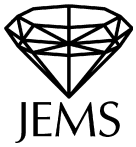
J. Eur. Math. Soc. 19, 1331–1359

DOI 10.4171/JEMS/694



Jean Bourgain · Alex Kontorovich

Beyond expansion II: low-lying fundamental geodesics

Received June 25, 2014

Abstract. A closed geodesic on the modular surface is "low-lying" if it does not travel "high" into the cusp. It is "fundamental" if it corresponds to an element in the class group of a real quadratic field. We prove the existence of infinitely many low-lying fundamental geodesics, answering a question of Einsiedler–Lindenstrauss–Michel–Venkatesh.

Keywords. Affine sieve, closed geodesics, fundamental discriminants, equidistribution

Contents

1. Introduction

In this paper, we answer a question of Einsiedler–Lindenstrauss–Michel–Venkatesh on the abundance of "low-lying" closed geodesics on the modular surface which are "fundamental" (see the definitions below). The main difficulty is to produce a strong "level of distribution" for a particular set coming from a "thin orbit."

J. Bourgain: School of Mathematics, Institute of Advanced Studies, Princeton, NJ 08540, USA; e-mail: bourgain@ias.edu

A. Kontorovich: Department of Mathematics, Yale University, New Haven, CT 06511, USA; current address: Department of Mathematics, Rutgers University, New Brunswick, NJ 08854, USA; e-mail: alex.kontorovich@yale.edu

Mathematics Subject Classification (2010): 11J70, 11N36, 37A45

1.1. Statement of the main theorem

Let $D > 0$ be a *fundamental discriminant*, that is, the discriminant of a real quadratic field $K_D = \mathbb{Q}(\sqrt{D})$, and let $\mathscr{C}_D$ be the class group of K_D, with class number $h_D = |\mathscr{C}_D|$. To each class $\gamma \in \mathscr{C}_D$, we associate in the standard way a closed geodesic (by abuse of notation also called γ) in the unit tangent bundle

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

of the modular surface. Not every closed geodesic on $\mathcal{X}$ corresponds to an element of the class group of a real quadratic field; we call those that do *fundamental*. The following rank-one question arose around 2004 in the work of Einsiedler–Lindenstrauss–Michel–Venkatesh on higher rank analogues of Duke's Theorem (see [ELMV09, §1.5] and the discussion below).

Question 1.1. *Does there exist a compact subset $\mathcal{Y} \subset \mathcal{X}$ which contains infinitely many fundamental geodesics?*

Geodesics confined to a compact region obviously never enter "high" in the cusp, and hence cannot equidistribute in $\mathcal{X}$; we refer to these as *low-lying* (in $\mathcal{Y}$). A natural set of candidate such, as observed by Sarnak, are the geodesics coming from Markov triples (see [Sar07, pp. 226, 234]), the difficulty being to understand when these are fundamental. (In fact, this very question initiated the study of the Affine Sieve [BGS06, BGS10, SGS13].) While we are unable to show the infinitude of fundamental Markov geodesics (which, if they exist, are extremely rare [Zag82]), our main goal (see Theorem 1.8 below) is to give an affirmative answer to Question 1.1, in a strong quantitative sense.

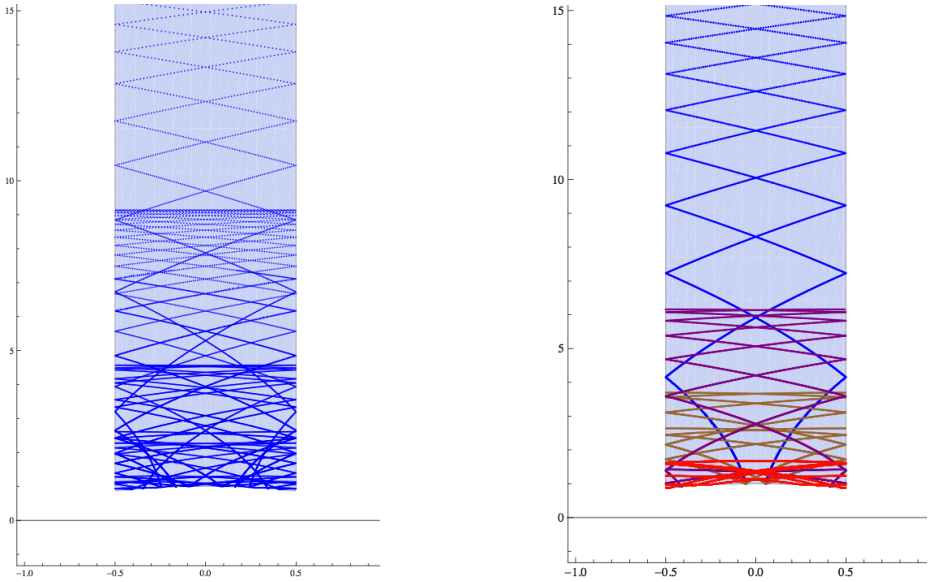
Before stating our result, we put Question 1.1 in perspective, by first recalling Duke's equidistribution theorem. Let $\mu_{\mathcal{X}}$ be the probability Haar measure on $\mathcal{X}$, and associate to each class $\gamma \in \mathscr{C}_D$ (or rather, the corresponding geodesic) the probability arc-length measure μ_γ. Then Duke's theorem [Duk88] asserts the equidistribution of μ_γ's to $\mu_{\mathcal{X}}$ *on average* over $\mathscr{C}_D$, for large discriminant:

$$\frac{1}{h_D} \sum_{\gamma \in \mathscr{C}_D} \mu_\gamma \xrightarrow{\text{weak}^*} \mu_{\mathcal{X}} \quad \text{as } D \to \infty. \tag{1.2}$$

The goal of asking Question 1.1 is to try to understand to what extent it is necessary to average over $\mathscr{C}_D$ in (1.2), or whether perhaps the equidistribution already happens at the level of individual closed geodesic orbits (as is expected in higher rank analogues from rigidity phenomena conjectured by Cassels/Swinnerton-Dyer, Furstenberg, Margulis, etc.). This question turns out to be quite subtle, as we indicate below.

It will be instructive to keep in mind the two examples illustrated in Figure 1. First recall some basic notions. Write $[A, B, C]$ for the binary quadratic form $Ax^2 + Bxy + Cy^2$, and $\{[A, B, C]\}$ for the corresponding class. The discriminant

$$D = B^2 - 4AC, \tag{1.3}$$



(a) $D = 1337$, $h_D = 1$
■ $= \{[19, 27, -8]\}$

(b) $D = 1365$, $h_D = 4$
■ $= \{[35, 35, -1]\}$, ■ $= \{[7, 35, -5]\}$,
■ $= \{[23, 33, -3]\}$, ■ $= \{[19, 23, -11]\}$

Fig. 1. Fundamental geodesics in $\mathscr{C}_D$

assumed throughout to be positive, is fundamental if either D is square-free (in which case $D \equiv 1 \pmod 4$), or $D \equiv 0 \pmod 4$, in which case $D/4$ is square-free and $D/4 \equiv 2, 3 \pmod 4$. To associate a closed geodesic to a class $\gamma = \{[A, B, C]\}$, connect the two real Galois-conjugate roots

$$\alpha, \bar{\alpha} = \frac{-B \pm \sqrt{D}}{2A} \tag{1.4}$$

by a geodesic in $T^1\mathbb{H}$, and project modulo $\mathrm{PSL}_2(\mathbb{Z})$. We first consider the situation in

Example I ($D = 1337 = 7 \times 191$, Figure 1a). While infinitely many closed geodesics are defined over $K = \mathbb{Q}(\sqrt{1337})$, only one of them is fundamental, because K has class number one, $h_{1337} = 1$. This one fundamental geodesic, corresponding to the class $\{[19, 27, -8]\}$, is not particularly "low-lying" (of course that depends on one's choice of the compact region $\mathcal{Y}$), illustrating the difficulty of Question 1.1.

In fact, whenever the field K_D has class number one (as conjecturally happens infinitely often), there is obviously no averaging in Duke's theorem (1.2), and the one geodesic in the class is individually becoming equidistributed. Moreover, Popa's refinement [Pop06] of Waldspurger's theorem, together with a subconvex bound for certain Rankin–Selberg L-functions (see [HM06]), implies that, as long as the class number is not too large,

$$h_D < D^\eta \quad \text{for some small } \eta > 0,$$

then *every* geodesic in $\mathscr{C}_D$ is *individually* becoming equidistributed, that is, without averaging over $\mathscr{C}_D$ as in (1.2). Assuming the Lindelöf hypothesis (which is a consequence of GRH) for such L-functions, Popa's work implies that the exponent η can be taken as large as $1/4-\varepsilon$. Meanwhile, it is widely believed that the same individual equidistribution holds with η as large as $1/2-\varepsilon$, for any fixed $\varepsilon > 0$. So to even have a chance of seeing any non-equidistributing behavior (as in Theorem 1.8), one must take the class number almost as large as possible,

$$h_D > D^{1/2-o(1)}. \tag{1.5}$$

On the other hand, such discriminants should be quite rare. Indeed, it is a longstanding open problem that the average class number satisfies, crudely,

$$\sum_{\substack{0<D<T\\ \text{fundamental}}} h_D \stackrel{?}{=} T^{1+o(1)} \quad (T\to\infty).$$

