On Zaremba's conjecture

By Jean Bourgain and Alex Kontorovich

Abstract

Zaremba's 1971 conjecture predicts that every integer appears as the denominator of a finite continued fraction whose partial quotients are bounded by an absolute constant. We confirm this conjecture for a set of density one.

Contents

1. Introduction

1.1. *Statements of the main theorems.* For a fixed finite set $\mathcal{A} \subset \mathbb{N}$, which we call an *alphabet*, let $\mathfrak{C}_\mathcal{A}$ denote the collection of all $x \in (0,1)$ whose continued fraction expansion

$$x = [a_1, a_2, \ldots, a_k, \ldots] = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_k + \ddots}}},$$

has all partial quotients a_j belonging to the alphabet $\mathcal{A}$. Any $x \in \mathfrak{C}_\mathcal{A}$ is uniformly badly approximable, in the sense that its partial quotients a_j are all bounded by

$$A := \max \mathcal{A}.$$

When A is an absolute constant, we call such a number *absolutely Diophantine* (of height A). That is, when we speak of numbers being absolutely Diophantine, the height A is fixed in advance.

Let $\mathfrak{R}_\mathcal{A}$ denote the set of partial convergents to $\mathfrak{C}_\mathcal{A}$, that is,

$$\mathfrak{R}_\mathcal{A} := \left\{ \frac{b}{d} = [a_1, a_2, \ldots, a_k] : 0 < b < d, \ (b, d) = 1, \text{ and } \forall j, \ a_j \in \mathcal{A} \right\},$$

and let $\mathfrak{D}_\mathcal{A} \subset \mathbb{N}$ be the set of denominators of fractions in $\mathfrak{R}_\mathcal{A}$,

$$\mathfrak{D}_\mathcal{A} := \left\{ d \in \mathbb{N} : \exists (b, d) = 1 \text{ with } \frac{b}{d} \in \mathfrak{R}_\mathcal{A} \right\}.$$

In 1971, S. K. Zaremba formulated the following assertion.

CONJECTURE 1.1 (Zaremba [Zar72, p. 76]). *Every positive integer is the denominator of a reduced absolutely Diophantine fraction.*

That is, the conjecture predicts the existence of some integer $A > 1$ so that

$$\mathfrak{D}_{\{1,2,\ldots,A\}} = \mathbb{N}.$$

In fact, Zaremba suggested that $A = 5$ may already be sufficient.

Zaremba's conjecture has important applications to numerical integration and pseudorandom number generation, producing collections of points of optimal discrepancy; see, e.g., the surveys [Nie78] and [Kon13]. Our main result is the following

THEOREM 1.2. *Almost every positive integer is the denominator of a reduced absolutely Diophantine fraction. That is, there exists an $A > 1$ so that*

$$\frac{1}{N} \#(\mathfrak{D}_{\{1,2,\ldots,A\}} \cap [1, N]) \to 1,$$

as $N \to \infty$. In particular, $A = 50$ suffices.

A more refined conjecture was stated by Hensley in 1996. The set $\mathfrak{C}_\mathcal{A} \subset (0, 1)$ is a Cantor-like fractal; let

$$\delta_\mathcal{A} := \text{H.dim}(\mathfrak{C}_\mathcal{A}) \ \in \ [0, 1]$$

be its Hausdorff dimension. This dimension can be 0 only if $|\mathcal{A}| = 1$; since we assume $\mathcal{A}$ is finite, $\delta_\mathcal{A} < 1$. Allowing a finite number of exceptions in Zaremba's conjecture, Hensley asserts the following

CONJECTURE 1.3 (Hensley [Hen96, Conj. 3, p. 16]). *The set of denominators $\mathfrak{D}_{\mathcal{A}}$ contains every sufficiently large integer if and only if the corresponding dimension $\delta_{\mathcal{A}}$ exceeds $1/2$.*

As stated, Hensley's conjecture is false. For example, consider the alphabet $\mathcal{A} = \{2, 4, 6, 8, 10\}$. By implementing an algorithm due to Jenkinson and Pollicott [JP01],[1] we have estimated its dimension to be $\delta_{\mathcal{A}} \approx 0.517 > 1/2$. Nevertheless, arbitrarily large numbers are missing from $\mathfrak{D}_{\mathcal{A}}$. Indeed, it is elementary to verify that

(1.4) $\mathfrak{D}_{\mathcal{A}} (\mathrm{mod}\, 4) \equiv \{0, 1, 2\};$

see Remark 1.33.

We propose the following alternative to Hensley's conjecture, borrowing language from Hilbert's 11th problem on representations of numbers by quadratic forms. We call an integer d *admissible* (for $\mathcal{A}$) if it passes all finite local obstructions:

(1.5) $\forall q > 1,\ d \in \mathfrak{D}_{\mathcal{A}} (\mathrm{mod}\, q).$

Remark 1.6. Admissibility can be checked using only one modulus $q = q(\mathcal{A})$; see Remark 1.33.

Let $\mathfrak{A}_{\mathcal{A}}$ denote the set of all admissible numbers,

$$\mathfrak{A}_{\mathcal{A}} := \{d \in \mathbb{Z} : (1.5) \text{ holds}\}.$$

We say d is *represented* (by $\mathcal{A}$) if $d \in \mathfrak{D}_{\mathcal{A}}$. The *multiplicity* of a denominator d is the number of coprime numerators $0 < b < d$ with $b/d \in \mathfrak{R}_{\mathcal{A}}$. Clearly d is represented if and only if its multiplicity is positive.

CONJECTURE 1.7. *If the dimension $\delta_{\mathcal{A}}$ exceeds $1/2$, then the set of denominators $\mathfrak{D}_{\mathcal{A}}$ contains every sufficiently large admissible integer.*

We interpret this conjecture as a local-global principle, where the dimension condition and "sufficiently large" are local obstructions at infinity. Theorem 1.2 follows from the following more refined approximation to Conjecture 1.7.

THEOREM 1.8. *There exists some $\delta_0 < 1$ (see (1.21)) so that, if the dimension $\delta_{\mathcal{A}}$ exceeds δ_0, then the set of denominators $\mathfrak{D}_{\mathcal{A}}$ contains almost every admissible integer. More precisely, there is a constant $c = c(\mathcal{A}) > 0$ so that*

(1.9) $\dfrac{\#(\mathfrak{D}_{\mathcal{A}} \cap [N/2, N])}{\#(\mathfrak{A}_{\mathcal{A}} \cap [N/2, N])} = 1 + O\left(e^{-c\sqrt{\log N}}\right),$

---

[1] The program is available at http://math.yale.edu/~avk23/maths/HausdorffZaremba.nb.

as $N \to \infty$. *Furthermore, each d produced above appears with multiplicity*

$$(1.10) \qquad\qquad multiplicity(d) \gg N^{2\delta_{\mathcal{A}} - \frac{1001}{1000}}.$$

The constant c in (1.9) is effectively computable, and the implied constants above depend only on $\mathcal{A}$.

Some remarks are in order.

Remark 1.11. There exist alphabets $\mathcal{A}$ with $\delta_{\mathcal{A}}$ arbitrarily close to 1, so Theorem 1.8 is not vacuous. Indeed, Hensley [Hen92] gives the asymptotic expansion

$$(1.12) \qquad \delta_{\{1,2,\dots,A\}} = 1 - \frac{6}{\pi^2 A} - \frac{72 \log A}{\pi^4 A^2} + O\left(\frac{1}{A^2}\right).$$

Remark 1.13. The number $1/2$ in Conjectures 1.3 and 1.7 cannot be reduced. Hensley [Hen89] showed that the truncated set of rationals

$$\mathfrak{R}_{\mathcal{A}}(N) := \left\{ \frac{b}{d} \in \mathfrak{R}_{\mathcal{A}} : (b,d) = 1, \ 0 < b < d < N \right\}$$

has cardinality

$$(1.14) \qquad\qquad \#\mathfrak{R}_{\mathcal{A}}(N) \asymp N^{2\delta_{\mathcal{A}}},$$

whence it follows immediately that

$$\#(\mathfrak{D}_{\mathcal{A}} \cap [1, N]) \ll N^{2\delta_{\mathcal{A}}}.$$

Thus if $\delta_{\mathcal{A}} < 1/2$, then certainly $\mathfrak{D}_{\mathcal{A}}$ is too thin a subset of the integers to contain even one admissible arithmetic progression.

Remark 1.15. The best previously known estimate,

$$(1.16) \qquad\qquad \#(\mathfrak{D}_{\mathcal{A}} \cap [1, N]) \gg N^{\delta_{\mathcal{A}}},$$

was proved by Hensley [Hen06, Th. 3.2] and follows easily from his estimate (1.14). In particular, as long as $|\mathcal{A}| > 1$, the set $\mathfrak{D}_{\mathcal{A}}$ grows at least polynomially. Moreover, taking $\mathcal{A}$ large so that $\delta_{\mathcal{A}} > 1 - \varepsilon$, one can already produce at least $N^{1-\varepsilon}$ denominators in $\mathfrak{D}_{\mathcal{A}}$ up to N.

Remark 1.17. We explain in Remark 1.33 below that for any $A \geq 2$, the alphabet $\{1, 2, \dots, A\}$ has no finite local obstructions; that is, $\mathfrak{A}_{\{1,2,\dots,A\}} = \mathbb{Z}$. This is why the statement of Theorem 1.2 needs no mention of admissibility. Moreover, the dimension $\delta_{\{1,2\}}$ is known [Goo41], [Bum85], [JP01] to be

$$(1.18) \qquad\qquad \delta_{\{1,2\}} \approx 0.531 \cdots,$$

which obviously exceeds $1/2$. Conjecture 1.7 then implies that $\mathfrak{D}_{\{1,2\}}$ already contains every sufficiently large number, as was conjectured by Hensley [Hen96].

Remark 1.19. An earlier version[2] of this paper also proved two weaker results made obsolete by Theorem 1.2, namely that for sufficiently large $\delta_\mathcal{A}$,

(i) $\mathfrak{D}_\mathcal{A}$ contains a positive proportion of numbers, and

(ii) that $\mathfrak{D}_\mathcal{A}$ contains almost every admissible number, without giving the rate in (1.9).

At the request of the referee to shorten the paper, we have removed these intermediary results (and of course the methods used to obtain them). We invite the interested reader to peruse the original arXiv posting for the details. Note also that some results of this paper have been announced in [BK11].

Remark 1.20. The value of δ_0 in Theorem 1.8 coming from our proof is

$$(1.21) \qquad\qquad \delta_0 = 307/312 \approx 0.984.$$

We have made no effort to optimize this quantity, as can surely be done with a modicum of effort. In fact, Frolenkov and Kan [FK13] have since sharpened our method to prove the weaker statement[3] that $\mathfrak{D}_\mathcal{A}$ contains a positive proportion of numbers whenever $\delta_\mathcal{A} > \delta_0$ with the improved range $\delta_0 = 5/6 \approx 0.833$. It does not seem likely that these methods can achieve the full range $\delta_0 = 1/2$ without significant new ideas. We have estimated the dimension $\delta_\mathcal{A}$ corresponding to the alphabet $\mathcal{A} = \{1, 2, \ldots, 49, 50\}$ to be about 0.986, exceeding (1.21), whereas the alphabet $\mathcal{A} = \{1, 2, 3, 4, 5\}$ is known [Jen04] to have dimension $\delta_\mathcal{A} > 5/6$.

Remark 1.22. The rate in (1.9) can be improved to a power savings, that is, an error term of the form $O(N^{-c})$; see [Bou12].

Although our main result requires large dimension, we are also able to sharpen the best previously known estimate (1.16) in the full range $\delta_\mathcal{A} > 1/2$.

THEOREM 1.23. *Write δ for $\delta_\mathcal{A}$. Then for any $\varepsilon > 0$,*

$$(1.24) \qquad\qquad \#(\mathfrak{D}_\mathcal{A} \cap [1, N]) \gg_\varepsilon N^{\delta + \frac{(2\delta-1)(1-\delta)}{5-\delta} - \varepsilon},$$

as $N \to \infty$. This bound improves on (1.16), as long as $\delta > 1/2$.

Remark 1.25. The improvement here is quite modest: for $\mathcal{A} = \{1, 2\}$, the exponent $\delta_\mathcal{A} \approx 0.531$ in (1.18) and (1.16) is replaced in (1.24) by 0.537. We have again made no attempt to optimize the exponent in (1.24), seeking just any power gain.

We state the multiplicity bound (1.10) to give another application to pseudorandom numbers. Specifically, in the (homogeneous) linear congruential

---

[2]http://arxiv.org/abs/1107.3776v1

[3]Added in press: Huang [Hua13] has combined the methods of Frolenkov-Kan with ours to obtain the full density one statement with $\delta_0 = 5/6$.

method, optimal conditions require a prime d and a primitive root $b(\mathrm{mod}\, d)$ so that the fraction b/d is absolutely Diophantine (see [Kon13]). Then the pseudorandom map with modulus d and multiplier b, that is, $x \mapsto bx(\mathrm{mod}\, d)$, has asymptotically optimal serial correlation of pairs.

THEOREM 1.26. *There exist infinitely many primes d with primitive roots $b(\mathrm{mod}\, d)$ so that the fractions b/d are absolutely Diophantine.*

The number of such prime d up to N provided by our proof is on the order of $N(\log N)^{-2}$. Theorem 1.26 is an easy corollary of Theorem 1.8. In fact, if $\mathcal{A} = \{1, 2, \ldots, A\}$ has dimension $\delta_{\mathcal{A}}$ exceeding δ_0 as in Theorem 1.8, then the fractions b/d produced in Theorem 1.26 can be taken to have all partial quotients bounded by $A + 1$.

1.2. *Reformulation and admissibility.* It is an old and trivial (but for our purposes crucial) observation that

$$\frac{b}{d} = [a_1, \ldots, a_k]$$

is equivalent to

$$(1.27) \qquad \begin{pmatrix} * & b \\ * & d \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & a_1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & a_2 \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & a_k \end{pmatrix}.$$

This observation will allow us to explain all local obstructions, as follows. In light of (1.27), let

$$\mathcal{G}_{\mathcal{A}} \subset \mathrm{GL}(2, \mathbb{Z})$$

be the semigroup generated by the matrices

$$(1.28) \qquad \qquad \qquad \begin{pmatrix} 0 & 1 \\ 1 & a \end{pmatrix}$$

for $a \in \mathcal{A}$. Then the orbit

$$(1.29) \qquad \qquad \qquad \mathcal{O}_{\mathcal{A}} := \mathcal{G}_{\mathcal{A}} \cdot e_2$$

of $e_2 = (0, 1)^t$ under $\mathcal{G}_{\mathcal{A}}$ corresponds to $\mathfrak{R}_{\mathcal{A}}$; that is, if $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then $\gamma \cdot e_2 = (b, d)^t$. Moreover, taking the inner product of this orbit with e_2 picks off the value of d, that is $\langle \gamma \cdot e_2, e_2 \rangle = d$, and

$$(1.30) \qquad \qquad \qquad \langle \mathcal{O}_{\mathcal{A}}, e_2 \rangle = \langle \mathcal{G}_{\mathcal{A}} \cdot e_2, e_2 \rangle$$

is precisely $\mathfrak{D}_{\mathcal{A}}$ (with multiplicity). Zaremba's conjecture can then be reformulated as: For some finite alphabet $\mathcal{A}$,

$$\mathbb{N} \subset \langle \mathcal{G}_{\mathcal{A}} \cdot e_2, e_2 \rangle.$$

For convenience we pass from $\mathcal{G}_{\mathcal{A}}$ to its determinant one subsemigroup

$$(1.31) \qquad \qquad \Gamma_{\mathcal{A}} = \mathcal{G}_{\mathcal{A}} \cap \mathrm{SL}_2 \subset \mathrm{SL}_2(\mathbb{Z}),$$

which is (freely and finitely) generated by the matrix products

$$(1.32) \qquad \begin{pmatrix} 0 & 1 \\ 1 & a \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ 1 & a' \end{pmatrix}$$

for $a, a' \in \mathcal{A}$. The orbit $\mathcal{O}_\mathcal{A}$ is recovered as a finite union of "coset" orbits

$$\mathcal{O}_\mathcal{A} = \Gamma_\mathcal{A} \cdot e_2 \ \cup \ \bigcup_{a \in \mathcal{A}} \Gamma_\mathcal{A} \cdot \begin{pmatrix} 0 & 1 \\ 1 & a \end{pmatrix} e_2.$$

Remark 1.33. It now follows from Strong Approximation [MVW84] and Goursat's Lemma (see the discussion in [Kon13, §2.2]) that the reduction of $\Gamma_\mathcal{A}$ mod q is all of $\mathrm{SL}_2(q)$ for all q coprime to a certain "bad" modulus $\mathfrak{B}$. Here $\mathfrak{B}$ is effectively computable and depends only on $\Gamma_\mathcal{A}$, that is, on $\mathcal{A}$. Moreover, $\mathfrak{B}$ can be chosen so that for any $q \equiv 0(\mathfrak{B})$, the reduction $\Gamma_\mathcal{A}(\mathrm{mod}\, q)$ is the full pre-image of $\Gamma_\mathcal{A}(\mathrm{mod}\, \mathfrak{B})$ under the projection map $\mathbb{Z}/q \to \mathbb{Z}/\mathfrak{B}$. From the mod $\mathfrak{B}$ reductions of $\Gamma_\mathcal{A}$, it is elementary to read off the reductions of $\mathcal{O}_\mathcal{A}$, and hence all finite local obstructions in $\mathfrak{D}_\mathcal{A}$; see Remark 1.6. Moreover, for the alphabet $\mathcal{A} = \{1, 2\}$, it is easy to see that $\mathfrak{B} = 1$, that is, $\Gamma_\mathcal{A}(\mathrm{mod}\, q)$ is already all of $\mathrm{SL}_2(q)$, for all $q > 1$; see Remark 1.17. Indeed, the *group* generated by $\Gamma_\mathcal{A}$ (that is, allowing inverses) is all of $\mathrm{SL}_2(\mathbb{Z})$, and the two have the same projections mod q. Finally, we note that this is precisely the phenomenon responsible for (1.4) in the failure of Hensley's Conjecture 1.3.

1.3. *An overview of the key ideas.* An observation which we had made in a slightly different context [BK10] is that there is a certain bilinear (in fact multilinear) structure to (1.30), making the problem amenable to the Hardy-Littlewood circle method via Vinogradov's techniques for estimating bilinear forms. We now outline the key steps.

From now on, we treat $\mathcal{A}$ as fixed, dropping it from subscripts, writing $\delta = \delta_\mathcal{A}$, $\Gamma = \Gamma_\mathcal{A}$, etc. In light of (1.30), we would like to study the exponential sum

$$(1.34) \qquad S_N(\theta) := \sum_{\substack{\gamma \in \Gamma \\ \|\gamma\| < N}} e(\theta \langle \gamma e_2, e_2 \rangle),$$

where $\theta \in [0, 1]$ and $\| \cdot \|$ is the Frobenius matrix norm, $\|\begin{pmatrix} a & b \\ c & d \end{pmatrix}\|^2 = a^2 + b^2 + c^2 + d^2$. Then the Fourier coefficient

$$(1.35) \qquad R_N(d) := \widehat{S}_N(d) = \int_0^1 S_N(\theta) e(-d\theta) d\theta = \sum_{\substack{\gamma \in \Gamma \\ \|\gamma\| < N}} \mathbf{1}_{\{\langle \gamma e_2, e_2 \rangle = d\}}$$

is just the "representation number" of d up to N, that is, its multiplicity. Of course if $R_N(d) > 0$, then $d \in \mathfrak{D}$.

Note that by (1.14),

$$(1.36) \qquad S_N(0) = \sum_d R_N(d) = \sum_{\substack{\gamma \in \Gamma \\ \|\gamma\| < N}} 1 \; \asymp \; N^{2\delta},$$

so if almost every $d \in [N/2, N]$ is to be represented without much bias, it should occur with multiplicity roughly $N^{2\delta-1}$.

Following the circle method, we decompose the integral in (1.35) into "major arcs" and "minor arcs," the former referring to frequencies θ quite near rationals with small denominators and the latter being the rest:

$$R_N(d) = \left(\int_{\mathfrak{M}} + \int_{[0,1]\setminus\mathfrak{M}} \right) S_N(\theta)e(-d\theta)d\theta = \mathcal{M}_N(d) + \mathcal{E}_N(d).$$

Here $\mathcal{M}_N$ is thought of as a "main" term and $\mathcal{E}_N$ is an "error" term, and the major arcs $\mathfrak{M} = \mathfrak{M}_\mathcal{Q}$ are given by

$$(1.37) \qquad \mathfrak{M}_\mathcal{Q} = \bigcup_{q < \mathcal{Q}} \bigcup_{(a,q)=1} \left[\frac{a}{q} - \frac{\mathcal{Q}}{N}, \frac{a}{q} + \frac{\mathcal{Q}}{N} \right],$$

where $\mathcal{Q}$ is roughly of size $e^{-c\sqrt{\log N}}$.

A key ingredient is to show that along the major arcs, $\theta = \frac{a}{q} + \beta \in \mathfrak{M}$, the function S_N essentially splits into two pieces,

$$(1.38) \qquad S_N\left(\frac{a}{q} + \beta \right) \sim \nu_q(a) \cdot \varpi(\beta).$$

Here ν_q is a purely modular term and ϖ is an archimedean one, which has the right order of magnitude on balls of certain size. It then follows that the main term $\mathcal{M}_N(d)$ also splits as a "singular series" $\mathfrak{S}$ times a "singular integral" Π,

$$\mathcal{M}_N(d) \sim \mathfrak{S}(d)\Pi_N(d),$$

where Π gives the expected archimedean contribution, roughly

$$\Pi_N(d) \gg N^{2\delta-1}$$

for $d \asymp N$, and the singular series $\mathfrak{S}$ controls the local obstructions. In particular, if $d \notin \mathfrak{A}$ is not admissible, then $\mathfrak{S}(d) = 0$; otherwise, we have roughly that

$$\mathfrak{S}(d) \asymp \prod_{p \nmid d} \left(1 + \frac{1}{p^2 - 1} \right) \prod_{p \mid d} \left(1 - \frac{1}{p+1} \right) \gg \frac{1}{\log \log d}.$$

The main ingredient in proving this is the renewal method in the thermodynamic formalism of Ruelle transfer operators (see Lalley [Lal89]), and the extension to "congruence" such established by Bourgain-Gamburd-Sarnak in [BGS11]. (We need here not just square-free but arbitrary moduli q, and we must also use the work of Bourgain-Varjú [BV12].)

With the major arcs controlled, if we could prove that the errors are individually bounded, $|\mathcal{E}_N(d)| \ll N^{2\delta-1-\varepsilon}$, say, then we would conclude the full

Conjecture 1.7. We are not able to establish control of this quality individually, but do succeed on average, proving essentially that

$$(1.39) \qquad \sum_{d \asymp N} |\mathcal{E}_N(d)|^2 \ll N^{4\delta-1} e^{-c\sqrt{\log N}},$$

from which Theorem 1.8 follows by a standard argument.

Bounds of this type will follow from bounds on

$$(1.40) \qquad \int_{W_{Q,K}} |S_N(\theta)|^2 d\theta,$$

where we have decomposed the minor arcs $[0,1] \setminus \mathfrak{M}$ into the dyadic regions

$$(1.41) \qquad W_{Q,K} := \left\{ \theta = \frac{a}{q} + \beta : q \asymp Q, (a,q) = 1, |\beta| \asymp \frac{K}{N} \right\}.$$

By Dirichlet's approximation theorem, the parameters Q and K vary in the range $Q < N^{1/2}$ and $K < \frac{N^{1/2}}{Q}$.

