, z)$—is given by

$$S(\mathcal{C}, z) := \#\mathcal{U}(\mathcal{C}, z) = \#\{c \in \mathcal{C} \;:\; p \mid c \;\Rightarrow\; p > z\}\,.$$

We let

$$w_n := 1_{\mathcal{A}}(n) - \frac{\kappa_{\mathcal{A}}\#\mathcal{A}}{\#\mathcal{B}}\,, \quad \kappa_{\mathcal{A}} := \begin{cases} \dfrac{10(\Phi(10)-1)}{9\Phi(10)}\,, & \text{if } (10, a_0) = 1 \\ \dfrac{10}{9}\,, & \text{otherwise}\,, \end{cases}$$

$$S_d(z) := \sum_{\substack{n<X/d \\ p\mid n \;\Rightarrow\; p>z}} w_{nd} = S(\mathcal{A}_d, z) - \frac{\kappa_{\mathcal{A}}\#\mathcal{A}}{X}\,S(\mathcal{B}_d, z)\,,$$

$$1_{\mathcal{A}}(n) \;\text{ is called the } \mathcal{A} - \text{part of } w_n\,,$$

$$S(\mathcal{A}_d, z) \;\text{ is called the } \mathcal{A} - \text{part of } S_d(z)\,,$$

The $\mathcal{B}$-parts are defined analogously.
We also define the exponential sums

$$S(\mathcal{C}, z, \theta) := \sum_{n \in \mathcal{U}(\mathcal{C}, z)} e(n\theta)\,,$$

$$S_d(z, \theta) := \sum_{\substack{n<X/d \\ p\mid n \;\Rightarrow\; p>z}} w_{nd}e(n\theta) = S(\mathcal{A}_d, z, \theta) - \frac{\kappa_{\mathcal{A}}\#\mathcal{A}}{X} S(\mathcal{B}_d, z, \theta)\,, \quad (\theta \in \mathbb{R})\,.$$

The essential idea of Harman's sieve is contained in Harman [4], Theorem 3.1 from [4] (The Fundamental Theorem).

Suppose that for any sequences of complex numbers, a_m, b_n, that satisfy $|a_m| \leq 1$, $|b_n| \leq 1$, we have for some $\lambda > 0$, $\alpha > 0$, $\beta \leq 1/2$, $M \geq 1$ that

$$\sum_{\substack{mn \in \mathcal{A} \\ m \leq M}} a_m = \lambda \sum_{\substack{mn \in \mathcal{B} \\ m \leq M}} a_m + O(Y) \tag{1}$$

and

$$\sum_{\substack{mn \in \mathcal{A} \\ X^{\alpha} \leq m \leq X^{\alpha+\beta}}} a_m b_n = \lambda \sum_{mn \in \mathcal{B}} a_m b_n + O(Y)\,, \tag{2}$$

where Y is a suitably chosen constant.

Let c_r be a sequence of complex numbers, such that $|c_r| \leq 1$, and if $c_r \neq 0$, then

$$p \mid r \;\Rightarrow\; p > x^{\epsilon}\ , \text{ for some } \epsilon > 0. \tag{3}$$

Then, if $X^{\alpha} < M$, $2R < \min(X^{\alpha}, M)$ and $M > X^{1-\alpha}$, if $2R > X^{\alpha+\beta}$, we have

$$\sum_{r \sim R} c_r S(\mathcal{A}_r, X^{\beta}) = \lambda \sum_{r \sim R} c_r S(B_r, X^{\beta}) + O(Y \log^3 X)\,. \tag{4}$$

Equation (1) is known as type I information, whereas (2) is known as type II information.

In the application of Theorem 3.1, information about a (complicated) set $\mathcal{A}$ is obtained from that of a (simple) set $\mathcal{B}$.

In Maynard's paper [12], the sets $\mathcal{A}$ and $\mathcal{B}$ are those from Definition 2.1. In a first step of the sieve decomposition, the counting function of interest $\#\{p \in \mathcal{A}\}$ is broken up as follows:

$$\begin{aligned}\#\{p \in \mathcal{A}\} &= \#\{p \in \mathcal{A} \;:\; p > X^{1/2}\} + O(X^{1/2}) \\ &= S_1(z_4) + (1 + o(1))\,\frac{\kappa_{\mathcal{A}} \#\mathcal{A}}{\log X} \ \text{(here } z_4 = X^{1/2}\text{)}\,.\end{aligned}$$

The function S_1 is now replaced in a series of steps by other terms of the form S_d. These steps consist in the application of *Buchstab's recursion*:

Let $u_1 < u_2$. Then,

$$S(\mathcal{C}, u_2) = S(\mathcal{C}, u_1) - \sum_{u_1 < p \leq u_2} S(\mathcal{C}_p, p)\,. \tag{5}$$

In the transformation of the form

$$S_d(z) = S(\mathcal{A}_d, z) - \frac{\kappa_{\mathcal{A}} \#\mathcal{A}}{\log X}\, S(\mathcal{B}_d, z)\,,$$

(5) is now applied separately with $C = \mathcal{A}_d$ and $C = \mathcal{B}_d$, and we get the recursion:

$$S_1(u_2) = S_1(u_1) - \sum_{u_1 < p \leq u_2} S_p(p)\,.$$

An important observation is that "the counting function version" of the Buchstab recursion is linked to a "characteristic function version."

$$1_{\mathcal{U}(\mathcal{C},u_2)}(n) = 1_{\mathcal{U}(\mathcal{C},u_1)}(n) - \sum_{u_1 < p \leq u_2} 1_{\mathcal{U}(\mathcal{C}_p,p)}(n)\,. \tag{6}$$

In our paper, the discrete circle method is applied, and therefore, we multiply (6) with the exponential function $e(n\theta)$ $(= e^{2\pi i n\theta})$ to obtain

$$1_{\mathcal{U}(\mathcal{C},u_2)}(n)e(n\theta) = 1_{\mathcal{U}(\mathcal{C},u_1)}(n)e(n\theta) - \sum_{u_1 < p \leq u_2} 1_{\mathcal{U}(\mathcal{C}_p,p)}(n)e(n\theta)\,. \tag{7}$$

We get the following version of Buchstab's recursion, which we state as

Lemma 2.2 *Let $u_1 < u_2$. Then,*

$$S(\mathcal{C}, u_2, \theta) = S(\mathcal{C}, u_1, \theta) - \sum_{u_1 < p \leq u_2} S(\mathcal{C}_p, p, \theta)\,. \tag{8}$$

We introduce another modification in our paper. Instead of considering all the integers in $\mathcal{A}$ as possible candidates for our representation of N_0, we now only choose the integers from a subset $\mathcal{A}^*$ of $\mathcal{A}$, which are contained in a short subinterval of $\mathcal{B}$.

Definition 2.3 Let $H \in \mathbb{N}$, $H \leq k$. For

$$n = \sum_{j=1}^{k} n_j 10^j, \quad (n_j \in \{0, \ldots, 9\}),$$

we write

$$n_{H,1} := \sum_{j=k-H+1}^{k} n_j 10^j =: \tilde{n}_H \cdot 10^{k-H+1}$$

and

$$n_{H,2} := \sum_{j=0}^{k-H} n_j 10^j\,.$$

Lemma 2.4 *Let $n = \tilde{n}_H \cdot 10^{k-H+1}$ as in Definition 2.3. Then,*

$$n \in \mathcal{A} \ \text{ if and only if } \ \tilde{n}_H \in \mathcal{A} \ \text{ and } \ n_{H,2} \in \mathcal{A} \tag{9}$$

There is an integer $\tilde{n}_H \in \mathcal{A} \cap [0, 10^{H-1}]$ such that for $n_H^* := \tilde{n}_H 10^{k-H+1}$ we have the following:

$$\left| n_H^* - 5 \cdot 10^{k-1} \right| \leq \frac{3}{2}\, 10^{k-2} \tag{10}$$

and for $n_{H,2} \in \mathcal{B}^* := [n_H^*, n_H^* + 10^{k-H})$, the following holds:

$$n_H^* + n_{H,2} \in \mathcal{A} \ \Rightarrow \ n_{H,2} \in \mathcal{A}\,. \tag{11}$$

Proof Equation (9) is obvious. To show (10) and (11), we consider the following cases:
Case 1: $a_0 = 5$, *Case 2*: $a_0 = 4$, *Case 3*: $a_0 \notin \{4, 5\}$.
Specifically, we have the following:
Case 1: Let $n_{k-i} \in \{0, \ldots, 9\} \setminus \{a_0\}$ for $2 \leq i \leq H - 1$. Then, we may take

$$n^* = 4,9 \cdot 10^k + \sum_{j=k-H+1}^{k-2} n_j 10^j\,.$$

Case 2: Let $n_{k-i} \in \{0, \ldots, 9\} \setminus \{a_0\}$ for $3 \leq i \leq H - 1$. Then, we may take

$$n^* = 5,09 \cdot 10^k + \sum_{j=k-H+1}^{k-3} n_j 10^j\,.$$

Case 3: The choices for n^* in cases 2 and 3 are both possible. □

Convention In the sequel, we have many estimates and definitions containing positive constants $C_1, C_2, \ldots$ (actually powers $(\log X)^{C_i}$). The C_i must satisfy certain conditions, which will be described. However, it will always be possible to choose the C_i, such that the $\min_i C_i$ is arbitrarily large. An estimate containing $O\left(D(x)(\log X)^{-\mathcal{A}}\right)$ ($D(x)$ a certain function of X) means that $A > 0$ may be taken arbitrarily large if $\min_i C_i$ is sufficiently large.

Definition 2.5 We define X by $2X \leq N_0 < 20X$. We then define

$$Int(N_0) = \left[\frac{N_0 - n_H^*}{2} - \frac{X}{8}, \ \frac{N_0 - n_H^*}{2} + \frac{X}{8} \right], \tag{12}$$

$$S_{c_0}(\theta) := \frac{1}{\gamma} \sum_{\substack{p \in Int(N_0) \\ p=[n^{1/\gamma}]}} (\log p)^{1-\gamma} e(p\theta) \, .$$

Let $S \subseteq [1, X]$ be a set of positive integers and $v(n)$ be a sequence of real numbers. For the exponential sum

$$E(\theta) := \sum_{n \in S} v(n) e(n\theta) \tag{13}$$

we define

$$J(E) := \frac{1}{X} \sum_{1 \leq a \leq X} E\left(\frac{a}{X}\right) S_{c_0}^2\left(\frac{a}{X}\right) e\left(-N_0 \frac{a}{X}\right) \tag{14}$$

$$J(E, \tau) := \frac{1}{X} \sum_{\frac{a}{X} \in \mathcal{T}} E\left(\frac{a}{X}\right) S_{c_0}^2\left(\frac{a}{X}\right) e\left(-N_0 \frac{a}{X}\right)$$

for a subset $\mathcal{T} \subseteq [0, 1]$ and the mean value

$$M(E) := \sum_{\substack{(m, p_2, p_3) \\ m \in S, p_2, p_3 \in \mathbb{P}_{c_0}, p_i \in Int(N_0) \\ m+p_2+p_3=N_0}} p_2^{1-\gamma} p_3^{1-\gamma} (\log p_2)(\log p_3) v(m) \, ,$$

The evaluation of $J(E)$ is also called the *a-variable* circle method.

Lemma 2.6 *We have $J(E) = M(E)$.*

Proof This follows by orthogonality. □

Instead of w_n, $S_d(z)$ from Definition 2.1, we now consider the expression given in

Definition 2.7 We determine H by $10^H = \lceil (\log X)^{C_1} \rceil$. Let $n^* = n_H^*$, which has been constructed in Lemma 2.4, $\mathcal{B}^*$ as in Lemma 2.4 and $\mathcal{A}^* = \mathcal{A} \cap \mathcal{B}^*$.
We let

$$w_n^* := 1_{\mathcal{A}^*}(n) - \frac{\kappa_{\mathcal{A}} \# \mathcal{A}^*}{\# \mathcal{B}^*}$$

$$S_d^*(z) := \sum_{\substack{n < X/d \\ p|n \Rightarrow p > z}} w_{nd}^* = S(\mathcal{A}_d^*, z) - \frac{\kappa_{\mathcal{A}} \# \mathcal{A}^*}{\# \mathcal{B}^*} S(B_d^*, z)$$

$1_{\mathcal{A}_d^*}(n)$ is called the $\mathcal{A}$-part of w_n^*.
$S(\mathcal{A}_d^*, z)$ is called the $\mathcal{A}$-part of $S_d^*(z)$.

The $\mathcal{B}$-parts are defined analogously. The analogue of Lemma 2.2 leads to an identity.

$$S_{\mathcal{A}^* \cap \mathbb{P}}(\theta) = \sum_j E_j(\theta)\,,$$

where the exponential sums $E_j(\theta)$ are extended over integers $n = p_1 \dots p_l$, defined by linear inequalities to be satisfied by the vector

$$\left(\frac{\log p_1}{\log X}\,, \dots, \frac{\log p_l}{\log X} \right) .$$

We also define the exponential sums

$$S_d^*(z,\theta) = \sum_{\substack{n \in \mathcal{B}^* \\ p \mid n \,\Rightarrow\, p > z}} w_{nd}^* = S(\mathcal{A}_d^*, z, \theta) - \frac{\kappa_{\mathcal{A}} \# \mathcal{A}^*}{\# \mathcal{B}^*} S(\mathcal{B}_d^*, z, \theta)\,.$$

For the evaluation of the sums $J(E)$ from (14) by the *a-variable* circle method, we partition the set $\{\frac{a}{X} \ : \ 1 \le a \le X\}$ into the two subsets of the major arcs and the minor arcs.

Definition 2.8 We set $Q_0 = (\log X)^3$. For $q \le X$, $1 \le c \le q$, $(c,q) = 1$ and $L \in [1, \infty)$, we set

$$I_{c,q}(L) := \left[\frac{c}{q} - q^{-1} X^{-1} L,\, \frac{c}{q} + q^{-1} X^{-1} L \right] .$$

We let $L_0 = (\log X)^{C_1}$, $L_1 = X^{1/5}$. The major arcs $\mathcal{M}$ are defined as

$$\mathcal{M} := \bigcup_{\substack{q \le Q_0 \\ (c,q)=1}} I_{c,q}(L_0)\,.$$

The minor arcs $\mathfrak{m}$ are defined as

$$\mathfrak{m} := [0,1] \setminus \mathcal{M}\,.$$

For the evaluation of $S_{c_0}\left(\frac{c}{q} + \xi\right)$, we apply the approach of Balog and Friedlander [1].

We now obtain a *local version* of the result of Maynard [12]. Instead of considering the sets C_d with $C = \mathcal{A}$ and $\mathcal{B}$ appearing in the Buchstab recursion in Lemma 2.2, we now consider the sets C_d with

$$C^* = C_{q,s} := \{ m \in C \ : \ m \equiv s \bmod q \}\,,$$

where $C = \mathcal{A}^*$ or $\mathcal{B}^*$ as defined in Definition 2.7.
We carry out the type I and type II estimates closely following Maynard [12], obtaining the contributions to $J(E)$ of the major arcs of the a-variable circle method. These type II estimates are based on the b-variable circle method:

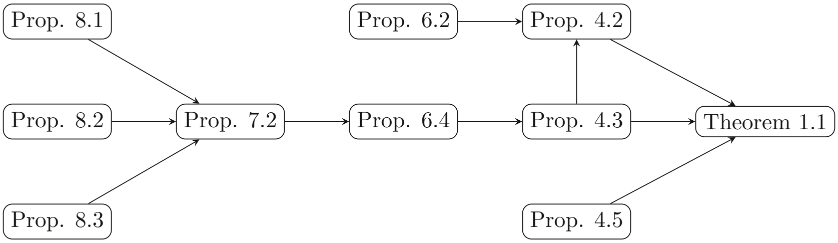
Let R be a subset of $\mathcal{B}^*$, $\mathcal{J} = \mathcal{A}^* \cap R$. Then, we have

$$S_{\mathcal{J}}(\theta) = \frac{1}{X} \sum_{1 \le b \le X} S_{\mathcal{A}}\left(\frac{b}{X}\right) S_R\left(-\left(\frac{b}{X} - \theta\right)\right) .$$

The minor arcs of the a-variable circle method finally are treated by estimates of large sieve type and by estimates of exponential sums over prime numbers.

3 Structure of the Paper

In Sect. 4, we carry out the sieve decomposition of the local version of Maynard [12] involving the exponential sums instead of counting functions and the sets $\mathcal{A}^*$ and $\mathcal{B}^*$ contained in short intervals.
We shall reduce the proof of Theorem 1.1 to the proof of three Propositions:
Proposition 4.2 our type I estimate, Proposition 4.3 our type II estimate, and Proposition 4.5 in which the $\mathcal{A}$-part is estimated trivially. All these propositions contain convolutions of the sums appearing in the Buchstab iterations with the Piatetski-Shapiro sums.
In Sect. 6, we reduce Propositions 4.2, 4.3, and 4.5 to the local version of Maynard's result, Propositions 6.2 and 6.4, which do not involve the Piatetski-Shapiro sum.
Proposition 6.2 is handled by a method from combinatorial sieve theory, replacing the Möbius function by functions with smaller support and Fourier analysis to fix locations and residue classes.
The proof of Proposition 4.3 is carried out by the classical circle method.
In Sect. 7, the ranges of summation are partitioned in small boxes.
These are now handled by the b-variable circle method, closely following Maynard [12]. The dependency graph between the main statements is as follows:



4 Sieve Decomposition and Proof of Theorem 1.1

We now describe the modification of Maynard's method of sieve decomposition and the reduction of the proof of Theorem 1.1 to the proof of three propositions, Propositions 4.2, 4.3, and 4.5, in which the various types of information are used: type I information in Proposition 4.2, type II information in Proposition 4.3, whereas in Proposition 4.5 neither type I nor type II information are used.