If true, this estimate and the above heuristic would imply that there can only be very few discriminants with such large class number,

$$\#\{0 < D \text{ fundamental} < T : h_D > D^{1/2-o(1)}\} \stackrel{?}{<} T^{1/2+o(1)}. \tag{1.6}$$

Despite this rarity, there does exist a standard way of making large class numbers, namely, by considering discriminants of the special form

$$D = t^2 - 4. \tag{1.7}$$

Then the fundamental solution to the Pellian equation

$$T^2 - DS^2 = 4$$

is $(T, S) = (t, 1)$, whence the fundamental unit ϵ_D is as small as possible,

$$\epsilon_D = \frac{t+\sqrt{D}}{2} \asymp \sqrt{D}.$$

Dirichlet's Class Number Formula and Siegel's (ineffective) Theorem then give, crudely,

$$h_D = \sqrt{D}\,\frac{L(1,\chi_D)}{\log \epsilon_D} > D^{1/2-o(1)}.$$

Not surprisingly, we will be looking for low-lying (and hence non-equidistributing) behavior among fundamental discriminants of the special form (1.7). (And since there are about $\sqrt{T}$ such up to T, we are not losing too much from (1.6).) This brings us to:

Example II ($D = 1365 = 37^2 - 4 = 3\times 5\times 7\times 13$, Figure 1b). The class number is $h_{1365} = 4$, and the behavior of the four fundamental geodesics defined over $\mathbb{Q}(\sqrt{1365})$ varies dramatically. The identity element of the class group $\mathscr{C}_{1365}$ is the class $\{[35, 35, -1]\}$, and the corresponding geodesic shoots high up into the cusp; meanwhile the geodesic corresponding to $\{[19, 23, -11]\}$ is very low-lying, not reaching above

$\mathfrak{Im}\, z = 2$. Nevertheless, the four geodesics taken *together* equidistribute about as well as the one geodesic in Example I, beautifully illustrating why one must *average* over $\mathscr{C}_D$ for Duke's theorem (1.2) to hold.

Returning to Question 1.1, we may now state our main result.

Theorem 1.8. *There exist infinitely many low-lying fundamental geodesics. More precisely, for each $\epsilon > 0$, there is a compact region $\mathcal{Y} = \mathcal{Y}(\epsilon) \subset \mathcal{X}$, and a set $\mathscr{D} = \mathscr{D}(\epsilon)$ of positive fundamental discriminants, such that:*

(1) *for each $D \in \mathscr{D}$, many of the geodesics in the corresponding class group are low-lying:*

$$\#\{\gamma \in \mathscr{C}_D : \gamma \subset \mathcal{Y}\} > |\mathscr{C}_D|^{1-\epsilon}, \tag{1.9}$$

(2) *compared to (1.6), there are many discriminants in $\mathscr{D}$:*

$$\#(\mathscr{D} \cap [1, T]) > T^{1/2-\epsilon} \quad (T \to \infty). \tag{1.10}$$

There are (at least) two ways to interpret this result. One can let $\epsilon \to 0$, so that the inequalities (1.9)–(1.10) give more and more "low-lying" fundamental geodesics; unfortunately the compact region $\mathcal{Y}(\epsilon)$ will then approach $\mathcal{X}$, giving less and less meaning to "low-lying." Alternatively, one can let ϵ be a fixed constant, say, $\epsilon = 1/100$; then $\mathcal{Y}$ is a fixed region containing infinitely many fundamental geodesics, giving an affirmative answer to Question 1.1.

Again, in light of (1.6), the estimate (1.10) is almost sharp. In the Appendix, we show (by more-or-less standard ergodic-theoretic techniques, combining Duke's theorem and mixing) that (1.9) is also essentially sharp, in the following sense.

Theorem 1.11. *For any compact region $\mathcal{Y} \subset \mathcal{X}$, there is an $\epsilon = \epsilon(\mathcal{Y}) > 0$ such that*

$$\#\{\gamma \in \mathscr{C}_D : \gamma \subset \mathcal{Y}\} < |\mathscr{C}_D|^{1-\epsilon}$$

as $D \to \infty$ through all non-square integers.

1.2. Ingredients

We now describe some of the tools going into the proof of Theorem 1.8, beginning with a series of reformulations.

1.2.1. Step 1: Convert to continued fractions. By the well-known connection [Hum16, Art24, Ser85] between continued fractions and the cutting sequence of the geodesic flow on $\mathcal{X}$, the condition that a geodesic γ be low-lying can be reformulated as a Diophantine property of the corresponding visual point α in (1.4), as follows. Given any $\alpha \in \mathbb{R}$, write its continued fraction expansion as

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots}} = [a_0, a_1, a_2, \dots],$$

where $a_0 \in \mathbb{Z}$ and $a_j \in \mathbb{Z}_{\geq 1}$ for $j \geq 1$; the numbers a_j are called the *partial quotients* of α. When α is a visual point of a closed geodesic, it is a quadratic irrational, and hence has an eventually periodic continued fraction expansion. By applying a $\mathrm{PSL}_2(\mathbb{Z})$ action, we may assume that α is *reduced*, meaning that $-1 < \bar{\alpha} < 0 < 1 < \alpha$; then the continued fraction expansion of α is exactly (as opposed to eventually) periodic.

Cutting off the cusp of $\mathcal{X}$ at some height $\mathcal{C} < \infty$ leaves a compact region $\mathcal{Y} = \mathcal{X} \cap \{\Im\mathfrak{m}\, z \leq \mathcal{C}\}$, and the condition that a geodesic γ is low-lying (in $\mathcal{Y}$) is essentially equivalent to its visual point α having all partial quotients bounded by some $\mathcal{A} = \mathcal{A}(\mathcal{C}) < \infty$.

To illustrate this fact, we return for a moment to Figure 1. In Example I, the one fundamental geodesic in the class $\mathscr{C}_{1337}$ corresponds to a reduced visual point α having the continued fraction expansion

$$\{[19, 27, -8]\} \leftrightsquigarrow [\overline{1, 1, 2, 17, 1, 8, 5, 8, 1, 17, 2, 1, 1, 3, 1, 35, 1, 3}].$$

The large partial quotient 35 is responsible for the high excursion of the geodesic in Figure 1a.

Meanwhile, the four geodesics in Example II correspond to the continued fraction expansions

$$\begin{aligned}
\{[35, 35, -1]\} &\leftrightsquigarrow [\overline{1, 35}],\\
\{[7, 35, -5]\} &\leftrightsquigarrow [\overline{5, 7}],\\
\{[23, 33, -3]\} &\leftrightsquigarrow [\overline{1, 1, 1, 11}],\\
\{[19, 23, -11]\} &\leftrightsquigarrow [\overline{1, 1, 1, 2, 1, 2}],
\end{aligned}$$

and the very small partial quotients of the last of these explain the corresponding low-lying geodesic in Figure 1b.

To ensure that a fundamental geodesic is low-lying, one can try to force its visual point to have all partial quotients bounded by some height $\mathcal{A} < \infty$. Alternatively, one can first consider all reduced quadratic irrationals with partial quotients bounded by $\mathcal{A}$, and try to understand when these come from fundamental geodesics. We will take the latter approach, which turns out to be a sieving problem on a certain "thin orbit."

1.2.2. Step 2: Convert to thin orbits. It is elementary that the matrix

$$\gamma = \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix}\cdots\begin{pmatrix} a_\ell & 1 \\ 1 & 0 \end{pmatrix} \tag{1.12}$$

fixes the quadratic irrational

$$\alpha = [\overline{a_0, a_1, \ldots, a_\ell}], \tag{1.13}$$

thus converting questions on continued fractions into ones about matrix products. In particular, we will be interested in the traces of matrices of the form (1.12), in light of

Lemma 1.14. *A sufficient condition for a closed geodesic $[\gamma]$ to be fundamental is that*

$$\mathrm{tr}(\gamma)^2 - 4 \text{ is square-free.} \tag{1.15}$$

Note that the corresponding discriminant $D = \mathrm{tr}(\gamma)^2 - 4$ is then fundamental and of the special form (1.7). Here is a quick proof: The fixed points of $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ are easily seen to be

$$\alpha, \bar{\alpha} = \frac{a - d \pm \sqrt{\mathrm{tr}(\gamma)^2 - 4}}{2c}. \tag{1.16}$$

Now assume that $D := \mathrm{tr}(\gamma)^2 - 4$ is square-free. Comparing (1.16) with (1.4), we set $B := d - a$ and $A := c$. Solving (1.3) for C gives $C = -b$. Then the equivalence class of the form $Q = [A, B, C]$ has fundamental discriminant D, and hence corresponds to the fundamental geodesic $[\gamma]$, as desired.

Thus to study the traces of matrix products of the form (1.12) with all $a_j \le \mathcal{A}$, we should introduce the semigroup of finite products of such matrices,[1]

$$\mathcal{G}_{\mathcal{A}} := \left\langle \begin{pmatrix} a & 1 \\ 1 & 0 \end{pmatrix} : a \le \mathcal{A} \right\rangle^{+} \subset \mathrm{GL}_2(\mathbb{Z}). \tag{1.17}$$

Preferring to work in SL_2, we immediately pass to the even-length (determinant-one) subsemigroup

$$\Gamma_{\mathcal{A}} := \mathcal{G}_{\mathcal{A}} \cap \mathrm{SL}_2(\mathbb{Z}), \tag{1.18}$$

which is (finitely) generated by the products $\left(\begin{smallmatrix} a & 1 \\ 1 & 0 \end{smallmatrix}\right) \cdot \left(\begin{smallmatrix} b & 1 \\ 1 & 0 \end{smallmatrix}\right)$ for $a, b \le \mathcal{A}$.

