Some results at the interface of combinatorics and number theory

Zachary Chase

The Queen's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2023

This thesis is dedicated to my family.

Acknowledgments

I would first like to thank the Oxford Mathematical Institute Fund (OMIF) for funding my studies throughout my time at Oxford. I am very grateful for their financial support, allowing me to focus on research.

I would also like to thank my friends, who made my time at Oxford much more enjoyable, as well the (fortunately) large analytic number theory group for the biweekly reading groups and weekly lunches, which were always a great source of fun.

Furthermore, I happily thank my advisor Ben Green for being a terrific PhD advisor, always willing to suggest problems whilst allowing me to independently explore my own research directions, and for putting up with some of my shenanigans.

Finally, I thank my family for always supporting me and being a constant source of joy in life.

Regarding Chapter 2, I would like to thank Po-Shen Loh for telling me the Gan-Loh-Sudakov conjecture. I also thank Daniel Korandi for a cleaner proof of Lemma 2 and for helpful suggestions on the paper's presentation.

Regarding Chapter 3, I would like to thank an anonymous referee for several helpful comments improving the presentation of the paper and pointing out a quick improvement to Proposition 3.3.1. I would also like to thank my advisor, Ben Green, for suggesting this problem to me. Finally, thanks are due to Juan Arias de Reyna for bringing to attention the dubious nature of the claim discussed in Section 3.8, and Hugh Williams for kindly responding to emails and helping resolve the situation.

Regarding Chapter 4, I would like to thank my advisor Ben Green for several helpful comments on the readability of the paper and Noah Golowich for pointing out a flaw in a claimed generalization of one of the propositions in a previous version of the paper.

Regarding Chapter 5, I would like to thank Shyam Narayanan for providing an extension to all $q \in (0,1)$, and Fedor Nazarov for an easier proof of Lemma 5.6.1.

Regarding Chapter 6, I would like to thank my advisor Ben Green for suggesting this problem to me and Fedor Nazarov for nearly solving the optimization problem by hand.

Regarding Chapter 7, we thank Ryan Alweiss, Brice Huang, and Mark Sellke for sharing their writeup with us.

Regarding Chapter 8, I would like to thank Jeffrey Shallit for striving to revive study of this problem.

I would like to thank Alex Scott and Julian Sahasrabudhe for helpful comments improving the presentation of this thesis.

Abstract

We present some results in the union and sometimes in the intersection of combinatorics and number theory.

Contents

Chapter 1

Introduction

This thesis comprises a collection of results in the fields of extremal graph theory, combinatorial number theory, combinatorics of strings, hypothesis testing, additive number theory, equidistribution, elementary number theory, and combinatorics of sets.

We hope this thesis supports the notion that “different” areas of math are not actually different, and that working on fundamental problems leads one to various interesting areas of mathematics.

In Chapter 2, we study the following question: what is the maximum number of triangles that a graph on n vertices can have, provided each vertex is adjacent to at most d others? It is not too difficult to see that if n is a multiple of $d+1$, then a disjoint union of K_{d+1}'s (i.e., cliques on $d+1$ vertices) is optimal. When n is not a multiple of $d+1$, however, the question was wide open, with the conjectured optimal graph being a disjoint union of as many K_{d+1}'s as possible, and then a clique on the remaining vertices. The conjecture, due to Gan, Loh, and Sudakov, received some attention with several interesting partial results. We resolve the conjecture. We do so by establishing a general identity, valid for all graphs, that concerns the number of triangles that a closed neighborhood intersects.

The answer resolves a conjecture of Gan, Loh, and Sudakov.

In Chapter 3, we study a random analogue of a conjecture of Gilbreath about the prime numbers.

2 3 5 7 11 13 17
1 2 2 4 2 4
1 0 2 2 2

1 2 0 0
1 2 0
1 2
1

To formulate Gilbreath's conjecture, consider the sequence of prime numbers, in increasing order. As the figure above indicates, we define a new sequence, where the j^{th} entry is the difference between the $j+1^{\text{st}}$ prime and the j^{th} prime. In general, given any sequence, we obtain a new one by replacing consecutive terms by the absolute value of their difference: i.e., given $(x_n)_{n=1}^{\infty}$, we obtain $(y_n)_{n=1}^{\infty} := (|x_{n+1} - x_n|)_{n=1}^{\infty}$, and analogously with finite sequences (whereby the length decreases by 1).

Gilbreath's conjecture is that, beginning with the sequence of prime numbers, if we repeatedly look at the sequence obtained by computing the absolute value of the difference of consecutive terms of the previous sequence, then the first term is always 1 (beginning with the sequence of consecutive prime differences).

Gilbreath's conjecture is rather old. Many have speculated throughout the years that the primes don't have much to do with the conjecture, that it should hold for any initial sequence that is "sufficiently random" and has small gaps. We prove such a random analogue.

A model problem whose solution will allow us to deduce a "random analogue" of Gilbreath's conjecture is the following.

Form an (initial) sequence of length M by letting each term be a uniformly randomly chosen element of $\{1, 2, \ldots, 100\}$. Is it true that after $M/2$ iterations of this consecutive differencing procedure, with high probability the obtained sequence consists solely of 0s and 1s?

We solve the model problem, with $\approx \log\log M$ in place of 100, in part by deriving a new result for bootstrapping monochromatic random walks on colored graphs.

In Chapter 4, we improve the upper bound on the "separating words problem". This problem concerns the ability of a deterministic finite automaton, one of the most basic models for computation, to distinguish between two given 0-1 strings of length n, one of the most basic computational tasks. In 1989, Robson showed that, for any distinct $x, y \in \{0,1\}^n$, there is a deterministic finite automaton with at most $Cn^{2/5+\epsilon}$

states that accepts x but not y. We improve the upper bound to $Cn^{1/3+\epsilon}$. We note that the lower bound still remains at $c\log n$.

In order to obtain the improvement, we solve a number-theoretic problem. As a warmup, one can show that, for any distinct sets $A, B \subseteq [n]$, there is some prime $p \leq C\sqrt{n \log n}$ and some $i \in [p]$ with

$$\#\{a \in A : a \equiv i \bmod p\} \neq \#\{b \in B : b \equiv i \bmod p\}.$$

The bound $C\sqrt{n \log n}$ is tight up to logarithms (and constants). However, if we are additionally given that A and B are sufficiently sparse sets, then this upper bound can be improved. Specifically, we show that if $A, B \subseteq [n]$ are distinct sets that are each $n^{1/3}$-separated (meaning every distinct $a, a' \in A$ satisfy $|a' - a| \geq n^{1/3}$), then there is some prime $p \leq Cn^{1/3+\epsilon}$ and some $i \in [p]$ so that

$$\#\{a \in A : a \equiv i \bmod p\} \neq \#\{b \in B : b \equiv i \bmod p\}.$$

We obtain this number-theoretic result by complex analytic methods, showing it suffices to prove that any distinct "separated" subsets of $[n]$ have a different m^{th} moment for a small m, which in turn is equivalent to a statement about the behavior of sparse Littlewood polynomials near 1, for which complex analytic tools can indeed be employed.

In Chapter 5, we improve the upper bound on the "trace reconstruction problem". In this problem, there's an unknown 0-1 string $x \in \{0, 1\}^n$ of length n. We see T independently generated "traces" of x, where a trace is a random string, formed by deleting each bit of x i.i.d. with probability 1/2. The question is to determine the smallest positive integer T, as a function of n, so that one can reconstruct x with high probability. The gaps between the lower and upper bounds on T are rather large, standing at $cn^{3/2-\epsilon}$ and $\exp\left(Cn^{1/3}\right)$ before our work. We improve the upper bound to $\exp\left(Cn^{1/5+\epsilon}\right)$. A large gap remains.

In Chapter 6, we show the sumset $A + B$ contains a perfect square if $A, B \subseteq [N]$ have $|A|, |B| \geq (\frac{3}{8} + \epsilon)N$, where the constant 3/8 is optimal. In the early 2000s, the analogous constant for the problem of $A + A$ containing a perfect square was proven to be $\frac{11}{32}$. We establish our "bipartite" result by first solving the "modular" version of the problem, namely when $A + B$ contains a quadratic residue mod q for two sets $A, B \subseteq \mathbb{Z}_q$, and then using basic fourier analysis to bootstrap the result to the natural numbers. We solve the modular problem via fourier analysis, quickly reducing to a quadratic optimization problem, that is solved by a mixture of hand and computer.

In Chapter 7, we improve the upper bound on the number of steps that a variant of the Euclidean algorithm lasts for. Specifically, we show that the process $a \mapsto n \pmod{a}$ reaches 0 after at most $Cn^{\frac{1}{3}-\frac{2}{177}+\epsilon}$ iterations, no matter the starting value $a = a_0$. The work in this chapter is joint with Mayank Pandey.

In Chapter 8, we build on a recent breakthrough of Gilmer to show that for any union-closed family of sets $\mathcal{F} \subseteq \mathcal{P}([n])$, there is some $x \in [n]$ that is in at least $\frac{3-\sqrt{5}}{2}$ proportion of sets $F \in \mathcal{F}$. We show that this result is optimal amongst families that are "approximately union-closed". The work in this chapter is joint with Shachar Lovett.

1.1 Notation

For functions f, g, we say $f = O(g)$ if there exists a constant C so that $|f(x)| \leq C|g(x)|$ for all x in the (common) domain of f and g. We write $f = \Omega(g)$ if there exists a constant $c > 0$ so that $|f(x)| \geq c|g(x)|$ for all x. We write $f = \widetilde{O}(g)$ if there exists a constant C so that $|f(x)| \leq C|g(x)| \log^C |g(x)|$ for all x. We write $f = \widetilde{\Omega}(g)$ if there exist constants $c, C > 0$ so that $|f(x)| \geq c|g(x)| \log^{-C} |g(x)|$ for all x. If f, g are defined on an ordered domain, we write $f = o(g)$ if, for every $\epsilon > 0$, it holds for all large x that $|f(x)| \leq \epsilon|g(x)|$.

We use the standard $[N] := \{1, \ldots, N\}$ and $e(\theta) := e^{2\pi i\theta}$ for $\theta \in \mathbb{R}$.

Chapter 2

A proof of the Gan-Loh-Sudakov conjecture

2.1 Summary

We prove that any graph on n vertices with max degree d has at most $q\binom{d+1}{3} + \binom{r}{3}$ triangles, where $n = q(d+1) + r$, $0 \le r \le d$. This resolves a conjecture of Gan, Loh, and Sudakov.

2.2 Introduction

Fix positive integers d and n with $d+1 \le n \le 2d+1$. Galvin [24] conjectured that the maximum number of cliques in an n-vertex graph with maximum degree d comes from a disjoint union $K_{d+1} \sqcup K_r$ of a clique on $d+1$ vertices and a clique on $r := n-d-1$ vertices. Cutler and Radcliffe [16] proved this conjecture. Engbers and Galvin [21] then conjectured that, for any fixed $t \ge 3$, the same graph $K_{d+1} \sqcup K_r$ maximizes the number of cliques of size t, over all $(d+1+r)$-vertex graphs with maximum degree d. Engbers and Galvin [21]; Alexander, Cutler, and Mink [1]; Law and McDiarmid [40]; and Alexander and Mink [2] all made progress on this conjecture before Gan, Loh, and Sudakov [26] resolved it in the affirmative. Gan, Loh, and Sudakov then extended the conjecture to arbitrary $n \ge 1$ (for any d).

Conjecture (Gan-Loh-Sudakov Conjecture)**.** *Any graph on n vertices with maximum degree d has at most $q\binom{d+1}{3} + \binom{r}{3}$ triangles, where $n = q(d+1) + r$, $0 \le r \le d$.*

They showed their conjecture implies that, for any fixed $t \ge 4$, any max-degree d graph on $n = q(d+1) + r$ vertices has at most $q\binom{d+1}{t} + \binom{r}{t}$ cliques of size t. In other words, considering triangles is enough to resolve the general problem of cliques of any fixed size.

The Gan-Loh-Sudakov conjecture (GLS conjecture) has attracted substantial attention. Cutler and Radcliffe [17] proved the conjecture for $d \leq 6$ and showed that a minimal counterexample, in terms of number of vertices, must have $q = O(d)$. Gan [25] proved the conjecture if $d + 1 - \frac{9}{4096}d \leq r \leq d$ (there are some errors in his proof, but they can be mended). Using fourier analysis, the author [11] proved the conjecture for Cayley graphs with $q \geq 7$. Kirsch and Radcliffe [36] investigated a variant of the GLS conjecture in which the number of edges is fixed instead of the number of vertices (with still a maximum degree condition).

In this chapter, we fully resolve the Gan-Loh-Sudakov conjecture.

Theorem 2.2.1. *For any positive integers $n, d \geq 1$, any graph on n vertices with maximum degree d has at most $q\binom{d+1}{3}+\binom{r}{3}$ triangles, where $n = q(d+1)+r$, $0 \leq r \leq d$.*

Analyzing the proof shows that $qK_{d+1} \sqcup K_r$ is the unique extremal graph if $r \geq 3$, and that $qK_{d+1} \sqcup H$, for any H on r vertices, are the extremal graphs if $0 \leq r \leq 2$.

The heart of the proof is the following Lemma, of independent interest, which says that, in any graph, we can find a closed neighborhood whose removal from the graph removes few triangles. Theorem 2.2.1 will follow from its repeated application.

Lemma 2.2.2. *In any graph G, there is a vertex v whose closed neighborhood meets at most $\binom{d(v)+1}{3}$ triangles.*

As mentioned above, Theorem 2.2.1, together with the work of Gan, Loh, and Sudakov [26], yields the general result, for cliques of any fixed size.

Theorem 2.2.3. *Fix $t \geq 3$. For any positive integers $n, d \geq 1$, any graph on n vertices with maximum degree d has at most $q\binom{d+1}{t} + \binom{r}{t}$ cliques of size t, where $n = q(d+1) + r$, $0 \leq r \leq d$.*

Theorem 2.2.3 gives another proof of (the generalization of) Galvin's conjecture (to $n \geq 2d+2$) that a disjoint union of cliques maximizes the total number of cliques in a graph with prescribed number of vertices and maximum degree.

Finally, the author would like to point out a connection to a related problem, that of determining the minimum number of triangles that a graph of fixed number of vertices n and prescribed minimum degree δ can have. The connection stems from a relation, reiterated in [2] and [26], between the number of triangles in a graph and the number of triangles in its complement:

$$|T(G)| + |T(G^c)| = \binom{n}{3} - \frac{1}{2}\sum_v d(v)[n-1-d(v)].$$

Lo [41] resolved this "dual" problem when $\delta \leq \frac{4n}{5}$. His results resolve the GLS conjecture for regular graphs for $q = 2, 3$, and the GLS conjecture implies his results, up to an additive factor of $O(\delta^2)$, for $q = 2, 3$, and yields an extension of his results for $q \geq 4$ — these are the optimal results asymptotically, in the natural regime of $\frac{\delta}{n}$ fixed, and $n \to \infty$.

2.3 Notation

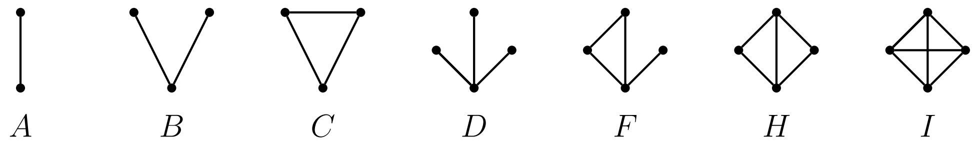
Denote by E the edge set of G; for two vertices u, v, we write "$uv \in E$" if there is an edge between u and v and "$uv \notin E$" otherwise — in particular, for any u, $uu \notin E$. For a vertex v, let $|T_{N[v]}|$ denote the number of triangles with at least one vertex in the closed neighborhood $N[v] := \{u : uv \in E\} \cup \{v\}$, and let $|T(G - N[v])|$ denote the number of triangles with all vertices in the graph $G - N[v]$ (the subgraph induced by the vertices not in $N[v]$). Finally, $d(v)$ denotes the degree of v.

2.4 Proof of Gan-Loh-Sudakov conjecture

For a graph G, let $W(G) = \{(x, u, v, w) : ux, vx, wx \in E, uv, uw, vw \notin E\}$.

Lemma 2.4.1. *For any graph G, $6\sum_v |T_{N[v]}| + |W(G)| = \sum_v d(v)^3$.*

Proof. Let $\Omega = \{(z, u, v, w) : uv, uw, vw \in E$ and $[zu \in E$ or $zv \in E$ or $zw \in E]\}$, $\Sigma = \{(x, u, v, w) : ux, vx, wx \in E\}$, and $W = W(G)$. Note that repeated vertices in the 4-tuples are allowed. Since there are 6 ways to order the vertices of a triangle, we have $\sum_v 6|T_{N[v]}| = |\Omega|$. Any 4-tuple in Σ, W, or Ω gives rise to one of the induced subgraphs shown below, since one vertex must be adjacent to all the others.



Since $|\Sigma| = \sum_v d(v)^3$, it thus suffices to show that for each of the induced subgraphs above, the number of times it comes from a 4-tuple in Σ is the sum of the number of times it comes from 4-tuples in Ω and W. Any fixed copy of A, say on vertices u and v, comes 0 times from a 4-tuple in Ω (since it has no triangles), and 2 times from each of W and Σ $((u, v, v, v), (v, u, u, u))$. Any fixed copy of B, say on vertices u, v, w with $vu, vw \in E$, comes 0 times from Ω, and 6 times from each of W and Σ $((v, u, u, w), (v, u, w, u), (v, u, w, w), (v, w, u, u), (v, w, u, w), (v, w, w, u))$. Any fixed copy of C comes 18 times from each of Ω and Σ (3 choices for the first vertex and then 6 for the ordered triangle), and 0 times from W. Similarly, any fixed copy of D comes 6 times from each of W and Σ, and 0 times from Ω; finally, F, H, I come $6, 12, 24$ times, respectively, from each of Ω and Σ, and 0 times from W. □

We now prove Lemma 2.2.2, repeated below for the reader's convenience.

Lemma 2.2.2: In any graph G, there is a vertex v whose closed neighborhood meets at most $\binom{d(v)+1}{3}$ triangles, i.e. $|T_{N[v]}| \le \binom{d(v)+1}{3}$.

Proof. By Lemma 2.4.1, since $|W(G)| \ge |\{(x, u, u, u) : ux \in E\}| = \sum_x d(x)$, we have $\sum_v |T_{N[v]}| \le \sum_v \frac{1}{6}[d(v)^3 - d(v)]$. By the pigeonhole principle, there is some v with

$$|T_{N[v]}| \le \frac{1}{6}[d(v)^3 - d(v)] = \binom{d(v)+1}{3}.$$

□

Lemma 2.4.2. *For any positive integers $a \ge b \ge 1$, it holds that $\binom{a}{3} + \binom{b}{3} \le \binom{a+1}{3} + \binom{b-1}{3}$. Consequently, for any positive integers a, b and any positive integer c with $\max(a, b) \le c \le a + b$, it holds that $\binom{a}{3} + \binom{b}{3} \le \binom{c}{3} + \binom{a+b-c}{3}$.*

Proof. $\binom{a+1}{3} - \binom{a}{3} = \binom{a}{2}$, and $\binom{b}{3} - \binom{b-1}{3} = \binom{b-1}{2}$. Iterate to get the consequence. □

We now finish the proof of Theorem 2.2.1. With a fixed d, we induct on n. For $n = 1$, the result is obvious. Take some $n \ge 2$, and suppose the theorem holds for all smaller values of n. Let G be a max-degree d graph on n vertices. By Lemma 2.2.2, we may take v with $|T_{N[v]}| \le \binom{d(v)+1}{3}$. Write $n = q(d + 1) + r$ for $0 \le r \le d$. Note $|T(G)| = |T(G - N[v])| + |T_{N[v]}|$. Since $G - N[v]$ has maximum degree (at most) d, if $d(v) + 1 \le r$, then induction and Lemma 2.4.2 give

$$|T(G)| \le q\binom{d+1}{3} + \binom{r - (d(v)+1)}{3} + \binom{d(v)+1}{3} \le q\binom{d+1}{3} + \binom{r}{3},$$

and if $d(v) + 1 > r$, then induction and Lemma 2.4.2 give

$$\begin{aligned} |T(G)| &\leq (q-1)\binom{d+1}{3} + \binom{d+1+r-(d(v)+1)}{3} + \binom{d(v)+1}{3} \\ &\leq q\binom{d+1}{3} + \binom{r}{3}. \end{aligned}$$

The maximum degree condition ensured $d+1+r-(d(v)+1) \geq 0$ and $d(v)+1 \leq d+1$.

Chapter 3

A random analogue of Gilbreath's conjecture

3.1 Summary

A well-known conjecture of Gilbreath, and independently Proth from the 1800s, states that if $a_{0,n} = p_n$ denotes the n^{th} prime number and $a_{i,n} = |a_{i-1,n} - a_{i-1,n+1}|$ for $i, n \geq 1$, then $a_{i,1} = 1$ for all $i \geq 1$. It has been postulated repeatedly that the property of having $a_{i,1} = 1$ for i large enough should hold for any choice of initial $(a_{0,n})_{n\geq1}$ provided that the gaps $a_{0,n+1} - a_{0,n}$ are not too large and are sufficiently random. We prove (a precise form of) this postulate.

3.2 Introduction

Given any sequence of non-negative integers $(a_n)_{n\geq1}$, we can form the sequence of non-negative integers $(|a_n - a_{n+1}|)_{n\geq1}$. Start with the primes as the initial sequence and iterate this consecutive differencing procedure. Gilbreath's conjecture is that the first term in every sequence, starting with the first iteration, is a 1. Precisely, if $a_{0,n} = p_n$ for $n \geq 1$ and $a_{i,n} = |a_{i-1,n} - a_{i-1,n+1}|$ for $i, n \geq 1$, then $a_{i,1} = 1$ for all $i \geq 1$. Below are the first few terms of the first few iterations.

2 3 5 7 11 13 17

1 2 2 4 2 4

1 0 2 2 2

1 2 0 0

1 2 0

1 2

1

Proth [54] conjectured (what later became known as) Gilbreath's conjecture in 1878, and Gilbreath independently made the same conjecture. Many sources claim Proth asserted he had a proof of the conjecture, and that his proof was wrong. However, we believe this claim is baseless. See Section 3.8 for more details. Odlyzko [49] verified Gilbreath's conjecture for $1 \leq i \leq \pi(10^{13}) \approx 3.34 \times 10^{11}$. One is led to wonder how special the primes are in Gilbreath's conjecture and whether any sequence beginning with 2 followed by an increasing sequence of odd numbers with small and "random" gaps between them will have first term 1 from some iteration onwards.

Odlyzko, at the end of Section 2 of [49], speculates that such a random sequence indeed will have first term 1 from some iteration onwards. Additionally, Problem 68 of [45] asks what gap or density properties of an initial sequence suffices to ensure the conclusion of Gilbreath's conjecture. Despite Gilbreath's conjecture being around for over a decade and several additional sources postulating that the conjecture should hold for initial sequences with small and random gaps, as of date, nothing has actually been *proven* along these lines, nor about Gilbreath's conjecture specifically.

In this chapter, we initiate a rigorous study of Gilbreath's conjecture by proving a random analogue of it.

Theorem 3.2.1. *Let $f : \mathbb{N} \to \mathbb{N}$ be an increasing function with $f(n) \leq \frac{1}{100} \frac{\log \log n}{\log \log \log n}$ for n large and $f(n) \geq 2$ for all $n \geq 1$. Let $a_1, a_2, \ldots$ be a random infinite sequence formed as follows. Let $a_1 = 2, a_2 = 3$, and for $n \geq 2$, $a_{n+1} = a_n + 2u_n$, where u_n is drawn uniformly at random from $\{0, 1, \ldots, f(n) - 1\}$, independent of the other u_i's. Then, with probability 1, there is some M_0 so that for all $M \geq M_0$, after M iterations of consecutive differencing, the first term of the sequence is a 1.*

Computations suggest that Gilbreath's conjecture holds because 0s and 2s form to the right of the leading 1 early on. We prove Theorem 3.2.1 by showing that our random initial sequence indeed has that property almost surely. Since the first iteration is $1, 2u_2, 2u_3, \ldots$, if we ignore the leading 1 and divide by 2, what we wish to show is encapsulated by the following theorem, which is the heart of the chapter.

Theorem 3.2.2. *For M large, for any C with $2 \leq C \leq \frac{1}{100} \frac{\log \log M}{\log \log \log M}$, if we form an initial sequence of length M by choosing numbers from $\{0, \ldots, C - 1\}$ independently and uniformly at random, then, with probability at least $1 - e^{-e^{\sqrt[20]{\log M}}}$, after $e^{\sqrt[5]{\log M}}$ iterations of consecutive differencing, everything is a 0 or 1.*

The randomness in Theorem 3.2.2 is certainly necessary. For example, if the initial sequence consists of only 0s and 3s, then after any number of iterations, everything is still a 0 or 3. However, there are more exotic examples of initial sequences

$$\begin{array}{c}
2\quad 0\quad 6\quad 0\quad 2\quad 2\quad 6\quad 5\quad 0\quad 0\quad 6\quad 1\quad 3\quad 2\quad 2\quad 3\quad 0\quad 6\quad 0\quad 5\\
2\quad 6\quad 6\quad 2\quad 0\quad 4\quad 1\quad 5\quad 0\quad 6\quad 5\quad 2\quad 1\quad 0\quad 1\quad 3\quad 6\quad 6\quad 5\\
4\quad 0\quad 4\quad 2\quad 4\quad 3\quad 4\quad 5\quad 6\quad 1\quad 3\quad 1\quad 1\quad 1\quad 2\quad 3\quad 0\quad 1\\
4\quad 4\quad 2\quad 2\quad 1\quad 1\quad 1\quad 1\quad 5\quad 2\quad 2\quad 0\quad 0\quad 1\quad 1\quad 3\quad 1\\
0\quad 2\quad 0\quad 1\quad 0\quad 0\quad 0\quad 4\quad 3\quad 0\quad 2\quad 0\quad 1\quad 0\quad 2\quad 2\\
2\quad 2\quad 1\quad 1\quad 0\quad 0\quad 4\quad 1\quad 3\quad 2\quad 2\quad 1\quad 1\quad 2\quad 0\\
0\quad 1\quad 0\quad 1\quad 0\quad 4\quad 3\quad 2\quad 1\quad 0\quad 1\quad 0\quad 1\quad 2\\
1\quad 1\quad 1\quad 1\quad 4\quad 1\quad 1\quad 1\quad 1\quad 1\quad 1\quad 1\quad 1\\
0\quad 0\quad 0\quad 3\quad 3\quad 0\quad 0\quad 0\quad 0\quad 0\quad 0\quad 0
\end{array}$$

for which all future iterations have only 0s and 3s (say). These exotic examples[1] suggest that we are far away from a proof of Gilbreath's conjecture.

3.3 A general bootstrapping argument

In this section, we prove a result about random walks on regular directed graphs that will be of use to proving Theorem 3.2.2.

Definition 3.3.1. A directed graph is *regular* if there is a positive integer d such that each vertex has in-degree and out-degree equal to d. We allow our graphs to have self-loops (but no multiple edges). For our discussion, a *simple random walk* on a regular directed graph of degree d is formed by choosing a starting point uniformly at random, and then walking along the directed edges, with each out-edge chosen with probability $1/d$, independent of the previous steps.

Proposition 3.3.2. *Let $G = (V, E)$ be a regular directed graph. Suppose V is red-blue colored such that the probability a simple random walk on G of length L consists entirely of red vertices is at least c. Then the probability a simple random walk on G of length $\lfloor (1 + \frac{1}{10}c)L \rfloor$ consists entirely of red vertices is at least $\frac{1}{10}c^2$.*

[1] To clarify, in the setting in which the primes are the initial sequence, the analogous situation to having only 0s and 3s is having only 0s and 6s past the first index, making the first index very likely to repeatedly change from 1 to 5 (see Lemma 3.4.5), thereby violating Gilbreath's conjecture.

Proof. Let $X_1, X_2, \ldots$ denote the steps of a simple random walk. Define functions $w_1, \ldots, w_L$ on V by $w_j(v) := \Pr(X_1, \ldots, X_L \text{ all red} | X_j = v)$. Note (by, e.g., induction on the number of steps) the regularity assumption implies

$$w_j(v) = |V| \Pr(X_1, \ldots, X_L \text{ all red}, X_j = v).$$

Thus, setting

$$\rho := \Pr(X_1, \ldots, X_L \text{ all red}),$$

we have by assumption for any j that

$$\sum_v w_j(v) = \sum_v |V| \Pr(X_1, \ldots, X_L \text{ all red}, X_j = v) = \rho|V|.$$

Let $K = \lceil \frac{3}{\rho} \rceil$, and let $k_1, \ldots, k_K$ be $k_j := \lfloor \frac{j}{K} L \rfloor$. By Cauchy-Schwarz,

$$\begin{aligned}\left(\sum_v \sum_j w_{k_j}(v)\right)^2 &\leq \left(\sum_v 1^2\right) \cdot \left(\sum_v \left(\sum_j w_{k_j}(v)\right)^2\right) \qquad (3.1)\\ &= |V| \left(\sum_j \sum_v w_{k_j}(v)^2 + \sum_{j \neq j'} \sum_v w_{k_j}(v) w_{k_{j'}}(v)\right).\end{aligned}$$

Note

$$\sum_v \sum_j w_{k_j}(v) = \sum_j \sum_v w_{k_j}(v) = K\rho|V|;$$

also, since $||w_j||_\infty \leq 1$, we have

$$\sum_j \sum_v w_{k_j}(v)^2 \leq \sum_j \sum_v w_{k_j}(v) = K\rho|V|.$$

So (3.1) implies

$$K^2\rho^2|V|^2 \leq |V| \left(K\rho|V| + \sum_{j \neq j'} \sum_v w_{k_j}(v) w_{k_{j'}}(v)\right),$$

and thus, since $K^2\rho^2|V| - K\rho|V|$ is increasing in K for (in particular) $K \geq 3/\rho$,

$$6|V| \leq \sum_{j \neq j'} \sum_v w_{k_j}(v) w_{k_{j'}}(v).$$

By the pigeonhole principle, there are $j \neq j'$ with

$$\sum_v w_{k_j}(v) w_{k_{j'}}(v) \geq \frac{1}{K^2} 6|V|.$$

Using

$$w_{k_j}(v) \leq \Pr(X_{k_j+1}, \dots, X_L \text{ all red}|X_{k_j} = v) = \Pr(X_{k_{j'}+1}, \dots, X_{L+k_{j'}-k_j} \text{ all red}|X_{k_{j'}} = v),$$

which is true merely due to translation invariance of the random walk, and the trivial

$$w_{k_{j'}}(v) \leq \Pr(X_1, \dots, X_{k_{j'}} \text{ all red}|X_{k_{j'}} = v),$$

we obtain

$$\begin{aligned}
\frac{1}{K^2}6|V| &\leq \sum_v \Pr(X_1, \dots, X_{k_{j'}} \text{ all red}|X_{k_{j'}} = v) \Pr(X_{k_{j'}+1}, \dots, X_{L+k_{j'}-k_j} \text{ all red}|X_{k_{j'}} = v) \\
&= |V| \sum_v \Pr(X_1, \dots, X_{k_{j'}} \text{ all red}, X_{k_{j'}} = v) \Pr(X_{k_{j'}+1}, \dots, X_{L+k_{j'}-k_j} \text{ all red}|X_{k_{j'}} = v) \\
&= |V| \sum_v \Pr(X_1, \dots, X_{L+k_{j'}-k_j} \text{ all red}, X_{k_{j'}} = v) \\
&= |V| \Pr(X_1, \dots, X_{L+k_{j'}-k_j} \text{ all red}),
\end{aligned}$$

yielding

$$\Pr(X_1, \dots, X_{L+k_{j'}-k_j} \text{ all red}) \geq \frac{1}{K^2}6.$$

Note $K \leq \frac{3}{\rho} + 1 \leq \frac{4}{\rho}$, so $\frac{1}{K^2}6 \geq \frac{6}{16}\rho^2 \geq \frac{1}{10}c^2$. Since the proposition is trivial if $L < 10/c$, we may assume $L \geq 10/c$ to obtain $k_{j'} - k_j \geq \frac{L}{K} - 1 \geq \frac{\rho}{4}L - 1 \geq \frac{c}{10}L$. □

Remark. It is natural to think that Proposition 3.3.2 can be extended, in some form, to arbitrary length increases. However, such an extension is not possible in general (note that iterating Proposition 3.3.2 results in only a summable geometric series of length increases). For example, consider $V = \{1, \dots, n\}, E = \{(1 \mapsto 2), \dots, (n-1 \mapsto n), (n \mapsto 1)\}$ with the vertices $\{1, \dots, \frac{1}{10}n\}$ colored red and the rest blue. Then with $L = \frac{1}{20}n$ and $c = \frac{1}{20}$, it holds that a simple random walk on G of length L will hit only red vertices with probability at least c. However, of course no simple (random) walk on G of length $5L = \frac{1}{2}n$ will hit only red vertices.

Examples of such "bad" colorings also exist on the graph we apply Proposition 3.3.2 to, namely a Debrujin graph. We don't think these colorings are actually the ones we need to address in our proof of Theorem 3.2.2, but we couldn't prove that.

3.4 A lower bound for ending with 0

We begin by exploiting the main property of the "dynamical system" of taking consecutive differences: the supremum never increases. In fact, we use that it quickly decreases provided there is no trivial obstruction to it doing so (Lemma 3.4.2).

Definition 3.4.1. We say non-negative integers $a_1, \ldots, a_i$ *come from* $\widetilde{a}_1, \ldots, \widetilde{a}_{i+1}$ if $|\widetilde{a}_j - \widetilde{a}_{j+1}| = a_j$ for each $1 \leq j \leq i$. Given $a_1, \ldots, a_i$ and a subset $E \subseteq \mathbb{Z}$, an *E-block* is a contiguous set of terms $a_{j_1+1}, \ldots, a_{j_1'}$ such that $a_j \in E$ for each $j_1 + 1 \leq j \leq j_1'$; the *length* of the block is $j_1' - j_1$.

Lemma 3.4.2. *Let $a_1, \ldots, a_i$ be non-negative integers with $d := \max_j a_j$. Let L denote the length of the longest $\{0, d\}$-block containing at least one d. If $L \leq i - 1$, then, after L iterations of consecutive differencing, the largest number is at most $d - 1$.*

Proof. We induct on L. For $L = 1$, the result is clear. Assume $L \geq 2$ and the result is true for all $L' < L$. It is easy to see that, since d is the maximum, any $\{0, d\}$-block containing a d after an iteration would have had to have come from a $\{0, d\}$-block of greater length containing a d, so the longest $\{0, d\}$-block containing a d after one iteration is at most $L - 1$, say L'. By induction, after L' more iterations, the largest number is at most $d - 1$. It follows that after L (total) iterations, the largest number is at most $d - 1$. □

So, to prove Theorem 3.2.2, "all" we need to do is argue that long $\{0, d\}$-blocks are unlikely to exist. In this next lemma, we observe that any large $\{0, d\}$-block essentially must have come from a block with no 0s.