Unfortunately, we do not know how to obtain such strong bounds for the function S_N as defined in (1.34). But taking a cue from Vinogradov (as we did in [BK10]), we work with a different function:

$$(1.42) \qquad S_N(\theta) = \sum_{\substack{\gamma_1 \in \Gamma \\ \|\gamma_1\| \asymp N^{1/2}}} \sum_{\substack{\gamma_2 \in \Gamma \\ \|\gamma_2\| \asymp N^{1/2}}} e(\theta \langle \gamma_1 \gamma_2 e_2, e_2 \rangle),$$

say. Since Γ is a semigroup, this modified function, or rather its Fourier transform, continues to capture elements of $\mathfrak{D}$. Moreover, the bilinear nature of the problem, namely that $\langle \gamma_1 \gamma_2 e_2, e_2 \rangle = \langle \gamma_2 e_2, {}^t\gamma_1 e_2 \rangle$, allows us to separate variables.

It is here in the separation of variables and application of Cauchy-Schwarz that we replace the thin semigroup Γ with all of $\mathrm{SL}_2(\mathbb{Z})$, a loss we can only tolerate if the dimension δ is large, at least some δ_0. We are then lead to a more classical setting, and in certain large ranges of the pair (Q, K) in (1.41), we can obtain the requisite cancelation. For slightly smaller values of (Q, K), it is beneficial to decompose the sum further as

$$S_N(\theta) = \sum_{\substack{\gamma_1 \in \Gamma \\ \|\gamma_1\| \asymp N^{1/2}}} \sum_{\substack{\gamma_2 \in \Gamma \\ \|\gamma_2\| \asymp N^{1/4}}} \sum_{\substack{\gamma_3 \in \Gamma \\ \|\gamma_3\| \asymp N^{1/4}}} e(\theta \langle \gamma_1 \gamma_2 \gamma_3 e_2, e_2 \rangle).$$

Continuing in this way, we handle every conceivable range of (Q, K) by considering a sum of the form

$$(1.43) \qquad S_N(\theta) = \sum_{\substack{\gamma_1 \in \Gamma \\ \|\gamma_1\| \asymp N^{1/2}}} \sum_{\substack{\gamma_2 \in \Gamma \\ \|\gamma_2\| \asymp N^{1/4}}} \cdots \sum_{\substack{\gamma_J \in \Gamma \\ \|\gamma_J\| \asymp N^{1/2^J}}} e(\theta \langle \gamma_1 \gamma_2 \cdots \gamma_J e_2, e_2 \rangle),$$

where $J \asymp \log \log N$, so that γ_J is of large but constant size (independent of N).[4] Unfortunately, another problem has crept up: we can no longer control the size of the long product $\gamma_1 \cdots \gamma_J$, which could have norm as large as $N(\log N)^C$.

To remedy this situation, we develop a bit of elementary linear algebra for Γ, showing that if the expanding vectors of two matrices are close, then their eigenvalues behave nearly multiplicatively. This forces us to concoct, for each $j = 1, \ldots, J$, a certain special subset $\Xi_j \subset \{\gamma \in \Gamma : \|\gamma\| \asymp N^{1/2^j}\}$, all the elements of which have expanding eigenvectors pointing near a common direction (independent of j). We then simply use the pigeonhole principle to make sure all elements of Ξ_j have almost the same eigenvalues. Moreover, we need to ensure that the representation $\gamma = \gamma_1 \gamma_2 \cdots \gamma_J$ in (1.43) is unique; that is, if

$$\gamma_1 \gamma_2 \cdots \gamma_J = \gamma_1' \gamma_2' \cdots \gamma_J'$$

with $\gamma_j, \gamma_j' \in \Xi_j$, then $\gamma_j = \gamma_j'$ for all j. We do this by forcing each $\gamma_j \in \Xi_j$ to have the same size in the wordlength metric, again by pigeonhole.

Then the large product ensemble

$$\Xi_1 \cdot \Xi_2 \cdots \Xi_J$$

is a good substitute for $\{\gamma \in \Gamma : \|\gamma\| \asymp N\}$ to handle the minor arcs. Unfortunately, the concocted sets Ξ_j are no longer amenable to the major arc methods! We rectify this by constructing a certain tiny set $\aleph \subset \Gamma$ with good modular/archimedean distribution properties and by prepending it to the product, forming

$$(1.44) \qquad\qquad\qquad \Omega_N = \aleph \, \tilde{\Xi}_1 \Xi_2 \cdots \, \Xi_J.$$

Here the size of Ξ_1 has been cut down a bit to $\tilde{\Xi}_1$ to make room for the set $\aleph$.

The "correct" definition of $S_N(\theta)$ is then to replace (1.34) by

$$(1.45) \qquad\qquad S_N(\theta) := \sum_{\gamma \in \Omega_N} e(\theta \langle \gamma e_2, e_2 \rangle),$$

from which the argument follows as described above. In the end, we prove Theorem 1.8, and hence Theorem 1.2. As already mentioned, Theorem 1.26 is an easy corollary to Theorem 1.8.

Remark 1.46. One may ponder the flexibility of our methods in applications to other problems. For one, in particular, McMullen [McM09] has formulated the problem of producing many closed geodesics in a compact subset of the modular surface, defined over a fixed real quadratic number field $\mathbb{Q}(\sqrt{f})$. This is the same as producing many elements $\gamma \in \mathcal{G}$ so that $\operatorname{tr}(\gamma)^2 - 4$ has square-free part f. Specifically, McMullen asks whether there is a finite

---

[4]We could take J even a bit smaller, but choose not to for the sake of exposition.

alphabet $\mathcal{A}$ so that the set of traces in $\mathcal{G}_{\mathcal{A}}$ contains every sufficiently large admissible integer. Our use of Vinogradov's bilinear estimates relies crucially on the structure in (1.30) and does not apply as it stands to the problem of traces. We plan to return to this problem in the future.

The proof of Theorem 1.23 follows along completely different lines, and is inspired by the recent advances in projection theorems [Bou10]. The observation here is that the set $\mathfrak{D}$ has a certain "sum-set" structure. Namely, if $b/d \in \mathfrak{R}$ is a reduced fraction and $a \in \mathcal{A}$, then clearly

$$(1.47) \qquad\qquad \frac{1}{a + \frac{b}{d}} = \frac{d}{b + ad} \in \mathfrak{R}.$$

This implies that $b + ad \in \mathfrak{D}$ whenever $b/d \in \mathfrak{R}$ and $a \in \mathcal{A}$; we exploit this sum-set structure to produce the bound (1.24). We note further that the Discretized Ring Theorem [Bou03] can be used to get an exponent gain over the lower bound (1.16) even when $\delta \leq 1/2$.

1.4. *Outline of the paper.* In Section 2, we study the multiplicative properties of expanding eigen-values and -vectors for matrices in Γ. We use Section 3 to construct the main ensemble Ω_N, reserving the construction of the leading set $\aleph$ for Section 8. The major arc analysis is carried out in Section 4, while the minor arc bounds are proved in Sections 5–6. Theorem 1.8 is then proved in Section 7, as is its corollary, Theorem 1.26. Lastly, we prove Theorem 1.23 in Section 9.

Notation. Throughout we use the following standard notation. We write $f \sim g$ to mean $f/g \to 1$. We use the Landau/Vinogradov notation $f = O(g)$ and $f \ll g$ synonymously to mean there exists an implied constant $C > 0$ such that for x sufficiently large, $f(x) \leq Cg(x)$. Moreover, $f \asymp g$ denotes $f \ll g \ll f$. We allow the implied constants to depend at most on the fixed alphabet $\mathcal{A}$, unless otherwise specified. We also use the shorthand $e(x) = e^{2\pi i x}$. The cardinality of a finite set S is denoted both as $\#S$ and $|S|$, and the Lebesgue measure of an interval $\mathcal{I}$ is also $|\mathcal{I}|$. Throughout there are some constants $c, C > 0$ which may change from line to line.

Acknowledgements. We thank Curt McMullen for bringing this problem to our attention, and Doug Hensley and Peter Sarnak for many helpful comments and suggestions regarding this work.

2. **Large matrix products**

In this section, we develop some tools in large matrix products, reminiscent of the avalanche principle; see, e.g., [Bou05, Ch. 6] or [GS01, §2]. Recall that $\Gamma = \Gamma_{\mathcal{A}}$ is the semigroup generated by even words in the matrices $a \in \mathcal{A}$. An

easy induction shows that for $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix} \right) \in \Gamma$, $\gamma \neq I$, we have

$$1 \leq a \leq \min(b, c) \leq \max(b, c) < d.$$

We use the Frobenius norm:

$$(2.1) \qquad \qquad \|\gamma\| := \sqrt{a^2 + b^2 + c^2 + d^2}.$$

Note that the trace and norm are comparable up to constants:

$$(2.2) \qquad \qquad \frac{1}{2}\|\gamma\| \leq \operatorname{tr}\gamma \leq 2\|\gamma\|,$$

as are the norm, sup-norm, and "second column" norm:

$$(2.3) \qquad \|\gamma\|_\infty = d < |\gamma e_2| = \sqrt{b^2 + d^2} < \|\gamma\| < 2|\gamma e_2| < 4\|\gamma\|_\infty.$$

For $\gamma \in \Gamma$, let the expanding and contracting eigenvalues of γ be $\lambda_+(\gamma)$ and $\lambda_-(\gamma) = 1/\lambda_+(\gamma)$, with corresponding normalized eigenvectors $v_+(\gamma)$ and $v_-(\gamma)$. Write $\lambda = \lambda_+$ for the expanding eigenvalue, so that

$$\lambda(\gamma) = \lambda_+(\gamma) = \frac{\operatorname{tr}(\gamma) + \sqrt{\operatorname{tr}(\gamma)^2 - 4}}{2}.$$

Remark 2.4. Note that for all $\gamma \in \Gamma$, the eigenvalues are real, and $\lambda > 1$ for $\gamma \neq I$. In particular, Γ has no parabolic elements.

The goal of this section is to prove the following elementary but very useful observation.

PROPOSITION 2.5. *The eigenvalues of two matrices $\gamma, \gamma' \in \Gamma$ with large norms behave essentially multiplicatively, subject to their expanding eigenvectors facing nearby directions. That is,*

$$(2.6) \qquad \lambda(\gamma\gamma') = \lambda(\gamma)\lambda(\gamma')\left[1 + O\left(\left|v_+(\gamma) - v_+(\gamma')\right| + \frac{1}{\|\gamma\|^2} + \frac{1}{\|\gamma'\|^2}\right)\right].$$

Moreover, the expanding eigenvector of the product $\gamma\gamma'$ faces a nearby direction to that of the first γ (and the same in reverse):

$$(2.7) \quad |v_+(\gamma\gamma') - v_+(\gamma)| \ll \frac{1}{\|\gamma\|^2} \qquad and \qquad |v_-(\gamma\gamma') - v_-(\gamma')| \ll \frac{1}{\|\gamma'\|^2}.$$

The implied constants above are absolute.

Proof. For γ large, we have

$$(2.8) \qquad \lambda(\gamma) = \frac{\operatorname{tr}(\gamma) + \sqrt{\operatorname{tr}(\gamma)^2 - 4}}{2} = \operatorname{tr}(\gamma) + O\left(\frac{1}{\|\gamma\|}\right)$$

and

$$v_+(\gamma) = \frac{(b, \lambda_+(\gamma) - a)}{\sqrt{b^2 + (\lambda_+(\gamma) - a)^2}} = \frac{(b, d)}{\sqrt{b^2 + d^2}} + O\left(\frac{1}{\|\gamma\|^2}\right),$$

$$v_-(\gamma) = \frac{(d - \lambda_-(\gamma), c)}{\sqrt{(d - \lambda_-(\gamma))^2 + c^2}} = \frac{(-d, c)}{\sqrt{c^2 + d^2}} + O\left(\frac{1}{\|\gamma\|^2}\right).$$

Note that for γ large,

$$(2.9) \qquad |\langle v_+(\gamma), v_-(\gamma)^\perp \rangle| = \frac{bc + d^2}{\sqrt{b^2 + d^2}\sqrt{c^2 + d^2}} + O\left(\frac{1}{\|\gamma\|^2}\right) \geq \frac{1}{2},$$

meaning that the angle between expanding and contracting vectors does not degenerate.

By (2.8), it is enough to show that the traces behave essentially multiplicatively. We compute

$$|\operatorname{tr}(\gamma\gamma') - \operatorname{tr}(\gamma)\operatorname{tr}(\gamma')| = |(aa' + bc' + cb' + dd') - (a + d)(a' + d')|$$

$$\leq \frac{d}{d'}\left|\frac{bc'd'}{d} - a'd'\right| + \frac{d'}{d}\left|\frac{cb'd}{d'} - ad\right|$$

$$\leq \frac{d}{d'}\left(1 + c'\left|\frac{bd'}{d} - b'\right|\right) + \frac{d'}{d}\left(1 + c\left|\frac{b'd}{d'} - b\right|\right)$$

$$= \frac{d}{d'} + \frac{d'}{d} + (cd' + c'd)\left|\frac{b}{d} - \frac{b'}{d'}\right|.$$

We clearly have

$$\left|\frac{b}{d} - \frac{b'}{d'}\right| \ll |v_+(\gamma) - v_+(\gamma')| + \frac{1}{\|\gamma\|^2} + \frac{1}{\|\gamma'\|^2},$$

and hence

$$|\operatorname{tr}(\gamma\gamma') - \operatorname{tr}(\gamma)\operatorname{tr}(\gamma')| \ll dd'\left(|v_+(\gamma) - v_+(\gamma')| + \frac{1}{\|\gamma\|^2} + \frac{1}{\|\gamma'\|^2}\right).$$

From this and (2.8), (2.6) follows easily. One proves (2.7) in a similar fashion. □

3. Construction of Ω_N

3.1. *The leading term $\aleph$.* In this subsection, we posit the existence and all necessary properties of the leading set $\aleph$ used in our construction of the main ensemble Ω_N. The proof of its existence is arguably the most technical part of the whole paper, so in the interest of exposition, we postpone it to Section 8.

Once and for all, we fix a density point $x \in \mathfrak{C}$ and let

$$(3.1) \qquad \mathfrak{v} = \mathfrak{v}_x := \frac{(x, 1)}{\sqrt{1 + x^2}}$$

be the corresponding unit vector. We will henceforth be largely concerned with elements of Γ whose expanding eigenvectors point in this direction.

For ease of exposition, we assume henceforth that for all $q \geq 1$, the reduction of Γ is full,

$$(3.2) \qquad \Gamma(\mathrm{mod}\, q) \cong \mathrm{SL}_2(q),$$

which is anyway the case for any alphabet $\mathcal{A}$ containing 1 and 2; see Remark 1.33. Minor modifications are needed in the general case.

For N large and δ exceeding δ_0 in (1.21), let

$$(3.3) \qquad \mathfrak{b} := \frac{1}{1000}(\delta - \delta_0) > 0,$$

and let $\alpha_0 > 0$ be a parameter to be chosen later in (8.20). Then we set

$$(3.4) \qquad B := N^{\mathfrak{b}}$$

and

$$(3.5) \qquad \mathcal{Q} := e^{\alpha_0 \sqrt{\log N}}.$$

Let

$$(3.6) \qquad \mathcal{U} \subset \left[\frac{1}{100}B, \frac{99}{100}B \right]$$

be an arithmetic progression of real numbers starting with $u_0 = \frac{1}{100}B$ having common difference

$$(3.7) \qquad |u - u'| = 2B/\mathcal{Q}^5,$$

for u, u' consecutive terms in $\mathcal{U}$, and ending with $u > (\frac{99}{100} - \frac{2}{\mathcal{Q}^5})B$. Then the cardinality of $\mathcal{U}$ is

$$(3.8) \qquad |\mathcal{U}| \asymp \mathcal{Q}^5.$$

PROPOSITION 3.9. *For each* $u \in \mathcal{U}$, *there are nonempty sets* $\aleph_u \subset \Gamma$, *all of the same cardinality*

$$(3.10) \qquad |\aleph_u| = |\aleph_{u'}|,$$

so that the following holds. For every $\mathfrak{a} \in \aleph_u$, *its expanding eigenvector is restricted by*

$$(3.11) \qquad |v_+(\mathfrak{a}) - \mathfrak{v}| < \mathcal{Q}^{-5}$$

and its expanding eigenvalue $\lambda(\mathfrak{a})$ *is restricted by*

$$(3.12) \qquad |\lambda(\mathfrak{a}) - u| < \frac{B}{\mathcal{Q}^5}.$$

In particular,

$$(3.13) \qquad \frac{1}{200}B < \lambda(\mathfrak{a}) < B$$

for N large. Moreover, for any $q < \mathcal{Q}$, any $\omega \in \mathrm{SL}_2(q)$, and any $u \in \mathcal{U}$, we have

(3.14) $$\# \{\mathfrak{a} \in \aleph_u : \mathfrak{a} \equiv \omega (\mathrm{mod}\, q)\} = \frac{|\aleph_u|}{|\mathrm{SL}_2(q)|}(1 + O(\mathcal{Q}^{-4})),$$

where the implied constant does not depend on q, ω, or u.

With the sets $\aleph_u$ as above, we define the main leading set $\aleph$ to be the union of the sets $\aleph_u$,

(3.15) $$\aleph := \bigsqcup_{u \in \mathcal{U}} \aleph_u.$$

Note that the sets $\aleph_u$ are disjoint by (3.12) and (3.7). We repeat that the proof of Proposition 3.9 will be postponed to Section 8.

3.2. *Sector counting.* In this section we give the following slight refinement of Hensley's estimate (1.14), which follows directly from Lalley's methods [Lal89] (see also Theorem 8.1 and Remark 8.6).

PROPOSITION 3.16. *There is a constant $\mathfrak{c} = \mathfrak{c}(\mathcal{A}) > 0$ so that as long as $H < e^{\mathfrak{c}\sqrt{\log T}}$, we have*

(3.17) $$\# \left\{\gamma \in \Gamma : \|\gamma\| < T \text{ and } |v_+(\gamma) - \mathfrak{v}| < \frac{1}{H}\right\} \gg \frac{T^{2\delta}}{H},$$

as $T \to \infty$.

Sketch of proof. Lalley [Lal89, Th. 9] proves the asymptotic formula

(3.18) $$\text{Left-hand side of (3.17)} \sim C \cdot T^{2\delta} \mu(\mathcal{I})$$

under the assumption that Γ is a nonelementary convex-cocompact sub*group* of $\mathrm{SL}_2(\mathbb{R})$. Here $\mathcal{I}$ is the interval of length $1/H$ about $\mathfrak{v}$, and μ is the δ-dimensional Hausdorff measure supported on the limit set $\mathfrak{C}$, lifted (by abuse of notation) to $\mathbb{P}^1$ via

$$d\mu(x, y) = d\mu(x/y),$$

$y \neq 0$. After setting up the symbolic dynamics, the requirement that Γ not contain parabolic elements is needed in the renewal method to make the distortion function eventually positive; see [Lal89, pp. 33, 41]. Our semigroup Γ has no parabolic elements (see Remark 2.4), so the only difference here between a group and (free) semigroup is that for the latter, the transition matrix (see [Lal89, pp. 5, 32]) is trivial; that is, all sequences are allowed in the symbolic dynamics. The rate in (3.18) can be determined directly from Lalley's method (see [BGS11, §12]), with the error crudely estimated as

(3.19) $$\ll T^{2\delta} e^{-c\sqrt{\log T}}.$$

Since $\mathfrak{v}$ in (3.1) corresponds to a density point in $\mathfrak{C}$, we have, again crudely, that

$$\mu(\mathcal{I}) \gg_{\varepsilon} H^{-\delta-\varepsilon} \gg H^{-1},$$

since $\delta < 1$. A sufficient condition for the main term, being bounded below by $CT^{2\delta}/H$, to dominate the error in (3.19), is that $H < e^{\mathfrak{c}\sqrt{\log T}}$ with $\mathfrak{c} < c$. $\qquad\square$

Remark 3.20. The methods of Dolgopyat [Dol98] and Naud [Nau05] might be used to prove (3.17) with H as large as T^{ε}, but this is not needed in our applications.

With this crude estimate in hand, we proceed in the next subsection to detail our construction of the special sets Ξ alluded to in Section 1.3.

3.3. *The set* $\Xi(M, H; L, k)$.

PROPOSITION 3.21. *Given* $M \gg 1$ *and* $H < e^{\mathfrak{c}\sqrt{\log M}}$, *there exists some* L *in the range*

$$(3.22) \qquad\qquad \frac{1}{4}M \le L \le 4M,$$

an integer $k \asymp \log M$, *and a set* $\Xi = \Xi(M, H; L, k) \subset \Gamma$ *having the following properties. For all* $\gamma \in \Xi$, *the expanding eigenvalues are controlled to within* $1/\log L$:

$$(3.23) \qquad\qquad L\left(1 - \frac{1}{\log L}\right) < \lambda(\gamma) < L,$$

the expanding eigenvectors are controlled to within $1/H$:

$$(3.24) \qquad\qquad |v_+(\gamma) - \mathfrak{v}| < \frac{1}{H},$$

and the wordlength metric ℓ (*in the generators* (1.32) *of* Γ) *is controlled exactly*:

$$(3.25) \qquad\qquad \ell(\gamma) = k.$$

Moreover, the cardinality of Ξ *is controlled by*

$$(3.26) \qquad\qquad L^{2\delta} \gg \#\Xi \gg \frac{L^{2\delta}}{H(\log L)^2}.$$

Recall again the the implied constants depend at most on $\mathcal{A}$, which is thought of as fixed throughout.

Proof. We proceed by the following algorithm.

(1) Let $S_1 \subset \Gamma$ be the set of $\gamma \in \Gamma$ of norm controlled by $\|\gamma\| \asymp M$ and for which the expanding vector $v_+(\gamma)$ is within $\frac{1}{H}$ of the fixed vector $\mathfrak{v}$:

$$S_1 := \left\{\gamma \in \Gamma : \frac{M}{2} < \|\gamma\| < M, |v_+(\gamma) - \mathfrak{v}| < \frac{1}{H}\right\}.$$

By (3.17), we have that
$$\#S_1 \gg \frac{M^{2\delta}}{H}.$$

(2) By (2.2) and (2.8), expanding eigenvalues $\lambda(\gamma)$ of $\gamma \in S_1$ satisfy
$$\frac{1}{4}M \leq \lambda(\gamma) \leq 4M.$$
Hence we can find (by pigeonhole) an L in this range so that
$$\#\{\gamma \in S_1 : L\left(1 - \frac{1}{\log L}\right) < \lambda(\gamma) < L\} \gg \frac{L^{2\delta}}{H \log L}.$$
Call the above set S_2; its expanding eigenvalues are all nearly of the same size.

(3) Lastly, note that the wordlength metric ℓ is commensurable with the archimedean one,
$$\ell(\gamma) \asymp \log \|\gamma\|,$$
with implied constant depending on $\mathcal{A}$. So (again by pigeonhole) we can find some k such that

(3.27)
$$\#\{\gamma \in S_2 : \ell(\gamma) = k\} \gg \frac{L^{2\delta}}{H(\log L)^2}.$$
Call this set S_3; then the elements of S_3 all have the same wordlength, in addition to the previous qualities.

We rename this last set S_3 to $\Xi = \Xi(M, H; L, k)$. $\square$

3.4. *Decomposing N and the ensemble Ω_N.* We return to our main parameter N and decompose it dyadically as follows. Recall that we have already presupposed the construction of a set $\aleph$ in (3.15), all of whose expanding eigenvectors are within $\mathcal{Q}^{-5}$ of $\mathfrak{v}$, and with eigenvalues of size B; see (3.13). Recall from (3.4) that $B = N^{\mathfrak{b}}$ and that $\mathcal{Q}$ is given by (3.5).