The reason we call $\Gamma_{\mathcal{A}}$ "thin" is the following. Let N be a growing parameter, and let

$$B_N \subset \mathrm{SL}_2(\mathbb{R}) \tag{1.19}$$

be a ball about the origin of size N in the Frobenius norm

$$\|g\|^2 = \mathrm{tr}({}^t g g).$$

A theorem of Hensley [Hen89] states that

$$\#(\Gamma_{\mathcal{A}} \cap B_N) \asymp_{\mathcal{A}} N^{2\delta_{\mathcal{A}}} \quad (N \to \infty), \tag{1.20}$$

where $\delta_{\mathcal{A}}$ is the Hausdorff dimension of the Cantor-like fractal of all numbers with partial quotients bounded by $\mathcal{A}$:

$$\delta_{\mathcal{A}} := \mathrm{H.dim}\,\{[a_0, a_1, a_2, \dots] : a_j \le \mathcal{A}\} \in (0, 1). \tag{1.21}$$

This dimension has been studied extensively, and it is known [Hen92] that it can be made arbitrarily close to 1 by taking $\mathcal{A}$ large,

$$\delta_{\mathcal{A}} = 1 - \frac{6}{\pi^2 \mathcal{A}} + o\left(\frac{1}{\mathcal{A}}\right) \quad (\mathcal{A} \to \infty). \tag{1.22}$$

On the other hand, the set of $\mathbb{Z}$-points in the Zariski closure of $\Gamma_{\mathcal{A}}$ is just $\mathrm{SL}_2(\mathbb{Z})$, and it is classical that $\#(\mathrm{SL}_2(\mathbb{Z}) \cap B_N) \asymp N^2$, instead of the much "thinner" count $N^{2\delta_{\mathcal{A}}}$ as in (1.20).

In light of Lemma 1.14 and (1.20), the main Theorem 1.8 will follow without much effort from

[1] The superscript $+$ in (1.17) denotes generation as a semigroup, that is, no inverses.

Theorem 1.23. *Many elements in $\Gamma_{\mathcal{A}}$ have traces satisfying (1.15). More precisely, for any $\eta > 0$, there is an $\mathcal{A} = \mathcal{A}(\eta) < \infty$ such that*

$$\#\{\gamma \in \Gamma_{\mathcal{A}} \cap B_N : \mathrm{tr}(\gamma)^2 - 4 \text{ is square-free}\} > N^{2\delta_{\mathcal{A}} - \eta} \quad (N \to \infty). \tag{1.24}$$

The problem is thus reduced to

1.2.3. Step 3: Try to execute a sieve. This subsection is purely expository and heuristic; we will give a rough discussion of the sieving procedure, deferring the precise (and somewhat technical) statements to §3.

To sieve for square-free values of $\mathrm{tr}(\gamma)^2 - 4$, we need to understand their distribution modulo $\mathfrak{q}$, as γ ranges roughly in $\Gamma_{\mathcal{A}} \cap B_N$, taking $\mathfrak{q}$ as large as possible relative to N. Since $\mathrm{tr}(\gamma)^2 - 4$ is of size N^2 when γ is of size N, we introduce the parameter

$$T = N^2. \tag{1.25}$$

Letting $\beta(\mathfrak{q})$ be the proportion of matrices in $\mathrm{SL}_2(\mathfrak{q})$ for which $\mathrm{tr}(\gamma)^2 - 4 \equiv 0 \pmod{\mathfrak{q}}$, one might expect that

$$r_{\mathfrak{q}}(T) := \sum_{\substack{\gamma \in \Gamma_{\mathcal{A}} \cap B_N \\ \mathrm{tr}(\gamma)^2 - 4 \equiv 0 \,(\mathfrak{q})}} 1 - \beta(\mathfrak{q}) \sum_{\gamma \in \Gamma_{\mathcal{A}} \cap B_N} 1$$

is a "remainder" term, which should be "small" in the following sense. We would like that for some large $\mathcal{Q}$, these remainders summed up to $\mathcal{Q}$ still do not exceed the total size,

$$\sum_{\mathfrak{q} < \mathcal{Q}} |r_{\mathfrak{q}}(N)| = o(\#(\Gamma_{\mathcal{A}} \cap B_N)). \tag{1.26}$$

If this can be rigorously established, then we call $\mathcal{Q}$ a *level of distribution* (for $\mathcal{A}$). Note that this is not a quantity intrinsic to our problem, but rather a function of what one can prove. The larger this quantity, the more control one has on the distribution of the set of traces on such arithmetic progressions. If moreover (1.26) can be confirmed with $\mathcal{Q}$ as large as a power of the parameter T,

$$\mathcal{Q} = T^{\alpha}, \quad \alpha > 0, \tag{1.27}$$

then we say that α is an *exponent of distribution* for $\mathcal{A}$.

The by-now "standard" Affine Sieve procedure applies in this context, and produces *some* (weak) exponent of distribution $\alpha > 0$. We briefly sketch the method now, before explaining why it is insufficient in our context. A theorem of Bourgain–Gamburd–Sarnak [BGS11] says very roughly (see Theorem 2.2 for a precise statement) that we do have equidistribution in $\Gamma_{\mathcal{A}}$ mod $\mathfrak{q}$, in the sense that there are constants

$$\Theta > 0 \tag{1.28}$$

and $C < \infty$ such that, for all $\mathfrak{q} \geq 1$ and all $\gamma_0 \in \mathrm{SL}_2(\mathfrak{q})$,

$$\left| \sum_{\substack{\gamma \in \Gamma_{\mathcal{A}} \cap B_N \\ \gamma \equiv \gamma_0 \,(\mathfrak{q})}} 1 - \frac{1}{|\mathrm{SL}_2(\mathfrak{q})|} \sum_{\gamma \in \Gamma_{\mathcal{A}} \cap B_N} 1 \right| \text{“} \ll \text{”} \, \mathfrak{q}^C N^{2\delta - \Theta}, \tag{1.29}$$

where the implied constant is independent of γ_0 and $\mathfrak{q}$. (We reiterate that the error in (1.29) is heuristic only; a statement of this strength is not currently known. That said, the true statement serves the same purpose in our application.) The positivity of Θ in (1.28) is called the "spectral gap" or "expander" property of $\Gamma_{\mathcal{A}}$, and follows from a resonance-free region for the resolvent of a certain "congruence" transfer operator (see §2.1). Summing (1.29) over those $\gamma_0 \in \mathrm{SL}_2(\mathfrak{q})$ with $\mathrm{tr}(\gamma_0)^2 - 4 \equiv 0 \pmod{\mathfrak{q}}$, and then over $\mathfrak{q}$ up to $\mathcal{Q}$, one proves (1.26) with $\mathcal{Q} = N^{\alpha}$ and exponent of distribution

$$\alpha = \Theta/C - \varepsilon \tag{1.30}$$

for any $\varepsilon > 0$. (The value of C may change from line to line.)

It turns out that this standard procedure is just shy of giving our main result! To successfully execute the sieve (that is, convert (1.26) into Theorem 1.23), one needs the exponent of distribution to be strong enough to overcome the thinness of $\Gamma_{\mathcal{A}}$, in the sense that we need something like

$$\alpha > 10(1 - \delta_{\mathcal{A}}) \tag{1.31}$$

(see §7.1). The term on the right can be made arbitrarily small (cf. (1.22)), so it seems that by taking $\mathcal{A}$ large enough, we should establish (1.31). Unfortunately, the spectral gap Θ in (1.28) coming from the proof in [BGS11] is *a priori* a function of $\mathcal{A}$, and it is an extremely important open problem to understand its behavior with respect to $\mathcal{A}$. Presumably Θ should not deteriorate to zero as $\mathcal{A}$ increases, but present methods are insufficient to show this, rendering the exponent (1.30) useless towards (1.31). Of course one can try to directly follow the proof in [BGS11], but then the $\mathcal{A}$ dependence will be abysmal, and insufficient relative to (1.22) to produce the required inequality (1.31).

Rather than attacking this difficult problem head-on, we circumvent it as follows.

1.2.4. Step 4: Prove an exponent of distribution beyond expansion. Instead of controlling the remainders (1.26) using only expansion (1.29), we seek to go beyond the direct procedure of the Affine Sieve, producing a stronger exponent of distribution to ensure that (1.31) is satisfied. We employ two novel techniques here, which appear in some form already in [BK10, BK11, BK14a, BK14b, BK15]. The first is to take inspiration from Vinogradov's method, replacing the full archimedean ball $\Gamma_{\mathcal{A}} \cap B_N$ by a product of several such, which more readily captures the semigroup structure, and moreover allows development of techniques from estimating "bilinear" forms. The second innovation is, for larger values of $\mathfrak{q}$, to capture the condition $\mathrm{tr}(\gamma)^2 - 4 \equiv 0 \pmod{\mathfrak{q}}$ by *abelian* harmonic analysis, rather than the "spectral" method in (1.29). One then faces various exponential sums over our thin semigroup $\Gamma_{\mathcal{A}}$, but after some applications of Cauchy–Schwarz, one uses positivity to replace $\Gamma_{\mathcal{A}}$ by the full ambient group $\mathrm{SL}_2(\mathbb{Z})$, allowing employment of more classical tools. This loss is acceptable as long as the dimension $\delta_{\mathcal{A}}$ of $\Gamma_{\mathcal{A}}$ is sufficiently near 1, that is, as long as $\mathcal{A}$ is large enough. In the end, we are able to produce the strong exponent of distribution (for $\mathcal{A}$ large) of

$$\alpha = 1/34. \tag{1.32}$$

In fact, our methods prove the exponent $\alpha = 1/32 - \varepsilon$ (and further refinements would give

$\alpha = 1/8 - \varepsilon$), but (1.32) is already more than sufficient for (1.31); for ease of exposition, we will not strive for optimal exponents.