Lemma 3.4.3. *Suppose that after i iterations, there is a $d\mathbb{Z}$-block of length L. Then either there was a $d\mathbb{Z}$-block of length $L + i$ in the initial sequence, or there is some i', $0 \leq i' \leq i - 1$, such that after i' iterations, there is a block of length $L + i - i'$ with no 0s.*

Proof. We prove by induction on i the statement for all L. For $i = 0$, the result is tautological. Take $i \geq 1$, and suppose the result holds for $i - 1$. The $d\mathbb{Z}$-block of length L had to come from either a $d\mathbb{Z}$-block of length $L + 1$ or a block of length $L + 1$ with no 0s (since everything will have the same residue modulo d), so we are done by the induction hypothesis. □

Another nice property of the consecutive differencing operation is that it "commutes" with reducing mod 2. This allows for a decently explicit formula for the parity of a term after a given number of iterations, merely in terms of the parities of the initial terms.

Definition 3.4.4. For non-negative integers a_1, a_2, define $f_1(a_1, a_2) = |a_1 - a_2|$, and for any $i \geq 2$ and non-negative $a_1, \ldots, a_{i+1}$, define $f_i(a_1, \ldots, a_{i+1}) = |f_{i-1}(a_1, \ldots, a_i) - f_{i-1}(a_2, \ldots, a_{i+1})|$. We say $a_1, \ldots, a_{i+1}$ *ultimately iterate* to $f_i(a_1, \ldots, a_{i+1})$.

Lemma 3.4.5. *For any $i \geq 1$, there is a subset $J_i \subseteq [i+1]$ containing 1 and $i+1$ so that for any non-negative integers $a_1, \dots, a_{i+1}$, $f_i(a_1, \dots, a_{i+1}) \equiv \sum_{j \in J_i} a_j \mod 2$.*

Proof. We induct on i. For $i = 1$, the result follows from $|a_1 - a_2| \equiv a_1 + a_2 \mod 2$. Assume $i \geq 2$ and the result is true for $i-1$. Note that $f_i(a_1, \dots, a_{i+1}) \equiv |f_{i-1}(a_1, \dots, a_i) - f_{i-1}(a_2, \dots, a_{i+1})| \equiv f_{i-1}(a_1, \dots, a_i) + f_{i-1}(a_2, \dots, a_{i+1}) \equiv$
$\sum_{j \in J_{i-1}} a_j + \sum_{j \in J_{i-1}} a_{j+1} \equiv \sum_{j \in J_{i-1} \triangle (J_{i-1}+1)} a_j \mod 2$. By induction, J_{i-1} contains 1 and i, and so $J_i := J_{i-1} \triangle (J_{i-1} + 1)$ contains 1 and $i+1$, as desired. □

We take a moment to note a useful immediate corollary of Lemma 3.4.5 which tells us that the parity of what $a_1, \dots, a_{i+1}$ ultimately iterate to depends linearly on each of the parities of a_1 and a_{i+1}.

Corollary 3.4.6. *Let $a_1, \dots, a_{i+1}$ be drawn independently, uniformly at random from $\{0, \dots, C-1\}$. Then, the probability $a_1, \dots, a_{i+1}$ ultimately iterate to an even integer is between $\frac{1}{3}$ and $\frac{2}{3}$. Furthermore, for any applicable j, T, the probability that all of $f_j(a_t, \dots, a_{t+j})$ are even for T consecutive values of t is exponentially small in T.*

Let $[C]_0 = \{0, \dots, C-1\}$.

The following proposition shows that 0s are never too rare, which will be useful in conjuction with Lemma 3.4.3. Before the proof, we introduce some notation for a given C and i. Define $i_0 = i$ and $i_{j+1} = \lfloor \frac{i_j}{1000C} \rfloor$ for $0 \leq j \leq C-3$. For $1 \leq j \leq C-2$, let E_j denote the event that after $i - i_{j-1}$ iterations there's a $\{0, C-j\}$-block of length (at least) $i_{j-1} - i_j$. For example, E_1 is the event that after 0 iterations, there's a $\{0, C-1\}$-block of length $i - i_1$, and E_2 is the event that after $i - i_1$ iterations, there's a $\{0, C-2\}$-block of length $i_1 - i_2$.

Proposition 3.4.7. *For any $C \geq 2$ and any $i \geq (2000C)^{2C}$, if $a_1, \dots, a_i$ are chosen independently and uniformly at random from $\{0, \dots, C-1\}$, then the probability they ultimately iterate to 0 is at least $\frac{1}{4000C^2}$.*

Proof. Fix $C \geq 2$ and $i \geq (2000C)^{2C}$. If $C = 2$, then Corollary 3.4.6 gives the result, so assume $C \geq 3$. We may suppose that the desired probability is at most 0.01. Let $\mathcal{B}_0$ denote all i-tuples in $[C]_0^i$ that ultimately iterate to something 0 mod 2; we say "conditional probability" when speaking of the conditional probability that $\mathcal{B}_0$ induces. Then, by Corollary 3.4.6, the conditional probability of ultimately iterating to 0 is at most 0.03, and so the conditional probability of not having only 0s and 1s after some iteration is at least 0.97.

Therefore, with notation as defined above Proposition 7.2.2, with conditional probability at least 0.97 some E_j occurs. Indeed, otherwise, repeated use of Lemma 3.4.2 shows that after $i-i_{C-2}$ iterations, everything is a 0 or 1: after $i-i_1$ iterations, there are no more $(C-1)$s and thus no $(C-1)$s ever again; after $i-i_2$ iterations, there are no more $(C-2)$s and thus no $(C-2)$s ever again, etc..

Consequently, by the pigeonhole principle, there is some j, $1 \leq j \leq C-2$, such that E_j occurs with conditional probability at least $\frac{0.97}{C-2}$. Clearly j cannot be 1, since we have the uniform distribution after 0 iterations. Also, j must be such that $C-j$ is odd, since by Corollary 3.4.6, the probability of having $i_{j-1}-i_j$ evens in a row is at most $(\frac{2}{3})^{i_{j-1}-i_j}$, which is at most $(\frac{2}{3})^{2(2000C)^C}$ since, as are easy to verify, $i_{j-1}-i_j \geq 2i_j$ and that $i \geq (2000C)^{2C}$ implies $i_j \geq i_{C-2} \geq (2000C)^C$ for each j. Since after $i-i_{j-1}$ iterations, there are only i_{j-1} indices, a block of length $i_{j-1}-i_j$ must contain the block $[i_j+1, i_{j-1}-i_j]$ (see Figure 1). So, with conditional probability at least $\frac{0.97}{C-2}$, all indices $i_j+\Delta$, for $1 \leq \Delta \leq i_{j-1}-2i_j$, will be 0 or $C-j$.

With $a_1, \ldots, a_i$ denoting the initial sequence, note that after $i-i_{j-1}$ iterations, none of the indices $i_j+\Delta$, $1 \leq \Delta \leq i_{j-1}-2i_j$, depend on a_1 or a_i (only the first and last indices do). Therefore, by Corollary 3.4.6, we see that with (unconditional) probability at least $\frac{1}{3}\frac{0.97}{C-2} \geq \frac{0.30}{C-2}$, all $i_j+\Delta$ will be 0 or $C-j$. Note that after $\bar{i} := i-i_{j-1}$ iterations, the integer at any index r is equal to $f_{\bar{i}}(a_r, a_{r+1}, \ldots, a_{r+\bar{i}})$.

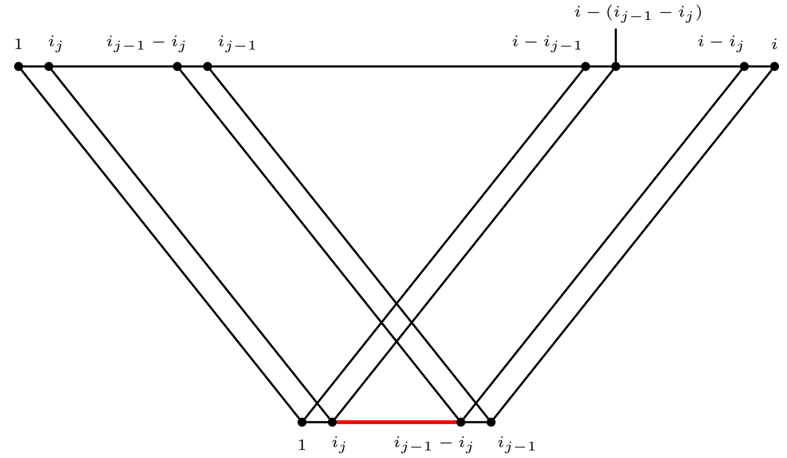


Figure 1: Indicates which initial indices (in $[i]$) a particular index after $\bar{i}$ iterations depends on.

Define a (regular) directed graph on $[C]_0^{\bar{i}+1}$ by $(x_1, \ldots, x_{\bar{i}+1}) \to (x_2, \ldots, x_{\bar{i}+1}, y)$ for any $x_1, \ldots, x_{\bar{i}+1}, y \in [C]_0$. Color a tuple $(x_1, \ldots, x_{\bar{i}+1}) \in [C]_0^{\bar{i}+1}$ "red" if and only if it ultimately iterates to 0 or $C-j$. The fact that, with probability at least $\frac{0.30}{C-2}$, all $f_{\bar{i}}(a_r, a_{r+1}, \ldots, a_{r+\bar{i}})$, for $i_j+1 \leq r \leq i_{j-1}-i_j$, are 0 or $C-j$ corresponds exactly to:

with probability at least $\frac{0.30}{C-2}$, a simple random walk in $[C]_0^{\bar{i}+1}$ of length $L := i_{j-1} - 2i_j$ consists entirely of red vertices.

Proposition 3.3.2 now tells us that with probability at least $\frac{1}{200C^2}$, a simple random walk of length[2] $(1 + \frac{1}{200C})L$ consists entirely of red vertices. Note $(1 + \frac{1}{200C})L \geq (1 + \frac{1}{400C})i_{j-1}$, since it is equivalent to $\frac{1}{400C}i_{j-1} \geq (2 + \frac{1}{200C})i_j$, which is true since $i_j \leq \frac{i_{j-1}}{1000C}$. We have thus shown that, if $a_1, \ldots, a_{(1+\frac{1}{400C})i_{j-1}+\bar{i}}$ are chosen independently and uniformly at random from $[C]_0$, then with probability at least $\frac{1}{200C^2}$, all $f_{\bar{i}}(a_r, \ldots, a_{r+\bar{i}})$ for $1 \leq r \leq (1 + \frac{1}{400C})i_{j-1}$ are either 0 or $C - j$.

We are nearly done, as, for $\ell := i_{j-1}$, we have that $(f_{\bar{i}}(a_r, \ldots, a_{r+\bar{i}}))_{1 \leq r \leq \ell}$ is the whole sequence after $\bar{i}$ iterations; since a $\{0, C-j\}$-block ultimately iterates to either 0 or $C-j$ and since $C-j$ is odd, we just need to additionally ensure that the ultimate iterate is even.

We now deduce that, if $a_1, \ldots, a_i$ are chosen independently and uniformly at random from $[C]_0$, then with probability at least $\frac{1}{4000C^2}$, they ultimately iterate to something 0 mod 2 and each $f_{\bar{i}}(a_r, \ldots, a_{r+\bar{i}})$, for $1 \leq r \leq \ell$, is either 0 or $C - j$. Let $\delta = \frac{1}{400C^2}$. By Corollary 3.4.6, the proportion of walks $(X_1, \ldots, X_{(1+\delta)\ell})$ in $[C]_0^{\bar{i}}$ of length $(1+\delta)\ell$ that have at most $\frac{\delta\ell}{8}$ values of $j \in [\delta\ell]$ with[3] $(X_{j+1}, X_{j+2}, \ldots, X_{j+\ell}) \in \mathcal{B}_0$ is at most $\frac{\delta\ell}{8}\binom{\delta\ell}{\delta\ell/8}(2/3)^{\delta\ell}$. Note that

$$\frac{\delta\ell}{8}\binom{\delta\ell}{\delta\ell/8}(2/3)^{\delta\ell} \leq \frac{1}{400C^2}; \tag{3.2}$$

indeed, the general

$$\binom{n}{k} \leq (\frac{en}{k})^k$$

implies

$$\begin{aligned}\frac{\delta\ell}{8}\binom{\delta\ell}{\delta\ell/8}(2/3)^{\delta\ell} &\leq \frac{\delta\ell}{8}\left(\frac{e\delta\ell}{\delta\ell/8}\right)^{\delta\ell/8}(2/3)^{\delta\ell} \\ &< \frac{\delta\ell}{8}(0.98)^{\delta\ell},\end{aligned}$$

which gives (3.2) since $\delta\ell \geq \frac{1}{80C^2}(400C^2)^C$. Therefore, since the proportion of walks $(X_1, \ldots, X_{(1+\delta)\ell})$ with $X_1, \ldots, X_{(1+\delta)\ell}$ all red is at least $\frac{1}{200C^2}$, if we let $\mathcal{A}$ denote the walks $(X_1, \ldots, X_{(1+\delta)\ell})$ such that $X_1, \ldots, X_{(1+\delta)\ell}$ are all red and such that there are

[2] To be light on notation, we suppress ceiling and floor functions in the rest of this section.

[3] Here we have abused notation, by associating the i-tuple that $X_{j+1}, \ldots, X_{j+\ell}$ form with $(X_{j+1}, \ldots, X_{j+\ell})$.

at least $\frac{\delta\ell}{8}$ values of j with $(X_{j+1}, X_{j+2}, \ldots, X_{j+\ell}) \in \mathcal{B}_0$, then the density of $\mathcal{A}$ is at least $\frac{1}{400C^2}$. So on one hand,

$$\sum_{(X_1,\ldots,X_{(1+\delta)\ell})\in\mathcal{A}} \sum_{j=1}^{\delta\ell} 1_{(X_{j+1},\ldots,X_{j+\ell})\in\mathcal{B}_0} \geq \frac{\delta\ell}{8}\frac{1}{400C^2}C^{\bar{i}}C^{(1+\delta)\ell-1},$$

while on another hand,

$$\begin{aligned}
&\sum_{(X_1,\ldots,X_{(1+\delta)\ell})\in\mathcal{A}} \sum_{j=1}^{\delta\ell} 1_{(X_{j+1},\ldots,X_{j+\ell})\in\mathcal{B}_0} \\
&\qquad = \sum_{j=1}^{\delta\ell} \sum_{(X_{j+1},\ldots,X_{j+\ell})\in\mathcal{B}_0} \sum_{\substack{X_1,\ldots,X_j,X_{j+\ell+1},\ldots,X_{(1+\delta)\ell} \\ (X_1,\ldots,X_{(1+\delta)\ell})\in\mathcal{A}}} 1 \\
&\qquad \leq \sum_{j=1}^{\delta\ell} \sum_{(X_{j+1},\ldots,X_{j+\ell})\in\mathcal{B}_0} C^{\delta\ell}\, 1_{X_{j+1},\ldots,X_{j+\ell} \text{ all red}} \\
&\qquad = \delta\ell C^{\delta\ell} \sum_{(X_1,\ldots,X_\ell)\in\mathcal{B}_0} 1_{X_1,\ldots,X_\ell \text{ all red}}.
\end{aligned}$$

We deduce that

$$\sum_{(X_1,\ldots,X_\ell)\in\mathcal{B}_0} 1_{X_l,\ldots,X_\ell \text{ all red}} \geq \frac{1}{3200C^2}C^{\bar{i}}C^{\ell-1},$$

which is what we wanted to deduce. □

Corollary 3.4.8. *For any $C \geq 2$ and any $i \geq 1$, if $a_1, \ldots, a_i$ are chosen independently and uniformly at random from $\{0, \ldots, C-1\}$, then the probability they ultimately iterate to 0 is at least $(\frac{1}{C})^{(2000C)^{2C}}$.*

Proof. For $i \geq (2000C)^{2C}$, Proposition 7.2.2 yields a lower bound of $\frac{1}{4000C^2}$, and for $1 \leq i < (2000C)^{2C}$, we use the trivial lower bound coming from $a_j = 0$ for all j. □

3.5 Finishing the proof of main theorem

We now finish the proof of Theorem 3.2.2, copied below for the reader's convenience.

Theorem 3.2.2: For M large, for any C with $2 \leq C \leq \frac{1}{100}\frac{\log\log M}{\log\log\log M}$, if we form an initial sequence of length M by choosing numbers from $\{0, \ldots, C-1\}$ independently and uniformly at random, then, with probability at least $1 - e^{-e^{\sqrt[20]{\log M}}}$, after $e^{\sqrt[5]{\log M}}$ iterations of consecutive differencing, everything is a 0 or 1.

Fix M large and C in the range $[3, \frac{1}{100}\frac{\log\log M}{\log\log\log M}]$ (the case $C = 2$ is trivial). Let E_1 denote[4] the event that after 0 iterations, there is a $\{0, C-1\}$-block of length $R := e^{\sqrt[10]{\log M}}$. Let E_2 be the event that after $2R$ iterations, there is a $\{0, C-2\}$-block of length R^2. Let E_3 be the event that after $2R^2$ iterations, there is a $\{0, C-3\}$-block of length R^3. In general, for $2 \leq j \leq C-2$, E_j is the event that after $2R^{j-1}$ iterations, there is a $\{0, C-j\}$-block of length R^j. Since $2R^{j-1} \geq 2R^{j-2}+R^{j-1}$ for $3 \leq j \leq C-1$, we see that, as before, by Lemma 3.4.2, if no E_j occurs, then after $2R^{C-2}$ iterations, everything is a 0 or a 1. Note that $2R^{C-2} \leq e^{\log^{1/5} M}$, so it suffices to show that the probability that some E_j occurs is at most $\exp\left(-e^{\log^{1/20} M}\right)$. By the union bound, it suffices to show $\Pr(E_j) \leq \exp\left(-e^{\log^{1/13} M}\right)$, say, for each $1 \leq j \leq C-2$.

Clearly, $\Pr(E_1) \leq M(\frac{2}{3})^R \leq \exp\left(-e^{\log^{1/13} M}\right)$, so fix some j with $2 \leq j \leq C-2$. By Lemma 3.4.3, if E_j occurs, either there is a $(C-j)\mathbb{Z}$-block of length R^j in the initial sequence or there is a block of length R^j in the first $2R^{j-1}-1$ iterations containing no 0s. Once again, the first option holds with probability at most $M(\frac{2}{3})^{R^j} \leq \frac{1}{2}\exp\left(-e^{\log^{1/13} M}\right)$, so by the union bound, it suffices to show that for each $0 \leq i \leq 2R^{j-1}-1$, the probability that there is a block of length $L := R^j = e^{j\log^{1/10} M}$ without 0s after i iterations is at most $\exp\left(-e^{\log^{1/12} M}\right)$, say.

So fix some $i \in [0, 2R^{j-1}-1]$. Let $b_1, \ldots, b_{M-i}$ denote the sequence after i iterations. Let's first focus on the block $b_1, \ldots, b_L$. Say the initial sequence is $a_1, \ldots, a_M$. Note that $b_{k(i+1)+1} = f_i(a_{k(i+1)+1}, \ldots, a_{(k+1)(i+1)})$ for $0 \leq k \leq \frac{1}{2}R - 1$. Since $(\frac{1}{2}R-1)(i+1)+1 \leq \frac{1}{2}R(i+1) \leq L$ and the sets $\{a_{k(i+1)+1}, \ldots, a_{(k+1)(i+1)}\}$ are disjoint as k ranges, by independence the probability that $b_1, \ldots, b_L$ are all nonzero is at most $\left(1-(\frac{1}{C})^{(400C^2)^{2C}}\right)^{R/2}$ by Corollary 3.4.8. Using the standard inequality $1-x \leq e^{-x}$, we see that

$$\begin{aligned}\left(1-(\frac{1}{C})^{(400C^2)^{2C}}\right)^{R/2} &\leq \exp\left(-\frac{R}{2}(\frac{1}{C})^{(400C^2)^{2C}}\right)\\ &\leq \exp\left(-\frac{R}{2}e^{-(\log C)e^{5C\log C}}\right)\\ &\leq \exp\left(-\frac{R}{2}e^{-(\log\log\log M)e^{\frac{1}{19}\log\log M}}\right)\\ &\leq \exp\left(-\frac{R}{2}e^{-\sqrt[15]{\log M}}\right)\\ &\leq \exp\left(-e^{\sqrt[11]{\log M}}\right).\end{aligned}$$

[4] To be light on notation, we suppress ceiling and floor functions in this section.

Therefore, by the union bound, the probability that there is some block of length L after i iterations containing no 0s is at most $M \exp\left(-e^{\log^{1/11} M}\right) \leq \exp\left(-e^{\log^{1/12} M}\right)$. The proof is thus complete. □

3.6 Proof of random analogue of Gilbreath's conjecture

In this section we deduce Theorem 3.2.1 from Theorem 3.2.2. We start with a lemma.

Lemma 3.6.1. *Take M large. Let $f : [M] \to \{2, 3, \ldots, \lfloor \frac{1}{100} \frac{\log \log M}{\log \log \log M} \rfloor\}$ be an increasing function. Form a random initial sequence $b_1, \ldots, b_M$ by choosing b_m uniformly at random from $\{0, 1, \ldots, f(m) - 1\}$, independently of the other b_i's. Then, with probability at least $1 - e^{-\frac{1}{20} \log^2 M}$, after $3 \frac{M}{\log^2 M}$ iterations of consecutive differencing, everything is a 0 or 1.*

Before proving Lemma 3.6.1, let's prove Theorem 3.2.1, copied below, assuming it.

Theorem 3.2.1: Let $f : \mathbb{N} \to \mathbb{N}$ be an increasing function with $f(n) \leq \frac{1}{100} \frac{\log \log n}{\log \log \log n}$ for n large and $f(n) \geq 2$ for all $n \geq 1$. Let $a_1, a_2, \ldots$ be a random infinite sequence formed as follows. Let $a_1 = 2, a_2 = 3$, and for $n \geq 2$, $a_{n+1} = a_n + 2u_n$, where u_n is drawn uniformly at random from $\{0, 1, \ldots, f(n) - 1\}$, independent of the other u_i's. Then, with probability 1, there is some M_0 so that for all $M \geq M_0$, after M iterations of consecutive differencing, the first term of the sequence is a 1.

Proof of Theorem 3.2.1. Let A_M denote the event that after M iterations, the first term is not a 1. We wish to show that, with probability 1, only finitely many A_M's occur. By Borel-Cantelli, it suffices to show that for all M large, the probability of A_M occurring is at most $e^{-\frac{1}{30} \log^2 M}$. Note that A_M is equivalent to $a_1, \ldots, a_{M+1}$ not ultimately iterating to 1. For M large enough, by Lemma 3.6.1, with probability at least $1 - e^{-\frac{1}{20} \log^2 M}$, after $3 \frac{M}{\log^2 M}$ iterations of consecutive differencing beginning with initial sequence $u_2, \ldots, u_M$, everything is a 0 or 1. Therefore, with probability at least $1 - e^{-\frac{1}{20} \log^2 M}$, after $3 \frac{M}{\log^2 M}$ iterations of consecutive differencing beginning with initial sequence $2u_2, \ldots, 2u_M$, everything is a 0 or 2. It follows that with probability at least $1 - e^{-\frac{1}{20} \log^2 M}$, after $1 + 3 \frac{M}{\log^2 M}$ iterations of consecutive differencing beginning with initial sequence $a_1, \ldots, a_{M+1}$, the obtained sequence starts off with an odd number at most $\frac{1}{100} \frac{\log \log M}{\log \log \log M}$ followed by only 0s and 2s. Since this odd number, whenever at least 3, decreases by 2 at each iteration in which the second (and adjacent) term of the sequence is 2, we wish to show that there are many 2s amongst the (only) 0s and 2s that follow; indeed, then the first term will become a 1 and consequently stay a 1 throughout the remaining iterations, since everything that follows is either a 0 or a 2.

By Corollary 3.4.6, with probability at least $1 - e^{-\frac{1}{10}\log^2 M}$, the second term of the sequence is congruent to 2 mod 4 at least $\frac{1}{3}\log^2 M$ times out of the $\log^2 M$ iterations following the $(1 + 3\frac{M}{\log^2 M})^{\text{th}}$ iteration. Therefore, with probability at least $1 - e^{-\frac{1}{20}\log^2 M} - e^{-\frac{1}{10}\log^2 M} \geq 1 - e^{-\frac{1}{30}\log^2 M}$, starting with $a_1, \ldots, a_{M+1}$, after $1 + 3\frac{M}{\log^2 M} + \log^2 M$ iterations, the first term will be a 1, and therefore will remain a 1 all the way until the final (i.e., M^{th}) iteration, since everything else is a 0 or 2. □

We finish by proving Lemma 3.6.1. We begin with a definition.

Definition 3.6.2. Let $a_1, \ldots, a_{M+1}$ be non-negative integers. We say that an index $i \in [M+1]$ *influenced* the index $j \in [M+1-t]$ after t iterations if $0 \leq i - j \leq t$. Recall that $f_t(a_j, \ldots, a_{j+t})$ is the value at index j after t iterations.

The idea of the proof of Lemma 3.6.1 is as follows. By Theorem 3.2.2, the blocks on which f is constant will become all 0s and 1s after not too many iterations. Although there are some indices that were influenced by initial indices on which f took different values, these indices are contained in not too many not too large intervals (since f is increasing), so we can let all the 0s and 1s drop the values at these "bad indices" with a few extra iterations.

We start by proving a lemma that allows us to isolate these "bad indices". For an interval $I \subseteq \mathbb{N}$, let $L(I)$ and $R(I)$ denote its left and right endpoints, respectively.

Lemma 3.6.3. *Suppose M is large, and let C_M be a positive integer with $C_M \leq \log\log M$. Let $I_1, \ldots, I_r \subseteq [M]$ be disjoint intervals with $r \leq C_M$ and $|I_t| \leq C_M e^{\sqrt[5]{\log M}}$ for each t. Then there are pairwise disjoint intervals $J_1, \ldots, J_s \subseteq [M]$, each containing some I_t, such that the following two hold.*

- *For all t, $1 \leq t \leq r$, there is some m with $I_t \subseteq J_m$.*
- *For any m, $1 \leq m \leq s$, if we let B_m denote the smallest interval containing all of the I_t's in J_m, then we have that either $L(B_m) - L(J_m) \geq (\log^2 M)^{C_M}|B_m|$ or $R(J_m) - R(B_m) \geq (\log^2 M)^{C_M}|B_m|$, with both being true if J_m contains neither 1 nor M.*

Proof. For a subset A of $[r]$, let B_A denote the smallest interval containing $\cup_{t\in A} I_t$, and let $J(A)$ denote the smallest interval containing $\cup_{t\in A} I_t$ such that either $L(B_A) - L(J(A)) \geq (\log^2 M)^{C_M}|B_A|$ or $R(J(A)) - R(B_A) \geq (\log^2 M)^{C_M}|B_A|$, with both required to be true if $J(A)$ contains neither 1 nor M; if no such interval exists, we let $J(A) = \emptyset$. We construct a finite sequence of sets $\mathcal{C}_0, \mathcal{C}_1, \ldots$ in an iterative manner as follows. Let $\mathcal{C}_0 = \{J(\{t\}) : 1 \leq t \leq r\}$. Assume we have defined $\mathcal{C}_j$ for $0 \leq j \leq i$. If $\mathcal{C}_i$ contains distinct intervals $J(A_1), J(A_2)$ that intersect, we define[5] $\mathcal{C}_{i+1}$ to be the same as $\mathcal{C}_i$, except we replace $J(A_1)$ and $J(A_2)$ with $J(A_1 \cup A_2)$; otherwise, we terminate the construction (of the $\mathcal{C}_j$'s). It is clear that the construction will terminate after at most r steps, say after k steps. Let $\mathcal{C}_0, \ldots, \mathcal{C}_{k-1}$ be the constructed collections. It is clear that if each element of $\mathcal{C}_{k-1}$ is non-empty, then the elements of $\mathcal{C}_{k-1}$ satisfy the conditions of Lemma 3.6.3. The largest diameter of an interval in $\mathcal{C}_0$ is at most $(2(\log^2 M)^{C_M} + 1)C_M e^{\sqrt[5]{\log M}} \leq 3(\log^2 M)^{C_M} C_M e^{\sqrt[5]{\log M}}$. If $J(A_1)$ and $J(A_2)$ each have diameter at most D and intersect, then the diameter of $J(A_1 \cup A_2)$ is at most $(2(\log^2 M)^{C_M} + 1)(2D) \leq 6(\log^2 M)^{C_M} D$. Therefore, each interval in any $\mathcal{C}_{i-1}$ has diameter at most $6^{i-1}(\log^2 M)^{(i-1)C_M} 3(\log^2 M)^{C_M} C_M e^{\sqrt[5]{\log M}} \leq 6^r(\log^2 M)^{rC_M} C_M e^{\sqrt[5]{\log M}} \leq e^{\sqrt[4]{\log M}}$. To finish the proof, it just remains to note that $J(A) \neq \emptyset$ if the diameter of $\cup_{t\in A} I_t$ is at most $e^{\sqrt[4]{\log M}}$. □

Proof of Lemma 3.6.1. Do $e^{\sqrt[5]{\log M}}$ iterations of consecutive differencing. For $2 \leq C \leq \frac{1}{100}\frac{\log\log M}{\log\log\log M} =: C_M$, we say that an index j is C-*pure* if f took the value C at all indices in the initial sequence that influenced j (after $e^{\sqrt[5]{\log M}}$ iterations). Let I denote the indices that are not C-pure for any C. Write $I = \sqcup_{t=1}^{r} I_t$ as a disjoint union of intervals with r minimal. Clearly $r \leq C_M$. Also, crudely, $|I_t| \leq C_M e^{\sqrt[5]{\log M}}$ for each t; indeed, since f is increasing and is always between 2 and C_M, there are at most C_M indices at which f strictly increased, and after $e^{\sqrt[5]{\log M}}$ iterations, there are thus at most $C_M e^{\sqrt[5]{\log M}}$ indices which were influenced by two indices at which f took different values.

Let $J_1, \ldots, J_s$ be the intervals guaranteed[6] by Lemma 3.6.3, and let $B_1, \ldots, B_s$ be as in Lemma 3.6.3. For any C, by[7] Theorem 3.2.2 applied to the (interval of) C-pure

[5] It does not matter, but $\mathcal{C}_{i+1}$ thus could depend on the choice of intersecting intervals.

[6] We are applying Lemma 3.6.3 with $M - e^{\sqrt[5]{\log M}}$ instead of M, but all bounds are essentially the same.

[7] As stated, Theorem 3.2.2 only applies to initial sequences of length M. However, given any shorter initial sequence, we can independently add elements uniformly chosen from $\{0, \ldots, C-1\}$ to obtain a sequence of length M, then do $e^{\sqrt[5]{\log M}}$ iterations, and then truncate the sequence to keep only indices influenced by the original initial sequence.

indices, the probability that all C-pure indices are 0 or 1 is at least $1 - e^{-e^{\sqrt[20]{\log M}}}$, and therefore the probability that all indices that are C-pure for some C are 0 or 1 is at least $1 - C_M e^{-e^{\sqrt[20]{\log M}}} \geq 1 - e^{-\sqrt[21]{\log M}}$. In particular, with probability at least $1 - e^{-\sqrt[21]{\log M}}$, all indices in $\cup_{m=1}^{s}(J_m \setminus B_m)$ are 0 or 1; going forward, we condition on this being the case. For $1 \leq m \leq s$ and $1 \leq j \leq C_M - 1$, let J_m^j denote the interval (of length $|J_m| - 2(\log^2 M)^j |B_m|$) whose indices after $2(\log^2 M)^j |B_m|$ iterations past the $e^{\sqrt[5]{\log M}}$th are influenced by indices only in J_m, and let B_m^j denote the interval (of length $|B_m| + 2(\log^2 M)^j |B_m|$) whose indices after $2(\log^2 M)^j |B_m|$ iterations past the $e^{\sqrt[5]{\log M}}$th are influenced by at least one index in B_m. Note that Lemma 3.6.3 implies $B_m^j \subseteq J_m^j$ for each $1 \leq j \leq C_M - 1$ (since $2(\log^2 M)^{C_M - 1}|B_m| \leq (\log^2 M)^{C_M}|B_m|$).

For $1 \leq m \leq s$, let E_m^0 denote the event that there is a $\{0, C_M\}$-block in J_m of length $(\log^2 M)|B_m|$ containing a C_M. For $1 \leq m \leq s$ and $1 \leq j \leq C_M - 2$, let E_m^j denote the event that, after $2(\log^2 M)^j |B_m|$ iterations (past the $e^{\sqrt[5]{\log M}}$th), there is a $\{0, C_M - j\}$-block in J_m^j of length $(\log^2 M)^{j+1}|B_m|$ containing a $C_M - j$. Fix m with $1 \leq m \leq s$. As in the proofs of Proposition 7.2.2 and Theorem 3.2.2, since $2(\log^2 M)^{i+1}|B_m| \geq (\log^2 M)^{i+1}|B_m| + 2(\log^2 M)^i |B_m|$, if none of $E_m^0, E_m^1, \ldots, E_m^{C_M - 2}$ occur, then after $2(\log^2 M)^{C_M - 1}$ iterations, the largest number in $J_m^{C_M - 1}$ is a 1.

We claim first that the probability E_m^0 occurs is at most $2(\frac{1}{2})^{\frac{1}{2}\log^2 M}$. Indeed, if J_m contains a $\{0, C_M\}$-block of length $(\log^2 M)|B_m|$, then at least $(\log^2 M - 1)|B_m|$ of that $\{0, C_M\}$-block must lie outside of B_m, and thus in $J_m \setminus B_m$, where everything is 0 or 1. Therefore, either to the left or to the right of B_m must be at least $\frac{1}{2}\log^2 M|B_m|$ consecutive 0's, so our claim follows from Corollary 3.4.6.

Similarly, the length of the longest $\{0, C_M - j\}$-block in J_m^j is at most the whole of B_m^j and 0s surrounding it, so the probability E_m^j occurs is at most $2(\frac{1}{2})^{\frac{1}{4}\log^2 M}$. Therefore, the probability that at least one of $E_m^0, \ldots, E_m^{C_M - 2}$ occurs is at most $2(\frac{1}{2})^{\frac{1}{2}\log^2 M} + (C_M - 2)2(\frac{1}{2})^{\frac{1}{4}\log^2 M} \leq e^{-\frac{1}{10}\log^2 M}$. Since $B_m^{C_M - 1} \subseteq J_m^{C_M - 1}$, if none of $E_m^0, \ldots, E_m^{C_m - 2}$ occur, then the elements of (the growing) B_m became 0 and 1 quickly enough to not affect anything outside of (the shrinking) J_m. In particular, if for each m, none of $E_m^0, \ldots, E_m^{C_M - 2}$ occur, then[8] after $2(\log^2 M)^{C_M - 1}\max_{1 \leq m \leq s}|B_m| \leq 2\frac{M}{\log^2 M}$ iterations past the $e^{\sqrt[5]{\log M}}$th, everything is a 0 or 1. Since the probability at least one E_m^j (over all j, m) occurs is at most $se^{-\frac{1}{10}\log^2 M} \leq e^{-\frac{1}{20}\log^2 M}$, Lemma 3.6.1 is established. □

[8] It is clear from Lemma 3.6.3 that $|B_m| \leq \frac{M}{(\log^2 M)^{C_M}}$ for each m.