Setup: We start by taking

(3.28)
$$M = \sqrt{N}/B = N^{1/2 - \mathfrak{b}}, \qquad H = \mathcal{Q}^5.$$

The exponent α_0 in the definition (3.5) of $\mathcal{Q}$ is chosen in (8.20) to be sufficiently small that $H < e^{\mathfrak{c}\sqrt{\log M}}$. Run the algorithm of the previous subsection to generate the set $\Xi(M, H; L, k)$. By (3.22), the returned parameter L satisfies
$$L = \alpha_1 M = \alpha_1 N^{1/2 - \mathfrak{b}},$$
with
$$\alpha_1 \in (1/4, 4).$$
Write
$$\tilde{N}_1 := L = \alpha_1 N^{1/2 - \mathfrak{b}}, \qquad N_1 := \alpha_1 N^{1/2} = B \cdot \tilde{N}_1,$$
and rename the returned set to $\tilde{\Xi}_1 = \Xi(M, H; L, k)$, also setting
$$\Xi_1 := \aleph \cdot \tilde{\Xi}_1.$$

Remark 3.29. Despite the wordlength in $\aleph$ being unrestricted, the wordlength in $\tilde{\Xi}_1$ is fixed in (3.25). So the representation of an element in Ξ_1 as a product of ones in $\aleph$ and $\tilde{\Xi}_1$ is still unique.

We have crudely that

$$(3.30) \qquad |\Xi_1| \geq |\tilde{\Xi}_1| \gg_\varepsilon \tilde{N}_1^{2\delta-\varepsilon} \gg N^{\delta-2\delta\mathfrak{b}-\varepsilon}.$$

(The cardinality of $\aleph$ is quite deficient relative to its norm, so we lose little from estimating trivially $|\aleph| \geq 1$.)

Step 1: Next we set

$$M = \frac{N_1^{1/2}}{\alpha_1} = \frac{N^{1/4}}{\alpha_1^{1/2}}, \qquad H = \log M,$$

and we generate another set $\Xi(M, H; L, k)$. Define

$$N_2 := L = \alpha_2 M = \frac{\alpha_2 N^{1/4}}{\alpha_1^{1/2}},$$

with $\alpha_2 \in (1/4, 4)$, and rename the returned set to Ξ_2. We have

$$|\Xi_2| \gg \frac{N_2^{2\delta}}{(\log N_2)^3}.$$

Iterate: Start with $j = 3$ and iterate up to $j = J - 1$, where

$$(3.31) \qquad 2^{J-1} = c \log N.$$

Here the constant $c > 0$ is absolute (independent of N), determined by (3.45). For each such j, set

$$(3.32) \qquad M := \frac{(N_{j-1})^{1/2}}{\alpha_{j-1}} = \frac{N^{1/2^j}}{\alpha_{j-1}^{1/2} \alpha_{j-2}^{1/4} \cdots (\alpha_1)^{1/2^{(j-1)}}}, \qquad H = \log M,$$

and use Proposition 3.21 to generate the set $\Xi(M, H; L, k)$. Define

$$(3.33) \qquad N_j := L = \alpha_j M = \frac{\alpha_j N^{1/2^j}}{\alpha_{j-1}^{1/2} \alpha_{j-2}^{1/4} \cdots (\alpha_1)^{1/2^{(j-1)}}},$$

with $\alpha_j \in (1/4, 4)$, and call the returned set Ξ_j. Note that

$$(3.34) \qquad |\Xi_j| \gg \frac{N_j^{2\delta}}{(\log N_j)^3}$$

and

$$(3.35) \qquad \frac{1}{16} N^{1/2^j} < N_j < 16 N^{1/2^j}.$$

End: For the last step, $j = J$, we set

$$M = \frac{N_{J-1}}{(\alpha_{J-1})^2} = \frac{N^{1/2^{(J-1)}}}{\alpha_{J-1}(\alpha_{J-2})^{1/2} \cdots \alpha_1^{1/2^{(J-2)}}}, \qquad H = \log M,$$

and we generate one last set $\Xi_J := \Xi(M, H; L, k)$. Define

$$N_J := L = \frac{\alpha_J N^{1/2^{(J-1)}}}{\alpha_{J-1} \cdots \alpha_1^{1/2^{(J-2)}}} \asymp N^{1/2^{(J-1)}} = e^{1/c} \ll 1,$$

where we used (3.31). Since $\frac{1}{4} < N_J/M = \alpha_J < 4$, we have

$$(3.36) \qquad \frac{1}{4} < \frac{N_1 N_2 \cdots N_J}{N} = \frac{B\tilde{N}_1 N_2 \cdots N_J}{N} < 4.$$

We now define the main ensemble Ω_N by concatenating the sets Ξ_j developed above.

$$(3.37) \qquad \Omega_N := \Xi_1 \cdot \Xi_2 \cdots \Xi_{J-1} \cdot \Xi_J = \aleph \cdot \tilde{\Xi}_1 \cdot \Xi_2 \cdots \Xi_{J-1} \cdot \Xi_J.$$

3.5. *Properties of Ω_N.* For $\gamma \in \Omega_N$, write

$$\gamma = \mathfrak{a} \cdot \tilde{\xi}_1 \ \xi_2 \cdots \xi_J$$

according to the decomposition (3.37). Note that by the fixed wordlength restriction (3.25), this decomposition is unique; see Remark 3.29. Recall that the expanding vectors v_+ all point nearly in the direction of $\mathfrak{v}$ in (3.1).

LEMMA 3.38. *For any $2 \le j_1 \le j_2 \le J$, and $\xi_{j_1} \in \Xi_{j_1}, \cdots \xi_{j_2} \in \Xi_{j_2}$, and any $\mathfrak{a} \in \aleph$, $\tilde{\xi}_1 \in \tilde{\Xi}_1$, we have the following control on expanding eigen-vectors and -values of large products:*

$$(3.39) \qquad |v_+(\tilde{\xi}_1 \cdot \xi_2 \cdots \xi_J) - \mathfrak{v}| \ll \mathcal{Q}^{-5},$$

$$(3.40) \qquad \frac{1}{2} < \frac{\lambda(\xi_{j_1} \xi_{j_1+1} \cdots \xi_{j_2-1} \xi_{j_2})}{N_{j_1} N_{j_1+1} \cdots N_{j_2-1} N_{j_2}} < 2,$$

$$(3.41) \qquad \frac{1}{2} < \frac{\lambda(\tilde{\xi}_1 \xi_2 \cdots \xi_{j_2-1} \xi_{j_2})}{\tilde{N}_1 N_2 \cdots N_{j_2-1} N_{j_2}} < 2,$$

and

$$(3.42) \qquad \frac{1}{2} < \frac{\lambda(\mathfrak{a}\tilde{\xi}_1 \xi_2 \cdots \xi_{j_2-1} \xi_{j_2})}{\lambda(\mathfrak{a})\tilde{N}_1 \cdot N_2 \cdots N_{j_2-1} N_{j_2}} < 2.$$

Proof. From (2.7), (3.24), and the choice of H in (3.28), we have that

$$|v_+(\tilde{\xi}_1 \cdot \xi_2 \cdots \xi_J) - \mathfrak{v}| \le |v_+(\tilde{\xi}_1 \cdot \xi_2 \cdots \xi_J) - v_+(\tilde{\xi}_1)| + |v_+(\tilde{\xi}_1) - \mathfrak{v}|$$
$$\ll \frac{1}{\|\tilde{\xi}_1\|^2} + \frac{1}{\mathcal{Q}^5},$$

whence (3.39) follows from (3.5).

Similarly, we have for $j \in [j_1, j_2] \subset [2, J]$ that

$$(3.43) \qquad |v_+(\xi_j \xi_{j+1} \cdots \xi_{j_2}) - \mathfrak{v}| \ll \frac{1}{\log N_j},$$

where we used the choice of H in (3.32).

We now prove by downward induction on j_1 that

(3.44)

$$\lambda(\xi_{j_1}\xi_{j_1+1}\cdots\xi_{j_2}) = N_{j_1}N_{j_1+1}\cdots N_{j_2}$$
$$\times \left[1 + O\left(\frac{1}{\log N_{j_1}} + \frac{1}{\log N_{j_1+1}} + \cdots + \frac{1}{\log N_{j_2}}\right)\right].$$

If $j_1 = j_2$, then (3.44) follows immediately from (3.23) and (3.33). If $j_1 = j_2-1$, then from (2.6), (3.23), (3.24), and (3.33), we have

$$\lambda(\xi_{j_2-1}\xi_{j_2}) = \lambda(\xi_{j_2-1})\lambda(\xi_{j_2})$$
$$\times \left[1 + O\left(|v_+(\xi_{j_2-1}) - v_+(\xi_{j_2})| + \frac{1}{\|\xi_{j_2-1}\|^2} + \frac{1}{\|\xi_{j_2}\|^2}\right)\right]$$
$$= N_{j_2-1}N_{j_2}\left[1 + O\left(\frac{1}{\log N_{j_2-1}} + \frac{1}{\log N_{j_2}}\right)\right],$$

as desired.

In general, we have by (3.43) that

$$\lambda(\xi_{j_1}\xi_{j_1+1}\cdots\xi_{j_2}) = \lambda(\xi_{j_1})\lambda(\xi_{j_1+1}\cdots\xi_{j_2})$$
$$\times \left[1 + O\left(|v_+(\xi_{j_1}) - v_+(\xi_{j_1+1}\cdots\xi_{j_2})| + \frac{1}{\|\xi_{j_1}\|^2} + \frac{1}{\lambda(\xi_{j_1+1}\cdots\xi_{j_2})^2}\right)\right]$$
$$= N_{j_1}\lambda(\xi_{j_1+1}\cdots\xi_{j_2})\left[1 + O\left(\frac{1}{\log N_{j_1}} + \frac{1}{\log N_{j_1+1}}\right)\right],$$

from which (3.44) follows by induction.

The rate in (3.44) may be replaced crudely by

(3.45) $$\left[1 + O\left(\frac{2^J}{\log N}\right)\right],$$

whence (3.40) follows on taking the constant c in (3.31) sufficiently small (independent of N). Estimates (3.41) and (3.42) are proved in the same way.   □

As a consequence of (3.42), (3.13), and (3.36), we have that for all $\gamma \in \Omega_N$,

(3.46) $$\|\gamma\| \le 2\lambda(\gamma) \le 16N,$$

so indeed the norms are all controlled.

Moreover, the size of Ω_N is not too much smaller than (1.36). Indeed, we have from (3.30), (3.34), and (3.31) that

(3.47) $$\#\Omega_N = \#\Xi_1 \cdot \#\Xi_2 \cdots \#\Xi_J$$
$$\gg_\varepsilon \tilde{N}_1^{2\delta-\varepsilon}\frac{(N_2)^{2\delta}}{(\log N_2)^3}\cdots\frac{(N_J)^{2\delta}}{(\log N_J)^3}$$
$$\gg N^{2\delta-2\delta\mathfrak{b}-\varepsilon}.$$

It also follows that for any $j \geq 2$,

$$(3.48) \qquad \#\Xi_j \cdot \#\Xi_{j+1} \cdot \cdots \cdot \#\Xi_J \gg (N_j N_{j+1} \cdots N_J)^{2\delta} \cdot e^{-c(J-j)\log\log N_j}$$

for an absolute constant $c > 0$.

With the set Ω_N constructed, we define our exponential sum S_N as in (1.45) and proceed with the circle method.

4. Major arcs analysis

In this section we estimate the major arcs contribution. First we use the set $\aleph$ described in Section 3.1 to prove that in the major arcs, our exponential sum S_N in (1.45) splits as a product of modular and archimedean components, as in (1.38). Then we prove that the major arcs contribution is of the correct order of magnitude.

4.1. *Splitting into modular and Archimedean aomponents.* Let $\mathcal{Q}$ be as in (3.5) and B as in (3.4). Recall from (1.37) that the major arcs of level $\mathcal{Q}$ are given by

$$\mathfrak{M}_{\mathcal{Q}} = \bigsqcup_{q < \mathcal{Q}} \bigsqcup_{(a,q)=1} \left[\frac{a}{q} - \frac{\mathcal{Q}}{N}, \frac{a}{q} + \frac{\mathcal{Q}}{N} \right].$$

Let $\nu_q : \mathbb{Z}/q\mathbb{Z} \to \mathbb{C}$ record the mod q distribution of $\mathfrak{D}$. That is, for $a \in \mathbb{Z}/q\mathbb{Z}$, set

$$(4.1) \qquad \nu_q(a) := \frac{1}{|\operatorname{SL}_2(q)|} \sum_{\omega \in \operatorname{SL}_2(q)} e\left(\frac{a}{q} \langle \omega e_2, e_2 \rangle \right).$$

THEOREM 4.2. *There exists a function* $\varpi_N : \mathbb{R}/\mathbb{Z} \to \mathbb{C}$, *given explicitly in* (4.19), *satisfying the following three conditions*:

(1) *The Fourier transform*

$$\widehat{\varpi}_N : \mathbb{Z} \to \mathbb{C} : n \mapsto \int_0^1 \varpi_N(\theta) e(-n\theta) d\theta$$

is real-valued and nonnegative, with

$$(4.3) \qquad \varpi_N(0) = \sum_n \widehat{\varpi}_N(n) \ll |\Omega_N|.$$

(2) *For* $\frac{1}{25}N < n < \frac{1}{5}N$, *we have*

$$(4.4) \qquad \widehat{\varpi}_N(n) \gg \frac{|\Omega_N|}{N}.$$

(3) *Moreover, we have on the major arcs* $\theta = \frac{a}{q} + \beta \in \mathfrak{M}_{\mathcal{Q}}$ *that*

$$(4.5) \qquad S_N\left(\frac{a}{q} + \beta \right) = \nu_q(a) \varpi_N(\beta) \left(1 + O(\mathcal{Q}^{-4}) \right).$$

Proof. We use the decomposition (3.37) in the form

$$(4.6) \qquad \Omega_N = \aleph \cdot \Omega'$$

with

$$\Omega' = \tilde{\Xi}_1 \Xi_2 \cdots \Xi_J,$$

so that

$$S_N(\theta) = \sum_{\mathfrak{a} \in \aleph} \sum_{\gamma \in \Omega'} e(\theta \langle \mathfrak{a}\gamma e_2, e_2 \rangle).$$

For $\mathfrak{a} \in \aleph$, recall from (3.13) that we have $\lambda(\mathfrak{a}) \asymp B$ and from (3.11) and (3.39) that

$$(4.7) \qquad |v_+(\mathfrak{a}) - \mathfrak{v}| < \mathcal{Q}^{-5}, \qquad |v_+(\gamma) - \mathfrak{v}| \ll \mathcal{Q}^{-5}.$$

To use the key property (3.12) in the construction of $\aleph$, we need to convert the expression $\langle \mathfrak{a}\gamma e_2, e_2 \rangle$ in S_N into one involving $\lambda(\mathfrak{a})$.

We will make regular use of the following elementary formula: For any two linearly independent vectors $v_+, v_- \in \mathbb{R}^2$, we can write any $w \in \mathbb{R}^2$ as

$$(4.8) \qquad w = \frac{\langle w, v_-^\perp \rangle}{\langle v_+, v_-^\perp \rangle} v_+ + \frac{\langle w, v_+^\perp \rangle}{\langle v_-, v_+^\perp \rangle} v_-.$$

Here $(x, y)^\perp = (-y, x)$. Recalling (2.9), it easily follows that for a unit vector w and any large $\xi \in \Gamma$,

$$(4.9) \qquad \xi w = \lambda(\xi) \frac{\langle w, v_-^\perp(\xi) \rangle}{\langle v_+(\xi), v_-^\perp(\xi) \rangle} v_+(\xi) \left(1 + O\left(\frac{1}{\|\xi\|^2} \right) \right),$$

whence

$$(4.10) \qquad \langle \xi e_2, e_2 \rangle = \lambda(\xi) \frac{\langle e_2, v_-^\perp(\xi) \rangle}{\langle v_+(\xi), v_-^\perp(\xi) \rangle} \langle v_+(\xi), e_2 \rangle \left(1 + O\left(\frac{1}{\|\xi\|^2} \right) \right).$$

Applied to our present situation, we have by (4.7) that

$$(4.11) \qquad \langle \gamma e_2, e_2 \rangle = \lambda(\gamma) \frac{\langle e_2, v_-^\perp(\gamma) \rangle}{\langle \mathfrak{v}, v_-^\perp(\gamma) \rangle} \langle \mathfrak{v}, e_2 \rangle \left(1 + O\left(\frac{1}{\mathcal{Q}^5} \right) \right)$$

and

$$(4.12) \qquad \langle \mathfrak{a}\gamma e_2, e_2 \rangle = \lambda(\mathfrak{a}\gamma) \frac{\langle e_2, v_-^\perp(\mathfrak{a}\gamma) \rangle}{\langle v_+(\mathfrak{a}\gamma), v_-^\perp(\mathfrak{a}\gamma) \rangle} \langle v_+(\mathfrak{a}\gamma), e_2 \rangle \left(1 + O\left(\frac{1}{N^2} \right) \right)$$

$$= \lambda(\mathfrak{a})\lambda(\gamma) \frac{\langle e_2, v_-^\perp(\gamma) \rangle}{\langle \mathfrak{v}, v_-^\perp(\gamma) \rangle} \langle \mathfrak{v}, e_2 \rangle \left(1 + O\left(\frac{1}{\mathcal{Q}^5} \right) \right),$$

where we also used (2.6) and (2.7).

Comparing (4.12) and (4.11), we have that

$$(4.13) \qquad \langle \mathfrak{a}\gamma e_2, e_2 \rangle = \lambda(\mathfrak{a}) \langle \gamma e_2, e_2 \rangle + O\left(N/\mathcal{Q}^5 \right).$$

For $\theta = \frac{a}{q} + \beta \in \mathfrak{M}_\mathcal{Q}$ with $|\beta| < \mathcal{Q}/N$, we insert (4.13) into S_N, giving

(4.14)

$$
S_N\left(\frac{a}{q} + \beta\right) = \sum_{\mathfrak{a} \in \aleph} \sum_{\gamma \in \Omega'} e\left(\frac{a}{q}\langle \mathfrak{a}\gamma e_2, e_2\rangle\right) e\left(\beta \langle \mathfrak{a}\gamma e_2, e_2\rangle\right)
$$

$$
= \sum_{\mathfrak{a} \in \aleph} \sum_{\gamma \in \Omega'} e\left(\frac{a}{q}\langle \mathfrak{a}\gamma e_2, e_2\rangle\right) e\left(\beta\lambda(\mathfrak{a})\langle \gamma e_2, e_2\rangle\right) + O\left(\mathcal{Q}^{-4}|\Omega|\right)
$$

$$
= \sum_{\gamma \in \Omega'} \sum_{\omega \in \mathrm{SL}_2(q)} e\left(\frac{a}{q}\langle \omega\gamma e_2, e_2\rangle\right) \sum_{\substack{\mathfrak{a} \in \aleph \\ \mathfrak{a} \equiv \omega (\mathrm{mod}\, q)}} e\left(\beta\lambda(\mathfrak{a})\langle \gamma e_2, e_2\rangle\right)
$$

$$
+ O\left(\mathcal{Q}^{-4}|\Omega_N|\right),
$$

where we decomposed the $\mathfrak{a}$ sum into residue classes ω in $\mathrm{SL}_2(q)$.

Next using (3.15), write the innermost sum above as

(4.15)
$$
\sum_{\substack{\mathfrak{a} \in \aleph \\ \mathfrak{a} \equiv \omega (\mathrm{mod}\, q)}} e\left(\lambda(\mathfrak{a})\beta \langle \gamma e_2, e_2\rangle\right)
$$

$$
= \sum_{u \in \mathcal{U}} \sum_{\substack{\mathfrak{a} \in \aleph_u \\ \mathfrak{a} \equiv \omega (\mathrm{mod}\, q)}} e\left(\lambda(\mathfrak{a})\beta \langle \gamma e_2, e_2\rangle\right)
$$

$$
= \sum_{u \in \mathcal{U}} e\left(\beta u \langle \gamma e_2, e_2\rangle\right) \left(\sum_{\substack{\mathfrak{a} \in \aleph_u \\ \mathfrak{a} \equiv \omega (\mathrm{mod}\, q)}} 1\right) \left(1 + O(\mathcal{Q}^{-4})\right),
$$

where we applied (3.12).

By (3.14) and (3.10) (that the cardinality of $\aleph_u$ is the same for all u), the innermost sum is

(4.16)
$$
\sum_{\substack{\mathfrak{a} \in \aleph_u \\ \mathfrak{a} \equiv \omega (\mathrm{mod}\, q)}} 1 = \frac{|\aleph|}{|\mathcal{U}| \cdot |\mathrm{SL}_2(q)|}\left(1 + O(\mathcal{Q}^{-4})\right),
$$

where the implied constant does not depend on u, ω, or q.

Returning to (4.14), inputting (4.16) and (4.15) gives

(4.17)

$$
S_N\left(\frac{a}{q} + \beta\right) = \frac{1}{|\mathrm{SL}_2(q)|} \sum_{\omega \in \mathrm{SL}_2(q)} e\left(\frac{a}{q}\langle \omega e_2, e_2\rangle\right) \frac{|\aleph|}{|\mathcal{U}|} \sum_{\gamma \in \Omega'} \sum_{u \in \mathcal{U}} e\left(\beta u \langle \gamma e_2, e_2\rangle\right)
$$

$$
\times \left(1 + O\left(\mathcal{Q}^{-4}\right)\right),
$$

where we used the fact that the ω sum runs over all of $\mathrm{SL}_2(q)$, so is independent of γ. Note that S_N has already split into modular and archimedean components, with the first piece being $\nu_q(a)$ as in (4.1).

We continue to massage the archimedean component. Fix γ and u. For any $m \in \mathbb{Z}$ with

$$|m - u \langle \gamma e_2, e_2 \rangle| \leq B \langle \gamma e_2, e_2 \rangle / \mathcal{Q}^5,$$

we clearly have

$$e\left(\beta u \langle \gamma e_2, e_2 \rangle\right) = e\left(\beta m\right)\left(1 + O(\mathcal{Q}^{-4})\right),$$

and there are $2B \langle \gamma e_2, e_2 \rangle / \mathcal{Q}^5 + O(1)$ integers m in this range. Hence

$$(4.18) \quad e\left(\beta u \langle \gamma e_2, e_2 \rangle\right) = \frac{\mathcal{Q}^5}{2B \langle \gamma e_2, e_2 \rangle} \sum_{\substack{m \in \mathbb{Z} \\ \left|\frac{m}{\langle \gamma e_2, e_2 \rangle} - u\right| \leq \frac{B}{\mathcal{Q}^5}}} e\left(\beta m\right)\left(1 + O(\mathcal{Q}^{-4})\right).$$

Reversing the u and m sums and inserting (4.18) into (4.17) gives

$$S_N\left(\frac{a}{q} + \beta\right) = \nu_q(a)\varpi_N(\beta)\left(1 + O\left(\mathcal{Q}^{-4}\right)\right),$$

where

$$(4.19) \qquad \varpi_N(\beta) := \frac{|\aleph|}{|\mathcal{U}|} \sum_{\gamma \in \Omega'} \frac{\mathcal{Q}^5}{2B \langle \gamma e_2, e_2 \rangle} \sum_{m \in \mathbb{Z}} e\left(\beta m\right) \sum_{u \in \mathcal{U}} \mathbf{1}_{\left|\frac{m}{\langle \gamma e_2, e_2 \rangle} - u\right| \leq \frac{B}{\mathcal{Q}^5}}.$$

Hence (4.5) is satisfied. Also the Fourier transform

$$(4.20) \qquad \widehat{\varpi}_N(n) = \frac{|\aleph|}{|\mathcal{U}|} \sum_{\gamma \in \Omega'} \frac{\mathcal{Q}^5}{2B \langle \gamma e_2, e_2 \rangle} \sum_{u \in \mathcal{U}} \mathbf{1}_{\left|\frac{n}{\langle \gamma e_2, e_2 \rangle} - u\right| \leq \frac{B}{\mathcal{Q}^5}}$$

is clearly real and nonnegative, so (1) is satisfied.