Definition 4.1 Let $\eta \in (0,1)$. Let $v(n,\eta)_{n\in\mathcal{S}}$ be a family of sequences of real numbers, indexed by the parameter η, $\mathcal{S}$ finite. The family of exponential sums

$$E(\theta;\eta) := \sum_{n\in\mathcal{S}} v(n,\eta)e(n\theta)$$

is called *negligible*, if

$$\lim_{\eta\to 0}\limsup_{k\to\infty}\frac{|J(E)|\log X}{(\#\mathcal{A}^*)X} = 0.$$

The term "*negligible*" will also be applied to an individual exponential sum $E(\theta)$ of the family $E(\theta,\eta)$.

Proposition 4.2 (Sieve Asymptotic Terms) *Let $\epsilon > 0$, $0 < \eta_0 \le \theta_2 - \theta_1$, $l = l(\eta_0)$ be fixed, where*

$$\theta_1 = \frac{9}{25} + 2\epsilon \ \text{ and } \ \theta_2 = \frac{17}{40} - 2\epsilon.$$

Let $\mathcal{L}$ be a set of $O_{\eta_0}(1)$ affine linear functions, $L : \mathbb{R}^l \to \mathbb{R}$. Let

$$E_0 := E_0(\theta,\eta_0) = \sideset{}{^\sim}\sum_{X^{\eta_0}\le p_1\le\cdots\le p_l} S^*_{p_1\cdots p_l}(X^{\eta_0},\theta)\ ,$$

where $\sum^\sim$ indicates that the summation is restricted by the condition

$$L\left(\frac{\log p_1}{\log X},\ldots,\frac{\log p_l}{\log X}\right) \ge 0\,,$$

for all $L \in \mathcal{L}$.
Then, E_0 is negligible.

Proposition 4.3 (Type II Terms) *Let $l = l(\eta_0)$, $\theta_1, \theta_2, \mathcal{L}$ be as in Proposition 4.2, and let $\mathcal{I} = \{1,\ldots,l\}$ and $j \in \{1,\ldots,l\}$,*

$$E_1(\theta, \eta_0) := \sum^{\sim}_{\substack{X^{\eta_0} \le p_1 \le \cdots \le p_l \\ X^{\theta_1} \le \prod_{i \in I} p_i \le X^{\theta_2} \\ p_1 \cdots p_l \le X/p_j}} S^*_{p_1 \cdots p_l}(p_j, \theta) ,$$

$$E_2(\theta, \eta_0) := \sum^{\sim}_{\substack{X^{\eta_0} \le p_1 \le \cdots \le p_l \\ X^{1-\theta_2} \le \prod_{i \in I} p_i \le X^{1-\theta_1} \\ p_1 \cdots p_l \le X/p_j}} S^*_{p_1 \cdots p_l}(p_j, \theta) ,$$

where $\sum^{\sim}$ *indicates the same restriction as in Proposition 4.2.*
Then, E_1 *and* E_2 *are negligible.*

Definition 4.4 The Buchstab function ω is defined by the delay-differential equation

$$\omega(u) = \frac{1}{u} , \ 1 \le u \le 2 ,$$

$$\omega'(u) = \omega(u-1) - \omega(u) , \ \ u > 2 .$$

For $\vec{p} = (p_1, \ldots, p_l)$, p_i primes for $1 \le i \le l$, let

$$Log(\vec{p}) = \left(\frac{\log p_1}{\log X}, \ldots, \frac{\log p_l}{\log X} \right) .$$

Let $\mathcal{C}$ be a set of $O(1)$ affine linear functions. Let the polytope $\mathcal{R}$ be defined by

$$\mathcal{R} = \{(u_1, \ldots, u_l) \in [0,1]^l \ : \ L(u_1, \ldots, u_l) \ge 0 \ \text{ for all } L \in \mathcal{L}\} .$$

Let

$$\Pi(\vec{p}) = p_1 \cdots p_l ,$$

$$\mathfrak{S}(N_0) = \prod_{p \nmid N_0} \left(1 + \frac{1}{(p-1)^3} \right) \prod_{p \mid N_0} \left(1 - \frac{1}{(p-1)^2} \right) .$$

Proposition 4.5 *Let* $l \in \mathbb{N}$, $\delta > 0$,

$$z \ : \ [0,1]^l \to [\delta, 1-\delta], \ \vec{u} = (u_1, \ldots, u_l) \to z(\vec{u}) = z(u_1, \ldots, u_l)$$

be continuous. Let

$$E(\theta) := \sum_{\vec{p}\,:\,Log(\vec{p})\in\mathcal{R}} S(\mathcal{B}^*_{\prod(\vec{p})}, X^{z(Log(\vec{p}))}, \theta)\,.$$

Then,

$$J(E(\theta)) = \frac{X(\#\mathcal{B}^*)}{4\log X}\mathfrak{S}_0(N_0)\int\cdots\int_{\mathcal{R}}\frac{\omega(1-u_1-\cdots-u_l)}{u_1\cdots u_l z(u_1,\ldots,u_l)}du_1\ldots du_l(1+o(1)).$$

We now replace each Buchstab recursion in the counting function version in Maynard [12] by its exponential sum version.
In each step in [11], in which Proposition 4.2 or Proposition 4.3 is applied, we deal with a negligible sum, which does not influence the asymptotics.
Each of the nine applications of Proposition 4.5 leads to a change of the estimate proportional to an integral I_i $(1 \leq i \leq 9)$. The main result of Theorem 1.1 is obtained by the estimate

$$I_1 + \cdots + I_9 \leq 0.996 < 1\,.$$

(Maynard [12] has included a Mathematica® file detailing this computation with his article on arxiv.org).
This computation is also applicable to our sequence of Buchstab recursions in the exponential sum version and leads to the proof of Theorem 1.1.

5 Fourier Estimates and Large Sieve Inequalities

An important question to be settled is the distribution of the elements with missing digits on congruence classes. Closely linked to that question is the estimate of exponential sums extended over these integers, which also appear in the application of the b-variable circle method.
These exponential sums are defined by the following:

Definition 5.1 Let

$$\mathcal{A}_1 := \left\{\sum_{0\leq i\leq k} n_i 10^i \;:\; n_i \in \{0,\ldots,9\}\setminus\{a_0\}, k \geq 0\right\}\,.$$

For Y an integral power of 10, we write

$$F_Y(\theta) := Y^{-\log 9/\log 10}\left|\sum_{n<Y} 1_{\mathcal{A}_1}(n)e(n\theta)\right|\,.$$

From the facts used about $F_Y(\theta)$, we just give Lemma 5.2 (Lemma 10.5 of [12]). For the full list, see [11].

Lemma 5.2 (Large Sieve Estimates) *We have*

$$\sup_{\beta\in\mathbb{R}} \sum_{c\le q} \sup_{|\eta|<\delta} F_Y\left(\frac{c}{q}+\beta+\eta\right) \ll (1+\delta q)\left(q^{27/77}+\frac{q}{Y^{50/77}}\right)$$

$$\sup_{\beta\in\mathbb{R}} \sum_{q\le Q} \sum_{\substack{0<c<q\\(c,q)=1}} \sup_{|\eta|<\delta} F_Y\left(\frac{c}{q}+\beta+\eta\right) \ll (1+\delta Q^2)\left(Q^{54/77}+\frac{Q^2}{Y^{50/77}}\right),$$

and for any integer d, we have

$$\sup_{\beta\in\mathbb{R}} \sum_{\substack{q\le Q\\d|q}} \sum_{\substack{0<c<q\\(c,q)=1}} \sup_{|\eta|<\delta} F_Y\left(\frac{c}{q}+\beta+\eta\right) \ll \left(1+\frac{\delta Q^2}{d}\right)\left(\left(\frac{Q^2}{d}\right)^{27/77}+\frac{Q^2}{dY^{50/77}}\right).$$

6 Local Versions of Maynard's Results

The propositions of Sect. 4 now are reduced to other facts to be proven later.

The proof of all three propositions employs the circle method: the discrete (a-variable) circle method for Propositions 4.2 and 4.3 and the classical continuous variable circle method for Proposition 4.5. Whereas the minor arcs contributions for all three cases are very similar, using estimates of the Piatetski-Shapiro sum due to Balog and Friedlander and of exponential sums over prime numbers, there are major differences in the treatment of the major arcs contributions.

In contrast to Propositions 4.3 and 4.5, where the sifted sets appear as a union of simpler sets, the set considered in Proposition 4.2 is obtained by a modification of the inclusion-exclusion principle appearing in the sieve of Eratosthenes. We recall the relation to the Möbius function $\mu(\cdot)$:

Let $\mathcal{C}$ be a set of integers and $\mathcal{P}$ a set of primes. Then,

$$\begin{aligned} S(\mathcal{C},\mathcal{P},z) &:= \#\{n\in\mathcal{C}\ :\ p\mid n,\ p\in\mathcal{P}\ \Rightarrow\ p>z\}\\ &= \sum_{n\in\mathcal{C}} \sum_{\substack{t|n\\t|P(z)}} \mu(t)\,, \quad \text{with } P(z) := \prod_{\substack{p\le z\\p\in\mathcal{P}}} p\,. \end{aligned}$$

A basic idea in the theory of combinatorial sieves is the replacement of the Möbius function by a function λ, having smaller support.

This is also done in the proof of Proposition 4.2: the Möbius function μ is replaced by two functions λ^+ and λ^-, and we consider the exponential sums

$$S(\mathcal{C}, z, \theta, \lambda) := \sum_{n \in \mathcal{C}} e(n\theta) \left(\sum_{\substack{t \mid n \\ t \mid P(z)}} \lambda(t) \right)$$

for $\lambda = \lambda^+$ and λ^-.

We recall the following result from combinatorial sieve theory:

Lemma 6.1 (Fundamental Lemma 6.3 of [7]) *Let $\kappa > 0$ and $y > 1$. There exist two sets of real numbers*

$$\Lambda^+ = (\lambda_d^+) \ \ \textit{and} \ \ \Lambda^- = (\lambda_d^-)$$

depending only on κ and y with the following properties:

$$\lambda_1^{\pm} = 1 \tag{15}$$

$$|\lambda_d^{\pm}| \le 1\,, \ \ \textit{if}\ 1 \le d < y \tag{16}$$

$$\lambda_d^{\pm} = 0\,, \ \ \textit{if}\ d \ge y$$

and for any integer $n > 1$,

$$\sum_{d \mid n} \lambda_d^- \le 0 \le \sum_{d \mid n} \lambda_d^+ \,. \tag{17}$$

Moreover, for any multiplicative function $g(d)$ with $0 \le g(p) < 1$ and satisfying the dimension conditions

$$\prod_{w \le p < z} (1 - g(p))^{-1} \le \left(\frac{\log z}{\log w} \right)^{\kappa} \left(1 + \frac{\kappa}{\log w} \right)$$

for all $2 \le w < y$, we have

$$\sum_{d \mid P(z)} \lambda_d^{\pm} g(d) = \left(1 + O\left(e^{-s} \left(1 + \frac{\kappa}{\log z} \right)^{10} \right) \right) \prod_{p < z} (1 - g(p))\,,$$

where $P(z)$ denotes the product of all primes $p < z$ and $s = \log y / \log z$. The implied constants depend only on κ.

We then show Proposition 6.2 and finally reduce the proof of Proposition 4.2 to Proposition 6.2 and Proposition 4.3.
We first give the statement and sketch of proof for Proposition 6.2.

Proposition 6.2 *Let* $\epsilon > 0$, $0 < \eta_0 \leq \theta_2 - \theta_1$, $l = l(\eta_0)$ *be fixed. Let* $\mathcal{L}$ *and the summation condition* $\sum^{\sim}$ *be as in Proposition 4.2,* $q \leq Q_0$, $(c, q) = 1$. *Let* $\lambda^{\pm}$ *satisfy the properties of Lemma 6.1 with* $y = X^{(\eta_0^{1/2})}$, *and let* $\lambda^{\pm}(t) = 0$, *if* $(t, 10) > 1$. *Then, we have for* $\lambda = \lambda^{-}$ *or* λ^{+}*:*

$$\sum_{X^{\eta_0} \leq p_1 \leq \cdots \leq p_l}^{\sim} \left(S\left(\mathcal{A}^*_{p_1 \cdots p_l}, X^{\eta_0}, \frac{c}{q}, \lambda \right) - \frac{\kappa_{\mathcal{A}} \# \mathcal{A}^*}{\# \mathcal{B}^*} S\left(\mathcal{B}^*_{p_1 \cdots p_l}, X^{\eta_0}, \frac{c}{q}, \lambda \right) \right)$$

$$= O\left((\# \mathcal{A}^*)(\log X)^{-A} \right) .$$

Proposition 6.2 is proved by the use of exponential sum estimates from Section 5 to the $\mathcal{A}$-part.
The same computation is now carried out for $\mathcal{B}^*$ instead of $\mathcal{A}^*$. One finds that the leading terms in the summation of Proposition 6.2 cancel.
From Proposition 6.2, we now deduce the following:

Lemma 6.3 *Let*

$$E_{0,\mathcal{A}^*,\lambda}(\theta) = \sum_{x^{\eta_0} \leq p_1 \leq \cdots \leq p_l}^{\sim} S(\mathcal{A}^*_{p_1 \cdots p_l}, X^{\eta_0}, \theta, \lambda)$$

$$E_{0,\mathcal{B}^*,\lambda}(\theta) = \sum_{x^{\eta_0} \leq p_1 \leq \cdots \leq p_l}^{\sim} S(\mathcal{B}^*_{p_1 \cdots p_l}, X^{\eta_0}, \theta, \lambda)$$

Then, for $\lambda = \lambda^{-}$ *or* λ^{+}, *we have*

$$\frac{1}{X} \sum_{1 \leq a \leq X} \left(E_{0,\mathcal{A}^*,\lambda}\left(\frac{a}{X}\right) - \kappa_{\mathcal{A}} \frac{\# \mathcal{A}^*}{\# \mathcal{B}^*} E_{0,\mathcal{B}^*,\lambda}\left(\frac{a}{X}\right) \right)$$

$$S_{c_0}^2\left(\frac{a}{X}\right) e\left(-N_0 \frac{a}{X}\right) = O\left(\# \mathcal{A}^* X (\log X)^{-A}\right).$$

From Lemma 6.1, we conclude

$$\limsup_{k \to \infty} \frac{|J(E(\theta, \eta^*, \mu))| \log X}{|\mathcal{A}^*| X} < \epsilon \ , \quad \text{for } \eta \geq \eta_0 \ .$$

We modify the analysis given in [12], p. 156, to pass from X^{η_0} to $X^{\theta_2 - \theta_1}$.

Given a set $\mathcal{C}$ and an integer d, we let

$$T_m(\mathcal{C}; d, \theta) := \sum_{\substack{X^\eta \le p'_m \le \cdots \le p'_1 \le X^\theta \\ d p'_1 \cdots p'_m \le X^{\theta_1}}} S(\mathcal{C}_{p'_1 \cdots p'_m}, X^\eta, \theta)$$

$$U_m(\mathcal{C}; d, \theta) := \sum_{\substack{X^\eta \le p'_m \le \cdots \le p'_1 \le X^\theta \\ d p'_1 \cdots p'_m \le X^{\theta_1}}} S(\mathcal{C}_{p'_1 \cdots p'_m}, p'_m X^\eta, \theta)$$

$$V_m(\mathcal{C}; d, \theta) := \sum_{X^\eta < p'_m \le \cdots \le p'_1 \le X^\theta} S(\mathcal{C}_{p'_1 \cdots p'_m}, p'_m, \theta).$$

Buchstab's identity shows that

$$U_m(\mathcal{C}; d, \theta) = T_m(\mathcal{C}; d, \theta) - U_{m+1}(\mathcal{C}; d, \theta) - V_{m+1}(\mathcal{C}; d, \theta)$$

The T_m-terms are now handled by Lemma 6.3, whereas the V_m-terms are reduced to Proposition 4.3.
Proposition 4.3 is deduced from the following:

Proposition 6.4 (Type II Terms, Local Version) *Let ϵ, η_0, l, $\mathcal{L}$, $\tilde{\Sigma}$, q, c, t be as in Proposition 6.2. Then, we have*

$$\sum_{X^{\eta_0} \le p_1 \le \cdots \le p_l}^{\sim} \left(S\left(\mathcal{A}^*_{p_1 \ldots p_l}, X^{\eta_0}, \frac{c}{q}, \lambda\right) - \frac{\kappa_{\mathcal{A}} \# \mathcal{A}^*}{\# \mathcal{B}^*} S\left(\mathcal{B}^*_{p_1 \ldots p_l}, X^{\eta_0}, \frac{c}{q}, \lambda\right)\right)$$

$$= O\left((\#\mathcal{A}^*)(\log X)^{-A}\right) .$$

As in the proof of Proposition 4.2, Proposition 4.3 is deduced from Proposition 6.4 by replacing the variable factors $e\left(n\left(\frac{c}{q}+\xi\right)\right)$ by $e(n_0\xi)e\left(n\frac{c}{q}\right)$ with $n_0 \in \mathcal{B}^*$.
Proposition 4.5 is proven by the classical circle method. Furthermore, one uses the connection between the Buchstab function and the number of integers free of small prime factors as well as the equidistribution of these integers on residue classes mod$q$.

7 Sieve Asymptotics for Local Version of Maynard

We now shall reduce Proposition 6.4 to Proposition 7.2 stated below. The range of the summation in Proposition 6.2 is defined by several sets of linear forms of the vectors

$$Log^{(1)}(n) := \left(\frac{\log \tilde{p}_1}{\log X}, \ldots, \frac{\log \tilde{p}_v}{\log X}\right) , \tag{18}$$

where $n = \tilde{p}_1 \cdots \tilde{p}_v$.

(1) The linear forms from $\mathcal{L}$, included by $\sum^{\sim}$.
(2) The linear forms related to the conditions

$$n \in \mathcal{A}^*_{p_1 \cdots p_v} , \quad p \mid n \;\Rightarrow\; p > p_j \; .$$

(3) The linear forms related to the chain of inequalities

$$p_1 \leq \cdots \leq p_v \; .$$

(4) The linear forms analogous to (3) related to the other prime factors.