3.7 Additional mathematical remarks

The proof of Theorem 3.2.2 can be relatively easily adapted to handle any distribution (not just the uniform distribution) on $\{0, \dots, C-1\}$ that gives not too large, positive weight to each of $0, \dots, C-1$ (one should create duplicate vertices in $[C]_0^i$ so that the obtained simple random walk models this different probability distribution).

In Theorem 3.2.2 we did not try to optimize $e^{-e^{\sqrt[20]{\log M}}}$ nor $e^{\sqrt[5]{\log M}}$. A proof allowing C to go all the way up to $\log^2 M$, or even a power of M, would be interesting. We expect that, in reality, the highest C can go is M, in that if $C = o(M)$, then with probability $1 - o(1)$, after $\frac{M}{2}$ iterations, everything is a 0 or 1, while if $C = \omega(M)$, with probability $o(1)$, after $\frac{M}{2}$ iterations, everything is a 0 or 1.

3.8 A historical remark

Various sources (websites, blog posts, etc.) have claimed that Proth believed he had proven Gilbreath's conjecture, and that his proof turned out to be wrong.

Not only do we currently have no evidence for this claim, the apparent source of this claim has retracted it.

The claim seemed plausible, for Proth did publish a paper [54] on (what later became known as) Gilbreath's conjecture and did, admittedly confusingly, call it a "theorem". However, a reading through the paper shows he did not seriously claim a proof. Indeed, Hugh Williams who made the claim about Proth without reference [61, p. 123], said "On rereading his actual paper ... I can find no support for my assertion. ... My apologies for seeming to have started a myth" [60].

We also take this time to correct another historical error, which actually is composed of two suberrors. The first suberror is that many sources incorrectly cited [53] when referring to Proth's discussion of Gilbreath's conjecture, referring to the correct title "Théorèmes sur les nombres premiers" but citing Comp. Rend. Acad. Sci. Paris, 85 (1877) instead of Comp. Rend. Acad. Sci. Paris, 87 (1877). The former actually corresponds to a completely unrelated paper of Pepin [50]. The second suberror is that, the intended reference, [53], didn't even discuss Gilbreath's conjecture! We were only able to find Proth discussing Gilbreath's conjecture in [54].

We refer the reader to [4] for more information concerning this situation.

Chapter 4

A new upper bound for separating words

4.1 Summary

We prove that for any distinct $x, y \in \{0,1\}^n$, there is a deterministic finite automaton with $\widetilde{O}(n^{1/3})$ states that accepts x but not y. This improves Robson's 1989 upper bound of $\widetilde{O}(n^{2/5})$.

4.2 Introduction

Given a positive integer n and two distinct 0-1 strings $x, y \in \{0,1\}^n$, let $f_n(x,y)$ denote the smallest positive integer m such that there exists a deterministic finite automaton with m states that accepts x but not y (of course, $f_n(x,y) = f_n(y,x)$). Define $f(n) := \max_{x \neq y \in \{0,1\}^n} f_n(x,y)$. The "separating words problem" is to determine the asymptotic behavior of $f(n)$. An easy example [28] shows $f(n) = \Omega(\log n)$, which is the best lower bound known to date. Goralcik and Koubek [28] in 1986 proved an upper bound of $f(n) = o(n)$, and Robson [55] in 1989 proved an upper bound of $f(n) = O(n^{2/5} \log^{3/5} n)$. Despite much attempt, there has been no further improvement to the upper bound to date.

In this chapter, we improve the upper bound on the separating words problem to $f(n) = \widetilde{O}(n^{1/3})$.

Theorem 4.2.1. *For any distinct $x, y \in \{0,1\}^n$, there is a deterministic finite automaton with $O(n^{1/3} \log^7 n)$ states that accepts x but not y.*

We made no effort to optimize the (power of the) logarithmic term $\log^7 n$.

4.3 Definitions and Notation

A *deterministic finite automaton* (DFA) M is a 4-tuple (Q, δ, q_1, F) consisting of a finite set Q, a function $\delta : Q \times \{0,1\} \to Q$, an element $q_1 \in Q$, and a subset $F \subseteq Q$. We call elements $q \in Q$ "states". We call q_1 the "initial state" and the elements of F the "accept states". We say M *accepts* a string $x = x_1, \ldots, x_n \in \{0,1\}^n$ if (and only if) the sequence defined by $r_1 = q_1, r_{i+1} = \delta(r_i, x_i)$ for $1 \leq i \leq n$, has $r_{n+1} \in F$.

We say a set $A \subseteq [n]$ is *d-separated* if $a, a' \in A, a \neq a'$ implies $|a - a'| \geq d$. For a set $A \subseteq [n]$, a prime p, and a residue $i \in [p] := \{1, \ldots, p\}$, let

$$A_{i,p} = \{a \in A : a \equiv i \pmod{p}\}.$$

For a string $x = x_1, \ldots, x_n \in \{0,1\}^n$ and a (sub)string $w = w_1, \ldots, w_l \in \{0,1\}^l$, let $\mathrm{pos}_w(x) := \{j \in \{1, \ldots, n-l+1\} : x_{j+k-1} = w_k \text{ for all } 1 \leq k \leq l\}$ denote the set of all (starting) positions at which w occurs as a (contiguous) substring in x.

For a positive integer n, we write $[n]$ for $\{1, \ldots, n\}$. We write $\sim$ as shorthand for $= (1 + o(1))$. In our inequalities, C and c refer to (large and small, respectively) absolute constants that sometimes change from line to line. For functions f and g, we say $f = \widetilde{O}(g)$ if $|f| \leq C|g| \log^C |g|$ for some constant C. We write $A \asymp B$ if $\frac{1}{2}B \leq A \leq B$.

4.4 An easy $\widetilde{O}(n^{1/2})$ bound, and motivation of our argument

In this section, we sketch an argument of an $\widetilde{O}(n^{1/2})$ upper bound for the separating words problem, and then how to generalize that argument to obtain $\widetilde{O}(n^{1/3})$. This argument also appears in [56] and in [59].

For any two distinct strings $x, y \in \{0,1\}^n$, the sets $\mathrm{pos}_1(x)$ and $\mathrm{pos}_1(y)$ are of course different. A natural way, therefore, to try to separate different strings x, y is to find a small prime p and a residue $i \in [p]$ so that $|\mathrm{pos}_1(x)_{i,p}| \neq |\mathrm{pos}_1(y)_{i,p}|$; if we can find such a p and i, then since[1] there will be a prime q of size $q = O(\log n)$ with $|\mathrm{pos}_1(x)_{i,p}| \not\equiv |\mathrm{pos}_1(y)_{i,p}| \pmod{q}$, there will be a deterministic finite automaton with $2pq = O(p \log n)$ states that accepts one string but not the other (see Lemma 4.5.2). We are thus led to the following (purely number-theoretic) problem.

[1] We make use of the fact that $q \mid a - b$ for all primes q in a set $\mathcal{Q}$ implies $\prod_{q \in \mathcal{Q}} q \mid a - b$, along with standard estimates on $\prod_{q \in \mathcal{Q}} q$ for $\mathcal{Q} = \{q \leq k : q \text{ prime}\}$.

Problem 4.4.1. For given n, determine the minimum k such that for any distinct $A, B \subseteq [n]$, there is some prime $p \leq k$ and some $i \in [p]$ for which $|A_{i,p}| \neq |B_{i,p}|$.

Problem 4.4.1 has been considered[2] in [56], [57], and [59] (and possibly other places) and was essentially solved in each. We present a simple solution, also discovered in [59].

Claim 4.4.2. *For any distinct $A, B \subseteq [n]$, there is some prime $p = O(\sqrt{n \log n})$ and some $i \in [p]$ for which $|A_{i,p}| \neq |B_{i,p}|$.*

Proof. (Sketch) Fix distinct $A, B \subseteq [n]$. Suppose k is such that $|A_{i,p}| = |B_{i,p}|$ for all primes $p \leq k$ and all $i \in [p]$. For a prime p, let $\Phi_p(x)$ denote the p^{th} cyclotomic polynomial, of degree $p-1$. Then since $\sum_{j=1}^n 1_A(j) e^{2\pi i \frac{aj}{p}} = \sum_{j=1}^n 1_B(j) e^{2\pi i \frac{aj}{p}}$ for all $p \leq k$ and all $a \in [p]$, the polynomials $\Phi_p(x)$, for $p \leq k$, divide $\sum_{j=1}^n (1_A(j) - 1_B(j)) x^j =: f(x)$. Therefore, $\prod_{p \leq k} \Phi_p(x)$ divides $f(x)$. Since $A \neq B$, f is not identically 0 and thus must have degree at least $\sum_{p \leq k}(p-1) \sim \frac{1}{2}\frac{k^2}{\log k}$. Since the degree of f is trivially at most n, we must have $(1+o(1))\frac{1}{2}\frac{k^2}{\log k} \leq n$. □

By a standard pigeonhole argument (see Section 7), the bound $\widetilde{O}(\sqrt{n})$ is sharp.

A natural idea to improve this $\widetilde{O}(\sqrt{n})$ bound for the separating words problem is to consider the sets $\text{pos}_w(x)$ and $\text{pos}_w(y)$ for longer w. The length of w is actually not important in terms of its "cost" to the number of states needed, just as long as it is at most p, where we will be considering $|\text{pos}_w(x)_{i,p}|$ and $|\text{pos}_w(y)_{i,p}|$ (see Lemma 4.5.2). One immediate benefit of considering longer w is that the sets $\text{pos}_w(x)$ and $\text{pos}_w(y)$ are *smaller* than $\text{pos}_1(x)$ and $\text{pos}_1(y)$; indeed, for example, it can be shown without much difficulty that for any distinct $x, y \in \{0,1\}^n$, there is some w of length $n^{1/3}$ such that $\text{pos}_w(x)$ and $\text{pos}_w(y)$ are distinct sets of size at most $n^{2/3}$. Thus, to get a bound of $\widetilde{O}(n^{1/3})$ on the separating words problem, it suffices to show the following.

Problem 4.4.3. For any distinct $A, B \subseteq [n]$ of sizes $|A|, |B| \leq n^{2/3}$, show there is some prime $p = \widetilde{O}(n^{1/3})$ and some $i \in [p]$ so that $|A_{i,p}| \neq |B_{i,p}|$.

As in the proof sketch above, this problem is equivalent to a statement about a product of cyclotomic polynomials dividing a sparse polynomial of small degree (see the last page of [59]). We were not able to solve Problem 4.4.3. However, we make the additional observation that we can take w so that $\text{pos}_w(x)$ and $\text{pos}_w(y)$ are *well-separated* sets. Indeed, if w has length $2n^{1/3}$ and has no period of length at most $n^{1/3}$, then $\text{pos}_w(x)$ and $\text{pos}_w(y)$ are $n^{1/3}$-separated sets. As we'll use later, Lemmas 1

[2] In the last reference, they look for an *integer* $m \leq k$ and some $i \in [m]_0$ for which $|A_{i,m}| \neq |B_{i,m}|$, which is of course more economical. We decided to restrict to primes for aesthetic reasons.

and 2 of [55] show that such w are common enough to ensure there is a choice with $\mathrm{pos}_w(x) \neq \mathrm{pos}_w(y)$. Our main technical theorem is thus the following[3].

Theorem 4.4.4. *Let A, B be distinct subsets of $[n]$ that are each $n^{1/3}$-separated. Then there is some prime $p = \widetilde{O}(n^{1/3})$ and some $i \in [p]$ so that $|A_{i,p}| \neq |B_{i,p}|$.*

Although Theorem 4.6.2 is also equivalent to a question about a product of cyclotomic polynomials dividing a certain type of polynomial, we were not able to make progress through number theoretic arguments. Rather, we reverse the argument of Scott [57], by noting that if there is some small m so that the m^{th}-moments of A and B differ, i.e. $\sum_{a\in A} a^m \neq \sum_{b\in B} b^m$, then there is some small p and some $i \in [p]$ so that $|A_{i,p}| \not\equiv |B_{i,p}| \bmod p$ (and thus $|A_{i,p}| \neq |B_{i,p}|$).[4]

The benefit of considering the "moments" problem is that it is more susceptible to complex analytic techniques. Borwein, Erdélyi, and Kós [9] use complex analytic techniques to show that for any distinct $A, B \subseteq [n]$, there is some $m \leq C\sqrt{n}$ with $\sum_{a\in A} a^m \neq \sum_{b\in B} b^m$. One proof of theirs was to show that any polynomial p of degree n with $|p(0)| = 1$ and coefficients bounded by 1 in absolute value must be at least $\exp(-C\sqrt{n})$ at some point close to 1. We were able to adapt this proof to find a small(er) m such that $\sum_{a\in A} a^m \neq \sum_{b\in B} b^m$ in the case that A, B are well-separated sets, and thus prove Theorem 4.6.2.

The adaptations we make are quite significant. See Lemma 5.6.2 and Lemma 4.7.5.

4.5 Deduction of main theorem from number theoretic statement

In this section, we quickly deduce Theorem 4.5.3 from our main number-theoretic theorem which we prove in Section 5. Recall we say $A \subseteq [n]$ is *d-separated* if $|a - a'| \geq d$ for any distinct $a, a' \in A$.

Theorem 4.5.1. *Let A, B be distinct subsets of $[n]$ that are each $n^{1/3}$-separated. Then there is some prime $p \asymp C'n^{1/3}\log^6 n$ and some $i \in [p]$ so that $|A_{i,p}| \neq |B_{i,p}|$. Here, $C' > 0$ is an absolute constant.*

[3] See page 4 for a more specific formulation.

[4] The implication just written is actually quite straightforward (see the deduction of Theorem 4.6.2 from Proposition 4.6.4); the implication of Scott, however, that some small p and some $i \in [p]$ with $|A_{i,p}| \not\equiv |B_{i,p}| \pmod p$ implies the existence of some small m with $\sum_{a\in A} a^m \neq \sum_{b\in B} b^m$ is less trivial, though basically just follows from the fact that $1_{x\equiv i \pmod p} \equiv 1 - (x-i)^{p-1} \pmod p$.

Recall that, for a string $x = x_1, \ldots, x_n \in \{0,1\}^n$ and a (sub)string $w = w_1, \ldots, w_l \in \{0,1\}^l$, we defined $\mathrm{pos}_w(x) := \{j \in \{1, \ldots, n-l+1\} : x_{j+k-1} = w_k \text{ for all } 1 \le k \le l\}$.

Lemma 4.5.2. *Let m, n be positive integers, $i \in [m]_0$ a residue mod m, q a prime number, $a \in [q]_0$ a residue mod q, and $w \in \{0,1\}^l$ a string of length $l \le m$. Then there is a determinsitic finite automaton with $2mq$ states that, for any string $x \in \{0,1\}^n$, accepts x if and only if $|\{j \in \mathrm{pos}_w(x) : j \equiv i \pmod{m}\}| \equiv a \pmod{q}$.*

Proof. Write $w = w_1, \ldots, w_l$. We assume $l > 1$; a minor modification to the following yields the result for $l = 1$. We interpret indices of w mod m, which we may, since $l \le m$. Let the states of the DFA be $\mathbb{Z}_m \times \{0,1\} \times \mathbb{Z}_q$. The initial state is $(1,0,0)$. If $j \not\equiv i \pmod{m}$ and $\epsilon \in \{0,1\}$, set $\delta((j,0,s),\epsilon) = (j+1,0,s)$. If $j \equiv i \pmod{m}$, set $\delta((j,0,s),w_1) = (j+1,1,s)$ and $\delta((j,0,s),1-w_1) = (j+1,0,s)$. If $j \not\equiv i+l-1 \pmod{m}$, set $\delta((j,1,s),w_{j-i+1}) = (j+1,1,s)$ and $\delta((j,1,s),1-w_{j-i+1}) = (j+1,0,s)$. Finally, if $j \equiv i+l-1 \pmod{m}$, set $\delta((j,1,s),w_l) = (j+1,0,s+1)$ and $\delta((j,1,s),1-w_l) = (j+1,0,s)$. The set of accept states is $\mathbb{Z}_m \times \{0,1\} \times \{a\}$. □

Theorem 4.5.3. *For any distinct $x, y \in \{0,1\}^n$, there is a deterministic finite automaton with $O(n^{1/3}\log^7 n)$ states that accepts x but not y.*

Proof. Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be two distinct strings in $\{0,1\}^n$. If $x_k \ne y_k$ for some $k < 2n^{1/3}$, then we are done[5], so we may suppose otherwise. Let $k \ge 2n^{1/3}$ be the first index with $x_k \ne y_k$. Let $w' = x_{k-2n^{1/3}+1}, \ldots, x_{k-1}$ be a (common sub)string of x and y of length $2n^{1/3} - 1$. By Lemma 1 and Lemma 2 of [55], there is some choice $w \in \{w'0, w'1\}$ for which $A := \mathrm{pos}_w(x)$ is $n^{1/3}$-separated and $B := \mathrm{pos}_w(y)$ is $n^{1/3}$-separated. By the choice of k, we have $A \ne B$, so Theorem 4.6.2 implies there is some prime $p \in [\frac{1}{2}C'n^{1/3}\log^6 n, C'n^{1/3}\log^6 n]$ and some $i \in [p]$ for which $|A_{i,p}| \ne |B_{i,p}|$. Since $|A_{i,p}|$ and $|B_{i,p}|$ are at most n, there is some prime $q = O(\log n)$ for which $|A_{i,p}| \not\equiv |B_{i,p}| \pmod{q}$. Since $|w| = 2n^{1/3} \le p$, by Lemma 4.5.2 there is a deterministic finite automaton with $2pq = O(n^{1/3}\log^7 n)$ states that accepts x but not y. □

[5] Simply use a DFA on $2n^{1/3}$ states that accepts exactly those strings starting with $x_1, \ldots, x_{2n^{1/3}}$.

4.6 Deduction of number theoretic statement from complex analytic statement

In this section, we deduce Theorem 4.6.2 from the following complex analytic theorem, which we prove in Section 6.

Let $\mathcal{P}_n$ denote the collection of all polynomials $p(x) = 1-\sigma x^d+\sum_{j=n^{1/3}}^{n} a_j x^j \in \mathbb{C}[x]$ such that $1 \le d < n^{1/3}$, $\sigma \in \{0,1\}$, and $|a_j| \le 1$ for each j.

Theorem 4.6.1. *There is some absolute constant $C_1 > 0$ so that for all $n \ge 2$ and all $p \in \mathcal{P}_n$, it holds that*

$$\max_{x\in[1-n^{-2/3},1]} |p(x)| \ge \exp(-C_1 n^{1/3}\log^5 n).$$

The deduction of Theorem 4.6.2 from Theorem 4.7.1 follows from first showing the polynomial $p(x) := \sum_{n\in A} x^n - \sum_{n\in B} x^n$ cannot be divisible by a large power of $x-1$. We will use part of Lemma 5.4 of [9], stated below.

Lemma 4.6.2. *Suppose the polynomial $f(x) = \sum_{j=0}^{n} a_j x^j \in \mathbb{C}[x]$ has $|a_j| \le 1$ for each j. If $(x-1)^k$ divides $f(x)$, then $\max_{1-\frac{k}{9n}\le x\le 1} |f(x)| \le (n+1)(\frac{e}{9})^k$.*

Proposition 4.6.3. *There exists an absolute constant $C > 0$ so that for all $n \ge 1$ and all $p(x) \in \mathcal{P}_n$, the polynomial $(x-1)^{\lfloor Cn^{1/3}\log^5 n\rfloor}$ does not divide $p(x)$.*

Proof. Take $C > 0$ large. Take $p(x) \in \mathcal{P}_n$. Suppose for the sake of contradiction that $(x-1)^{Cn^{1/3}\log^5 n}$ divided $p(x)$. Then, by Lemma 4.6.2 and Theorem 4.7.1,

$$\begin{aligned}(n+1)(\frac{e}{9})^{Cn^{1/3}\log^5 n} &\ge \max_{x\in[1-\frac{C}{9}n^{-2/3}\log^5 n,1]} |p(x)| \\ &\ge \max_{x\in[1-n^{-2/3},1]} |p(x)| \\ &\ge e^{-C_1 n^{1/3}\log^5 n},\end{aligned}$$

which is a contradiction if C is large enough. □

We now exploit the (well-known) equivalence between common moments and a large vanishing of the associated polynomial at $x = 1$.

Proposition 4.6.4. *Let A, B be distinct subsets of $[n]$ that are each $n^{1/3}$-separated. Then there is some non-negative integer $m = O(n^{1/3}\log^5 n)$ such that $\sum_{a\in A} a^m \ne \sum_{b\in B} b^m$.*

Proof. Let $f(x) = \sum_{j=0}^{n} \epsilon_j x^j$, where $\epsilon_j := 1_A(j) - 1_B(j)$. Let $\tilde{f}(x) = \frac{f(x)}{x^r}$, where r is maximal with respect to $\epsilon_0, \ldots, \epsilon_{r-1} = 0$. We may assume without loss of generality that $\tilde{f}(0) = 1$. Then the fact that A, B are $n^{1/3}$-separated implies $\tilde{f}(x) \in \mathcal{P}_n$. By Proposition 4.6.3, $(x-1)^{Cn^{1/3}\log^5 n}$ does not divide $\tilde{f}(x)$ and thus does not divide $f(x)$. This means that there is some non-negative integer $k \leq Cn^{1/3}\log^5 n - 1$ so that $f^{(k)}(1) \neq 0$. Take a minimal such k. If $k = 0$, we're of course done. Otherwise, since $f^{(m)}(1) = \sum_{j=0}^{n} j(j-1)\ldots(j-m+1)\epsilon_j$ for $m \geq 1$, it's easy to inductively see that $\sum_{j\in A} j^m = \sum_{j \in B} j^m$ for all $0 \leq m \leq k-1$ and then $\sum_{j \in A} j^k \neq \sum_{j\in B} j^k$. □

We can now deduce Theorem 4.6.2.

Theorem 4.6.2. *Let A, B be distinct subsets of $[n]$ that are each $n^{1/3}$-separated. Then there is some prime $p \asymp C'n^{1/3}\log^6 n$ and some $i \in [p]$ so that $|A_{i,p}| \neq |B_{i,p}|$. Here, $C' > 0$ is an absolute constant.*

Proof. By Proposition 4.6.4, take $m = O(n^{1/3}\log^5 n)$ such that $\sum_{a\in A} a^m \neq \sum_{b \in B} b^m$. Since $\left|\sum_{a\in A} a^m - \sum_{b\in B} b^m\right| \leq n n^m \leq \exp(O(n^{1/3}\log^6 n))$, there is some prime $p \in [\frac{1}{2}C'n^{1/3}\log^6 n, C'n^{1/3}\log^6 n]$ such that $\sum_{a\in A} a^m \not\equiv \sum_{b\in B} b^m \pmod{p}$. Noting that $\sum_{a \in A} a^m \equiv \sum_{i=0}^{p-1} |A_{i,p}| i^m \pmod{p}$ and $\sum_{b\in B} b^m \equiv \sum_{i=0}^{p-1} |B_{i,p}| i^m \pmod{p}$, we see that there is some $i \in [p]$ for which $|A_{i,p}| \not\equiv |B_{i,p}| \pmod{p}$. □

4.7 Proof of complex analytic statement

In this section, we finish off the proof of Theorem 4.5.3 by proving the needed theorem about sparse Littlewood polynomials being "large" somewhere near 1.

Recall that $\mathcal{P}_n$ denotes the collection of all polynomials $p(x) = 1 - \sigma x^d + \sum_{j=n^{1/3}}^{n} a_j x^j$ in $\mathbb{C}[x]$ such that $1 \leq d < n^{1/3}$, $\sigma \in \{0, 1\}$, and $|a_j| \leq 1$ for each j.

Theorem 4.7.1. *There is some absolute constant $C_1 > 0$ so that for all $n \geq 2$ and all $p \in \mathcal{P}_n$, it holds that*

$$\max_{x \in [1-n^{-2/3}, 1]} |p(x)| \geq \exp(-C_1 n^{1/3}\log^5 n).$$

For $a > 0$, define $\widetilde{E}_a$ to be the ellipse with foci at $1-a$ and $1 - a + \frac{1}{4}a$ and with major axis $[1 - a - \frac{a}{32}, 1 - a + \frac{9a}{32}]$. We borrow[6] Corollary 5.3 from [9]:

[6] They state Lemma 4.7.2 for $p \in \mathcal{S}$, where they define $\mathcal{S}$ to be the set of all analytic functions f on the (open) unit disk such that $|f(z)| \leq \frac{1}{1-|z|}$ for each $z \in \mathbb{D}$. It is clear $\mathcal{P}_n \subseteq \mathcal{S}$ for each n.

Lemma 4.7.2. *For every $n \geq 1$, $p \in \mathcal{P}_n$, and $a > 0$, we have*

$$\left(\max_{z \in \widetilde{E}_a} |p(z)|\right)^2 \leq \frac{64}{39a} \max_{x \in [1-a,1]} |p(x)|.$$

By Lemma 4.7.2, in order to prove Theorem 4.7.1 it suffices to show:

Proposition 4.7.3. *There is an absolute constant $C > 0$ so that for every $n \geq 1$ and every $p \in \mathcal{P}_n$, it holds that $\left(\max_{z \in \widetilde{E}_{n^{-2/3}}} |p(z)|\right)^2 \geq \exp(-Cn^{1/3} \log^5 n)$.*

While [9] certainly uses that $\widetilde{E}_a$ is an ellipse, all we will use is about $\widetilde{E}_a$ (besides using Lemma 4.7.2 as a black box) is that the interior of $\widetilde{E}_a$, denoted $\widetilde{E}_a^\circ$, contains a ball of radius $\frac{a}{10^{10}}$ centered at $1-a$. We begin with two lemmas.

In the proof of Theorem 5.1 of [9], the authors use the function $h(z) = (1-a)\frac{z+z^2}{2}$ for a maximum modulus principle argument to lower bound the quantity $\left(\max_{z \in \widetilde{E}_a} |p(z)|\right)^2$. For $z = e^{2\pi i t}$ for small t, the magnitude $|h(e^{2\pi i t})|$ is quadratically in t less than 1. For our purposes, we need a linear deviation of $|h(e^{2\pi i t})|$ from 1. This motivates the following lemma.

Lemma 4.7.4. *There are absolute constants $c_4, c_5, C_6 > 0$ such that the following holds for $a > 0$ small enough. Let $\tilde{h}(z) = \sum_{j=1}^r d_j z^j$ for*

$$d_j := \frac{\lambda_a}{j^2 \log^2(j+3)}$$

and $r := a^{-1}$, where $\lambda_a \in (1,2)$ is such that $\sum_{j=1}^r d_j = 1$. Let $h(z) = (1-a)\tilde{h}(z)$. Then $h(0) = 0$, $|h(e^{2\pi i t})| \leq 1-a$ for each t, $h(e^{2\pi i t}) \in \widetilde{E}_a^\circ$ for $t \in [-c_4 a, c_4 a]$, and

$$|h(e^{2\pi i t})| \leq 1 - c_5 \frac{|t|}{\log^2(a^{-1})}$$

for $t \in [-\frac{1}{2}, \frac{1}{2}] \setminus [-C_6 a, C_6 a]$.

Proof. Clearly $h(0) = 0$ and $|h(e^{2\pi i t})| \leq 1-a$ for each t. Now, for any $t \in \mathbb{R}$,

$$|\tilde{h}(e^{2\pi i t}) - 1| = \left|\sum_{j=1}^r d_j (e^{2\pi i t j} - 1)\right| \leq \sum_{j=1}^r d_j 2\pi t j = 2\pi t \sum_{j=1}^r \frac{\lambda_a}{j \log^2(j+3)} \leq C_4 t$$

for C_4 absolute. Thus,

$$|h(e^{2\pi i t}) - (1-a)| = (1-a)|\tilde{h}(e^{2\pi i t}) - 1| \leq C_4 t.$$

If $|t| \leq c_4 a$ for $c_4 > 0$ sufficiently small, we conclude $h(e^{2\pi i t}) \in \widetilde{E}_a^{\circ}$.

We now go on to showing the last inequality in the statement of Lemma 5.6.2.

By summation by parts, for any $z \in \mathbb{C}$, we have

$$\sum_{j=1}^{r} \frac{\lambda_a z^j}{j^2 \log^2(j+3)} = \frac{\lambda_a \sum_{j=1}^{r} z^j}{r^2 \log^2(r+3)} + 2\lambda_a \int_1^r (\sum_{j \leq x} z^j) g(x) dx. \tag{4.1}$$

There, and what follows, we denote(d)

$$g(x) := \frac{\log(x+3) + \frac{x}{x+3}}{x^3 \log^3(x+3)}.$$

Quickly note that, for $z = 1$, (4.1) gives

$$1 = \frac{\lambda_a}{r \log^2(r+3)} + 2\lambda_a \int_1^r \lfloor x \rfloor g(x) dx. \tag{4.2}$$

Trivially, for any $z \in \partial\mathbb{D}$, we have

$$\left| \frac{\lambda_a \sum_{j=1}^{r} z^j}{r^2 \log^2(r+3)} \right| \leq \frac{\lambda_a}{r \log^2(r+3)}. \tag{4.3}$$

Note that, for any $x \geq 1$,

$$\left| \sum_{j \leq x} z^j \right| = \left| z \frac{1 - z^{\lfloor x \rfloor}}{1 - z} \right| \leq \frac{2}{|1 - z|} \leq t^{-1} \tag{4.4}$$

for all $z = e^{2\pi i t}$ with $t \in (0, \frac{1}{2}]$. Take $C_6 > 3$ to be chosen later. Note $t \in (C_6 a, \frac{1}{2}]$ implies $3t^{-1} < r$. For $z = e^{2\pi i t}$ with $C_6 a < t \leq \frac{1}{2}$, (4.4) and (4.2) imply

$$\left| 2\lambda_a \int_1^r (\sum_{j \leq x} z^j) g(x) dx \right| \leq$$

$$\begin{aligned} & 2\lambda_a \int_1^{3t^{-1}} \lfloor x \rfloor g(x) dx + 2\lambda_a \int_{3t^{-1}}^{r} t^{-1} g(x) dx \\ &= 1 - 2\lambda_a \int_{3t^{-1}}^{r} \left(\lfloor x \rfloor - t^{-1} \right) g(x) dx - \frac{\lambda_a}{r \log^2(r+3)}. \end{aligned} \tag{4.5}$$

Observe $\lfloor x \rfloor - t^{-1} \geq \frac{1}{2} x$ for $x \geq 3t^{-1}$. Therefore,

$$\begin{aligned} 2\lambda_a \int_{3t^{-1}}^{r} \left(\lfloor x \rfloor - t^{-1} \right) g(x) dx &\geq \lambda_a \int_{3t^{-1}}^{r} \frac{1}{x^2 \log^2(x+3)} dx \\ &\geq \frac{\lambda_a}{\log^2(r+3)} \int_{3t^{-1}}^{r} \frac{1}{x^2} dx \end{aligned}$$

$$= \frac{\lambda_a t}{3\log^2(r+3)} - \frac{\lambda_a}{r\log^2(r+3)}. \quad (4.6)$$

Combining (4.1), (4.3), (4.5), and (4.6), we conclude that, for any $t \in (C_6 a, \frac{1}{2}]$,

$$\left|\tilde{h}(e^{2\pi it})\right| = \left|\sum_{j=1}^{r} \frac{\lambda_a e^{2\pi ijt}}{j^2\log^2(j+3)}\right| \leq 1 - \frac{\lambda_a t}{3\log^2(r+3)} + \frac{\lambda_a}{r\log^2(r+3)}. \quad (4.7)$$

Taking C_6 to be much larger than 3, (4.7) gives the bound

$$|\tilde{h}(e^{2\pi it})| \leq 1 - c_5\frac{t}{\log^2(a^{-1})}$$

for $t \in (C_6 a, \frac{1}{2}]$, for suitable $c_5 > 0$. By symmetry, the proof is complete. □

We from now on fix some $n \geq 1$ and some $p \in \mathcal{P}_n$ (defined at the beginning of the section). Let $\tilde{p}$ be the truncation of p to terms of degree less than $n^{1/3}$; either $\tilde{p} = 1$ or $\tilde{p} = 1 - x^d$ for some $1 \leq d < n^{1/3}$. Take $a = n^{-2/3}$, and let h be as in Lemma 5.6.2. Let $m = c_4^{-1}n^{2/3}$. Let $J_1 = c_5^{-1}n^{-1/3}m\log^4 n$ and $J_2 = m - J_1$.

In the proof below of Proposition 4.7.3, we will need to upper bound the product $\prod_{j=J_1}^{J_2-1} |\tilde{p}(h(e^{2\pi i\frac{j}{m}}))|$ by $\exp(\widetilde{O}(n^{1/3}))$. We must be careful in doing so, as the trivial upper bound on each term is 2 and there are approximately $n^{2/3}$ terms. However, we expect the argument of $h(e^{2\pi i\frac{j}{m}})$ to behave as if it were random, and thus we expect $|\tilde{p}(h(e^{2\pi i\frac{j}{m}}))|$ to sometimes be smaller than 1. The fact that the cancellation between terms smaller than 1 and terms greater than 1 is nearly perfect comes from the fact that $\log|\tilde{p}(h(w))|$ is harmonic, which we make crucial use of below.

Lemma 4.7.5. *For any $t \in [0,1]$, we have $|\tilde{p}(h(e^{2\pi it}))| \geq \frac{1}{2}n^{-2/3}$. For any $\delta \in [0,1)$, we have $\prod_{j=J_1}^{J_2-1} |\tilde{p}(h(e^{2\pi i\frac{j+\delta}{m}}))| \leq \exp(Cn^{1/3}\log^5 n)$ for some absolute $C > 0$.*

Proof. Clearly both inequalities hold if $\tilde{p} = 1$, so suppose $\tilde{p}(x) = 1 - x^d$ for some $1 \leq d < n^{1/3}$. For the first inequality, we use

$$|\tilde{p}(h(e^{2\pi it}))| = |1 - h(e^{2\pi it})^d| \geq 1 - |h(e^{2\pi it})|^d \geq 1 - (1-a)^d \geq \frac{1}{2}ad \geq \frac{1}{2}n^{-2/3}.$$

We now move on to the second inequality. Define $g(t) = 2\log|\tilde{p}(h(e^{2\pi i(t+\frac{\delta}{m})}))|$. For notational ease, we assume $\delta = 0$; the argument about to come works for all $\delta \in [0,1)$.