Now, combining (4.11), (2.9), (2.2), (2.3), (2.8), (3.41), and (3.36), we have

$$(4.21) \qquad \qquad \frac{1}{4}\frac{N}{B} < \langle \gamma e_2, e_2 \rangle < 4\frac{N}{B},$$

and hence for $\frac{1}{25}N < n < \frac{1}{5}N$, we have, crudely, that

$$\frac{1}{100}B < \frac{n}{\langle \gamma e_2, e_2 \rangle} < \frac{99}{100}B.$$

Hence by the spacing in (3.7) of $u \in \mathcal{U}$ in this range, the innermost sum in (4.20) is guaranteed to have at least one contribution, giving

$$\widehat{\varpi}_N(n) \gg \frac{|\aleph|}{|\mathcal{U}|} \sum_{\gamma \in \Omega'} \frac{\mathcal{Q}^5}{2B \langle \gamma e_2, e_2 \rangle} \gg \frac{|\aleph||\Omega'|}{N} = \frac{|\Omega_N|}{N},$$

where we used (4.21), (3.8), and (4.6). So (4.4) is satisfied, and the proof of Theorem 4.2 is complete. $\qquad\qquad\square$

4.2. *The major arcs contribution.* Equipped with (4.5), it is now straight-forward to produce the necessary major arcs contribution. For technical reasons, we need a smoothed cutoff, and we introduce the triangle function, ψ, given by

$$(4.22) \qquad \psi(x) := \begin{cases} 1+x, & \text{if } -1 < x < 0, \\ 1-x, & \text{if } 0 \le x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

It is well known that the Fourier transform is nonnegative:

$$(4.23) \qquad \widehat{\psi}(y) = \left(\frac{\sin(\pi y)}{\pi y} \right)^2.$$

Let ψ_N be the function localized at level $\mathcal{Q}/N$ near the origin:

$$\psi_N(x) := \psi\left(\frac{N}{\mathcal{Q}} x \right).$$

Periodize ψ_N to Ψ_N on $\mathbb{R}/\mathbb{Z}$:

$$\Psi_N(\theta) := \sum_{m \in \mathbb{Z}} \psi_N(\theta + m),$$

and put each such spike at a major arc:

$$(4.24) \qquad \Psi_{\mathcal{Q},N}(\theta) := \sum_{q < \mathcal{Q}} \sum_{(a,q)=1} \Psi_N\left(\theta - \frac{a}{q} \right).$$

Note that the support of $\Psi_{\mathcal{Q},N}$ is $\mathfrak{M}_{\mathcal{Q}}$.

As in (1.35), write the representation number

$$R_N(n) := \widehat{S}_N(n) = \int_0^1 S_N(\theta) e(-n\theta) d\theta,$$

and decompose it into a (smoothed) major arcs contribution and an error

$$(4.25) \qquad R_N(n) = \mathcal{M}_N(n) + \mathcal{E}_N(n),$$

where

$$(4.26) \qquad \mathcal{M}_N(n) := \int_0^1 \Psi_{\mathcal{Q},N}(\theta) S_N(\theta) e(-n\theta) d\theta$$

and

$$(4.27) \qquad \mathcal{E}_N(n) := \int_0^1 (1 - \Psi_{\mathcal{Q},N}(\theta)) S_N(\theta) e(-n\theta) d\theta.$$

The ultimate goal of this section is to prove the following

THEOREM 4.28. *For $\frac{1}{20} N \le n < \frac{1}{10} N$,*

$$(4.29) \qquad \mathcal{M}_N(n) \gg \frac{1}{\log \log N} \frac{|\Omega_N|}{N}.$$

Proof. Fix $\frac{1}{20}N \leq n < \frac{1}{10}N$. Starting with (4.26), insert (4.24) and (4.5) (recall supp $\Psi_{\mathcal{Q},N} \subset \mathfrak{M}_{\mathcal{Q}}$), and make the change of variables $\beta = \theta - a/q$:

$$(4.30) \qquad \mathcal{M}_N(n) = \sum_{q < \mathcal{Q}} \sum_{(a,q)=1} \nu_q(a) e\left(-n\frac{a}{q}\right)$$

$$\times \int_0^1 \Psi_N(\beta)\, \varpi_N(\beta)\, e(-n\beta)\, d\beta$$

$$+ O\left(\mathcal{Q}\mathcal{Q}\frac{\mathcal{Q}}{N}|\Omega_N|\mathcal{Q}^{-4}\right),$$

where we used (4.3).

Note that $\mathcal{M}_N$ has already split (up to acceptable error) into the product of the singular series

$$(4.31) \qquad \mathfrak{S}_{\mathcal{Q}}(n) := \sum_{q < \mathcal{Q}} \sum_{(a,q)=1} \nu_q(a) e\left(-n\frac{a}{q}\right)$$

and the singular integral

$$(4.32) \quad \Pi_N(n) := \int_0^1 \Psi_N(\beta)\, \varpi_N(\beta)\, e(-n\beta)\, d\beta = \sum_{m \in \mathbb{Z}} \widehat{\Psi}_N(n-m)\widehat{\varpi}_N(m)$$

$$= \frac{\mathcal{Q}}{N} \sum_{m \in \mathbb{Z}} \widehat{\psi}\left(\frac{\mathcal{Q}}{N}(n-m)\right) \widehat{\varpi}_N(m).$$

First we sketch an analysis of the singular series, which is standard. Insert (4.1) into (4.31):

$$\mathfrak{S}_{\mathcal{Q}}(n) = \sum_{q < \mathcal{Q}} \frac{1}{|\operatorname{SL}_2(q)|} \sum_{\gamma \in \operatorname{SL}_2(q)} c_q\left(\langle \gamma e_2, e_2 \rangle - n\right),$$

where c_q is the classical Ramanujan sum

$$c_q(m) = \sum_{(a,q)=1} e(am/q).$$

Recall that c_q is multiplicative in q and that $c_q(m) = \mu(q)$ if $(m,q) = 1$. (Here μ is the Möbius function.) Hence we may extend the range of the sum $q < \mathcal{Q}$ to $q < \infty$ with a negligible error, obtaining a sum which factors into an Euler product. At each place, the contribution from prime powers is negligible. We are left to analyze

$$(4.33) \qquad \mathfrak{S}_{\mathcal{Q}}(n) \gg \mathfrak{S}(n) \gg \prod_p \left(1 + \frac{1}{|\operatorname{SL}_2(p)|} \sum_{\gamma \in \operatorname{SL}_2(p)} c_p\left(\langle \gamma e_2, e_2 \rangle - n\right)\right)$$

$$= \prod_{p \nmid n} \left(1 + \frac{1}{p^2 - 1}\right) \prod_{p | n} \left(1 - \frac{1}{p+1}\right) \gg \frac{1}{\log \log n}.$$

Returning to (4.32), we now analyze the singular integral. By positivity and using (4.23) that $\widehat{\psi}(y) > 2/5$ for $|y| < 1/2$, we have

$$\Pi_N(n) = \frac{\mathcal{Q}}{N} \sum_{m \in \mathbb{Z}} \widehat{\psi}\left(\frac{\mathcal{Q}}{N}(n-m)\right) \widehat{\varpi}_N(m) \geq \frac{2}{5} \frac{\mathcal{Q}}{N} \sum_{|m-n| < N/(2\mathcal{Q})} \widehat{\varpi}_N(m).$$

For N (and hence $\mathcal{Q}$) sufficiently large, the ranges $n/N \in [\frac{1}{20}, \frac{1}{10}]$ and $|m-n| < N/(2\mathcal{Q})$ force $m/N \in [\frac{1}{25}, \frac{1}{5}]$, so (4.4) applies, giving

$$(4.34) \qquad \Pi_N(n) \gg \frac{\mathcal{Q}}{N} \frac{N}{2\mathcal{Q}} \frac{|\Omega_N|}{N} \gg \frac{|\Omega_N|}{N}.$$

Inserting (4.34) and (4.33) into (4.30) gives (4.29), as claimed. $\qquad \square$

5. Minor arcs analysis I

We keep all the notation from the previous section. Having dealt with the main term (4.26), we are now tasked with estimating the error $\mathcal{E}_N$ in (4.27). As discussed in (1.39)–(1.41), the key goal is to estimate

$$\sum_{n \in \mathbb{Z}} |\mathcal{E}_N(n)|^2 = \int_0^1 |1 - \Psi_{\mathcal{Q},N}(\theta)|^2 |S_N(\theta)|^2 d\theta,$$

where we applied Parseval's formula. For θ outside the major arcs, $\Psi_{\mathcal{Q},N}$ vanishes, and we decompose the above integral into regions

$$W_{Q,K} := \left\{ \theta = \frac{a}{q} + \beta : \frac{1}{2}Q \leq q < Q, (a,q) = 1, \frac{K}{2N} \leq |\beta| < \frac{K}{N} \right\}.$$

Here the parameters Q and K range dyadically in

$$(5.1) \qquad Q < N^{1/2}, \qquad K < N^{1/2}/Q.$$

If $K = O(1)$, we replace the condition $\frac{1}{2}K/N \leq |\beta| < K/N$ in $W_{Q,K}$ by just $|\beta| < K/N$, and any appearances of K should be replaced by 1.

In this section, we give two bounds for $S_N(\theta)$, similar to Theorems 5.1 and 6.1 of [BK10]. These will suffice as long as Q or K is large.

5.1. *The bound for K large.* We will first bound $\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta$ by pulling out the largest value of the integrand and multiplying by the measure of the domain, which $\ll Q^2 \frac{K}{N}$. To get the desired bound of $N^{4\delta-1}$ (see (1.39)), we need to bound the sup norm of S_N on $W_{Q,K}$ by a bit less than $N^{2\delta}/(K^{1/2}Q)$. We will win by an extra $K^{1/2}$.

PROPOSITION 5.2. *Let N, Q, K be as above, and write $\theta = \frac{a}{q} + \beta \in W_{Q,K}$. Then*

$$(5.3) \qquad |S_N(\theta)| \ll N^{2\delta} \left(\frac{N^{1-\delta}}{KQ}\right),$$

as $N \to \infty$.

Proof. This is a simplified version of Theorem 5.1 in [BK10]; we repeat the arguments. By (3.37), we decompose

$$(5.4) \qquad \Omega_N = \Xi_1 (\Xi_2 \Xi_3 \cdots \Xi_J) = \Xi_1 \cdot \Omega'.$$

Then by (2.3), (2.2), (2.8), Lemma 3.38, and (3.35), we have for $\gamma \in \Xi_1$ and $\omega \in \Omega'$ that

$$(5.5) \qquad |{}^t\gamma e_2|, |\omega e_2| < 50N^{1/2}.$$

Note also from (3.26) that

$$(5.6) \qquad \#\Xi_1, \#\Omega' \ll N^\delta.$$

Then we can rewrite $S_N(\theta)$ as

$$(5.7) \qquad S_N(\theta) = \sum_{x \in \mathbb{Z}^2} \sum_{y \in \mathbb{Z}^2} \mu(x)\nu(y)e(\theta \langle x, y\rangle),$$

where μ and ν are image measures in $\mathbb{Z}^2$ defined by

$$\mu(x) := \sum_{\gamma \in \Xi_1} \mathbf{1}_{\{x = {}^t\gamma \cdot e_2\}},$$

and similarly

$$\nu(y) := \sum_{\omega \in \Omega'} \mathbf{1}_{\{y = \omega \cdot e_2\}}.$$

The projection $\omega \mapsto \omega \cdot e_2$ in ν is one-to-one, since it is well known that the continued fraction of a rational number, if restricted to have even length, is unique. The map μ is also one-to-one, since $\mathcal{G}$ is preserved under transposition, ${}^tg \in \mathcal{G}$ for $g \in \mathcal{G}$ (since its generators (1.28) are fixed by transposition). Hence we have

$$(5.8) \qquad \|\mu\|_\infty \leq 1, \qquad \|\nu\|_\infty \leq 1.$$

Note that for any $y, y' \in \operatorname{supp} \nu$, we have from (5.5) that $|y-y'| < 100N^{1/2}$. Decompose ν into 100000 blocks $\nu = \sum_\alpha \nu^{(\alpha)}$ so that for each α and any $y, y' \in \operatorname{supp} \nu^{(\alpha)}$,

$$(5.9) \qquad |y - y'| < \frac{1}{2}N^{1/2}.$$

Write $|S_N(\theta)| \leq \sum_\alpha |S_N^{(\alpha)}(\theta)|$, where

$$S_N^{(\alpha)}(\theta) := \sum_x \sum_y \mu(x)\nu^{(\alpha)}(y)e(\theta \langle x, y\rangle).$$

We will bound each such $S_N^{(\alpha)}$ independently of α, so we drop the superscripts α.

Let $\Upsilon : \mathbb{R}^2 \to \mathbb{R}_+$ be a smooth test function which exceeds 1 on the square $[-1, 1] \times [-1, 1]$ and has Fourier transform supported in a ball of radius 1 about

the origin. Apply Cauchy-Schwarz in the x variable, insert Υ, and open the squares:

$$|S_N(\theta)| \ll \left(\sum_x \mu^2(x)\right)^{1/2} \left(\sum_x \Upsilon\left(\frac{x}{50N^{1/2}}\right) \sum_y \nu(y) \sum_{y'} \nu(y') e(\langle x, y - y'\rangle \theta)\right)^{1/2}.$$

The first parentheses contribute $N^{\delta/2}$ by (5.6) and (5.8). To the last sum on x apply Poisson summation, recalling the support of $\widehat{\Upsilon}$:

$$(5.10) \qquad |S_N(\theta)| \ll N^{\delta/2} \left(\sum_y \nu(y) \sum_{y'} \nu(y') \ N \ \mathbf{1}_{\{\|(y-y')\theta\| < \frac{1}{50N^{1/2}}\}}\right)^{1/2}.$$

Here $\|\cdot\|$ is the distance to the nearest lattice point in $\mathbb{Z}^2$. For such y, y', θ, we have

$$\|(y - y')\frac{a}{q}\| \le \|(y - y')\theta\| + |y - y'||\beta| < \frac{1}{50N^{1/2}} + \frac{1}{2}N^{1/2}\frac{K}{N} < \frac{1}{Q},$$

where we used (5.9) and (5.1). Then $q < Q$ forces $\|(y - y')\frac{a}{q}\| = 0$, or

$$y \equiv y'(q).$$

This being the case, we now have

$$\frac{1}{50N^{1/2}} > \|(y - y')\theta\| = |(y - y')\beta|;$$

that is,

$$|y - y'| \ll \frac{N^{1/2}}{K}.$$

In summary, we have

$$|S_N(\theta)| \ll N^{(\delta+1)/2} \left(\sum_y \nu(y) \sum_{y'} \mathbf{1}_{\left\{\substack{y \equiv y'(q) \\ |y-y'| \ll \frac{N^{1/2}}{K}}\right\}}\right)^{1/2},$$

where we used (5.8). Using $Q < \frac{N^{1/2}}{K}$ and the crudest bound on the y' sum gives

$$|S_N(\theta)| \ll N^{(\delta+1)/2} \left(\sum_y \nu(y) \left(\frac{N^{1/2}}{QK}\right)^2\right)^{1/2} \ll \frac{N^{\delta+1}}{QK},$$

as claimed. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

The bound (5.3) is already conclusive if K is a bit larger than $N^{2(1-\delta)}$.

THEOREM 5.11. *Assume $Q < N^{1/2}$ and $K < N^{1/2}/Q$. Then*

$$(5.12) \qquad\qquad \int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll \frac{(\#\Omega_N)^2}{N}\left\lceil\frac{N^{2(1-\delta)+4\mathfrak{b}}}{K}\right\rceil.$$

Proof. We bound trivially using (5.3):

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll \frac{K}{N} Q^2 \left(\frac{N^{\delta+1}}{QK}\right)^2 \ll N^{4\delta-1}\left[\frac{N^{2(1-\delta)}}{K}\right],$$

and the claim follows from (3.47), on crudely using $\delta < 1$. □

5.2. *Another bilinear forms estimate.* Next we introduce the cross-section of $W_{Q,K}$ for given β:

$$P_{Q,\beta} := \left\{\theta = \frac{a}{q} + \beta : \frac{1}{2}Q \le q < Q, (a,q) = 1\right\}.$$

We will bound using (5.3), giving essentially

$$\int_{W_{Q,K}} |S_N|^2 \ll \sup |S_N| \frac{K}{N} \sup_\beta \sum_{P_{Q,\beta}} |S_N| \ll \frac{N^{2\delta+}}{KQ} \frac{K}{N} \sup_\beta \sum_{P_{Q,\beta}} |S_N|.$$

The trivial bound on $\sum_{P_{Q,\beta}} |S_N|$ is of course $N^{2\delta}Q^2$, so we need to save a little more than a power of Q to get our target bound of less than $N^{4\delta-1}$. This is achieved by exploiting the extra structure in the a and q sums, as follows.

PROPOSITION 5.13. *Let the notation be as above. Then for all $\varepsilon > 0$,*

$$(5.14) \qquad \sum_{\theta \in P_{Q,\beta}} |S_N(\theta)| \ll_\varepsilon N^{2\delta}Q^2 N^{1-\delta+\varepsilon}\left[\frac{1}{Q^{3/2}} + \frac{1}{QN^{1/8}}\right].$$

Proof. The proof is nearly identical to that of Theorem 6.1 in [BK10], but we reproduce it for the reader's convenience. We again use (3.37) to decompose Ω_N into pieces, now grouping by

$$\Omega_N = (\Xi_1 \Xi_2)(\Xi_3 \cdots \Xi_J) = \Omega' \cdot \Omega''.$$

As before, we have for $\gamma \in \Omega'$ and $\omega \in \Omega''$ that

$$(5.15) \qquad |{}^t\gamma e_2| < 300N^{3/4} \qquad \text{and} \qquad |\omega e_2| < 2000N^{1/4}.$$

Also from (3.26), we have

$$(5.16) \qquad \#\Omega' \ll N^{3\delta/2} \qquad \text{and} \qquad \#\Omega'' \ll N^{\delta/2}.$$

Again we define the measures μ and ν on $\mathbb{Z}^2$ by

$$\mu(x) := \sum_{\gamma \in \Omega'} \mathbf{1}_{\{x = {}^t\gamma e_2\}},$$

$$\nu(y) := \sum_{\omega \in \Omega''} \mathbf{1}_{\{y = \omega e_2\}},$$

with $\mu, \nu \le 1$. For any two elements y, y' in the support of ν, we have $|y - y'| < 4000N^{1/4}$. Hence we again decompose ν into $O(1)$ pieces, $\nu = \sum_\alpha \nu^{(\alpha)}$, so as

to make the difference

$$(5.17) \qquad |y - y'| < \frac{1}{10000} N^{1/4}$$

for y, y' in the support of $\nu^{(\alpha)}$. Writing

$$S_N^{(\alpha)}(\theta) = \sum_x \sum_y \mu(x) \nu^{(\alpha)}(y) e(\theta \langle x, y \rangle),$$

and dropping the superscripts α, we proceed to bound

$$\sum_{\theta \in P_{Q,\beta}} |S_N(\theta)| = \sum_{q \asymp Q} \sum_{(a,q)=1} \zeta(\theta) S_N(\theta)$$

$$= \sum_{q \asymp Q} \sum_{(a,q)=1} \zeta(\theta) \sum_x \sum_y \mu(x) \nu(y) e(\theta \langle x, y \rangle),$$

where ζ has modulus 1. Recall the bump function Υ which is at least one on $[-1, 1]^2$; assume now that its Fourier transform is supported in a ball of radius $1/40$ about the origin. Apply Cauchy-Schwarz in the x sum and (5.16), insert the function Υ, reverse orders, and apply Poisson summation:

$$(5.18) \qquad \sum_{\theta \in P_{Q,\beta}} |S_N(\theta)|$$

$$\ll N^{3\delta/4} \left(\sum_x \Upsilon\left(\frac{x}{300 N^{3/4}}\right) \left| \sum_{q \asymp Q} \sum_{(a,q)=1} \zeta(\theta) \sum_y \nu(y) e(\theta \langle x, y \rangle) \right|^2 \right)^{1/2}$$

$$\ll N^{3(\delta+1)/4} \, \mathcal{X}^{1/2},$$

where

$$(5.19) \qquad \mathcal{X} = \mathcal{X}_{Q,\beta} := \sum_q \sum_{q'} \sum_a \sum_{a'} \sum_y \sum_{y'} \nu(y) \nu(y') \mathbf{1}_{\left\{ \|y\theta - y'\theta'\| < \frac{1}{12000 N^{3/4}} \right\}}.$$

Here $\theta' = \frac{a'}{q'} + \beta$; note that β is the same for θ and θ'.

Write $y = (y_1, y_2)$ and the same with y'. Consider the innermost condition in (5.19):

$$(5.20) \qquad \|y_1\theta - y_1'\theta'\|, \|y_2\theta - y_2'\theta'\| < \frac{1}{12000 N^{3/4}}.$$

Recall that $y = \gamma e_2$ for some (nonidentity) $\gamma \in \Gamma$, and the same for y'; hence we have

$$y_1 y_2 y_1' y_2' \neq 0.$$

Also note using (5.20), (5.17), and $|\beta| < K/N < 1/(N^{1/2}Q)$ that

$$(5.21)$$

$$\left\| y_1 \frac{a}{q} - y_1' \frac{a'}{q'} \right\| \leq \|y_1\theta - y_1'\theta'\| + |(y_1 - y_1')\beta| < \frac{1}{12000 N^{3/4}} + \frac{N^{1/4}}{10000 N^{1/2}Q},$$

and similarly with y_2, y_2'.

Let $Y := \begin{pmatrix} y_1 & y_1' \\ y_2 & y_2' \end{pmatrix}$, so that

(5.22) $$\mathcal{Y} := \det(Y) = y_1 y_2' - y_1' y_2.$$

Observe then by (5.21), (5.15), and $Q < N^{1/2}$ that

$$\left\| \mathcal{Y} \frac{a}{q} \right\| \leq \left\| y_2' \left(y_1 \frac{a}{q} - y_1' \frac{a'}{q'} \right) \right\| + \left\| y_1' \left(y_2' \frac{a'}{q'} - y_2 \frac{a}{q} \right) \right\|$$

$$< 2000 N^{1/4} \left(\frac{1}{12000 N^{3/4}} + \frac{N^{1/4}}{10000 N^{1/2} Q} \right) \times 2$$

$$< \frac{1}{Q}.$$

Of course this forces $\mathcal{Y} \equiv 0 (\mathrm{mod}\, q)$. The same argument gives $\mathcal{Y} \equiv 0 (\mathrm{mod}\, q')$, and hence we have

(5.23) $$\mathcal{Y} \equiv 0 (\mathrm{mod}\, \mathfrak{q}),$$

where $\frac{1}{2} Q \leq \mathfrak{q} < Q^2$ is the least common multiple of q and q'.

Decompose $\mathcal{X}$ in (5.19) as $\mathcal{X} = \mathcal{X}_1 + \mathcal{X}_2$ according to whether $\mathcal{Y} = 0$ or not; we handle the two contributions separately. We will prove the following two lemmata.