All the linear forms from (1) to (4) now form a set

$$\mathcal{L}^* := \bigcup_v \tilde{\mathcal{L}}(v) , \tag{19}$$

where v denotes the total number of prime factors.
To be able to describe the set of integers satisfying these linear inequalities by a polytope, we pass from the vector $Log^{(1)}$ in (19) to the vector

$$Log^{(2)}(n) := \left(\frac{\log \tilde{p}_1}{\log n}, \ldots, \frac{\log \tilde{p}_v}{\log n}\right) . \tag{20}$$

Obviously,

$$Log^{(2)}(n) \in Q_v(\eta) := \{(x_1, \ldots, x_n) \in \mathbb{R}^v ,\; \eta \leq x_1 \leq \cdots \leq x_v, x_1 + \cdots + x_v = 1\}.$$

By a closed convex polytope in $\mathbb{R}^v$, we mean a region R defined by a finite number of non-affine linear inequalities in the coordinates (equivalently, this is the convex hull of a finite set of points in $\mathbb{R}^v$).
Given a closed convex polytope $R \subseteq Q_l(\eta)$, we let

$$1_R(n) := \begin{cases} 1 , & \text{if } n = p_1 \cdots p_v \text{ with } Log^{(2)}(n) \in R^v \\ 0 , & \text{otherwise} , \end{cases}$$

We now let $\overline{R} \subseteq [\eta, 1]^{v-1}$ denote the projection of R onto the first $v - 1$ coordinates (which is also a convex polytope).

Definition 7.1 Fix $\eta > 0$ and let $v \in \mathbb{Z}$ satisfy $1 \leq v \leq 2/q$. Let $\gamma > 0$ and let

$$\vec{a} := (a_1, a_2, \dots, a_{v-1})$$

be a sequence of real numbers. Let

$$\vec{p} := (p_1, \dots, p_v)$$

be an l-tuplet of prime numbers, $\prod(\vec{p}) = p_1 \cdots p_v$. Then, we define

$$\mathcal{C}(\vec{a}, \gamma) := \left\{\vec{p} = (p_1, \dots, p_v) \ : \ p_j \in (X^{a_j}, X^{a_j+\gamma}), 1 \leq j \leq v_1, \prod(\vec{p}) \in \mathcal{B}^*\right\}$$

and

$$\mathcal{C}(\vec{a}, \gamma, q, s) := \left\{\vec{p} \in \mathcal{C}(\vec{a}, \gamma) \ : \ \prod(\vec{p}) \equiv s \bmod q\right\} .$$

The sequence $\vec{a}$ and the box $\mathcal{C}(\vec{a}, \gamma)$ are called *normal*, if $a_j + \gamma < a_{j+1}$, for $1 \leq j \leq v - 2$.

Proposition 7.2 *Let $\mathcal{C}(\vec{a}, \gamma)$ be as defined in Definition 7.1, $\gamma = (\log X)^{-C_3}$ for $C_3 > 0$ fixed. Let $q \leq Q_0$, $(s, q) = 1$. Then,*

$$\sum_{n \in \mathcal{C}(\vec{a}, \gamma, q, s)} w_n = O\left(\left(\frac{1}{\phi(q)} \sum_{n \in \mathcal{C}(\vec{a}, \gamma, q, s)} 1\right) (\log X)^{-A}\right) .$$

Proof of Proposition 6.4 assuming Proposition 7.3

Definition 7.3 Let $\delta_0 := (\log X)^{-C_3}$. We cover $[\eta, 1]^{v-1}$ by $O_\eta(\delta_0^{-(v-1)})$ disjoint hypercubes $\mathcal{C}(\vec{a}, \delta_0)$. We partition the $\vec{a} \in \overline{R}$ into two disjoint sets:

$$\mathcal{Y}_1 := \{\vec{a} \in \overline{R} \ : \ \mathcal{C}(\vec{a}, \delta_0) \subseteq \overline{R}\}$$

$$\mathcal{Y}_2 := \{\vec{a} \in \overline{R} \ : \ \mathcal{C}(\vec{a}, \delta_0) \cap bd\overline{R} \neq \emptyset\} .$$

Since the set $\mathcal{L}^*$ of linear forms defining R imply

$$\frac{\log p_i}{\log n} \neq \frac{\log p_j}{\log n} \ , \ \text{for } i \neq j \ ,$$

$\mathcal{C}(\vec{a}, \gamma) \subseteq R$ implies that $\mathcal{C}(\vec{a}, \gamma)$ is normal.

We have thus by Proposition 7.2 that

$$\sum_{\vec{a}\,:\,\mathcal{C}(\vec{a},\delta_0)\subseteq \overline{R}} \ \sum_{n\in\mathcal{C}(\vec{a},\delta)} e\left(n\frac{c}{q}\right) w_n^* = \sum_{\substack{s \bmod q\\ (s,q)=1}} e\left(\frac{sc}{q}\right) \sum_{\vec{a}\,:\,\mathcal{C}(\vec{a},\delta_0)} \ \sum_{n\in\mathcal{C}(\vec{a},\delta_0,q,s)} w_n^*$$

By the prime number theorem for short intervals and arithmetic progressions, we have for any s_0 with $(s_0, q) = 1$

$$\sum_{n\in\mathcal{C}(\vec{a},\delta_0,q,s)} 1 = \left(\sum_{n\in\mathcal{C}(\vec{a},\delta_0,q,s_0)} 1 \right) \left(1 + O(\log X)^{-A}\right) .$$

Thus, we obtain by Proposition 7.2

$$\sum_{\vec{a}\,:\,\mathcal{C}(\vec{a},\delta_0)\in \overline{R}} \ \sum_{n\in C^+(u,\gamma)} e\left(n\frac{c}{q}\right) w_n^* = O\left(\sum_{\vec{a}\,:\,\mathcal{C}(\vec{a},\delta_0)\subseteq \overline{R}} |\mathcal{C}(\vec{a}, \delta_0)| \right) (\log X)^{-A} .$$

For the contribution of $\mathcal{Y}_2$, we estimate the total volume of the $\mathcal{C}(\vec{a}, \delta_0)$ and treat the $\mathcal{A}^*$-part and the $\mathcal{B}^*$-part separately.
Proposition 6.4 thus has been reduced to Proposition 7.2.

8 *b*-Variable Circle Method

In this section, we state propositions needed in the estimate of type II expressions by the b-variable circle method. We then derive Proposition 7.2 from them.

Proposition 8.1 *Fix $\eta > 0$ and let $v \in \mathbb{Z}$ satisfy $1 \le v \le 2/\eta$. Let*

$$\mathcal{C} := \mathcal{C}(\vec{a}, r, q, s)$$

be as in Definition 7.1. Let $q \le Q_0$. Let $\mathcal{M}^{(b)} = \mathcal{M}^{(b)}(C_4)$ be given by

$$\mathcal{M}^{(b)} := \left\{ 0 \le b < X \ : \ \left| \frac{b}{X} - \frac{d}{r} \right| \le \frac{(\log X)^{C_4}}{X} \right\}$$

for some integers d, r with $r \le (\log X)^{C_4}$, $r \mid X$.
Then, if C_4 is chosen sufficiently large,

$$\frac{1}{X}\sum_{\substack{0\le b<X\\ b\in\mathcal{M}}} \mathcal{S}_{\mathcal{A}^*}\left(\frac{b}{X}\right) S_{\mathcal{C}}\left(-\frac{b}{X}\right) - \frac{\kappa_{\mathcal{A}}\#\mathcal{A}^*}{\#\mathcal{B}^*}\,\#\mathcal{C}(\vec{a}, r, q, s) = O\left(\frac{\#\mathcal{A}^*}{(\log X)^A}\right).$$

The implied constants depend on A, but not on η, v, and the a_j.

Proposition 8.2 (Generic Minor Arcs) *Let $\mathcal{C}$ and $\mathcal{M}(C_4)$ be as in Proposition 8.1. Then, there is some exceptional set*

$$\mathcal{E} := \mathcal{E}(\mathcal{C}) \subseteq [0, X]\,, \quad \textit{with } \#\mathcal{E} \le X^{23/40}\,,$$

such that

$$\frac{1}{X}\sum_{\substack{b<X\\ b\notin\mathcal{E}}} \left| S_{\mathcal{A}^*}\left(\frac{b}{X}\right) S_{\mathcal{C}}\left(-\frac{b}{X}\right)\right| = O\left(\frac{\#\mathcal{A}^*}{X^{\epsilon}}\right).$$

The implied constant depends on η but not on the a_j.

Proposition 8.3 (Exceptional Minor Arcs) *Let $\mathcal{C}$ and $\mathcal{M} = \mathcal{M}(C_4)$ be as given in Proposition 8.1. Let $a_1, \ldots, a_{v-1}$ in the definition of $\mathcal{C}(\vec{a}, r, q, s)$ satisfy*

$$\sum_{i\in\mathcal{I}} a_i \in \left[\frac{9}{40}+\frac{\epsilon}{2},\ \frac{16}{25}-\frac{\epsilon}{2}\right] \cup \left[\frac{23}{40}+\frac{\epsilon}{2},\ \frac{16}{25}-\frac{\epsilon}{2}\right]$$

for some $\mathcal{I} \subseteq \{1, \ldots, v-1\}$, and let C_4 be sufficiently large. Let $\mathcal{E} \subseteq [0, X]$ be any set, such that $\#\mathcal{E} \le X^{23/40}$. Then, we have

$$\frac{1}{X}\sum_{\substack{b\in\mathcal{E}\\ b\notin\mathcal{M}}} S_{\mathcal{A}^*}\left(\frac{b}{X}\right) S_{\mathcal{C}}\left(-\frac{b}{X}\right) = O\left(\frac{\#\mathcal{A}^*}{(\log X)^A}\right).$$

The implied constant depends on η but not on the $a_1, \ldots, a_{v-1}$.

Proof of Proposition 7.2
By orthogonality, we have

$$\#(\mathcal{C}\cap\mathcal{A}^*) = \frac{1}{X}\sum_{1\le b\le X} S_{\mathcal{A}^*}\left(\frac{b}{X}\right) S_{\mathcal{C}}\left(-\frac{b}{X}\right).$$

Proposition 7.2 now follows by the partition given by Propositions 8.1, 8.2, and 8.3.

9 *b*-Variable Major Arcs

The principles for the treatment of the major arcs are the same as in [11]. $\mathcal{M}^{(b)}$ is partitioned into three disjoint sets $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ which are defined by the fraction $\frac{d}{r}$ approximating $\frac{b}{X}$.
For two sets $\mathcal{M}_1$, $\mathcal{M}_2$, the $\mathcal{A}^*$- and the $\mathcal{B}^*$-estimates—as in [12]—give negligible results.
In the $\mathcal{A}^*$-estimate for the set $\mathcal{M}_3$, we again succeed to identify the main contributions as originating in the fractions $\frac{d}{r}$ with $r \mid 10$.
We apply the same techniques as in [11], which however are complicated by the appearance of congruences mod$q$. We now give details.
We split $\mathcal{M}^{(b)}$ up as three disjoint sets.

$$\mathcal{M}^{(b)} = \mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3 \, ,$$

where

$$\mathcal{M}_1 := \left\{ b \in \mathcal{M}^{(b)} \; : \; \left| \frac{b}{X} - \frac{d}{r} \right| \le \frac{(\log X)^{C_2}}{X} \text{ for some } d, r \le (\log X)^{C_3}, \; r \nmid X \right\} \, ,$$

$$\mathcal{M}_2 := \left\{ b \in \mathcal{M}^{(b)} \; : \; \frac{b}{X} = \frac{d}{r} + v \text{ for some } d, r \le (\log X)^{C_3}, \; r \mid X, \; 0 < |v| \le \frac{(\log X)^{C_3}}{X} \right\} \, ,$$

$$\mathcal{M}_3 := \left\{ b \in \mathcal{M}^{(b)} \; : \; \frac{b}{X} = \frac{d}{r} + v \text{ for some } d, r \le (\log X)^{C_3}, \; r \mid X \right\} \, .$$

By Lemma 5.2 and recalling X is a power of 10, we have

$$\sup_{b \in \mathcal{M}_1} \left| S_{\mathcal{A}^*}\left(\frac{b}{X}\right) \right| = \#\mathcal{A}^* \sup_{b \in \mathcal{M}_1} F_{10^{k-H}}\left(\frac{b}{X}\right) = O\left(\#\mathcal{A}^* \exp(-(\log X)^{-1/2+\epsilon})\right) \, .$$

Using the trivial bound

$$S_{\mathcal{C}(\vec{a},r,q,s)} = O(X(\log X)^B)$$

and noting that

$$\mathcal{M}_1 \ll (\log X)^{3B} \, ,$$

we obtain

$$\frac{1}{X} \sum_{b \in \mathcal{M}_1} S_{\mathcal{A}^*}\left(\frac{b}{X}\right) S_{\mathcal{C}}\left(-\frac{b}{X}\right) = O\left(\frac{\#\mathcal{A}^*}{(\log X)^A}\right) \, . \tag{21}$$

This gives the result for $\mathcal{M}_1$. We now consider $\mathcal{M}_2$.
For $\vec{p} = (p_1, \dots, p_v)$ we write $\vec{p}_{v-1} = (p_1, \dots, p_{v-1})$, $\prod_{v-1}(\vec{p}) = p_1 \cdots p_{v-1}$:

$$S_{\mathcal{C}} = \sum_{\substack{\vec{p}_{v-1}=(p_1,\dots,p_{v-1}) \\ p_j \in (X^{a_j}, X^{a_j+\gamma})}} \sum_{\substack{\prod(\vec{p}_{v-1})p_v \in \mathcal{B}^* \\ \prod(\vec{p}_{v-1})p_v \equiv s \bmod q}} e\left(\frac{b \prod_{v-1}(\vec{p}) p_v}{X}\right) .$$

We note that if $b \in \mathcal{M}_2$, then

$$\frac{b}{X} = \frac{d}{r} + \frac{c}{X}, \text{ for some integers } b, r, |c| \leq (\log X)^{C_4}, \ (c \text{ is an integer since } r \mid X).$$

We now chose $C_5 > 0$, $C_5 \in \mathbb{Z}$, so large, that - after $C_1, \dots, C_4$ have been chosen - the following considerations are true and set

$$\Delta := \lceil \log X \rceil^{-C_5} .$$

We remark that Δ^{-1} is an integer.
We separate the sum $S_{\mathcal{C}}\left(\frac{b}{X}\right)$ by putting the prime variable p_v in short intervals of length

$$\Delta(\#\mathcal{B}^*)/(p_1 \cdots p_{v-1})$$

and in arithmetic progressions mod$[q, r]$. Thus, we have

$$\left| S_{\mathcal{C}}\left(\frac{b}{X}\right) \right| = \sum_{\vec{p}_{v-1} \,:\, p_j \in (X^{a_j}, X^{a_j+\gamma})} \sum_{p_v < \frac{|\mathcal{B}^*|}{p_1 \cdots p_{v-1}}} e(p_1 \cdots p_{v-1} p_v) .$$

If $mp = j\Delta x + O(\Delta x)$ and $p \equiv u \bmod d$, then we have

$$e\left(mp\left(\frac{d}{r} + \frac{c}{X}\right)\right) = e\left(\frac{dum}{r}\right) e(jcx) + O(\Delta (\log X)^{C_4}) .$$

By the prime number theorem in short intervals and arithmetic progressions, we have

$$\sum_{p \in [j\Delta X/m, (j+1)\Delta X/m]} 1 = E \frac{\Delta X}{m} (1 + O((\log X)^{-A}) ,$$

where $E = 1$ if the system

$$\begin{cases} p \equiv u \bmod r \\ p \equiv s \bmod q , \end{cases}$$

is solvable and $E = 0$ otherwise,
with

$$E\frac{\Delta X}{m}(1+O((\log X)^{-A}) \leq \Delta|\mathcal{B}^*| \sup_{\substack{d\leq(\log X)^C \\ r\leq(\log X)^C}} \sum_{u\equiv l \bmod q} e\left(\frac{dum}{r}\right) \sum_{1\leq j<\Delta^{-1}} e(j\Delta c)\,.$$

We have

$$\sum_{1\leq j<\Delta^{-1}} e(j\Delta c) = e(-c) = -1 = O(1)\,.$$

We finally obtain

$$\frac{1}{X}\sum_{b\in\mathcal{M}_2} S_{\mathcal{A}^*}\left(\frac{b}{X}\right) S_{\mathcal{C}}\left(-\frac{b}{X}\right) = O\left(\frac{\#\mathcal{A}^*}{(\log X)^A}\right)\,,$$

where the implied elements depend on η and γ, but not on the a_j.
Finally, we consider $\mathcal{M}_3$.
For $(d, r) = 1$, we have

$$S_{\mathcal{C}}\left(\frac{d}{r}\right) = \sum_{0\leq u\leq r} e\left(\frac{du}{r}\right) \sum_{\substack{n\in\mathcal{B}^* \\ n\equiv u(\bmod r) \\ n\equiv s(\bmod q)}} 1 = \frac{1}{\phi([q,r])}\left(\sum_{n\in\mathcal{L}} 1\right) \sum_{\substack{0<u<r \\ (u,r)=1 \\ r\equiv s(\bmod (q,r))}} e\left(\frac{du}{r}\right).$$

The solution set of

$$\begin{cases} n \equiv u \bmod r \\ n \equiv s \bmod q\ , \end{cases}$$

is non-empty if and only if for the square-free kernels r_0 of r the solution set of

$$\begin{cases} n \equiv u \bmod r_0 \\ n \equiv s \bmod q\ , \end{cases}$$

is non-empty.
For the exponential sum

$$\sum_{\substack{0<u<r \\ (u,r)=1}} e\left(\frac{du}{r}\right),$$

we have

$$\sum_{g=0}^{\frac{r}{r_0}-1} e\left(\frac{d(s+gr_0)}{r}\right) = e\left(\frac{du}{r}\right)\sum_{g=0}^{\frac{r}{r_0}-1} e\left(\frac{g}{r/r_0}\right) = \begin{cases} 0, \text{ if } r_0 < r \\ 1, \text{ if } r_0 = r \end{cases}$$

We finally obtain

$$S_{\mathcal{C}(\vec{a},r,q,s)}\left(\frac{d}{r}\right) = \sum_{\substack{0<u<r \\ (u,r)=1}} e\left(\frac{du}{r}\right)\frac{\phi(q)}{\phi([q,r])}\sum_{n\in\mathcal{C}(\vec{a},r,q,s)} 1_{\mathcal{C}(\vec{a},r,q,s)}(u)\left(1+O((\log X)^{-A})\right)$$

$$= \frac{\phi(q)}{\phi([q,r])}\sum_{n\in\mathcal{C}(\vec{a},r,q,s)} 1 \sum_{\substack{0<u<r \\ (u,r)=1}} e\left(\frac{du}{r}\right)\left(1+O((\log X)^{-A})\right)$$

$$= \mu(r)\frac{\phi(q)}{\phi([q,r])}\sum_{n\in\mathcal{C}(\vec{a},r,q,s)} 1\,.$$

Since $\mu(r) = 0$ for $r \mid 10^k$, unless $r \in \{1, 2, 5, 10\}$, the estimate can easily be concluded.

