BEYOND EXPANSION, III: RECIPROCAL GEODESICS

JEAN BOURGAIN and ALEX KONTOROVICH

Abstract

We prove the existence of infinitely many low-lying and fundamental closed geodesics on the modular surface which are reciprocal, that is, invariant under time reversal. The method combines ideas from parts I and II of this series, namely, the dispersion method in bilinear forms, as applied to thin semigroups coming from restricted continued fractions.

1. Introduction

We quote from Sarnak's lecture [18] regarding the genesis of the affine sieve (see [1], [2], [19]):

> For me the starting point of this investigation was in 2005 when Michel and Venkatesh asked me about the existence of poorly distributed closed geodesics on the modular surface. It was clear that Markov's constructions of his geodesics using his Markov equation provided what they wanted but they preferred quadratic forms with square free discriminant. This raised the question of sieving in this context of an orbit of a group of (nonlinear) morphisms of affine space.

The initial question, which arose in Einsiedler, Lindenstrauss, Michel, and Venkatesh's investigations into higher-rank analogues of Duke's theorem [9], asked (see the discussion below [10, Theorem 1.10]) for an infinitude of low-lying (i.e., being poorly distributed by not entering the cusp) *fundamental* geodesics (i.e., those corresponding to fundamental classes of binary quadratic forms). This problem was solved in part II of our series (see [7], [15] for a detailed discussion). But the question of an infinitude of fundamental Markov geodesics (for a discussion of Markov geodesics, see, e.g., [17, p. 226]) remains wide open, despite recent progress on the "strong approximation" aspect in [4] and [5]. Such geodesics are all reciprocal, that is, equivalent to themselves under time reversal of the geodesic flow. In this paper, we

DUKE MATHEMATICAL JOURNAL
Vol. 168, No. 18, © 2019 DOI 10.1215/00127094-2019-0056
Received 9 August 2018. Revision received 8 May 2019.
First published online 15 November 2019.
2010 *Mathematics Subject Classification*. Primary 11J70; Secondary 11N36, 37A45.

relax Markov geodesics to just low-lying ones and solve the problem of producing an infinitude of low-lying, fundamental, and reciprocal geodesics.

1.1. Statement of the main theorem

Before stating our main result, we give precise definitions of *low-lying*, *fundamental*, and *reciprocal*. By *closed geodesic*, we always mean primitive.

Definition 1.1

Given a compact subset $\mathcal{Y}$ of the unit tangent bundle of the modular surface

$$\mathcal{X} = T^1\big(\mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H}\big) \cong \mathrm{PSL}_2(\mathbb{Z})\backslash \mathrm{PSL}_2(\mathbb{R}),$$

a closed geodesic γ on $\mathcal{X}$ is called *low-lying* (with respect to $\mathcal{Y}$) if $\gamma \subset \mathcal{Y}$.

Definition 1.2

As is well known, closed geodesics on $\mathcal{X}$ are in one-to-one correspondence with primitive conjugacy classes of hyperbolic elements of $\mathrm{PSL}_2(\mathbb{Z})$ and with equivalence classes of indefinite binary quadratic forms (see, e.g., [15]). The latter have discriminants, and we say that a closed geodesic has discriminant D if its corresponding class does. The *trace* of a closed geodesic is that of its corresponding conjugacy class. Recall that a nonsquare discriminant D is called fundamental if it is the discriminant of the real quadratic field $\mathbb{Q}(\sqrt{D})$. We call a closed geodesic *fundamental* if its discriminant is.

Definition 1.3

The time-reversal symmetry on $\mathcal{X}$ corresponds to replacing all tangent vectors by their negatives; if a closed geodesic is invariant under this involution, it is called *reciprocal*.

Recall that the total number of all primitive closed geodesics, ordered by trace (which is equivalent to ordering by length), has the following well-known asymptotic:

$$\#\{\text{closed geodesics with trace} < X\} \sim \frac{X^2}{2\log X}.$$

There are about square-root as many reciprocal geodesics, which makes intuitive sense, as the geodesic has to spend the second half of its life undoing the twists of its first half.

THEOREM 1.4 (Sarnak [17, Theorem 2])

$$\#\{\textit{reciprocal geodesics with trace} < X\} \sim \frac{3}{8}X.$$

Our main result produces almost as many low-lying, fundamental, and reciprocal geodesics.

THEOREM 1.5 (Main theorem)
For any $\eta > 0$, there is a compact subset $\mathcal{Y} = \mathcal{Y}(\eta) \subset \mathcal{X}$ so that

$$\#\{\textit{low-lying, fundamental, reciprocal geodesics with trace} < X\} \gg_\eta X^{1-\eta}.$$

1.2. Ingredients

As in part II of our series (see [7]), we must study restricted continued fractions, and, to understand these, we use the semigroup

$$\Gamma_{\mathcal{A}} := \left\langle \begin{pmatrix} a & 1 \\ 1 & 0 \end{pmatrix} : a \le \mathcal{A} \right\rangle^{+} \cap \mathrm{SL}_2 \qquad (1.6)$$

of even-length words in the generators displayed. Write B_N for the archimedean ball in $\mathrm{SL}_2(\mathbb{R})$ with respect to the Frobenius metric:

$$B_N := \left\{ g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R}) : \mathrm{tr}(g^\dagger g) = a^2 + b^2 + c^2 + d^2 < N^2 \right\}.$$

Hensley [12] estimates the size of an archimedean ball in $\Gamma_{\mathcal{A}}$ to be

$$\#\Gamma_{\mathcal{A}} \cap B_N \asymp N^{2\delta_{\mathcal{A}}}, \qquad (1.7)$$

where $\delta_{\mathcal{A}}$ is the Hausdorff dimension of the limiting Cantor set,

$$\mathfrak{C}_{\mathcal{A}} := \{[0, a_1, a_2, \ldots] : a_j \le \mathcal{A} \text{ for all } j\}.$$

Here we are using the standard notation $x = [a_0, a_1, a_2, \ldots]$ for the continued fraction

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots}}.$$

These fractal dimensions are known to tend to 1 as $\mathcal{A} \to \infty$; indeed, Hensley [13] has shown that

$$\delta_{\mathcal{A}} = 1 - \frac{6}{\pi^2 \mathcal{A}} + o\Big(\frac{1}{\mathcal{A}}\Big). \qquad (1.8)$$

The following lemmas give sufficient conditions for a closed geodesic—represented by a hyperbolic conjugacy class $[\gamma]$ with $\gamma \in \mathrm{SL}_2(\mathbb{Z})$—to be fundamental and reciprocal.

LEMMA 1.9 ([7, Lemma 1.14])
A sufficient condition for a closed geodesic $[\gamma]$ *to be fundamental is that*

$$\mathrm{tr}(\gamma)^2 - 4 \textit{ is square-free.} \tag{1.10}$$

LEMMA 1.11 (see [17])
A sufficient condition for a closed geodesic $[\gamma_1]$ *to be reciprocal is that it is of the form* $\gamma_1 = \gamma^\dagger \gamma$, *for some* $\gamma \in \mathrm{SL}_2(\mathbb{Z})$.

We then reduce Theorem 1.5 to the following sieving result.

THEOREM 1.12
For any $\eta > 0$, *there is an* $\mathcal{A} = \mathcal{A}(\eta) < \infty$ *so that*

$$\#\{\gamma \in \Gamma_{\mathcal{A}} \cap B_N : \mathrm{tr}(\gamma^\dagger\gamma)^2 - 4 \textit{ is square-free}\} \gg N^{2-\eta}.$$

Remark 1.13
As in part II (see [7]), we cannot simply execute the affine sieve, because the "spectral gap" is insufficiently robust relative to the growth exponent δ_A, and we must produce an "exponent of distribution" going beyond that arising from expansion alone (see Remark 6.6). To do this, we again create certain "bilinear forms" and substitute "resonance" harmonics with abelian theory, which is much more tractable. Unlike in part II, the direct approach fails due to the nature of the quadratic forms arising in the error terms, and therefore a version of Linnik's "dispersion method" is needed. Fortunately, we recently developed such in the "orbital sieve" context in part I of our series (see [6]), which comes to the rescue here.