LEMMA 5.24. *For any $\varepsilon > 0$,*

$$\mathcal{X}_1 \ll_\varepsilon N^{\delta/2+\varepsilon} Q^4 \left[\frac{1}{N^{3/4}} + Q^{-2} \right].$$

LEMMA 5.25. *For any $\varepsilon > 0$,*

$$\mathcal{X}_2 \ll_\varepsilon N^{\delta+\varepsilon} Q.$$

We momentarily postpone the proofs of these two lemmata, first using them to finish the proof of Proposition 5.13. Returning to (5.18), we have

$$\sum_{\theta \in P_{Q,\beta}} |S_N(\theta)| \ll_\varepsilon N^{3(\delta+1)/4+\varepsilon} \left[N^{\delta/2} Q^4 \left(\frac{1}{N^{3/4}} + Q^{-2} \right) + N^\delta Q \right]^{1/2},$$

from which the claim follows using $Q < N^{1/2}$. □

Now we establish the lemmata separately.

5.2.1. *Bounding $\mathcal{X}_2$: the case $\mathcal{Y} \neq 0$.*

Proof of Lemma 5.25. Note from (5.15), (5.17), and (5.22) that

$$|\mathcal{Y}| \leq |y_1(y_2' - y_2)| + |(y_1 - y_1')y_2| < 2000 N^{1/4} \frac{1}{10000} N^{1/4} \times 2 < N^{1/2}.$$

Since $\mathfrak{q} \mid \mathcal{Y}$ and $\mathcal{Y} \neq 0$, we have

$$\mathfrak{q} \leq \min(Q^2, N^{1/2}) \leq Q N^{1/4}.$$

Then (5.21) and $Q < N^{1/2}$ forces

(5.26)
$$y_1 \frac{a}{q} - y_1' \frac{a'}{q'} \equiv 0 (\mathrm{mod}\, 1),$$

and the same holds for y_2, y_2'. Let $\tilde{q} := (q, q')$ and $q = q_1 \tilde{q}$, $q' = q_1' \tilde{q}$ so that $\mathfrak{q} = q_1 q_1' \tilde{q}$. Then (5.26) becomes

$$y_1 a q_1' \equiv y_1' a' q_1 (\mathrm{mod}\, \mathfrak{q}),$$

and the same for y_2, y_2'. Recall a and q are coprime, as are a' and q'. It then follows that $q_1 \mid y_1$, and similarly, $q_1 \mid y_2$. But since the coordinates of y are coprime, $(y_1, y_2) = 1$, this forces $q_1 = 1$. The same argument applies to q_1', so we have $q = q' = \mathfrak{q}$. Then (5.26) now reads

(5.27)
$$y_1 a \equiv y_1' a' (\mathrm{mod}\, q),$$

and similarly for y_2, y_2'.

Hence, once we fix $y, y' \in \Omega'' e_2$, the value of $\mathcal{Y}$ is determined, and $q \mid \mathcal{Y}$ leaves at most N^{ε} choices for q. Then there are at most Q choices for a, from which a' is determined by (5.27) (again using the coprimality of the coordinates of y and y').

Then using (5.16), $\mathcal{X}_2$ is bounded by

(5.28)
$$\mathcal{X}_2 \ll \sum_y \nu(y) \sum_{y'} \nu(y') \sum_{\substack{q \mid \mathcal{Y} \\ \frac{1}{2} Q \leq q < Q}} \sum_{a (\mathrm{mod}\, q)} 1$$
$$\ll_{\varepsilon} \left(N^{\delta/2} \right)^2 N^{\varepsilon} Q,$$

as claimed. $\square$

5.2.2. *Bounding $\mathcal{X}_1$: the case $\mathcal{Y} = 0$.*

Proof of Lemma 5.24. The condition $\mathcal{Y} = 0$ implies $y_1/y_2 = y_1'/y_2'$. Recall that rationals have unique continued fraction expansions (of even length), and thus $y = y'$. The bottom line savings from this fact is at most $N^{1/4}$, whereas we need to save a bit more than Q, which can be as large as $N^{1/2}$.

Let $N' := \frac{1}{12000} N^{3/4}$. The condition (5.20) then becomes

(5.29)
$$\left\| y_1 \left(\frac{a}{q} - \frac{a'}{q'} \right) \right\| < \frac{1}{N'}.$$

Let $(y_1, q) = v$ and $(y_1, q') = v'$ with $q = vr$. Assume without loss of generality that $v \leq v'$. Fix y (for which there are $N^{\delta/2}$ choices) and $v, v' \mid y_1$ (at most N^{ε} choices). There are $\ll Q/v'$ choices for $q' \equiv 0 (\mathrm{mod}\, v')$, and then $\ll Q$ choices for $(a', q') = 1$. Write ψ for $y_1 a'/q' (\mathrm{mod}\, 1)$, which is now fixed, and write $y_1 = vz$ with $(z, r) = 1$. Then (5.29) becomes

$$\left\| z \frac{a}{r} - \psi \right\| < \frac{1}{N'}.$$

Let $\mathcal{U}_z$ be the set of possible fractions $\frac{za}{r} (\mathrm{mod}\, 1)$ as r varies in $Q/(2v) \leq r < Q/v$, and a ranges up to Q subject to $(a, vr) = 1$. Note that distinct points

$u \in \mathcal{U}_z$ are separated by a distance of at least v^2/Q^2. Hence the size of the intersection of $\mathcal{U}_z$ with the interval

$$\left[\psi - \frac{1}{N'}, \psi + \frac{1}{N'}\right]$$

contains at most $\frac{Q^2}{v^2 N'} + 1$ points. Once $u = f/r \in \mathcal{U}_z$ is determined, so is its denominator; that is, r is determined. Also $a(\mathrm{mod}\, r)$ is determined (to be f), hence $a(\mathrm{mod}\, q)$ has v possible values (recall $q = rv$).

In summary, we use (5.16) again to bound $\mathcal{X}_1$ by

$$\mathcal{X}_1 \ll \sum_y \nu(y) \sum_{\substack{v,v'|y_1 \\ v \leq v'}} \sum_{q' \equiv 0 (\mathrm{mod}\, v')} \sum_{(a',q')=1} \sum_{\substack{f/r \in \mathcal{U}_z \cap [\psi - \frac{1}{N'}, \psi + \frac{1}{N'}] \\ q=rv}} \sum_{\substack{a<q \\ a \equiv f(\mathrm{mod}\, r)}} 1$$

$$\ll \sum_y \nu(y) \sum_{\substack{v,v'|y_1 \\ v \leq v'}} \frac{Q}{v'} Q \left(\frac{Q^2}{v^2 N'} + 1\right) v$$

$$\ll_\varepsilon N^{\delta/2} N^\varepsilon Q^2 \left(\frac{Q^2}{N^{3/4}} + 1\right),$$

as claimed. $\qquad\square$

With the lemmata established, we have completed the proof of Proposition 5.13.

5.3. *The bound for Q large.* Lastly, we input this bound to get another bound on the main integral, one which is favorable as long as Q is a bit bigger than $N^{4(1-\delta)}$.

THEOREM 5.30. *Assume that $Q < N^{1/2}$ and $KQ < N^{1/2}$. Then*

$$(5.31) \qquad \int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll \frac{(\#\Omega_N)^2}{N} N^{2(1-\delta)} N^{4\mathfrak{b}} \left(\frac{1}{Q^{1/2}} + \frac{1}{N^{1/8}}\right).$$

Proof. Write

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll \sup_{\theta \in W_{Q,K}} |S_N(\theta)| \cdot \frac{K}{N} \sup_{|\beta| \asymp \frac{K}{N}} \sum_{\theta \in P_{Q,\beta}} |S_N(\theta)|$$

$$\ll_\varepsilon N^{2\delta} \left(\frac{N^{1-\delta}}{KQ}\right) \cdot \frac{K}{N} \left(N^{2\delta} Q^2 N^{1-\delta+\varepsilon} \left[\frac{1}{Q^{3/2}} + \frac{1}{QN^{1/8}}\right]\right)$$

$$\ll N^{4\delta-1} N^{2(1-\delta)+\varepsilon} \left(\frac{1}{Q^{1/2}} + \frac{1}{N^{1/8}}\right),$$

where we used (5.3) and (5.14). The claim follows from (3.47), again crudely using $\delta < 1$. $\qquad\square$

It remains to handle the regions when both K and Q are very small, less than N^ε for ε near zero.

6. **Minor arcs analysis II**

We now push the methods of the previous section down to the level of Q and K being of constant size. We again do this in two stages. But first we record the following counting bound.

LEMMA 6.1. *For $(qK)^{13/5} < Y < X$, and vectors $\eta, \eta' \in \mathbb{Z}^2$ having coprime coordinates with $|\eta| \asymp X/Y$ and $|\eta'| \asymp Y$,*

$$\# \left\{ \gamma \in \mathrm{SL}_2(\mathbb{Z}) : \|\gamma\| \asymp Y, |\gamma\eta - \eta'| < \frac{X}{YK}, \ \text{and } \gamma\eta \equiv \eta' (\mathrm{mod}\, q) \right\} \ll \left(\frac{Y}{qK} \right)^2.$$

The implied constant is absolute, depending on the implied constants above.

Sketch of proof. Write $G(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z})$, and let $G_\eta(q)$ be the stabilizer of η mod q:

$$G_\eta(q) := \{ \gamma \in G(\mathbb{Z}) : \gamma\eta \equiv \eta(q) \}.$$

Then $G(\mathbb{Z}) \cong (G(\mathbb{Z})/G_\eta(q)) \times G_\eta(q)$. Let $R = R_{Y,K}$ denote the region

$$R := \{ g \in \mathrm{SL}_2(\mathbb{R}) : \|g\| \asymp Y, |g\eta - \eta'| < X/(YK) \}.$$

The methods in [Goo83] (see also [BKS10]) give an estimate of the form

$$\sum_{\gamma \in G(\mathbb{Z})} \mathbf{1}_{\{\gamma \in R\}} \mathbf{1}_{\gamma\eta \equiv \eta'(\mathrm{mod}\, q)} = \sum_{\omega \in G(\mathbb{Z})/G_\eta(q)} \mathbf{1}_{\omega\eta \equiv \eta'(\mathrm{mod}\, q)} \sum_{\gamma' \in G_\eta(q)} \mathbf{1}_{\{\omega\gamma' \in R\}}$$

$$\ll_\varepsilon \sum_{\omega \in G(\mathbb{Z})/G_\eta(q)} \mathbf{1}_{\omega\eta \equiv \eta'(\mathrm{mod}\, q)} \left(\left(\frac{Y}{qK} \right)^2 + Y^{2\Theta + \varepsilon} \right)$$

$$\ll \left(\frac{Y}{qK} \right)^2 + Y^{2\Theta + \varepsilon},$$

where $\Theta = 1/2 + 7/64$ is the best known bound towards the Ramanujan conjectures [Kim03]. (We apply the argument to a smoothed sum.) The first term dominates as long as $(qK)^2 < Y^{25/32 - \varepsilon}$, and the claim follows using $64/25 + \varepsilon < 65/25 = 13/5$. $\qquad\square$

Remark 6.2. Recall that we are not interested here in optimizing the final value of δ_0 in Theorem 1.8, so we allow ourselves to be a bit crude in the above for the sake of exposition.

6.1. *The bound for K at least a small power of Q.* We return to the approach of Section 5.1, that is just bounding the sup norm and needing to win more than $K^{1/2}Q$ off the trivial bound. Now we use the fact that KQ is quite small to improve on the trivial bound by $(KQ)^{1-\varepsilon}$ with ε small (depending on the distance from δ to 1). Then we will have, roughly

$$\int_{W_{Q,K}} |S_N|^2 \ll Q^2 \frac{K}{N} \left(\frac{\#\Omega_N}{(KQ)^{1-\varepsilon}} \right)^2 = \frac{(\#\Omega_N)^2}{N} \left(\frac{Q^\varepsilon}{K^{1-\varepsilon}} \right),$$

which is a savings as long as K is at least a small power of Q. Note now for K and Q small that we must be careful with the loss in the size of Ω_N in the lower bound (3.47). We make all of this precise below.

PROPOSITION 6.3. *Assume* $\theta \in W_{Q,K}$ *with*

$$(6.4) \qquad\qquad 1 \ll KQ < N^{5/52}.$$

Then

$$(6.5) \qquad\qquad |S_N(\theta)| \ll \#\Omega_N \left(\frac{e^{c(\log\log(KQ))^2}}{(KQ)^{1-(1-\delta)52/5}} \right).$$

Proof. Recalling (3.35) that $N_j \asymp N^{1/2^j}$, we can find a $1 \leq j \leq J$ so that

$$(6.6) \qquad\qquad \frac{1}{100}(QK)^{13/5} < N_j < (QK)^{26/5},$$

say. Here we used (6.4) that $(QK)^{26/5} < N^{1/2}$.

Define the sets

$$(6.7) \qquad\qquad \Omega^{(1)} := \Xi_1 \Xi_2 \cdots \Xi_{j-1},$$
$$\Omega^{(2)} := \Xi_j,$$
$$\Omega^{(3)} := \Xi_{j+1}\Xi_{j+2} \cdots \Xi_J.$$

Hence for $g_i \in \Omega^{(i)}$,

$$(6.8) \qquad\qquad \lambda(g_3) \sim N_{j+1}N_{j+2} \cdots N_J =: M,$$
$$(6.9) \qquad\qquad \lambda(g_2) \sim N_j,$$
$$(6.10) \qquad\qquad \lambda(g_1) \asymp \frac{N}{M \, N_j}.$$

Note that

$$(6.11) \qquad\qquad \frac{N_j}{\log N_j} \ll M \ll N_j \log N_j$$

and that from (3.48) and (3.26) we have

$$(6.12) \qquad |\Omega^{(3)}| \gg \frac{M^{2\delta}}{e^{c(\log\log M)^2}}, \qquad |\Omega^{(2)}| \gg \frac{(N_j)^{2\delta}}{(\log N_j)^3}.$$

In the above, we used that $J - j \asymp \log\log M$.

Estimate

$$(6.13) \qquad |S_N(\theta)| \ll \sum_{g_1 \in \Omega^{(1)}} \sum_{g_3 \in \Omega^{(3)}} \left| \sum_{g_2 \in \Omega^{(2)}} e(\langle g_3 e_2, {}^t g_2 {}^t g_1 e_2 \rangle \theta) \right|.$$

Fix g_1 and set $\eta = {}^t g_1 e_2$. Note that

$$(6.14) \qquad\qquad |\eta| \asymp \frac{N}{MN_j}.$$

Estimate as in (5.10):

(6.15)

$$\sum_{g_3 \in \Omega^{(3)}} \left| \sum_{g_2 \in \Omega^{(2)}} e(\langle g_3 e_2, {}^t g_2 \eta \rangle \theta) \right|$$

$$\ll (\#\Omega^{(3)})^{1/2} \, M \left[\# \left\{ (g,g') \in {}^t\Omega^{(2)} \times {}^t\Omega^{(2)} : \begin{matrix} \| \langle (g-g')\eta, e_1 \rangle \| \ll \frac{1}{M} \\ \| \langle (g-g')\eta, e_2 \rangle \| \ll \frac{1}{M} \end{matrix} \right\} \right]^{1/2} ,$$

where we extended the sum over g_3 to $g_3 e_2 \in \{z \in \mathbb{Z}^2 : |z| \ll M\}$. Write

(6.16) $$\left\| \langle (g-g')\eta, e_i \rangle \frac{a}{q} \right\| = \| \langle (g-g')\eta, e_i \rangle \theta \| + | \langle (g-g')\eta, e_i \rangle \beta |,$$

where

$$| \langle (g-g')\eta, e_i \rangle \beta | \ll N_j \frac{N}{MN_j} \frac{K}{N} = \frac{K}{M}.$$

From (6.6) and (6.11) we clearly have $\frac{K}{M} < \frac{1}{Q}$, so (6.16) forces

(6.17) $$(g-g')\eta \equiv 0(q)$$

and

(6.18) $$|(g-g')\eta| \ll \frac{1}{M|\beta|} \ll \frac{N}{KM}.$$

Fix g', and enlarge $g \in {}^t\Omega^{(2)}$ to $\{g \in \mathrm{SL}_2(\mathbb{Z}) : \|g\| \ll N_j\}$. Applying Lemma 6.1 with $\eta' = g'\eta$, $X = N/M$, and $Y = N_j$, the g cardinality contributes

$$\ll \left(\frac{N_j}{KQ} \right)^2.$$

Thus we have by (6.12) and (6.11) that

$$(6.15) \ll \left(\#\Omega^{(3)} \cdot \#\Omega^{(2)} \right)^{1/2} \frac{N_j^2}{KQ}$$

$$\ll \#\Omega^{(3)} \cdot \#\Omega^{(2)} \frac{(MN_j)^{1-\delta} e^{c(\log\log M)^2} (\log N_j)^3}{KQ}.$$

Hence by (6.6) and (6.11),

$$(6.13) \ll \#\Omega_N \frac{(KQ)^{(1-\delta)52/5} e^{c(\log\log(KQ))^2}}{KQ},$$

as claimed. $\qquad\square$

Inserting this bound into the main integral and estimating trivially gives

THEOREM 6.19. *Assuming* (6.4),

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll \frac{(\#\Omega_N)^2}{N} \frac{Q^{(1-\delta)104/5} e^{c(\log\log(KQ))^2}}{K^{1-(1-\delta)104/5}}.$$

This bound is sufficient for our purposes unless K is much less than

$$(6.20) \qquad Q^{\frac{104/5(1-\delta)}{1-104/5(1-\delta)}} \approx Q^{\varepsilon}.$$

6.2. *The bound for K even smaller.* In this last section, we give the final bound for minor arcs, which we apply to the remaining range of K. Recall the approach of Section 5.2: we bound $\int_{W_{Q,K}} |S_N|^2$ by the sup norm times K/N times $\sum_{P_{\beta,Q}} |S_N|$. The sup norm has already gained almost KQ, so we need to gain more than a power of Q off of the last summation. We proceed as follows.

PROPOSITION 6.21. *Recall the cross section $P_{Q,\beta}$ for a given $|\beta| \asymp \frac{K}{N}$:*

$$P_{Q,\beta} := \left\{ \theta = \frac{a}{q} + \beta : q \asymp Q, (a,q) = 1 \right\}.$$

Then assuming (6.4), *we have*

$$(6.22) \qquad \sum_{\theta \in P_{Q,\beta}} |S_N(\theta)| \ll \#\Omega_N \, Q^2 \left(\frac{(KQ)^{(1-\delta)52/5} e^{c(\log\log(KQ))^2}}{Q^{3/2}} \right).$$

Proof. This argument is similar to Proposition 5.13, and we sketch the proof. Using the same decomposition (6.7), we follow (5.18) and bound the left-hand side of (6.22) by

(6.23)

$$\ll \sum_{g_1 \in \Omega^{(1)}} \left(\#\Omega^{(3)} \right)^{1/2} M$$

$$\times \left[\# \left\{ (\theta, \theta', g, g') \in P_\beta \times P_\beta \times {}^t\Omega^{(2)} \times {}^t\Omega^{(2)} : \|(g\theta - g'\theta')\eta\| \ll \frac{1}{M} \right\} \right]^{1/2},$$

where $\eta = {}^t g_1 e_2$. The innermost condition guarantees $q = q'$ and

$$a(g\eta) \equiv a'(g'\eta) \pmod{q}.$$

The number of choices for g' given g, q, a, and a' is

$$\ll \left(\frac{N_j}{Q} \right)^2,$$

hence

$$(6.23) \ll \#\Omega^{(1)} \left(\#\Omega^{(3)} \right)^{1/2} M \left[QQ^2 \#\Omega^{(2)} \left(\frac{N_j}{Q} \right)^2 \right]^{1/2}$$

$$\ll \#\Omega_N \, Q^2 \left(\frac{(KQ)^{(1-\delta)52/5} e^{c(\log\log(KQ))^2}}{Q^{3/2}} \right),$$

as claimed. $\qquad\qquad\square$

Using (6.22) and (6.5), we now have the bound

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll \frac{K}{N} \#\Omega_N \left(\frac{e^{c(\log\log(KQ))^2}}{(KQ)^{1-(1-\delta)52/5}} \right)$$

$$\times \#\Omega_N \, Q^2 \left(\frac{(KQ)^{(1-\delta)52/5} e^{c(\log\log(KQ))^2}}{Q^{3/2}} \right),$$

from which we immediately have

THEOREM 6.24. *Assuming* (6.4),

$$\int_{W_{Q,K}} |S_N(\theta)|^2 d\theta \ll \frac{(\#\Omega_N)^2}{N} \left(\frac{(KQ)^{(1-\delta)104/5} e^{c(\log\log(KQ))^2}}{Q^{1/2}} \right).$$

7. **Proofs of Theorems 1.8 and 1.25**

Keeping all the notation of previous sections, we now prove the main minor arcs estimate, analogous to (1.39), before completing a proof of Theorem 1.8.

THEOREM 7.1. *Assume*

$$\delta > \delta_0, \tag{7.2}$$

with δ_0 given by (1.21). *Then for some $c > 0$,*

$$\sum_{n\in\mathbb{Z}} |\mathcal{E}_N(n)|^2 \ll \frac{|\Omega_N|^2}{N} \mathcal{Q}^{-c}. \tag{7.3}$$

Proof. By Parseval, we have

$$\sum_{n\in\mathbb{Z}} |\mathcal{E}_N(n)|^2 = \int_0^1 |1 - \Psi_{\mathcal{Q},N}(\theta)|^2 |S_N(\theta)|^2 d\theta = \int_{\mathfrak{M}_{\mathcal{Q}}} + \int_{\mathfrak{m}},$$

where we decomposed the integral into the major arcs $\mathfrak{M}_{\mathcal{Q}}$ and the complementary minor arcs $\mathfrak{m} = [0,1] \setminus \mathfrak{M}_{\mathcal{Q}}$.

On the major arcs, note from (4.22) that $1 - \psi(x) = |x|$ on $[-1,1]$. Then using (6.5) with $K \asymp N|\beta|$ gives

$$\int_{\mathfrak{M}_{\mathcal{Q}}} \ll \sum_{q<\mathcal{Q}} \sum_{(a,q)=1} \int_{|\beta|<\mathcal{Q}/N} \left| \frac{N}{\mathcal{Q}}\beta \right|^2 \left(|\Omega_N| \left(\frac{1}{(N|\beta|\mathcal{Q})^{1-c}} \right) \right)^2 d\beta$$

$$\ll \frac{|\Omega_N|^2}{N} \frac{1}{\mathcal{Q}^{1-4c}}.$$

Here $0 < c < (1-\delta)52/5 < 1/4$ by (7.2), so renaming the constant $c > 0$, we are done with the major arcs.