10 Generic Minor Arcs

In this section, we establish Proposition 8.2 and obtain some bounds on the exceptional set $\mathcal{E}$ by using the estimates of Lemma 5.2.

Lemma 10.1 *Let $\mathcal{C} = \mathcal{C}(\vec{a}, \gamma, q, s)$ as in Definition 7.1. We have that*

$$\#\left\{0 \le b < X \;:\; \left|S_c\left(\frac{b}{X}\right)\right| \sim \frac{X}{C}\right\} \ll \frac{C^2|\mathcal{C}|}{X}\,.$$

Proof We have

$$\sum_{b\,:\,\left|S_{\mathcal{C}}\left(\frac{b}{X}\right)\right|^2 \ge \frac{|\mathcal{C}|^2}{10C^2}} \left|S_{\mathcal{C}}\left(\frac{b}{X}\right)\right|^2 \ge \frac{|\#\mathcal{C}|^2}{10C^2}\,\#\left\{b \;:\; \left|S_{\mathcal{C}}\left(\frac{b}{X}\right)\right| \ge \frac{\#\mathcal{C}}{10C}\right\}\,.$$

Thus,

$$\#\left\{b \;:\; |S_{\mathcal{C}(\vec{a},\gamma)}| \ge \frac{|\mathcal{C}|}{10C}\right\} \le \frac{10C^2}{X^2}\sum_{b\le X}\left|S_{\mathcal{C}}\left(\frac{b}{X}\right)\right|^2 = \frac{10C^2}{X^2}\,X|\mathcal{C}|\,,$$

the last identity following Parseval's equation. □

Lemma 10.2 *Let*

$$\mathcal{E} := \left\{ 0 \le b \le X \ : \ F_X\left(\frac{b}{X}\right) \ge \frac{1}{X^{23/80}} \right\} .$$

Then,

$$\#\mathcal{E} \ll X^{23/40-\epsilon} ,$$

$$\sum_{b\in\mathcal{E}} F_X\left(\frac{b}{X}\right) \ll X^{23/80-\epsilon} ,$$

and

$$\frac{1}{X}\sum_{\substack{b<X\\ b\notin\mathcal{E}}} \left| F_X\left(\frac{b}{X}\right) S_{\mathcal{C}}\left(-\frac{b}{X}\right)\right| \ll \frac{1}{X^{\epsilon}} .$$

Proof The first bound on the size of $\mathcal{E}$ follows from Section 5. For the second bound, we see from Lemma 5.2 that

$$\begin{aligned}\sum_{b\in\mathcal{E}} F_X\left(\frac{b}{X}\right) &\ll \sum_{\substack{j\ge 0\\ 2^j\le X^{23/80}}} \#\left\{0 \le b < X \ : \ F_X\left(\frac{b}{X}\right) \sim 2^{-j}\right\} \\ &\ll \sum_{\substack{j\ge 0\\ 2^j\le X^{23/80}}} 2^{(235/154-1)j} X^{59/433} \ll X^{59/433+(23\times 235)/(80\times 154)-23/80} ,\end{aligned}$$

and so the calculation above gives the result.

It remains to bound the sum over $b \notin \mathcal{E}$. We divide the sum into $O((\log X)^2)$ subsums, where we restrict to these b, such that

$$F_X\left(\frac{b}{X}\right) \sim \frac{1}{B} \quad \text{and} \quad \left|S_{\mathcal{C}}\left(\frac{b}{X}\right)\right| \sim \frac{|\mathcal{C}|}{C}$$

for some $B \ge X^{23/80}$ and $C \le X^2$ (terms with $C > X^2$ make a contribution O(1/X)). This gives

$$\frac{1}{X}\sum_{\substack{b<X\\ b\notin\mathcal{E}}} \left| F_X\left(\frac{b}{X}\right) S_{\mathcal{C}}\left(-\frac{b}{X}\right)\right|$$

$$\ll \sum_{\substack{X^{23/80} \leq B \\ 1 \leq C \leq X^2}} \frac{(\log X)^2}{X} \sum_{\substack{b<X \\ F_X\left(\frac{b}{X}\right) \asymp \frac{1}{B} \\ S_{\mathcal{C}}\left(-\frac{b}{X}\right) \sim \frac{X}{C}}} \left| F_X\left(\frac{b}{X}\right) S_{\mathcal{C}}\left(-\frac{b}{X}\right) \right| + \frac{1}{X^2} \, .$$

□

11 Exceptional Minor Arcs

In [12] in the treatment of the exceptional minor arcs, methods from the geometry of numbers are of great importance. The key result, the Bilinear Sum Bound, Lemma 13.1, is derived by them.
In our paper, it is enough to have knowledge of this result.

Lemma 11.1 (Bilinear Sum Bound) *Let $N, M, R \geq 1$ and E satisfy*

$$X^{9/25} \leq N \leq X^{17/40}, \;\; R \leq X^{1/2}, NM \leq 1000X, \;\; \textit{and } E \leq 100 \, \frac{X^{1/2}}{R}$$

$$\textit{and either } E \geq \frac{1}{X} \textit{ or } E = 0.$$

Let $\mathcal{F} := \mathcal{F}(R, E)$ be given by

$$\mathcal{F} := \left\{ b < X \; : \; \frac{b}{X} = \frac{d}{r} + v \, \textit{for some } (d, r) = 1 \textit{ with } r \asymp R, \;\; v = \frac{E}{X} \right\} .$$

Then, for any 1-bounded complex sequences $\alpha_n, \beta_n, \gamma_b$, we have

$$\sum_{b \in \mathcal{F} \cap \mathcal{E}} \sum_{\substack{n \sim N \\ m \sim M}} \alpha_n \beta_m \gamma_b e\left(-\frac{bnm}{X}\right) \ll \frac{X(\log X)^{O(1)})}{(R+E)^{\epsilon/10}} \, .$$

Proof This is Lemma 13.1 of [12]. □

We now derive Proposition 8.3 from Lemma 13.1.

Proof of Proposition 8.3
By symmetry, we may assume that $\mathcal{I} = \{1, \ldots, l_1\}$ for some $l_1 < l$. By Dirichlet's theorem on diophantine approximation, any $b \in [0, X]$ has a representation

$$\frac{b}{X} = \frac{d}{r} + \nu$$

for some integers $(d, r) = 1$ with $r \leq X^{1/2}$ and some real $|\nu| \leq 1/X^{1/2}r$.
Thus, we can partition $[0, X]$ into $O((\log X)^2)$ sets $\mathcal{F}(R, E)$ as defined by Lemma 10.1 for different parameters R, E satisfying

$$1 \leq R \leq X^{1/2} \text{ and } E = 0 \text{ or } \frac{1}{X} \leq E \leq \frac{100X^2}{R} .$$

Moreover, if $b \notin \mathcal{M}^{(b)}$, then $b \in \mathcal{F} = \mathcal{F}(R, E)$ for some R, E with

$$R + E \geq (\log X)^{C_3} .$$

Thus, provided C_3 is sufficiently large, we see that it is sufficient to show that

$$\frac{1}{X}\left|\sum_{b\in\mathcal{F}\cap\mathcal{E}} S_{\mathcal{A}}\left(\frac{b}{X}\right) S_{\mathcal{C}_{q,s}}\left(-\frac{b}{X}\right)\right| \ll \frac{\#\mathcal{A}}{(R+E)^{\epsilon/20}} . \tag{22}$$

Recalling Definition 7.1

$$\mathcal{C}_{q,s} := \{\vec{p} = (p_1, p_2, \ldots, p_l) \ : \ p_i \in I_i \ , \Pi(\vec{p}) \equiv s \ (\mathrm{mod} q)\},$$

let

$$\mathcal{C}^{(I)} := \bigtimes_{j\in\mathcal{I}} I_{i_j} \ , \ \mathcal{C}^{(II)} := \bigtimes_{j\notin\mathcal{I}} I_{i_j} \ ,$$

such that

$$n \in \mathcal{C}_i^{(I)} \ \Rightarrow \ X^{9/25} \leq n \leq X^{17/40}.$$

We have (with $t^{-1}t \equiv 1 \ (\mathrm{mod} q)$):

$$\mathcal{C}_{q,s} = \bigcup_{\substack{t \mod q\\(t,q)=1}} (\mathcal{C}_{q,t}^{(I)} \times \mathcal{C}_{q,t^{-1}s}^{(II)}),$$

and thus,

$$\begin{aligned}&\frac{1}{X}\sum_{n\in\mathcal{E}} S_{\mathcal{A}}\left(\frac{b}{X}\right) S_{\mathcal{C}_{q,s}}\left(-\frac{b}{X}\right)\\ &\ll \sum_{\substack{t \bmod q\\(t,q)=1}} \ \sum_{b\in\mathcal{F}\cap\mathcal{E}} S_{\mathcal{A}}\left(\frac{b}{X}\right) \sum_{\substack{n_1\sim N_1\\n_2\sim N_2}} \alpha_{n_1}\beta_{n_2}\, e\left(-\frac{bn_1n_2}{X}\right) ,\end{aligned}$$

where

$$\alpha_{n_1} := \begin{cases} 1\,, & \text{if } n_1 \in \mathcal{C}_{q,t}^{(I)} \\ 0\,, & \text{otherwise}\,, \end{cases} \quad \beta_{n_2} := \begin{cases} 1\,, & \text{if } n_2 \in \mathcal{C}_{q,t^{-1}s}^{(II)} \\ 0\,, & \text{otherwise}\,. \end{cases}$$

Thus, it suffices to show that

$$\frac{1}{X} \sum_{b \in \mathcal{F} \cap \mathcal{E}} S_{\mathcal{A}}\left(\frac{b}{X}\right) \sum_{n \sim N} \alpha_n \sum_{m \sim M} \beta_m\, e\left(-\frac{bnm}{X}\right) \ll \frac{\#\mathcal{A}^*}{(\log X)^A}, \tag{23}$$

Let γ_b be the 1-bounded sequence, satisfying

$$S_{\mathcal{A}}\left(\frac{b}{X}\right) = \#\mathcal{A}\gamma_b F_X\left(\frac{b}{X}\right).$$

After substituting this expression for $S_{\mathcal{A}}$, we see that (23) follows immediately from Lemma 10.1, if the parameter C_3 is chosen sufficiently large.

12 The Ternary Goldbach Problem with a Prime with a Missing Digit, a Piatetski-Shapiro Prime, and a Prime of Another Special Type

We conclude this paper by sketching the proof of a modification of Theorem 1.1. One of the two Piatetski-Shapiro primes is replaced by a prime p of the form

$$p = x^2 + y^2 + 1\,.$$

For details of the proof, we refer to a forthcoming paper.

Theorem 12.1 *Assume the GRH. Let γ^*, γ_0, c_0, a_0 as in Theorem 1.1. Then, each sufficiently large odd integer N_0 can be represented in the form*

$$N_0 = p_1 + p_2 + p_3\,,$$

where the p_i are of the form $p_2 = [n_2^{c_0}]$,

$$p_3 = x_3^2 + y_3^2 + 1\,,$$

for some $x_3, y_3 \in \mathbb{Z}$, and the decimal expansion of p_1 does not contain the digit a_0.

Sketch of Proof We again apply the local version of Maynard's result.
We then follow Hooley [6], introduce the generating function

$$S_Q(\theta) = \sum_{\substack{(x,y)\in\mathbb{Z} \\ x^2+y^2+1=p \text{ prime} \\ \frac{N_0}{2}-\frac{X}{4}<p\le\frac{N_0}{2}-\frac{X}{8}}} r(p-1)e(p\theta)\log p,$$

and use the non-principal Dirichlet character χ mod 4, to obtain the following decompositions:
Let $D := X^{1/2}(\log X)^{-C_0}$, $C_0 > 0$ and

$$S_Q^{(1)}(\theta) := \sum_{d\le D} \chi(d) \sum_{\substack{p\in Int(N_0) \\ p\equiv 1 \bmod d}} e(p\theta)\log p\ ,$$

$$S_Q^{(2)}(\theta) := \sum_{D<d\le\frac{X}{D}} \chi(d) \sum_{\substack{p\in Int(N_0) \\ p\equiv 1 \bmod d}} e(p\theta)\log p\ ,$$

$$S_Q^{(3)}(\theta) := \sum_{\frac{X}{D}<d\le X} \chi(d) \sum_{\substack{p\in Int(N_0) \\ p\equiv 1 \bmod d}} e(p\theta)\log p\ ,$$

with

$$Int(N_0) = \left[\frac{N_0-n_H^*}{2}-\frac{X}{8}, \frac{N_0-n_H^*}{2}+\frac{X}{8}\right].$$

We define

$$J^{(i)}(E) = \frac{1}{X}\sum_{1\le a\le X} E\left(\frac{a}{X}\right) S_{c_0}\left(\frac{a}{X}\right) S_Q^{(i)}\left(\frac{a}{X}\right)$$

and

$$J(E) = J^{(1)}(E) + J^{(3)}(E)\ .$$

For each exponential sum E appearing in the Buchstab recursion in the exponential sum version, we evaluate $J(E)$ asymptotically. $J^{(2)}(E)$, which is of smaller order of magnitude, will be estimated from above.

Acknowledgments M. Th. Rassias: I would like to express my gratitude to Artur Avila and Ashkan Nikeghbali for their essential support throughout the preparation of this work.

References

1. Balog, A., Friedlander, J.: A hybrid of theorems of Vinogradov and Piatetski-Shapiro. Pacific J. Math. **156**(1), 45–62 (1992)
2. Dimitrov, S. I.: The ternary Goldbach problem with prime numbers of a mixed type. Notes Number Theory Discrete Math. **24**(2), 6–20 (2018)
3. Dimitrov, S. I.: Prime triples p_1, p_2, p_3 in arithmetic progressions such that $p_1 = x^2 + y^2 + 1$, $p_3 = [n^c]$. Notes Number Theory Discrete Math. **23**(4), 22–33 (2017)
4. Harman, G.: Prime-detecting sieves, vol. 33. London Mathematical Society Monographs Series. Princeton University Press, Princeton, NJ (2007)
5. Helfgott, H.A.: Major Arcs for Goldbach's Theorem. Ann. of Math. Studies, Princeton, to appear. See also http://arxiv.org/abs/1305.2897v1
6. Hooley, C.: Applications of Sieve Methods to the Theory of Numbers. Cambridge Univ. Press (1976)
7. Iwaniec, H., Kowalski, E.: Analytic Number Theory, Vol. 53. Amer. Math. Soc. Colloq., Providence, RI (2004)
8. Jia, C.H.: On the Piatetski-Shapiro-Vinogradov theorem. Acta Arith. **73**, 1–28 (1995)
9. Kumchev, A.: On the Piatetski-Shapiro-Vinogradov theorem. Journal de Théorie des Nombres de Bordeaux **9**, 11–23 (1997)
10. Maier, H., Rassias, M.Th.: The ternary Goldbach problem with a prime and two isolated primes. Proc. Steklov Inst. Math. **296**, 183–197 (2017). Also translated in Russian and published in Trudy Matematich. Instituta im. V.A. Steklova **296**, 192–206 (2017)
11. Maier, H., Rassias, M.Th.: The ternary Goldbach problem with two Piatetski-Shapiro primes and a prime with a missing digit. Commun. Contemp. Math. (to appear). arXiv:2006.07873
12. Maynard, J.: Primes with restricted digits. Inventiones Mathematicae **217**, 127–218 (2019)
13. Rassias, M.Th.: Goldbach's Problem: Selected Topics. Springer (2017)
14. Teräväinen, J.: The Goldbach problem for primes that are sums of two squares plus one. Mathematika **64**(1), 20–70 (2018)
15. Vinogradov, I.M.: Representation of an odd number as the sum of three primes. Dokl. Akad. Nauk SSSR **15**, 291–294 (1937)

A Note on Harmonious Sets

Yves François Meyer

En hommage respectueux à Jean Bourgain

Abstract A flaw in *algebraic numbers and harmonic analysis*, Elsevier (1972), is corrected.

1 A Wrong Lemma Is Revisited

Harmonious sets are playing a seminal role in the mathematical theory of quasi-crystals. This was showed by Robert Vaughan Moody in [5]. Harmonious sets generalize lattices. The union $\Lambda \cup M$ between two harmonious sets Λ and M is not harmonious in general. The simplest counterexample is $\Lambda = \mathbb{Z}$ and $M = \alpha\,\mathbb{Z}$ when $\alpha \notin \mathbb{Q}$. If Λ is a harmonious set and if F is a finite set, then $F \cup \Lambda$ is still harmonious. This is Theorem II, page 45, Chapter II of [4]. Unfortunately, the proof given in [4] is wrong. This wrong proof is based on Lemma 5, page 45. Lemma 5 is doubtful and its proof is incorrect. Two correct proofs of Theorem II are given in this note. In both proofs, Lemma 5 of [4] is replaced by a weaker result which implies Theorem II. Moreover, the second proof bridges the gap between the theory of harmonious sets with some remarkable results by Nikolaï Bogolyubov and Erling Følner [1, 2].

In this note, G is a locally compact abelian (l.c.a.) group. A function on G will be a real or complex valued function. The given topology on G is denoted by $\mathcal{T}_0$. Four definitions are needed.