Remark 1.14
The main result in part II was proved unconditionally but would also follow immediately from a certain "local-global conjecture for thin orbits" (see the discussion in [15]). In contradistinction, Theorem 1.5 does *not* follow from this conjecture, because the function

$$\mathrm{SL}_2(\mathbb{Z}) \to \mathbb{Z} : \gamma \mapsto \mathrm{tr}(\gamma^\dagger\gamma)$$

is quadratic in the entries and so cannot be surjective when restricted to any $\Gamma_{\mathcal{A}}$; the image is itself thin! (For a definition of thinness in this context, see [14, p. 954].)

1.3. Organization

The rest of the paper is organized as follows. After some preliminary calculations in Section 2, we state the sieving theorem and construct the bilinear forms in Section 3 before analyzing the "main term" in Section 4. The error terms are analyzed in Section 5, after which the sieving theorem is proved in Section 6. Finally, putting together the above ingredients, we prove Theorem 1.5 in Section 7.

1.4. Notation

The transpose of a matrix γ is written ${}^\dagger\gamma$. When a calculation involves modular arithmetic, an overbar, $\bar{a}$, shall denote the multiplicative inverse of a. The constants C, c are absolute but may change from line to line. We use the notation $f \ll g$ and $f = O(g)$ to mean $f(x) \le Cg(x)$ for all $x > C$, where C is an implied constant. We write $f \asymp g$ for $g \ll f \ll g$. Unless otherwise specified, implied constants depend at most on $\mathcal{A}$, which is treated as fixed, and possibly on an arbitrarily small $\varepsilon > 0$.

2. Preliminaries

We recommend the technical estimates in this section be omitted on a first reading. They are only referenced as needed in the proof, which begins in Section 3.

2.1. Local estimates

We begin with some elementary computations.

LEMMA 2.1

For p an odd prime,

$$\#\{(x,y) \in \mathbb{F}_p^2 : x^2 + y^2 = 0\} = \begin{cases} 2p-1 & \text{if } p \equiv 1(4), \\ 1 & \text{if } p \equiv 3(4). \end{cases}$$

Moreover, for $\ell \not\equiv 0(p)$,

$$\#\{(x,y) \in \mathbb{F}_p^2 : x^2 + y^2 = \ell\} = \begin{cases} p-1 & \text{if } p \equiv 1(4), \\ p+1 & \text{if } p \equiv 3(4). \end{cases}$$

Proof

Elementary. □

For $\varpi = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, define

$$\mathfrak{f}(\varpi) := \operatorname{tr}(\varpi^\dagger \varpi) = a^2 + b^2 + c^2 + d^2, \tag{2.2}$$

and for $\epsilon = \pm 1$, set

$$\rho(p) := \frac{1}{|\,\mathrm{SL}_2(p)|} \sum_{\gamma \in \mathrm{SL}_2(p)} \mathbf{1}_{\{\mathfrak{f}(\gamma) \equiv 2\epsilon(p)\}}. \tag{2.3}$$

Extend the definition of ρ to all square-free q by multiplicativity. A priori, ρ seems to depend on ϵ, although the next lemma shows that it does not.

LEMMA 2.4
For p an odd prime,

$$\rho(p) = \begin{cases} \frac{2p-1}{p(p+1)} & \text{if } p \equiv 1(4), \\ \frac{1}{p(p-1)} & \text{if } p \equiv 3(4). \end{cases}$$

Also, $\rho(2) = 1/3$.

Proof
For $p = 2$, two of the six matrices in $\mathrm{SL}_2(2)$ have $\mathfrak{f} = 0$, so $\rho(2) = 2/6$. Now assume that $p \geq 3$. We need to count the number of $(a, b, c, d) \in \mathbb{F}_p^4$ with

$$a^2 + b^2 + c^2 + d^2 = 2\epsilon \qquad \text{and} \qquad ad - bc = 1.$$

We make the following linear change of variables:

$$a = x + y, \qquad d = x - y, \qquad b = w + z, \qquad c = w - z, \tag{2.5}$$

which is invertible since $p \neq 2$. The equations become

$$x^2 + y^2 + z^2 + w^2 = \epsilon \qquad \text{and} \qquad x^2 - y^2 + z^2 - w^2 = 1$$

or, equivalently,

$$x^2 + z^2 = 1 + y^2 + w^2 = 1 + \bar{2}(\epsilon - 1) = \begin{cases} 1 & \text{if } \epsilon = 1, \\ 0 & \text{if } \epsilon = -1. \end{cases} \tag{2.6}$$

Using Lemma 2.1 and $|\,\mathrm{SL}_2(p)| = p(p-1)(p+1)$ gives the claim. □

Given $n \in \mathbb{Z}$, define $\Xi(q; n)$ on square-free q by the expression

$$\Xi(p; n) := \mathbf{1}_{\{n \equiv 0(p)\}} - \rho(p) \tag{2.7}$$

on primes p, and extend multiplicatively to q.

LEMMA 2.8
For any $\omega \in \mathrm{SL}_2(p)$ *and* $\epsilon = \pm 1$,

$$\frac{1}{|\,\mathrm{SL}_2(p)|} \sum_{\gamma \in \mathrm{SL}_2(p)} \Xi\big(p; \mathfrak{f}(\gamma\omega) - 2\epsilon\big) = 0.$$

Proof
The coset ω plays no role since the γ sum is over all of $\mathrm{SL}_2(p)$. The lemma follows from definition (2.3) of ρ. □

The key estimate of this subsection is the following.

PROPOSITION 2.9
Let $\omega, \omega' \in \mathrm{SL}_2(p)$ *and* $\epsilon, \epsilon' \in \{\pm 1\}$. *Then*

$$\frac{1}{|\,\mathrm{SL}_2(p)|} \sum_{\gamma \in \mathrm{SL}_2(p)} \Xi\big(p; \mathfrak{f}(\gamma\omega) - 2\epsilon\big) \Xi\big(p; \mathfrak{f}(\gamma\omega') - 2\epsilon'\big) \ll \begin{cases} \frac{1}{p} & \text{if } \omega \in \omega' \cdot \mathrm{PO}_2(p), \\ \frac{1}{p^2} & \text{otherwise.} \end{cases} \tag{2.10}$$

Here we have defined

$$\mathrm{PO}_2(p) := \big\{k \in \mathrm{SL}_2(p) : k^\dagger k \equiv \pm I(p)\big\} = \left\{\begin{pmatrix} a & b \\ -b & a \end{pmatrix} : a^2 + b^2 \equiv \pm 1(p)\right\}. \tag{2.11}$$

Proof
Expanding Ξ and using the definition (2.3) of ρ, we must estimate

$$\frac{1}{|\,\mathrm{SL}_2(p)|} \sum_{\gamma \in \mathrm{SL}_2(p)} \mathbf{1}_{\{\mathfrak{f}(\gamma\omega) \equiv 2\epsilon(p)\}} \mathbf{1}_{\{\mathfrak{f}(\gamma\omega') \equiv 2\epsilon'(p)\}} - \rho(p)^2. \tag{2.12}$$

The second term is plainly $\ll p^{-2}$ by Lemma 2.4.

If $p \equiv 3(4)$, then we may trivially bound $\mathbf{1}_{\{\mathfrak{f}(\gamma\omega') \equiv 2\epsilon'(p)\}} \leq 1$, whence the first term is $\rho(p) = 1/(p(p-1)) \ll 1/p^2$, as desired. Thus we may restrict to $p \equiv 1(4)$.