The first inequality implies g is C^1, so by the mean value theorem,

$$\begin{aligned}\left|\frac{1}{m}\sum_{j=J_1}^{J_2-1} g\left(\frac{j}{m}\right) - \int_{J_1/m}^{J_2/m} g(t)dt\right| &= \left|\sum_{j=J_1}^{J_2-1}\int_{j/m}^{(j+1)/m}\left(g(t)-g\left(\frac{j}{m}\right)\right)dt\right| \\ &\leq \sum_{j=J_1}^{J_2-1}\int_{j/m}^{(j+1)/m}\left(\max_{\frac{j}{m}\leq y\leq\frac{j+1}{m}}|g'(y)|\right)\frac{1}{m}dt \\ &\leq \frac{1}{m^2}\sum_{j=J_1}^{J_2-1}\max_{\frac{j}{m}\leq y\leq\frac{j+1}{m}}|g'(y)|. \qquad (4.8)\end{aligned}$$

Since $w \mapsto \log|\tilde{p}(h(w))|$ is harmonic and $\log|\tilde{p}(h(0))| = \log|\tilde{p}(0)| = 0$, we have

$$\int_0^1 g(t)dt = 2\int_0^1 \log|\tilde{p}(h(e^{2\pi it}))|dt = 0,$$

and therefore

$$\left|\int_{J_1/m}^{J_2/m} g(t)dt\right| \leq \left|\int_0^{J_1/m} g(t)dt\right| + \left|\int_{J_2/m}^1 g(t)dt\right|. \qquad (4.9)$$

Since

$$\frac{1}{2}n^{-2/3} \leq \left|\tilde{p}(h(e^{2\pi it}))\right| \leq 1$$

for each t, we have

$$\left|\int_0^{J_1/m} g(t)dt\right| + \left|\int_{J_2/m}^1 g(t)dt\right| \leq 2\left(\frac{J_1}{m} + (1-\frac{J_2}{m})\right)\log n \leq C\frac{\log^5 n}{n^{1/3}}. \qquad (4.10)$$

By (5.4), (5.5), and (5.6), we have

$$\left|\frac{1}{m}\sum_{j=J_1}^{J_2-1} g(\frac{j}{m})\right| \leq C\frac{\log^5 n}{n^{1/3}} + \frac{1}{m^2}\sum_{j=J_1}^{J_2-1}\max_{\frac{j}{m}\leq t\leq\frac{j+1}{m}}|g'(t)|.$$

Multiplying through by m, changing C slightly, and exponentiating, we obtain

$$\prod_{j=J_1}^{J_2-1}\left|\tilde{p}(h(e^{2\pi i\frac{j}{m}}))\right|^2 \leq \exp\left(Cn^{1/3}\log^5 n + \frac{1}{m}\sum_{j=J_1}^{J_2-1}\max_{\frac{j}{m}\leq t\leq\frac{j+1}{m}}|g'(t)|\right). \qquad (4.11)$$

Note

$$g'(t_0) = \frac{\frac{\partial}{\partial t}\left[|\tilde{p}(h(e^{2\pi it}))|^2\right]_{t=t_0}}{|\tilde{p}(h(e^{2\pi it_0}))|^2}.$$

We first show

$$\frac{\partial}{\partial t}\left[|\tilde{p}(h(e^{2\pi it}))|^2\right]_{t=t_0} \leq 100d$$

for each $t_0 \in [0,1]$. We start by noting

$$\left|\tilde{p}(h(e^{2\pi it}))\right|^2 = 1 + (1-a)^{2d}\left|\sum_{j=1}^{r} d_j e^{2\pi itj}\right|^{2d} - 2\operatorname{Re}\left((1-a)\sum_{j=1}^{r} d_j e^{2\pi itj}\right)^d.$$

Let

$$f_1(t) = (1-a)^{2d}\left|\sum_{j=1}^{r} d_j e^{2\pi itj}\right|^{2d}.$$

Then,

$$\begin{aligned} f_1'(t) &= (1-a)^{2d} d\left|\sum_{j=1}^{r} d_j e^{2\pi itj}\right|^{2(d-1)} \frac{\partial}{\partial t}\left|\sum_{j=1}^{r} d_j e^{2\pi itj}\right|^{2} \\ &= (1-a)^{2d} d\left|\sum_{j=1}^{r} d_j e^{2\pi itj}\right|^{2(d-1)} \sum_{1\le j_1,j_2\le r} d_{j_1} d_{j_2} 2\pi i(j_1-j_2)e^{2\pi i(j_1-j_2)t}. \end{aligned}$$

Since $\sum_{j=1}^{r} d_j = 1$, we therefore have

$$\begin{aligned} |f_1'(t)| &\le 2\pi d \sum_{1\le j_1,j_2\le r} \lambda_a^2 \frac{j_1+j_2}{j_1^2 j_2^2 \log^2(j_1+3)\log^2(j_2+3)} \\ &= 4\pi d\left(\sum_{j_1=1}^{r} \frac{\lambda_a}{j_1\log^2(j_1+3)}\right)\left(\sum_{j_2=1}^{r} \frac{\lambda_a}{j_2^2\log^2(j_2+3)}\right) \\ &\le 50d. \end{aligned}$$

Now, let

$$f_2(t) = -2\operatorname{Re}\left((1-a)\sum_{j=1}^{r} d_j e^{2\pi itj}\right)^d$$

and note

$$\begin{aligned} f_2'(t) &= \frac{\partial}{\partial t}\left[-2(1-a)^d \sum_{1\le j_1,\dots,j_d\le r} d_{j_1}\dots d_{j_d}\cos(2\pi t(j_1+\dots+j_d))\right] \\ &= 4\pi(1-a)^d \sum_{1\le j_1,\dots,j_d\le r} d_{j_1}\dots d_{j_d}(j_1+\dots+j_d)\sin(2\pi t(j_1+\dots+j_d)), \end{aligned}$$

yielding

$$\begin{aligned} |f_2'(t)| &\le 4\pi \sum_{1\le j_1,\dots,j_d\le r} \lambda_a^d \frac{j_1+\dots+j_d}{j_1^2\dots j_d^2\log^2(j_1+3)\dots\log^2(j_d+3)} \\ &= 4\pi d\left(\sum_{j_1=1}^{r} \frac{\lambda_a}{j_1\log^2(j_1+3)}\right)\left(\sum_{j=1}^{r} \frac{\lambda_a}{j^2\log^2(j+3)}\right)^{d-1} \\ &\le 50d. \end{aligned}$$

We have thus shown

$$\frac{\partial}{\partial t}\Big[|\tilde{p}(h(e^{2\pi it}))|^2\Big]_{t=t_0} \leq 100d$$

for each $t_0 \in [0,1]$.

Recall

$$|\tilde{p}(h(e^{2\pi it}))| = |1 - h(e^{2\pi it})^d| \geq 1 - |h(e^{2\pi it})|^d.$$

For $j \in [J_1, J_2] \subseteq [C_6am, (1-C_6a)m]$, we use

$$|h(e^{2\pi i\frac{j}{m}})| \leq 1 - c_5\frac{\min(\frac{j}{m}, 1-\frac{j}{m})}{\log^2 n}$$

to obtain

$$\frac{1}{m}\sum_{j=J_1}^{J_2-1}\max_{\frac{j}{m}\leq t\leq\frac{j+1}{m}}|g'(t)| \leq \frac{1}{m}\sum_{j=J_1}^{J_2-1}100d\left(1-\Big(1-c_5\min(\frac{j}{m},1-\frac{j}{m})\log^{-2}n\Big)^d\right)^{-2}.$$

Up to a factor of 2, we may deal only with $j \in [J_1, \frac{m}{2}]$. Let $J_* = c_5^{-1}d^{-1}m\log^2 n$. Note that $j \leq J_*$ implies $c_5\frac{j}{m\log^2 n} \leq d^{-1}$ and $j \geq J_*$ implies $c_5\frac{j}{m\log^2 n} \geq d^{-1}$. Thus, using $(1-x)^d \leq 1 - \frac{1}{2}xd$ for $x \leq \frac{1}{d}$, we have

$$\begin{aligned}\frac{1}{m}\sum_{j=J_1}^{\min(J_*,\frac{m}{2})}\frac{100d}{\Big(1-(1-c_5\frac{j}{m\log^2 n})^d\Big)^2} &\leq \frac{100d}{m}\sum_{j=J_1}^{\min(J_*,\frac{m}{2})}\frac{1}{\Big(\frac{1}{2}c_5\frac{j}{m\log^2 n}d\Big)^2}\\ &= \frac{400m\log^4 n}{c_5^2 d}\sum_{j=J_1}^{\min(J_*,\frac{m}{2})}\frac{1}{j^2}\\ &\leq \frac{400m\log^4 n}{c_5^2 d}\frac{2}{J_1}\\ &\leq Cn^{1/3}. \qquad (4.12)\end{aligned}$$

Finally, since there is some $c > 0$ such that $(1-x)^l \leq 1-c$ for all $l \in \mathbb{N}$ and $x \in [l^{-1}, 1]$, using the notation $\sum_{i=a}^b x_i = 0$ if $a > b$, we see

$$\begin{aligned}\frac{1}{m}\sum_{j=\min(J_*,\frac{m}{2})+1}^{m/2}\frac{100d}{\Big(1-(1-c_5\frac{j}{m\log^2 n})^d\Big)^2} &\leq \frac{100d}{m}\sum_{j=\min(J_*,\frac{m}{2})+1}^{m/2}c^{-2}\\ &\leq Cd\\ &\leq Cn^{1/3}. \qquad (4.13)\end{aligned}$$

Combining (4.12) and (4.13), we obtain

$$\frac{1}{m}\sum_{j=J_1}^{J_2-1}\max_{\frac{j}{m}\leq\frac{j+1}{m}}|g'(t)| \leq Cn^{1/3}.$$

Plugging this upper bound into (5.7) yields the desired result. □

Proof of Proposition 4.7.3. Define $g(z) = \prod_{j=0}^{m-1} p(h(e^{2\pi i \frac{j}{m}} z))$. Fix $z \in \partial\mathbb{D}$; say $z = e^{2\pi i(\frac{j_0}{m}+\delta)}$ for some $j_0 \in \{0, \ldots, m-1\}$ and $\delta \in [0, \frac{1}{m})$. For ease of notation, we assume $j_0 = 0$; the argument about to come is to any j_0. Then, $e^{2\pi i \frac{j}{m}} z$ is in $\{e^{2\pi i t} : -c_4 a \leq t < c_4 a\}$ if $j \in \{0, m-1\}$. Therefore, together with the maximum modulus principle (p is analytic), we see

$$\begin{aligned} |g(z)| &\leq \left(\max_{w \in \widetilde{E}_a^\circ} |p(w)| \right)^2 \prod_{j \notin \{0,m-1\}} |p(h(e^{2\pi i \frac{j}{m}} z))| \\ &\leq \left(\max_{w \in \widetilde{E}_a} |p(w)| \right)^2 \prod_{j \notin \{0,m-1\}} |p(h(e^{2\pi i \frac{j}{m}} z))|. \end{aligned} \tag{4.14}$$

Let $I = [J_1, J_2 - 1] \cap \mathbb{Z}$. For $j \notin I$, using the bound $|p(w)| \leq \frac{1}{1-|w|}$ for each $w \in \partial\mathbb{D}$, we see

$$|p(h(e^{2\pi i \frac{j}{m}} z))| \leq \frac{1}{1 - |h(e^{2\pi i \frac{j}{m}} z)|} \leq \frac{1}{1-(1-a)} = n^{2/3},$$

thereby obtaining

$$\prod_{j \notin I \cup \{0,m-1\}} |p(h(e^{2\pi i \frac{j}{m}} z))| \leq (n^{2/3})^{(J_1-1)+(m-J_2+1)} \leq (n^{2/3})^{Cn^{1/3}\log^4 n} \leq e^{Cn^{1/3}\log^5 n}. \tag{4.15}$$

Now, for $j \in I$, since

$$|h(e^{2\pi i \frac{j}{m}} z)| \leq 1 - c_5 \frac{\min\left(\frac{j}{m}+\delta, 1-(\frac{j}{m}+\delta)\right)}{\log^2 n} \leq 1 - c' n^{-1/3} \log^2 n,$$

we have

$$\left| p\left(h(e^{2\pi i \frac{j}{m} z})\right) - \tilde{p}\left(h(e^{2\pi i \frac{j}{m} z})\right) \right| \leq n e^{-c' \log^2 n} \leq e^{-c \log^2 n}.$$

Therefore,

$$\prod_{j \in I} |p(h(e^{2\pi i \frac{j}{m}} z))| \leq \prod_{j \in I} \left(|\tilde{p}(h(e^{2\pi i \frac{j}{m}} z))| + e^{-c\log^2 n} \right). \tag{4.16}$$

By both parts of Lemma 4.7.5, we obtain

$$
\begin{aligned}
\prod_{j\in I}\left(|\tilde{p}(h(e^{2\pi i\frac{j}{m}}z))| + e^{-c\log^2 n}\right) &= \sum_{I'\subseteq I} e^{-c(\log^2 n)|I'|}\prod_{j\in I\setminus I'}|\tilde{p}(h(e^{2\pi i\frac{j}{m}}z))| \\
&= \sum_{I'\subseteq I}\left(\prod_{j\in I}|\tilde{p}(h(e^{2\pi i\frac{j}{m}}z))|\right)\left(\prod_{j\in I'}|\tilde{p}(h(e^{2\pi i\frac{j}{m}}z))|\right)^{-1} e^{-c(\log^2 n)|I'|} \\
&\le e^{Cn^{1/3}\log^5 n}\sum_{I'\subseteq I}(2n^{2/3})^{|I'|}e^{-c(\log^2 n)|I'|} \\
&\le e^{Cn^{1/3}\log^5 n}\sum_{I'\subseteq I} e^{-c'(\log^2 n)|I'|} \\
&\le e^{Cn^{1/3}\log^5 n}\sum_{k=0}^{|I|}\binom{|I|}{k}e^{-c'k\log^2 n} \\
&\le 2e^{Cn^{1/3}\log^5 n}. \qquad (4.17)
\end{aligned}
$$

Combining (4.14), (4.15), (4.16), and (4.17), we've shown

$$|g(z)| \le \left(\max_{z\in\widetilde{E}_a}|p(z)|\right)^2 e^{Cn^{1/3}\log^5 n}.$$

As this holds for all $z\in\partial\mathbb{D}$, we have

$$\max_{z\in\partial\mathbb{D}}|g(z)| \le \left(\max_{z\in\widetilde{E}_a}|p(z)|\right)^2 e^{Cn^{1/3}\log^5 n}.$$

To finish, note that $|g(0)| = |p(h(0))|^m = |p(0)|^m = 1$, so, as g is clearly analytic, the maximum modulus principle implies $\max_{z\in\partial\mathbb{D}}|g(z)|\ge 1$. □

4.8 Tightness of our methods

In this section, we prove the following, showing that our methods cannot be pushed further in some sense. We denote $\{0,1\}^{\le p} := \bigcup_{j=1}^{p}\{0,1\}^j$.

Proposition 4.8.1. *For all n large, there are distinct strings $x, y\in\{0,1\}^n$ such that for all $p\le\frac{1}{10}n^{1/3}$, $i\in[p]$, and $w\in\{0,1\}^{\le p}$, it holds that $|\mathrm{pos}_w(x)_{i,p}| = |\mathrm{pos}_w(y)_{i,p}|$.*

We begin by showing Theorem 4.6.2 is tight, via a standard pigeonhole argument that has been used in a variety of other papers.

Proposition 4.8.2. *For all n large, there are distinct $n^{1/3}$-separated subsets A, B of $[n]$ such that $|A_{i,p}| = |B_{i,p}|$ for all $p\le cn^{1/3}\log^{1/2} n$ and all $i\in[p]$.*

Proof. Let Σ denote the collection of subsets $A \subseteq [n]$ that have at most one number from each of the intervals $[1, n^{1/3}], [2n^{1/3}, 3n^{1/3}], [4n^{1/3}, 5n^{1/3}], \dots$. Note $|\Sigma| \geq (n^{1/3})^{\frac{1}{3}n^{2/3}} = e^{\frac{1}{9}n^{2/3}\log n}$. On the other hand, for any $A \subseteq [n]$, the number of possible tuples $(|A_{i,p}|)_{\substack{p \leq k \\ i \in [p]}}$ is at most $\prod_{p \leq k} n^p \leq e^{\frac{k^2}{\log k}\log n}$. Taking $k = cn^{1/3}\log^{1/2} n$ yields $\frac{k^2}{\log k}\log n < \frac{1}{9}n^{2/3}\log n$, meaning there are distinct $A, B \in \Sigma$ with the same tuple, i.e. $|A_{i,p}| = |B_{i,p}|$ for all $p \leq k$ and $i \in [p]$. As A, B are $n^{1/3}$-separated, the proof is complete. □

Proof of Proposition 4.8.1. For a large n, let $A, B \subseteq [n/2]$ be the sets guaranteed by Proposition 4.8.2. Let $x = (1_A(j - \frac{n}{4}))_{j=1}^n, y = (1_B(j - \frac{n}{4}))_{j=1}^n \in \{0,1\}^n$ be the strings with 1s at indices in A and B then padded at the beginning and end by 0s. Fix $p \leq \frac{1}{10}n^{1/3}$ and $i \in [p]$. Since A, B are $\frac{1}{10}n^{1/3}$-separated, we have $|\text{pos}_w(x)_{i,p}| = |\text{pos}_w(y)_{i,p}| = 0$ for all $w \in \{0,1\}^{\leq p}$ with at least two 1s. Since

$$\text{pos}_{0^l}(x) = [n - l + 1] \setminus \sqcup_{s=0}^{l-1}\text{pos}_{0^s 1 0^{l-1-s}}(x),$$

it suffices to show $|\text{pos}_w(x)_{i,p}| = |\text{pos}_w(y)_{i,p}|$ for all $w \in \{0,1\}^{\leq p}$ with exactly one 1. Fix such a w; say $w = 0^s 1 0^{l-1-s}$ for some $l \leq p$ and $s \in \{0, \dots, l-1\}$. Then, due to the padding preventing boundary issues, $\text{pos}_w(x) = \{j : x_{j+s} = 1\} = \{j : 1_A(j + s - \frac{n}{4}) = 1\} = A - s + \frac{n}{4}$ and thus $|\text{pos}_w(x)_{i,p}| = |A_{i+s-\frac{n}{4},p}|$. Similarly, $|\text{pos}_w(y)_{i,p}| = |B_{i+s-\frac{n}{4},p}|$. Since $p \leq c(n/2)^{1/3}\log^{1/2}(n/2)$, the proof is complete. □

Chapter 5

New upper bounds for trace reconstruction

5.1 Summary

We show that any n-bit string can be recovered with high probability from $\exp(\widetilde{O}(n^{1/5}))$ independent random subsequences.

5.2 Introduction

Given a string $x \in \{0,1\}^n$, a *trace* of x is a random string obtained by deleting each bit of x with probability q, independently, and concatenating the remaining string. For example, a trace of 11001 could be 101, obtained by deleting the second and third bits. The goal of the trace reconstruction problem is to determine an unknown string x, with high probability, by looking at as few independently generated traces of x as possible.

More precisely, fix $\delta, q \in (0,1)$. Take n large. For each $x \in \{0,1\}^n$, let μ_x be the probability distribution on $\cup_{j=0}^{n}\{0,1\}^j$ given by $\mu_x(w) = (1-q)^{|w|}q^{n-|w|}f(w;x)$, where $f(w;x)$ is the number of times w appears as a subsequence in x, that is, the number of strictly increasing tuples $(i_0, \dots, i_{|w|-1})$ such that $x_{i_j} = w_j$ for $0 \leq j \leq |w|-1$. The problem is to determine the minimum value of $T = T_{q,\delta}(n)$ for which there exists a function $F : (\cup_{j=0}^{n}\{0,1\}^j)^T \to \{0,1\}^n$ satisfying $\mathbb{P}_{\mu_x^T}[F(U^1, \dots, U^T) = x] \geq 1-\delta$ for each $x \in \{0,1\}^n$ (where the U^j denote the T independent traces).

Supressing the dependence on q and δ, Holenstein, Mitzenmacher, Panigrahy, and Wieder [32] established an upper bound, that $\exp(\widetilde{O}(n^{1/2}))$ traces suffice. Nazarov and Peres [48] and De, O'Donnell, and Servedio [19] simultaneously obtained the (previous) best upper bound known, that $\exp(O(n^{1/3}))$ random traces suffice.

In this chapter, we improve the upper bound on trace reconstruction to $\exp(\widetilde{O}(n^{1/5}))$.

Theorem 5.2.1. *For any deletion probability $q \in (0,1)$ and any $\delta > 0$, there exists $C > 0$ so that any unkown string $x \in \{0,1\}^n$ can be reconstructed with probability at least $1-\delta$ from $T = \exp(Cn^{1/5}\log^5 n)$ i.i.d. traces of x.*

Batu, Kannan, Khanna, and McGregor [7] proved a lower bound of $\Omega(n)$, which was improved to $\widetilde{\Omega}(n^{5/4})$ by Holden and Lyons [30], which was then improved to $\widetilde{\Omega}(n^{3/2})$ by the author [12].

A variant of the trace reconstruction problem requires one to, instead of reconstruct any string x from traces of it, reconstruct a string x chosen uniformly at random from traces of it. For a formal statement of the problem, see Section 1.2 of [30]. Peres and Zhai [51] obtained an upper bound of $\exp(O(\log^{1/2} n))$ for $q < \frac{1}{2}$, which was then improved to $\exp(O(\log^{1/3} n))$ for all (constant) q by Holden, Pemantle, Peres, and Zhai [31].

Holden and Lyons [30] proved a lower bound for this random variant of $\widetilde{\Omega}(\log^{9/4} n)$, which was then improved by the author [12] to $\widetilde{\Omega}(\log^{5/2} n)$.

Several other variants of the trace reconstruction problem have been considered. The interested reader should refer to [5], [6], [18], [15], [10], [46], [37], [47].

In a previous version of this chapter, we proved Theorem 5.2.1 only for $q \in (0, \frac{1}{2}]$. Shyam Narayanan found a short argument extending our methods to get all $q \in (0,1)$. He kindly allowed us to use his argument in this chapter.

We made no effort to optimize the (power of the) logarithmic term $\log^5 n$ in Theorem 5.2.1.

5.3 Notation

We index starting at 0. For strings w and x, we sometimes write $1_{x_{k+i}=w_i}$ as shorthand for $\prod_{i=0}^{|w|-1} 1_{x_{k+i}=w_i}$. Let $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$. The symbol $\mathbb{E}_x$ denotes the expectation under the probability distribution over traces generated by the string x. For a trace U, we define $U_j = 2$ for $j > |U|$; this is simply to make "$U_j = 0$" and "$U_j = 1$" both false. We use $0^0 := 1$. For a positive integer n, denote $[n] := \{1, \ldots, n\}$. For a function f and a set E, denote $||f||_E := \max_{z\in E}|f(z)|$. We say $A \subseteq \{0,\dots,n-1\}$ is *d-separated* if distinct $a, a' \in A$ have $|a-a'| \geq d$.

5.4 Sketch of Argument

The upper bound of $\exp(O(n^{1/3}))$ was obtained by analyzing the polynomial $\sum_k [x_k - y_k]z^k$ whose value can be well enough approximated from a sufficient number of traces. In this chapter, we analyze the polynomial $\sum_k [1_{x_{k+i}=w_i} - 1_{y_{k+i}=w_i}]z^k$, for a well-chosen (sub)string w; its value can be well enough approximated from a sufficient number of traces, provided $q \leq 1/2$. The benefit of this polynomial is that for certain choices of w, it is far sparser than the more general $\sum_k [x_k - y_k]z^k$. In the author's paper [13] improving the upper bound on the separating words problem, lower bounds were obtained for (the absolute value of) these sparser polynomials near 1 on the real axis that were superior to those for the more general $\sum_k [x_k - y_k]z^k$. We use the methods developed in that paper and methods used in [8] to obtain superior lower bounds for points on a small arc of the unit circle centered at 1.

5.5 Deduction of main theorem from complex analytic statement

Fix $q \in (0,1)$, and let $p = 1-q$. The following 'single bit statistics' identity was proven in [48, Lemma 2.1]; in it, U denotes a random trace of x.

$$\mathbb{E}_x\left[p^{-1}\sum_{0\leq j\leq n-1} 1_{U_j=1}\left(\frac{z-q}{p}\right)^j\right] = \sum_{0\leq k\leq n-1} 1_{x_k=1}z^k.$$

We shall use a generalization of this identity to approximate a weighted count (by position) of subsequence appearances in x rather than a weighted count (by position) of appearances of 1. Choosing variables appropriately will recover a weighted count of *(contiguous) substring* appearances in x. An unweighted version was used in [14].

Proposition 5.5.1. *For any $x \in \{0,1\}^n, l \geq 1, w \in \{0,1\}^l$, and $z_0,\dots,z_{l-1} \in \mathbb{C}$, we have*

$$\mathbb{E}_x\left[p^{-1}\sum_{j_0<\dots<j_{l-1}}\left(\prod_{i=0}^{l-1} 1_{U_{j_i}=w_i}\right)\left(\frac{z_0-q}{p}\right)^{j_0}\left(\prod_{i=1}^{l-1}\left(\frac{z_i-q}{p}\right)^{j_i-j_{i-1}-1}\right)\right]$$

$$= \sum_{k_0<\dots<k_{l-1}}\left(\prod_{i=0}^{l-1} 1_{x_{k_i}=w_i}\right) z_0^{k_0}\left(\prod_{i=1}^{l-1} z_i^{k_i-k_{i-1}-1}\right).$$

Proof. By basic combinatorics, the left hand side of the above is

$$p^{-l}\sum_{j,\Delta_1,\dots,\Delta_{l-1}}\sum_{k_0<\dots<k_{l-1}}\left(\prod_{i=0}^{l-1}1_{x_{k_i}=w_i}\right)\binom{k_0}{j}\binom{k_1-k_0-1}{\Delta_1-1}$$
$$\times\binom{k_2-k_1-1}{\Delta_2-1}\times\dots\times\binom{k_{l-1}-k_{l-2}-1}{\Delta_{l-1}-1}$$
$$\times p^{j+\Delta_1+\dots+\Delta_{l-1}+1}q^{k_{l-1}+1-(j+\Delta_1+\dots+\Delta_{l-1}+1)}$$
$$\times(\frac{z_0-q}{p})^j(\frac{z_1-q}{p})^{\Delta_1-1}\dots(\frac{z_{l-1}-q}{p})^{\Delta_{l-1}-1}$$
$$=\sum_{k_0<\dots<k_{l-1}}\left(\prod_{i=0}^{l-1}1_{x_{k_i}=w_i}\right)\left(\sum_j\binom{k_0}{j}(z_0-q)^jq^{k_0-j}\right)$$
$$\times\left(\sum_{\Delta_1}\binom{k_1-k_0-1}{\Delta_1-1}(z_1-q)^{\Delta_1-1}q^{k_1-k_0-1-(\Delta_1-1)}\right)\times\dots$$
$$\dots\times\left(\sum_{\Delta_{l-1}}\binom{k_{l-1}-k_{l-2}-1}{\Delta_{l-1}-1}(z_{l-1}-q)^{\Delta_{l-1}-1}q^{k_{l-1}-k_{l-2}-1-(\Delta_{l-1}-1)}\right).$$

The binomial theorem finishes the proof. □

Let $\mathcal{P}_n$ be the set of all polynomials[1] $p(z)=1-\sigma z^d+\sum_{j=n^{1/5}}^n c_jz^j\in\mathbb{C}[z]$ with $1\le d<n^{1/5},\sigma\in\{0,1\}$, and $|c_j|\le 1$ for each j.

We prove the following theorem in the next section. We assume it to be true until then.

Theorem 5.5.2. *There is some $C>0$ so that for any $n\ge 2$ and any $p\in\mathcal{P}_n$,*

$$\max_{|\theta|\le n^{-2/5}}|p(e^{i\theta})|\ge\exp(-Cn^{1/5}\log^5 n).$$

Proposition 5.5.3. *For any distinct $x,y\in\{0,1\}^n$ with $x_i=y_i$ for all $0\le i<2n^{1/5}-1$, there are $w\in\{0,1\}^{2n^{1/5}}$ and $z_0\in\{e^{i\theta}:|\theta|\le n^{-2/5}\}$ such that*

$$\left|\sum_k[1_{x_{k+i}=w_i}-1_{y_{k+i}=w_i}]z_0^k\right|\ge\exp(-Cn^{1/5}\log^5 n).$$

[1] Throughout the chapter, we omit floor functions when they don't meaningfully affect anything.

Proof. Let $i \geq 2n^{1/5} - 1$ be the first index with $x_i \neq y_i$. Let $w' = x_{i-2n^{1/5}+1}, \ldots, x_{i-1}$. As used in [13], Lemmas 1 and 2 of [55] imply that there is some choice $w \in \{w'0, w'1\}$ such that the indices k for which $x_{k+i} = w_i$ for all $0 \leq i \leq 2n^{1/5} - 1$ are $n^{1/5}$-separated, and such that the indices k for which $y_{k+i} = w_i$ for all $0 \leq i \leq 2n^{1/5} - 1$ are $n^{1/5}$-separated. Therefore, if $p(z) := \sum_k [1_{x_{k+i}=w_i} - 1_{y_{k+i}=w_i}] z^k$, then $\epsilon \frac{p(z)}{z^m} \in \mathcal{P}_n$ for some $\epsilon \in \{-1, 1\}$ and $0 \leq m \leq n-1$. Thus, by Theorem 5.5.2, there is some $\theta \in [-n^{-2/5}, n^{-2/5}]$ such that $\exp(-Cn^{1/5} \log^5 n) \leq |\epsilon \frac{p(e^{i\theta})}{e^{im\theta}}| = |p(e^{i\theta})|$. Take $z_0 = e^{i\theta}$. □

In a previous version of this chapter, we used Proposition 5.5.1 with $z_1, \ldots, z_{l-1} = 0$ and z_0 chosen according to Proposition 5.5.3 to prove Theorem 5.2.1, which only worked for $q \leq 1/2$, since, for $q > 1/2$, the quantity $(-q/p)^{j_i - j_{i-1}}$ would be too large in magnitude (for $j_i - j_{i-1} \approx n$), leading to too large a variance to well-enough approximate $\sum_k [1_{x_{k+i}=w_i} - 1_{y_{k+i}=w_i}] z_0^k$ with few traces. Following an idea of Shyam Narayanan, we choose $z_1, \ldots, z_{l-1}$ close to 1 so that $(\frac{z_i - q}{p})^{j_i - j_{i-1}}$ would no longer be too large in magnitude, while also keeping the right hand side of Proposition 5.5.1 not too small. The following corollary, due to him, establishes the existence of such $z_1, \ldots, z_{l-1}$.

Corollary 5.5.4. *For any distinct $x, y \in \{0,1\}^n$ with $x_i = y_i$ for all $0 \leq i < l - 1 := 2n^{1/5} - 1$, there are $w \in \{0,1\}^l$, $z_0 \in \{e^{i\theta} : |\theta| \leq n^{-2/5}\}$, and $z_1, \ldots, z_{l-1} \in [1-2p, 1]$ such that*[2]

$$\left| \sum_{k_0 < \cdots < k_{l-1}} [1_{x_{k_i}=w_i} - 1_{y_{k_i}=w_i}] z_0^{k_0} z_1^{k_1 - k_0 - 1} \ldots z_{l-1}^{k_{l-1} - k_{l-2} - 1} \right| \geq \exp(-C' n^{1/5} \log^5 n).$$

Proof. Let w and z_0 be those guaranteed by Proposition 5.5.3. Let

$$f(z_1) = \binom{n}{2n^{1/5}}^{-1} \sum_{k_0 < \cdots < k_{l-1}} [1_{x_{k_i}=w_i} - 1_{y_{k_i}=w_i}] z_0^{k_0} z_1^{k_{l-1} - k_0 - (l-1)}.$$

Note that f is a polynomial in z_1 with each coefficient trivially upper bounded by 1 in absolute value. Therefore, by Theorem 5.1 of [9],

$$\begin{aligned}
\binom{n}{2n^{1/5}} \max_{z_1 \in [1-2p, 1]} |f(z_1)| &\geq \binom{n}{2n^{1/5}} |f(0)|^{c_1/(2p)} e^{-c_2/(2p)} \\
&\geq \binom{n}{2n^{1/5}} \left(\binom{n}{2n^{1/5}}^{-1} \exp(-Cn^{1/5} \log^5 n) \right)^{c_1/(2p)} e^{-c_2/(2p)} \\
&\geq \exp(-C' n^{1/5} \log^5 n).
\end{aligned}$$

[2] We similarly abuse notation by writing $1_{x_{k_i}=w_i}$ to denote $\prod_{i=0}^{l-1} 1_{x_{k_i}=w_i}$.

The corollary then follows by taking a z_1 realizing this maximum and then setting $z_2, \ldots, z_{l-1} = z_1$. $\square$

We are now ready to establish our main theorem. We encourage the reader to first read the proof of the $\exp(O(n^{1/3}))$ upper bound in [48].

Proof of Theorem 5.2.1. Take distinct $x, y \in \{0,1\}^n$. By padding the beginning of x and y with $2n^{1/5}$ many 0s if needed, we may assume the first i for which $x_i \neq y_i$ satisfies $i \geq 2n^{1/5} - 1$. Let $w, z_0, z_1, \ldots, z_{2n^{1/5}-1}$ be those guaranteed by Corollary 5.5.4. Since $z_1, \ldots, z_{2n^{1/5}-1} \in [1-2p, 1]$, each of $\frac{z_i - q}{p}$, $1 \leq i \leq 2n^{1/5} - 1$, is between -1 and 1, and so the expression in brackets in Proposition 5.5.1 has magnitude upper bounded by $n|\frac{z_0-q}{p}|^n 2^{2n^{1/5}}$, which, by the choice of z_0, is upper bounded by $n \exp(C\frac{n}{n^{4/5}})2^{2n^{1/5}}$ (see [48, (2.3)] for details). Therefore, since the expression in brackets in Proposition 5.5.1 is a function of just the observed traces, by Corollary 5.5.4 and a standard Höeffding inequality argument (see [48] for details; note the pigeonhole is not necessary), we see that $\exp(C'''n^{1/5}\log^5 n)$ traces suffice to distinguish between x and y. As explained in [48], this "pairwise upper bound" in fact suffices to establish Theorem 5.2.1. $\square$

5.6 Proof of complex analytic statement

We may of course assume n is large.