Definition 1.1 Let f be a continuous and bounded function on G, and let $0 \le \epsilon < 2$. An ϵ almost period τ of f is defined by

Y. F. Meyer (✉)
CMLA, ENS-Cachan, Université Paris-Saclay, CNRS, France
e-mail: yves.meyer305@orange.fr

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_15

$$\sup_{x \in G} |f(x+\tau) - f(x)| \leq \epsilon \, \|f\|_\infty \,. \tag{1}$$

Definition 1.2 A set $E \subset G$ is relatively dense in G if there exists a compact set K such that $E + K = G$. A set $E \subset G$ is relatively dense with respect to a finite set F if $E + F = G$.

If M is relatively dense in the usual sense, then for every neighborhood V of 0 in G there exists a finite set F such that $M + V$ is relatively dense with respect to F. Indeed, we know that there exists a compact set K such that $M + K = G$ and there exists a finite set F such that $K \subset F + V$. Then, $M + V + F = G$.

Definition 1.3 A continuous and bounded function on G is Bohr almost periodic if for any $\epsilon \in [0, 2)$ the set M_ϵ of ϵ almost period τ of f is relatively dense.

The Bohr compactification $\widetilde{G}$ of G is defined by the three following properties: (a) $\widetilde{G}$ is a compact abelian group, (b) G is a dense subgroup of $\widetilde{G}$, and (c) the Bohr almost periodic functions on G are the restriction to G of the continuous functions on $\widetilde{G}$. The topology on G which is induced by the topology of $\widetilde{G}$ is denoted by $\mathcal{T}$. The topology $\mathcal{T}$ on G is the weakest topology on G for which the Bohr almost periodic functions on G are continuous. In particular, $\mathcal{T}$ is weaker than $\mathcal{T}_0$. If E and F are two subsets of G, $E - F$ denotes the set of all differences $x - y$, $x \in E$, $y \in F$, similarly for $E + F$. A subset E of G is closed for the topology $\mathcal{T}$ if and only if $E = K \cap G$ where $K \subset \widetilde{G}$ is a compact set. Equivalently, E is closed if and only if there exists a Bohr almost periodic function f such that $E = \{x \in G;\, f(x) = 0\}$. If $G = \mathbb{R}$, $\mathbb{N}$ is not closed for the topology $\mathcal{T}$ since any Bohr almost periodic function f vanishing on $\mathbb{N}$ vanishes on $\mathbb{Z}$. Similarly, $\mathbb{Z} \setminus \{0\}$ is not closed for the topology $\mathcal{T}$. However, $\mathbb{Z}$ is closed for the topology $\mathcal{T}$. Similarly, if $G = \mathbb{R}$, the open interval $(0, 1)$ is not open for the topology $\mathcal{T}$. However, if $K \subset G$ is a compact set for the the topology $\mathcal{T}_0$, the topologies $\mathcal{T}_0$ and $\mathcal{T}$ coincide on K. For instance, if $G = \mathbb{R}$, the compact interval $[0, 1]$ is also compact for the topology $\mathcal{T}$.

Lemma 1.1 *If $\Lambda \subset G$ is closed for the topology $\mathcal{T}$ and if $K \subset G$ is a compact set for the topology $\mathcal{T}_0$, then $\Lambda + K$ is closed for the topology $\mathcal{T}$.*

Indeed, we have $\Lambda = L \cap G$ where $L \subset \widetilde{G}$ is a compact set. Since K is contained in the group G, we have $\Lambda + K = (L + K) \cap G$. However, $L + K$ is a compact subset of $\widetilde{G}$ which ends the proof.

Let $\Omega \subset G$ be an open set for the topology $\mathcal{T}$. Then, either Ω is the empty set or Ω is relatively dense. That explains why the open interval $(0, 1)$ is not open for the topology $\mathcal{T}$.

We now reach the definition of harmonious sets. If $0 \leq \epsilon < 2$, the ϵ-dual of a set $\Lambda \subset \mathbb{R}^n$ is the closed set $\Lambda^*_\epsilon \subset \mathbb{R}^n$ defined by

$$\Lambda^*_\epsilon = \{x;\, |\exp(2\pi i x \cdot y) - 1| \leq \epsilon, \;\; \forall y \in \Lambda\}. \tag{2}$$

If $\Lambda \subset \mathbb{R}^n$ is a lattice and if $0 \leq \epsilon < \sqrt{3}$, then Λ_ϵ^* is the dual lattice defined by $\Lambda^* = \{x;\ \exp(2\pi i x \cdot y) = 1,\ \forall y \in \Lambda\}$. The definition of the ϵ-dual extends naturally to the general case of a l.c.a. group Γ. Let us begin with a definition:

Definition 1.4 Let $\mathbb{T}$ be the multiplicative group $\{z \in \mathbb{C};\ |z| = 1\}$. A character χ on a l.c.a. group Γ is a homomorphism $\chi : \Gamma \mapsto \mathbb{T}$.

The continuous characters on a l.c.a. group are playing the role of the trigonometric functions $\chi_y(x) = \exp(2\pi i x \cdot y)$ on $\mathbb{R}^n$. A continuous character χ_y on $\Gamma = \mathbb{R}^n$ is indexed by $y \in \mathbb{R}^n$ and is given by $\chi_y(x) = \exp(2\pi i x \cdot y)$. Similarly, a continuous character χ_y on a l.c.a group Γ is indexed by an element y of the dual group G of Γ. In other terms, $G = \Gamma^*$ is the multiplicative group consisting of all *continuous* characters on Γ. If G is the dual group of Γ, then the dual group of G is Γ.

Definition 1.5 Let Γ be a l.c.a. group and let G be the dual group of Γ. If $0 \leq \epsilon < 2$, the ϵ-dual of a set $\Lambda \subset \Gamma$ is the closed set $\Lambda_\epsilon^* \subset G$ defined by

$$\Lambda_\epsilon^* = \{\chi \in G\,;\, |\chi(y) - 1| \leq \epsilon,\ \forall y \in \Lambda\}. \tag{3}$$

The ϵ-dual Λ_ϵ^* is closed in for the topology $\mathcal{T}$, we have $\Lambda_\epsilon^* = -\Lambda_\epsilon^*$, and

$$\Lambda_\epsilon^* \pm \Lambda_\epsilon^* \subset \Lambda_{2\epsilon}^*. \tag{4}$$

Definition 1.6 Let $F \subset \Gamma$ be a finite set and let $0 < \epsilon < 2$. The Bohr set $B(F, \epsilon) \subset G$ is defined by $B(F, \epsilon) = \{\chi \in G\,;\, |\chi(y) - 1| \leq \epsilon,\ \forall y \in F\}$.

In other terms, $B(F, \epsilon) = F_\epsilon^*$ is the ϵ-dual of the finite set F. Bohr sets are fundamental neighborhoods of 0 for the topology $\mathcal{T}$. Any Bohr set $B(F, \epsilon) \subset G$ is closed for the topology $\mathcal{T}$.

Definition 1.7 A set $\Lambda \subset \Gamma$ is harmonious if for any positive ϵ the set $\Lambda_\epsilon^* \subset G$ is relatively dense in G.

A harmonious set is uniformly discrete [4]: there exists a neighborhood V of 0 such that for $\lambda \in \Lambda, \lambda' \in \Lambda$ and $\lambda \neq \lambda'$ we have $(\lambda + V) \cap (\lambda' + V) = \emptyset$. Let H be the additive subgroup of Γ generated by Λ. Then, Λ is harmonious if and only if for any character $\chi : H \mapsto \mathbb{T}$ and any positive ϵ there exists a continuous character ξ on Γ such that $\sup_{x \in \Lambda} |\chi(x) - \xi(x)| \leq \epsilon$. This is proved in [4], Chapter II.

Our goal is to prove the following theorem:

Theorem 1.1 *If Γ is a l.c.a. group, if $\Lambda \subset \Gamma$ is harmonious, and if $F \subset \Gamma$ is finite, then $\Lambda \cup F$ is still harmonious.*

Two proofs of Theorem 1.1 are given. They both rely on the interplay between the additive properties of relatively dense sets and the topology $\mathcal{T}$.

Here begins the first proof. Theorem 1.1 is an easy consequence of the following proposition:

Proposition 1.1 *Let $M \subset G$ be a relatively dense set and let $\Omega \subset G$ be a neighborhood of* 0 *for the topology $\mathcal{T}$. If M is closed for the topology $\mathcal{T}$, then the intersection $M_2 = (M - M) \cap \Omega$ is relatively dense in G.*

It will be proved that Proposition 1.1 implies Theorem 1.2 which is stronger than Theorem 1.1. Lemma 5 of [4] is a similar statement, but the assumption that M is closed for the topology $\mathcal{T}$ is forgotten. Moreover, the proof of Lemma 5 given in [4] is wrong even if M is closed. Before proving Proposition 1.1, let us give an important corollary.

Theorem 1.2 *Let $\Lambda \subset \Gamma$ be a set of points. Let us assume that for any positive ϵ there exists a finite subset F_ϵ of Λ such that the ϵ-dual of $\Lambda \setminus F_\epsilon$ is relatively dense. Then, Λ is harmonious.*

Theorem 1.2 obviously implies Theorem 1.1. To prove Theorem 1.2, it suffices to show that for any positive ϵ the 2ϵ-dual $\Lambda^*_{2\epsilon}$ of Λ is relatively dense. Let η be a positive real number. The η-dual of $\Lambda \setminus F_\epsilon$ is denoted by $Q_{(\eta,\epsilon)}$, and the 2ϵ-dual of F_ϵ is denoted by Ω_ϵ. On one hand, Ω_ϵ is a neighborhood of 0 for the topology $\mathcal{T}$. On the other hand, $Q_{(\epsilon,\epsilon)}$ is closed for the topology $\mathcal{T}$ and $Q_{(\epsilon,\epsilon)} - Q_{(\epsilon,\epsilon)} \subset Q_{(2\epsilon,\epsilon)}$. By assumption, $Q_{(\epsilon,\epsilon)}$ is relatively dense. We now apply Proposition 1.1 to $Q_{(\epsilon,\epsilon)}$ and conclude that $Q_{(2\epsilon,\epsilon)} \cap \Omega_\epsilon$ is relatively dense. However, $Q_{(2\epsilon,\epsilon)} \cap \Omega_\epsilon = \Lambda^*_{2\epsilon}$ which ends the proof. A similar statement is given by the following theorem:

Theorem 1.3 *Let $\Lambda \subset \Gamma$ be a set of points. Let us assume that for any positive ϵ there exists a finite set F_ϵ and a subset Λ_ϵ of Λ such that $\Lambda \subset F_\epsilon + \Lambda_\epsilon$ and that the ϵ-dual of Λ_ϵ is relatively dense. Then, Λ is harmonious.*

The proof of Theorem 1.3 is the same as the proof of Theorem 1.2.

Let us return to Proposition 1.1. Proposition 1.1 is a corollary of Lemma 1.2. The notations used in Proposition 1.1 are kept here.

Lemma 1.2 *If $M \subset G$ is relatively dense in G and is closed for the topology $\mathcal{T}$, then for every $V \subset G$ which is a neighborhood of* 0 *for $\mathcal{T}_0$, $M - M + V$ is a neighborhood of* 0 *for the topology $\mathcal{T}$.*

Before proving Lemma 1.2, let us show that Lemma 1.2 implies Proposition 1.1. We begin with two simple observations.

Lemma 1.3 *Any non-empty set $\Omega \subset G$ which is open for the topology $\mathcal{T}$ is relatively dense in G.*

Lemma 1.4 *If $M \subset G$ and if there exists a compact set K such that $M + K$ is relatively dense in G, then M is relatively dense in G.*

We are ready to show that Lemma 1.2 implies Proposition 1.1. Let $B(F, \epsilon)$ be a Bohr set and $M_2 = (M - M) \cap B(F, \epsilon)$. Let $V \subset B(F, \epsilon/2)$ be a compact neighborhood of 0 for the topology $\mathcal{T}_0$. Then,

$$(M - M + V) \cap B(F, \epsilon/2) \subset M_2 + V. \tag{5}$$

Indeed if $x \in (M - M + V) \cap B(F, \epsilon/2)$, we have $x = y + z$, $y \in M - M$, $z \in V$ and $x \in B(F, \epsilon/2)$. It implies $y = x - z \in B(F, \epsilon/2) - V \subset B(F, \epsilon)$ and finally $y \in (M - M) \cap B(F, \epsilon) = M_2$ which ends the proof of (5). By Lemma 1.2, $M - M + V$ is a neighborhood of 0 for the topology $\mathcal{T}$. Therefore, $(M - M + V) \cap B(F, \epsilon/2)$ is also a neighborhood of 0 for the topology $\mathcal{T}$. Finally, $M_2 + V$ is relatively dense in G and so is M_2.

We now prove Lemma 1.2 under the following form:

Lemma 1.5 *Let us assume that M is relatively dense in G. If M is closed in G for the topology $\mathcal{T}$, then for every $V \subset G$ which is a neighborhood of* 0 *for the topology $\mathcal{T}_0$, the interior of $M + V$ for the topology $\mathcal{T}$ is non-empty.*

In other terms, there exists a $x_0 \in G$ such that $M + V - x_0$ is a neighborhood of 0 for the topology $\mathcal{T}$. Let us show that Lemma 1.5 implies Lemma 1.2. Without losing generality, it can be assumed that $V = W - W$ where $W \subset G$ is a compact neighborhood of 0 for $\mathcal{T}_0$. We know that there exists a $x_0 \in G$ such that $\Omega = M + W - x_0$ is a neighborhood of 0 for the topology $\mathcal{T}$. In particular, we have $0 = m + w - x_0$ where $m \in M$ and $w \in W$. Therefore, $\Omega = M - m + W - w$. But $M - m + W - w \subset M - M + W - W$ which ends the proof of Lemma 1.2.

We now prove Lemma 1.5. Without losing generality, it can be assumed that $V = W - W$ where $W \subset G$ is a compact neighborhood of 0 for $\mathcal{T}_0$. Since M is relatively dense, there exists a finite set F such that $M + W + F = G$. By Lemma 1.1, the set $E = M + W$ is closed for $\mathcal{T}$, and we have $E + F = G$. We then use an obvious remark:

Remark 1.1 Let X be a topological space, let $A \subset X$, and let $B \subset X$ be two subsets of X such that $A \cup B = X$. If the interior of A is empty and if B is closed, then $B = X$.

Since G is the finite union of the sets $E + y$, $y \in F$, Remark 1.1 implies that the interior of E for $\mathcal{T}$ is non-empty. Lemma 1.5 is proved.

2 Bogolyobov's Approach

The second proof of Theorem 1.2 relies on an improved version of Proposition 1.1 which was discovered by Nikolaï Bogolyubov and by Erling Følner [1, 2]. They proved that Proposition 1.1 remains valid when M is any relatively dense set. The assumption that M is closed for the topology $\mathcal{T}$ is no more needed, but the simple difference $M - M$ shall be replaced by an iterated difference. This does not affect the proof of Theorem 1.2.

Let us begin with Bogolyubov's theorem. It uses the following definition:

Definition 2.1 The lower and upper Banach densities of a set A of integers are defined by $d_*(A) = \lim_{n \to \infty} \min_x \#(A \cap [x + 1, x + n])/n$ and $d^*(A) = \lim_{n \to \infty} \sup_x \#(A \cap [x + 1, x + n])/n$.

For instance, the upper Banach density of the union A of the intervals $[2^k, 2^k + k]$, $k \in \mathbb{N}$, is 1, and the lower Banach density of A is 0. Bogolyubov proved the following result [1]:

Theorem 2.1 *If $A \subset \mathbb{Z}$ and $d^*(A) > 0$, then there exist k real numbers $\alpha_1, \dots, \alpha_k$, and a positive ϵ with k and ϵ depending only on $d^*(A) > 0$, such that the Bohr set $B(\alpha_1, \dots, \alpha_k, \epsilon)$ is contained in $A - A + A - A$.*

Følner generalized Bogolyubov's theorem to arbitrary l.c.a. groups.

Theorem 2.2 *If $M \subset G$ is relatively dense with respect to a finite set F and if $V \subset G$ is a neighborhood of 0 in G for the topology $\mathcal{T}_0$, then the set $M_4 = M - M + M - M + V$ is a neighborhood of 0 for the topology $\mathcal{T}$.*

As above, Følner's theorem implies the following result:

Corollary 2.1 *If $M \subset G$ is relatively dense and if Ω is a neighborhood of 0 for the topology $\mathcal{T}$, then the set $(M - M + M - M) \cap \Omega$ is relatively dense.*

This improves on Proposition 1.1 since it is no longer assumed that M is closed. However, a second difference is needed. Følner's theorem was already known and published when I wrote my wrong proof of Lemma 5 of [3]. I was just ignorant.

Corollary 2.1 implies Theorem 1.2 by the argument used above. The only difference being that 2ϵ is replaced by 4ϵ.

3 New Examples of Harmonious Sets

Using Theorem 1.2, we give new examples of harmonious sets. Let ω_j, $j \in \mathbb{N}$, be an increasing sequence of positive real numbers, and let m_j, $j \in \mathbb{N}$, be a sequence of natural integers. Let us assume that $m_j\omega_j/\omega_{j+1} \to 0$ as $j \to \infty$. Then we have the following:

Theorem 3.1 *If the two sequences ω_j, m_j, $j \in \mathbb{N}$, are defined as above, the set*

$$\Lambda = \{m\,\omega_j;\ 1 \le m \le m_j,\ 1 \le j\} \tag{6}$$

is harmonious.