Decompose the minor arcs $\mathfrak{m}$ into dyadic regions

$$\int_{\mathfrak{m}} |S_N(\theta)|^2 d\theta \ll \sum_{\substack{Q < N^{1/2} \\ \text{dyadic}}} \sum_{\substack{K < \frac{N^{1/2}}{Q} \\ \text{dyadic}}} \mathcal{I}_{Q,K},$$

where at least one of Q or K exceeds $\mathcal{Q}$, and

$$\mathcal{I}(Q, K) := \int_{W_{Q,K}} |S_N(\theta)|^2 d\theta.$$

Write $Q = N^\alpha$, $K = N^\kappa$, with the parameters (α, κ) ranging in

(7.4) $$0 \le \alpha < 1/2 \text{ and } 0 \le \kappa < 1/2 - \alpha.$$

It will be convenient to define

(7.5) $$\eta := (1 - \delta)104/5.$$

Assume that $1 - \delta < 5/208$, so that $0 < \eta < 1/2$. We break the summation into the following four ranges:

$$\begin{aligned}
\mathcal{R}_1 &:= \{(\alpha, \kappa) : \kappa > 2(1 - \delta) + 4\mathfrak{b}\}, \\
\mathcal{R}_2 &:= \{(\alpha, \kappa) : \alpha > 4(1 - \delta) + 8\mathfrak{b}\}, \\
\mathcal{R}_3 &:= \{(\alpha, \kappa) : \eta(\alpha + \kappa) < \kappa \text{ and } \alpha + \kappa < 5/52\}, \\
\mathcal{R}_4 &:= \{(\alpha, \kappa) : \eta(\alpha + \kappa) < \tfrac{1}{2}\alpha \text{ and } \alpha + \kappa < 5/52\}.
\end{aligned}$$

We need to show that these four regions cover the entire range (7.4). Using (3.3) and (7.2) with (1.21) guarantees that the regions $\mathcal{R}_1$ and $\mathcal{R}_2$ certainly cover the range $\alpha + \kappa \ge 5/52$. In the complimentary range, $\mathcal{R}_3$ and $\mathcal{R}_4$ give two regions: the region below the line through the origin with slope $\eta/(1 - \eta)$, and the region above the line through the origin with slope $(1/2 - \eta)/\eta$. These two regions overlap when the slopes overlap, that is, when $\eta < 1/3$. Then (7.5) explains the value of δ_0 in (1.21).

Since the four regions cover the full range (7.4), we now just collect the results of the previous two sections. In the range $\mathcal{R}_1$, we apply Theorem 5.11, getting

(7.6) $$\mathcal{I}(Q, K) \ll \frac{(\#\Omega_N)^2}{N} K^{-c}$$

for some $c > 0$. In $\mathcal{R}_2$, we apply Theorem 5.30, getting

(7.7) $$\mathcal{I}(Q, K) \ll \frac{(\#\Omega_N)^2}{N} Q^{-c}.$$

In the range $\mathcal{R}_3$ with $K > \mathcal{Q}$, Theorem 6.19 gives (7.6), and in $\mathcal{R}_4$ with $Q > \mathcal{Q}$, Theorem 6.24 gives (7.7). Combining these estimates completes the proof of (7.3). $\qquad\square$

It is now standard to derive Theorem 1.8 from (4.29) and (7.3).

Proof of Theorem 1.8, *assuming Proposition* 3.9. In light of (4.29), the proof of which uses Proposition 3.9, we have for $n \asymp N$ that

$$\mathcal{M}_N(n) \gg |\Omega_N|/(N \log \log N)$$
$$\gg N^{2\delta - 1 - 1/1000},$$

where we crudely used (3.47) and (3.3). Hence we expect the same for $R_N(n)$. If this is not the case, it means that

$$|\mathcal{E}_N(n)| = |R_N(n) - \mathcal{M}_N(n)| \gg \frac{1}{\log \log N} \frac{|\Omega_N|}{N}.$$

Let $\mathfrak{E}(N)$ denote the set of $n \asymp N$ which have a small representation number $R_N(n)$:

$$\mathfrak{E}(N) := \left\{ \frac{1}{20} N \le n < \frac{1}{10} N : R_N(n) < \frac{1}{2} \mathcal{M}_N(n) \right\}.$$

Then assuming (7.2), we have

$$\#\mathfrak{E}(N) \ll \sum_{\frac{1}{20} N \le n < \frac{1}{10} N} \mathbf{1}_{\{|\mathcal{E}_N(n)| \gg \frac{|\Omega_N|}{N \log \log N}\}}$$
$$\ll \frac{N^2 (\log \log N)^2}{|\Omega_N|^2} \sum_n |\mathcal{E}_N(n)|^2$$
$$\ll \frac{N^2 (\log \log N)^2}{|\Omega_N|^2} \frac{|\Omega_N|^2}{N} \mathcal{Q}^{-c} \ll N e^{-c\sqrt{\log N}},$$

using (7.3) and (3.5). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

This completes the proof of Theorem 1.8, modulo the construction of the leading set $\aleph$, which is taken up in the next section. First we give a quick

7.1. *Proof of Theorem* 1.25. Let $\mathcal{P} = \mathcal{P}_N$ be the set of primes p up to N which are $3 \pmod 4$, so that $(p-1)/2$ is a 10-almost-prime; that is,

$$\mathcal{P} := \{ p < N : p \equiv 3 \pmod 4, \text{ and } m \mid (p-1)/2 \Longrightarrow m > N^{1/10} \}.$$

A standard sieve argument shows that $\mathcal{P}$ has cardinality $\gg \frac{N}{(\log N)^2}$. By (1.9), the cardinality $N e^{-c\sqrt{\log N}}$ of the exceptional set is much smaller, and so $\mathfrak{D}_\mathcal{A}(N) \cap \mathcal{P}$ is unbounded in N for $\delta_\mathcal{A} > \delta_0$.

By (1.10), each $p = d$ in the intersection appears with multiplicity at least $N^{2\delta - 1.001} > N^{10/11}$, say. That is, there are distinct $b_1, \ldots, b_L$ so that $b_j/d \in \mathfrak{R}_\mathcal{A}$, $j = 1, \ldots, L$, and $L > N^{10/11}$. Let r be any primitive root mod d. For $j = 1, \ldots, L$, let k_j be defined by $b_j \equiv r^{k_j} \pmod d$, and let $K = \{k_1, \ldots, k_L\}$. Of course b_j is a primitive root mod d if and only if $(k_j, d-1) = 1$.

Consider the subset K' of $k \in K$ for which $(k, d-1) > 2$. Since $d \in \mathcal{P}$, each such k has a prime factor of size $N^{1/10}$, and hence the cardinality of K'

is $\ll N^{9/10}$. This is less than the cardinality of K, so we may safely discard K' from K, leaving a nonempty set K''.

Consider $b \equiv r^k (\mathrm{mod}\, d)$ with $k \in K''$. If $(k, d-1) = 1$, we are done, since b is a root mod d and $b/d \in \mathfrak{R}_\mathcal{A}$. The only other possibility is $(k, d-1) = 2$, whence b is a square mod d. Set $b' := d - b$, so $b' \equiv -r^k (\mathrm{mod}\, d)$; since $d \equiv 3 (\mathrm{mod}\, 4)$, b' is now a primitive root. It is elementary to verify that $b/d \in \mathfrak{R}_{\{1,2,\dots,A\}}$ implies that $b'/d = 1 - b/d \in \mathfrak{R}_{\{1,2,\dots,A+1\}}$. That is, these quotients are still absolutely Diophantine, completing the proof.

8. Construction of ℵ

In this section, we arrange the special leading set ℵ in the ensemble Ω_N as described in Section 3.1. We need two pieces of background, using Section 8.1 to extract some modular/archimedean counting statements from [BGS11], and spending Section 8.2 proving a certain "randomness extraction argument." Finally, we proceed in Section 8.3 to construct ℵ, thereby proving Proposition 3.9 and finalizing the proof of Theorem 1.8.

8.1. *Congruence counting theorems.* Recall from Section 3.2 that μ is the δ-dimensional Hausdorff measure supported on the limit set $\mathfrak{C}$, lifted to $\mathbb{P}^1$. Extending the work of Lalley's [Lal89] to the congruence setting, Bourgain-Gamburd-Sarnak [BGS11, Th. 1.5] proved the following theorem, adapted to our present context.

THEOREM 8.1 ([BGS11]). *Let $\Gamma = \Gamma_\mathcal{A}$ be the semigroup in* (1.31). *There exist an integer*

$$(8.2) \qquad\qquad \mathfrak{B} = \mathfrak{B}(\mathcal{A}) \geq 1$$

and a constant

$$(8.3) \qquad\qquad \mathfrak{c} = \mathfrak{c}(\mathcal{A}) > 0$$

so that the following holds. For any $(q, \mathfrak{B}) = 1$, any $\omega \in \mathrm{SL}_2(q)$, any $\gamma_0 \in \Gamma$, and parameters $T, H \to \infty$ with $H < e^{c\sqrt{\log T}}$, there is a constant $C(\gamma_0) > 0$ so that

$$(8.4) \qquad \# \left\{ \gamma \in \Gamma : \gamma \equiv \omega (\mathrm{mod}\, q), |v_+(\gamma) - \mathfrak{v}| < \frac{1}{H},\ \text{and}\ \frac{\|\gamma\gamma_0\|}{\|\gamma_0\|} \leq T \right\}$$

$$= C(\gamma_0) \cdot T^{2\delta} \frac{\mu(\mathcal{I})}{|\mathrm{SL}_2(q)|} + O\left(T^{2\delta} e^{-\mathfrak{c}\sqrt{\log T}}\right), \qquad \text{as } T \to \infty.$$

Here $\mathcal{I}$ is the interval of length $1/H$ about $\mathfrak{v}$, and the implied constant does not depend on T, H, q, ω, or γ_0.

With the same conditions as above, except for a modulus q with $\mathfrak{B} \mid q$, we have

(8.5)
$$\# \left\{ \gamma \in \Gamma : \gamma \equiv \omega(\mathrm{mod}\, q), v_+(\gamma) \in \mathcal{I}, \ and \ \frac{\|\gamma\gamma_0\|}{\|\gamma_0\|} \leq T \right\}$$
$$= \frac{|\mathrm{SL}_2(\mathfrak{B})|}{|\mathrm{SL}_2(q)|} \cdot \# \left\{ \gamma \in \Gamma : \gamma \equiv \omega(\mathrm{mod}\, \mathfrak{B}), v_+(\gamma) \in \mathcal{I}, \ and \ \frac{\|\gamma\gamma_0\|}{\|\gamma_0\|} \leq T \right\}$$
$$+ O\left(T^{2\delta} e^{-\mathfrak{c}\sqrt{\log T}} \right).$$

Remark 8.6. Recall that throughout, the appearance of constants c and C may change from line to line. The special constant $\mathfrak{c}$ in (8.3) is in contradistinction with this principle, being the same constant wherever it appears. Moreover, Proposition 3.16 follows immediately from (8.4), so the constant $\mathfrak{c}$ appearing there can be taken to be the same as the one here.

Remark 8.7. In Theorem 8.1, we have stated the result only for the extreme cases $(\mathfrak{B}, q) = 1$ and $\mathfrak{B} \mid q$. Of course intermediate cases can be obtained by summing over suitable arithmetic progressions. This introduces no extra error since the number of terms is $\ll_\mathfrak{B} 1$, and our implied constants may depend on the fixed alphabet $\mathcal{A}$ (recall $\mathfrak{B}$ depends only on $\mathcal{A}$).

Remark 8.8. The proof of Theorem 8.1 is the same as that of [BGS11, Th. 1.5], with the following caveats. The latter was proved under the further assumptions that the modulus q is square-free and that Γ is a convex-cocompact sub*group* of $\mathrm{SL}_2(\mathbb{Z})$. As discussed in Section 3.2, the proof also works when the group is replaced by our free semigroup Γ with no parabolic elements. As for the level, taking q square-free was enough for the sieving purposes in [BGS11], but for the circle method used here, we must take arbitrary q. The main ingredient in analyzing the modular aspect (see [BGS11, Lemma 7.2]) was the spectral gap (expansion property) proved using methods of additive combinatorics and an L^2-flattening lemma in [BG08], [BGS10], again for square-free q. The relevant results have since been established for arbitrary modulus; see [BV12] and [GV12, Rem. 30]. With this input, [BGS11, Prop. 12.1] holds for arbitrary q, from which one can derive Theorem 8.1.

Remark 8.9. A final caveat is that there is a minor inaccuracy in the statement of [BGS11, Th. 1.5], in that one must "unsmoothe" the smoothed estimate given in [BGS11, Prop. 12.1] to obtain [BGS11, Th. 1.5]. Applying a Tauberian theorem, the removal of smooth weights results in a loss in the stated error term where the smooth error term $T^{-c/\log\log T}$ must be replaced in the sharp cutoff by $e^{-c\sqrt{\log T}}$; hence the error terms stated in Theorem 8.1.

The condition $\|\gamma\gamma_0\|/\|\gamma_0\| < T$ arises naturally in the above through the renewal method; in fact, this condition is essentially equivalent to

$$d_{\mathbb{H}}(\gamma\gamma_0 i, i) - d_{\mathbb{H}}(\gamma_0 i, i) < C \log T,$$

where $d_{\mathbb{H}}(\cdot, \cdot)$ denotes hyperbolic distance in the upper half-plane $\mathbb{H}$. As in Section 3.2, one typically sets $\gamma_0 = I$, but we will use γ_0 for a different purpose. Namely, we will need to control both the expanding direction $v_+(\gamma)$, and its expanding eigenvalue $\lambda(\gamma)$, but taking $\gamma_0 = I$ gives us control only on the norm $\|\gamma\|$ (which can be off by a constant from the eigenvalue). So we instead do the following.

It is easy to see from (4.10) that

$$(8.10) \qquad \|\gamma\| = \frac{\lambda(\gamma)}{|\langle v_+(\gamma), v_-^{\perp}(\gamma)\rangle|}\left(1 + O\left(\frac{1}{\|\gamma\|^2}\right)\right).$$

Assume now that both γ and γ_0 lie in $\mathcal{I}$, the interval of length $1/H$ about $\mathfrak{v}$, and assume $\|\gamma\|, \|\gamma_0\| > H$. Then applying (8.10) to γ_0 and $\gamma\gamma_0$ gives

$$\|\gamma_0\| = \frac{\lambda(\gamma_0)}{|\langle \mathfrak{v}, v_-^{\perp}(\gamma_0)\rangle|}\left(1 + O\left(\frac{1}{H}\right)\right)$$

and

$$\begin{aligned}\|\gamma\gamma_0\| &= \frac{\lambda(\gamma\gamma_0)}{|\langle v_+(\gamma\gamma_0), v_-^{\perp}(\gamma\gamma_0)\rangle|}\left(1 + O\left(\frac{1}{\|\gamma\gamma_0\|^2}\right)\right)\\ &= \frac{\lambda(\gamma)\lambda(\gamma_0)}{|\langle \mathfrak{v}, v_-^{\perp}(\gamma_0)\rangle|}\left(1 + O\left(\frac{1}{H}\right)\right),\end{aligned}$$

where we used (2.6) and (2.7). On dividing, we obtain

$$\frac{\|\gamma\gamma_0\|}{\|\gamma_0\|} = \lambda(\gamma)\left(1 + O\left(\frac{1}{H}\right)\right).$$

We can thus convert statements restricting norms into ones controlling eigenvalues, without losing constants.

The constant $C(\gamma_0)$ in (8.4) approaches a constant $C(\mathfrak{v}) > 0$ as $\|\gamma_0\| \to \infty$ with $v_+(\gamma_0) \to \mathfrak{v}$; indeed, $C(\gamma_0)$ is obtained by evaluating a certain Gibbs measure. (See [BGS11, §10] or [Lal89, (2.5)], where his x plays the role of our γ_0.) Hence (8.4) can be replaced by

$$\#\left\{\gamma \in \Gamma : \gamma \equiv \omega(\mathrm{mod}\, q), |v_+(\gamma) - \mathfrak{v}| < \frac{1}{H}, \text{ and } \lambda(\gamma) \le T\right\}$$

$$= C(\mathfrak{v}) \cdot T^{2\delta}\frac{\mu(\mathcal{I})}{|\mathrm{SL}_2(q)|}\left(1 + O\left(\frac{1}{H}\right)\right) + O\left(T^{2\delta}e^{-\mathfrak{c}\sqrt{\log T}}\right)$$

and a similar expression analogous to (8.5).

Finally, we can restrict $\lambda(\gamma)$ to a smaller range, $\lambda(\gamma) = T(1 + O(1/H_1))$, for a smaller parameter H_1; in applications we take $H_1 = H^{1/2}$. In summary, we have the following

COROLLARY 8.11. *With notation as above, we have for any* $T, H, H_1 \to \infty$ *with* $H_1 = o(H)$ *and* $H < e^{c\sqrt{\log T}}$, *and any* $(q, \mathfrak{B}) = 1$, $\omega \in \mathrm{SL}_2(q)$, *that*

$$(8.12) \quad \#\left\{\gamma \in \Gamma : \gamma \equiv \omega(\mathrm{mod}\, q), |v_+(\gamma) - \mathfrak{v}| < \frac{1}{H}, |\lambda(\gamma) - T| < \frac{T}{H_1}\right\}$$

$$= C(\mathfrak{v}) \cdot \frac{T^{2\delta}}{H_1} \frac{\mu(\mathcal{I})}{|\mathrm{SL}_2(q)|} \left(1 + O\left(\frac{1}{H_1} + \frac{H_1}{H}\right)\right) + O\left(T^{2\delta} e^{-\mathfrak{c}\sqrt{\log T}}\right).$$

The implied constants are independent of T, H, H_1, q, *and* ω.

For a modulus $q \equiv 0(\mathrm{mod}\,\mathfrak{B})$, *we have*

$$(8.13)$$

$$\#\left\{\gamma \in \Gamma : \gamma \equiv \omega(\mathrm{mod}\, q), v_+(\gamma) \in \mathcal{I}, |\lambda(\gamma) - T| < \frac{T}{H'}\right\}$$

$$= \frac{|\mathrm{SL}_2(\mathfrak{B})|}{|\mathrm{SL}_2(q)|} \cdot \#\left\{\gamma \in \Gamma : \gamma \equiv \omega(\mathrm{mod}\,\mathfrak{B}), v_+(\gamma) \in \mathcal{I}, |\lambda(\gamma) - T| < \frac{T}{H'}\right\}$$

$$\times \left(1 + O\left(\frac{1}{H_1} + \frac{H_1}{H}\right)\right) \quad + \quad O\left(T^{2\delta} e^{-\mathfrak{c}\sqrt{\log T}}\right).$$

8.2. *A randomness extraction argument.* Corollary 8.11 gives us good modular/achimedean control away from the modulus $\mathfrak{B}$, but we need $\aleph$ to have good distribution properties for all moduli. So in the next subsection we will concoct certain special sets engineered to have good equidistribution mod $\mathfrak{B}$. But in so doing, we will potentially ruin the distribution away from $\mathfrak{B}$. To recover this distribution, we apply a certain more-or-less standard "randomness extraction" argument, which states roughly that if a large set has good modular distribution, then so does a sufficiently large random subset of it. We will need to have the flexibility to stay away from a modulus q_0, which in applications is either 1 or $\mathfrak{B}$.

LEMMA 8.14. *Let* $\mu = \mu_S$ *be the normalized (probability) measure of a finite subset* $S \subset \mathrm{SL}(2, \mathbb{Z})$,

$$\mu(\gamma) = \frac{1}{|S|} \sum_{s \in S} \mathbf{1}_{\{s = \gamma\}},$$

and fix $\eta > 0$. *Let* $q_0 < Q$ *be a fixed modulus, let* $\omega_0 \in \mathrm{SL}_2(q_0)$ *be a fixed element, and let* $\mathfrak{Q} = \mathfrak{Q}_{q_0} \subset [1, Q]$ *be the set of moduli* $q < Q$ *with* $q_0 \mid q$. *Assume that for all* $q \in \mathfrak{Q}$ *and all* $\omega \in \mathrm{SL}_2(q)$ *with* $\omega \equiv \omega_0(\mathrm{mod}\, q_0)$, *the projection*

$$\pi_q[\mu](\omega) = \sum_{\gamma \equiv \omega(\mathrm{mod}\, q)} \mu(\gamma)$$

is near the uniform measure on $\mathrm{SL}_2(q)$ conditioned on being $\equiv \omega_0 (\mathrm{mod}\, q_0)$,

(8.15)
$$
\left\| \pi_q[\mu] - \frac{|\mathrm{SL}_2(q_0)|}{|\mathrm{SL}_2(q)|} \right\|_{L^\infty}\Bigg|_{\equiv \omega_0(\mathrm{mod}\, q_0)} = \max_{\substack{\omega \in \mathrm{SL}(2,q) \\ \omega \equiv \omega_0(\mathrm{mod}\, q_0)}} \left| \pi_q[\mu](\omega) - \frac{|\mathrm{SL}_2(q_0)|}{|\mathrm{SL}_2(q)|} \right| < \eta.
$$

Then for any T with

(8.16)
$$
\eta^{-2} \log Q \;<\; T \;=\; o(|S|^{1/2}),
$$

there exist T distinct points $\gamma_1, \ldots, \gamma_T \in S = \operatorname{supp} \mu$ such that the probability measure $\nu = \nu_{T, \gamma_1, \ldots, \gamma_T}$ defined by

(8.17)
$$
\nu = \frac{1}{T}\left(\mathbf{1}_{\gamma_1} + \cdots + \mathbf{1}_{\gamma_T}\right)
$$

has the same property. That is, for all $q \in \mathfrak{Q}$, projection $\pi_q[\nu]$ is also nearly uniform:

(8.18)
$$
\max_{q \in \mathfrak{Q}} \left(\left\| \pi_q[\nu] - \frac{|\mathrm{SL}_2(q_0)|}{|\mathrm{SL}_2(q)|} \right\|_{L^\infty}\Bigg|_{\equiv \omega_0(\mathrm{mod}\, q_0)} \right) \ll \eta.
$$

The implied constant above is absolute.

Proof. This is a standard argument, so we give a sketch. Take ν as in (8.17). Let $\mathcal{D}$ be the expectation with respect to μ of the left-hand side of (8.18),

$$
\mathcal{D} := \sum_{\gamma \in \mathrm{SL}_2(\mathbb{Z})^T} \max_{q \in \mathfrak{Q}} \max_{\substack{\omega \in \mathrm{SL}_2(q) \\ \omega \equiv \omega_0(\mathrm{mod}\, q_0)}} \left| \frac{1}{T} \sum_{j=1}^{T} \mathbf{1}_{\{\gamma_j \equiv \omega(q)\}} - \frac{|\mathrm{SL}_2(q_0)|}{|\mathrm{SL}_2(q)|} \right| \mu^{(T)}(\gamma),
$$

where $\mu^{(T)}$ is the product measure on $\mathrm{SL}_2(\mathbb{Z})^T$ and $\gamma = (\gamma_1, \ldots, \gamma_T)$. Using (8.15), we have

$$
\mathcal{D} < \eta + \sum_{\gamma \in \mathrm{SL}_2(\mathbb{Z})^T} \sum_{\xi \in \mathrm{SL}_2(\mathbb{Z})^T} \max_{q \in \mathfrak{Q}} \max_{\omega \in \mathrm{SL}_2(q)} \left| \frac{1}{T} \sum_{j=1}^{T} f_\omega(\gamma_j, \xi_j) \right| \mu^{(T)}(\gamma) \mu^{(T)}(\xi),
$$

where

$$
f_\omega(\gamma_j, \xi_j) := \mathbf{1}_{\{\gamma_j \equiv \omega(q)\}} - \mathbf{1}_{\{\xi_j \equiv \omega(q)\}},
$$

and we extended the max over ω to all of $\mathrm{SL}_2(q)$.