Let $a = n^{-2/5}$ and $r = a^{-1/2}$. Let $r_* \in [r]$ be such that

$$\sum_{j=1}^{r_*} \frac{1}{\log^2(j+3)} - \sum_{j=r_*+1}^{r} \frac{1}{\log^2(j+3)} \in [20, 21];$$

such an r_* clearly exists. Let

$$\begin{cases} \epsilon_j = +1 & \text{if } 1 \leq j \leq r_* \\ \epsilon_j = -1 & \text{if } r_* + 1 \leq j \leq r \end{cases}.$$

Let $\lambda_a \in (1,2)$ be such that

$$\sum_{j=1}^{r} \frac{\lambda_a}{j^2 \log^2(j+3)} = 1.$$

Let

$$d_j = \frac{\lambda_a}{j^2 \log^2(j+3)}.$$

Define

$$\widetilde{h}(z) = \widetilde{\lambda}_a \sum_{j=1}^{r} \epsilon_j d_j z^j,$$

where $\widetilde{\lambda}_a \in (1, 2)$ is such that $\widetilde{h}(1) = 1$. Define

$$h(z) = (1 - a^{10})\widetilde{h}(z).$$

Let

$$\alpha = e^{ia}, \beta = e^{-ia},$$

and

$$I_t = \{z \in \mathbb{C} : \arg(\frac{\alpha - z}{z - \beta}) = t\}$$

for $t \geq 0$. Note that I_0 is the line segment connecting α and β and $I_a = \{e^{i\theta} : |\theta| \leq a\}$ is the set on which we wish to lower bound p at some point. Let

$$G_a = \{z \in \mathbb{C} : \arg(\frac{\alpha - z}{z - \beta}) \in (\frac{a}{2}, a)\}$$

be the open region bounded by $I_{a/2}$ and I_a.

As in [13], we needed our choice of h to satisfy (i) $|h(e^{2\pi it})| \leq 1 - c|t|$ for $|t| > a^{1/2}$ (up to logs). In this chapter, we need (ii) $|h(e^{2\pi it})| \geq 1 - Ca^2$ for $|t| \approx a$; in [13], we instead had $|h(e^{2\pi it})| \approx 1 - a$ for $|t| \approx a$. Some thought shows that a polynomial with positive coefficients will not work. We therefore had roughly half of our coefficients be -1 so that (ii) holds; changing those coefficients doesn't affect (i) since the corresponding degrees are large. However, due to our required normalization that $h(1)$ is basically 1, the negative coefficients make it so that h might no longer map into the unit disk, which is highly problematic for later application. Luckily, though, $\widetilde{h}$, and thus h, *does* map into the unit disk. We prove that in the appendix.

Lemma 5.6.1. *For any $t \in [-\pi, \pi]$, $\widetilde{h}(e^{it}) \in \overline{\mathbb{D}}$.*

Lemma 5.6.2. *There are absolute constants $c_4, c_5, C_6 > 0$ such that the following hold for $a > 0$ small enough. First, $h(e^{2\pi it}) \in G_a$ for $|t| \leq c_4 a$. Second, $|h(e^{2\pi it})| \leq 1 - c_5 \frac{|t|}{\log^2(a^{-1})}$ for $t \in [\frac{-1}{2}, \frac{1}{2}] \setminus [-C_6 a^{1/2}, C_6 a^{1/2}]$.*

Proof. Take $|t| \leq a$. Then,

$$\begin{aligned}\widetilde{h}(e^{2\pi it}) = \widetilde{\lambda}_a \sum_{j=1}^{r_*} \frac{\lambda_a}{j^2 \log^2(j+3)}(1 + 2\pi itj - 2\pi^2 t^2 j^2 + O(t^3 j^3)) \\ - \widetilde{\lambda}_a \sum_{j=r_*+1}^{r} \frac{\lambda_a}{j^2 \log^2(j+3)}(1 + 2\pi itj - 2\pi^2 t^2 j^2 + O(t^3 j^3)).\end{aligned}$$

By our choice of r_*, $h(e^{2\pi it}) = 1 - \delta + \epsilon i$ for $\delta := c_1 t^2 + a^{10} + O(\frac{t^3 r^2}{\log^2 r})$ and $\epsilon := c_2 t + O(\frac{t^3 r^2}{\log^2 r})$, where c_1, c_2 are bounded positive quantities that are bounded away from 0. By multiplying the denominator by its conjugate, we have

$$\arg\left(\frac{e^{ia} - (1 - \delta + \epsilon i)}{(1 - \delta + \epsilon i) - e^{-ia}}\right) = \arg\Big(\big(e^{ia} - (1 - \delta + \epsilon i)\big) \cdot \big((1 - \delta - \epsilon i) - e^{ia}\big)\Big).$$

The ratio of the imaginary part to the real part of the term inside $\arg(\cdot)$ is

$$\frac{2(1 - \delta - \cos a)\sin a}{-\cos^2 a + 2(1 - \delta)\cos a - (1 - \delta)^2 + \sin^2 a - \epsilon^2}.$$

Writing $\cos a = 1 - \frac{1}{2}a^2 + O(a^4)$ and $\sin a = a + O(a^3)$, and using $\delta = O(a^2)$, the above simplifies to

$$\frac{a^3 - 2a\delta + O(a^4)}{a^2 - \epsilon^2 + O(a^3)}.$$

If $|t| \leq c_4 a$, then, as $\delta = c_1 t^2 + a^{10} + O(\frac{t^3 r^2}{\log^2 r}), \epsilon = c_2 t + O(\frac{t^3 r^2}{\log^2 r})$, the inverse tangent of the above is at least $\frac{a}{2}$; the arctangent is at most a, since, by Lemma 5.6.1, $h(e^{2\pi it})$ lies in the unit disk (alternatively, one may note $2a\delta > \epsilon^2$).

We now establish the second part of the lemma. What [13] shows is

$$\left|\sum_{j=1}^{m} \frac{\lambda_a e^{2\pi itj}}{j^2 \log^2(j + 3)}\right| \leq 1 - \frac{\lambda_a |t|}{3\log^2(m + 3)} + \frac{\lambda_a}{m\log^2(m + 3)}$$

for any $m \geq 1$ and $t \in [-\frac{1}{2}, \frac{1}{2}] \setminus [-3m^{-1}, 3m^{-1}]$. For $m = r_*$, if $|t| > C_6 a^{1/2}$, for say $C_6 = 100$, then certainly $3|t|^{-1} < m$, and so we have

$$\left|\sum_{j=1}^{r_*} \frac{\lambda_a e^{2\pi itj}}{j^2 \log^2(j + 3)}\right| \leq 1 - c\frac{|t|}{\log^2(a^{-1})}. \tag{5.1}$$

We can crudely bound

$$\left|\sum_{j=r_*+1}^{r} \frac{\lambda_a e^{2\pi itj}}{j^2 \log^2(j + 3)}\right| \leq \frac{4}{\log^2(a^{-1})}\frac{1}{r_*}. \tag{5.2}$$

Combining (5.1) and (5.2), we obtain

$$\left|\sum_{j=1}^{r} \frac{\lambda_a \epsilon_j e^{2\pi itj}}{j^2 \log^2(j + 3)}\right| \leq 1 - c_5' \frac{|t|}{\log^2(a^{-1})}$$

for $|t| \geq C_6 r^{-1}$, with $c_5' > 0$ small and C_6 large enough. Now, since

$$\begin{aligned}
\widetilde{\lambda}_a^{-1} &= \sum_{j=1}^{r_*} \frac{\lambda_a}{j^2 \log^2(j+3)} - \sum_{j=r_*+1}^{r} \frac{\lambda_a}{j^2 \log^2(j+3)} \\
&= 1 - 2 \sum_{j=r_*+1}^{r} \frac{\lambda_a}{j^2 \log^2(j+3)} \\
&\geq 1 - 2 \frac{2}{\log^2(a^{-1})} \frac{2}{r_*} \\
&\geq 1 - \frac{20}{r \log^2(a^{-1})},
\end{aligned}$$

we see

$$\left| \widetilde{\lambda}_a \sum_{j=1}^{r} \frac{\lambda_a \epsilon_j e^{2\pi i t j}}{j^2 \log^2(j+3)} \right| \leq 1 - c_5 \frac{|t|}{\log^2(a^{-1})}$$

for $|t| \geq C_6 r^{-1}$, provided C_6 is large enough. Since $1 - a^{10} \leq 1$, we are done. □

Let $m = c_4^{-1} n^{2/5}$, $J_1 = c_5^{-1} n^{-1/5} m \log^4 n$, and $J_2 = m - J_1$. A minor adapation of the relevant proof in [13] proves the following.

Lemma 5.6.3. *Suppose $\widetilde{p}(z) = 1 - z^d$ for some $d \leq n^{1/5}$. Then*

$$\prod_{j=J_1}^{J_2-1} \left| \widetilde{p}\big(h(e^{2\pi i \frac{j+\delta}{m}})\big) \right| \leq \exp(C n^{1/5} \log^5 n)$$

for any $\delta \in [0,1)$.

By adapating the proof of the above lemma, we prove the following.

Lemma 5.6.4. *Suppose $u(z) = z - \zeta$ for some $\zeta \in \partial\mathbb{D}$. Then, for any $\delta \in [0,1)$, we have*

$$\prod_{j=J_1}^{J_2-1} \left| u\big(h(e^{2\pi i \frac{j+\delta}{m}})\big) \right| \leq \exp(C n^{1/5} \log^5 n).$$

Proof. First note that

$$|u(h(e^{2\pi i \theta}))| \geq 1 - |h(e^{2\pi i \theta})| \geq a^{10}. \tag{5.3}$$

Define $g(t) = 2 \log |u(h(e^{2\pi i (t + \frac{\delta}{m})}))|$. For notational ease, we assume $\delta = 0$; the argument about to come works for all $\delta \in [0,1)$. Since (5.3) implies g is C^1, by the mean value theorem we have

$$\begin{aligned}\left|\frac{1}{m}\sum_{j=J_1}^{J_2-1} g\left(\frac{j}{m}\right)-\int_{J_1/m}^{J_2/m} g(t)dt\right| &= \left|\sum_{j=J_1}^{J_2-1}\int_{j/m}^{(j+1)/m}\left(g(t)-g\left(\frac{j}{m}\right)\right)dt\right| \\ &\le \sum_{j=J_1}^{J_2-1}\int_{j/m}^{(j+1)/m}\left(\max_{\frac{j}{m}\le y\le\frac{j+1}{m}}|g'(y)|\right)\frac{1}{m}dt \\ &\le \frac{1}{m^2}\sum_{j=J_1}^{J_2-1}\max_{\frac{j}{m}\le y\le\frac{j+1}{m}}|g'(y)|. \qquad (5.4)\end{aligned}$$

Since $w \mapsto \log|u(h(w))|$ is harmonic and $\log|u(h(0))| = \log|u(0)| = 0$, we have

$$\int_0^1 g(t)dt = 2\int_0^1 \log|u(h(e^{2\pi it}))|dt = 0,$$

and therefore

$$\left|\int_{J_1/m}^{J_2/m} g(t)dt\right| \le \left|\int_0^{J_1/m} g(t)dt\right| + \left|\int_{J_2/m}^1 g(t)dt\right|. \qquad (5.5)$$

Since

$$a^{10} \le \left|u(h(e^{2\pi it}))\right| \le 2$$

for each t, we have

$$\left|\int_0^{J_1/m} g(t)dt\right| + \left|\int_{J_2/m}^1 g(t)dt\right| \le 20\left(\frac{J_1}{m}+(1-\frac{J_2}{m})\right)\log n \le C\frac{\log^5 n}{n^{1/5}}. \qquad (5.6)$$

By (5.4), (5.5), and (5.6), we have

$$\left|\frac{1}{m}\sum_{j=J_1}^{J_2-1} g(\frac{j}{m})\right| \le C\frac{\log^5 n}{n^{1/5}} + \frac{1}{m^2}\sum_{j=J_1}^{J_2-1}\max_{\frac{j}{m}\le t\le\frac{j+1}{m}}|g'(t)|.$$

Multiplying through by m, changing C slightly, and exponentiating, we obtain

$$\prod_{j=J_1}^{J_2-1}\left|u(h(e^{2\pi i\frac{j}{m}}))\right|^2 \le \exp\left(Cn^{1/5}\log^5 n + \frac{1}{m}\sum_{j=J_1}^{J_2-1}\max_{\frac{j}{m}\le t\le\frac{j+1}{m}}|g'(t)|\right). \qquad (5.7)$$

Note

$$g'(t_0) = \frac{\frac{\partial}{\partial t}\left[|u(h(e^{2\pi it}))|^2\right]\Big|_{t=t_0}}{|u(h(e^{2\pi it_0}))|^2}.$$

We first show

$$\frac{\partial}{\partial t}\left[|u(h(e^{2\pi it}))|^2\right]\Big|_{t=t_0} \le 500 \qquad (5.8)$$

for each $t_0 \in [0,1]$. Let $\widetilde{d_j} = d_j$ for $j \leq r_*$ and $\widetilde{d_j} = -d_j$ for $j > r_*$ so that $h(e^{2\pi it}) = (1-a^{10})\sum_{j=1}^{r} \widetilde{d_j} e^{2\pi itj}$. Then,

$$\left|u\left(h(e^{2\pi it})\right)\right|^2 = \left|(1-a^{10})\sum_{j=1}^{r} \widetilde{d_j} e^{2\pi ijt} - \zeta\right|^2$$

$$= (1-a^{10})^2 \left|\sum_{j=1}^{r} \widetilde{d_j} e^{2\pi ijt}\right|^2 - 2\operatorname{Re}\left[(1-a^{10})\zeta \sum_{j=1}^{r} \widetilde{d_j} e^{2\pi ijt}\right] + 1. \tag{5.9}$$

The derivative of the first term is

$$(1-a^{10})^2 \sum_{j_1,j_2=1}^{r} \widetilde{d_{j_1}} \widetilde{d_{j_2}} 2\pi(j_1 - j_2) e^{2\pi i(j_1-j_2)t}.$$

Since

$$\sum_{j=1}^{r} |\widetilde{d_j}| \leq 4$$

and

$$\sum_{j=1}^{r} j|\widetilde{d_j}| \leq 4,$$

we get an upper bound of 250 for the absolute value of the derivative of the first term of (5.9). The derivative of the second term, if $\zeta = e^{i\theta}$, is

$$2(1-a^{10})\sum_{j=1}^{r} \widetilde{d_j} \sin(2\pi jt + \theta) 2\pi j,$$

which is also clearly upper bounded by (crudely) 250. We've thus shown (5.8).

Recall $|u(h(e^{2\pi i\theta}))| \geq 1 - |h(e^{2\pi i\theta})|$. For $j \in [J_1, J_2] \subseteq [C_6 a^{1/2} m, (1 - C_6 a^{1/2})m]$, we use (by Lemma 5.6.2)

$$|h(e^{2\pi i \frac{j}{m}})| \leq 1 - c_5 \frac{\min(\frac{j}{m}, 1 - \frac{j}{m})}{\log^2 n}$$

to obtain

$$\frac{1}{m}\sum_{j=J_1}^{J_2-1} \max_{\frac{j}{m} \leq t \leq \frac{j+1}{m}} |g'(t)| \leq \frac{1}{m}\sum_{j=J_1}^{J_2-1} \frac{500}{\left(c_5 \frac{\min(\frac{j}{m}, 1-\frac{j}{m})}{\log^2 n})\right)^2}.$$

Up to a factor of 2, we may deal only with $j \in [J_1, \frac{m}{2}]$. Then we obtain

$$\begin{aligned}\frac{1}{m}\sum_{j=J_1}^{J_2-1} \max_{\frac{j}{m} \leq t \leq \frac{j+1}{m}} |g'(t)| &\leq \frac{1}{m}\sum_{j=J_1}^{m/2} \frac{500 m^2 \log^4 n}{c_5^2 j^2} \\ &\leq \frac{500 m \log^4 n}{c_5^2} \frac{2}{J_1} \\ &\leq C n^{1/5}.\end{aligned}$$

□

Let $\mathcal{Q}_n$ denote all polynomials of the form $(z-\alpha)(z-\beta)p(z)$ for $p \in \mathcal{P}_n$.

Corollary 5.6.5. *For any $q \in \mathcal{Q}_n$ and $\delta \in [0,1)$, we have*

$$\prod_{j \notin \{0,m-1\}} |q(h(e^{2\pi i \frac{j+\delta}{m}} z))| \le \exp(Cn^{1/5}\log^5 n).$$

Proof. Take $q \in \mathcal{Q}_n$; say $q(z) = (z-\alpha)(z-\beta)p(z)$ for $p \in \mathcal{P}_n$. For $j \in \{1,\dots,J_1-1\}$ and for $j \in \{J_2,\dots,m-2\}$, by Lemma 5.6.1 we can bound $|q(h(e^{2\pi i \frac{j}{m}}z))| \le 4n$, to obtain

$$\prod_{j \notin \{J_1,\dots,J_2-1\}} |q(h(e^{2\pi i \frac{j+\delta}{m}}))| \le (4n)^{J_1-1+m-J_2-1} \le e^{Cn^{1/5}\log^5 n}. \tag{5.10}$$

By applying Lemma 5.6.4 to $u(z) := z-\alpha$ and to $u(z) := z-\beta$ and multiplying the results, we see

$$\prod_{j=J_1}^{J_2-1} |\overline{u}(h(e^{2\pi i \frac{j+\delta}{m}}))| \le e^{Cn^{1/5}\log^5 n}, \tag{5.11}$$

where $\overline{u}(z) := (z-\alpha)(z-\beta)$. Let $\widetilde{p}(z) \in \{1, 1-z^d\}$ be the truncation of p to terms of degree less than $n^{1/5}$. Then, since Lemma 5.6.2 gives

$$|h(e^{2\pi i \frac{j+\delta}{m}})| \le 1 - c_5 \frac{\min\left(\frac{j}{m}+\delta, 1-(\frac{j}{m}+\delta)\right)}{\log^2 n} \le 1 - c'n^{-1/5}\log^2 n$$

for $j \in \{J_1,\dots,J_2-1\}$, we see

$$\left| p\Big(h(e^{2\pi i \frac{j+\delta}{m}})\Big) - \widetilde{p}\Big(h(e^{2\pi i \frac{j+\delta}{m}})\Big) \right| \le n e^{-c'\log^2 n} \le e^{-c\log^2 n}. \tag{5.12}$$

Lemma 5.6.3 implies

$$\prod_{j=J_1}^{J_2-1} |\widetilde{p}(h(e^{2\pi i \frac{j+\delta}{m}}))| \le e^{Cn^{1/5}\log^5 n}. \tag{5.13}$$

By an easy argument given in [13], (5.12) and (5.13) combine to give

$$\prod_{j=J_1}^{J_2-1} |p(h(e^{2\pi i \frac{j+\delta}{m}}))| \le e^{C'n^{1/5}\log^5 n}. \tag{5.14}$$

Combining (5.10), (5.11), and (5.14), the proof is complete. □

Proposition 5.6.6. *For any $q \in \mathcal{Q}_n$, it holds that*

$$\max_{w \in G_a} |q(w)| \ge \exp(-Cn^{1/5}\log^5 n).$$

Proof. Let $g(z) = \prod_{j=0}^{m-1} q(h(e^{2\pi i \frac{j}{m}} z))$. For $z = e^{2\pi i\theta}$, with, without loss of generality, $\theta \in [0, \frac{1}{m})$, we have by Lemma 5.6.2 and Corollary 5.6.5

$$|g(z)| \leq \left(\max_{w\in G_a} |q(w)|\right)^2 \prod_{j\notin\{0,m-1\}} |q(h(e^{2\pi i(\frac{j}{m}+\theta)}))| \leq \left(\max_{w\in G_a} |q(w)|\right)^2 \exp(Cn^{1/5}\log^5 n).$$

Thus, $(\max_{w\in G_a} |q(w)|)^2 \exp(Cn^{1/5}\log^5 n) \geq \max_{z\in\partial\mathbb{D}} |g(z)| \geq |g(0)| = 1$, where the last inequality used the maximum modulus principle (clearly g is analytic). □

The following lemma was proven in [8].

Lemma 5.6.7. *Suppose g is an analytic function in the open region bounded by I_0 and I_a, and suppose g is continuous on the closed region between I_0 and I_a. Then,*

$$\max_{z\in I_{a/2}} |g(z)| \leq \left(\max_{z\in I_0} |g(z)|\right)^{1/2} \left(\max_{z\in I_a} |g(z)|\right)^{1/2}.$$

Proof of Theorem 5.5.2. Take $f \in \mathcal{P}_n$, and let $g(z) = (z-\alpha)(z-\beta)f(z)$. A straightforward geometric argument yields

$$|g(z)| \leq \frac{|(z-\alpha)(z-\beta)|}{1-|z|} \leq \frac{2}{\sin a} \leq 3n^{2/5}$$

for $z \in I_0$. Letting $L = ||g||_{I_a}$, Lemma 5.6.7 then gives

$$\max_{z\in I_{a/2}} |g(z)| \leq (3Ln^{2/5})^{1/2}.$$

Since we then have

$$\max_{z\in I_{a/2}\cup I_a} |g(z)| \leq \max(L, (3Ln^{2/5})^{1/2}),$$

the maximum modulus principle implies

$$\max_{z\in G_a} |g(z)| \leq \max(L, (3Ln^{2/5})^{1/2}).$$

By Proposition 5.6.6, we conclude

$$\exp(-Cn^{1/5}\log^5 n) \leq \max\left(L, (3Ln^{2/5})^{1/2}\right).$$

Thus,

$$||f||_{I_a} \geq \frac{1}{4}||g||_{I_a} = \frac{L}{4} \geq \exp(-C'n^{1/5}\log^5 n),$$

as desired. □

5.7 Appendix: proof of lemma 5.6.1

We thank Fedor Nazarov for a simpler proof of Lemma 5.6.1, which we include below.

Claim 5.7.1. *Let $\mathcal{F}$ be a compact family of (uniformly) bounded real Lipschitz functions on $[0,1]$ such that $\int_0^{1/2} f < \int_{1/2}^1 f$ for every $f \in \mathcal{F}$. Then there exist $M, \epsilon > 0$ so that for all $m > M$, $m_* \in ((\frac{1}{2}-\epsilon)m, (\frac{1}{2}+\epsilon)m)$, and $f \in \mathcal{F}$, it holds that*

$$\sum_{j=1}^{m_*} \frac{1}{\log^2(j+3)} f\left(\frac{j}{m}\right) < \sum_{j=m_*+1}^{m} \frac{1}{\log^2(j+3)} f\left(\frac{j}{m}\right). \tag{5.15}$$

Proof. By compactness, there exists $\epsilon > 0$ so that for all $\gamma \in (\frac{1}{2}-\epsilon, \frac{1}{2}+\epsilon)$ and all $f \in \mathcal{F}$, we have

$$\int_0^{\gamma} f(x)dx < \int_{\gamma}^{1} f(x)dx - \epsilon. \tag{5.16}$$

Quickly note, for $C > 0$ a uniform upper bound on $\max_{x\in[0,1]} |f(x)|$, $f \in \mathcal{F}$, we have

$$\begin{aligned} \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{\log^2(j+3)} - \frac{1}{\log^2(m+3)}\right]\left|f\left(\frac{j}{m}\right)\right| &\le \frac{C}{m}\left[\sum_{j=1}^{\frac{m}{\log^3(m+3)}} 1 + \sum_{j=\frac{m}{\log^3(m+3)}}^{m} \frac{\log\log(m+3)}{\log^3(m+3)}\right] \qquad (5.17)\\ &\le 2C\frac{\log\log(m+3)}{\log^3(m+3)} \\ &= o(\frac{1}{\log^2(m+3)}) \end{aligned}$$

as $m \to \infty$. As (5.15) is equivalent to

$$\frac{\log^2(m+3)}{m}\sum_{j=1}^{m_*} \frac{1}{\log^2(j+3)} f\left(\frac{j}{m}\right) < \frac{\log^2(m+3)}{m}\sum_{j=m_*+1}^{m} \frac{1}{\log^2(j+3)} f\left(\frac{j}{m}\right),$$

by (5.17) it suffices to prove

$$\frac{1}{m}\sum_{j=1}^{m_*} f\left(\frac{j}{m}\right) < \frac{1}{m}\sum_{j=m_*+1}^{m} f\left(\frac{j}{m}\right) - \frac{\epsilon}{2}, \tag{5.18}$$

say (for m large enough and $m_* \in ((\frac{1}{2}-\epsilon)m, (\frac{1}{2}+\epsilon), m))$. But the LHS becomes arbitrarily close to $\int_0^{m_*/m} f(x)dx$, and the RHS becomes arbitrarily close to $\int_{m_*/m}^1 f(x)dx - \frac{\epsilon}{2}$, so (5.18) is established by (5.16). □

Now, letting $f(x) = \frac{1}{2} - \frac{1}{2}\left(\frac{\sin(x/2)}{x/2}\right)^2$ for $x \in (0,1]$ and $f(0) = 0$, and then setting $f_c(x) = c^{-4}f(cx)$ for $c > 0$ and $x \in [0,1]$ and $f_0(x) = \frac{x^4}{24}$, we will apply

Claim 5.7.1 to the family $\mathcal{F} := \{f_c : c \in [0, C]\}$, for a suitable absolute $C > 0$. An easy computation shows that $\mathcal{F}$ is indeed a compact family of bounded Lipschitz functions. The condition that $\int_0^{1/2} f_c < \int_{1/2}^1 f_c$ for all $c \in [0, C]$ is equivalent to $\int_0^a f(x)dx < \int_a^{2a} f(x)$ for all $a > 0$, which is equivalent to

$$\int_0^b \left(\frac{\sin x}{x}\right)^2 dx > \int_b^{2b} \left(\frac{\sin x}{x}\right)^2 dx$$

for all $b > 0$, which is easily verified[3].

Proof of Lemma 5.6.1. The proof of Lemma 5.6.2 shows that $\tilde{h}(e^{it}) \in \overline{D}$ if $t \in [-\pi, \pi] \setminus [-\frac{1}{100}, \frac{1}{100}]$, say. So we may assume $|t| \leq \frac{1}{100}$. First note that

$$\begin{aligned} \left|\mathrm{Im}[\widetilde{h}(e^{it})]\right| &= \widetilde{\lambda}_a \sum_{j=1}^{r} \epsilon_j d_j \sin(jt) \\ &\leq \widetilde{\lambda}_a \sum_{j=1}^{r} d_j j |t| \\ &\leq 2|t|. \end{aligned} \tag{5.19}$$

Also,

$$\begin{aligned} \mathrm{Re}[\widetilde{h}(e^{it})] &= \widetilde{\lambda}_a \sum_{j=1}^{r} \epsilon_j d_j \cos(jt) \\ &\geq \widetilde{\lambda}_a \sum_{j=1}^{r} \epsilon_j d_j \left(1 - \frac{j^2 t^2}{2}\right) \\ &= 1 - \frac{1}{2} t^2 \widetilde{\lambda}_a \sum_{j=1}^{r} \epsilon_j j^2 d_j \\ &\geq 1 - \frac{1}{2} t^2 \widetilde{\lambda}_a \cdot 21 \\ &> 0. \end{aligned} \tag{5.20}$$

Finally, using the identity

$$\frac{\cos x - 1 + \frac{x^2}{2}}{x^2} = \frac{1}{2} - \frac{1}{2}\left(\frac{\sin(x/2)}{x/2}\right)^2,$$

we see that

$$\mathrm{Re}[\widetilde{h}(e^{it})] = \widetilde{\lambda}_a \left(\sum_{j=1}^{r_*} \frac{1}{\log^2(j+3)} \left(\frac{1}{j^2} - \frac{t^2}{2}\right) - \sum_{j=r_*+1}^{r} \frac{1}{\log^2(j+3)} \left(\frac{1}{j^2} - \frac{t^2}{2}\right) \right)$$

[3] As $\frac{\sin x}{x}$ decreases on $[0, \pi]$, the case $b \leq \frac{\pi}{2}$ is immediate. For $b > \frac{\pi}{2}$, we can do $\int_b^{2b} (\frac{\sin x}{x})^2 dx < \int_{\pi/2}^{\infty} \frac{1}{x^2} dx = \frac{2}{\pi}$, which suffices since, by monotonicity, $\int_0^b (\frac{\sin x}{x})^2 dx > \int_0^{\pi/2} (\frac{\sin x}{x})^2 dx \geq \frac{\pi}{2} (\frac{2}{\pi})^2 = \frac{2}{\pi}$.

$$+\widetilde{\lambda}_a r^4 t^6 \left(\sum_{j=1}^{r_*} \frac{1}{\log^2(j+3)} f_{tr}(\frac{j}{r}) - \sum_{j=r_*+1}^{r} \frac{1}{\log^2(j+3)} f_{tr}(\frac{j}{r}) \right).$$

By Claim 5.7.1, we then see

$$\mathrm{Re}[\widetilde{h}(e^{it})] \leq \widetilde{\lambda}_a \left(\sum_{j=1}^{r_*} \frac{1}{\log^2(j+3)} \left(\frac{1}{j^2} - \frac{t^2}{2} \right) - \sum_{j=r_*+1}^{r} \frac{1}{\log^2(j+3)} \left(\frac{1}{j^2} - \frac{t^2}{2} \right) \right),$$

which is at most $1 - 10t^2$ by our choice of r_*. Combining with (5.20) and (5.19), we see

$$\begin{aligned} \left| \widetilde{h}(e^{it}) \right|^2 &= \left(\mathrm{Re}[\widetilde{h}(e^{it})] \right)^2 + \left(\mathrm{Im}[\widetilde{h}(e^{it})] \right)^2 \\ &\leq (1 - 10t^2)^2 + 4t^2 \\ &\leq 1, \end{aligned}$$

as desired. □

Chapter 6

On sumsets containing a perfect square

6.1 Summary

We show $A+B$ contains a perfect square if $A, B \subseteq \{1, \ldots, N\}$ have $|A|, |B| \geq (\frac{3}{8}+\epsilon)N$. The constant $\frac{3}{8}$ is optimal.

6.2 Introduction

Let A, B be subsets of the first N positive integers. What are the maximum possible sizes of A and B if $A + B$ does not contain a perfect square?

Let us first discuss the history of the related question of the largest size of a subset $A \subseteq \{1, \ldots, N\}$ with $A+A$ not containing a perfect square, originally raised by Erdős and Silverman [22, p. 87, 107]. Erdős initially conjectured that the answer is roughly $\frac{1}{3}N$, coming from

$$A := \{n \leq N : n \equiv 1 \text{ mod } 3\}.$$

However, Massias [43] noted that

$$A := \{n \leq N : n \text{ mod } 32 \in \{1, 5, 9, 13, 14, 17, 21, 25, 26, 29, 30\}\}$$

gives the larger size of roughly $\frac{11}{32}N$. The two mentioned sets A indeed have the property that $A + A$ does not contain a perfect square, since the sumset of $\{1\} \subseteq \mathbb{Z}/3\mathbb{Z}$ with itself does not contain a quadratic residue (in $\mathbb{Z}/3\mathbb{Z}$), and the sumset of $\{1, 5, 9, 13, 14, 17, 21, 25, 26, 29, 30\} \subseteq \mathbb{Z}/32\mathbb{Z}$ with itself avoids quadratic residues.

Given that these two examples come from "lifting up" a set $A \subseteq \mathbb{Z}/q\mathbb{Z}$ for some $q \in \mathbb{N}$, and that any perfect square must be a quadratic residue mod q, it is natural to first solve the "modular" version of the problem: for given $q \in \mathbb{N}$, what is the largest size of a set $A \subseteq \mathbb{Z}/q\mathbb{Z}$ such that $A + A$ does not contain a quadratic residue?

In 1982, Lagarias, Odlyzko, and Shearer [39] showed the answer is $\frac{11}{32}q$ (which is tight if $32 \mid q$). In 1983, they released a companion paper [38] proving that if $A \subseteq [N]$ has $|A| \geq 0.475N$ then $A + A$ contains a perfect square. Finally, in 2001, Khalfalah, Lodha, and Szemerédi [35] resolved the Erdős-Silverman problem, by showing that for all $\epsilon > 0$, if N is sufficiently large, then any $A \subseteq [N]$ with $A + A$ avoiding perfect squares must have $|A| \leq (\frac{11}{32} + \epsilon)N$.

In this chapter, we solve the aformentioned "bipartite" version of the Erdős-Silverman question. Our result is asymptotically optimal.

Theorem 6.2.1. *For any $\epsilon > 0$, if N is sufficiently large and $A, B \subseteq [N]$ have $|A|, |B| \geq (\frac{3}{8} + \epsilon)N$, then $A + B$ contains a perfect square.*

An example achieving roughly $\frac{3}{8}N$ is

$$A := \{n \leq N : n \bmod 8 \in \{0, 1, 5\}\}$$

$$B := \{n \leq N : n \bmod 8 \in \{2, 5, 6\}\},$$

which works since the $\mathbb{Z}/8\mathbb{Z}$-sumset $\{0, 1, 5\} + \{2, 5, 6\}$ avoids quadratic residues.

We prove Theorem 6.5.1 by first resolving the associated "modular" version of the problem. While the methods of [39], solving the modular problem for $A + A$, are highly graph-theoretic, our methods use Fourier analysis to reduce (in one direction) to solving some optimization problem in 48 variables. Interestingly, the paper [39] also involved solving some optimization problems, specifically various integer programs. It is plausible our methods could solve the modular $A + A$ problem, though the number of variables in the obtained optimization problem would be significantly too large.

We then obtain the result in the integers by basic Fourier-analytic arguments. While [35], solving the $A + A$ problem in the integers, introduced a novel "shifting method" and a sort of low-level strong arithmetic regularity lemma with tower-type bounds, our Fourier arguments amount to a rather basic arithmetic regularity lemma with only singly exponential bounds. In rough terms, we approximate the characteristic function of $A \subseteq [N]$ (and of B) by its best modulo Q weight function approximation on η^{-1} intervals each of length ηN, where η^{-1} and $\log Q$ are polynomials of ϵ^{-1}. Counting the number of perfect squares "in" the convolution of these weight functions essentially reduces to the modular problem. For details, see Section 6.5.

6.3 Notation

We use the standard $[N] := \{1, \ldots, N\}$ and $e(\theta) := e^{2\pi i\theta}$. Let $\frac{1}{\mathbb{N}} := \{\frac{1}{n} : n \geq 1\}$. Let $\mathbb{T} = \mathbb{R}/\mathbb{Z}$. For $f : [N] \to \mathbb{C}$, define $\widehat{f} : \mathbb{T} \to \mathbb{C}$ by

$$\widehat{f}(\theta) := \sum_{n \leq N} f(n)e(-n\theta).$$

For $f : \mathbb{Z}/q\mathbb{Z} \to \mathbb{C}$, define $\widehat{f} : \mathbb{Z}/q\mathbb{Z} \to \mathbb{C}$ by

$$\widehat{f}(r) := \frac{1}{q} \sum_{x \in \mathbb{Z}/q\mathbb{Z}} f(x)e\left(-\frac{rx}{q}\right).$$

Define the weighted indicator function of the quadratic residues $f_q : \mathbb{Z}/q\mathbb{Z} \to \mathbb{R}$ by

$$f_q(t) := |\{x \in \mathbb{Z}/q\mathbb{Z} : x^2 = t\}|.$$

For functions $f, g : \mathbb{Z}/q\mathbb{Z} \to \mathbb{C}$, define the convolution of f, g as

$$(f * g)(x) := \frac{1}{q} \sum_{a \in \mathbb{Z}/q\mathbb{Z}} f(a)g(x - a),$$

while for finitely supported functions $f, g : \mathbb{Z} \to \mathbb{C}$, we define the convolution as

$$(f * g)(x) := \sum_{n \in \mathbb{Z}} f(n)g(x - n).$$

6.4 The Modular Problem

In this section, we prove the following, a (doubly) weighted, quantitative version of the statement that $A+B$ contains a quadratic residue if $A, B \subseteq \mathbb{Z}/q\mathbb{Z}$ have $|A|, |B| > \frac{3}{8}q$.