Let $\epsilon \in (0, 2)$ and let us define $N = N_\epsilon$ by

$$\omega_{j+1} \ge 20\, m_j\, \omega_j/\epsilon,\ \forall j \ge N. \tag{7}$$

Next we define Λ_N by imposing $j \ge N$ in (6). We shall prove that if $N \ge N_\epsilon$, the ϵ-dual of Λ_N is relatively dense. Then, Theorem 1.2 implies Theorem 3.1. To prove it we define

$$V_j = \{x; |\exp(2\pi x\omega_j) - 1| \leq \epsilon/m_j\}. \tag{8}$$

Then, V_j is $1/\omega_j$ periodic and is a union of intervals of length $\epsilon(2\pi\, m_j\omega_j)^{-1}$. Using (7), one observes that each interval of V_j contains at least an interval of V_{j+1}. Therefore, $\cap_{j\geq N} V_j$ is relatively dense which proves our claim. A similar proof applies to the following theorem:

Theorem 3.2 *Let $\omega_j \geq 1,\ j = 1, 2, \dots$ be an increasing sequence of real numbers such that $\omega_{j+1} \geq 2^j\omega_j,\ j = 1, 2, \dots$. Then, the set Λ of all finite sums $\lambda = \sum_{j\geq 1}\alpha_j\omega_j,\ \alpha_j \in \{0, 1\}$, is harmonious.*

We apply Theorem 1.3. For any $\epsilon > 0$, we chose N such that $2^N > 20/\epsilon$ and define $F_\epsilon = \sum_1^N \alpha_j\omega_j,\ \alpha_j \in \{0, 1\}$ and $\Lambda_\epsilon = \sum_N^\infty \alpha_j\omega_j,\ \alpha_j \in \{0, 1\}$. By Theorem 1.3, it suffices to prove that the ϵ-dual of Λ_ϵ is relatively dense. We define

$$W_j = \{x; |\exp(2\pi x\omega_{j+N}) - 1| \leq \epsilon/2^j\}, \tag{9}$$

and we observe that for any $\lambda \in \Lambda_\epsilon$ we have

$$|\exp(2\pi x\lambda) - 1| \leq \sum_1^\infty |\exp(2\pi x\omega_{j+N}) - 1|. \tag{10}$$

If $x \in \cap_{j\geq 1} W_j$, the right hand side of (10) does not exceed ϵ and x belongs to the ϵ-dual of Λ_ϵ. Then, it remains to prove that $\cap_{j\geq 1} W_j$ is relatively dense. This is achieved by the argument which was used to prove Theorem 3.1.

4 The Union of Two Harmonious Sets

We now concentrate on the case $G = \mathbb{R}^n$. If Λ is harmonious so are $\Lambda \pm \Lambda$. This is obvious from the definition. Indeed, if χ is a weak character on $\Lambda \pm \Lambda$, its restriction to Λ is a weak character on Λ. Therefore, χ is a uniform limit on Λ of a sequence $\chi_j(x) = \exp(2\pi i\omega_j \cdot x)$ of strong characters. We have $\chi(x + y) = \chi(x)\chi(y)$, and it implies that χ is a uniform limit on $\Lambda + \Lambda$ of the same sequence χ_j. The same observation applies to $\Lambda - \Lambda$.

Theorem 1.1 implies the following property:

Lemma 4.1 *If Λ is a harmonious set and if F is a finite set, then $\Lambda + F$ is still harmonious.*

This remark leads to the following equivalence relation between harmonious sets and to partial order on the collection of harmonious sets $\Lambda \subset \mathbb{R}^n$.

Definition 4.1 If $\Lambda, \Lambda' \subset \mathbb{R}^n$, one writes $\Lambda \simeq \Lambda'$ if there exist two finite sets F and F' such that $\Lambda \subset \Lambda' + F'$ and $\Lambda' \subset \Lambda + F$.

The partial order is defined now.

Definition 4.2 If $\Lambda, \Lambda' \subset \mathbb{R}^n$, one writes $\Lambda \prec \Lambda'$ if there exists a finite set F such that $\Lambda \subset \Lambda' + F$.

If $\Lambda \prec \Lambda'$ and $\Lambda' \prec \Lambda$, we have $\Lambda \simeq \Lambda'$.

Definition 4.3 A harmonious set Λ is maximal if it is maximal for the partial order of Definition 4.2.

If Λ is maximal, any harmonious set M containing Λ satisfies $M \subset \Lambda + F$ where F is finite. A set $\Lambda \subset \mathbb{R}^n$ is called a Delone set if it is uniformly discrete and relatively dense. We then have the following:

Theorem 4.1 *Any harmonious set Λ which is a Delone set is maximal.*

We shall prove that if M is a harmonious set, if F is a finite set, and if $\Lambda \subset M + F$, then there exists a finite set F_1 such that $M \subset \Lambda + F_1$. We first observe that $M' = M + F$ is harmonious. Then, it suffices to prove that there exists a finite set F' such that $M' \subset \Lambda + F'$. In other words, we forget M and F and focus on M'. We know that $M' - M'$ is also harmonious. It implies that $M' - M'$ is locally finite. Since Λ is a Delone set, there exists a constant $R > 0$ with the following property: for any $m \in M'$, there exists a $\lambda \in \Lambda$ with $|m - \lambda| \leq R$. Let B_R be the ball centered at 0 with radius R. The set $F' = (M' - M') \cap B_R$ is finite since $M' - M'$ is locally finite. We have $\Lambda \subset M'$ which implies $m - \lambda \in F'$. Therefore, $M' \subset \Lambda + F'$ which ends the proof.

Corollary 4.1 *Let Λ and M be two harmonious sets. If Λ is a Delone set and if $\Lambda \cup M$ is harmonious, then there exists a finite set F such that $M \subset \Lambda + F$.*

Conversely, let Λ and M be two harmonious sets. Then, $M \subset \Lambda + F$ implies that $\Lambda \cup M$ is harmonious. To prove Corollary 4.1, it suffices to apply Theorem 4.1 to Λ. It implies that Λ is maximal. Since $M_1 = \Lambda \cup M$ is harmonious and contains Λ, we have $M_1 \subset \Lambda + F$ where F is finite.

How does one construct maximal harmonious sets? The “cut and projection" scheme is an answer [4]. Here is the recipe. A lattice $\Gamma \subset \mathbb{R}^N$ is a discrete subgroup such that the quotient group $\mathbb{R}^N / \Gamma$ is compact. Equivalently, $\Gamma = A(\mathbb{Z}^N)$ where A is an invertible $N \times N$ matrix. Let $m \geq 1,\ N = n+m,\ \mathbb{R}^N = \mathbb{R}^n \times \mathbb{R}^m$. Let $\Gamma \subset \mathbb{R}^N$ be a lattice. For $X = (x, y) \in \mathbb{R}^n \times \mathbb{R}^m$, one sets $x = p_1(X)$ and $y = p_2(X)$. We now assume that $p_1 : \Gamma \to p_1(\Gamma)$ is a one-to-one mapping and that $p_2(\Gamma)$ is dense in $\mathbb{R}^m$. Recall that a compact set $K \subset \mathbb{R}^m$ is Riemann integrable if its boundary has a zero Lebesgue measure. We are now ready to define “model sets."

Definition 4.4 Let $K \subset \mathbb{R}^m$ be a Riemann integrable compact set with a non-empty interior. Then, the model set $\Lambda = \Lambda(\Gamma, K)$ defined by Γ and K is

$$\Lambda = \{\lambda = p_1(\gamma);\ \gamma \in \Gamma,\ p_2(\gamma) \in K\} \tag{11}$$

Then, $\Lambda(\Gamma, K)$ is a harmonious set [4]. It is a Delone set. Therefore, it is a maximal harmonious set. Conversely, if a harmonious set Λ is a Delone set, there exists a model set M and a finite set F such that $\Lambda \subset M + F$. Are there other maximal harmonious sets ?

The harmonious set defined by Theorem 3.1 is not maximal. Keeping the notations of Theorem 3.1, we define M by the same recipe with the only modification that m_j is replaced by $2m_j$. Then, we cannot have $M \subset \Lambda + F$ where F is finite. To reach a contradiction, we consider the interval $I_j = [(4/3)m_j\,\omega_j, (3/2)m_j\,\omega_j]$. We have $(\Lambda + F) \cap I_j = \emptyset$ if j is sufficiently large. However, $M \cap I_j = \{m\omega_j; (4/3)m_j \leq m \leq (3/2)m_j\}$.

Acknowledgments This work was supported by a grant from the Simons Foundation (601950, YM).

References

1. Bogolyubov, N.N.: Some algebraical properties of almost periods. Zap. Kafedry Mat. Fiziki Kiev **4**, 185–194 (1939)
2. Følner, E.: Generalization of a theorem of Bogoliuboff to topological abelian groups. With an appendix on Banach mean values in non-abelian groups. Math. Scand. **2**, 5–18 (1954)
3. Følner, E.: Note on a generalization of a theorem of Bogoliuboff. Math. Scand. **2**, 224–226 (1954)
4. Meyer, Y.: Algebraic Numbers and Harmonic Analysis. Elsevier (1972)
5. Moody, R.V.: Meyer sets and their duals. The Mathematics of aperiodic order. In: Moody, R.V. (ed.) Proceedings of the Nato Advanced Study Institute on Long-range Aperiodic Order, pp. 403–411. Nato ASI Series C489. Kluwer Acad. Press (1997)

On the Multiplicative Group Generated by Two Primes in $\mathbf{Z}/Q\mathbf{Z}$

Péter P. Varjú

Dedicated to the memory of Jean Bourgain.

Abstract We study the action of the multiplicative group generated by two prime numbers in $\mathbf{Z}/Q\mathbf{Z}$. More specifically, we study returns to the set $([-Q^{\varepsilon}, Q^{\varepsilon}] \cap \mathbf{Z})/Q\mathbf{Z}$. This is intimately related to the problem of bounding the greatest common divisor of S-unit differences, which we revisit. Our main tool is the S-adic subspace theorem.

1 Introduction

In this note, we study the multiplicative group $\{p^m q^n : m, n \in \mathbf{Z}\}$ acting on $\mathbf{Z}/Q\mathbf{Z}$, where p and q are prime numbers and $Q \in \mathbf{Z}_{\geq 2}$ with $\gcd(Q, pq) = 1$. We are interested in returns to the set $([-Q^{\beta}, Q^{\beta}] \cap \mathbf{Z})/Q\mathbf{Z}$ for some $\beta \in (0, 1)$. That is, we aim to describe the set of $(m, n) \in \mathbf{Z}^2$ such that $p^m q^n a = b$ for some $a, b \in ([-Q^{\beta}, Q^{\beta}] \cap \mathbf{Z})/Q\mathbf{Z}$. If a and b lifts to integers that are much smaller than Q^{β}, then small perturbations of (m, n) will also satisfy the same property. To eliminate this triviality, we restrict our attention to the following subset of $([-Q^{\beta}, Q^{\beta}] \cap \mathbf{Z})/Q\mathbf{Z}$.

Definition 1 We write $B(\beta, Q)$ for the set of residues $a \in \mathbf{Z}/Q\mathbf{Z}$ that have lifts $\widetilde{a}$ in $[-Q^{\beta}, Q^{\beta}] \cap \mathbf{Z}$ with $\gcd(\widetilde{a}, pq) = 1$.

The author has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 803711). The author was supported by the Royal Society.

P. P. Varjú (✉)
University of Cambridge, Centre for Mathematical Sciences, Cambridge, UK
e-mail: pv270@dpmms.cam.ac.uk

A. Avila et al. (eds.), *Analysis at Large*, https://doi.org/10.1007/978-3-031-05331-3_16

Notice that $B(\beta, Q)$ contains a canonical representative of each "short orbit segment" intersecting $([-Q^{\beta}, Q^{\beta}] \cap \mathbf{Z})/Q\mathbf{Z}$ in the following sense. Given $a \in ([-Q^{\beta}, Q^{\beta}] \cap \mathbf{Z})/Q\mathbf{Z}$, there is $a' \in B(\beta, Q)$ and $m', n' \in \mathbf{Z}$ such that $a = p^{m'} q^{n'} a'$ and

$$|m'| \leq \beta \log Q / \log p,$$
$$|n'| \leq \beta \log Q / \log q.$$

The choice of a' is unique provided $\beta < 1/3$. This means that, in a sense, to understand returns to $([-Q^{\beta}, Q^{\beta}] \cap \mathbf{Z})/Q\mathbf{Z}$, it is enough to understand returns to $B(\beta, Q)$.

Our first main result is the following:

Theorem 2 *Let p and q be two prime numbers, and let $K \in \mathbf{Z}_{\geq 1}$. Then, there is $C \in \mathbf{R}_{>1}$ and $\beta \in \mathbf{R}_{>0}$ such that for all $Q \in \mathbf{Z}_{>C}$, the set of $(m, n) \in \mathbf{Z}^2$ satisfying the conditions*

- $|m| \leq K \log Q / \log |p|$,
- $|n| \leq K \log Q / \log |q|$,
- *there are $a, b \in B(\beta, Q)$ such that $p^m q^n a = b$*

is contained in a line.

The constant C is ineffective, but β can be made explicit. In particular, the theorem holds with $\beta = (147K)^{-1}$ with some C that is suitably large depending on p, q, and K.

As can be seen from the proof, the result remains valid if we require only that p and q are multiplicatively independent integers instead of being primes. However, in that more general setting, it is less natural to restrict our study to the set $B(\beta, Q)$. Instead, one might formulate a result in terms of the set $([-Q^{\beta}, Q^{\beta}] \cap \mathbf{Z})/Q\mathbf{Z}$ in place of $B(\beta, Q)$ and replace the conclusion by saying that the resulting set of (m, n) will be contained in a suitable neighborhood of a line. We leave this to the interested reader.

Theorem 2 has the following corollary:

Corollary 3 *Let p and q be two prime numbers. For an integer $Q \in \mathbf{Z}_{\geq 2}$ with* $\gcd(pq, Q) = 1$, *we write* $\operatorname{ord}(Q)$ *for the order of the multiplicative group generated by p and q in $\mathbf{Z}/Q\mathbf{Z}$. Then,*

$$\lim_{Q \to \infty} \frac{\operatorname{ord}(Q)}{(\log Q)^2} = \infty.$$

Again, this remains valid if we replace the condition of primality for p and q by multiplicative independence.

Corollary 3 is not a new result. It is well known to follow from a result of Hernández, Luca [7] and Corvaja, Zannier [3], which we will recall below.

Before that, we discuss how Corollary 3 follows from Theorem 2. We observe that the set

$$\Lambda = \{(m,n) \in \mathbf{Z}^2 : p^m q^n \equiv 1 \mod Q\}$$

is a sublatice of $\mathbf{Z}^2$ and its index is $\mathrm{ord}(Q)$. We write λ_1 for the first and λ_2 for the second minima of Λ. If $(m,n) \in \Lambda$ is non-zero, then necessarily

$$|m| \log p + |n| \log q \geq \log Q,$$

so $\lambda_1 \geq c \log Q$ for some constant c that depends on p, q and our choice for the norm with respect to which the minima are defined. By Theorem 2, $\lambda_2 / \log Q \to \infty$ as $Q \to \infty$. Corollary 3 now follows from Minkowski's theorem on successive minima.

Now we discuss some relevant results from the literature. Bugeaud, Corvaja, and Zannier [2, Theorem 1] proved that

$$\gcd(a^n - 1, b^n - 1) \leq \max(a^n, b^n)^{-\varepsilon}$$

for all pair of multiplicatively independent integers a, b and for all $\varepsilon > 0$ provided n is sufficiently large depending on a, b, and ε. This has been extended both by Hernández, Luca [7, Theorem 2.1] and Corvaja, Zannier [3, Remark 1] to the case when a^n and b^n are replaced by two multiplicatively independent integers u and v containing prime factors only from a previously fixed set of primes S. They proved that the inequality

$$\gcd(u-1, v-1) \leq \max(|u|, |v|)^{-\varepsilon}$$

holds provided $\max(|u|, |v|)$ is sufficiently large depending on S and ε. This result is well known to imply Corollary 3. See also Corvaja, Rudnick, and Zannier [4] for a related application of these methods to the multiplicative order of integer matrices mod Q, which contains Corollary 3. A further extension was obtained by Luca [9, Theorem 2.1], who allows u and v to be rational numbers that may contain prime factors outside S, provided their product (with multiplicities) is small compared to $\max(|u|, |v|)$. Furthermore, in this work, u and v are allowed to be multiplicatively dependent, provided they have no multiplicative relation with small exponents.

Corvaja and Zannier [5] developed these ideas in another direction to estimate the greatest common divisors of rational functions evaluated at S units. These results have been extended by Levin [8] to higher dimension.

See also the books of Zannier [10] and Corvaja, Zannier [6], where some of these results are discussed further.

We introduce some notation. We fix a set S, which consists of a finite number of (finite) primes and the symbol ∞. We write $S_f = S \backslash \{\infty\}$. For each $v \in S$, we define a valuation $|\cdot|_v$ on $\mathbf{Q}$. If v is finite and $x \in \mathbf{Z}$, then we set $|x|_v = v^{-m}$, where m is the largest integer with $v^m | x$, and we extend $|\cdot|_v$ to $\mathbf{Q}$ multiplicatively. This is the standard v-adic absolute value. We define $|\cdot|_\infty$ to be the standard Archimedean absolute value. We write $\mathcal{S}$ for the set of positive integers all of whose prime factors are contained in S_f.

Now we can state our second main result, which extends the abovementioned result of Luca [9].

Theorem 4 *For all $\varepsilon > 0$ and S as above, there are $C \in \mathbf{R}_{>1}$, $\alpha \in \mathbf{R}_{>0}$ and $N \in \mathbf{Z}_{>0}$ such that the following holds:*

Let $a_1, b_1, a_2, b_2 \in \mathbf{Z}$ be numbers that are not divisible by any prime in S_f. Let $s_1, t_1, s_2, t_2 \in \mathcal{S}$. Assume

$$\gcd(a_1 s_1, b_1 t_1) = \gcd(a_2 s_2, b_2 t_2) = 1.$$

Let

$$H = \max(s_1, t_1, s_2, t_2).$$

Assume further that

$$\gcd(a_1 s_1 - b_1 t_1, a_2 s_2 - b_2 t_2) \geq H^\varepsilon. \tag{1}$$

Then at least one of the following three items holds:

(a) *$H \leq C$,*
(b) $\max(a_1, b_1, a_2, b_2) \geq H^\alpha$*,*
(c) *There are $n_1, n_2 \in \mathbf{Z}$ not both 0 such that $|n_1|, |n_2| \leq N - 1$ and*

$$\left(\frac{a_1 s_1}{b_1 t_1}\right)^{n_1} = \left(\frac{a_2 s_2}{b_2 t_2}\right)^{n_2}.$$

The constant C is ineffective, but α and N can be made explicit. The theorem always holds (with a suitably large C depending on ε and S) provided

$$N = \left\lfloor \frac{32}{7\varepsilon} \right\rfloor, \qquad \alpha = \frac{7}{512}\varepsilon^2.$$

In fact, we will use in the proof only that ε, N, and α satisfy the inequalities

$$(N+1)\varepsilon > 2N^2\alpha + 4, \tag{2}$$

$$\varepsilon > 16(N-1)\alpha. \tag{3}$$

This result improves on [9, Theorem 2.1] in the following aspects:

- The result in [9] is not applicable when s_1, t_2, s_2, t_2 are of comparable size.