Inserting the strong exponent of distribution in (1.32) into (1.31), we are able to sieve down to square-free values of $\mathrm{tr}(\gamma)^2 - 4$, thus establishing Theorem 1.23, from which Theorem 1.8 follows easily.

The proof of Theorem 1.11 is by completely different methods, namely, a combination of mixing, Duke's theorem, and standard tools in ergodic theory.

1.3. Organization

In §2, we collect some preliminaries needed in the sieve analysis. We spend §3 constructing the main "bilinear" set $\Pi \subset \Gamma_{\mathcal{A}}$ used for sieving, and stating the main sieving theorems. The main term is analyzed in §4, while the errors are handled in §5. We prove the main sieving theorem (see Theorem 3.15) in §6, and derive Theorem 1.8 in §7. Finally, the appendix proves Theorem 1.11.

1.4. Notation

We use the following notation throughout. Set $e(x) = e^{2\pi i x}$ and $e_q(x) = e(x/q)$. We use $f \sim g$ to mean $f/g \to 1$. The symbols $f \ll g$ and $f = O(g)$ are used interchangeably to mean the existence of an implied constant $C > 0$ such that $f(x) \leq Cg(x)$ for all $x > C$; moreover $f \asymp g$ means $f \ll g \ll f$. The letters c, C denote positive constants, not necessarily the same at each occurrence. Unless otherwise specified, implied constants may depend at most on $\mathcal{A}$, which is treated as fixed. The letter $\varepsilon > 0$ is an arbitrarily small constant, not necessarily the same at each occurrence. When it appears in an inequality, the implied constant may also depend on ε without further specification. The symbol $\mathbf{1}_{\{\cdot\}}$ is the indicator function of the event $\{\cdot\}$. The trace of a matrix γ is denoted $\mathrm{tr}(\gamma)$. The number of divisors of n is denoted $\tau(n)$. The greatest common divisor of n and m is written (n, m) and their least common multiple is $[n, m]$. The symbol $\nu(n)$ denotes the number of distinct prime factors of n. The cardinality of a finite set S is denoted $|S|$ or $\#S$. The transpose of a matrix g is written ${}^t g$. When there can be no confusion, we use the shorthand $a \equiv b\ (q)$ for $a \equiv b \pmod q$. The prime symbol $'$ in $\sum'_{r \bmod q}$ means the range of $r \bmod q$ is restricted to $(r, q) = 1$.

2. Preliminaries

In this section, we state two results that are needed later, namely Propositions 2.9 and 2.17. We recommend the reader to skip the proofs on the first pass, instead proceeding directly to §3.

2.1. Expansion

In this subsection, we make precise the "expansion" theorem heuristically stated in (1.29). We will only require expansion for the fixed value $\mathcal{A}_0 = 2$, so as to make the expan-

sion constants absolute, and not dependent on $\mathcal{A}$ (see the discussion after (1.31) and Remark 6.8).

To this end, let $\Gamma_2 \subset \mathrm{SL}_2(\mathbb{Z})$ be the semigroup as in (1.18) corresponding to $\mathcal{A}_0 := 2$. It is easy to see that Γ_2 is free, that every non-identity matrix $\gamma \in \Gamma_2$ is hyperbolic, and that

$$\operatorname{tr}\gamma \asymp \|\gamma\|.$$

Note also that the *group* $\langle\Gamma_2\rangle$ generated by the semigroup Γ_2 is all of $\mathrm{SL}_2(\mathbb{Z})$. This immediately implies that for any $q \geq 1$, the mod q reduction of Γ_2 is everything,

$$\Gamma_2 \bmod q \cong \mathrm{SL}_2(q). \tag{2.1}$$

The following theorem is a consequence of the general expansion theorem proved by Bourgain–Gamburd–Sarnak [BGS11].

Theorem 2.2 ([BGS11]). *Let Γ_2 be the semigroup above. There exists an absolute square-free integer*

$$\mathfrak{B} \geq 1, \tag{2.3}$$

absolute constants $c, C > 0$, and an absolute "spectral gap"

$$\Theta = \Theta(\mathcal{A}_0) > 0, \tag{2.4}$$

such that, for any square-free $q \equiv 0\ (\mathfrak{B})$ and any $\omega \in \mathrm{SL}_2(q)$, as $Y \to \infty$, we have

$$\begin{aligned}\#\{\gamma \in \Gamma_2 \cap B_Y : \gamma \equiv \omega\ (q))\} = \frac{|\mathrm{SL}_2(\mathfrak{B})|}{|\mathrm{SL}_2(q)|}|\{\gamma \in \Gamma_2 \cap B_Y : \gamma \equiv \omega\ (\mathfrak{B})\}| \\ + O(\#\{\gamma \in \Gamma_2 : \|\gamma\| < Y\} \cdot \mathfrak{E}(Y; q)),\end{aligned} \tag{2.5}$$

where

$$\mathfrak{E}(Y; q) := \begin{cases} e^{-c\sqrt{\log Y}} & \text{if } q \leq C \log Y, \\ q^C Y^{-\Theta} & \text{if } q > C \log Y. \end{cases} \tag{2.6}$$

Remark 2.7. This theorem is proved in [BGS11, Theorem 1.5] under the assumption that, instead of Γ_2, one is given a convex-cocompact sub*group* of $\mathrm{SL}_2(\mathbb{Z})$. But the proof is the same when the group is replaced by our free semigroup Γ_2; we emphasize again that Γ_2 has no parabolic elements. The error term (2.6) is consequence of a Tauberian argument applied to a resonance-free region [BGS11, Theorem 9.1] of the form

$$\sigma > \delta_{\mathcal{A}_0} - C \min\left\{1, \frac{\log q}{\log(1 + |t|)}\right\}, \qquad \sigma + it \in \mathbb{C}, \tag{2.8}$$

for the resolvent of a certain "congruence" transfer operator (see [BGS11, §12] for details). For small q, we only obtain a "Prime Number Theorem"-quality error (given here in crude form), while for larger q, (2.8) is as good as a resonance-free strip.

We have stated the result only for the case $\mathfrak{B} \mid q$. The distribution modulo $\mathfrak{B}$ cannot be obtained directly from present methods, even though all reductions of Γ_2 are surjective (see (2.1)). Nevertheless, one can construct a set which has the desired equidistribution for all q, as claimed in

Proposition 2.9. *Given any $Y \gg 1$, there is a non-empty subset $\aleph = \aleph(Y) \subset \Gamma_2 \cap B_Y$ such that, for all square-free q and all $\mathfrak{a}_0 \in \mathrm{SL}_2(q)$,*

$$\left| \frac{\#\{\mathfrak{a} \in \aleph : \mathfrak{a} \equiv \mathfrak{a}_0 \ (q)\}}{|\aleph|} - \frac{1}{|\mathrm{SL}_2(q)|} \right| \ll \mathfrak{E}(Y; q). \tag{2.10}$$

Here $\mathfrak{E}$ is given in (2.6).

Note that we do not have particularly good control on the cardinality of $\aleph$; regardless, the estimate (2.10) is only nontrivial if $q < Y^{\Theta/C}$. The construction of the set $\aleph$ proceeds in a similar way to [BK14a, §8]; we give a sketch for the reader's convenience.

Proof of Proposition 2.9 (sketch). Let the constants $\mathfrak{B}$, c, C, and Θ be as in Theorem 2.2; they depend only on $\mathcal{A}_0 = 2$, that is, they are absolute.

Let U be a parameter to be chosen later relative to Y. Let

$$R := |\mathrm{SL}_2(\mathfrak{B})| \asymp 1, \qquad \mathcal{S}(U) := \{\gamma \in \Gamma_2 : \|\gamma\| < U\}.$$

From (1.20), we have

$$\#\mathcal{S}(U) \gg U^{2\delta_2},$$

where $\delta_2 = \delta_{\mathcal{A}_0}$ is the corresponding Hausdorff dimension. Then by the pigeonhole principle, there exists some $\mathfrak{s}_U \in \mathcal{S}(U)$ such that the set

$$\mathcal{S}'(U) := \{\gamma \in \Gamma_2 : \|\gamma\| < U, \ \gamma \equiv \mathfrak{s}_U \ (\mathfrak{B})\}$$

has cardinality

$$\#\mathcal{S}'(U) \geq \frac{1}{R} \#\mathcal{S}(U) \gg U^{2\delta_2}.$$

Observe that the elements in $\mathcal{S}'(U) \cdot \mathfrak{s}_U^{R-1}$ are all congruent to the identity mod $\mathfrak{B}$. Write $\mathrm{SL}_2(\mathfrak{B}) = \{\gamma_1, \dots, \gamma_R\}$, and find $\mathfrak{x}_1, \dots, \mathfrak{x}_R \in \Gamma$ with $\mathfrak{x}_j \equiv \gamma_j \ (\mathfrak{B})$. Such $\mathfrak{x}_j$ can be found of size $\|\mathfrak{x}_j\| \ll 1$.