If $\omega \in \omega' \cdot \mathrm{PO}_2(p)$, then $\mathfrak{f}(\gamma\omega) = \pm\mathfrak{f}(\gamma\omega')$, so if the signs ϵ, ϵ' align, then the first term in (2.12) could be exactly $\rho(p) = (2p-1)/(p(p+1)) \asymp 1/p$. Thus we cannot do better than $1/p$ in this case. Now we seek extra cancellation when $\omega \notin \omega' \cdot \mathrm{PO}_2(p)$.

Write

$$(\omega^{-1}\omega')^\dagger(\omega^{-1}\omega') =: \begin{pmatrix} U & V \\ V & W \end{pmatrix}.$$

Changing $\gamma \mapsto \gamma\omega^{-1}$ in (2.12) and using (2.2), we must bound

$$\frac{1}{|\operatorname{SL}_2(p)|} \sum_{\gamma \in \operatorname{SL}_2(p)} \mathbf{1}_{\{\mathfrak{f}(\gamma) \equiv 2\epsilon\}} \mathbf{1}_{\left\{\operatorname{tr}\left(\gamma^\dagger\gamma\left(\begin{smallmatrix} U & V \\ V & W \end{smallmatrix}\right)\right) \equiv 2\epsilon'\right\}}.$$

Writing $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, the last equation becomes

$$U(a^2+b^2) + 2V(ac+bd) + W(c^2+d^2) \equiv 2\epsilon'.$$

Apply the same change of variables as in (2.5); then the equations become (2.6) and

$$U\big(1+2(xy+zw)\big) + 4V(xz-yw) + W\big(1-2(xy+zw)\big) \equiv 2\epsilon'. \tag{2.13}$$

Now suppose that $\epsilon = 1$ (the case $\epsilon = -1$ being similar). Then $x^2+z^2 = 1$ and $p \equiv 1(4)$, so there are $p-1$ choices of (x,z) by Lemma 2.1. With (x,z) fixed, (2.13) becomes linear in (y,w); we isolate y:

$$2\big[(U-W)x - 2Vw\big]y \equiv 2\epsilon' - (U+W) - 4Vxz - 2(U-W)zw.$$

Square and add $(2[(U-W)x-2Vw]w)^2$ to both sides to take advantage of $y^2+w^2 = 0$. This gives a quartic equation in w with everything else determined:

$$\begin{aligned} 0 &= \big[2\epsilon' - (U+W) - 4Vxz - 2(U-W)zw\big]^2 \\ &\quad + \big(2\big[(U-W)x - 2Vw\big]w\big)^2 \\ &= \big[2\epsilon' - (U+W) - 4Vxz\big]^2 - 4\big[2\epsilon' - (U+W) - 4Vxz\big](U-W)zw \\ &\quad + 4(U-W)^2w^2 - 16(U-W)Vxw^3 + 16V^2w^4. \end{aligned}$$

This equation has at most four solutions in w, unless all the coefficients vanish, in which case $V = 0$ and $U = W$. But $\det\left(\begin{smallmatrix} U & V \\ V & W \end{smallmatrix}\right) = UW - V^2 = 1$, so $V = 0$ implies $U = \overline{W} = W$. Hence $U = W = \pm 1$, which means that $\omega^{-1}\omega' \in \operatorname{PO}_2(p)$. Since we have already dealt with this case, we may assume that the coefficients do not all vanish, whence there are at most four choices for w, from which y is determined. In summary, there are $\ll p$ choices for (x,z) and a bounded number of choices of (y,w), while $|\operatorname{SL}_2(p)| \asymp p^3$; the ratio is $\ll 1/p^2$, as claimed. □

2.2. Spectral and automorphic estimates

We import here some lemmas from [7], the first of which is an automorphic estimate in $\mathrm{SL}_2(\mathbb{Z})$.

LEMMA 2.14 ([7, Lemma 2.13])

Let $X \gg 1$ be an increasing parameter. Then there is a smooth bump function $\varphi_X : \mathrm{SL}_2(\mathbb{R}) \to \mathbb{R}_{\geq 0}$ with the following properties:

- *It gives support to the norm-X ball: if $\|g\| := \sqrt{\mathrm{tr}(g^\dagger g)} < X$, then*

$$\varphi_X(g) \geq 1. \tag{2.15}$$

- *Furthermore,*

$$\sum_{\gamma \in \mathrm{SL}_2(\mathbb{Z})} \varphi_X(\gamma) \ll X^2. \tag{2.16}$$

- *Finally, φ_X is evenly distributed in progressions: for any square-free q and any $\gamma_0 \in \mathrm{SL}_2(q)$,*

$$\sum_{\substack{\gamma \in \mathrm{SL}_2(\mathbb{Z}) \\ \gamma \equiv \gamma_0 (q)}} \varphi_X(\gamma) = \frac{1}{|\mathrm{SL}_2(q)|} \sum_{\gamma \in \mathrm{SL}_2(\mathbb{Z})} \varphi_X(\gamma) + O(X^{3/2}). \tag{2.17}$$

All implied constants above are absolute.

Remark 2.18

The error $X^{3/2}$ in (2.17) comes from using Selberg's $3/16$ spectral gap (see [20]); we are striving for the simplest explicit exponents here, not optimal ones, and so we do not bother using the best available exponents.

Finally, we will need super-approximation in our thin semigroup $\Gamma_{\mathcal{A}}$. As discussed in Remark 1.13, we need this spectral gap to be absolute, and so we pick a fixed parameter $\mathcal{A}_0 = 2$; then $\Gamma_{\mathcal{A}_0} = \Gamma_2$.

LEMMA 2.19

For any $Y \gg 1$, there is a nonempty subset

$$\aleph \subset \{\gamma \in \Gamma_2 : \|\gamma\| < Y\}$$

and "spectral gap"

$$\Theta > 0, \tag{2.20}$$

so that, for any q and any $\mathfrak{a}_0 \in \mathrm{SL}_2(q)$,

$$\#\{\mathfrak{a} \in \aleph : \mathfrak{a} \equiv \mathfrak{a}_0(q)\} = \frac{1}{|\,\mathrm{SL}_2(q)|}|\aleph| + O\big(|\aleph| q^C Y^{-\Theta}\big). \tag{2.21}$$

Here C, Θ, and the implied constant are all absolute.

Proof
A nearly identical statement is proved in [7, Proposition 2.9] with a weaker error term. The main ingredient there is a "prime number theorem"-type resonance-free region as proved in [3]. Now a resonance-free strip is available (and does not require q to be square-free) due to Magee, Oh, and Winter [16] and Bourgain, Kontorovich, and Magee [8]; substituting this result into the proof of [7, Proposition 2.9] gives the above claim. □

Remark 2.22
Actually, the weaker statement [7, Proposition 2.9] using only [3] would already suffice for our purposes (see the treatment in [7]). The resonance-free strip slightly simplifies the exposition, and so we use it here.

3. Construction of Π and the sieving theorem

3.1. Construction of the set Π

We create here a certain subset $\Pi \subset \Gamma_{\mathcal{A}}$, all elements $\varpi \in \Pi$ being of size $\|\varpi\| \ll N$, a growing parameter, with Π exhibiting a mutlilinear structure. First we break the parameter N as

$$XYZ = N, \tag{3.1}$$

and take the set $\aleph$ from Lemma 2.19 with parameter Y.

The elements $\gamma \in \Gamma_{\mathcal{A}}$ of size $\|\gamma\| < X$ all have word-length $\ell(\gamma) \asymp \log X$ in the generators (1.6). By the pigeonhole principle, there is therefore a subset Ω_X of $\Gamma_{\mathcal{A}} \cap B_X$ of size

$$\#\Omega_X \gg \frac{X^{2\delta}}{\log X}, \tag{3.2}$$

(cf. (1.7)) all having the same word-length. (We henceforth write δ for $\delta_{\mathcal{A}}$, treating $\mathcal{A}$ as fixed.) In the same way we construct the set Ω_Z to parameter Z.