Theorem 6.4.1. *For any $\epsilon > 0$ there is some $c(\epsilon) > 0$ so that for any $q \geq 1$, if $w_A, w_B : \mathbb{Z}/q\mathbb{Z} \to [0, 1]$ have $\sum_{t \in \mathbb{Z}/q\mathbb{Z}} w_A(t), \sum_{t \in \mathbb{Z}/q\mathbb{Z}} w_B(t) \geq (\frac{3}{8} + \epsilon)q$, then*

$$\sum_{t \in \mathbb{Z}/q\mathbb{Z}} (w_A * w_B)(t)f_q(t) \geq \frac{1}{\sqrt{5}}\epsilon q.$$

Our approach is Fourier-analytic. We start by noting the Fourier representation of this weighted count of quadratic residues "in" the convolution of w_A and w_B.

Lemma 6.4.2. *For any $w_A, w_B : \mathbb{Z}/q\mathbb{Z} \to \mathbb{R}$, we have*

$$\frac{1}{q} \sum_{t \in \mathbb{Z}/q\mathbb{Z}} (w_A * w_B)(t)f_q(t) = \sum_{m \in \mathbb{Z}/q\mathbb{Z}} \widehat{w_A}(m)\widehat{w_B}(m)\widehat{f_q}(-m).$$

Proof. The right hand side is, by definition, equal to

$$\sum_{m\in\mathbb{Z}/q\mathbb{Z}} \frac{1}{q^3} \sum_{x,y,z\in\mathbb{Z}/q\mathbb{Z}} w_A(x)w_B(y)f_q(z)e\left(\frac{m(z-x-y)}{q}\right).$$

Interchanging summations and using the orthogonality condition

$$\sum_{m\in\mathbb{Z}/q\mathbb{Z}} e\left(\frac{mr}{q}\right) = \begin{cases} q & \text{if } r\equiv 0 \bmod q \\ 0 & \text{if } r\not\equiv 0 \bmod q \end{cases}$$

finishes the proof. □

Remark. Let us take a moment to motivate the arguments to come. Suppose for now q is divisible by 8. We (a posteriori) expect $\sum_{t\in\mathbb{Z}/q\mathbb{Z}}(w_A * w_B)(t)f_q(t)$ to be minimized by weights w_A, w_B that are "lift-ups" of weights $\overline{w}_A, \overline{w}_B : \mathbb{Z}/8\mathbb{Z} \to [0,1]$ in the sense[1] $w_A(t) = \overline{w}_A(t \bmod 8)$ and $w_B(t) = \overline{w}_B(t \bmod 8)$. If w_A and w_B were indeed of this form, then, as one may easily check, we would have $\widehat{w}_A(m), \widehat{w}_B(m) = 0$ for each $m \in \mathbb{Z}/q\mathbb{Z}$ with $\frac{q}{\gcd(q,m)} \nmid 8$. Therefore, in our setting (in which w_A, w_B might not be exactly of that form), it's natural to separate[2],

$$\sum_{m\in\mathbb{Z}/q\mathbb{Z}} \widehat{w_A}(m)\widehat{w_B}(m)\widehat{f_q}(-m) = \sum_{q\mid 8m} \widehat{w_A}(m)\widehat{w_B}(m)\widehat{f_q}(-m) + \sum_{q\nmid 8m} \widehat{w_A}(m)\widehat{w_B}(m)\widehat{f_q}(-m).$$

The latter term we shall upper-bound in magnitude, using that $\widehat{f_q}(-m)$ is small for all m with $q \nmid 8m$ (this follows from quadratic Gauss sum bounds). And the first term actually turns out to be just the weighted count of mod 8 quadratic residues in the weighted sumset of the mod 8 projections of the weight functions w_A, w_B.

For technical reasons, we work mod 24 instead of mod 8.

To set some notation, if $q \in \mathbb{N}$ is a multiple of 24 and $w_A, w_B : \mathbb{Z}/q\mathbb{Z} \to [0,1]$ are two (weight) functions, we let $a, b : \mathbb{Z}/24\mathbb{Z} \to [0,1]$ denote the (mod 24)-projections of w_A and w_B:

$$a(k) := \frac{1}{q/24} \sum_{\substack{x\in\mathbb{Z}/q\mathbb{Z} \\ x\equiv k \bmod 24}} w_A(x) \tag{6.1}$$

$$b(k) := \frac{1}{q/24} \sum_{\substack{x\in\mathbb{Z}/q\mathbb{Z} \\ x\equiv k \bmod 24}} w_B(x). \tag{6.2}$$

[1] Note "mod 8" makes sense since $8 \mid q$.

[2] Note that $\frac{q}{\gcd(q,m)} \nmid 8$ is equivalent to $q \nmid 8m$.

Lemma 6.4.3. *Let $q \in \mathbb{N}$ be a multiple of* 24*. Let $w_A, w_B : \mathbb{Z}/q\mathbb{Z} \to [0,1]$ be two (weight) functions, and let $a, b : \mathbb{Z}/24\mathbb{Z} \to [0,1]$ be the mod* 24*-projections of w_A, w_B, as in* (6.1),(6.2)*. Then one has*

$$\sum_{\substack{m \in \mathbb{Z}/q\mathbb{Z} \\ q \mid 24m}} \widehat{w_A}(m)\widehat{w_B}(m)\widehat{f_q}(-m) = \frac{1}{24} \sum_{t \in \mathbb{Z}/24\mathbb{Z}} (a * b)(t) f_{24}(t).$$

Proof. Noting $q \mid 24m$ if and only if $m = \frac{rq}{24}$, we may write the LHS as

$$\sum_{r=0}^{23} \frac{1}{q^3} \sum_{x,y,z \in \mathbb{Z}/q\mathbb{Z}} w_A(x) w_B(y) f_q(z) e\left(\frac{rq}{24} \, \frac{z-x-y}{q}\right),$$

which by orthogonality (mod 24) is equal to

$$\frac{24}{q^3} \sum_{\substack{x,y,z \in \mathbb{Z}/q\mathbb{Z} \\ x+y \equiv z \bmod 24}} w_A(x) w_B(y) f_q(z).$$

Splitting into cases mod 24, we may write the above as

$$\frac{24}{q^3} \sum_{i,j \in \mathbb{Z}/24\mathbb{Z}} \Bigg(\sum_{\substack{x \in \mathbb{Z}/q\mathbb{Z} \\ x \equiv i \bmod 24}} w_A(x) \Bigg) \Bigg(\sum_{\substack{y \in \mathbb{Z}/q\mathbb{Z} \\ y \equiv j \bmod 24}} w_B(y) \Bigg) \Bigg(\sum_{\substack{z \in \mathbb{Z}/q\mathbb{Z} \\ z \equiv i+j \bmod 24}} f_q(z) \Bigg). \tag{6.3}$$

Noting

$$\sum_{\substack{z \in \mathbb{Z}/q\mathbb{Z} \\ z \equiv i+j \bmod 24}} f_q(z) = \sum_{\substack{z \in \mathbb{Z}/q\mathbb{Z} \\ z \equiv i+j \bmod 24}} \sum_{v \in \mathbb{Z}/q\mathbb{Z}} 1_{v^2 \equiv z \bmod q} = \sum_{v \in \mathbb{Z}/q\mathbb{Z}} 1_{v^2 \equiv i+j \bmod 24} = \frac{q}{24} f_{24}(i+j),$$

and using the definitions of a, b, we may write (6.3) as

$$\frac{1}{24^2} \sum_{i,j \in \mathbb{Z}/24\mathbb{Z}} a(i) b(j) f_{24}(i+j) = \frac{1}{24} \sum_{t \in \mathbb{Z}/24\mathbb{Z}} (a * b)(t) f_{24}(t),$$

as desired. □

We now go on to handle the other Fourier term, $\sum_{q \nmid 24m} \widehat{w}_A(m)\widehat{w}_B(m)\widehat{f_q}(-m)$.

Lemma 6.4.4. *Let $q \in \mathbb{N}$ be a multiple of* 24*. Then for any $m \in \mathbb{Z}$ with $q \nmid 24m$, one has*

$$\left|\widehat{f_q}(-m)\right| \leq \frac{1}{\sqrt{5}}.$$

Proof. By definition,

$$\widehat{f}_q(-m) = \frac{1}{q}\sum_{t\in\mathbb{Z}/q\mathbb{Z}}\left(\sum_{x\in\mathbb{Z}/q\mathbb{Z}} 1_{x^2\equiv t}\right) e(\frac{mt}{q}) = \frac{1}{q}\sum_{x\in\mathbb{Z}/q\mathbb{Z}} e\left(\frac{mx^2}{q}\right) = \frac{1}{q/g}\sum_{x\in\mathbb{Z}/\frac{q}{g}\mathbb{Z}} e\left(\frac{\frac{m}{g}x^2}{q/g}\right),$$

where $g := \gcd(m,q)$. Thus, by standard quadratic Gauss sum estimates (e.g., [33]),

$$\left|\widehat{f}_q(-m)\right| \le \begin{cases} \sqrt{\frac{1}{q/g}} & \text{if } q/g \in \{1,3\} \bmod 4 \\ \sqrt{\frac{2}{q/g}} & \text{if } q/g \equiv 0 \bmod 4 \\ 0 & \text{if } q/g \equiv 2 \bmod 4. \end{cases}$$

Now, $q \nmid 24m$ implies $\frac{q}{g} \nmid 24$. This implies, firstly, that $\frac{q}{g} \ge 5$, giving $\sqrt{\frac{1}{q/g}} \le \frac{1}{\sqrt{5}}$, and, secondly, that if $\frac{q}{g} \equiv 0 \bmod 4$, then $\frac{q}{g} \ge 16$, giving $\sqrt{\frac{2}{q/g}} \le \frac{1}{\sqrt{8}} \le \frac{1}{\sqrt{5}}$. □

Lemma 6.4.5. *Let $q \in \mathbb{N}$ be a multiple of 24. Let $w_A, w_B : \mathbb{Z}/q\mathbb{Z} \to [0,1]$ be two (weight) functions, and let $a, b : \mathbb{Z}/24\mathbb{Z} \to [0,1]$ be the projections of w_A, w_B mod 24 as in Lemma 6.4.3. Then,*

$$\left|\sum_{\substack{m\in\mathbb{Z}/q\mathbb{Z}\\ q\nmid 24m}} \widehat{w_A}(m)\widehat{w_B}(m)\widehat{f}_q(-m)\right| \le \frac{1}{24\sqrt{5}}\left(\sum_{k\in\mathbb{Z}/24\mathbb{Z}} a(k)-a(k)^2\right)^{1/2}\left(\sum_{k\in\mathbb{Z}/24\mathbb{Z}} b(k)-b(k)^2\right)^{1/2}.$$

Proof. By Lemma 6.4.4 and Cauchy-Schwarz, we have

$$\begin{aligned}\left|\sum_{\substack{m\in\mathbb{Z}/q\mathbb{Z}\\ q\nmid 24m}} \widehat{w_A}(m)\widehat{w_B}(m)\widehat{f}_q(-m)\right| &\le \left(\sup_{\substack{m\in\mathbb{Z}/q\mathbb{Z}\\ q\nmid 24m}} |\widehat{f}_q(-m)|\right)\left(\sum_{\substack{m\in\mathbb{Z}/q\mathbb{Z}\\ q\nmid 24m}} |\widehat{w_A}(m)|\,|\widehat{w_B}(m)|\right) \\ &\le \frac{1}{\sqrt{5}}\sqrt{\sum_{\substack{m\in\mathbb{Z}/q\mathbb{Z}\\ q\nmid 24m}} |\widehat{w_A}(m)|^2}\sqrt{\sum_{\substack{m\in\mathbb{Z}/q\mathbb{Z}\\ q\nmid 24m}} |\widehat{w_B}(m)|^2}.\end{aligned}$$

The following two (in)equalities (and their analogues for B) finish the proof:

$$\begin{aligned}\sum_{\substack{m\in\mathbb{Z}/q\mathbb{Z}\\ q\mid 24m}} |\widehat{w_A}(m)|^2 &= \sum_{r=0}^{23}\frac{1}{q^2}\sum_{x,y\in\mathbb{Z}/q\mathbb{Z}} w_A(x)w_A(y)e\left(\frac{r(x-y)}{24}\right) \\ &= \frac{24}{q^2}\sum_{i\in\mathbb{Z}/24\mathbb{Z}}\left(\sum_{\substack{x\in\mathbb{Z}/q\mathbb{Z}\\ x\equiv i \bmod 24}} w_A(x)\right)^2 = \frac{1}{24}\sum_{k\in\mathbb{Z}/24\mathbb{Z}} a(k)^2.\end{aligned}$$

$$\begin{aligned}\sum_{m\in\mathbb{Z}/q\mathbb{Z}} |\widehat{w_A}(m)|^2 &= \sum_{m\in\mathbb{Z}/q\mathbb{Z}} \frac{1}{q^2} \sum_{x,y\in\mathbb{Z}/q\mathbb{Z}} w_A(x)w_A(y)e\left(\frac{m(x-y)}{q}\right) \\ &= \frac{1}{q}\sum_{x\in\mathbb{Z}/q\mathbb{Z}} w_A(x)^2 \le \frac{1}{q}\sum_{x\in\mathbb{Z}/q\mathbb{Z}} w_A(x) = \frac{1}{24}\sum_{k\in\mathbb{Z}/24\mathbb{Z}} a(k).\end{aligned}$$

□

Combining Lemmas 6.4.2, 6.4.3, and 6.4.5 (and multiplying through by 24) yields

$$\begin{aligned}\frac{24}{q}\sum_{t\in\mathbb{Z}/q\mathbb{Z}} (w_A * w_B)(t)f_q(t) \ge \sum_{t\in\mathbb{Z}/24\mathbb{Z}} (a*b)(t)f_{24}(t) \qquad (6.4)\\ -\frac{1}{\sqrt{5}}\sqrt{\sum_{k\in\mathbb{Z}/24\mathbb{Z}} a(k)-a(k)^2}\sqrt{\sum_{k\in\mathbb{Z}/24\mathbb{Z}} b(k)-b(k)^2}.\end{aligned}$$

Note that $a(k)\in[0,1]$ for each k and that

$$\sum_{k\in\mathbb{Z}/24\mathbb{Z}} a(k) = 24\cdot\frac{1}{q}\sum_{x\in\mathbb{Z}/q\mathbb{Z}} w_A(x),$$

implying $\sum_{k\in\mathbb{Z}/24\mathbb{Z}} a(k) \ge 9+24\epsilon$ if $\sum_{x\in\mathbb{Z}/q\mathbb{Z}} w_A(x) \ge (\frac{3}{8}+\epsilon)q$. We prove the following proposition in Section 6.6. We assume it to be true for the rest of this section. In it, we use the notation $a(i) := a_i, b(i) := b_i$. We emphasize that it is "merely" a (quadratic) optimization problem in 48 variables.

Proposition 6.4.6. *For any $\epsilon > 0$, the following holds. For all $a_0,\dots,a_{23},b_0,\dots,b_{23} \in [0,1]$ with $\sum_{i=0}^{23} a_i \ge 9+\epsilon, \sum_{i=0}^{23} b_i \ge 9+\epsilon$, one has*

$$\sum_{t\in\mathbb{Z}/24\mathbb{Z}} (a*b)(t)f_{24}(t) \ge \frac{1}{\sqrt{5}}\epsilon + \frac{1}{\sqrt{5}}\sqrt{\sum_i a_i - \sum_i a_i^2}\sqrt{\sum_i b_i - \sum_i b_i^2}.$$

Proof of Theorem 6.4.1 assuming Proposition 6.4.6. If $24 \mid q$, then Theorem 6.4.1 follows immediately from (6.4) and Proposition 6.4.6. Otherwise, we use a simple "lift-up" argument to reduce to the case $q \mid 24$. Define $\widetilde{w}_A, \widetilde{w}_B : \mathbb{Z}/24q\mathbb{Z} \to [0,1]$ by $\widetilde{w}_A(x) := \frac{1}{24}\sum_{\substack{y\in\mathbb{Z}/24\mathbb{Z}\\ y\equiv x \bmod q}} w_A(y), \widetilde{w}_B(x) := \frac{1}{24}\sum_{\substack{y\in\mathbb{Z}/24\mathbb{Z}\\ y\equiv x \bmod q}} w_B(y)$. Then

$$\frac{1}{q}\sum_{t\in\mathbb{Z}/q\mathbb{Z}} (w_A * w_B)(t)f_q(t) = \frac{1}{24q}\sum_{t\in\mathbb{Z}/24q\mathbb{Z}} (\widetilde{w}_A * \widetilde{w}_B)(t)f_{24q}(t)$$

and

$$\frac{1}{24q}\sum_{x\in\mathbb{Z}/24q\mathbb{Z}} \widetilde{w}_A(x) = \frac{1}{q}\sum_{x\in\mathbb{Z}/q\mathbb{Z}} w_A(x)$$

$$\frac{1}{24q}\sum_{x\in\mathbb{Z}/24q\mathbb{Z}}\widetilde{w}_B(x) = \frac{1}{q}\sum_{x\in\mathbb{Z}/q\mathbb{Z}} w_B(x).$$

□

6.5 Converting to Integers

In this section, we "boost" the solution to the modular problem (Theorem 6.4.1) to the integers to establish our main theorem (Theorem 6.5.1). For subsets $A, B \subseteq [N]$ with $|A|, |B| \geq (\frac{3}{8}+\epsilon)N$ we shall, as in the modular problem, look at the number of squares in the weighted sumset of A and B:

$$\sum_{n\geq 1}(1_A * 1_B)(n)1_S(n),$$

where $S \subseteq \mathbb{N}$ is the set of perfect squares,

$$S := \{m^2 : m \in \mathbb{N}\}.$$

Our approach is inspired by the arithmetic regularity lemma (see, e.g., [20, 29]), though a much lower-tech version suffices for our purposes; the dependence on the relevant parameters will be singly-exponential rather than tower-type.

Definition 6.5.1. Fix (parameters) $Q \in \mathbb{N}$ and $\eta \in \frac{1}{\mathbb{N}}$. For $k \in \{0, 1, \ldots, \eta^{-1}-1\}$, let

$$I_{\eta,k} = \Big(k\eta N, (k+1)\eta N\Big] \cap \mathbb{N}.$$

For $N \in \mathbb{N}$ (large) and $A \subseteq [N]$, define[3] the function $w^A_{Q;\eta,k} : I_{\eta,k} \to [0,1]$ by

$$w^A_{Q;\eta,k}(n) := \frac{\#\{m \in I_{\eta,k} : m \in A \text{ and } m \equiv n \bmod Q\}}{\#\{m \in I_{\eta,k} : m \equiv n \bmod Q\}}.$$

Finally, define the function $w^A_{Q;\eta} : \mathbb{N} \to [0,1]$ by

$$w^A_{Q;\eta} := \sum_{k=0}^{\eta^{-1}-1} w^A_{Q;\eta,k}1_{I_{\eta,k}}.$$

Remark. One should think of the function $w^A_{Q;\eta,k}$ as the best mod Q approximation to A, or as a "smoothed out" version of A modulo Q, on $I_{\eta,k}$. Indeed, for $n \in I_{\eta,k}$,

[3] Extend (the domain of) $w^A_{Q;\eta,k}$ to $\mathbb{N}$ by setting $w^A_{Q;\eta,k} = 0$ outside $I_{\eta,k}$.

the function $w^A_{Q;\eta,k}(n)$ just depends on the residue of n modulo Q, and, immediately from the definition, for any $r \in \{0, \dots, Q-1\}$, one has

$$\sum_{\substack{n \in I_{\eta,k} \\ n \equiv r \bmod Q}} w^A_{Q;\eta,k}(n) = \sum_{\substack{n \in I_{\eta,k} \\ n \equiv r \bmod Q}} 1_A(n). \tag{6.5}$$

The use of $w^A_{Q;\eta}$ comes from the fact that its Fourier transform models that of A nearly perfectly on rationals with denominator dividing Q. As long as Q is sufficiently composite (which we will choose it to be), we don't need to care much about other rationals, since the Fourier transform of the indicator function of the squares will be sufficiently small there.

For the following lemma, fix $Q, N \in \mathbb{N}, \eta \in \frac{1}{\mathbb{N}}$, and $A \subseteq [N]$.

Definition 6.5.2. Define the *balanced function* $f^A_{Q;\eta} : \mathbb{N} \to \mathbb{R}$ by $f^A_{Q;\eta} := 1_A - w^A_{Q;\eta}$.

Lemma 6.5.3. *Take some $a, q \in \mathbb{N}$ with $q \mid Q$. Then, for any $\beta \in \mathbb{R}$, it holds that*

$$\left| \widehat{f^A_{Q;\eta}} \left(\frac{a}{q} + \beta \right) \right| \leq 2|\beta|\eta N^2.$$

Proof. For $k \in \{0, \dots, \eta^{-1} - 1\}$, define $f^A_{Q;\eta,k} := f^A_{Q;\eta} 1_{I_{\eta,k}} = 1_{I_{\eta,k}} 1_A - w^A_{Q;\eta,k}$ so that

$$f^A_{Q;\eta} = \sum_{k=0}^{\eta^{-1}-1} f^A_{Q;\eta,k}. \tag{6.6}$$

Fix $a, q \in \mathbb{N}$ with $q \mid Q$, and fix $\beta \in \mathbb{R}$. By (6.6), linearity of the fourier transform, and the triangle inequality, to prove Lemma 6.5.3 it suffices to show

$$\left| \widehat{f^A_{Q;\eta,k}} (\frac{a}{q} + \beta) \right| \leq 2|\beta|\eta N |I_{\eta,k}|$$

for each $k \in \{0, \dots, \eta^{-1} - 1\}$. So fix some such k. By definition,

$$\widehat{f^A_{Q;\eta,k}} (\frac{a}{q} + \beta) = \sum_{n \in I_{\eta,k}} 1_A(n) e\left((\frac{a}{q} + \beta) n \right) - \sum_{n \in I_{\eta,k}} w^A_{Q;k}(n) e\left((\frac{a}{q} + \beta) n \right). \tag{6.7}$$

Letting $L = \lfloor k\eta N \rfloor + 1$ denote the left endpoint of $I_{\eta,k}$, we trivially from (6.7) have

$$\left| \widehat{f^A_{Q;\eta,k}} (\frac{a}{q} + \beta) \right| = \left| \sum_{n \in I_{\eta,k}} 1_A(n) e\left((\frac{a}{q} + \beta)(n - L) \right) - \sum_{n \in I_{\eta,k}} w^A_{Q;\eta,k}(n) e\left((\frac{a}{q} + \beta)(n - L) \right) \right|.$$

The reason for shifting the phase by L is that if we now use

$$\sum_{n\in I_{\eta,k}} 1_A(n)e\left(\frac{a(n-L)}{q}\right) - \sum_{n\in I_{\eta,k}} w^A_{Q;\eta,k}(n)e\left(\frac{a(n-L)}{q}\right) = 0$$

(which follows from (6.5) and that $q \mid Q$) to write

$$\begin{aligned}\left|\widehat{f^A_{Q;\eta,k}}(\frac{a}{q}+\beta)\right| = \Bigg| &\sum_{n\in I_{\eta,k}} 1_A(n)\left[e\left((\frac{a}{q}+\beta)(n-L)\right) - e\left(\frac{a(n-L)}{q}\right)\right] \\ &- \sum_{n\in I_{\eta,k}} w^A_{Q;\eta,k}(n)\left[e\left((\frac{a}{q}+\beta)(n-L)\right) - e\left(\frac{a(n-L)}{q}\right)\right]\Bigg|,\end{aligned}$$

then the trivial $|e(x)-e(y)| \leq |x-y|$ is strong enough to give the sufficient bound

$$\begin{aligned}\left|\widehat{f^A_{Q;\eta,k}}(\frac{a}{q}+\beta)\right| &\leq \sum_{n\in I_{\eta,k}} 1_A(n)|\beta|(n-L) + \sum_{n\in I_{\eta,k}} |w^A_{Q;\eta,k}(n)|\,|\beta|\,(n-L) \\ &\leq 2|\beta|\eta N|I_{\eta,k}|,\end{aligned}$$

the last inequality using that $n - L \leq \eta N$ for each $n \in I_{\eta,k}$. □

Remark. The plan to prove Theorem 6.5.1 is to decompose

$$1_A * 1_B = w^A_{Q;\eta} * w^B_{Q;\eta} + f^A_{Q;\eta} * w^B_{Q;\eta} + w^A_{Q;\eta} * f^B_{Q;\eta} + f^A_{Q;\eta} * f^B_{Q;\eta}$$

and use Lemma 6.5.3 to argue that the "number" of squares "in" $1_A * 1_B$ is approximately the same as that in $w^A_{Q;\eta} * w^B_{Q;\eta}$. The latter, involving the convolution of two functions constant on residues modulo Q, is more easily calculable and comes down to the weighted number of mod Q quadratic residues in the convolution of the natural mod Q projections of $w^A_{Q;\eta}, w^B_{Q;\eta}$. The following (with Lemma 6.5.3) will be used to prove the validity of the approximation.

Proposition 6.5.4. *Let $f, g : [N] \to [-1,1]$ be (1-bounded) functions. Suppose $\delta > 0$ is such that $\left|\widehat{f}(\frac{a}{q}+\beta)\right| \leq \delta|\beta|N^2$ for each $a, q \leq \lambda^{-2}$ and*[4] *$\beta \in \mathbb{R}$. Then we have*

$$\left|\sum_{n\geq 1}(f*g)(n)1_S(n)\right| \leq 10(\delta\lambda^{-8}+\lambda)N^{3/2}.$$

In order to prove Proposition 6.5.4, we import the needed "minor arc" estimate from [42]:

[4] We will only need the condition for $|\beta| \leq \frac{\lambda^{-2}}{2N}$.

Lemma 6.5.5 ([42], Proposition 1)**.** *For any $\lambda > 0$, if $N \in \mathbb{N}$ is sufficiently large and $\theta \in \mathbb{T}$ is such that $|\theta - \frac{a}{q}| > \frac{\lambda^{-2}}{N}$ for each $a, q \leq \lambda^{-2}$, then $|\widehat{1_{S_N}}(\theta)| \leq 5\lambda N^{1/2}$.*

Proof of Proposition 6.5.4. We may replace S by $S_{2N} := \{m^2 : m \in \mathbb{N}, m^2 \leq 2N\}$ and write

$$\sum_{n\geq 1}(f * g)(n)1_{S_{2N}}(n) = \int_{\mathbb{T}} \widehat{f}(\theta)\widehat{g}(\theta)\widehat{1_{S_{2N}}}(-\theta)d\theta. \tag{6.8}$$

This lemma together with Cauchy-Schwarz and Plancherel immediately gives

$$\begin{aligned}\left|\int_{\mathfrak{m}} \widehat{f}(\theta)\widehat{g}(\theta)\widehat{1_{S_{2N}}}(-\theta)d\theta\right| &\leq 5\lambda\sqrt{2N}\int_{\mathfrak{m}} |\widehat{f}(\theta)||\widehat{g}(\theta)|d\theta \\ &\leq 10\lambda N^{1/2}\left(\int_{\mathbb{T}}|\widehat{f}(\theta)|^2 d\theta\right)^{1/2}\left(\int_{\mathbb{T}}|\widehat{g}(\theta)|^2 d\theta\right)^{1/2} \\ &= 10\lambda N^{1/2}\left(\sum_{n\leq N} f(n)^2\right)^{1/2}\left(\sum_{n\leq N} g(n)^2\right)^{1/2} \\ &\leq 10\lambda N^{3/2},\end{aligned}$$

where $\mathfrak{m}$ is defined so that

$$\mathbb{T}\setminus\mathfrak{m} := \bigcup_{q=1}^{\lambda^{-2}} \bigcup_{\substack{1\leq a\leq q \\ (a,q)=1}} \left\{\theta \in \mathbb{T} : \left|\theta - \frac{a}{q}\right| \leq \frac{\lambda^{-2}}{2N}\right\}.$$

Letting $\beta_* = \frac{\lambda^{-2}}{2N}$ for notational ease, we handle the "major arc" as follows:

$$\begin{aligned}\left|\int_{\mathbb{T}\setminus\mathfrak{m}} \widehat{f}(\theta)\widehat{g}(\theta)\widehat{1_{S_{2N}}}(-\theta)d\theta\right| &\leq \sum_{q=1}^{\lambda^{-2}}\sum_{1\leq a\leq q}\left|\int_{\frac{a}{q}-\beta_*}^{\frac{a}{q}+\beta_*} \widehat{f}(\theta)\widehat{g}(\theta)\widehat{1_{S_{2N}}}(-\theta)d\theta\right| \\ &\leq \sum_{q=1}^{\lambda^{-2}}\sum_{1\leq a\leq q}\int_{-\beta_*}^{\beta_*}\left(\delta|\beta|N^2\right)(N)\left(\sqrt{2N}\right)d\beta \\ &\leq \sqrt{2}\delta N^{7/2}\sum_{q=1}^{\lambda^{-2}}\sum_{1\leq a\leq q} 2\beta_*^2 \\ &\leq 10\delta\lambda^{-8}N^{3/2}.\end{aligned}$$

(The bound "10" here is loose and used for simplicity.) We're done by (6.8). □

To complete the plan outlined in Remark 6.5, we need to argue that $w^A_{Q;\eta} * w^B_{Q;\eta}$ "contains" many squares. We start by focusing on particular intervals. We abstract out from our exact the situation the relevant property of $w^A_{Q;\eta,k}$ and $w^B_{Q;\eta,k}$.

Proposition 6.5.6. *Fix $\epsilon > 0$ and $Q \geq 1$. Let functions $\overline{w}_1, \overline{w}_2 : \mathbb{Z}/Q\mathbb{Z} \to [0,1]$ satisfy*

$$\sum_{t \in \mathbb{Z}/Q\mathbb{Z}} \overline{w}_i(t) \geq \left(\frac{3}{8} + \epsilon\right) Q$$

for $i = 1, 2$. For large $M \in \mathbb{N}$ and intervals $I_i = [k_i M, (k_i+1)M]$, $i = 1, 2$, define

$$w_i(n) := 1_{I_i}(n)\,\overline{w}_i(n \bmod Q)$$

for $i = 1, 2$. Then we have the lower bound

$$\sum_{n \geq 1} (w_1 * w_2)(n) 1_S(n) \geq \frac{1}{200} c(\epsilon) \frac{M^{3/2}}{\sqrt{k_1 + k_2}},$$

where $c(\epsilon) > 0$ is the constant guaranteed by Theorem 6.4.1.

Proof. Let

$$J = \left[(k_1 + k_2 + 1)M - \frac{1}{10}M, (k_1 + k_2 + 1)M + \frac{1}{10}M\right]$$

so that for any $n \in J$ and $a \in \{0, \ldots, Q-1\}$, it holds that

$$\#\left\{m \in I_1 : m \equiv a \bmod Q \text{ and } n - m \in I_2\right\} \geq \frac{1}{10}\frac{M}{Q}$$

(provided M is large enough). Therefore,

$$\begin{aligned}
\sum_{n \geq 1} (w_1 * w_2)(n) 1_S(n) &\geq \sum_{n \in J} \sum_{\substack{m \in I_1 \\ n - m \in I_2}} w_1(m) w_2(n-m) 1_S(n) \\
&= \sum_{\substack{n \in J \\ n \in S}} \sum_{a=0}^{Q-1} \overline{w}_1(a)\overline{w}_2(n - a \bmod Q) \sum_{\substack{m \equiv a \bmod Q \\ m \in I_1 \\ n - m \in I_2}} 1 \\
&\geq \frac{M}{10Q} Q \sum_{\substack{n \in J \\ n \in S}} (\overline{w}_1 * \overline{w}_2)(n \bmod Q) \\
&= \frac{M}{10} \sum_{t=0}^{Q-1} (\overline{w}_1 * \overline{w}_2)(t) \cdot \#\{m \in \mathbb{N} : m^2 \in J,\, m^2 \equiv t \bmod Q\}.
\end{aligned}$$

Note that, for $\overline{J} := \{m \in \mathbb{N} : m^2 \in J\}$, we have as $M \to \infty$ that

$$\#\{m \in \mathbb{N} : m^2 \in J,\, m^2 \equiv t \bmod Q\} = (1 + o(1))\, f_Q(t) \frac{|\overline{J}|}{Q}.$$

We lower-bound

$$\begin{aligned}|\overline{J}| &\geq \frac{1}{2}\left(\sqrt{(k_1+k_2+1)M+\frac{1}{10}M}-\sqrt{(k_1+k_2+1)M-\frac{1}{10}M}\right)\\ &= \frac{1}{2}\frac{\frac{2}{10}M}{\sqrt{(k_1+k_2+1)M+\frac{1}{10}M}+\sqrt{(k_1+k_2+1)M-\frac{1}{10}M}}\\ &\geq \frac{1}{2}\frac{\frac{1}{10}M}{\sqrt{(k_1+k_2)M}}.\end{aligned}$$

Combining everything, we obtain

$$\sum_{n\geq 1}(w_1 * w_2)(n)1_S(n) \geq \frac{M}{10Q}\frac{\sqrt{M}}{20\sqrt{k_1+k_2}}\sum_{t=0}^{Q-1}(\overline{w}_1 * \overline{w}_2)(t)f_Q(t).$$

By the assumptions of the current theorem, Theorem 6.4.1 finishes the proof. □

Back to our specific setting, we can now handle $w^A_{Q;\eta} * w^B_{Q;\eta}$.

Proposition 6.5.7. *Fix $\epsilon > 0, Q \in \mathbb{N}$, and $\eta \in \frac{1}{\mathbb{N}}$. Then for all large $N \in \mathbb{N}$ and any $A, B \subseteq [N]$ with $|A|, |B| \geq (\frac{3}{8}+\epsilon)N$, we have*

$$\sum_{n\geq 1}(w^A_{Q;\eta} * w^B_{Q;\eta})(n)1_S(n) \geq \frac{\epsilon^2}{5000}c\left(\frac{\epsilon}{2}\right)N^{3/2},$$

where $c(\epsilon) > 0$ is the constant guaranteed by Theorem 6.4.1.

Proof. It is easy to see that $|A| \geq (\frac{3}{8}+\epsilon)N$ implies there are at least $\frac{\epsilon}{3}\eta^{-1}$ values of $k \in \{0, \ldots, \eta^{-1}-1\}$ with $|A \cap I_{\eta,k}| \geq (\frac{3}{8}+\frac{3\epsilon}{4})|I_{\eta,k}|$. Therefore, by taking N large enough, if we let[5]

$$J^A := \left\{k \in \{0, \ldots, \eta^{-1}-1\} : \sum_{n=\lfloor k\eta N\rfloor+1}^{\lfloor k\eta N\rfloor+Q} w^A_{Q;\eta,k}(n) \geq \left(\frac{3}{8}+\frac{\epsilon}{2}\right)Q\right\},$$

then we have $|J^A| \geq \frac{\epsilon}{4}\eta^{-1}$. Defining J^B in the analogous way, we by symmetry have

[5] The choice of summing n over $[\lfloor k\eta N\rfloor+1, \lfloor k\eta N\rfloor+Q]$ is arbitrary; any Q numbers in $I_{\eta,k}$, all distinct modulo Q, would of course be equivalent.