For each $j = 1, \dots, R$, let

$$\mathcal{S}'_j(U) := \mathcal{S}'(U) \cdot \mathfrak{s}_U^{R-1} \cdot \mathfrak{x}_j.$$

This is a subset of Γ_2 in which each element $\mathfrak{s}$ has size

$$\|\mathfrak{s}\| \ll U^R.$$

Choose $U \asymp Y^{1/R}$ so that all elements $\mathfrak{s} \in \bigcup_j \mathcal{S}'_j(U)$ satisfy $\|\mathfrak{s}\| < Y$. Then we claim that

$$\aleph := \bigsqcup_{j=1}^{R} \mathcal{S}'_j(U)$$

gives the desired special set.

Indeed, applying Theorem 2.2 shows that for each $j = 1, \dots, R$, any q with $q \equiv 0\ (\mathfrak{B})$, and any $\omega \in \mathrm{SL}_2(q)$ with $\omega \equiv \mathfrak{x}_j\ (\mathfrak{B})$, we have

$$\begin{aligned}\#\{\mathfrak{s} \in \mathcal{S}_j'(U) : \mathfrak{s} \equiv \omega\ (q)\} &= \#\{\mathfrak{s} \in \mathcal{S}'(U) : \mathfrak{s} \equiv \omega(\mathfrak{s}_U^{R-1}\mathfrak{x}_j)^{-1}\ (q)\} \\ &= \#\{\mathfrak{s} \in \mathcal{S}(U) : \mathfrak{s} \equiv \omega(\mathfrak{s}_U^{R-1}\mathfrak{x}_j)^{-1}\ (q)\} \\ &= \frac{|\mathrm{SL}_2(\mathfrak{B})|}{|\mathrm{SL}_2(q)|}\#\{\mathfrak{s} \in \mathcal{S}(U) : \mathfrak{s} \equiv \mathfrak{s}_U\ (\mathfrak{B})\} + O(\mathfrak{E}(U; q)) \\ &= \frac{|\mathrm{SL}_2(\mathfrak{B})|}{|\mathrm{SL}_2(q)|}\#\mathcal{S}_j'(U) + O(\mathfrak{E}(U; q)).\end{aligned}$$

Then the sets $\mathcal{S}_j'(U)$ each have good modular distribution properties for distinct residues mod $\mathfrak{B}$. Note that they also all have the same cardinality, namely that of $\mathcal{S}'(U)$. Moreover, after renaming constants, we have $\mathfrak{E}(U; q) \ll \mathfrak{E}(Y; q)$.

The equidistribution (2.10) is now clear for any $q \equiv 0\ (\mathfrak{B})$, while the same for other q is obtained by summing over suitable arithmetic progressions. □

2.2. *An exponential sum over* $\mathrm{SL}_2(\mathbb{Z})$

In this subsection, we state an estimate, showing roughly that there is cancellation in a certain exponential sum over $\mathrm{SL}_2(\mathbb{Z})$ in a ball. We identify $\mathbb{Z}^4$ with $M_{2\times 2}(\mathbb{Z})$, and observe that for $A, B \in M_{2\times 2}(\mathbb{Z})$,

$$\mathrm{tr}({}^tAB) = A \cdot B, \tag{2.11}$$

where the operation on the right is the dot product in $\mathbb{Z}^4$. We first give the following local result.

Lemma 2.12. *For any square-free $q \geq 1$, any vector $\mathbf{s} \in \mathbb{Z}^4$ with $(\mathbf{s}, q) = 1$, and any $\varepsilon > 0$,*

$$\Big|\sum_{\gamma \in \mathrm{SL}_2(q)} e_q(\gamma \cdot \mathbf{s})\Big| \ll q^{3/2+\varepsilon}.$$

Proof. We could appeal to Deligne, but in fact the estimate is elementary, involving only Weil's bound for Kloosterman sums. The left hand side is multiplicative and q is square-free, so we may consider just the case of $q = p$ prime. Writing $\mathbf{s} = (x, y, z, w)$, we may assume that, say, $y \not\equiv 0\ (p)$. Writing $\gamma \in \mathrm{SL}_2(p)$ as $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, we break the sum according to whether or not $c \equiv 0$. The former case contributes

$$\sideset{}{'}\sum_{a \bmod p} \sum_{b \bmod p} e_p(ax + by + \bar{a}w) = 0,$$

since $y \not\equiv 0\ (p)$. When $c \not\equiv 0$, we have

$$\begin{aligned}&\sideset{}{'}\sum_{c \bmod p} \sum_{a,d \bmod p} e_p\big(ax + \bar{c}(ad-1)y + cz + dw\big) \\ &\qquad = \sideset{}{'}\sum_{c \bmod p} e_p(cz - \bar{c}y) \sum_{a \bmod p} e_p(ax) \sum_{d \bmod p} e_p(d(\bar{c}ay + w)).\end{aligned}$$

The d sum vanishes except for the one value of $a \equiv -c\bar{y}w \ (p)$, in which case it contributes p. What remains is a Kloosterman sum in c, which is bounded by $2\sqrt{p}$, since $y \not\equiv 0 \ (p)$. □

The next result we record is well-known (see, e.g., a special case of [BK15, Theorem 2.9]).

Lemma 2.13. *Let $X \gg 1$ be a growing parameter. There exists a function*

$$\varphi_X : \mathrm{SL}_2(\mathbb{R}) \to \mathbb{R}_{\geq 0}$$

which approximates the indicator of an archimedean ball, by which we mean the following. We have the lower bound

$$\varphi_X(g) \geq 1 \tag{2.14}$$

on $\|g\| \leq X$, and the upper bound

$$\sum_{\xi \in \mathrm{SL}_2(\mathbb{Z})} \varphi_X(\xi) \ \ll \ X^2. \tag{2.15}$$

Moreover, for any $q \geq 1$, and any $\gamma_0 \in \mathrm{SL}_2(q)$,

$$\sum_{\substack{\xi \in \mathrm{SL}_2(\mathbb{Z}) \\ \xi \equiv \gamma_0 \ (q)}} \varphi_X(\xi) = \frac{1}{|\mathrm{SL}_2(q)|} \sum_{\xi \in \mathrm{SL}_2(\mathbb{Z})} \varphi_X(\xi) + O(X^{3/2}). \tag{2.16}$$

The error term in (2.16) comes from applying Selberg's 3/16th spectral gap [Sel65]; of course better estimates are now known [KS03], but since we are not optimizing exponents, we will use the simplest results which suffice.

Combining the previous two lemmata, we obtain the main result of this subsection

Proposition 2.17. *Let φ_X be as in Lemma 2.13. For any square-free $q \geq 1$, any vector $\mathbf{s} \in \mathbb{Z}^4$ with $(\mathbf{s}, q) = 1$, and any $\varepsilon > 0$,*

$$\Big| \sum_{\xi \in \mathrm{SL}_2(\mathbb{Z})} \varphi_X(\xi) e_q(\xi \cdot \mathbf{s}) \Big| \ll q^{-3/2+\varepsilon} X^2 + q^3 X^{3/2} \quad (X \to \infty). \tag{2.18}$$

Proof. Decompose the left side of (2.18) according to the residue class of ξ in $\mathrm{SL}_2(q)$, and apply (2.16) and (2.15), to get

$$\begin{aligned} \Big| \sum_{\xi \in \mathrm{SL}_2(\mathbb{Z})} \varphi_X(\xi) e_q(\xi \cdot \mathbf{s}) \Big| &= \Big| \sum_{\gamma \in \mathrm{SL}_2(q)} e_q(\gamma \cdot \mathbf{s}) \sum_{\substack{\xi \in \mathrm{SL}_2(\mathbb{Z}) \\ \xi \equiv \gamma \ (q)}} \varphi_X(\xi) \Big| \\ &\ll \Big| \sum_{\gamma \in \mathrm{SL}_2(q)} e_q(\gamma \cdot \mathbf{s}) \Big| \frac{X^2}{|\mathrm{SL}_2(q)|} + O(q^3 X^{3/2}). \end{aligned}$$

The estimate follows from Lemma 2.12. □

3. Construction of Π and the sieving theorem

3.1. Construction of the set Π

The first goal in this subsection is to construct an appropriate subset Π of $\Gamma_{\mathcal{A}} \cap B_N$ in which to execute our sieve. Let $\mathcal{A} < \infty$ be fixed, let $\Gamma_{\mathcal{A}}$ be the semigroup in (1.18), and let $\delta_{\mathcal{A}}$ be the corresponding dimension (1.21), assumed to be near 1. Since $\mathcal{A}$ is fixed, we henceforth drop the subscripts, writing $\Gamma = \Gamma_{\mathcal{A}}$ and $\delta = \delta_{\mathcal{A}}$. Recall also that implied constants may depend on $\mathcal{A}$ without further specification.

Recall that N is our main growing parameter, and let

$$X = N^x, \quad Y = N^y, \quad Z = N^z, \quad x, y, z > 0 \tag{3.1}$$

be some parameters to be chosen later; they will decompose N, in the sense that

$$N = XYZ, \quad \text{that is,} \quad x + y + z = 1. \tag{3.2}$$

We think of X as large, $X > N^{1-\eta}$, and Y and Z as tiny.