Then the set

$$\Pi := \Omega_X \cdot \aleph \cdot \Omega_Z \tag{3.3}$$

is a genuine subset (as opposed to a multiset) of $\Gamma_{\mathcal{A}}$, since each

$$\varpi = \xi \cdot \mathfrak{a} \cdot \omega, \qquad \begin{cases} \xi \in \Omega_X, \\ \mathfrak{a} \in \aleph, \\ \omega \in \Omega_Z, \end{cases}$$

is uniquely represented.

3.2. *The sieving theorem*

In light of Lemmas 1.9 and 1.11, we define Π_{AP} to be the set of $\varpi \in \Pi$ for which $\operatorname{tr}(\varpi^\dagger \varpi)^2 - 4$ has no small prime factors:

$$\Pi_{AP} := \left\{\varpi \in \Pi : p \,\middle|\, \left(\operatorname{tr}(\varpi^\dagger \varpi)^2 - 4\right) \Longrightarrow p > N^{1/350}\right\}. \tag{3.4}$$

An easy consequence of the main sieving theorem stated below is the following.

THEOREM 3.5
For any small $\eta > 0$, there is an $\mathcal{A} = \mathcal{A}(\eta)$, sufficiently large, and a choice of parameters X, Y, Z in (3.1) so that

$$\#\Pi_{AP} > N^{2\delta - \eta}, \tag{3.6}$$

as $N \to \infty$.

The aforementioned sieving theorem is the following "level of distribution" result.

Recalling $\mathfrak{f}$ defined in (2.2), our sifting sequence is $\mathfrak{A} = \{a_N(n)\}$ with

$$a_N(n) := \sum_{\varpi \in \Pi} \mathbf{1}_{\mathfrak{f}(\varpi)^2 - 4 = n}.$$

Note that $\mathfrak{A}$ is supported on $n < T$, where

$$T \asymp N^4. \tag{3.7}$$

For square-free $\mathfrak{q} \geq 1$, write

$$|\mathfrak{A}_\mathfrak{q}| := \sum_{n \equiv 0(\mathfrak{q})} a_N(n),$$

which measures the distribution of a_N on multiplies of $\mathfrak{q}$.

THEOREM 3.8 (The sieving theorem)
For any small $\eta > 0$, there is a sufficiently large $\mathcal{A} = \mathcal{A}(\eta)$ and a choice of the parameters X, Y, Z so that the following holds. Given a square-free $\mathfrak{q}$, there is a decomposition

$$|\mathfrak{A}_{\mathfrak{q}}| = \beta(\mathfrak{q}) \cdot |\Pi| + r(\mathfrak{q}). \tag{3.9}$$

The function β is multiplicative and satisfies the "quadratic sieve" condition:

$$\prod_{w \le p < z} \big(1 - \beta(p)\big)^{-1} \le C \cdot \Big(\frac{\log z}{\log w}\Big)^2. \tag{3.10}$$

Moreover, the "remainder" term $r(\mathfrak{q})$ is controlled by

$$\sum_{\substack{\mathfrak{q} < \mathcal{Q} \\ squarefree}} \big|r(\mathfrak{q})\big| \ll_K \frac{|\Pi|}{\log^K N}, \quad \text{for any } K < \infty, \tag{3.11}$$

where the "level of distribution" $\mathcal{Q}$ can be taken as large as

$$\mathcal{Q} = T^{1/72-\eta}. \tag{3.12}$$

Finally, the set Π is large:

$$|\Pi| > N^{2\delta-\eta}. \tag{3.13}$$

The deduction of Theorem 3.5 from Theorem 3.8 is completely standard, so we give a quick sketch.

Sketch
The sifting sequence $\mathfrak{A}$ has "sieve dimension" $\kappa = 2$, and any exponent of distribution $\alpha < 1/72$. Taking, say, $\alpha = 1/73$ (again, we are not striving for optimal exponents), and using the crudest Brun sieve (see, e.g., [11, Theorem 6.9]), one shows that

$$\sum_{\substack{n \\ (n,P_z)=1}} a_N(n) \gg \frac{|\Pi|}{(\log N)^2}, \tag{3.14}$$

where $P_z = \prod_{p<z} p$ and z does not exceed $T^{\alpha/(9\kappa+1)} = T^{1/1387} = N^{4/1387}$; we take $z = N^{4/1400} = N^{1/350}$. Of course, any $n = \mathrm{tr}(\varpi)^2 - 4$ coprime to P_z has no prime factors below z. Then (3.14) and (3.13) confirm (3.6) after renaming constants. □

We focus henceforth on establishing Theorem 3.8.

3.3. *The decomposition and dispersion*

To prepare the proof, write, for square-free $\mathfrak{q} \ge 1$,

$$|\mathfrak{A}_{\mathfrak{q}}| := \sum_{n \equiv 0(\mathfrak{q})} a_N(n) = \sum_{\substack{\tau \bmod \mathfrak{q} \\ \tau^2 \equiv 4(\mathfrak{q})}} \sum_{\varpi \in \Pi} \mathbf{1}_{\{\mathfrak{f}(\varpi) - \tau \equiv 0(\mathfrak{q})\}}.$$

To apply the "dispersion" method, we write

$$\mathbf{1}_{n\equiv 0(p)} = \Xi(p;n) + \rho(p),$$

with ρ and Ξ defined in (2.3) and (2.7), respectively.

Then

$$\begin{aligned}|\mathfrak{A}_\mathfrak{q}| &= \sum_{\substack{\tau \bmod \mathfrak{q}\\ \tau^2\equiv 4(\mathfrak{q})}} \sum_{\varpi\in\Pi} \prod_{p|\mathfrak{q}} \bigl(\Xi\bigl(p;\mathfrak{f}(\varpi)-\tau\bigr)+\rho(p)\bigr)\\ &= \sum_{q|\mathfrak{q}} \sum_{\substack{\tau \bmod \mathfrak{q}\\ \tau^2\equiv 4(\mathfrak{q})}} \sum_{\varpi\in\Pi} \Xi\bigl(q;\mathfrak{f}(\varpi)-\tau\bigr)\rho\Bigl(\frac{\mathfrak{q}}{q}\Bigr). \qquad (3.15)\end{aligned}$$

To give a decomposition toward (3.9), we break the sum

$$|\mathfrak{A}_\mathfrak{q}| = \mathcal{M}_\mathfrak{q} + r(\mathfrak{q}) \qquad (3.16)$$

according to whether $q < Q_0$ or not. The two contributions are dealt with separately in the next two sections.

4. Main-term analysis

From the decomposition (3.16) of $\mathfrak{A}_\mathfrak{q}$ in (3.15) the "main" term is

$$\mathcal{M}_\mathfrak{q} = \sum_{\substack{q|\mathfrak{q}\\ q<Q_0}} \sum_{\substack{\tau(\mathfrak{q})\\ \tau^2\equiv 4}} \sum_{\varpi\in\Pi} \Xi\bigl(q;\mathfrak{f}(\varpi)-\tau\bigr)\rho\Bigl(\frac{\mathfrak{q}}{q}\Bigr). \qquad (4.1)$$

The main goal of this section is to prove the following.