$|J^B| \geq \frac{\epsilon}{4}\eta^{-1}$. The point is that Proposition 6.5.6 (with $M = \eta N$) then lets us bound

$$\begin{aligned}\sum_{n\geq 1}(w^A_{Q;\eta} * w^B_{Q;\eta})(n)1_S(n) &= \sum_{k_1,k_2=0}^{\eta^{-1}-1}\sum_{n\geq 1}(w^A_{Q;\eta,k_1} * w^B_{Q;\eta,k_2})(n)1_S(n)\\ &\geq \sum_{\substack{k_1\in J^A\\ k_2\in J^B}}\sum_{n\geq 1}(w^A_{Q;\eta,k_1} * w^B_{Q;\eta,k_2})(n)1_S(n)\\ &\geq \sum_{\substack{k_1\in J^A\\ k_2\in J^B}}\frac{1}{200}c\left(\frac{\epsilon}{2}\right)\frac{(\eta N)^{3/2}}{\sqrt{k_1+k_2}}\\ &\geq \frac{1}{200}c\left(\frac{\epsilon}{2}\right)\frac{(\eta N)^{3/2}}{\sqrt{2\eta^{-1}}}|J^A||J^B|.\end{aligned}$$

The proof is complete by inserting the lower bounds $|J^A|, |J^B| \geq \frac{\epsilon}{4}\eta^{-1}$. □

We now put everything together to obtain (a more quantitative version of) our main theorem.

Theorem 6.5.1. *For any $\epsilon > 0$, if N is sufficiently large and $A, B \subseteq [N]$ have $|A|, |B| \geq (\frac{3}{8}+\epsilon)N$, then $A + B$ contains a perfect square. In fact, we have the quantitative*

$$\#\{(a,b) \in A \times B : a + b \in S\} \geq 10^{-6}\epsilon^3 N^{3/2}.$$

Proof. Let $\eta \in \frac{1}{\mathbb{N}}, \overline{Q} \in \mathbb{N}$ be parameters (based on ϵ) to be determined, and set $Q := \operatorname{lcm}(1, \ldots, \overline{Q})$. Take N sufficiently large and $A, B \subseteq [N]$ with $|A|, |B| \geq (\frac{3}{8}+\epsilon)N$. As remarked earlier, we decompose

$$1_A * 1_B = w^A_{Q;\eta} * w^B_{Q;\eta} + f^A_{Q;\eta} * w^B_{Q;\eta} + w^A_{Q;\eta} * f^B_{Q;\eta} + f^A_{Q;\eta} * f^B_{Q;\eta}.$$

Proposition 6.5.7 gives

$$\sum_{n\geq 1}(w^A_{Q;\eta} * w^B_{Q;\eta})(n)1_S(n) \geq \frac{\epsilon^2}{5000}c\left(\frac{\epsilon}{2}\right)N^{3/2},$$

and Proposition 6.5.4 together with Lemma 6.5.3 gives

$$\left|\sum_{n\geq 1}(f^A_{Q;\eta} * w^B_{Q;\eta})(n)1_S(n)\right| \leq 10\left(2\eta\overline{Q}^4 + \overline{Q}^{-1/2}\right)N^{3/2},$$

and the same bound for the analogous inequalities involving $w^A_{Q;\eta} * f^B_{Q;\eta}$ and $f^A_{Q;\eta} * f^B_{Q;\eta}$. Therefore,

$$\sum_{n\geq 1}(1_A * 1_B)(n)1_S(n) \geq \frac{\epsilon^2}{5000}c\left(\frac{\epsilon}{2}\right)N^{3/2} - 30\left(2\eta\overline{Q}^4 + \overline{Q}^{-1/2}\right)N^{3/2}.$$

Setting $\eta = \overline{Q}^{-9/2}$ and using $c(\epsilon) \geq \epsilon/3$, we obtain

$$\sum_{n\geq 1}(1_A * 1_B)(n)1_S(n) \geq \left(\frac{\epsilon^3}{30000} - 90\overline{Q}^{-1/2}\right) N^{3/2}.$$

Choosing $\overline{Q}$ a perfect square (merely so that $\eta \in \frac{1}{\mathbb{N}}$) with $\overline{Q}^{-1/2} \leq 10^{-7}\epsilon^3$, say, finishes the proof. □

6.6 Solving the Optimization Problem

We finish the chapter by proving the inequality that Theorem 6.4.1 relied upon. It could be verified directly by a computer but would take quite a bit of time.

For $a_0, \ldots, a_{23} \in [0,1]$, we let $a : \mathbb{Z}/24\mathbb{Z} \to [0,1]$ be given by $a(i) = a_i$. Recall, for $a, b \in \mathbb{Z}/24\mathbb{Z}$ and $t \in \mathbb{Z}/24\mathbb{Z}$, we define

$$(a * b)(t) := \frac{1}{24}\sum_{i\in\mathbb{Z}/24\mathbb{Z}} a(i)b(t-i)$$

$$f_{24}(t) := \#\{j \in \mathbb{Z}/24\mathbb{Z} : j^2 \equiv t \bmod 24\}.$$

In this section, we prove the following, stated previously in Section 6.4.

Theorem 6.6.1. *For any $\epsilon > 0$, there is some $c'(\epsilon) > 0$ so that the following holds. For all $a_0, \ldots, a_{23}, b_0, \ldots, b_{23} \in [0,1]$ with $\sum_{i=0}^{23} a_i \geq 9 + \epsilon, \sum_{i=0}^{23} b_i \geq 9 + \epsilon$, we have*

$$\sum_{t\in\mathbb{Z}/24\mathbb{Z}} (a * b)(t) f_{24}(t) \geq c'(\epsilon) + \frac{1}{\sqrt{5}}\sqrt{\sum_i a_i - \sum_i a_i^2}\sqrt{\sum_i b_i - \sum_i b_i^2}.$$

In fact, one can take $c'(\epsilon) = \frac{1}{\sqrt{5}}\epsilon$.

The proof, with $c'(\epsilon) = \frac{1}{\sqrt{5}}\epsilon$, will follow from the proof of the "$\epsilon = 0$" case, in which we also identify the extremizers. We say a is a *lift-up* of a subset A of $\mathbb{Z}/8\mathbb{Z}$ if: $a_i = 1$ if and only if $i \bmod 8 \in A$, and $a_i = 0$ otherwise.

Proposition 6.6.2. *For all $a_0, \ldots, a_{23}, b_0, \ldots, b_{23} \in [0,1]$ with $\sum_{i=0}^{23} a_i \geq 9, \sum_{i=0}^{23} b_i \geq 9$, we have*

$$\sum_{t\in\mathbb{Z}/24\mathbb{Z}} (a * b)(t) f_{24}(t) \geq \frac{1}{\sqrt{5}}\sqrt{\sum_i a_i - \sum_i a_i^2}\sqrt{\sum_i b_i - \sum_i b_i^2}$$

with equality if and only if there is some $x \in \mathbb{Z}/8\mathbb{Z}$ so that a, b are lift-ups of $\{0,1,5\} + x, \{2,5,6\} - x \subseteq \mathbb{Z}/8\mathbb{Z}$.

We prove Proposition 6.6.2 by first massaging the desired inequality into a homogeneous quadratic form. It is of course easy to check the "if" implication of the equality part of Proposition 6.6.2; the "only if" direction will follow from equality needing to hold at each step of the proof and equality holding only for the claimed extremizers at the end of the proof.

By the arithmetic-geometric inequality, it suffices to show

$$\sum_{t\in\mathbb{Z}/24\mathbb{Z}} (a*b)(t)f_{24}(t) \geq \frac{1}{2\sqrt{5}}\left(\sum_i a_i - \sum_i a_i^2 + \sum_i b_i - \sum_i b_i^2\right)$$

for all $a_i, b_i \in [0,1]$ with $\sum_i a_i, \sum_i b_i \geq 9$. Since[6] $\frac{2}{9}xy \geq x+y$ if $x, y \geq 9$, it suffices to show

$$\sum_{t\in\mathbb{Z}/24\mathbb{Z}} (a*b)(t)f_{24}(t) \geq \frac{1}{2\sqrt{5}}\left(\frac{2}{9}(\sum_i a_i)(\sum_i b_i) - \sum_i a_i^2 - \sum_i b_i^2\right)$$

for all $a_i, b_i \in [0,1]$ with $\sum_i a_i, \sum_i b_i \geq 9$. Of course it then suffices to prove the inequality for any non-negative reals a_i, b_i.

Proposition 6.6.3. *For any $a_0, b_0, \ldots, a_{23}, b_{23} \in [0,\infty)$ one has*

$$\sum_{t\in\mathbb{Z}/24\mathbb{Z}} (a*b)(t)f_{24}(t) \geq \frac{1}{2\sqrt{5}}\left(\frac{2}{9}(\sum_i a_i)(\sum_i b_i) - \sum_i a_i^2 - \sum_i b_i^2\right).$$

We will present a proof of Proposition 6.6.3 due to Fedor Nazarov. The (quite ingenious) proof significantly reduces the computational power needed.

Proof.

Step 1: Reduction to a norm inequality in a single (non-negative) variable.

Using that

$$\sum_{t\in\mathbb{Z}/24\mathbb{Z}} (a*b)(t)f_{24}(t) = \sum_{t\in\mathbb{Z}/24\mathbb{Z}} (\widetilde{a}*f_{24})(t)b(t),$$

where $\widetilde{a}(i) := a(-i)$ and

$$\left(\sum_i a_i\right)\left(\sum_i b_i\right) = 24\sum_{t\in\mathbb{Z}/24\mathbb{Z}} (\widetilde{a}*\mathbb{1})(t)b(t),$$

[6] If $x, y \geq 9+\epsilon$, then $\frac{2}{9}xy \geq x+y+2\epsilon$, which is why $c'(\epsilon) := \frac{1}{\sqrt{5}}\epsilon$ suffices.

where $\mathbb{1} : \mathbb{Z}/24\mathbb{Z} \to [0,1]$ is the constant function $\equiv 1$, we wish to prove

$$\sum_{t\in\mathbb{Z}/24\mathbb{Z}} \left(\widetilde{a} * (\frac{16}{3}\mathbb{1} - 2\sqrt{5}f_{24})\right)(t)\, b(t) \le \sum_{t\in\mathbb{Z}/24\mathbb{Z}} \left[a(t)^2 + b(t)^2\right].$$

We may, of course, ignore the distinction between a and $\widetilde{a}$, so we drop the $\widetilde{}$ from here on[7]. Since $2xy \le x^2 + y^2$ for all $x, y \in \mathbb{R}$, it suffices to show

$$\sum_{t\in\mathbb{Z}/24\mathbb{Z}} \left(a * (\frac{16}{3}\mathbb{1} - 2\sqrt{5}f_{24})\right)(t)\, b(t) \le 2\left(\sum_{t\in\mathbb{Z}/24\mathbb{Z}} a(t)^2\right)^{1/2} \left(\sum_{t\in\mathbb{Z}/24\mathbb{Z}} b(t)^2\right)^{1/2},$$

which we write more compactly as

$$\langle a * \varphi, b\rangle \le 2\|a\|_2\|b\|_2,$$

with $\varphi := \frac{16}{3}\mathbb{1} - 2\sqrt{5}f_{24}$. Since $b(t) \ge 0$ for each t, it suffices to prove

$$\Big\langle (a * \varphi)_+, b\Big\rangle \le 2\|a\|_2\|b\|_2.$$

By Cauchy-Schwarz, it then suffices to prove

$$\|(a * \varphi)_+\|_2 \le 2\|a\|_2$$

for each $a : \mathbb{Z}/24\mathbb{Z} \to [0,\infty)$.

Step 2: Showing the maximizer is an eigenvector of a related operator.

By compactness, let $a = \widehat{a}$ be a maximizer of $\|(a * \varphi)_+\|_2$ subject to $\|a\|_2 = 1$ and $a \ge 0$ (pointwise). Let $\widehat{\sigma} : \mathbb{Z}/24\mathbb{Z} \to \mathbb{R}$ satisfy $|\widehat{\sigma}(t)| < \widehat{a}(t)$ whenever $\widehat{a}(t) > 0$ (think $\widehat{\sigma} \to 0$). Then

$$\begin{aligned}\left\|((\widehat{a}+\widehat{\sigma}) * \varphi)_+\right\|_2^2 - \|(\widehat{a} * \varphi)_+\|_2^2 &= \sum_{t\in\mathbb{Z}/24\mathbb{Z}} \left[\Big((\widehat{a} * \varphi)(t) + (\widehat{\sigma} * \varphi)(t)\Big)_+^2 - \Big((\widehat{a} * \varphi)(t)\Big)_+^2\right] \\ &= 2\sum_{t\in\mathbb{Z}/24\mathbb{Z}} \Big((\widehat{a} * \varphi)(t)\Big)_+ (\widehat{\sigma} * \varphi)(t) + O\left(\|\widehat{\sigma}\|^2\right) \\ &= 2\Big\langle (\widehat{a} * \varphi)_+\, ,\, \widehat{\sigma} * \varphi\Big\rangle + O\left(\|\widehat{\sigma}\|^2\right) \\ &= 2\Big\langle (\widehat{a} * \varphi)_+ * \widetilde{\varphi}\, ,\, \widehat{\sigma}\Big\rangle + O\left(\|\widehat{\sigma}\|^2\right),\end{aligned}$$

where the second equality used the fact that $(x+y)_+^2 - x_+^2 = 2yx_+ + O(y^2)$ for any reals x, y with $|y| < |x|$, and in the last equality, we again use the notation $\widetilde{\varphi}(\cdot) := \varphi(-\cdot)$. Let $\widehat{v} : \mathbb{Z}/24\mathbb{Z} \to \mathbb{R}$ be $\widehat{v} := (\widehat{a} * \varphi)_+ * \widetilde{\varphi}$ so that

[7] However, the reader should keep in mind that we are "mirroring" the extremizers.

$$\left\|((\widehat{a}+\widehat{\sigma})*\varphi)_+\right\|_2^2 - \|(\widehat{a}*\varphi)_+\|_2^2 = 2\langle\widehat{v},\widehat{\sigma}\rangle + O\left(\|\widehat{\sigma}\|^2\right).$$

We see that no $t \in \mathbb{Z}/24\mathbb{Z}$ can satisfy $\widehat{a}(t) = 0$ and $\widehat{v}(t) > 0$, for otherwise we could let $\widehat{\sigma}(t) = +\alpha$ for some (very) small $\alpha > 0$, $\widehat{\sigma}(t') = -\delta$ for some t' with $\widehat{a}(t') > 0$ and appropriate $\delta > 0$ (which will be $O(\alpha^2)$), and $\widehat{\sigma} = 0$ elsewhere, to have

$$||\widehat{a}+\widehat{\sigma}||_2 = 1 \quad \text{and} \quad \left\|((\widehat{a}+\widehat{\sigma})*\varphi)_+\right\| > \left\|(\widehat{a}*\varphi)_+\right\|,$$

contradicting the maximality of $\widehat{a}$. And similarly no $t \in \mathbb{Z}/24\mathbb{Z}$ can satisfy $\widehat{a}(t) > 0$ and $\widehat{v}(t) \leq 0$. Therefore, $\widehat{v}_+$ is positive exactly when $\widehat{a}$ is, and each are 0 otherwise. This implies

$$\widehat{v}_+ \equiv \lambda\widehat{a}$$

for some $\lambda > 0$, for otherwise one could make $2\langle\widehat{v},\widehat{\sigma}\rangle + O(\|\widehat{\sigma}\|^2)$ negative for suitable small $\widehat{\sigma}$, contradicting the maximality of $\widehat{a}$. To end this step, quickly note

$$\begin{aligned}\|(\widehat{a}*\varphi)_+\|_2^2 &= \Big\langle(\widehat{a}*\varphi)_+,(\widehat{a}*\varphi)_+\Big\rangle \\ &= \Big\langle(\widehat{a}*\varphi)_+,\widehat{a}*\varphi\Big\rangle \\ &= \langle\widehat{v},\widehat{a}\rangle \\ &= \langle\widehat{v}_+,\widehat{a}\rangle \\ &= \lambda.\end{aligned} \tag{6.9}$$

Step 3: Choosing a convenient norm.

We are given $\widehat{a} : \mathbb{Z}/24\mathbb{Z} \to [0,\infty)$ satisfying

$$\left((\widehat{a}*\varphi)_+ * \widetilde{\varphi}\right)_+ \equiv \lambda\widehat{a}$$

and, by (6.9), we wish to show $\lambda \leq 4$. It suffices to find a function ("norm") $N : [0,\infty)^{\mathbb{Z}/24\mathbb{Z}} \to [0,\infty)$ satisfying the multiplicativity condition

$$N(\gamma a) = \gamma N(a) \tag{6.10}$$

for all $\gamma \in [0,\infty)$ and $a : \mathbb{Z}/24\mathbb{Z} \to [0,\infty)$, and the two (dual) norm bounds

$$N\left((a*\varphi)_+\right) \leq 2N(a) \tag{6.11}$$

$$N\left((a*\widetilde{\varphi})_+\right) \leq 2N(a) \tag{6.12}$$

for all $a : \mathbb{Z}/24\mathbb{Z} \to [0, \infty)$. Indeed, with such a norm N, we have

$$\lambda N(\widehat{a}) = N(\lambda\widehat{a}) = N\Big(((\widehat{a} * \varphi)_+ * \widetilde{\varphi})_+ \Big) \leq 2N\Big((\widehat{a} * \varphi)_+\Big) \leq 4N(\widehat{a}).$$

Motivated by the (conjectured) extremizers, we use the norm

$$N(a) := \max\left(9\|a\|_\infty, \|a\|_1\right).$$

Step 4: Showing the desired norm bounds.

It is clear that N satisfies condition (6.10). To prove (6.11), we may normalize to $N(a) = 9$ so that it suffices to show

$$\begin{Bmatrix} \|a\|_\infty \leq 1 \\ \|a\|_1 \leq 9 \end{Bmatrix} \implies \begin{Bmatrix} \|(a * \varphi)_+\|_\infty \leq 2 \\ \|(a * \varphi)_+\|_1 \leq 18 \end{Bmatrix},$$

where, to recall,

$$\varphi = \frac{16}{3}\mathbb{1} - 2\sqrt{5} f_{24}.$$

So take $a : \mathbb{Z}/24\mathbb{Z} \to [0, \infty)$ with $\|a\|_\infty \leq 1$ and $\|a\|_1 \leq 9$. Then we easily have

$$\|(a * \varphi)_+\|_\infty \leq \max_{t \in \mathbb{Z}/24\mathbb{Z}} \frac{1}{24} \sum_{j \in \mathbb{Z}/24\mathbb{Z}} a(j)\varphi(t - j) \leq \frac{1}{24} \cdot \frac{16}{3} \cdot 9 = 2.$$

As $a \mapsto \|(a * \varphi)_+\|_1$ is convex, it simply suffices to check that $\|(a * \varphi)_+\|_1 \leq 18$ for all $a \in \{0, 1\}^{24} \subseteq [0, 1]^{\mathbb{Z}/24\mathbb{Z}}$. We may assume WLOG that $a_0 = 1$, so that there are only $\sum_{k=0}^{8} \binom{23}{k} < 10^6$ cases to check, which is easily handled by a computer.

We do everything analogous to establish (6.12) as well.

Below is the python code, presented in two columns to save space.

```
import math
import itertools

f = []
for t in range(0,24):
    sum1 = 0
    for j in range(0,24):
        if ((j*j)%24 == t):
            sum1 = sum1+1
    f.append(sum1)
phi = []
for t in range(0,24):
    phi.append(16/3-2*math.sqrt(5)*f[t])
phit = []
for t in range(0,24):
    phit.append(phi[23-t])

def h(a,psi):
    sum1 = 0
    for t in range(0,24):
        sum2 = 0
        for j in range(0,24):
            sum2=sum2+a[j]*psi[(t-j)%24]
        sum2 = sum2/24
        sum2 = max(sum2,0)
        sum1 = sum1+sum2
    return sum1
```

```
c = []
for j in range(1,24):
    c.append(j)
max1 = 0
max2 = 0
for k in range(0,9):
    for A in itertools.combinations(c,k):
        A = list(A)
        A.insert(0,0)
        a = []
        for j in range(0,24):
            if (j in A):
                a.append(1)
            else:
                a.append(0)
        v1 = h(a,phi)
        v2 = h(a,phit)
        max1 = max(max1,v1)
        max2 = max(max2,v2)
        if (v1 >= 17.99):
            print ("extremizer - "+str(a))
        if (v2 >= 17.99):
            print ("extremizer for dual - "+str(a))
print (max1)
print (max2)
```

The output of the python code is as follows.

```
extremizer - [1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0]
extremizer for dual - [1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0]
extremizer for dual - [1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0]
extremizer - [1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1]
extremizer - [1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0]
extremizer for dual - [1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1]
18.000000000000004
18.000000000000004
```

Since we printed all a for which $\|(a*\varphi)_+\|_1, \|(a*\widetilde{\varphi})_+\|_1 \geq 17.99$ and the ones printed have $\|(a*\varphi)_+\|_1, \|(a*\widetilde{\varphi})_+\|_1 = 18$, the $+4\cdot 10^{-15}$ (added to 18) is merely a computer-induced rounding error.

We finish by analyzing the extremizers. We obtained only 3 of the 8 conjectured extremizers; however, we assumed WLOG that $a_0 = 1$. Translating the outputted extremizers indeed recovers all 8 conjectured extremizers for a. Since such a have $\sum_i a_i - \sum_i a_i^2 = 0$, the only extremizing b, for a given a, must satisfy $\sum_t (a*b)(t) f_{24}(t) = 0$, i.e., $a+b$ "contains" no squares. Since all extremizers a are translates of one another, we may focus on a particular extremizer a. Then, as is easily checked, b is uniquely determined merely by "process of elimination". $\square$

Chapter 7

On the length of Pierce expansions

7.1 Summary

For a given positive integer n, how long can the process $x \mapsto n \pmod{x}$ last before reaching 0? We improve Erdős and Shallit's upper bound of $O(n^{\frac{1}{3}+\varepsilon})$ to $O(n^{\frac{1}{3}-\frac{2}{177}+\varepsilon})$ for any $\varepsilon > 0$.

7.2 Introduction

The continued fraction expansion of a real number $x \in (0,1)$, given by

$$x = \frac{1}{a_1 + \frac{1}{a_2 + \ldots}}$$

plays an important role throughout number theory, where the terms a_i can be extracted, for example, from the iterated process $t \mapsto \frac{1}{t} \pmod{1}$ beginning with $t = x$. It is well-known and not difficult to see that the continued fraction expansion of a real number x is finite if and only if x is a rational number. And if x is rational, the sequence of terms a_i produced are exactly the quotients produced by the classic Euclidean algorithm applied to the numerator and denominator.

In this chapter, we are concerned with the *Pierce expansion* of a real number $x \in (0,1)$, introduced by Pierce [52] and named by Shallit [58]. Here, the expansion is of the form

$$x = \frac{1}{b_1} - \frac{1}{b_1 b_2} + \frac{1}{b_1 b_2 b_3} - \ldots,$$

where now the terms b_i can be extracted from the iterated process $t \mapsto 1 \pmod{t}$ beginning with $t = x$. It is also not difficult to see that the Pierce expansion of a real number x is finite if and only if x is rational (see, e.g., [58]). And if x is rational, the sequence of terms b_i produced are exactly the quotients produced by an algorithm that at first glance appears similar to Euclid's algorithm.

Let us give an example of the algorithm. Say $x = \frac{13}{35}$. We start with 13 and repeatedly obtain successive integers by reducing 35 modulo the current number. For example,

$$35 = 2 \cdot 13 + 9$$

$$35 = 3 \cdot 9 + 8$$

$$35 = 4 \cdot 8 + 3$$

$$35 = 11 \cdot 3 + 2$$

$$35 = 17 \cdot 2 + 1$$

$$35 = 35 \cdot 1 + 0$$

gives rise to

$$\frac{13}{35} = \frac{1}{2} - \frac{1}{2 \cdot 3} + \frac{1}{2 \cdot 3 \cdot 4} - \frac{1}{2 \cdot 3 \cdot 4 \cdot 11} + \frac{1}{2 \cdot 3 \cdot 4 \cdot 11 \cdot 17} - \frac{1}{2 \cdot 3 \cdot 4 \cdot 11 \cdot 17 \cdot 35}.$$

Motivated by the known fact that the Euclidean algorithm used to divide a positive integer a by a positive integer q terminates after $O(\log q)$ steps (which is sharp), it is natural to ask how quickly must the above algorithm terminate, as a function of the denominator, no matter what the numerator is.

To this end, for positive integers $a, n \in \mathbb{N}$, define $P(a, n)$ to be the first positive integer k such that $a_k = 0$, where $a_0 := a$ and $a_{j+1} = n \pmod{a_j} \in \{0, 1, \ldots, a_j - 1\}$ for $j \geq 0$. In the above example we have $P(a, n) = P(13, 35) = 6$. Since we only concern ourselves with the "length" of the algorithm, one need not keep track of quotients and may compress for instance the above example to

$$35 \pmod{13} = 9$$

$$35 \pmod{9} = 8$$

$$35 \pmod{8} = 3$$

$$35 \pmod{3} = 2$$

$$35 \pmod{2} = 1$$

$$35 \pmod{1} = 0.$$

Noting $P(a,n) = 2$ if $a > n$, we set

$$P(n) := \max_{1 \le a \le n} P(a,n).$$

The problem we consider is to obtain bounds on $P(n)$. Shallit [58] proved, using purely "archimedean" arguments, that $P(n) \ll n^{\frac{1}{2}+\varepsilon}$, while $P(n) = \Omega(\frac{\log n}{\log\log n})$ for infinitely many n. The upper bound was improved by Erdős and Shallit [23] who leveraged "arithmetic" arguments to combine with the previous "archimedean" ones. They established $P(n) \ll n^{\frac{1}{3}+\varepsilon}$ (see Section 7.3 for our conventions regarding Vinogradov notation) and they also improved the lower bound to $P(n) = \Omega(\log n)$ for infinitely many n. These bounds have since remained the state of the art, with the exponent 1/3 representing a natural barrier.

In this chapter, we improve the upper bound on $P(n)$, (slightly) pushing past the 1/3 barrier.

Theorem 7.2.1. *We have*

$$P(n) \ll n^{\frac{1}{3}-\frac{2}{177}+\varepsilon}.$$

We did not put substantial effort into optimizing the exponent gain achieved in Theorem 7.2.1; we could not, however, see a way to improve the upper bound to $O(n^{\varepsilon})$ using our techniques (or any others).

Secondly, we establish a lower bound that applies to all $n \in \mathbb{N}$. As we can tell, the best bound known prior was $\Omega(\log\log n)$.

Theorem 7.2.2. *We have the lower bound*

$$P(n) \gg \frac{\log n}{\log\log n}$$

for all sufficiently large $n \ge n_0$.

As one can see, there is an exponential gap between the best known lower and upper bounds on $P(n)$. We hope this chapter will reignite interest in determining the true asymptotics and related questions. In this same vein, since the problem is quite simple-to-state and is potentially prone to elementary methods, we strive to make this chapter as self-contained as possible.

In Section 2, we define the notation we use throughout the chapter. In Section 3, we give the proof of our main theorem, Theorem 7.2.1. In Section 4, we give the proof of the lower bound, Theorem 7.2.2. Finally, we provide in the appendix the definitions and theorems we import from the analytic number theory literature.

7.3 Notation

We use the standard Vinogradov notation, in which, we write $A \ll B$ (and equivalently $B \gg A$) to denote that $|A| \leq C|B|$ for some $C > 0$ that is either absolute or depends only on ε. We write $A \asymp B$ to denote that both $A \ll B$ and $B \ll A$ hold. We further, for a parameter β, use the notation $\ll_\beta$ and $\asymp_\beta$ to mean that the implicit "constant" C in the inequalities can depend on β. For brevity, throughout the chapter we do not explicitly say "for all $\varepsilon > 0$" in statements involving ε. For positive integers $a, A \in \mathbb{N}$, we write $a \sim A$ to denote $A < a \leq 2A$.

7.4 Tools from analytic number theory

In this section, we prove some results that are standard and often used in analytic number theory.

7.4.1 Construction of a bump function

In this subsection, we give an explicit example of a bump function used to prove Theorem 7.2.1. Precisely, we construct a C^∞ function $w : \mathbb{R} \to \mathbb{R}$ satisfying

$$\mathbb{1}_{[-1,1]}(x) \leq w(x) \leq \mathbb{1}_{[-2,2]}(x)$$

for all $x \in \mathbb{R}$.

Define $\psi : [0,1] \to [0,1]$ by

$$\psi(x) := \frac{\int_0^x \exp\left(-\frac{1}{t(1-t)}\right) dt}{\int_0^1 \exp\left(-\frac{1}{t(1-t)}\right) dt}.$$

Clearly $\psi(0) = 0$ and $\psi(1) = 1$. Furthermore,

$$\psi'(x) = \frac{\exp\left(-\frac{1}{x(1-x)}\right)}{\int_0^1 \exp\left(-\frac{1}{t(1-t)}\right) dt}$$

shows ψ' and all of its derivatives vanish at $x = 0$ and at $x = 1$.

Therefore, we can take our bump function w to be

$$w(x) := \begin{cases} 0 & x < -2 \\ \psi(x+2) & -2 \leq x \leq -1 \\ 1 & -1 < x < 1 \\ \psi(2-x) & 1 \leq x \leq 2 \\ 0 & x > 2. \end{cases}$$

7.4.2 Poisson summation

In this subsection, we state the Poisson summation formula, which we make critical use of in the proof of Theorem 7.2.1. We first define functions that behave nicely enough for Poisson summation to apply.

Definition 7.4.1. We that a function $f : \mathbb{R} \to \mathbb{C}$ is *Schwartz* if f is infinitely differentiable and satisfies

$$\left|f^{(j)}(x)\right| \ll_{j,N} (1+|x|)^{-N}$$

for all $j, N \geq 0$. If f is Schwartz (for example), we can define its *Fourier transform* to be the function $\hat{f} : \mathbb{R} \to \mathbb{C}$ given by

$$\hat{f}(\xi) := \int_{-\infty}^{\infty} f(x) e(-\xi x) dx.$$

Proposition 7.4.2 (Poisson Summation). *Let $f : \mathbb{R} \to \mathbb{C}$ be a Schwartz function. Then we have the identity*

$$\sum_{k\in\mathbb{Z}} f(k) = \sum_{r\in\mathbb{Z}} \hat{f}(r).$$

We make use of the following corollary, obtained by writing $h = kb - n$ and applying Poisson summation to $k \mapsto f(kb - n)$.

Corollary 7.4.3. *Let $f : \mathbb{R} \to \mathbb{C}$ be a Schwartz function. Then, for any positive integers $n, b \in \mathbb{N}$, we have*

$$\sum_{\substack{h\in\mathbb{Z} \\ h\equiv -n(b)}} f(h) = \frac{1}{b}\sum_{r\in\mathbb{Z}} e\left(\frac{rn}{b}\right) \hat{f}\left(\frac{r}{b}\right).$$

7.5 Proof of main theorem

In this section, we prove our main theorem, that $P(n) \ll n^{\frac{1}{3}-\frac{2}{177}+\varepsilon}$. We do this by establishing bounds for the amount of time the process (algorithm) spends in dyadic intervals.

For the rest of this section, fix a (large) positive integer n and a positive integer a_0, letting $a_{j+1} = n \pmod{a_j}$ for $j \geq 0$.

Write

$$T(A) := \#\Big\{j \geq 0 : a_j \sim A\Big\}.$$

The first bound we present on $T(A)$ was proven in [58] and is due to "archimedean" reasons.

Lemma 7.5.1. *We have* $T(A) \leq \frac{n}{2A} + 2$.

Proof. For $j \geq 0$, let $b_j = \lfloor \frac{n}{a_j} \rfloor$, so that $\frac{n}{b_j+1} < a_j \leq \frac{n}{b_j}$. We claim that $b_{j+1} > b_j$ for each $j \geq 0$. Indeed, if not, $n = b_j a_j + a_{j+1}$, so $a_{j+1} > \frac{n}{b_j+1}$ implies $n(b_j+1) - b_j a_j (b_j+1) > n$, which yields $a_j < \frac{n}{b_j+1}$, a contradiction. Therefore, since $a_j \sim A$ implies $b_j \in [\frac{n}{2A} - 1, \frac{n}{A})$, the desired bound follows. □

Note that Lemma 7.5.1 combined with the trivial $T(a) \leq a$ already establishes the bound $P(n) \ll n^{1/2}$. The second bound we present improves this trivial bound, by taking advantage of "arithmetic" properties of the iterative process. It was proven in [23].

Lemma 7.5.2. *For* $1 \leq A \leq n$, *we have the bound*

$$T(A) \ll A^{\frac{1}{2}} n^{\varepsilon}.$$

Proof. If $T(A) \leq 1$, we are done, so suppose that $T(A) \geq 2$. Let

$$\mathcal{J} = \left\{ j \geq 0 : a_j \sim A, a_j - a_{j+1} \leq \frac{4A}{T(A)} \right\}.$$

Note that

$$\sum_{\substack{j \geq 0 \\ a_j, a_{j+1} \sim A}} 1 = T(A) - 1 \geq \frac{1}{2} T(A), \quad \sum_{\substack{j \geq 0 \\ a_j, a_{j+1} \sim A}} (a_j - a_{j+1}) \leq A.$$

It follows that

$$\# \left\{ j \geq 0 : a_j \sim A, a_j - a_{j+1} > \frac{4A}{T(A)} \right\} < \frac{1}{4} T(A),$$

so $\#J \geq \frac{1}{4} T(A)$. Now, note that for all j,

$$a_{j+1} \equiv n \mod (a_j) \implies a_j | n + a_j - a_{j+1}.$$

We obtain that

$$T(A) \ll \#\mathcal{J} \leq \# \left\{ (a, h) : 0 < h \leq \frac{4A}{T(A)}, a \sim A \right\} \leq \sum_{h \leq \frac{4A}{T(A)}} \sum_{\substack{a \sim A \\ a | n+h}} 1.$$

By the divisor bound, $\sum_{\substack{a\sim A\\ a|n+h}} 1 \leq d(n+h) \ll n^{\varepsilon}$, so we obtain

$$T(A) \ll \frac{A}{T(A)} n^{\varepsilon}.$$

Rearranging yields the desired result. □

Together, Lemmas 7.5.1, 7.5.2 applied to the ranges $A \geq n^{2/3}, A \leq n^{2/3}$, respectively, give the bound $P(n) \ll n^{\frac{1}{3}+\varepsilon}$. To obtain a bound of $n^{\frac{1}{3}-\delta+\varepsilon}$, it therefore suffices to show that $T(A) \ll n^{\frac{1}{3}-\delta+\varepsilon}$ for $A \in [n^{\frac{2}{3}-2\delta}, n^{\frac{2}{3}+\delta}]$. This is the content of Proposition 7.5.3 for sufficiently small δ. To do this, we use of the arithmetic information obtained by analyzing two consecutive jumps. In the end, we are reduced, roughly, to obtaining a power saving over the trivial bound for the sum

$$\sum_{b\sim n^{1/3}} e\left(\frac{n}{b}\right).$$

Such bounds follow from standard exponential sum bounds. In our case, we use the exponent pair $\left(\frac{13}{84}+\varepsilon, \frac{55}{84}+\varepsilon\right)$ of Bourgain [?]. Much simpler methods would have also worked, to give a slightly worse saving over the trivial bound (the van der Corput A-process, followed by the B-process, for example).