Note that for fixed ω, $f_\omega(\gamma_j, \xi_j)$ are independent, mean zero random variables and bounded by 1. Hence the contraction principle gives

(8.19)
$$
\mathcal{D} < \eta + \sum_{\gamma} \sum_{\xi} \mathcal{D}(\gamma, \xi) \mu^{(T)}(\gamma) \mu^{(T)}(\xi),
$$

where

$$\mathcal{D}(\gamma,\xi) := \frac{1}{2^T} \sum_{\varepsilon\in\{\pm1\}^T} \max_{q\in\mathfrak{Q}} \max_{\omega\in\mathrm{SL}_2(q)} \left| \frac{1}{T}\sum_{j=1}^{T} \varepsilon_j f_\omega(\gamma_j,\xi_j) \right|.$$

Replace the max by an L^p norm with p to be chosen later:

$$\mathcal{D}(\gamma,\xi) \le \frac{1}{2^T} \sum_{\varepsilon\in\{\pm1\}^T} \left(\sum_{q\in\mathfrak{Q}}\sum_{\omega\in\mathrm{SL}_2(q)} \left| \frac{1}{T}\sum_{j=1}^{T} \varepsilon_j f_\omega(\gamma_j,\xi_j) \right|^p \right)^{1/p}$$

$$\ll \left(\sum_{q\in\mathfrak{Q}}\sum_{\omega\in\mathrm{SL}_2(q)} \frac{1}{2^T} \sum_{\varepsilon\in\{\pm1\}^T} \left| \frac{1}{T}\sum_{j=1}^{T} \varepsilon_j f_\omega(\gamma_j,\xi_j) \right|^p \right)^{1/p}$$

$$\ll \left(\sum_{q\in\mathfrak{Q}}\sum_{\omega\in\mathrm{SL}_2(q)} p^{p/2} \left(\sum_{j=1}^{T} \left| \frac{f_\omega(\gamma_j,\xi_j)}{T} \right|^2 \right)^{p/2} \right)^{1/p}$$

$$\ll Q^{4/p}p^{1/2}T^{-1/2},$$

where we applied Khintchine's inequality [Haa81]. (The implied constant is absolute.) Now we choose $p = \log Q$, so that

$$\mathcal{D}(\gamma,\xi) \ll (\log Q)^{1/2}T^{-1/2}.$$

Inserting this into (8.19) and setting $T > \eta^{-2}\log Q$ gives

$$\mathcal{D} \ll \eta.$$

The number of T-tuples γ with distinct γ_j is $|S|!/(|S|-T)!$, which is asymptotic to $|S|^T$ for $T = o(|S|^{1/2})$. This completes the proof. $\square$

Equipped with this randomness extraction argument, we proceed with the

8.3. *Proof of Proposition* 3.9. Recalling the parameters $\mathfrak{b}$ in (3.3), $\mathfrak{c}$ from (8.3), and $\mathfrak{B}$ from (8.2), we set $R := |\mathrm{SL}_2(\mathfrak{B})|$ and define α_0 in (3.5) by

$$(8.20) \qquad \alpha_0 := \frac{\beta\mathfrak{c}}{40R}.$$

For a parameter

$$(8.21) \qquad T = N^{c_1}$$

with small c_1 to be determined in (8.30), let $H = \mathcal{Q}^{12}$, $H_1 = \mathcal{Q}^6$, and set

$$(8.22) \qquad \mathcal{S}(T) := \left\{ \gamma\in\Gamma : |v_+(\gamma)-v| < \frac{1}{H}, |\lambda(\gamma)-T| < \frac{T}{H_1} \right\}.$$

By (8.12) with $q = 1$, crudely using $\delta < 1$ we have that

$$(8.23) \qquad \#\mathcal{S}(T) \gg T^{2\delta}/\mathcal{Q}^{18} + O(T^{2\delta}e^{-\mathfrak{c}\sqrt{\log T}}).$$

We have from (8.21), (8.20), and (3.5) that

$$e^{-\mathfrak{c}\sqrt{\log T}} = e^{-\mathfrak{c}\sqrt{c_1}\sqrt{\log N}} = \mathcal{Q}^{-\sqrt{c_1}40R/\beta}.$$

So as long as

(8.24) $$c_1 > \left(\frac{3\mathfrak{b}}{4R}\right)^2,$$

the estimate (8.23) is significant, with an error of size $\ll T^{2\delta}/\mathcal{Q}^{30}$.

By the pigeonhole principle, there is some element $\mathfrak{s}_T \in \mathcal{S}(T)$ so that

$$\mathcal{S}'(T) := \{s \in \mathcal{S}(T) : s \equiv \mathfrak{s}_T(\mathrm{mod}\,\mathfrak{B})\}$$

satisfies

(8.25) $$\#\mathcal{S}'(T) \geq \frac{1}{|\,\mathrm{SL}_2(\mathfrak{B})|}\#\mathcal{S}(T) \gg T^{2\delta}/\mathcal{Q}^{18}.$$

(Recall our implied constants may depend implicitly on $\mathcal{A}$, and $\mathfrak{B}$ depends only on $\mathcal{A}$.)

For this set, the counting statement (8.23) remains significant even with a modular restriction: for any $q < \mathcal{Q}$ with $\mathfrak{B} \mid q$ and any $\omega \in \mathrm{SL}_2(q)$ with $\omega \equiv \mathfrak{s}_T(\mathrm{mod}\,\mathfrak{B})$, applying (8.13) gives

(8.26)

$$\#\{s \in \mathcal{S}'(T) : s \equiv \omega(\mathrm{mod}\,q)\} = \#\{s \in \mathcal{S}(T) : s \equiv \omega(\mathrm{mod}\,q)\}$$

$$= \frac{|\,\mathrm{SL}_2(\mathfrak{B})|}{|\,\mathrm{SL}_2(q)|}\#\{s \in \mathcal{S}(T) : s \equiv \omega \equiv \mathfrak{s}_T(\mathrm{mod}\,\mathfrak{B})\}\,(1 + O(\mathcal{Q}^{-6})) + O(T^{2\delta}\mathcal{Q}^{-30})$$

$$= \frac{|\,\mathrm{SL}_2(\mathfrak{B})|}{|\,\mathrm{SL}_2(q)|}\#\mathcal{S}'(T)\,(1 + O(\mathcal{Q}^{-6})) + O(T^{2\delta}\mathcal{Q}^{-30}).$$

From (8.25), the main term is $\gg T^{2\delta}/\mathcal{Q}^{21}$, dominating the error.

We just need to play with $\mathcal{S}'(T)$ to get good distribution modulo $\mathfrak{B}$. Recall $R = |\,\mathrm{SL}_2(\mathfrak{B})|$. Then every element of the "coset"

$$\gamma \in \mathcal{S}'(T) \cdot \mathfrak{s}_T^{R-1}$$

satisfies $\gamma \equiv I(\mathrm{mod}\,\mathfrak{B})$. Next write $\mathrm{SL}_2(\mathfrak{B}) = \{\gamma_1, \gamma_2, \ldots, \gamma_R\}$, and take $x_1, \ldots, x_R \in \Gamma$ so that

(8.27) $$x_r \equiv \gamma_r(\mathrm{mod}\,\mathfrak{B}), \qquad r = 1, \ldots, R.$$

(Recall we had assumed in Section 3.1 that $\Gamma(\mathrm{mod}\,q)$ is all of $\mathrm{SL}_2(q)$ for all q, so such x_r exist.) Such x_r can be found of size $\asymp_{\mathcal{A}} 1$.

Note that any element

$$\gamma \in \mathcal{S}'(T) \cdot \mathfrak{s}_T^{R-1} \cdot x_r$$

has $\gamma \equiv \gamma_r (\mathrm{mod}\,\mathfrak{B})$. Unfortunately, this triple-product does not work since we do not have control on the expanding vector of x_r. To remedy this, we take a single fixed element $\mathfrak{f}_0 \in \Gamma$ of size

$$(8.28) \qquad\qquad\qquad\qquad \lambda(\mathfrak{f}_0) \asymp B^{1/100},$$

say, with

$$(8.29) \qquad\qquad\qquad\qquad |v_+(\mathfrak{f}_0) - \mathfrak{v}| < \mathcal{Q}^{-6}.$$

Then from (2.7), $v_+(\mathfrak{f}_0 x_r) = \mathfrak{v}(1 + O(\mathcal{Q}^{-6}))$, and for any $s \in \mathcal{S}'(T)$,

$$v_+(s \cdot \mathfrak{s}_T^{R-1} \cdot \mathfrak{f}_0 x_r) = \mathfrak{v}(1 + O(\mathcal{Q}^{-6})).$$

Moreover, from (8.22) and (2.6), we have

$$\lambda(s \cdot \mathfrak{s}_T^{R-1} \cdot \mathfrak{f}_0 x_r) = \lambda(s)\lambda(\mathfrak{s}_T)^{R-1}\lambda(\mathfrak{f}_0 x_r)\left(1 + O(\mathcal{Q}^{-6})\right)$$
$$= T^R \lambda(\mathfrak{f}_0 x_r)\left(1 + O(\mathcal{Q}^{-6})\right).$$

Now for each $u \in \mathcal{U}$, $u \asymp B$, and each $r = 1, \ldots, R$, take $T = T_{u,r}$ so that

$$T^R \lambda(\mathfrak{f}_0 x_r) = u;$$

that is, let

$$(8.30) \qquad\qquad T_{u,r} := \left(\frac{u}{\lambda(\mathfrak{f}_0 x_r)}\right)^{1/R} \asymp B^{99/(100R)} = N^{99\mathfrak{b}/(100R)}$$

which, by (3.4), determines c_1 in (8.21). Note that (8.24) is easily satisfied.

Thus for each u and r, we have sets

$$\mathcal{B}_{u,r} := \mathcal{S}'(T_{u,r}) \cdot (\mathfrak{s}_{T_{u,r}})^{R-1} \cdot \mathfrak{f}_0 \cdot x_r \subset \Gamma,$$

so that for all $\mathfrak{a} \in \mathcal{B}_{u,r}$, the expanding vector is controlled,

$$|v_+(\mathfrak{a}) - v| \ll \mathcal{Q}^{-6},$$

and the eigenvalue is controlled,

$$\lambda(\mathfrak{a}) = u(1 + O(\mathcal{Q}^{-6})).$$

Since we have saved an extra $\mathcal{Q}$, we can use it to set the implied constant to 1, getting (3.11) and (3.12).

Note that by (8.26), for all $q < \mathcal{Q}$ with $\mathfrak{B} \mid q$, and all $\omega \in \mathrm{SL}_2(q)$ with $\omega \equiv \mathfrak{f}_0 x_r (\mathrm{mod}\,\mathfrak{B})$, we have, crudely, that

$$(8.31) \qquad \#\{\mathfrak{a} \in \mathcal{B}_{u,r} : \mathfrak{a} \equiv \omega(\mathrm{mod}\,q)\} = \frac{|\mathrm{SL}_2(\mathfrak{B})|}{|\mathrm{SL}_2(q)|} \#\mathcal{B}_{u,r}\left(1 + O(\mathcal{Q}^{-5})\right).$$

Recall also from (8.23) that the cardinality of $\mathcal{B}_{u,r}$ is

$$(8.32) \qquad\qquad\qquad\qquad \gg (T_{u,r})^{2\delta}/\mathcal{Q}^{18} \gg N^c,$$

using (8.30).

Hence for fixed u, we may apply the randomness extraction argument in Lemma 8.14 to $\mathcal{B}_{u,r}$, with $\eta = \mathcal{Q}^{-5}$ and $q_0 = \mathfrak{B}$. This gives sets $\mathcal{B}'_{u,r} \subset \mathcal{B}_{u,r}$ of size $\gg N^c$, for which (8.31) continues to hold and, moreover, we can force them all to have exactly the same cardinality independently of r,

$$|\mathcal{B}'_{u,r}| = |\mathcal{B}'_{u,r'}|.$$

Set

$$\tilde{\aleph}_u := \bigsqcup_{r=1}^{R} \mathcal{B}'_{u,r},$$

and note that for $\mathfrak{B} \mid q < \mathcal{Q}$ and $\omega \in \mathrm{SL}_2(q)$,

$$\#\{\mathfrak{a} \in \tilde{\aleph}_u : \mathfrak{a} \equiv \omega (\mathrm{mod}\, q)\} = \frac{|\mathrm{SL}_2(\mathfrak{B})|}{|\mathrm{SL}_2(q)|} \#\mathcal{B}'_{u,r} \left(1 + O(\mathcal{Q}^{-5})\right),$$

where r is the index for which $\omega \equiv \mathfrak{f}_0 x_r (\mathrm{mod}\, \mathfrak{B})$. Since $\#\mathcal{B}'_{u,r} = |\tilde{\aleph}_u|/R$, and $R = |\mathrm{SL}_2(\mathfrak{B})|$, we have that for each u, $\tilde{\aleph}_u$ satisfies

$$(8.33) \qquad \#\{\mathfrak{a} \in \tilde{\aleph}_u : \mathfrak{a} \equiv \omega (\mathrm{mod}\, q)\} = \frac{|\tilde{\aleph}_u|}{|\mathrm{SL}_2(q)|} \left(1 + O(\mathcal{Q}^{-5})\right).$$

We can now also drop the condition $\mathfrak{B} \mid q$ in (8.33) by summing along certain arithmetic progressions; since $\mathfrak{B} \ll_A 1$, the implied constant still depends only on A (cf. Remark 8.7).

Now we apply Lemma 8.14 again to $\tilde{\aleph}_u$, with $\eta = \mathcal{Q}^{-5}$ and $q_0 = 1$, giving sets

$$\aleph_u \subset \tilde{\aleph}_u$$

for which (8.33) still holds, that is (3.14) holds, and which all have the same cardinality, giving (3.10).

This completes the proof of Proposition 3.9.

9. Proof of Theorem 1.22

Recall that $\mathfrak{R}_\mathcal{A}(N)$ is the set of rationals b/d with partial quotients bounded in the alphabet $\mathcal{A}$ with $d < N$, $\mathfrak{D}_\mathcal{A}(N)$ is the set of continuants up to N, and $\mathfrak{C}_\mathcal{A}$ is the limit set of $\mathfrak{R}_\mathcal{A}$ with Hausdorff dimension $\delta = \delta_\mathcal{A}$. Recall the sum-set structure (1.47), that if $a \in \mathcal{A}$ and $b/d \in \mathfrak{R}_\mathcal{A}$, then d and $b + ad$ are in $\mathfrak{D}_\mathcal{A}$. The same holds for another $a' \in \mathcal{A}$; that is, all three of d, $b + ad$, and $b + a'd$ are in $\mathfrak{D}_\mathcal{A}$.

We wish to show (1.24) that for any $\varepsilon > 0$,

$$\#(\mathfrak{D}_\mathcal{A} \cap [1, N]) \gg_\varepsilon N^{\delta + (2\delta - 1)(1-\delta)/(5-\delta) - \varepsilon}.$$

For ease of notation, we lose no generality by specializing from now on to the case $1, 2 \in \mathcal{A}$; whence $b/d \in \mathfrak{R}_\mathcal{A}(N)$ implies that

$$(9.1) \qquad\qquad d, b + d, b + 2d \in \mathfrak{D}_\mathcal{A}(3N).$$

And again, we can drop the subscript $\mathcal{A}$ from $\mathfrak{D}$, $\mathfrak{R}$, and $\mathfrak{C}$.

Our new ensemble of focus is the collection $\Omega = \Omega_N$ of intervals given by

$$(9.2) \qquad \Omega := \bigcup_{\substack{d \in \mathfrak{D} \\ N/2 \leq d < 3N}} \left[\frac{d}{N}, \frac{d+1}{N} \right] \subset [1/2, 3],$$

so that

$$(9.3) \qquad |\Omega| \ll \frac{\#(\mathfrak{D} \cap [1, 3N])}{N}.$$

Hensley's conjecture implies, assuming $\delta > 1/2$, that we should have

$$(9.4) \qquad |\Omega| \asymp 1.$$

A priori, we do not even know that $|\Omega| > 1$, and the bound (1.16) follows from

$$(9.5) \qquad |\Omega| \gg N^{-(1-\delta)},$$

so this is what we must improve upon.

Note from (9.1) that

$$(9.6) \qquad \mathbf{1}_\Omega(x)\mathbf{1}_\Omega(y)\mathbf{1}_\Omega(x+y)\mathbf{1}_\mathfrak{R}\left(\frac{y}{x} - 1 \right) = 1$$

if $x = d/N$ and $y = (b+d)/N$ with $b/d \in \mathfrak{R}$. Just as we thickened $\mathbf{1}_\mathfrak{D}$ to some intervals $\mathbf{1}_\Omega$ in (9.2), we wish to thicken $\mathbf{1}_\mathfrak{R}$ to some intervals.

By Frostman's theorem [Fro35], there is a probability measure μ supported on the Cantor set $\mathfrak{C}$, so that for any interval $\mathcal{I} \subset [0,1]$, we have

$$(9.7) \qquad \mu(\mathcal{I}) \ll_\varepsilon |\mathcal{I}|^{\delta - \varepsilon}.$$

The most naive thickening of $\mathfrak{R}$ we could take is at the scale of $1/N^2$; namely, for each $x \in \mathfrak{C}$, there is (by Dirichlet's approximation theorem and properties of continued fractions) some $b/d \in \mathfrak{R}(N)$ with

$$\left| x - \frac{b}{d} \right| < \frac{1}{N^2}.$$

So we can find a collection of $\asymp N^{2\delta}$ intervals of length $\asymp 1/N^2$ which cover $\mathfrak{C}$, each of which has a point in $\mathfrak{R}(N)$. Note also that the spacing between consecutive points in $\mathfrak{R}(N)$ is $\geq 1/N^2$. Instead it will be more fruitful to collect points at square-root this scale, $1/N$, as follows.

By (9.7), there is a collection $\{\mathcal{I}_\ell\}_{\ell \leq L}$ of $L \asymp N^\delta$ disjoint intervals $\mathcal{I}_\ell \subset [1/2, 1]$, each of length $1/N$, so that

$$(9.8) \qquad \mu(\mathcal{I}_\ell) \gg_\varepsilon N^{-\delta - \varepsilon}.$$

Denote their union by

$$(9.9) \qquad \tilde{\mathcal{R}} := \bigsqcup_{\ell=1}^{L} \mathcal{I}_\ell.$$

Subdividing each $\mathcal{I}_\ell$ into intervals $\mathcal{I}_{\ell,n}$ of length $1/N^2$, it follows that there are at least $\gg N^{\delta-\varepsilon}$ of them intersecting $\mathfrak{C}$. For each such ℓ, n, the intersection $\mathcal{I}_{\ell,n} \cap \mathfrak{R}(N)$ is also nonempty, possibly after replacing $\mathcal{I}_{\ell,n}$ with a doubling. Hence the cardinality of

$$\tilde{\mathcal{R}}(N) := \left\{ \frac{b}{d} \in \mathfrak{R} \cap \tilde{\mathcal{R}} : (b,d) = 1, N/2 < b < d < N \right\}$$

is of the right order

(9.10) $$\#\tilde{\mathcal{R}}(N) \gg_\varepsilon N^{2\delta-\varepsilon}.$$

We thicken these intervals just a little further, setting

(9.11) $$\mathcal{R} := \bigcup_{\ell=1}^{L} \left(\mathcal{I}_\ell + \left[-\frac{2}{N}, \frac{2}{N} \right] \right).$$

Note that since $L \ll N^\delta$, we have

(9.12) $$|\mathcal{R}| \ll N^{-(1-\delta)}.$$

Note further that $\mathfrak{C}$, and hence $\mathcal{R}$, is contained strictly inside $[0,1]$; that is, for some $\nu = \nu_A > 0$, we have

(9.13) $$\mathcal{R} \subset [\nu, 1 - \nu].$$

Consider now our main integral $\mathcal{J}$, motivated by (9.6), defined by

(9.14) $$\mathcal{J} := \iint_{\mathbb{R}^2} \mathbf{1}_\Omega(x)\mathbf{1}_\Omega(y)\mathbf{1}_\Omega(x+y)\mathbf{1}_\mathcal{R}\left(\frac{y}{x} - 1 \right) dy\,dx.$$

By (9.2), the domain of integration above is supported in the box $[1/2, 3] \times [1/2, 3]$.

For each $b/d \in \tilde{\mathcal{R}}(N)$, the intervals

$$\frac{d}{N} \le x \le \frac{d+1/2}{N}, \qquad \frac{b+d}{N} \le y \le \frac{b+d+1/2}{N}$$

belong to Ω, by (9.1), as does the interval

$$\frac{b+2d}{N} \le x + y \le \frac{b+2d+1}{N}.$$

Moreover, the interval

$$\frac{b}{d} - \frac{2}{N} < \frac{b-1}{d} \le \frac{y}{x} - 1 \le \frac{b+1}{d} < \frac{b}{d} + \frac{2}{N}$$

is in $\mathcal{R}$ by the thickening in (9.11). That is, these intervals contribute $1/N^2$ to $\mathcal{J}$ for each b/d, and hence by (9.10), we have established the following lower bound.

PROPOSITION 9.15.

(9.16) $$\mathcal{J} \gg_\varepsilon N^{-2(1-\delta)-\varepsilon}.$$

The trivial bound $\mathcal{J} \ll |\Omega|^2$ from (9.14) recovers (9.5), which is what we must improve. The rest of the appendix is devoted to establishing the following

PROPOSITION 9.17.

(9.18) $$\mathcal{J} \ll_\varepsilon N^{-\frac{1-\delta}{2}+\varepsilon}|\Omega|^{2-\frac{3(1-\delta)}{2(2-\delta)}}.$$

Then Theorem 1.23 follows immediately from (9.16), (9.18), and (9.3).

9.1. *Proof of Proposition* 9.17. Let $M > 0$ be a parameter to be chosen later (it will be a little less than $N^{-(1-\delta)}$), and decompose

(9.19) $$\mathcal{J} = \mathcal{J}_1 + \mathcal{J}_2,$$

where

(9.20) $$\mathcal{J}_1 := \iint_{(\mathbf{1}_\Omega * \mathbf{1}_{-\Omega})(-x) < M} \cdots dx dy,$$

and

(9.21) $$\mathcal{J}_2 := \iint_{(\mathbf{1}_\Omega * \mathbf{1}_{-\Omega})(-x) \geq M} \cdots dx dy.$$

Then writing $\mathbf{1}_\mathcal{R} \leq 1$, we have

(9.22) $$\mathcal{J}_1 \leq \iint_{(\mathbf{1}_\Omega * \mathbf{1}_{-\Omega})(-x) < M} \mathbf{1}_\Omega(x)\mathbf{1}_\Omega(y)\mathbf{1}_{-\Omega}(-x - y) dx dy$$

$$\leq \int_{(\mathbf{1}_\Omega * \mathbf{1}_{-\Omega})(-x) < M} \mathbf{1}_\Omega(x)\left(\mathbf{1}_\Omega * \mathbf{1}_{-\Omega}(-x)\right) dx$$

$$< M|\Omega|.$$

It is clear already that to get a gain on $|\Omega|$, we must take M a power less than $N^{-(1-\delta)}$.

We are left to analyze $\mathcal{J}_2$. Note that in the domain of $\mathcal{J}_2$, we have

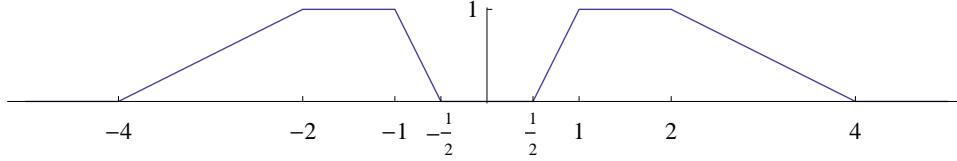$$\mathbf{1}_\Omega(x + y) \leq 1 \leq \frac{1}{M}(\mathbf{1}_\Omega * \mathbf{1}_{-\Omega})(-x).$$

Hence we can write

$$\mathcal{J}_2 \leq \frac{1}{M} \iint_{\mathbb{R}^2} \eta(x) \, (\mathbf{1}_\Omega * \mathbf{1}_{-\Omega})(-x) \, \mathbf{1}_\Omega(y) \, \mathbf{1}_\mathcal{R}\left(\frac{y}{x} - 1\right) dy dx,$$

where we have bounded $\mathbf{1}_\Omega(x)$ by a smooth bump function $\eta(x)$ with support in $[1/4, 4]$, say, and $\eta \geq 1$ on $[1/2, 3]$ to recall (9.2) that $x \in [1/2, 3]$.