THEOREM 4.2
We have that

$$\mathcal{M}_\mathfrak{q} = \beta(\mathfrak{q})|\Pi| + r^{(1)}(\mathfrak{q}), \qquad (4.3)$$

where β is a multiplicative function defined on the primes by

$$\beta(p) := \begin{cases} \frac{1}{3} & \text{if } p = 2,\\ \frac{2(2p-1)}{p(p+1)} & \text{if } p\equiv 1(4),\\ \frac{2}{p(p-1)} & \text{if } p \equiv 3(4),\end{cases} \qquad (4.4)$$

and the "remainder" term $r^{(1)}$ satisfies

$$\sum_{\mathfrak{q}<\mathcal{Q}} \bigl|r^{(1)}(\mathfrak{q})\bigr| \ll |\Pi|\mathcal{Q}^{\varepsilon} Q_0^{C} Y^{-\Theta}. \qquad (4.5)$$

To begin the proof, insert the construction (3.3) of Π into (4.1), writing $\varpi = \xi \mathfrak{a} \omega$. Since $\Xi(q;*)$ depends only on $*$ mod q, we decompose the $\mathfrak{a}$ sum along progressions mod q,

$$\mathcal{M}_{\mathfrak{q}} = \sum_{\substack{q \mid \mathfrak{q} \\ q<Q_0}} \rho\Big(\frac{\mathfrak{q}}{q}\Big) \sum_{\substack{\tau(\mathfrak{q}) \\ \tau^2 \equiv 4}} \sum_{\xi \in \Omega_X} \sum_{\omega \in \Omega_Z} \sum_{\mathfrak{a}_0 \in \mathrm{SL}_2(q)} \Xi\big(q; \mathfrak{f}(\xi \mathfrak{a}_0 \omega) - \tau\big) \Big[\sum_{\mathfrak{a} \in \aleph} \mathbf{1}_{\mathfrak{a} \equiv \mathfrak{a}_0(q)}\Big],$$

and apply expansion (2.21),

$$= \mathcal{M}_{\mathfrak{q}}^{(1)} + r^{(1)}(\mathfrak{q}),$$

where

$$\mathcal{M}_{\mathfrak{q}}^{(1)} := |\Pi| \sum_{\substack{\tau(\mathfrak{q}) \\ \tau^2 \equiv 4}} \sum_{\substack{q \mid \mathfrak{q} \\ q<Q_0}} \rho\Big(\frac{\mathfrak{q}}{q}\Big) \Big[\frac{1}{|\mathrm{SL}_2(q)|} \sum_{\mathfrak{a}_0 \in \mathrm{SL}_2(q)} \Xi\big(q; \mathfrak{f}(\mathfrak{a}_0) - \tau\big)\Big] \tag{4.6}$$

and

$$\begin{aligned} \big|r^{(1)}(\mathfrak{q})\big| &\ll \sum_{\substack{q \mid \mathfrak{q} \\ q<Q_0}} \rho\Big(\frac{\mathfrak{q}}{q}\Big) \sum_{\substack{\tau(\mathfrak{q}) \\ \tau^2 \equiv 4}} \sum_{\xi \in \Omega_X} \sum_{\omega \in \Omega_Z} \sum_{\mathfrak{a}_0 \in \mathrm{SL}_2(q)} \big|\Xi\big(q; \mathfrak{f}(\xi \mathfrak{a}_0 \omega) - \tau\big)\big| |\aleph| q^C Y^{-\Theta} \\ &\ll \frac{\mathfrak{q}^{\varepsilon}}{\mathfrak{q}} |\Pi| Q_0^C Y^{-\Theta}. \end{aligned}$$

Here we used $|\Xi| \le 1$ and Lemma 2.4 that $\rho(q) \ll q^{\varepsilon}/q$. Then (4.5) is immediately satisfied.

Returning to $\mathcal{M}_{\mathfrak{q}}^{(1)}$ in (4.6), the bracketed term vanishes unless $q = 1$ by Lemma 2.8, and so we are left with

$$\mathcal{M}_{\mathfrak{q}}^{(1)} = |\Pi| 2^{\nu(\mathfrak{q}) - \mathbf{1}_{\{2 \mid \mathfrak{q}\}}} \rho(\mathfrak{q}).$$

Here we elementarily evaluated the contribution from the τ summation (see [7, Lemma 4.1]). Inserting Lemma 2.4, we see that (4.4) is verified, completing the proof of Theorem 4.2.

5. Error-term analysis

The remainder term from (3.16) is

$$r(\mathfrak{q}) := \sum_{\substack{q \mid \mathfrak{q} \\ q>Q_0}} \sum_{\substack{t \bmod q \\ t^2 \equiv 4(q)}} \sum_{\varpi \in \Pi} \Xi\big(q; \mathfrak{f}(\varpi) - t\big) \sum_{\substack{\tau \bmod \mathfrak{q} \\ \tau^2 \equiv 4(\mathfrak{q}), \tau \equiv t(q)}} \rho\Big(\frac{\mathfrak{q}}{q}\Big),$$

and the total error is

$$\mathcal{E} := \sum_{\mathfrak{q}<\mathcal{Q}} \bigl|r(\mathfrak{q})\bigr|;$$

we need to save a little more than $\mathcal{Q}$ off of the trivial bound.

The goal of this section is to prove the following.

THEOREM 5.1
We have that

$$\mathcal{E} \ll T^{\varepsilon}|\Pi|(XZ)^{1-\delta}\Bigl[\frac{\mathcal{Q}^4}{X^{1/4}} + \frac{1}{Q_0^{1/2}} + \frac{\mathcal{Q}^{1/2}}{Z^{1/4}}\Bigr]. \tag{5.2}$$

First write $\mathcal{E}$ as

$$\mathcal{E} = \sum_{\mathfrak{q}<\mathcal{Q}} \zeta(\mathfrak{q}) r(\mathfrak{q}),$$

where $\zeta(\mathfrak{q}) = \overline{r(\mathfrak{q})}/|r(\mathfrak{q})| = \operatorname{sgn} r(\mathfrak{q})$. Expanding gives

$$\mathcal{E} = \sum_{Q_0<q<\mathcal{Q}} \sum_{\substack{t \bmod q \\ t^2\equiv 4(q)}} \sum_{\varpi\in\Pi} \Xi\bigl(q; \mathfrak{f}(\varpi) - t\bigr)\zeta_1(q,t),$$

where

$$\zeta_1(q,t) := \sum_{\substack{\mathfrak{q}<\mathcal{Q} \\ \mathfrak{q}\equiv 0(q)}} \zeta(\mathfrak{q}) \sum_{\substack{\tau \bmod \mathfrak{q} \\ \tau^2\equiv 4(\mathfrak{q}), \tau\equiv t(q)}} \rho\Bigl(\frac{\mathfrak{q}}{q}\Bigr) \ll T^{\varepsilon} \sum_{\mathfrak{q}<\mathcal{Q}/q} \frac{1}{\mathfrak{q}} \ll T^{\varepsilon}.$$

Decomposing Π as $\Omega_X \aleph \Omega_Z$ gives

$$\begin{aligned}\mathcal{E} &= \sum_{Q_0<q<\mathcal{Q}} \sum_{\substack{t \bmod q \\ t^2\equiv 4(q)}} \sum_{\gamma\in\Omega_X} \sum_{\mathfrak{a}\in\aleph} \sum_{\omega\in\Omega_Z} \Xi\bigl(q; \mathfrak{f}(\gamma\mathfrak{a}\omega) - t\bigr)\zeta_1(q,t), \\ &\ll \sum_{\substack{Q_0<Q<\mathcal{Q} \\ \text{dyadic}}} \sum_{\mathfrak{a}\in\aleph} \bigl|\mathcal{E}_1(\mathfrak{a}, Q)\bigr|,\end{aligned}$$

where

$$\mathcal{E}_1(\mathfrak{a}, Q) := \sum_{q\asymp Q} \sum_{\substack{t \bmod q \\ t^2\equiv 4(q)}} \sum_{\gamma\in\Omega_X} \sum_{\omega\in\Omega_Z} \Xi\bigl(q; \mathfrak{f}(\gamma\mathfrak{a}\omega) - t\bigr)\zeta_1(q,t).$$

Theorem 5.1 follows immediately from the next result.