Proposition 7.5.3. *Suppose that $\delta, \eta > 0$ are such that*

$$\delta < \frac{1}{18}, \eta \leq \frac{1}{3} - \delta.$$

Then, for $n^{\frac{2}{3}-2\delta} \leq A \leq n^{\frac{2}{3}+\delta}$, we have

$$T(A) \ll n^{\frac{1}{3}+2\delta-\eta} + n^{\frac{1}{3}-\delta} + n^{\frac{1}{3}-\frac{4}{63}+\frac{349}{84}\delta+\frac{13}{84}\eta+\varepsilon}$$

Before proving Proposition 7.5.3, let us first quickly spell out how our main theorem, Theorem 7.2.1, follows.

Proof of Theorem 7.2.1 assuming Proposition 7.5.3. Take

$$\delta = \frac{2}{177}, \eta = 3\delta.$$

It is easy to check that δ, η satisfy the hypotheses of Proposition 7.5.3. We have that

$$P(n) \leq 1 + \sum_{A\leq n} T(A),$$

where the sum over A runs over powers of 2. The contribution of $A > n^{\frac{2}{3}+\delta}$ is, by Lemma 7.5.1,

$$\ll \sum_{n^{\frac{2}{3}+\delta}<A\leq n} \frac{n}{A} \ll n^{\frac{1}{3}-\delta}.$$

By Lemma 7.5.2, the contribution of $A < n^{\frac{2}{3}-2\delta}$ is

$$\ll n^{\varepsilon} \sum_{A<n^{\frac{2}{3}-2\delta}} A^{\frac{1}{2}} \ll n^{\frac{1}{3}-\delta+\varepsilon},$$

For $n^{\frac{2}{3}-2\delta} \leq A \leq n^{\frac{2}{3}+\delta}$, by Proposition 7.5.3, we have that

$$T(A) \ll n^{\frac{1}{3}-\delta},$$

since

$$\frac{2}{177} = \delta = \eta - 2\delta = \frac{4}{63} - \frac{349}{84}\delta - \frac{13}{84}\eta.$$

Then, summing over A in $[n^{\frac{2}{3}-2\delta}, n^{\frac{2}{3}+\delta}]$ at the harmless cost of $O(\log n)$, the the desired result follows. □

Proof of Proposition 7.5.3. Suppose that $T(A) \geq T_0 = n^{\frac{1}{3}-\delta}$, for we are done otherwise. Let m be so that $a_{m+T(A)} \leq A < a_{m+T(A)-1} < \cdots < a_m \leq 2A$. Then, for a positive proportion of $m+2 \leq j < m+T(A)$, we have that

$$a_{j-2} - a_j \leq H := \frac{4A}{T_0}.$$

We record the bound $n^{\frac{1}{3}-\delta} \ll H \ll n^{\frac{1}{3}+2\delta}$. Call the set of such j $\mathcal{J}$. Consider some $j \in \mathcal{J}$, and write $a = a_{j-2}, a-h = a_{j-1}, a-h-h' = a_j$. Then, as in the proof of Lemma 7.5.2, we have

$$a|n+h, a-h|n+h'.$$

In particular, there exist $b \asymp n/A, k$ such that $ab = n+h, (a-h)(b+k) = n+h'$. Dividing these two, we obtain that

$$1 + \frac{k}{b} = 1 + O\left(\frac{H}{A}\right),$$

so $|k| \ll \frac{Hn}{A^2}$. Also, note that

$$|(b+k)h - ak| = |ab - (a-h)(b+k)| \ll H.$$

For $k \neq 0$, we have that $(b+k)h = ak + O(H)$, so

$$h = \frac{ak}{b+k} + O\left(\frac{AH}{n}\right) = \frac{abk}{b(b+k)} + O\left(\frac{AH}{n}\right) = \frac{nk}{b(b+k)} + O\left(\frac{AH}{n}\right)$$

since $Hk/B^2 \ll H/B = AH/n$. Write $H_0 = \frac{nk}{b(b+k)}$. Recall that $\eta \leq \frac{1}{3} - \delta$, so

$$H_0 n^{-\eta} \geq \frac{H_0}{T_0} \gg \frac{A^2}{nT_0} \asymp \frac{AH}{n}$$

It follows that for some sufficiently large $C > 0$, $\mathbb{1}_{|h-H_0| \ll AH/n} \leq \mathbb{1}_{|h-H_0| \leq L}$ with $L = CH_0 n^{-\eta}$. The reason for this maneuver is to lower the "analytic conductor" of the phase in the resulting exponential sum so that we may get superior savings when we execute the sum over b. □

Theorem 7.5.4. *Let n be a (large) positive integer. Let A, k be positive integers satisfying*

$$n^{\frac{2}{3}-2\delta} \leq A \leq n^{\frac{2}{3}+\delta} \qquad 1 \leq k \ll \frac{Hn}{A^2}.$$

Then, letting

$$H = \frac{4A}{n^{\frac{1}{3}-\delta}} \qquad H_0(b) = \frac{nk}{b(b+k)} \qquad L = CH_0(\frac{n}{A})n^{-\eta},$$

we have

$$\sum_{|k| \ll Hn/A^2} \sum_{h \leq H} \sum_{\substack{b|n+h \\ b \asymp n/A}} \mathbb{1}_{|h-H_0(b)| \leq L} \ll n^{\frac{89}{126}+\frac{13}{84}\eta+\frac{13}{84}\delta} A^{-\frac{55}{84}} n^{\varepsilon}$$

Proof. Take some smooth w so that $\mathbb{1}_{[-1,1]} \leq w \leq \mathbb{1}_{[-2,2]}$ (see Section 7.4). Then, we have

$$\sum_{h \leq H} \sum_{\substack{b|n+h \\ b \asymp n/A}} \mathbb{1}_{|h-H_0| \leq L} \leq \sum_{|k| \ll Hn/A^2} \sum_{b \asymp n/A} \sum_{h \equiv -n(b)} w\left(\frac{h-H_0}{L}\right).$$

By Poisson summation (see Proposition 7.4.2),

$$\begin{aligned}\sum_{b \asymp n/A} \sum_{h \equiv -n(b)} w\left(\frac{h-H_0}{L}\right) &= \sum_{b \asymp n/A} \frac{L}{b} \sum_{r \in \mathbb{Z}} e\left(\frac{r(n+H_0)}{b}\right) \hat{w}\left(\frac{Lr}{b}\right) \\ &= \sum_{r \in \mathbb{Z}} \sum_{b \asymp n/A} \frac{L}{b} e\left(\frac{r(n+H_0)}{b}\right) \hat{w}\left(\frac{Lr}{b}\right)\end{aligned}$$

The contribution of the zero frequency, $r = 0$, is

$$\ll \frac{Hn}{A^2} \cdot \frac{n}{A} \cdot \frac{Hn^{-\lambda}}{n/A} \ll n^{\frac{1}{3}+2\delta-\lambda},$$

which is acceptable. It remains to bound the contribution of $|r| > 0$, and we suppose from now on that $|r| > 0$. Defining

$$H_0(x,k) := \frac{nk}{x(x+k)},$$

a quick computation shows, for $x_0 \asymp n/A$, that

$$\frac{\mathrm{d}^j}{\mathrm{d}x^j} \left. \frac{r(n + H_0(x,k))}{x} \right|_{x=x_0} \asymp_j rAx_0^{-j}.$$

By Theorem 6 of [?], we have the exponent pair $(\frac{13}{84} + \varepsilon, \frac{55}{84} + \varepsilon)$ (see §8.4 of [34] for a definition; note that in the notation of [34], we instead have the exponent pair $(13/84, 13/84)$). By partial summation (see, e.g., [44, Lemma 2.2]), the fact that $\hat{w}, (\hat{w})'$ are Schwartz, and that $|r| > 0$ (which implies that $|r|A \gg n/A$, so (8.56) of [34] holds), we have for some $c > 0$ that

$$\begin{aligned} &\Bigg| \sum_{b \asymp n/A} e\left(\frac{r(n + H_0(b,k))}{b}\right) \frac{n/A}{b} \hat{w}\left(\frac{Lr}{b}\right) \Bigg| \\ &\qquad \ll \left(1 + \left|\frac{Lr}{n/A}\right|\right)^{-2022} \sup_{t \ll n/A} \Bigg| \sum_{cn/A < b \le t} e\left(\frac{r(n + H_0(b,k))}{b}\right) \Bigg| \\ &\qquad \ll \left(1 + \left|\frac{Lr}{n/A}\right|\right)^{-2022} \left(\frac{A^2|r|}{n}\right)^{\frac{13}{84}+\varepsilon} \left(\frac{n}{A}\right)^{\frac{55}{84}+\varepsilon}. \end{aligned}$$

Putting this all together, we obtain that

$$\begin{aligned} &\Bigg| \sum_{|k| \ll Hn/A^2} \sum_{b \asymp n/A} \frac{L}{n/A} \sum_{r \neq 0} e\left(\frac{r(n + H_0)}{b}\right) \frac{n/A}{b} \hat{w}\left(\frac{Lr}{b}\right) \Bigg| \\ &\ll \frac{Hn}{A^2} \cdot \left(\frac{An^\eta}{H}\right)^{\frac{13}{84}} \left(\frac{n}{A}\right)^{\frac{55}{84}} \cdot n^\varepsilon \ll \frac{n}{AT_0} n^{\frac{13}{84}\eta} T_0^{\frac{13}{84}} \left(\frac{n}{A}\right)^{\frac{55}{84}} n^\varepsilon \\ &\ll n^{3\delta} \cdot n^{\frac{13}{84}\eta} n^{\frac{1}{3}\cdot\frac{13}{84} - \frac{13}{84}\delta} n^{\frac{1}{3}\cdot\frac{55}{84} + \frac{55}{84}\cdot 2\delta} \\ &\ll n^{\frac{1}{3} - \frac{4}{63} + \frac{349}{84}\delta + \frac{13}{84}\eta + \varepsilon}. \end{aligned}$$

The desired result follows. □

7.6 Proof of improved lower bound

In this section, we prove Theorem 7.2.2, repeated below for the reader's convenience.

Theorem 7.6.2. *There exists $c > 0$ so that*

$$P(n) \geq c\frac{\log n}{\log\log n}$$

for all $n \geq 1$.

Shallit [58] and Erdős-Shallit established lower bounds for $P(n)$ of $c\frac{\log n}{\log\log n}$ and $c\log n$, respectively, (only) for positive integers n such that $n+1$ is divisible by all sufficiently small positive integers. Such positive integers n will cause the process $x \mapsto n \pmod{x}$ to repeatedly decrement by 1 at the end. We establish a lower bound that is valid for all positive integers by choosing a starting number based on n that causes the process $x \mapsto n \pmod{x}$ to repeatedly decrement by 1 at the beginning, for "archimedean" reasons rather than "arithmetic" ones.

We will need the following elementary lemma.

Lemma 7.6.3. *There exists $c > 0$ so that the following holds for sufficiently large $n \in \mathbb{N}$. For any $k \in \mathbb{N}$ with $k \leq c\frac{\log n}{\log\log n}$, one has*

$$(-1)^k k!\left(\sum_{j=0}^{k}\frac{(-1)^j}{j!} - \frac{1}{e}\right)n > \frac{n}{k+2} + k!.$$

Proof. Note, from the power series for e^{-1}, that

$$(-1)^k(k+2)k!\left(\sum_{j=0}^{k}\frac{(-1)^j}{j!} - \frac{1}{e}\right) = 1+\frac{1}{(k+3)(k+1)} - O\left(\frac{1}{k^3}\right) = 1+\frac{1}{k^2} - O\left(\frac{1}{k^3}\right),$$

which is greater than $1+\frac{k!(k+2)}{n}$ for sufficiently large n, by assumption. □

Proof of Theorem 7.2.2. By adjusting the implied constant, we may assume n is sufficiently large. Let $a = \lfloor(1-\frac{1}{e})n\rfloor$, $a_0 = a$, and $a_{k+1} = n \pmod{a_k}$ for $k \geq 0$. Let $b_0 = (1-\frac{1}{e})n$ and $b_k = (-1)^k k!\left(\sum_{j=0}^{k}\frac{(-1)^j}{j!} - \frac{1}{e}\right)n$ for $k \geq 1$.

We show $P(a,n) \geq c\frac{\log n}{\log\log n}$, where $c > 0$ is as in Lemma 7.6.3.

We prove inductively that $|a_k - b_k| \leq k!$ and $a_k = n - ka_{k-1}$. For $k = 0$, the first is clearly true. The second is true for $k = 1$ and thus so is the first. Now assume they are both true for some $k \geq 1$. We have by Lemma 7.6.3 that $a_k \geq b_k - k! > \frac{n}{k+2}$. Since $\lfloor \frac{n}{a_k} \rfloor$ must strictly increase, we have $a_k < \frac{n}{k+1}$. Therefore, $a_{k+1} = n - (k+1)a_k$ and thus $|a_{k+1} - b_{k+1}| = |(n - (k+1)a_k) - (n - (k+1)b_k)| = (k+1)\,|a_k - b_k| \leq (k+1)!$. We have thus shown $a_k > \frac{n}{k+2} > 0$ as long as $k \leq c\frac{\log n}{\log \log n}$. □

Chapter 8

Approximate union-closed conjecture

8.1 Summary

A set system is called union closed if for any two sets in the set system their union is also in the set system. Gilmer recently proved that in any union closed set system, some element belongs to at least a 0.01 fraction of sets and conjectured that his technique can be pushed to the constant $\frac{3-\sqrt{5}}{2}$. We verify his conjecture; show that it extends to approximate union closed set systems, where for nearly all pairs of sets their union belongs to the set system; and show that for such set systems this bound is optimal.

8.2 Introduction

The union closed conjecture is a well-known conjecture in combinatorics.

Definition 8.2.1 (Union closed set system). A set system $\mathcal{F}$ is *union closed* if for all $A, B \in \mathcal{F}$ we have $A \cup B \in \mathcal{F}$.

Frankl introduced the conjecture that for any finite union closed set system $\mathcal{F}$, there is an element in at least $\frac{1}{2}$ of the sets of $\mathcal{F}$. Recently, Gilmer [27] established the first constant lower bound for this conjecture, obtaining $\frac{1}{100}$ in place of $\frac{1}{2}$. Gilmer conjectured that his technique can be sharpened to give the constant $\psi := \frac{3-\sqrt{5}}{2} \approx 0.38$. Below we verify his conjecture, and also show that it is optimal for "approximate" union closed set systems.

Definition 8.2.2 (Approximate union closed set system). Let $0 \leq c \leq 1$. A set system $\mathcal{F}$ is c-approximate union closed if for at least a c-fraction of the pairs $A, B \in \mathcal{F}$ we have $A \cup B \in \mathcal{F}$.

Informally, we say that $\mathcal{F}$ is approximate union closed if it is $1-o(1)$ approximate union closed. The following theorem shows that in any approximate union closed set system, some element is in a $\psi - o(1)$ fraction of sets.

Theorem 8.2.3. *Let $\mathcal{F}$ be a $(1-\varepsilon)$-approximate union closed set system, where $\varepsilon < 1/2$. Then there is an element which is contained in a $\psi - \delta$ fraction of sets in $\mathcal{F}$, where $\delta = 2\varepsilon\left(1 + \frac{\log(1/\varepsilon)}{\log|\mathcal{F}|}\right)$.*

The threshold of ψ is optimal for approximate union closed set systems, as the following example shows.

Example 8.2.4. Let n be large enough, and define the following set systems over $[n]$:

$$\mathcal{F}_1 = \{x \in \{0,1\}^n : |x| = \psi n + n^{2/3}\}, \mathcal{F}_2 = \{x \in \{0,1\}^n : |x| \geq (1-\psi)n\}, \mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2.$$

One can verify that: (i) $\mathcal{F}$ is $1 - o(1)$ approximate union closed (using the fact that $1 - \psi = 2\psi - \psi^2$); (ii) that $|\mathcal{F}_2| = o(|\mathcal{F}_1|)$; and (iii) that hence each element $i \in [n]$ is in at most $\psi + o(1)$ fraction of sets in $\mathcal{F}$.

8.3 Preliminaries

All logarithms are in base two. Let $h(x) = -(x \log x + (1-x)\log(1-x))$ be the binary entropy function. Let $\varphi = 1 - \psi = \frac{\sqrt{5}-1}{2}$ be the positive root of $x^2 + x - 1 = 0$. We will rely on the following analytic claim which we verified using a computer simulation. It has been proven rigorously in [3].

Claim 8.3.1. *The minimum of $\frac{h(x^2)}{xh(x)}$ for $x \in [0,1]$ is obtained at $x = \varphi$.*

8.4 Analytic claims

Let $f : [0,1]^2 \to \mathbb{R}_{\geq 0}$ be defined as

$$f(x,y) := \frac{h(xy)}{h(x)y + h(y)x}$$

for $(x,y) \in (0,1)^2$ and extended (continuously) to $[0,1]^2$ by setting $f(x,y) = 1$ if $x \in \{0,1\}$ or $y \in \{0,1\}$.

Claim 8.4.1. *The function f is minimized at (φ, φ). At this point $f(\varphi, \varphi) = \frac{1}{2\varphi}$.*

Proof. First, by routine calculations one can verify that f is indeed continuous on $[0,1]^2$ and that $f(x,y) < 1$ for $(x,y) \in (0,1)^2$. Thus, the minimum of f is attained in $(0,1)^2$. Next, let $g(x) = \frac{h(x)}{x}$, which is defined on $(0,1)$, and note that

$$f(x,y) = \frac{g(xy)}{g(x) + g(y)}.$$

We first show that f is minimized on the diagonal, namely at some point (x,x). Assume that f is minimized at some point (x^*, y^*), and let $\alpha = f(x^*, y^*)$. Define

$$F(x,y) = g(xy) - \alpha(g(x) + g(y)).$$

Then $F(x,y) \geq 0$ for all $x, y \in (0,1)^2$ and $F(x^*, y^*) = 0$. Thus the partial derivatives of F must be zero at the minimum point:

$$\frac{\partial F}{\partial x}(x^*, y^*) = \frac{\partial F}{\partial y}(x^*, y^*) = 0.$$

Evaluating the derivatives gives

$$\frac{\partial F}{\partial x}(x,y) = g'(xy) \cdot y - \alpha g'(x), \qquad \frac{\partial F}{\partial y}(x,y) = g'(xy) \cdot x - \alpha g'(y).$$

Define $G(x) = xg'(x)$ and note that we obtained that $G(x^*) = G(y^*)$. A direct calculation gives $g'(x) = \frac{\log(1-x)}{x^2}$, which implies that G is monotonically decreasing, and so we must have $x^* = y^*$.

Finally, restricting to $x = y$, we have

$$f(x,x) = \frac{h(x^2)}{2xh(x)}.$$

Claim 8.3.1 gives that $f(x,x)$ is minimized at $x = \varphi$. Since $\varphi^2 = 1 - \varphi$ we have $h(\varphi^2) = h(\varphi)$ and hence

$$f(\varphi, \varphi) = \frac{1}{2\varphi}.$$

□

Corollary 8.4.2. *For $x, y \in [0,1]$ we have*

$$h(xy) \geq \frac{1}{2\varphi}\Big(xh(y) + yh(x)\Big).$$

8.5 Proof of the main theorem

Claim 8.5.1. *Let A, B be two independent random variables taking values in $\{0,1\}^n$. Assume for all $i \in [n]$ that $\Pr[A_i = 0] \geq p$ and $\Pr[B_i = 0] \geq p$. Then*

$$H(A \cup B) \geq \frac{p}{2\varphi}\Big(H(A) + H(B)\Big).$$

Proof. The chain rule and data processing inequality yield

$$H(A \cup B) = \sum_{i \in [n]} H(A_i \cup B_i | (A \cup B)_{<i}) \geq \sum_{i \in [n]} H(A_i \cup B_i | A_{<i}, B_{<i}).$$

Let $p(x) = \Pr[A_i = 0 | A_{<i} = x]$ and $q(y) = \Pr[B_i = 0 | B_{<i} = y]$. Then by Corollary 8.4.2

$$H\Big(A_i \cup B_i | A_{<i} = x, B_{<i} = y\Big) = h\Big(p(x)q(y)\Big) \geq \frac{1}{2\varphi}\Big(p(x)h(q(y)) + q(y)h(p(x))\Big).$$

Averaging over $A_{<i}, B_{<i}$ which are independent gives

$$\begin{aligned} H(A_i \cup B_i | A_{<i}, B_{<i}) &\geq \frac{1}{2\varphi}\Big(\mathbb{E}_{A_{<i}}[p(A_{<i})] \cdot \mathbb{E}_{B_{<i}}[h(q(B_{<i}))] + \mathbb{E}_{B_{<i}}[q(B_{<i})] \cdot \mathbb{E}_{A_{<i}}[h(p(A_{<i}))]\Big) \\ &= \frac{1}{2\varphi}\Big(\Pr[A_i = 0] \cdot H(B_i | B_{<i}) + \Pr[B_i = 0] \cdot H(A_i | A_{<i})\Big). \end{aligned}$$

Using the assumption that $\Pr[A_i = 0] \geq p$ and $\Pr[B_i = 0] \geq p$ gives

$$H(A_i \cup B_i) \geq \frac{p}{2\varphi}\Big(H(A_i | A_{<i}) + H(B_i | B_{<i})\Big).$$

The claim follows by summing over $i \in [n]$. □

Proof of Theorem 8.2.3. Let $\mathcal{F}$ be a $(1-\varepsilon)$-approximate union closed family over $[n]$. Let $p = \min_{i \in [n]} \Pr_{A \in \mathcal{F}}[A_i = 0]$, where our goal is to lower bound $1 - p$. Let $A, B \in \mathcal{F}$ be uniformly and independently chosen. Claim 8.5.1 then gives

$$H(A \cup B) \geq \frac{p}{2\varphi}\Big(H(A) + H(B)\Big) = \frac{p}{\varphi} \log |\mathcal{F}|.$$

Next we show that $H(A \cup B)$ cannot be much larger than $\log |\mathcal{F}|$. Let I be the indicator for the event $A \cup B \in \mathcal{F}$, where by assumption $\Pr[I = 1] \geq 1 - \varepsilon$. Then

$$H(A \cup B) \leq H(A \cup B, I) = H(I) + H(A \cup B | I = 0) \Pr[I = 0] + H(A \cup B | I = 1) \Pr[I = 1].$$

We bound the terms one by one. First, since I is binary and $\Pr[I = 0] \leq \varepsilon < 1/2$ we have $H(I) \leq h(\varepsilon) \leq 2\varepsilon \log(1/\varepsilon)$. Next, when $I = 0$, we use the naive bound $H(A \cup B | I = 0) \leq H(A, B | I = 0) \leq 2 \log |\mathcal{F}|$. Finally, when $I = 1$ we have that

$A \cup B | I = 1$ is a distribution supported on $\mathcal{F}$ and so $H(A \cup B | I = 1) \leq \log |\mathcal{F}|$. Putting these together gives

$$\frac{p}{\varphi} \log |\mathcal{F}| \leq H(A \cup B) \leq 2\varepsilon \log(1/\varepsilon) + (1 + 2\varepsilon) \log |\mathcal{F}|.$$

We thus obtain

$$1 - p \geq 1 - \varphi - 2\varepsilon \left(1 + \frac{\log(1/\varepsilon)}{\log |\mathcal{F}|}\right).$$

The proof follows, as $1 - \varphi = \frac{3-\sqrt{5}}{2} = \psi$. □

Bibliography

[1] James Alexander, Jonathan Cutler, and Tim Mink. Independent sets in graphs with given minimum degree. *Electron. J. Combin.*, 19(3):Paper 37, 11, 2012.

[2] James Alexander and Tim Mink. A new method for enumerating independent sets of a fixed size in general graphs. *J. Graph Theory*, 81(1):57–72, 2016.

[3] Ryan Alweiss, Brice Huang, and Mark Sellke. Improved lower bound for the union-closed sets conjecture. *arXiv preprint arXiv:2211.11731*, 2022.

[4] Juan Arias de Reyna. Gilbreath's conjecture. https://institucional.us.es/blogimus/en/2020/07/gilbreaths-conjecture/.

[5] Frank Ban, Xi Chen, Adam Freilich, Rocco A. Servedio, and Sandip Sinha. Beyond trace reconstruction: population recovery from the deletion channel. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science*, pages 745–768. IEEE Comput. Soc. Press, Los Alamitos, CA, [2019].

[6] Frank Ban, Xi Chen, Rocco A. Servedio, and Sandip Sinha. Efficient average-case population recovery in the presence of insertions and deletions. In *Approximation, randomization, and combinatorial optimization. Algorithms and techniques*, volume 145 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 44, 18. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2019.

[7] Tuğkan Batu, Sampath Kannan, Sanjeev Khanna, and Andrew McGregor. Reconstructing strings from random traces. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 910–918. ACM, New York, 2004.

[8] P. Borwein and T. Erdélyi. Littlewood-type problems on subarcs of the unit circle. *Indiana Univ. Math. J.*, 46(4):1323–1346, 1997.

[9] Peter Borwein, Tamás Erdélyi, and Géza Kós. Littlewood-type problems on $[0, 1]$. *Proc. London Math. Soc. (3)*, 79(1):22–46, 1999.

[10] Joshua Brakensiek, Ray Li, and Bruce Spang. Coded trace reconstruction in a constant number of traces. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*, pages 482–493. IEEE Computer Soc., Los Alamitos, CA, [2020].

[11] Zachary Chase. The maximum number of three term arithmetic progressions, and triangles in cayley graphs. *arXiv preprint arXiv:1809.03729*, 2018.

[12] Zachary Chase. New lower bounds for trace reconstruction. *Ann. Inst. Henri Poincaré Probab. Stat.*, 57(2):627–643, 2021.

[13] Zachary Chase. Separating words and trace reconstruction. In *STOC '21—Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 21–31. ACM, New York, [2021].

[14] Xi Chen, Anindya De, Chin Ho Lee, Rocco A. Servedio, and Sandip Sinha. Polynomial-time trace reconstruction in the smoothed complexity model. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 54–73. [Society for Industrial and Applied Mathematics (SIAM)], Philadelphia, PA, 2021.

[15] Mahdi Cheraghchi, Ryan Gabrys, Olgica Milenkovic, and João Ribeiro. Coded trace reconstruction. *IEEE Trans. Inform. Theory*, 66(10):6084–6103, 2020.

[16] Jonathan Cutler and A. J. Radcliffe. The maximum number of complete subgraphs in a graph with given maximum degree. *J. Combin. Theory Ser. B*, 104:60–71, 2014.

[17] Jonathan Cutler and A. J. Radcliffe. The maximum number of complete subgraphs of fixed size in a graph with given maximum degree. *J. Graph Theory*, 84(2):134–145, 2017.

[18] Sami Davies, Miklós Z. Rácz, and Cyrus Rashtchian. Reconstructing trees from traces. *Ann. Appl. Probab.*, 31(6):2772–2810, 2021.

[19] Anindya De, Ryan O'Donnell, and Rocco A. Servedio. Optimal mean-based algorithms for trace reconstruction. In *STOC'17—Proceedings of the 49th Annual*

ACM SIGACT Symposium on Theory of Computing, pages 1047–1056. ACM, New York, 2017.

[20] Sean Eberhard. The abelian arithmetic regularity lemma. *arXiv preprint arXiv:1606.09303*, 2016.

[21] John Engbers and David Galvin. Counting independent sets of a fixed size in graphs with a given minimum degree. *J. Graph Theory*, 76(2):149–168, 2014.

[22] P. Erdős and R. L. Graham. *Old and new problems and results in combinatorial number theory*, volume 28 of *Monographies de L'Enseignement Mathématique [Monographs of L'Enseignement Mathématique]*. Université de Genève, L'Enseignement Mathématique, Geneva, 1980.

[23] P. Erdős and J. O. Shallit. New bounds on the length of finite Pierce and Engel series. *Sém. Théor. Nombres Bordeaux (2)*, 3(1):43–53, 1991.

[24] David Galvin. Two problems on independent sets in graphs. *Discrete Math.*, 311(20):2105–2112, 2011.

[25] Wenying Gan. *Several Problems in Extremal Combinatorics.* ProQuest LLC, Ann Arbor, MI, 2014. Thesis (Ph.D.)–University of California, Los Angeles.

[26] Wenying Gan, Po-Shen Loh, and Benny Sudakov. Maximizing the number of independent sets of a fixed size. *Combin. Probab. Comput.*, 24(3):521–527, 2015.

[27] Justin Gilmer. A constant lower bound for the union-closed sets conjecture. *arXiv preprint arXiv:2211.09055*, 2022.

[28] P. Goralčík and V. Koubek. On discerning words by automata. In *Automata, languages and programming (Rennes, 1986)*, volume 226 of *Lecture Notes in Comput. Sci.*, pages 116–122. Springer, Berlin, 1986.

[29] Ben Green and Terence Tao. An arithmetic regularity lemma, an associated counting lemma, and applications. In *An irregular mind*, volume 21 of *Bolyai Soc. Math. Stud.*, pages 261–334. János Bolyai Math. Soc., Budapest, 2010.

[30] Nina Holden and Russell Lyons. Lower bounds for trace reconstruction. *Ann. Appl. Probab.*, 30(2):503–525, 2020.

[31] Nina Holden, Robin Pemantle, Yuval Peres, and Alex Zhai. Subpolynomial trace reconstruction for random strings and arbitrary deletion probability. *Math. Stat. Learn.*, 2(3-4):275–309, 2019.

[32] Thomas Holenstein, Michael Mitzenmacher, Rina Panigrahy, and Udi Wieder. Trace reconstruction with constant deletion probability and related results. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 389–398. ACM, New York, 2008.

[33] Kenneth Ireland and Michael Rosen. *A classical introduction to modern number theory*, volume 84 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1990.

[34] Henryk Iwaniec and Emmanuel Kowalski. *Analytic number theory*, volume 53. American Mathematical Soc., 2021.

[35] A. Khalfalah, S. Lodha, and E. Szemerédi. Tight bound for the density of sequence of integers the sum of no two of which is a perfect square. *Discrete Math.*, 256(1-2):243–255, 2002.

[36] Rachel Kirsch and A. J. Radcliffe. Many triangles with few edges. *Electron. J. Combin.*, 26(2):Paper No. 2.36, 23, 2019.

[37] Akshay Krishnamurthy, Arya Mazumdar, Andrew McGregor, and Soumyabrata Pal. Trace reconstruction: generalized and parameterized. *IEEE Trans. Inform. Theory*, 67(6, part 1):3233–3250, 2021.

[38] J. C. Lagarias, A. M. Odlyzko, and J. B. Shearer. On the density of sequences of integers the sum of no two of which is a square. I. Arithmetic progressions. *J. Combin. Theory Ser. A*, 33(2):167–185, 1982.

[39] J. C. Lagarias, A. M. Odlyzko, and J. B. Shearer. On the density of sequences of integers the sum of no two of which is a square. II. General sequences. *J. Combin. Theory Ser. A*, 34(2):123–139, 1983.

[40] Hiu-Fai Law and Colin McDiarmid. On independent sets in graphs with given minimum degree. *Combin. Probab. Comput.*, 22(6):874–884, 2013.

[41] Allan Lo. Cliques in graphs with bounded minimum degree. *Combin. Probab. Comput.*, 21(3):457–482, 2012.

[42] Neil Lyall. A new proof of Sárközy's theorem. *Proc. Amer. Math. Soc.*, 141(7):2253–2264, 2013.

[43] J.P. Massias. Sur les suites dont les sommes des terms deux a deux ne sont pas des carrés. *Publications du Département de Mathématiques de Limoges*, 1982.

[44] Kaisa Matomäki, Maksym Radziwiłł, and Terence Tao. Correlations of the von Mangoldt and higher divisor functions I. Long shift ranges. *Proc. Lond. Math. Soc. (3)*, 118(2):284–350, 2019.

[45] Hugh L. Montgomery. *Ten lectures on the interface between analytic number theory and harmonic analysis*, volume 84 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, 1994.

[46] Shyam Narayanan. Improved algorithms for population recovery from the deletion channel. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1259–1278. [Society for Industrial and Applied Mathematics (SIAM)], Philadelphia, PA, 2021.

[47] Shyam Narayanan and Michael Ren. Circular trace reconstruction. In *12th Innovations in Theoretical Computer Science Conference*, volume 185 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 18, 18. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2021.

[48] Fedor Nazarov and Yuval Peres. Trace reconstruction with $\exp(O(n^{1/3}))$ samples. In *STOC'17—Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1042–1046. ACM, New York, 2017.

[49] Andrew M. Odlyzko. Iterated absolute values of differences of consecutive primes. *Math. Comp.*, 61(203):373–380, 1993.

[50] Théophile Pépin. Sur la formule $2^{2^n}+1$. *CR Acad. Sci. Paris*, 85:329–331, 1877.

[51] Yuval Peres and Alex Zhai. Average-case reconstruction for the deletion channel: subpolynomially many traces suffice. In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, pages 228–239. IEEE Computer Soc., Los Alamitos, CA, 2017.

[52] T. A. Pierce. On an Algorithm and Its Use in Approximating Roots of Algebraic Equations. *Amer. Math. Monthly*, 36(10):523–525, 1929.

[53] E Proth. Theoremes sur les nombres premiers, c'. r. *Acad. Sci. Paris Ser. I*, 1877.

[54] François Proth. Sur la série des nombres premiers. *Nouv. Corresp. Math*, 4:236–240, 1878.

[55] J. M. Robson. Separating strings with small automata. *Inform. Process. Lett.*, 30(4):209–214, 1989.

[56] J. M. Robson. Separating words with machines and groups. *RAIRO Inform. Théor. Appl.*, 30(1):81–86, 1996.

[57] A. D. Scott. Reconstructing sequences. *Discrete Math.*, 175(1-3):231–238, 1997.

[58] J. O. Shallit. Metric theory of Pierce expansions. *Fibonacci Quart.*, 24(1):22–40, 1986.

[59] M. N. Vyalyĭ and R. A. Gimadeev. On separating words by the occurrences of subwords. *Diskretn. Anal. Issled. Oper.*, 21(1):3–14, 106, 2014.

[60] Hugh Williams. Personal communication.

[61] Hugh C. Williams. *Édouard Lucas and primality testing*, volume 22 of *Canadian Mathematical Society Series of Monographs and Advanced Texts*. John Wiley & Sons, Inc., New York, 1998. A Wiley-Interscience Publication.