- The bound on $\max(a_1, b_1, a_2, b_2)$ in [9] is of the form $H^{\alpha/\log\log H}$. (Note that H signifies a different quantity in the notation of [9], and we translated the bound to our notation.)
- We make the value of N explicit.

It was observed by Bugeaud, Corvaja, and Zannier that there are infinitely many values of n such that

$$\gcd(a^n - 1, b^n - 1) \geq \exp(\exp(c \log n / \log\log n)),$$

where a and b are multiplicatively independent integers and $c > 0$ is an absolute constant; see the second remark after Theorem 1 in [2]. This significantly limits the extent of any possible improvement over (1). However, in this example, the greatest common divisor is highly composite, and it is not clear how large a common prime factor of $s_1 - 1$ and $s_2 - 1$ can be for some $s_1, s_2 \in \mathcal{S}$. This question is of particular interest in the context of Corollary 3 if we restrict Q to be a prime.

It follows by the box principle that for any $Q \in \mathbf{Z}_{\geq 1}$ and for any $s \in \mathbf{Z}_{\geq 1}$, there are $a, b \in \mathbf{Z}$ with $|a|, |b| \leq Q^{1/2}$ such that $Q|as - b$. This shows that we cannot hope to take α larger than $C\varepsilon$ in Theorem 4 for some constant $C > 0$. However, this still leaves significant room for improvement.

Theorem 2 easily follows from Theorem 4, which we show now.

Proof of Theorem 2 Assuming Theorem 4 Suppose there are $a_1, b_1, a_2, b_2 \in B(\beta, Q)$ and $(m_1, n_1), (m_2, n_2) \in \mathbf{Z}^2$ that are not collinear such that

$$|m_1|, |m_2| \leq \frac{K \log Q}{\log p}, \qquad |n_1|, |n_2| \leq \frac{K \log Q}{\log q}$$

and

$$p^{m_1} q^{n_1} a_1 = b_1, \qquad p^{m_2} q^{n_2} a_2 = b_2.$$

We show that Q must be bounded by a constant depending on p, q, and K only.

To this end, we set $S = \{p, q, \infty\}$ and define $s_1, t_1, s_2, t_2 \in \mathcal{S}$ such that $s_1/t_1 = p^{m_1} q^{n_1}$ and $s_2/t_2 = p^{m_2} q^{n_2}$. We denote by the same symbols the unique lifts of a_1, b_1, a_2, b_2 in $\mathbf{Z} \cap [-Q^\beta, Q^\beta]$. We assume without loss of generality that

$$\gcd(a_1, b_1) = \gcd(a_2, b_2) = \gcd(s_1, t_1) = \gcd(s_2, t_2) = 1.$$

Since $a_1, b_1, a_2, b_2 \in B(\beta, Q)$, their lifts (denoted by the same symbol) are not divisible by p or q, so we get

$$\gcd(a_1 s_1, b_1 t_1) = \gcd(a_2 s_2, b_2 t_2) = 1.$$

We also note that

$$\gcd(a_1s_1 - b_1t_1, a_2s_2 - b_2t_2) \geq Q \geq H^{1/2K},$$

where

$$H = \max(s_1, t_1, s_2, t_2) \leq \max(p^{|m_1|}q^{|n_1|}, p^{|m_2|}q^{|n_2|}).$$

Now we see that all the assumptions of Theorem 4 hold with

$$\varepsilon := \log Q / \log H \geq 1/2K.$$

Item (b) of the conclusion cannot hold, because

$$\max(|a_1|, |b_1|, |a_2|, |b_2|) \leq Q^\beta = H^{\beta\varepsilon},$$

provided β is small enough so that $\beta\varepsilon \leq \alpha$.

Item (c) also cannot hold, because (m_1, n_1) and (m_2, n_2) are not collinear, and this implies that a_1s_1/b_1t_1 and a_2s_2/b_2t_2 are multiplicatively independent. This means that item (a) must hold, which is precisely what we wanted to prove.

For this argument to work, we only need that β is not larger than α/ε. With $\alpha = (7/512)\varepsilon^2$ and $\varepsilon \geq 1/2K$, we see that $\beta = 1/147K$ is sufficient. □

We prove Theorem 4 in the next section. The proof uses Schlickewei's S-adic generalization of Schmidt's subspace theorem. The general approach goes back to the paper of Bugeaud, Corvaja, and Zannier [2], which has been developed further subsequently in [3–5, 7–9]. Our proof makes use of the new construction introduced by Levin [8] to choose the linear forms for which the subspace theorem is applied.

1.1 Notation

Throughout the paper, we fix a finite set S that consists of some prime numbers and the symbol ∞. We write $S_f = S\backslash\{\infty\}$. We write $\mathcal{S}$ for the set of positive integers, all of whose prime divisors are in S_f.

When we have a notation similar to $X_1, \ldots, X_n$, we sometimes write $X_\bullet$ to refer to the whole sequence, or to a generic element of the sequence. The exact meaning will always be clear from the context.

The height of an integer vector $x \in \mathbf{Z}^d$ is defined as

$$H(x_1, \ldots, x_d) = \max(|x_1|_\infty, \ldots, |x_d|_\infty),$$

where $|\cdot|_\infty$ is the standard Archimedean absolute value on $\mathbf{Q}$.

2 Proof of Theorem 4

The purpose of this section is the proof of Theorem 4. Our main tool is Schmidt's subspace theorem in the following generalized form due to Schlickewei:

Theorem 5 (S-Adic Subspace Theorem) *Let $d \in \mathbf{Z}_{\geq 2}$. For each $v \in S$, let $L_1^{(v)}, \ldots, L_d^{(v)} \in \mathbf{Q}[x_1, \ldots, x_d]$ be linearly independent linear forms. Then for all $\varepsilon > 0$, the solutions $(x_1, \ldots, x_d) \in \mathbf{Z}^d$ of the inequality*

$$\prod_{v \in S} \prod_{j=1}^{d} |L_j^{(v)}(x_1, \ldots, x_d)|_v \leq H(x_1, \ldots, x_d)^{-\varepsilon} \tag{4}$$

lie in a finite union of proper subspaces of $\mathbf{Q}^d$.

See [1, Corollary 7.2.5] for a proof of this result. In our applications, we will use the subspace theorem in a finite-dimensional vector space V over $\mathbf{Q}$, and to facilitate the application of the subspace theorem, we need to fix an isomorphism from V to $\mathbf{Q}^d$. In these applications, there will be no natural choice for this isomorphism, and its exact choice will be largely immaterial. For this reason, we reformulate the subspace theorem in the following equivalent form:

Theorem 6 *Let V be a $d \in \mathbf{Z}_{\geq 2}$ dimensional vector space over $\mathbf{Q}$. For each $v \in S$, let $\Lambda_1^{(v)}, \ldots, \Lambda_d^{(v)}$ be a basis of the dual space V^*. Furthermore, let $\Lambda_1^{(0)}, \ldots, \Lambda_d^{(0)}$ be another basis of V^*. Then, for all ε, there is a finite set $\Phi_1, \ldots, \Phi_m \in V^*_{\neq 0}$ such that every solution of*

$$\prod_{v \in S} \prod_{j=1}^{d} |\Lambda_j^{(v)}(x)|_v \leq H(\Lambda_1^{(0)}(x), \ldots, \Lambda_d^{(0)}(x))^{-\varepsilon}$$

for $x \in V$ with $\Lambda_j^{(0)}(x) \in \mathbf{Z}$ for all $j = 1, \ldots, d$ satisfies $\Phi_i(x) = 0$ for some $i \in \{1, \ldots, m\}$.

In our proof of Theorem 4, the first application of the subspace theorem will yield a finite collection of polynomials in two variables depending only on ε and S such that one of the polynomials must vanish at the point $(a_1 s_1/b_1 t_1, a_2 s_2/b_2 t_2)$ for any putative counterexample to the theorem. After this, a second application of the subspace theorem will be needed to conclude the proof. This second part of the proof amounts to proving the following statement:

Proposition 7 *For all $\varepsilon > 0$ and S as above, there are $\alpha \in \mathbf{R}_{>0}$ and $N \in \mathbf{Z}_{>0}$ such that the following holds: Fix a polynomial $P \in \mathbf{Q}[x_1, x_2]$ of degree at most $N - 1$. Then, there is C (depending on P, S, and ε) such that the following holds:*

Let $a_1, b_1, a_2, b_2 \in \mathbf{Z}$ be numbers that are not divisible by any prime in S_f. Let $s_1, t_1, s_2, t_2 \in \mathcal{S}$. Assume

$$\gcd(a_1s_1, b_1t_1) = \gcd(a_2s_2, b_2t_2) = 1$$

and

$$P\Big(\frac{a_1s_1}{b_1t_1}, \frac{a_2s_2}{b_2t_2}\Big) = 0.$$

Let

$$H = \max(s_1, t_1, s_2, t_2).$$

Assume further that

$$\gcd(a_1s_1 - b_1t_1, a_2s_2 - b_2t_2) \geq H^{\varepsilon}.$$

Then at least one of the following three items holds:

(a) *$H \leq C$,*
(b) $\max(a_1, b_1, a_2, b_2) \geq H^{\alpha}$*,*
(c) *There are $n_1, n_2 \in \mathbf{Z}$ not both* 0 *such that $|n_1|, |n_2| \leq N - 1$ and*

$$\Big(\frac{a_1s_1}{b_1t_1}\Big)^{n_1} = \Big(\frac{a_2s_2}{b_2t_2}\Big)^{n_2}.$$

The constant C is ineffective, but α and N can be made explicit. The proposition always holds (with a suitably large C depending on ε, S and P) provided

$$N = \Big\lfloor \frac{32}{7\varepsilon} \Big\rfloor, \qquad \alpha = \frac{7}{512}\varepsilon^2.$$

Notice that this is just a restatement of Theorem 4 with the additional assumption that the point $(a_1s_1/b_1t_1, a_2s_2/b_2t_2)$ is restricted to a curve. This result is unlikely to be either new or optimal. However, it suffices for our purposes, and the proof is simple, so we include it after we showed how Theorem 4 can be reduced to it.

The construction of the linear forms in the following proof is essentially a special case of the construction of Levin [8, Proof of Theorem 3.2].

Proof of Theorem 4 Assuming Proposition 7 Let $\varepsilon \in \mathbf{R}_{>0}$, and let $\alpha \in \mathbf{R}_{>0}$ and $N \in \mathbf{Z}_{>0}$ satisfy (2)–(3). We also fix some $a_1, b_1, a_2, b_2, s_1, t_1, s_2, t_2$ that satisfy all hypotheses of Theorem 4 and which fail items (b) and (c) of the conclusion. We aim to show that item (a) of the conclusion holds, that is, $H \leq C$ for some constant C depending only on ε and S.

We let

$$Q = \gcd(a_1s_1 - b_1t_1, a_2s_2 - b_2t_2).$$

We assume, as we may, that Q is not divisible by any prime in S_f. If we had $p|Q$ for some $p \in S_f$, then necessarily $p \nmid s_1 t_1 s_2 t_2$, and we could just omit p from S.

In what follows, we consider the space $\mathbf{Q}^{N^2} \equiv \mathbf{Q}^{\{0,\ldots,N-1\}^2}$ and write

$$y = (y_{l_1,l_2})_{l_1=0,\ldots,N-1,l_2=0,\ldots,N-1}$$

for its typical element. We will apply the subspace theorem for the quotient space

$$V = \mathbf{Q}^{\{0,\ldots,N-1\}^2}/\{(z,z,\ldots,z)\}.$$

We will evaluate the linear forms at the point $\widetilde{y} \in V$ whose coordinates are

$$\widetilde{y}_{l_1,l_2} = \frac{a_1^{l_1} s_1^{l_1} b_1^{N-1-l_1} t_1^{N-1-l_1} a_2^{l_2} s_2^{l_2} b_2^{N-1-l_2} t_2^{N-1-l_2}}{Q}.$$

Strictly speaking, this specifies a point in $\mathbf{Q}^{\{0,\ldots,N-1\}^2}$, but we do not distinguish $\widetilde{y}$ from its projection to V in our notation.

For each $v \in S$, let $(l_1^{(v)}, l_2^{(v)})$ be such that $|\widetilde{y}_{l_1,l_2}|_v$ is minimal for $(l_1,l_2) = (l_1^{(v)}, l_2^{(v)})$. We define the set of linear forms $\Lambda_\bullet^{(v)} \in V^*$ to be an enumeration of the forms

$$y \mapsto y_{l_1,l_2} - y_{l_1^{(v)},l_2^{(v)}}$$

for $(l_1,l_2) \in \{0,\ldots,N-1\}^2 \backslash (l_1^{(v)}, l_2^{(v)})$. It is easy to verify that these are indeed in V^*, that is they are constant on cosets of the line $\{(z,z,\ldots,z)\}$, and that they also form a basis.

We also define $\Lambda_\bullet^{(0)} = \Lambda_\bullet^{(\infty)}$. We note that

$$\widetilde{y}_{l_1,l_2} - \widetilde{y}_{l_1',l_2'} \in \mathbf{Z}$$

for all $l_1, l_2, l_1', l_2' \in \{0,\ldots,N-1\}$, since

$$a_1 s_1 \equiv b_1 t_1, \quad a_2 s_2 \equiv b_2 t_2 \mod Q,$$

and hence

$$a_1^{l_1} s_1^{l_1} b_1^{N-1-l_1} t_1^{N-1-l_1} a_2^{l_2} s_2^{l_2} b_2^{N-1-l_2} t_2^{N-1-l_2} \mod Q$$

is independent of l_1 and l_2. For this reason,

$$\Lambda_\bullet^{(0)}(\widetilde{y}) \in \mathbf{Z}^{N^2-1}.$$

We observe that

$$|\widetilde{y}_{l_1,l_2} - \widetilde{y}_{l_1^{(v)},l_2^{(v)}}|_v \le C_v|\widetilde{y}_{l_1,l_2}|_v$$

for each (l_1, l_2) and $v \in S$, where $C_v = 1$ if v is finite and $C_\infty = 2$. This means that we have

$$\prod_{v\in S}\prod\nolimits^{\bullet}|\Lambda_{\bullet}^{(v)}(\widetilde{y})|_v \le 2^{N^2-1}\frac{\prod_{v\in S}\prod_{l_1=0}^{N-1}\prod_{l_2=0}^{N-1}|\widetilde{y}_{l_1,l_2}|_v}{\prod_{v\in S}|\widetilde{y}_{l_1^{(v)},l_2^{(v)}}|_v}. \qquad (5)$$

Here $\prod^{\bullet}$ signifies multiplication over the index suppressed by the $\bullet$ notation.

We first estimate the numerator in (5). For each $(l_1, l_2) \in \{0, \dots, N-1\}^2$, we have

$$\prod_{v\in S}|\widetilde{y}_{l_1,l_2}|_v = \frac{|a_1^{l_1}b_1^{N-1-l_1}a_2^{l_2}b_2^{N-1-l_2}|_\infty}{|Q|_\infty} \le H^{2(N-1)\alpha}Q^{-1}.$$

This gives us

$$\prod_{v\in S}\prod_{l_1=0}^{N-1}\prod_{l_2=0}^{N-1}|\widetilde{y}_{l_1,l_2}|_v \le H^{2(N-1)N^2\alpha}Q^{-N^2}.$$

Next, we estimate the denominator in (5). We note that

$$|\widetilde{y}_{l_1^{(v)},l_2^{(v)}}|_\infty \ge Q^{-1}.$$

Furthermore, we have

$$|\widetilde{y}_{l_1^{(v)},l_2^{(v)}}|_v \ge |s_1^{N-1}t_1^{N-1}s_2^{N-1}t_2^{N-1}|_v$$

for all finite $v \in S$; hence

$$\prod_{v\in S_f}|\widetilde{y}_{l_1^{(v)},l_2^{(v)}}|_v \ge s_1^{-N+1}t_1^{-N+1}s_2^{-N+1}t_2^{-N+1} \ge H^{-4(N-1)}.$$

(Here we used that Q is not divisible by any prime in S.)

Combining our estimates for the numerator and denominator in (5), we get

$$\begin{aligned}\prod_{v\in S}\prod\nolimits^{\bullet}|\Lambda_{\bullet}^{(v)}(\widetilde{y})|_v \le & 2^{N^2-1}Q^{-N^2+1}H^{2(N-1)N^2\alpha+4(N-1)}\\ \le & 2^{N^2-1}H^{2(N-1)N^2\alpha+4(N-1)-(N^2-1)\varepsilon}.\end{aligned}$$

We write

$$2\delta = (N^2 - 1)\varepsilon - 2(N - 1)N^2\alpha - 4(N - 1),$$

which is positive by (2). We assume as we may that $2^{N^2-1} \le H^{\delta}$; for otherwise $H \le 2^{\delta^{-1}(N^2-1)}$, and we see that item (a) of the conclusion holds. Therefore, we have

$$\prod_{v\in S}\prod\nolimits^{\bullet} |\Lambda_{\bullet}^{(v)}(\widetilde{y})|_v \le H^{-\delta}.$$

We observe that

$$H(\Lambda_{\bullet}^{(0)}(\widetilde{y})) \le 2\max_{l_1,l_2} |\widetilde{y}_{l_1,l_2}|_{\infty} \le H^{2(1+\alpha)(N-1)},$$

and hence

$$\prod_{v\in S}\prod\nolimits^{\bullet} |\Lambda_{\bullet}^{(v)}(\widetilde{y})|_v \le H(\Lambda_{\bullet}^{(0)}(\widetilde{y}))^{-\frac{\delta}{2(1+\alpha)(N-1)}}.$$

This means that the subspace theorem applies, and we conclude that there is a finite collection of linear forms $\Phi_{\bullet}$ such that $\Phi_j(\widetilde{y}) = 0$ for some j. It may appear that the set of linear forms $\Phi_{\bullet}$ depends on $\widetilde{y}$, for the linear forms $\Lambda_{\bullet}^{(v)}$ were chosen in a manner depending on it. However, there are only finitely many possibilities, and if we take $\Phi_{\bullet}$ to be the union of all linear forms that we obtain from each possible application of the subspace theorem, then it is independent of $\widetilde{y}$.