Let $\aleph = \aleph(Y) \subset \Gamma_2 \subset \Gamma$ be the set constructed in Proposition 2.9, and let

$$\Xi_0 := \{\xi \in \Gamma : \|\xi\| < X\}, \quad \Omega_0 := \{\omega \in \Gamma : \|\omega\| < Z\} \tag{3.3}$$

be norm balls in Γ. While we do not have good control on the size of $|\aleph|$, recall from (1.20) that

$$|\Xi_0| \asymp X^{2\delta}, \quad |\Omega_0| \asymp Z^{2\delta}. \tag{3.4}$$

We will want the products

$$\xi_0 \cdot \mathfrak{a} \cdot \omega_0$$

to be unique for $\xi_0 \in \Xi_0$, $\mathfrak{a} \in \aleph$, $\omega_0 \in \Omega_0$; since $\Gamma_{\mathcal{A}}$ is a free finitely-generated semigroup, this will be the case if the wordlength $\ell(\cdot)$ in the generators (1.17) is fixed in each norm ball. It is easy to see that the wordlength metric is commensurable to the log-norm,

$$\ell(\gamma) \asymp \log \|\gamma\| . \tag{3.5}$$

Then by the pigeonhole principle and (3.4), there is some $\ell_X \asymp \log X$ such that, defining

$$\Xi := \{\gamma \in \Xi_0 : \ell(\gamma) = \ell_X\},$$

we have

$$\#\Xi \gg X^{2\delta}/\log X. \tag{3.6}$$

Similarly, there is some $\ell_Z \asymp \log Z$ such that if we define Ω to be the subset of Ω_0 having wordlength exactly ℓ_Z, then

$$\#\Omega \gg Z^{2\delta}/\log Z. \tag{3.7}$$

Then the product

$$\Pi := \Xi \cdot \aleph \cdot \Omega \tag{3.8}$$

is a subset (and not a multi-set, since the products are unique) of Γ. By (3.2), we clearly have

$$\Pi \subset \Gamma \cap B_{100N}. \tag{3.9}$$

The set Π will have our desired "bilinear" (in fact, multi-linear) structure, suitable for sieving.

3.2. Statement of the sieving theorem

We can finally state the main sieving theorems for Π.

Theorem 3.10. *Let Π_{AP} denote the set of elements $\varpi \in \Pi$ for which $\mathrm{tr}(\varpi)^2 - 4$ is "almost prime," in particular having no "small" prime factors,*

$$\Pi_{\mathrm{AP}} := \{\varpi \in \Pi : p \mid (\mathrm{tr}(\varpi)^2 - 4) \Rightarrow p > N^{1/350}\}.$$

Then for any sufficiently small $\eta > 0$, there is an $\mathcal{A} = \mathcal{A}(\eta)$ sufficiently large, and a choice of the parameters X, Y, Z in (3.1), such that

$$\#\Pi_{\mathrm{AP}} > N^{2\delta-\eta} \quad (N \to \infty). \tag{3.11}$$

Theorem 3.10 will easily imply Theorem 1.23, and will itself be easily implied by the following "level of distribution" result.

Define the sifting sequence $\mathfrak{A} = \{a_N\}$ by

$$a_N(n) := \sum_{\varpi \in \Pi} \mathbf{1}_{\{\mathrm{tr}(\varpi)^2-4=n\}}. \tag{3.12}$$

Note that, by (3.9),

$$\mathrm{supp}\,\mathfrak{A} \subset \{n \ll T\}, \tag{3.13}$$

where $T = N^2$ (see (1.25)). For a parameter $\mathcal{Q}$ and any square-free $\mathfrak{q} < \mathcal{Q}$, we define

$$|\mathfrak{A}_{\mathfrak{q}}| := \sum_{n \equiv 0\,(\mathfrak{q})} a_N(n). \tag{3.14}$$

Theorem 3.15. *For any sufficiently small $\eta > 0$, there is an $\mathcal{A} = \mathcal{A}(\eta)$ sufficiently large such that the sequence $\mathfrak{A}$ has level of distribution*

$$\mathcal{Q} = T^{1/32-\eta}. \tag{3.16}$$

More precisely, there is a multiplicative function $\beta : \mathbb{N} \to \mathbb{R}$ satisfying the "quadratic sieve" condition

$$\prod_{w \le p < z} (1 - \beta(p))^{-1} \le C\left(\frac{\log z}{\log w}\right)^2 \tag{3.17}$$

for some $C > 1$ and any $2 \le w < z$; and a decomposition

$$|\mathfrak{A}_{\mathfrak{q}}| = \beta(\mathfrak{q})|\Pi| + r(\mathfrak{q}) \tag{3.18}$$

such that, for all $K < \infty$,

$$\sum_{\substack{\mathfrak{q} < \mathcal{Q} \\ \text{square-free}}} |r(\mathfrak{q})| \ll_K |\Pi|(\log N)^{-K}. \tag{3.19}$$

Moreover, we can choose

$$X = N^{1-\eta} \tag{3.20}$$

in the decomposition (3.8) of Π, so that

$$\#\Pi \gg N^{2\delta-\eta}. \tag{3.21}$$

We now give a quick

Sketch of Theorem 3.10 assuming Theorem 3.15. The deduction is standard. The content of the latter is that the sifting sequence $\mathfrak{A}$ has "sieve dimension" $\kappa = 2$, and any exponent of distribution $\alpha < 1/32$; this confirms the discussion below (1.32). Taking $\alpha = 1/34$, say (again, we are not striving for optimal exponents), and using the crudest Brun sieve (see, e.g., [FI10, Theorem 6.9]), one shows that

$$\sum_{\substack{n \\ (n, P_z)=1}} a_N(n) \gg |\Pi|/(\log N)^2, \tag{3.22}$$

where $P_z = \prod_{p<z} p$ and z does not exceed $T^{\alpha/(9\kappa+1)} = T^{1/646} = N^{1/323}$; we take $z = N^{1/350}$. Of course any $n = \mathrm{tr}(\varpi)^2 - 4$ coprime to P_z has no prime factors below z. Then (3.22) and (3.21) confirm (3.11) after renaming constants. □

We focus henceforth on establishing Theorem 3.15.

3.3. The decomposition

The decomposition (3.18) is determined as follows. Inserting (3.12) into (3.14) gives

$$|\mathfrak{A}_\mathfrak{q}| = \sum_{\varpi \in \Pi} \mathbf{1}_{\{\mathrm{tr}(\varpi)^2 - 4 \equiv 0 \ (\mathfrak{q})\}} = \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q} \\ \mathfrak{t}^2 \equiv 4 \ (\mathfrak{q})}} \sum_{\varpi \in \Pi} \mathbf{1}_{\{\mathrm{tr}(\varpi) \equiv \mathfrak{t} \ (\mathfrak{q})\}}.$$

Rather than applying expansion (that is, the analogue of (1.29)) directly, we first capture the indicator function by *abelian* harmonic analysis, writing

$$|\mathfrak{A}_\mathfrak{q}| = \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q} \\ \mathfrak{t}^2 \equiv 4 \ (\mathfrak{q})}} \sum_{\varpi \in \Pi} \frac{1}{\mathfrak{q}} \sum_{q | \mathfrak{q}} \sideset{}{'}\sum_{r \bmod q} e_q(r(\mathrm{tr}(\varpi) - \mathfrak{t})).$$

Introducing a new parameter $Q_0 < \mathcal{Q}$, we obtain the decomposition (3.18) from breaking the penultimate sum above according to whether $q < Q_0$ or not. To this end, we write

$$|\mathfrak{A}_\mathfrak{q}| = \mathcal{M}_\mathfrak{q} + r(\mathfrak{q}), \tag{3.23}$$

say, where

$$\mathcal{M}_\mathfrak{q} := \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q} \\ \mathfrak{t}^2 \equiv 4 \ (\mathfrak{q})}} \sum_{\varpi \in \Pi} \frac{1}{\mathfrak{q}} \sum_{\substack{q | \mathfrak{q} \\ q < Q_0}} \sideset{}{'}\sum_{r \bmod q} e_q(r(\mathrm{tr}(\varpi) - \mathfrak{t})) \tag{3.24}$$

will be treated as a "main" term, the remainder $r(\mathfrak{q})$ being an error. The two terms are handled separately in the next two sections.

4. Main term analysis

First we wish to record the following elementary calculation. Recall that $\nu(n)$ is the number of distinct prime factors of n.