PROPOSITION 5.3
We have that

$$\big|\mathcal{E}_1(\mathfrak{a},Q)\big| \ll T^{\varepsilon}|\Omega_X||\Omega_Z|(XZ)^{1-\delta}\Big[\frac{Q^4}{X^{1/4}}+\frac{1}{Q^{1/2}}+\frac{Q^{1/2}}{Z^{1/4}}\Big]. \tag{5.4}$$

To begin the proof, apply the Cauchy–Schwarz inequality in the γ variable and insert the smooth bump function φ_X from Lemma 2.14:

$$\begin{aligned}\big|\mathcal{E}_1(\mathfrak{a},Q)\big|^2 &\ll |\Omega_X|\cdot\sum_{\gamma\in\mathrm{SL}_2(\mathbb{Z})}\varphi_X(\gamma)\Big|\sum_{q\asymp Q}\sum_{\substack{t \bmod q\\ t^2\equiv 4(q)}}\sum_{\omega\in\Omega_Z}\Xi\big(q;\mathfrak{f}(\gamma\mathfrak{a}\omega)-t\big)\zeta_1(q,t)\Big|^2\\ &\ll |\Omega_X|\cdot T^{\varepsilon}\sum_{q,q'\asymp Q}\sum_{\substack{t \bmod q\\ t^2\equiv 4(q)}}\sum_{\substack{t' \bmod q'\\ (t')^2\equiv 4(q')}}\sum_{\omega,\omega'\in\Omega_Z}\\ &\qquad\times\Big|\sum_{\gamma\in\mathrm{SL}_2(\mathbb{Z})}\varphi_X(\gamma)\Xi\big(q;\mathfrak{f}(\gamma\mathfrak{a}\omega)-t\big)\Xi\big(q';\mathfrak{f}(\gamma\mathfrak{a}\omega')-t'\big)\Big|. \end{aligned}\tag{5.5}$$

Having applied Cauchy–Schwarz, we now need to save a little more than Q^2. We first address the innermost γ sum.

LEMMA 5.6
Let

$$\mathfrak{q}_1=\mathfrak{q}_1(\omega,\omega';q):=\max_{\pm}\big(\gcd\big(q,(\omega^{-1}\omega')^{\dagger}(\omega^{-1}\omega')\mp I\big)\big),$$

so that $\mathfrak{q}_1\mid q$ *is the largest modulus for which* $\omega^{-1}\omega'\in\mathrm{PO}_2(\mathfrak{q}_1)$, *the group defined in (2.11). Then*

$$\begin{aligned}&\Big|\sum_{\gamma\in\mathrm{SL}_2(\mathbb{Z})}\varphi_X(\gamma)\Xi\big(q;\mathfrak{f}(\gamma\mathfrak{a}\omega)-t\big)\Xi\big(q';\mathfrak{f}(\gamma\mathfrak{a}\omega')-t'\big)\Big|\\ &\qquad\ll Q^6X^{3/2}+\mathbf{1}_{\{q=q'\}}Q^{\varepsilon}X^2\frac{\mathfrak{q}_1}{q^2}.\end{aligned}\tag{5.7}$$

Remark 5.8
The first term above is a savings of $X^{1/2}$ against the loss of some powers of Q, which is more than the requisite Q^2 savings, as long as Q is not too large relative to X. The second term is a savings of Q from the $q=q'$ restriction and a second factor of $Q^2/\mathfrak{q}_1$ from the $\mathfrak{q}_1/q^2$ term. If $\mathfrak{q}_1$ is small, then this already saves more than Q^2, but if $\mathfrak{q}_1$ is of size q, then the net savings is Q^2 but no more. In that case, we will

need just a bit of extra savings from the fact that $\omega^{-1}\omega' \in \mathrm{PO}_2(\mathfrak{q}_1)$ with such a large modulus $\mathfrak{q}_1$.

Proof of Lemma 5.6
Let

$$\bar{q} := [q,q'] = \operatorname{lcm}(q,q'), \qquad \tilde{q} = (q,q') = \gcd(q,q'), \qquad q = q_1\tilde{q}, \qquad q' = q_1'\tilde{q},$$

with q_1, q_1', and $\tilde{q}$ pairwise coprime. Because $\Xi(q,n)$ depends only on the residue of $n \bmod q$, we break the innermost γ sum into progressions, obtaining

$$\begin{aligned}\Big|\sum_{\gamma\in \mathrm{SL}_2(\mathbb{Z})}\Big| &= \sum_{\gamma_0\in \mathrm{SL}_2(\bar{q})} \Xi\big(q;\mathfrak{f}(\gamma_0\mathfrak{a}\omega)-t\big)\,\Xi\big(q';\mathfrak{f}(\gamma_0\mathfrak{a}\omega')-t'\big)\Big[\sum_{\substack{\gamma\in \mathrm{SL}_2(\mathbb{Z})\\ \gamma\equiv\gamma_0(\bar{q})}} \varphi_X(\gamma)\Big]\\ &\ll X^2\Big|\frac{1}{|\,\mathrm{SL}_2(\bar{q})|}\sum_{\gamma_0\in \mathrm{SL}_2(\bar{q})} \Xi\big(q;\mathfrak{f}(\gamma_0\mathfrak{a}\omega)-t\big)\,\Xi\big(q';\mathfrak{f}(\gamma_0\mathfrak{a}\omega')-t'\big)\Big|\\ &\quad + O(\bar{q}^3X^{3/2}),\end{aligned} \tag{5.9}$$

where we used (2.17) and (2.16). Since $\bar{q} \ll Q^2$, the last term contributes $Q^6X^{3/2}$ to (5.7).

Now, the remaining γ_0 sum in (5.9) is multiplicative, and so decomposing $\bar{q} = q_1q_1'\tilde{q}$, we can write it as

$$\begin{aligned}\Big|\sum_{\gamma_0\in \mathrm{SL}_2(\bar{q})}\Big| &= \Big|\frac{1}{|\,\mathrm{SL}_2(q_1)|}\sum_{\gamma_0\in \mathrm{SL}_2(q_1)} \Xi\big(q_1;\mathfrak{f}(\gamma_0\mathfrak{a}\omega)-t\big)\Big|\\ &\quad\times\Big|\frac{1}{|\,\mathrm{SL}_2(q_1')|}\sum_{\gamma_0\in \mathrm{SL}_2(q_1')} \Xi\big(q_1';\mathfrak{f}(\gamma_0\mathfrak{a}\omega')-t'\big)\Big|\\ &\quad\times\Big|\frac{1}{|\,\mathrm{SL}_2(\tilde{q})|}\sum_{\gamma_0\in \mathrm{SL}_2(\tilde{q})} \Xi\big(\tilde{q};\mathfrak{f}(\gamma_0\mathfrak{a}\omega)-t\big)\,\Xi\big(\tilde{q};\mathfrak{f}(\gamma_0\mathfrak{a}\omega')-t'\big)\Big|.\end{aligned}$$

From Lemma 2.8, we see that the first two terms completely vanish unless $q_1 = q_1' = 1$, that is, $q = q' = \tilde{q} = \bar{q}$. For the third term, we apply the key Proposition 2.9; then every $p \mid q$ contributes either $1/p$ or $1/p^2$, depending on whether $\omega^{-1}\omega' \in \mathrm{PO}_2(p)$ or not. This savings is exactly captured by $Q^{\varepsilon}\mathfrak{q}_1/q^2$, completing the proof. □