Now Φ_j lifts to a nonzero linear form on $\mathbf{Q}^{N\times N}$, and it induces a non-zero polynomial $P_j \in \mathbf{Q}[x_1, x_2]$ such that

$$P_j\Big(\frac{a_1 s_1}{b_1 t_1}, \frac{a_2 s_2}{b_2 t_2}\Big) = 0.$$

We can now apply Proposition 7 for each polynomial P_j that arises in this way, and we conclude the proof. □

We turn to the proof of Proposition 7. It requires the following simple lemma:

Lemma 8 *Let $y_1 \ne y_2 \in \mathbf{Z}$, $Q \in \mathbf{Z}_{\ne 0}$ be such that $Q | y_1 - y_2$ and Q is not divisible by any primes in S. Then,*

$$\prod_{v\in S} \min(|y_1|_v, |y_2|_v) \le \frac{2}{Q} \cdot \prod_{v\in S} |y_1 y_2|_v$$

Proof It is clear that neither the assumptions nor the conclusion of the lemma changes if we divide both y_1 and y_2 by a divisor of $\gcd(y_1, y_2)$ all of whose prime factors are in S. For this reason, we may assume that $\gcd(y_1, y_2)$ contains no prime factor that is in S.

We have

$$\prod_{v\in S}\min(|y_1|_v,|y_2|_v)=\prod_{v\in S}\frac{|y_1y_2|_v}{\max(|y_1|_v,|y_2|_v)}.$$

Since $\gcd(y_1,y_2)$ contains no prime factor that is in S, we have $\max(|y_1|_v,|y_2|_v)=1$ for all finite places $v\in S$. In addition, we have $\max(|y_1|_\infty,|y_2|_\infty)\geq Q/2$, because y_1 and y_2 are distinct integers whose difference is divisible by Q. Plugging these observations into the above identity, we get the claim of the lemma. □

Proof of Proposition 7 Let $\varepsilon\in\mathbf{R}_{>0}$, and let $\alpha\in\mathbf{R}_{>0}$ and $N\in\mathbf{Z}_{>0}$ satisfy (2)–(3). Let P be as in Proposition 7. We also fix some $a_1,b_1,a_2,b_2,s_1,t_1,s_2,t_2$ that satisfy all hypotheses of the proposition and which fail items (b) and (c) of the conclusion. We aim to show that item (a) of the conclusion holds.

We assume without loss of generality that P is irreducible. Write d_1 and d_2 for the degrees of P in x_1 and x_2, respectively, and let

$$P(x_1,x_2)=\sum_{j_1=0}^{d_1}\sum_{j_2=0}^{d_2}\alpha_{j_1,j_2}x_1^{j_1}x_2^{j_2}.$$

We note that $d_1,d_2\leq N-1$ by assumption.

We also assume without loss of generality that $d_1,d_2\geq 1$. Indeed, if we had $d_2=0$, say, then there would be only finitely many possibilities for a_1s_1/b_1t_1 such that $P(a_1s_1/b_1t_1,\cdot)=0$ holds, and this in turn restricts a_1,s_1,b_1,t_1 to a finite set. This imposes an upper bound on $\gcd(a_1s_1-b_1t_1,a_2s_2-b_2t_2)$ and hence on H unless $a_1s_1-b_1t_1=0$. However, this latter case is not possible, because item (c) of the conclusion would hold with

$$\Big(\frac{a_1s_1}{b_1t_1}\Big)^1=\Big(\frac{a_2s_2}{b_2t_2}\Big)^0.$$

We see that $d_1=0$ or $d_2=0$ implies that item (c) of the conclusion holds, so we can indeed assume $d_1,d_2\geq 1$.

We also note that at least one in each of the four sets $\alpha_{0,\bullet}$, $\alpha_{d_1,\bullet}$, $\alpha_{\bullet,0}$, and $\alpha_{\bullet,d_2}$ of coefficients does not vanish. (Here we used that P is irreducible and $P\neq x_1$ and $P\neq x_2$.)

In what follows, we consider the space $\mathbf{Q}^{2d_1\times 2d_2}$, whose typical element is denoted by

$$y=(y_{l_1,l_2})_{l_1=0,\ldots,2d_1-1,l_2=0,\ldots,2d_2-1}.$$

For $m_1=0,\ldots,d_1-1$ and $m_2=0,\ldots,d_2-1$, we write

$$\Psi_{m_1,m_2}(y) = \sum_{j_1=0}^{d_1} \sum_{j_2=0}^{d_2} \alpha_{j_1,j_2} y_{j_1+m_1,j_2+m_2},$$

which is a linear form on $\mathbf{Q}^{2d_1\times 2d_2}$. We observe that a point $(x_1, x_2) \in \mathbf{Q}^2_{\neq 0}$ satisfies $P(x_1, x_2)$ if and only if

$$\Psi_{m_1,m_2}((x_1^{l_1} x_2^{l_2})_{l_1=0,\dots,2d_1-1,l_2=0,\dots,2d_2-1}) = 0$$

holds for at least one and hence for all m_1, m_2 in the relevant range. We write V for the $3d_1d_2$-dimensional subspace of $\mathbf{Q}^{2d_1\times 2d_2}$ on which all Ψ_{m_1,m_2} vanish.

We consider the point $\widetilde{y} \in V$ given by

$$\widetilde{y}_{l_1,l_2} = a_1^{l_1} s_1^{l_1} b_1^{2d_1-1-l_1} t_1^{2d_1-1-l_1} a_2^{l_2} s_2^{l_2} b_2^{2d_2-1-l_2} t_2^{2d_2-1-l_2}.$$

To verify that $\Psi_{m_1,m_2}(\widetilde{y}) = 0$, we note that

$$b_1^{-2d_1+1} t_1^{-2d_1+1} b_2^{-2d_2+1} t_2^{-2d_2+1} \cdot \widetilde{y}_{l_1,l_2} = \left(\frac{a_1s_1}{b_1t_1}\right)^{l_1} \left(\frac{a_2s_2}{b_2t_2}\right)^{l_2}.$$

In what follows, we use the subspace theorem to show that there is a finite collection $\Phi_\bullet \in V^*_{\neq 0}$ such that $\Phi_j(\widetilde{y}) = 0$ for some j, and this collection of linear forms is independent of the choice of $a_1, b_1, a_2, b_2, s_1, t_1, s_2, t_2$. Each Φ_j can be lifted to a linear form on $\mathbf{Q}^{2d_1\times 2d_2}$, which is not in the span of the Ψ_{m_1,m_2}. We denote this linear form with the same symbol. Then, the polynomial

$$Q_j(x_1, x_2) = \Phi_j((x_1^{l_1} x_2^{l_2})_{l_1=0,\dots,2d_1-1,l_2=0,\dots,2d_2-1})$$

is not in the ideal generated by P, but

$$Q_j\left(\frac{a_1s_1}{b_1t_1}, \frac{a_2s_2}{b_2t_2}\right) = 0.$$

Each such Q_j has only finitely many common solutions with P. This means that the point

$$\left(\frac{a_1s_1}{b_1t_1}, \frac{a_2s_2}{b_2t_2}\right)$$

must belong to a certain finite set, which depends only on P and S, and this means that item (a) in the conclusion holds with some C that depends only on P and S. This will complete the proof.

The next step is to choose the families of linear forms on V needed for the application of the subspace theorem. For each place $v \in S$, we choose a set

$\mathcal{L}_v \subset \{0, \dots, 2d_1 - 1\} \times \{0, \dots, 2d_2 - 1\}$ of cardinality $\dim V = 3d_1d_2$. We then define $\Lambda_\bullet^{(v)}$ to be an enumeration of the linear forms $y \mapsto y_{l_1,l_2}$ for $(l_1, l_2) \in \mathcal{L}_v$.

Let i be the smallest and let k be the largest index such that $\alpha_{0,i} \neq 0$ and $\alpha_{d_1,k} \neq 0$, respectively. (Recall that α_{j_1,j_2} are the coefficients of P.) Each of the sets $\mathcal{L}_v$ will be either

$$\{0, \dots, 2d_1 - 1\} \times \{0, \dots, 2d_2 - 1\} \backslash \{d_1, \dots, 2d_1 - 1\} \times \{k, \dots, k + d_2 - 1\} \tag{6}$$

or

$$\{0, \dots, 2d_1 - 1\} \times \{0, \dots, 2d_2 - 1\} \backslash \{0, \dots, d_1 - 1\} \times \{i, \dots, i + d_2 - 1\}. \tag{7}$$

We first show that the resulting linear forms $\Lambda_\bullet^{(v)}$ form a basis of V^* in either case. In fact, we show this only in the case of (6), because the case of (7) can be treated in a similar fashion. Since $|(6)| = \dim V$, it is enough to show that the linear forms $y \mapsto y_{l_1,l_2}$ for $(l_1, l_2) \in (6)$ span V^*. To that end, it is enough to show that $y \mapsto y_{l_1,l_2}$ is in the span for all $(l_1, l_2) \in \{d_1, \dots, 2d_1 - 1\} \times \{k, \dots, k + d_1 - 1\}$. Fix some $(l_1', l_2') \in \{d_1, \dots, 2d_1 - 1\} \times \{k, \dots, k + d_1 - 1\}$. We observe that

$$y_{l_1',l_2'} = - \sum_{(j_1,j_2) \neq (d_1,k)} \frac{\alpha_{j_1,j_2}}{\alpha_{d_1,k}} y_{j_1+l_1'-d_1, j_2+l_2'-k}$$

for all $y \in V$. This means that $y \mapsto y_{l_1',l_2'}$ is in the span of the linear forms $y \mapsto y_{l_1,l_2}$ for

$$(l_1, l_2) \in \{0, \dots, l_1' - 1\} \times \{0, \dots, 2d_2 - 1\} \cup \{l_1'\} \times \{0, \dots, l_2' - 1\}.$$

Using this observation, we can prove that (l_1', l_2') is in the span of $y \mapsto y_{l_1,l_2}$ for $(l_1, l_2) \in (6)$ by induction first on l_1' and then on l_2'.

For each $v \in S$, we define $\Lambda_\bullet^{(v)}$ using (6) if $|\widetilde{y}_{d_1,k}|_v \geq |\widetilde{y}_{0,i}|_v$, and we use (7) otherwise. We write $\mathcal{A} = (6) \cap (7)$ and $\mathcal{B} = (6) \backslash (7)$. We observe that $\{0, \dots, 2d_1 - 1\} \times \{0, \dots, 2d_2 - 1\}$ is the disjoint union of the sets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{B} + (d_1, k - i)$. For each $v \in S$, $\Lambda_\bullet^{(v)}$ contains $y \mapsto y_{l_1,l_2}$ for all $(l_1, l_2) \in \mathcal{A}$, and it also contains precisely one of $y \mapsto y_{l_1,l_2}$ or $y \mapsto y_{l_1+d_1,l_2+k-i}$ for each $(l_1, l_2) \in \mathcal{B}$, and it contains the one which gives a smaller or equal $|\cdot|_v$ value to $\widetilde{y}$. This means that

$$\begin{aligned}\prod_{v \in S} \prod^\bullet |\Lambda_\bullet^{(v)}(\widetilde{y})|_v = & \prod_{v \in S} \prod_{(l_1,l_2) \in \mathcal{A}} |\widetilde{y}_{l_1,l_2}|_v \\ & \times \prod_{v \in S} \prod_{(l_1,l_2) \in \mathcal{B}} \min(|\widetilde{y}_{l_1,l_2}|_v, |\widetilde{y}_{l_1+d_1,l_2+k-i}|_v).\end{aligned}$$

Here $\prod^\bullet$ signifies multiplication over the index suppressed by the $\bullet$ notation.

We note that $\widetilde{y}_{l_1,l_2} \neq \widetilde{y}_{l_1+d_1,l_2+k-i}$ for each $(l_1, l_2) \in \mathcal{B}$ follows from

$$\left(\frac{a_1 s_1}{b_1 t_1}\right)^{d_1} \neq \left(\frac{a_2 s_2}{b_2 t_2}\right)^{i-k},$$

which in turn follows from our assumption that item (c) in the conclusion does not hold. Therefore, we can apply Lemma 8 for each pair $\widetilde{y}_{l_1,l_2}, \widetilde{y}_{l_1+d_1,l_2+k-i}$ for $(l_1, l_2) \in \mathcal{B}$ and get

$$\prod_{v\in S}\prod{}^{\bullet} |\Lambda_{\bullet}^{(v)}(\widetilde{y})|_v \leq \left(\frac{2}{Q}\right)^{|\mathcal{B}|} \cdot \prod_{v\in S}\prod_{l_1=0}^{2d_1-1}\prod_{l_2=0}^{2d_2-1} |\widetilde{y}_{l_1,l_2}|_v.$$

We note that

$$\prod_{v\in S} |\widetilde{y}_{l_1,l_2}|_v = |a_1^{l_1} b_1^{2d_1-1-l_1} a_2^{l_2} b_2^{2d_2-1-l_2}|_\infty \leq H^{(2d_1+2d_2-2)\alpha}.$$

This and $Q \geq H^{\varepsilon}$ gives

$$\prod_{v\in S}\prod{}^{\bullet} |\Lambda_{\bullet}^{(v)}(\widetilde{y})|_v \leq 2^{d_1 d_2} H^{4d_1 d_2(2d_1+2d_2-2)\alpha - d_1 d_2 \varepsilon}.$$

We write

$$2\delta = \varepsilon - 8(d_1 + d_2 - 1)\alpha,$$

which is positive by assumption (3). We assume as we may that $2 \leq H^{\delta}$, for otherwise $H \leq 2^{1/\delta}$, and item (a) of the conclusion holds. We have, therefore,

$$\prod_{v\in S}\prod{}^{\bullet} |\Lambda_{\bullet}^{(v)}(\widetilde{y})|_v \leq H^{-\delta d_1 d_2}.$$

We apply the subspace theorem with the linear forms $\Lambda_{\bullet}^{(v)}$ defined above and with $\Lambda_{\bullet}^{(0)} = \Lambda_{\bullet}^{(\infty)}$. We note that $\Lambda_{\bullet}^{(0)}(\widetilde{y}) \in \mathbf{Z}^{3d_1 d_2}$ and

$$H(\Lambda_{\bullet}^{(0)}(\widetilde{y})) \leq \max_{l_1,l_2} |\widetilde{y}_{l_1,l_2}|_\infty \leq H^{(2d_1+2d_2-2)(1+\alpha)}.$$

We have, therefore,

$$\prod_{v\in S}\prod{}^{\bullet} |\Lambda_{\bullet}^{(v)}(\widetilde{y})|_v \leq H(\Lambda_{\bullet}^{(0)}(\widetilde{y}))^{-\frac{\delta d_1 d_2}{(2d_1+2d_2-2)(1+\alpha)}}.$$

This means that the subspace theorem applies, and hence there is a finite collection of linear forms $\Phi_{\bullet} \in V_{\neq 0}^{*}$ such that $\Phi_j(\widetilde{y}) = 0$ for some j.

It may appear that the linear forms $\Phi_\bullet$ depend on $\widetilde{y}$, because the choice of $\Lambda_\bullet^{(v)}$ for each $v \in S$ depends on it. However, there are only finitely many possibilities we need to consider, so we can simply take the union of the linear forms that result from each possible application of the subspace theorem. As we discussed above, this completes the proof. □

Acknowledgments I am grateful to Elon Lindenstrauss for the very helpful discussions on the subject of this note and to Yann Bugeaud, Pietro Corvaja, Umberto Zannier, and the anonymous referee for the very helpful comments on an earlier version of this note.

References

1. Bombieri, E., Gubler, W.: Heights in Diophantine geometry. New Mathematical Monographs, vol. 4. Cambridge University Press, Cambridge (2006). MR2216774 ↑6
2. Bugeaud, Y., Corvaja, P., Zannier, U.: An upper bound for the G.C.D. of $a^n - 1$ and $b^n - 1$. Math. Z. **243**(1), 79–84 (2003). MR1953049 ↑3 4 5
3. Corvaja, P., Zannier, U.: On the greatest prime factor of (ab + 1)(ac + 1). Proc. Am. Math. Soc. **131**(6), 1705–1709 (2003). MR1955256 ↑2, 3, 5
4. Corvaja, P., Rudnick, Z., Zannier, U.: A lower bound for periods of matrices. Comm. Math. Phys. **252**(1–3), 535–541 (2004). MR2104888 ↑3, 5
5. Corvaja, P., Zannier, U.: A lower bound for the height of a rational function at S-unit points. Monatsh. Math. **144**(3), 203–224 (2005). MR2130274 ↑3, 5
6. Corvaja, P., Zannier, U.: Applications of Diophantine approximation to integral points and transcendence. Cambridge Tracts in Mathematics, vol. 212. Cambridge University Press, Cambridge (2018). MR3793125 ↑3
7. Hernández, S., Luca, F.: On the largest prime factor of $(ab{+}1)(ac{+}1)(bc{+}1)$. Bol. Soc. Mat. Mexicana (3) **9**(2), 235–244 (2003). MR2029272 ↑2, 3, 5
8. Levin, A.: Greatest common divisors and Vojta's conjecture for blowups of algebraic tori. Invent. Math. **215**(2), 493–533 (2019). MR3910069 ↑3, 5, 7
9. Luca, F.: On the greatest common divisor of $u-1$ and $v-1$ with u and v near $\mathcal{S}$-units. Monatsh. Math. **146**(3), 239–256 (2005). MR2184226 ↑3, 4, 5
10. Zannier, U.: Some problems of unlikely intersections in arithmetic and geometry. Annals of Mathematics Studies, vol. 181. Princeton University Press, Princeton, NJ (2012). With appendixes by David Masser. MR2918151 ↑3