Lemma 4.1. *For $\mathfrak{q}$ square-free,*

$$\#\{\mathfrak{t} \in \mathbb{Z}/\mathfrak{q}\mathbb{Z} : \mathfrak{t}^2 \equiv 4\ (\mathfrak{q})\} \ = \ 2^{\nu(\mathfrak{q})-\mathbf{1}_{\{2|\mathfrak{q}\}}}. \tag{4.2}$$

Proof. Since $\mathfrak{q}$ is square-free, the equation is multiplicative. If $\mathfrak{q}$ is prime, then $\mathfrak{t}^2 \equiv 4$ implies $\mathfrak{t} \equiv \pm 2$, which has two solutions unless $\mathfrak{q} = 2$. □

We now analyze the $\mathcal{M}_\mathfrak{q}$ term in (3.24), proving the following

Proposition 4.3. *Let β be the multiplicative function given at primes by*

$$\beta(p) := \frac{1 + \mathbf{1}_{\{p\neq 2\}}}{p}\left(1 + \frac{1}{p^2-1}\right). \tag{4.4}$$

There is a decomposition

$$\mathcal{M}_\mathfrak{q} = \beta(\mathfrak{q})|\Pi| + r^{(1)}(\mathfrak{q}) + r^{(2)}(\mathfrak{q}), \tag{4.5}$$

where

$$\sum_{\mathfrak{q}<\mathcal{Q}} |r^{(1)}(\mathfrak{q})| \ll |\Pi|(\log \mathcal{Q})^2 \left(\frac{1}{e^{c\sqrt{\log Y}}} + \mathcal{Q}_0^C Y^{-\Theta}\right), \tag{4.6}$$

and

$$\sum_{\mathfrak{q}<\mathcal{Q}} |r^{(2)}(\mathfrak{q})| \ \ll \ |\Pi|\frac{\mathcal{Q}^\varepsilon}{\mathcal{Q}_0}. \tag{4.7}$$

Proof. Inserting the definition (3.8) of Π into (3.24) gives

$$\begin{aligned}\mathcal{M}_\mathfrak{q} &= \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q}\\ \mathfrak{t}^2\equiv 4\,(\mathfrak{q})}} \sum_{\xi\in\Xi}\sum_{\omega\in\Omega} \frac{1}{\mathfrak{q}} \sum_{\substack{q|\mathfrak{q}\\ q<\mathcal{Q}_0}} \sum_{r \bmod q}{}' \sum_{\mathfrak{a}\in\aleph} e_q(r(\mathrm{tr}(\xi\mathfrak{a}\omega) - \mathfrak{t}))\\ &= \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q}\\ \mathfrak{t}^2\equiv 4\,(\mathfrak{q})}} \sum_{\xi\in\Xi}\sum_{\omega\in\Omega} \frac{1}{\mathfrak{q}} \sum_{\substack{q|\mathfrak{q}\\ q<\mathcal{Q}_0}} \sum_{r \bmod q}{}' \sum_{\mathfrak{a}_0\in \mathrm{SL}_2(q)} e_q(r(\mathrm{tr}(\xi\mathfrak{a}_0\omega) - \mathfrak{t}))\Big[\sum_{\substack{\mathfrak{a}\in\aleph\\ \mathfrak{a}\equiv\mathfrak{a}_0\,(q)}} 1\Big].\end{aligned}$$

Applying (2.10) to the innermost sum gives

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

say, where

$$\mathcal{M}_\mathfrak{q}^{(1)} := \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q}\\ \mathfrak{t}^2\equiv 4\,(\mathfrak{q})}} |\Pi| \cdot \frac{1}{\mathfrak{q}} \sum_{\substack{q|\mathfrak{q}\\ q<\mathcal{Q}_0}} \sum_{r \bmod q}{}' \frac{1}{|\mathrm{SL}_2(q)|} \sum_{\mathfrak{a}_0\in \mathrm{SL}_2(q)} e_q(r(\mathrm{tr}(\mathfrak{a}_0) - \mathfrak{t})),$$

and

$$|r^{(1)}(\mathfrak{q})| \ll \tau(\mathfrak{q})|\Pi| \cdot \frac{1}{\mathfrak{q}} \sum_{\substack{q|\mathfrak{q} \\ q<Q_0}} q^4 \mathfrak{E}(Y;q).$$

Here we have used (4.2), and the error $\mathfrak{E}$ is as given in (2.6). (Recall that $\tau(n)$ is the number of divisors of n.) We estimate

$$\begin{aligned}\sum_{\mathfrak{q}<\mathcal{Q}} |r^{(1)}(\mathfrak{q})| &\ll |\Pi| \sum_{q<Q_0} q^4 \mathfrak{E}(Y;q) \sum_{\mathfrak{q}<\mathcal{Q}} \frac{\tau(\mathfrak{q})}{\mathfrak{q}} \\ &\ll |\Pi|(\log \mathcal{Q})^2[(\log Y)^C e^{-c\sqrt{\log Y}} + Q_0^C Y^{-\Theta}],\end{aligned}$$

thus proving (4.6).

Returning to $\mathcal{M}_{\mathfrak{q}}^{(1)}$, we add back in the large divisors q of $\mathfrak{q}$, writing

$$\mathcal{M}_{\mathfrak{q}}^{(1)} = \mathcal{M}_{\mathfrak{q}}^{(2)} + r^{(2)}(\mathfrak{q}),$$

say, where

$$\mathcal{M}_{\mathfrak{q}}^{(2)} := \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q} \\ \mathfrak{t}^2 \equiv 4\,(\mathfrak{q})}} |\Pi| \cdot \frac{1}{\mathfrak{q}} \sum_{q|\mathfrak{q}} \sideset{}{'}\sum_{r \bmod q} \frac{1}{|\mathrm{SL}_2(q)|} \sum_{\mathfrak{a}_0 \in \mathrm{SL}_2(q)} e_q(r(\mathrm{tr}(\mathfrak{a}_0) - \mathfrak{t})),$$

(That is, the condition $q < Q_0$ has been dropped in $\mathcal{M}_{\mathfrak{q}}^{(2)}$.) Given $\mathfrak{t}$, let $\rho_{\mathfrak{t}}(q)$ be the multiplicative function given at primes by

$$\rho_{\mathfrak{t}}(p) := \frac{1}{|\mathrm{SL}_2(p)|} \sum_{\gamma \in \mathrm{SL}_2(p)} \sideset{}{'}\sum_{r \bmod p} e_p(r(\mathrm{tr}(\gamma) - \mathfrak{t})),$$

so that

$$\mathcal{M}_{\mathfrak{q}}^{(2)} = \sum_{\substack{\mathfrak{t} \bmod \mathfrak{q} \\ \mathfrak{t}^2 \equiv 4\,(\mathfrak{q})}} |\Pi| \cdot \frac{1}{\mathfrak{q}} \prod_{p|\mathfrak{q}} (1 + \rho_{\mathfrak{t}}(p)).$$

Since $\mathfrak{t}^2 \equiv 4\ (\mathfrak{q})$ and $p \mid \mathfrak{q}$, we have $\mathfrak{t} \equiv \pm 2\ (p)$. By an elementary computation, we then evaluate explicitly

$$\rho_{\mathfrak{t}}(p) = \frac{1}{p^2 - 1}.$$

Using (4.2), we obtain

$$\mathcal{M}_{\mathfrak{q}}^{(2)} = |\Pi| \beta(\mathfrak{q}),$$

with β as given in (4.4).

Lastly, we deal with $r^{(2)}$. Since we crudely have $|\rho_{\mathfrak{t}}(q)| \le 1/q$, we obtain the bound

$$|r^{(2)}(\mathfrak{q})| \ll \tau(\mathfrak{q})|\Pi| \cdot \frac{1}{\mathfrak{q}} \sum_{\substack{q|\mathfrak{q} \\ q \ge Q_0}} \frac{1}{q} \ll |\Pi| \frac{\mathfrak{q}^{\varepsilon}}{\mathfrak{q}} \frac{1}{Q_0}.$$

The estimate (4.7) follows immediately, completing the proof. □

Remark 4.8. Since Y in (3.1) is a small power of N, the first error term in (4.6) saves an arbitrary power of $\log N$, as required in (3.19). For the rest of the paper, all other error terms will be power saving. In particular, if we set

$$Q_0 = N^{\alpha_0}, \quad \alpha_0 > 0, \tag{4.9}$$

the error in (4.7) is already a power saving, while the second term in (4.6) requires that

$$\alpha_0 < y\Theta/C. \tag{4.10}$$

It is here that we crucially use the expander property for Γ (in fact, it is only needed for $\Gamma_2 \subset \Gamma$); of course our whole point is to make the final level of distribution much larger.

5. Error term analysis

Returning to the decomposition (3.23), it remains to control the error term $r(\mathfrak{q})$ on average. Define

$$\mathcal{E} := \sum_{\mathfrak{q}<\mathcal{Q}} |r(\mathfrak{q})| = \sum_{\mathfrak{q}<\mathcal{Q}} \Big| \sum_{\mathfrak{t}^2\equiv 4\,(\mathfrak{q})} \sum_{\pi\in\Pi} \frac{1}{\mathfrak{q}} \sum_{\substack{q|\mathfrak{q}\\ q\geq Q_0}} \sum_{r \bmod q}{}' e_q(r(\mathrm{tr}(\xi\mathfrak{a}\omega) - \mathfrak{t})) \Big|. \tag{5.1}$$

Recall the decomposition $N = XYZ$ from (3.2). Our main result is

Theorem 5.2. *For any $\varepsilon > 0$, and any $1 \ll Q_0 < \mathcal{Q} < N \to \infty$,*

$$\mathcal{E} \ll N^{\varepsilon}|\Pi|(XZ)^{1-\delta}\left[\frac{1}{Q_0^{1/4}} + \frac{1}{Z^{1/4}} + \frac{\mathcal{Q}^4}{X^{1/4}}\right]. \tag{5.3}$$