For a smooth, nonnegative, even function Υ with compact support and $\int \Upsilon = 1$, let $\Upsilon_N(y) := 10N\Upsilon(10Ny)$, and dominate $\mathbf{1}_\Omega(y)$, up to constant, by the smooth function

$$\mathcal{S}_\Omega := \mathbf{1}_\Omega * \Upsilon_N.$$

Figure 1. The Fourier multiplier λ_0.

So we have

(9.23) $$\mathcal{J}_2 \ll \frac{1}{M} \cdot \mathcal{J}_2',$$

where

(9.24) $$\mathcal{J}_2' := \iint_{\mathbb{R}^2} \eta(x)(\mathbf{1}_\Omega * \mathbf{1}_{-\Omega})(-x) \, \mathcal{S}_\Omega(y) \, \mathbf{1}_\mathcal{R}\left(\frac{y}{x} - 1\right) dy dx.$$

Note that, by the smoothness of Υ, the Fourier spectrum of $\mathcal{S}_\Omega$ is contained, up to negligible error, in $[-N^{1+\varepsilon}, N^{1+\varepsilon}]$. So we can decompose $\mathcal{S}_\Omega(y)$ by the technique of "slicing." That is, introduce a certain dyadic partition of unity via the Fourier multipliers $\lambda_k(\xi)$, defined as follows. Let $\lambda_0(\xi)$ be even, $\equiv 1$ on $[1,2]$, and decaying piecewise-linearly to 0 at $\xi = 1/2$ and $\xi = 4$; see Figure 1. For integers k ranging in

(9.25) $$0 < 2^{2k} < N^{1+\varepsilon},$$

define

$$\lambda_k(\xi) := \lambda_0(\xi \cdot 2^{-2k}).$$

Let Λ_k be the Fourier inverse of λ_k, so

$$\Lambda_k(y) = 2^{2k}\Lambda_0(2^{2k}y) = (\mathcal{D}_{2^{2k}}\Lambda_0)(y),$$

where

$$\Lambda_0(y) := \frac{\sin^2\left(\frac{3}{2}\pi y\right)}{\pi^2 y^2}\left(2\cos\left(2\pi y\right) - \cos\left(\pi y\right) + \cos\left(5\pi y\right)\right),$$

and $\mathcal{D}_u$ is the dilation representation:

$$(\mathcal{D}_u f)(y) := u f(uy).$$

We also introduce $\mathcal{T}_u$, the translation representation,

$$\mathcal{T}_u f(y) := f(y + u).$$

Of course $\widehat{\Lambda_k * f} = \lambda_k \cdot \widehat{f}$, so we have

$$\mathcal{S}_\Omega = \sum_{2^{2k} < N^{1+\varepsilon}} (\Lambda_k * \mathcal{S}_\Omega) + \mathrm{Err},$$

where Err is bounded by an arbitrarily large power of $1/N$, and will henceforth be ignored.

Then we can bound (9.24) as

$$(9.26) \qquad |\mathcal{J}_2'| \ll \sum_k \left|\mathcal{J}_2^{(k)}\right|,$$

where

$$(9.27) \qquad \mathcal{J}_2^{(k)} := \iint \eta(x)\,(\mathbf{1}_\Omega * \mathbf{1}_{-\Omega})(-x)\,(\Lambda_k * \mathcal{S}_\Omega)(y)\,\mathbf{1}_\mathcal{R}\left(\frac{y}{x}-1\right)dy dx.$$

Since the Fourier spectrum of $\mathcal{S}_\Omega$ is now controlled, so is that of $\mathbf{1}_\mathcal{R}$, as follows. Write $\mathbf{1}_\mathcal{R}\left(\frac{y}{x}-1\right) = \mathbf{1}_{x\mathcal{R}}\left(y-x\right) = (\mathcal{T}_{-x}\mathbf{1}_{x\mathcal{R}})(y)$. Then the y integral can be written as

$$(9.28) \qquad \int_\mathbb{R}(\Lambda_k * \mathcal{S}_\Omega)(y)\,(\mathcal{T}_{-x}\mathbf{1}_{x\mathcal{R}})(y)dy$$

$$= \int_\mathbb{R}\lambda_k(\xi)\widehat{\mathcal{S}}_\Omega(\xi)\,\overline{(\widehat{\mathcal{T}_{-x}\mathbf{1}_{x\mathcal{R}}})(\xi)}d\xi$$

$$= \int_\mathbb{R}\lambda_k(\xi)\widehat{\mathcal{S}}_\Omega(\xi)\,\lambda_k'(x\xi)\overline{(\widehat{\mathcal{T}_{-x}\mathbf{1}_{x\mathcal{R}}})(\xi)}d\xi,$$

where we inserted another bump function λ_k' which is smooth in addition to other properties of λ_k. Namely, let λ_0' be even, $\equiv 1$ on $\pm[1/16, 16]$, and decay smoothly to 0 outside of $\pm[1/32, 32]$; then set $\lambda_k'(\xi) := \lambda_0'(\xi 2^{-2k})$. The point is that $\lambda_k'(x\xi) \equiv 1$ on the support of λ_k, since $x \in [1/4, 4]$ by the support of η, so the above equality holds.

Then writing Λ_k' for the inverse transform of λ_k', we have

$$(9.29) \quad \mathcal{J}_2^{(k)} = \iint \eta(x)\,(\mathbf{1}_\Omega * \mathbf{1}_{-\Omega})(-x)\,(\Lambda_k * \mathcal{S}_\Omega)(y)\left[\Lambda_k' * \mathbf{1}_\mathcal{R}\right]\left(\frac{y}{x}-1\right)\,dy\,dx.$$

Now we handle two ranges of k separately. We introduce a cutoff parameter $\mathcal{K}$ to be chosen later, see (9.41).

9.1.1. *The range $k \le \mathcal{K}$.* We wish to prove

LEMMA 9.30. *For $k \le \mathcal{K}$,*

$$(9.31) \qquad \left|\mathcal{J}_2^{(k)}\right| \ll_\varepsilon |\Omega|^3\,2^{2k(1-\delta)}\,N^{-(1-\delta)+\varepsilon}.$$

This is a gain of a power of $|\Omega|$. (Recall we are assuming $|\Omega| < 1$.)

Proof. Estimate (9.29) by

$$(9.32) \qquad \left|\mathcal{J}_2^{(k)}\right| \ll \|\mathbf{1}_\Omega * \mathbf{1}_{-\Omega}\|_1 \cdot \|\Lambda_k * \mathcal{S}_\Omega\|_1 \cdot \left\|\Lambda_k' * \mathbf{1}_\mathcal{R}\right\|_\infty.$$

The first factor contributes $|\Omega|^2$, and the second is $\ll |\Omega|$, since Λ_0 is integrable. For the last term, write

$$(9.33) \qquad \left\| \Lambda_k' * \mathbf{1}_{\mathcal{R}} \right\|_\infty \ll \sup_z \int_{\mathbb{R}} 2^{2k} |\Lambda_0'(u2^{2k})| \mathbf{1}_{\mathcal{R}}(z-u) du$$

$$\ll \sum_{m \geq 0} 2^{2(k-m)} \sup_{|\mathcal{U}|=2^{m+1-2k}} |\mathcal{R} \cap \mathcal{U}|,$$

where the supremum is taken over intervals $\mathcal{U}$. Here we used that Λ_0' has rapid decay, so certainly

$$\Lambda_0'(y) \ll \mathbf{1}_{|y|<1} + \frac{1}{2^2} \mathbf{1}_{|y|<2} + \frac{1}{4^2} \mathbf{1}_{|y|<4} + \frac{1}{8^2} \mathbf{1}_{|y|<8} + \cdots.$$

Note by (9.25) that

$$|\mathcal{U}| \geq 2^{-2k+1} \geq \frac{2}{N^{1+\varepsilon}}.$$

We have yet to exploit the structure of $\mathcal{R}$ and do so now. This requires the following

LEMMA 9.34. *For any interval $\mathcal{U}$ of length at least $1/N^{1+\varepsilon}$, we have*

$$(9.35) \qquad |\mathcal{R} \cap \mathcal{U}| \ll N^{-(1-\delta)+\varepsilon} |\mathcal{U}|^{\delta+\varepsilon}.$$

Postponing the proof of this lemma, we see that applying (9.35) in (9.33) gives

$$(9.36) \qquad \left\| \Lambda_k' * \mathbf{1}_{\mathcal{R}} \right\|_\infty \ll_\varepsilon 2^{2k(1-\delta)} N^{-(1-\delta)+\varepsilon}.$$

Putting (9.36) into (9.32) gives (9.31), as claimed. $\square$

It remains to establish (9.35).

Proof of Lemma 9.34. From the structure of $\mathcal{R}$ in (9.11), we have that

$$|\mathcal{R} \cap \mathcal{U}| \leq \sum_\ell \left| \mathcal{U} \cap \left(\mathcal{I}_\ell + [-2/N, 2/N] \right) \right|$$

$$\ll \frac{1}{N} \# \left\{ \ell \leq L : \mathcal{U} \cap \left(\mathcal{I}_\ell + [-2/N, 2/N] \right) \neq \emptyset \right\}$$

$$\ll_\varepsilon \frac{1}{N} \frac{\mu(\mathcal{U} + [0, 1/N])}{N^{-\delta-\varepsilon}}$$

$$\ll_\varepsilon N^{-(1-\delta)+\varepsilon} |\mathcal{U}|^{\delta+\varepsilon},$$

where we used (9.8) and (9.7) in the penultimate and final lines, respectively.

$\square$

9.1.2. *The range $k > \mathcal{K}$.* In this range, we will establish

LEMMA 9.37. *For $k > \mathcal{K}$ and any $\varepsilon > 0$,*

(9.38)
$$\left| \mathcal{J}_2^{(k)} \right| \ll_\varepsilon |\Omega|^{3/2} \; 2^{-k\delta} \; N^{-(1-\delta)+\varepsilon}.$$

Proof. Changing variables $y \mapsto yx$ in (9.29), we have

$$\left| \mathcal{J}_2^{(k)} \right| \ll \left| \iint \mathfrak{f}(x)(\Lambda_k * \mathcal{S}_\Omega)(xy) \; \eta(y) \left(\Lambda'_k * \mathbf{1}_\mathcal{R} \right)(y-1) \; dy \; dx \right|,$$

where we set

(9.39)
$$\mathfrak{f}(x) := x \; \eta(x)(\mathbf{1}_\Omega * \mathbf{1}_{-\Omega})(-x).$$

By the rapid decay of Λ'_k and (9.13), we may restrict the integral to $y \asymp 1$ with a negligible error. Now reverse orders, apply Parseval in x, reverse orders again, use the definition of the Fourier multipliers λ_k, apply Cauchy-Schwarz in y, change variables $y \mapsto \xi/y$, and estimate

$$\left| \mathcal{J}_2^{(k)} \right| \ll \left| \int_{y \asymp 1} \left(\int_{|\xi|/y \asymp 2^{2k}} \widehat{\mathfrak{f}}(\xi) \; \overline{\widehat{\mathcal{S}_\Omega}(\xi/y)} \frac{1}{y} d\xi \right) \left(\Lambda'_k * \mathbf{1}_\mathcal{R} \right)(y-1) \, dy \right|$$

$$\ll \int_{|\xi| \asymp 2^{2k}} |\widehat{\mathfrak{f}}(\xi)| \; \left(\frac{1}{|\xi|} \int |\widehat{\mathcal{S}_\Omega}(y)|^2 dy \right)^{1/2} \|\Lambda'_k * \mathbf{1}_\mathcal{R}\|_2 \; d\xi$$

$$\ll_\varepsilon \left(2^{-k} \int_{\xi \in \mathbb{R}} |\widehat{\mathfrak{f}}(\xi)| d\xi \right) |\Omega|^{1/2} \; 2^{k(1-\delta)} N^{-(1-\delta)+\varepsilon},$$

where we estimated the last piece by

$$\begin{aligned} \|\Lambda'_k * \mathbf{1}_\mathcal{R}\|_2 &\leq \|\Lambda'_k * \mathbf{1}_\mathcal{R}\|_\infty^{1/2} \|\Lambda'_k * \mathbf{1}_\mathcal{R}\|_1^{1/2} \\ &\ll_\varepsilon 2^{k(1-\delta)} N^{-(1-\delta)/2+\varepsilon} |\mathcal{R}|^{1/2} \\ &\ll_\varepsilon 2^{k(1-\delta)} N^{-(1-\delta)+\varepsilon}, \end{aligned}$$

using the $\mathcal{L}^\infty$ bound in (9.36) and (9.12). We easily estimate from (9.39) that $\|\widehat{\mathfrak{f}}\|_1 \ll |\Omega|$, giving (9.38), as claimed. $\square$

9.1.3. *Completion of Proof.* It is now a simple matter to establish Proposition 9.17. Putting (9.31), (9.38), and (9.26) into (9.23) gives

(9.40)
$$\begin{aligned} \mathcal{J}_2 &\ll_\varepsilon \frac{1}{M} N^{-(1-\delta)+\varepsilon} \left(|\Omega|^3 \; 2^{2\mathcal{K}(1-\delta)} \; + |\Omega|^{3/2} \; 2^{-\mathcal{K}\delta} \; \right) \\ &\ll_\varepsilon \frac{1}{M} N^{-(1-\delta)+\varepsilon} \; |\Omega|^{3/(2-\delta)} \; , \end{aligned}$$

on setting

(9.41)
$$\mathcal{K} := \frac{-3 \log_2 |\Omega|}{2(2-\delta)}.$$

Combining (9.40) with (9.22) and choosing

$$M = N^{-(1-\delta)/2+\varepsilon}|\Omega|^{(1+\delta)/(4-2\delta)}$$

gives (9.18), as claimed. This competes the proof of Theorem 1.23.

References

[Bou03]   J. BOURGAIN, On the Erdős-Volkmann and Katz-Tao ring conjectures, *Geom. Funct. Anal.* **13** (2003), 334–365. MR 1982147. Zbl 1115.11049. http://dx.doi.org/10.1007/s000390300008.

[Bou05]   J. BOURGAIN, *Green's Function Estimates for Lattice Schrödinger Operators and Applications*, *Ann. of Math. Stud.* **158**, Princeton Univ. Press, Princeton, NJ, 2005. MR 2100420. Zbl 1137.35001.

[Bou10]   J. BOURGAIN, The discretized sum-product and projection theorems, *J. Anal. Math.* **112** (2010), 193–236. MR 2763000. Zbl 1234.11012. http://dx.doi.org/10.1007/s11854-010-0028-x.

[Bou12]   J. BOURGAIN, Partial quotients and representation of rational numbers, *C. R. Math. Acad. Sci. Paris* **350** (2012), 727–730. MR 2981341. Zbl 06100614. http://dx.doi.org/10.1016/j.crma.2012.09.002.

[BG08]   J. BOURGAIN and A. GAMBURD, Uniform expansion bounds for Cayley graphs of $SL_2(\mathbb{F}_p)$, *Ann. of Math.* **167** (2008), 625–642. MR 2415383. Zbl 1216.20042. http://dx.doi.org/10.4007/annals.2008.167.625.

[BGS10]   J. BOURGAIN, A. GAMBURD, and P. SARNAK, Affine linear sieve, expanders, and sum-product, *Invent. Math.* **179** (2010), 559–644. MR 2587341. Zbl 1239.11103. http://dx.doi.org/10.1007/s00222-009-0225-3.

[BGS11]   J. BOURGAIN, A. GAMBURD, and P. SARNAK, Generalization of Selberg's $\frac{3}{16}$ theorem and affine sieve, *Acta Math.* **207** (2011), 255–290. MR 2892611. Zbl 1276.11081. http://dx.doi.org/10.1007/s11511-012-0070-x.

[BK10]   J. BOURGAIN and A. KONTOROVICH, On representations of integers in thin subgroups of $SL_2(\mathbb{Z})$, *Geom. Funct. Anal.* **20** (2010), 1144–1174. MR 2746949. Zbl 1230.11050. http://dx.doi.org/10.1007/s00039-010-0093-4.

[BK11]   J. BOURGAIN and A. KONTOROVICH, On Zaremba's conjecture, *C. R. Math. Acad. Sci. Paris* **349** (2011), 493–495. MR 2802911. Zbl 1215.11005. http://dx.doi.org/10.1016/j.crma.2011.03.023.

[BKS10]   J. BOURGAIN, A. KONTOROVICH, and P. SARNAK, Sector estimates for hyperbolic isometries, *Geom. Funct. Anal.* **20** (2010), 1175–1200. MR 2746950. Zbl 1208.22007. http://dx.doi.org/10.1007/s00039-010-0092-5.

[BV12]   J. BOURGAIN and P. P. VARJÚ, Expansion in $SL_d(\mathbf{Z}/q\mathbf{Z})$, q arbitrary, *Invent. Math.* **188** (2012), 151–173. MR 2897695. Zbl 1247.20052. http://dx.doi.org/10.1007/s00222-011-0345-4.

[Bum85]   R. T. BUMBY, Hausdorff dimension of sets arising in number theory, in *Number Theory* (New York, 1983–84), *Lecture Notes in Math.* **1135**,

Springer-Verlag, New York, 1985, pp. 1–8. MR 0803348. Zbl 0575.28004. http://dx.doi.org/10.1007/BFb0074599.

[Dol98]   D. Dolgopyat, On decay of correlations in Anosov flows, *Ann. of Math.* **147** (1998), 357–390. MR 1626749. Zbl 0911.58029. http://dx.doi.org/10.2307/121012.

[FK13]   D. A. Frolenkov and I. D. Kan, A reinforcement of the Bourgain-Kontorovich's theorem by elementary methods II, 2013. arXiv 1303.3968.

[Fro35]   O. Frostman, Potentiel d'equilibre et capacité des ensembles avec quelques applications à la théorie des fonctions, *Meddel. Lunds Univ. Math. Sen.* **3** (1935), 1–118. JFM 61.1262.02.

[GS01]   M. Goldstein and W. Schlag, Hölder continuity of the integrated density of states for quasi-periodic Schrödinger equations and averages of shifts of subharmonic functions, *Ann. of Math.* **154** (2001), 155–203. MR 1847592. Zbl 0990.39014. http://dx.doi.org/10.2307/3062114.

[GV12]   A. S. Golsefidy and P. P. Varjú, Expansion in perfect groups, *Geom. Funct. Anal.* **22** (2012), 1832–1891. MR 3000503. Zbl 06132223. http://dx.doi.org/10.1007/s00039-012-0190-7.

[Goo83]   A. Good, *Local Analysis of Selberg's Trace Formula*, *Lecture Notes in Math.* **1040**, Springer-Verlag, New York, 1983. MR 0727476. Zbl 0525.10013. http://dx.doi.org/10.1007/BFb0073074.

[Goo41]   I. J. Good, The fractional dimensional theory of continued fractions, *Proc. Cambridge Philos. Soc.* **37** (1941), 199–228. MR 0004878. Zbl 0061.09408. http://dx.doi.org/10.1017/S030500410002171X.

[Haa81]   U. Haagerup, The best constants in the Khintchine inequality, *Studia Math.* **70** (1981), 231–283 (1982). MR 0654838. Zbl 0501.46015.

[Hen89]   D. Hensley, The distribution of badly approximable numbers and continuants with bounded digits, in *Théorie des Nombres* (Quebec, PQ, 1987), de Gruyter, Berlin, 1989, pp. 371–385. MR 1024576. Zbl 0689.10042. http://dx.doi.org/10.1515/9783110852790.371.

[Hen92]   D. Hensley, Continued fraction Cantor sets, Hausdorff dimension, and functional analysis, *J. Number Theory* **40** (1992), 336–358. MR 1154044. Zbl 0745.28005. http://dx.doi.org/10.1016/0022-314X(92)90006-B.

[Hen06]   D. Hensley, *Continued Fractions*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2006. MR 2351741. Zbl 1161.11028. http://dx.doi.org/10.1142/9789812774682.

[Hen96]   D. Hensley, A polynomial time algorithm for the Hausdorff dimension of continued fraction Cantor sets, *J. Number Theory* **58** (1996), 9–45. MR 1387719. Zbl 0858.11039. http://dx.doi.org/10.1006/jnth.1996.0058.

[Hua13]   S. Huang, An improvement on Zaremba's conjecture, 2013. arXiv 1310.3772.

[Jen04]   O. Jenkinson, On the density of Hausdorff dimensions of bounded type continued fraction sets: the Texan conjecture, *Stoch. Dyn.* **4** (2004), 63–76. MR 2069367. Zbl 1089.28006. http://dx.doi.org/10.1142/S0219493704000900.

[JP01]   O. Jenkinson and M. Pollicott, Computing the dimension of dynam-
           ically defined sets: E_2 and bounded continued fractions, *Ergodic Theory
           Dynam. Systems* **21** (2001), 1429–1445. MR 1855840. Zbl 0991.28009.
           http://dx.doi.org/10.1017/S0143385701001687.

[Kim03]   H. H. Kim, Functoriality for the exterior square of GL_4 and the symmetric
           fourth of GL_2, *J. Amer. Math. Soc.* **16** (2003), 139–183. MR 1937203.
           Zbl 1018.11024. http://dx.doi.org/10.1090/S0894-0347-02-00410-1.

[Kon13]   A. Kontorovich, From Apollonius to Zaremba: local-global phenomena
           in thin orbits, *Bull. Amer. Math. Soc.* **50** (2013), 187–228. MR 3020826.
           Zbl 06157470. http://dx.doi.org/10.1090/S0273-0979-2013-01402-2.

[Lal89]   S. P. Lalley, Renewal theorems in symbolic dynamics, with applications
           to geodesic flows, non-Euclidean tessellations and their fractal limits, *Acta
           Math.* **163** (1989), 1–55. MR 1007619. Zbl 0701.58021. http://dx.doi.org/
           10.1007/BF02392732.

[MVW84]   C. R. Matthews, L. N. Vaserstein, and B. Weisfeiler, Congru-
           ence properties of Zariski-dense subgroups. I, *Proc. London Math. Soc.*
           48 (1984), 514–532. MR 0735226. Zbl 0551.20029. http://dx.doi.org/10.
           1112/plms/s3-48.3.514.

[McM09]   C. T. McMullen, Uniformly Diophantine numbers in a fixed real qua-
           dratic field, *Compos. Math.* **145** (2009), 827–844. MR 2521246. Zbl 1176.
           11032. http://dx.doi.org/10.1112/S0010437X09004102.

[Nau05]   F. Naud, Expanding maps on Cantor sets and analytic continuation of zeta
           functions, *Ann. Sci. École Norm. Sup.* **38** (2005), 116–153. MR 2136484.
           Zbl 1110.37021. http://dx.doi.org/10.1016/j.ansens.2004.11.002.

[Nie78]   H. Niederreiter, Quasi-Monte Carlo methods and pseudo-random num-
           bers, *Bull. Amer. Math. Soc.* **84** (1978), 957–1041. MR 0508447. Zbl 0404.
           65003. http://dx.doi.org/10.1090/S0002-9904-1978-14532-7.

[Zar72]   S. K. Zaremba, La méthode des "bons treillis" pour le calcul des intégrales
           multiples, in *Applications of Number Theory to Numerical Analysis* (Proc.
           Sympos., Univ. Montreal, Montreal, Que., 1971), Academic Press, New
           York, 1972, pp. 39–119. MR 0343530. Zbl 0246.65009.

(Received: August 2, 2011)
(Revised: August 8, 2013)

Institute for Advanced Study, Princeton, NJ
E-mail: bourgain@ias.edu

Stony Brook University, Stony Brook, NY
Current address: Yale University, New Haven, CT
E-mail: alex.kontorovich@yale.edu