Proof of Proposition 5.3
Inserting (5.7) into (5.5) gives

$$\begin{aligned}\big|\mathcal{E}_1(\mathfrak{a},Q)\big|^2 &\ll T^\varepsilon|\Omega_X|Q^2|\Omega_Z|^2Q^6X^{3/2}\\ &\quad + T^\varepsilon|\Omega_X|\sum_{q\asymp Q}\sum_{\omega\in\Omega_Z}\sum_{\mathfrak{q}_1|q}X^2\frac{\mathfrak{q}_1}{q^2}\Big[\sum_{\substack{\omega'\in\mathrm{SL}_2(\mathbb{Z})\\ \omega^{-1}\omega'\in\mathrm{PO}_2(\mathfrak{q}_1)}}\varphi_Z(\omega')\Big], \qquad (5.10)\end{aligned}$$

where we extended the ω' sum to all of $\mathrm{SL}_2(\mathbb{Z})$ and again inserted the bump function φ from Lemma 2.14. Now break the innermost ω' sum in (5.10) into progressions mod $\mathfrak{q}_1$, and apply (2.17) and (2.16) to obtain

$$\begin{aligned}\Big[\sum_{\omega'\in\mathrm{SL}_2(\mathbb{Z})}\Big] &= \sum_{\omega_0'\in\omega\cdot\mathrm{PO}_2(\mathfrak{q}_1)}\sum_{\substack{\omega'\in\mathrm{SL}_2(\mathbb{Z})\\ \omega'\equiv\omega_0'(\mathfrak{q}_1)}}\varphi_Z(\omega') \ll \big|\mathrm{PO}_2(\mathfrak{q}_1)\big|\Big[\frac{Z^2}{\mathfrak{q}_1^3}+Z^{3/2}\Big]\\ &\ll \mathfrak{q}_1^\varepsilon\Big[\frac{Z^2}{\mathfrak{q}_1^2}+\mathfrak{q}_1Z^{3/2}\Big],\end{aligned}$$

since $|\mathrm{PO}_2(\mathfrak{q}_1)| \ll \mathfrak{q}_1^{1+\varepsilon}$. The contribution of this to (5.10) is then

$$\begin{aligned}&\ll T^\varepsilon|\Omega_X|\sum_{q\asymp Q}\sum_{\omega\in\Omega_Z}\sum_{\mathfrak{q}_1|q}X^2\frac{\mathfrak{q}_1}{q^2}\Big[\frac{Z^2}{\mathfrak{q}_1^2}+\mathfrak{q}_1Z^{3/2}\Big]\\ &\ll T^\varepsilon|\Omega_X|X^2|\Omega_Z|Z^2\Big[\frac{1}{Q}+\frac{Q}{Z^{1/2}}\Big].\end{aligned}$$

Combined with the first term of (5.10) and (3.2), this gives (5.4), as claimed. Theorem 5.1 follows immediately. □

6. Proof of the sieving theorem

We proceed now to prove Theorem 3.8. Combining (4.3) with (3.16) gives the decomposition (3.9). The content of (3.10) is, roughly, that $\beta(p)\sim 2/p$ on average; indeed, from (4.4) we have that

$$\beta(p)=\begin{cases}\frac{4}{p}+O(p^{-2}) & \text{if } p\equiv 1(4),\\ O(p^{-2}) & \text{if } p\equiv 3(4),\end{cases}$$

so (3.10) is elementarily verified. Combining (4.5) and (5.2) gives (3.11), as long as the following inequalities are satisfied:

$$C\alpha_0 < \Theta y, \qquad (6.1)$$

$$4\alpha + (1-\delta)(x+z) < x/4, \qquad (6.2)$$

$$(1-\delta)(x+z) < \alpha_0/2, \qquad (6.3)$$

$$\alpha/2 + (1-\delta)(x+z) < z/4. \qquad (6.4)$$

Here

$$Q_0 = N^{\alpha_0}, \qquad \mathcal{Q} = N^{\alpha}, \qquad X = N^{x}, \qquad Y = N^{y}, \qquad Z = N^{z}.$$

Remark 6.5
Treating $1-\delta$ as 0 and $x+z$ as 1, one quickly sees that the best one can do is the choice $\alpha \approx 1/18$, $x \approx 8/9$, $z \approx 1/9$.

Let $\eta > 0$ be given, and set

$$\alpha = 1/18 - \eta.$$

Since $\mathcal{Q} = N^{\alpha}$ and $N \asymp T^{1/4}$ (see (3.7)), this gives the exponent of distribution $1/72$ claimed in (3.12). Next we set

$$x = \frac{8}{9} - \eta,$$

and assume at first that $1-\delta < \eta$ (more stringent restrictions on δ will follow). Then since $x+z<1$, we have

$$x = 16\alpha + 15\eta > 16\alpha + 4\eta > 16\alpha + 4(1-\delta)(x+z).$$

That is, (6.2) is satisfied. Similarly, we set

$$z = \frac{1}{9} - \eta,$$

whence (6.4) holds once $1-\delta < \eta/4$. This means that $y = 2\eta$, so (6.1) is satisfied when

$$\alpha_0 = \frac{\Theta\eta}{C}.$$

Finally, for (6.3) to hold, we need

$$\delta > 1 - \frac{\Theta\eta}{2C}(1-2\eta)^{-1}.$$

Recalling that $\delta = \delta_{\mathcal{A}}$, this can be achieved (cf. (1.8)) by taking $\mathcal{A}$ sufficiently large.

Remark 6.6
It is here that we are crucially using that the parameters Θ and C coming from the "spectral gap" estimate (2.21) are independent of $\mathcal{A}$, and only depend on the fixed quantity $\mathcal{A}_0 = 2$; that is, they are absolute.

7. Proof of Theorem 1.5

7.1. Proof of Theorem 1.12

LEMMA 7.1
As $t \to \infty$,

$$\#\{\gamma \in \mathrm{SL}_2(\mathbb{Z}) : \mathrm{tr}(\gamma^\dagger \gamma) = t\} \ll t^{\varepsilon}. \tag{7.2}$$

Proof
One must count the number of $(a,b,c,d) \in \mathbb{Z}^4$ having $ad - bc = 1$ and $a^2 + b^2 + c^2 + d^2 = t$. Changing variables to $a = x + y$, $d = x - y$, $b = w + z$, and $c = w - z$ gives the equations $x^2 - y^2 + z^2 - w^2 = 1$ and $2(x^2 + y^2 + z^2 + w^2) = t$. That is, any solution to the former equations in integers gives one to the latter equations. It is elementary to see that there are at most t^{ε} solutions to the latter. □

In this section, write $\alpha = 1/350$, so that we can write (3.4) as

$$\Pi_{AP} = \{\gamma \in \Pi : p \mid (\mathrm{tr}(\gamma^\dagger\gamma)^2 - 4) \Longrightarrow p > N^{\alpha}\}.$$

Theorem 1.12 asks us to count

$$\begin{aligned} &\#\{\gamma \in \Gamma_{\mathcal{A}} \cap B_N : \mathrm{tr}(\gamma^\dagger\gamma)^2 - 4 \text{ is square-free}\} && (7.3)\\ &\quad \geq \#\{\gamma \in \Pi_{AP} : \mathrm{tr}(\gamma^\dagger\gamma)^2 - 4 \text{ is square-free}\} \\ &\quad > N^{2\delta - \eta} - \#\Pi_{AP}^{\square}, && (7.4)\end{aligned}$$

where we used (3.6) and defined

$$\Pi_{AP}^{\square} := \{\gamma \in \Pi_{AP} : \mathrm{tr}(\gamma^\dagger\gamma)^2 - 4 \text{ is not square-free}\}.$$

Now, for each $\gamma \in \Pi_{AP}^{\square}$, there is a prime p with $p^2 \mid (\mathrm{tr}(\gamma^\dagger\gamma)^2 - 4)$. Since $\gamma \in \Pi_{AP}$, we thus have that $p > N^{\alpha}$, and, moreover, p^2 divides either $\mathrm{tr}(\gamma^\dagger\gamma) + 2$ or $\mathrm{tr}(\gamma^\dagger\gamma) - 2$; in particular, $p \ll N$. Therefore, reversing orders and applying (7.2), we have

$$\begin{aligned} \#\Pi_{AP}^{\square} &\leq \sum_{N^{\alpha} < p \ll N} \sum_{\substack{t < N^2 \\ t^2 - 4 \equiv 0 (p^2)}} \#\{\gamma \in \Gamma_{\mathcal{A}} \cap B_N : \mathrm{tr}(\gamma^\dagger\gamma) = t\} \\ &\ll \sum_{N^{\alpha} < p \ll N} \frac{N^2}{p^2} N^{\varepsilon} \ll N^{2 - \alpha + \varepsilon}. \end{aligned}$$

Since $\alpha = 1/350$ is fixed, it is clear that, by making $\delta = \delta_{\mathcal{A}}$ sufficiently near 1 (by taking $\mathcal{A}$ large), one gets the desired main term from (7.4). This completes the proof of Theorem 1.12.

7.2. *Proof of Theorem 1.5*

Again, this will be an easy consequence of Theorem 1.12. Each $\gamma \in \Gamma_{\mathcal{A}} \cap B_N$ arising in (7.3) gives a hyperbolic conjugacy class $[\gamma^{\dagger}\gamma]$ of trace at most N^2, for which the corresponding geodesic is low-lying (relative to $\mathcal{A}$), reciprocal, and fundamental. The only two issues are (a) that the class need not be primitive, and (b) different γ's can give rise to the same geodesic. Since the word-length metric is commensurate with the logarithm of the archimedean metric, the number of imprimitive classes (i.e., the γ's, which, as symbols in the generators of $\Gamma_{\mathcal{A}}$, have a repeating sequence) is easily bounded by $N^{1+\varepsilon}$; these can safely be discarded from (7.3) without affecting the cardinality. The latter (b) happens when the symbols generating γ and γ', say, are the same up to a cyclic permutation. This adds at most $\log N$ to the multiplicity of (7.3) and can thus also be safely discarded. In summary, we have produced $N^{2-\eta}$ low-lying, fundamental, and reciprocal closed geodesics with trace bounded by N^2, as claimed. This completes the proof.

Acknowledgments. It is our pleasure to thank Nick Katz and Zeev Rudnick for stimulating discussions. The second-named author expresses his gratitude to the Institute for Advanced Study, where much of this work was written.

Bourgain's work was partially supported by National Science Foundation (NSF) grant DMS-1301619. Kontorovich's work was partially supported by NSF CAREER grants DMS-1254788 and DMS-1455705, NSF FRG grant DMS-1463940, an Alfred P. Sloan Research Fellowship, and a grant from the United States–Israel Binational Foundation.

References

[1] J. BOURGAIN, A. GAMBURD, and P. SARNAK, *Sieving and expanders*, C. R. Math. Acad. Sci. Paris **343** (2006), no. 3, 155–159. MR 2246331. DOI 10.1016/j.crma.2006.05.023. *(3413)*

[2] ———, *Affine linear sieve, expanders, and sum-product*, Invent. Math. **179** (2010), no. 3, 559–644. MR 2587341. DOI 10.1007/s00222-009-0225-3. *(3413)*

[3] ———, *Generalization of Selberg's 3/16th theorem and affine sieve*, Acta Math **207** (2011), no. 2, 255–290. MR 2892611. DOI 10.1007/s11511-012-0070-x. *(3422)*

[4] ———, *Markoff triples and strong approximation*, C. R. Math. Acad. Sci. Paris **354** (2016), no. 2, 131–135. MR 3456887. DOI 10.1016/j.crma.2015.12.006. *(3413)*

[5] ———, *Markoff surfaces and strong approximation*, preprint, arXiv:1607.01530. *(3413)*

[6] J. BOURGAIN and A. KONTOROVICH, *The affine sieve beyond expansion, I: Thin hypotenuses*, Int. Math. Res. Not. IMRN **2015**, no. 19, 9175–9205. MR 3431590. DOI 10.1093/imrn/rnu222. *(3416)*

[7] ———, *Beyond expansion, II: Low-lying fundamental geodesics*, J. Eur. Math. Soc. (JEMS) **19** (2017), no. 5, 1331–1359. MR 3635355. DOI 10.4171/JEMS/694. *(3413, 3415, 3416, 3421, 3422, 3426)*

[8] J. BOURGAIN, A. KONTOROVICH, and M. MAGEE, *Thermodynamic expansion to arbitrary moduli*, appendix to [16]. *(3422)*

[9] W. DUKE, *Hyperbolic distribution problems and half-integral weight Maass forms*, Invent. Math. **92** (1988), no. 1, 73–90. MR 0931205. DOI 10.1007/BF01393993. *(3413)*

[10] M. EINSIEDLER, E. LINDENSTRAUSS, P. MICHEL, and A. VENKATESH, *Distribution of periodic torus orbits on homogeneous spaces*, Duke Math. J. **148** (2009), no. 1, 119–174. MR 2515103. DOI 10.1215/00127094-2009-023. *(3413)*

[11] J. FRIEDLANDER and H. IWANIEC, *Opera de Cribro*, Amer. Math. Soc. Colloq. Publ. **57**, Amer. Math. Soc., Providence, 2010. MR 2647984. DOI 10.1090/coll/057. *(3424)*

[12] D. HENSLEY, "The distribution of badly approximable numbers and continuants with bounded digits" in *Théorie des nombres (Quebec, PQ, 1987)*, de Gruyter, Berlin, 1989, 371–385. MR 1024576. *(3415)*

[13] ———, *Continued fraction Cantor sets, Hausdorff dimension, and functional analysis*, J. Number Theory **40** (1992), no. 3, 336–358. MR 1154044. DOI 10.1016/0022-314X(92)90006-B. *(3415)*

[14] A. KONTOROVICH, *Levels of distribution and the affine sieve*, Ann. Fac. Sci. Toulouse Math. (6) **23** (2014), no. 5, 933–966. MR 3294598. DOI 10.5802/afst.1432. *(3416)*

[15] ———, "Applications of thin orbits" in *Dynamics and Analytic Number Theory*, London Math. Soc. Lecture Note Ser. **437**, Cambridge Univ. Press, Cambridge, 2016, 289–317. MR 3618792. *(3413, 3414, 3416)*

[16] M. MAGEE, H. OH, and D. WINTER, *Uniform congruence counting for Schottky semigroups in* $\mathrm{SL}_2(\mathbf{Z})$, with an appendix by J. Bourgain, A. Kontorovich, and M. Magee, J. Reine Angew. Math. **753** (2019), 89–135. MR 3987865. DOI 10.1515/crelle-2016-0072. *(3422, 3434)*

[17] P. SARNAK, "Reciprocal geodesics" in *Analytic Number Theory*, Clay Math. Proc. **7**, Amer. Math. Soc., Providence, 2007, 217–237. MR 2362203. *(3413, 3415, 3416)*

[18] ———, *Affine sieve lecture slides*, 2010, http://publications.ias.edu/sites/default/files/Affine%20sieve%20summer%202010_0.pdf. *(3413)*

[19] A. SALEHI GOLSEFIDY and P. SARNAK, *The affine sieve*, J. Amer. Math. Soc. **26** (2013), no. 4, 1085–1105. MR 3073885. DOI 10.1090/S0894-0347-2013-00764-X. *(3413)*

[20] A. SELBERG, *On the estimation of Fourier coefficients of modular forms*, Proc. Sympos. Pure Math. **VII** (1965), 1–15. MR 0182610. *(3421)*

Bourgain
School of Mathematics, Institute for Advance Study, Princeton, New Jersey, USA

Kontorovich
Department of Mathematics, Rutgers University, New Brunswick, New Jersey, USA;
alex.kontorovich@rutgers.edu