As a first step, we massage $\mathcal{E}$ into a more convenient form. Let $\zeta(\mathfrak{q}) := |r(\mathfrak{q})|/r(\mathfrak{q})$ be the complex unit corresponding to the absolute value in (5.1), and rearrange terms as

$$\mathcal{E} = \sum_{Q_0\leq q<\mathcal{Q}} \frac{1}{q} \sum_{r \bmod q}{}' \sum_{\varpi\in\Pi} e_q(r\,\mathrm{tr}(\varpi)) \cdot \zeta_1(q, r),$$

where we have set

$$\zeta_1(q, r) := q \sum_{\substack{\mathfrak{q}<\mathcal{Q}\\ \mathfrak{q}\equiv 0\,(q)}} \frac{\zeta(\mathfrak{q})}{\mathfrak{q}} \sum_{\mathfrak{t}^2\equiv 4\,(\mathfrak{q})} e_q(-r\mathfrak{t}).$$

Decomposing Π as in (3.8) and leaving the special set $\aleph$ alone, we break the q-sum into dyadic pieces. This gives

$$\mathcal{E} \ll \sum_{\mathfrak{a}\in\aleph} \sum_{\substack{Q_0\leq Q<\mathcal{Q}\\ \text{dyadic}}} \frac{1}{Q} |\mathcal{E}_1(Q; \mathfrak{a})|, \tag{5.4}$$

where we have defined

$$\mathcal{E}_1(Q;\mathfrak{a}) := \sum_{q\asymp Q}\Big|\sideset{}{'}\sum_{r \bmod q}\zeta_1(q,r)\sum_{\xi\in\Xi}\sum_{\omega\in\Omega}e_q(r\,\mathrm{tr}(\xi\mathfrak{a}\omega))\Big|. \tag{5.5}$$

Theorem 5.2 follows immediately from the following estimate on $\mathcal{E}_1(Q;\mathfrak{a})$.

Theorem 5.6. *We have*

$$|\mathcal{E}_1(Q;\mathfrak{a})| \ll N^{\varepsilon}Q|\Xi|\,|\Omega|(XZ)^{1-\delta}\Big[\frac{1}{Q^{1/4}}+\frac{1}{Z^{1/4}}+\frac{Q^4}{X^{1/4}}\Big]. \tag{5.7}$$

Proof. To begin, capture the absolute value in (5.5) by another factor $|\zeta_2(q)| = 1$, and apply Cauchy–Schwarz in the "long" variable ξ in (5.5), obtaining

$$|\mathcal{E}_1(Q;\mathfrak{a})|^2 \ll |\Xi|\sum_{\xi\in\mathrm{SL}_2(\mathbb{Z})}\varphi_X(\xi)\Big|\sum_{q\asymp Q}\zeta_2(q)\sideset{}{'}\sum_{r \bmod q}\zeta_1(q,r)\sum_{\omega\in\Omega}e_q(r\,\mathrm{tr}(\xi\mathfrak{a}\omega))\Big|^2.$$

Here we have used positivity and (2.14) to insert the weighting function φ_X from Proposition 2.17 and extend the ξ-sum to all of $\mathrm{SL}_2(\mathbb{Z})$. Since the trace of a product is a dot-product (on identifying $\mathbb{Z}^4$ with $M_{2\times 2}(\mathbb{Z})$ as in (2.11)), it is linear, and hence when we open the square, we obtain

$$|\mathcal{E}_1(Q;\mathfrak{a})|^2 \ll \mathcal{Q}^{\varepsilon}|\Xi|\sum_{q,q'\asymp Q}\sum_{\omega,\omega'}\sideset{}{'}\sum_{\substack{r \bmod q\\ r' \bmod q'}}\Big|\sum_{\xi\in\mathrm{SL}_2(\mathbb{Z})}\varphi_X(\xi)\,e\Big(\xi\cdot\Big[\frac{r}{q}\mathfrak{a}\omega-\frac{r'}{q'}\mathfrak{a}\omega'\Big]\Big)\Big|. \tag{5.8}$$

Here we have used the crude estimate $|\zeta_1(q,r)| \ll \mathcal{Q}^{\varepsilon}$.

Write the bracketed expression in lowest terms as

$$\frac{\mathbf{s}}{q_0} := \frac{r}{q}\mathfrak{a}\omega - \frac{r'}{q'}\mathfrak{a}\omega', \tag{5.9}$$

with $\mathbf{s} = \mathbf{s}(q,q',r,r',\omega,\omega',\mathfrak{a}) \in \mathbb{Z}^4$ being coprime to $q_0 \geq 1$; here q_0 depends on the same parameters as $\mathbf{s}$. To study this expression in greater detail, we introduce some more notation. All variables labelled q, however decorated, denote square-free numbers.

Write

$$\widetilde{q} := (q,q'), \quad q = q_1\widetilde{q}, \quad q' = q_1'\widetilde{q}, \quad \widehat{q} := [q,q'] = q_1q_1'\widetilde{q},$$

and observe from (5.9) that $q_1q_1' \mid q_0$ and $q_0 \mid \widehat{q}$. Hence we can furthermore write

$$\widetilde{q}_0 := (q_0,\widetilde{q}), \quad \widehat{q} = q_0\widehat{q}_0 = q_1q_1'\widetilde{q}_0\widehat{q}_0,$$

whence $q_0 = q_1q_1'\widetilde{q}_0$. Note also that $Q \ll \widehat{q} \ll Q^2$.

Observe further that (5.9) implies

$$q_1'r\omega \equiv q_1r'\omega' \ \ (\widehat{q}_0),$$

and using $\det\omega = \det\omega' = 1$ gives

$$(q_1' r)^2 \equiv (q_1 r')^2 \ \ (\widehat{q}_0).$$

Since $(q_1 r', \widehat{q}_0) = 1 = (q_1' r, \widehat{q}_0)$, we obtain

$$q_1' r \equiv u q_1 r' \ \ (\widehat{q}_0), \tag{5.10}$$

where $u^2 \equiv 1\ (\widehat{q}_0)$. There are at most $2^{\nu(\widehat{q}_0)} \ll N^\varepsilon$ such $u \bmod \widehat{q}_0$, where $\nu(m)$ is the number of distinct prime factors of m. It follows that

$$\omega \equiv u\omega' \ \ (\widehat{q}_0). \tag{5.11}$$

To make full use of this last condition, we extend the ω-summation to all of $\mathrm{SL}_2(\mathbb{Z})$, again inserting the smoothing function φ, now to parameter Z. In summary, we have

$$|\mathcal{E}_1(Q,\mathfrak{a})|^2 \ll |\Xi| \sum_{Q \ll \widehat{q} \ll Q^2} \sum_{\substack{q_1 q_1' \widetilde{q}_0 \widehat{q}_0 = \widehat{q} \\ q := q_1 \widetilde{q}_0 \widehat{q}_0 \asymp Q \\ q' := q_1' \widetilde{q}_0 \widehat{q}_0 \asymp Q \\ q_0 := q_1 q_1' \widetilde{q}_0}} \sum_{\substack{u \bmod \widehat{q}_0 \\ u^2 \equiv 1\ (\widehat{q}_0)}} \sum_{r \bmod q}{}' \sum_{\substack{r' \bmod q' \\ q_1' r \equiv u q_1 r'\ (\widehat{q}_0)}}{}' \sum_{\omega' \in \Omega}$$
$$\sum_{\substack{\omega \in \mathrm{SL}_2(\mathbb{Z}),\ \omega \equiv u\omega'\ (\widehat{q}_0) \\ s := q_0(\frac{r}{q}\mathfrak{a}\omega - \frac{r'}{q'}\mathfrak{a}\omega') \in \mathbb{Z}^4 \\ (s, q_0) = 1}} \varphi_Z(\omega) \Big| \sum_{\xi \in \mathrm{SL}_2(\mathbb{Z})} \varphi_X(\xi) e_{q_0}(\xi \cdot s) \Big|.$$

Working from the inside out, apply (2.18) to the innermost ξ-sum, and (2.16) to the ω-sum, estimating the ω'-sum trivially. There are at most $q'/\widehat{q}_0$ values for r', and at most q values for r; note that

$$\frac{qq'}{\widehat{q}_0} = \frac{qq' q_0}{\widehat{q}} \ll \frac{Q^2 q_0}{\widehat{q}}.$$

The u-sum contributes N^ε, as does the sum over divisors of $\widehat{q}$. Putting everything together gives

$$|\mathcal{E}_1(Q,\mathfrak{a})|^2 \ll |\Xi| N^\varepsilon \sum_{Q \ll \widehat{q} \ll Q^2} \sum_{q_0 \widehat{q}_0 = \widehat{q}} \frac{Q^2 q_0}{\widehat{q}} |\Omega| \Big[\frac{Z^2}{\widehat{q}_0^3} + Z^{3/2}\Big]\Big[\frac{X^2}{q_0^{3/2}} + q_0^3 X^{3/2}\Big]$$
$$\ll N^\varepsilon Q^2 |\Xi|^2 |\Omega|^2 (XZ)^{2(1-\delta)} \sum_{Q \ll \widehat{q} \ll Q^2} \frac{1}{\widehat{q}} \Big[\frac{1}{\widehat{q}^{\,1/2}} + \frac{1}{Z^{1/2}} + \frac{Q^8}{X^{1/2}}\Big],$$

where we have used (3.6) and (3.7). Theorem 5.6 follows immediately, as does Theorem 5.2. □

6. Proof of the sieving theorem

In this section, we combine the analyses of the previous two to prove Theorem 3.15.

Let $\mathfrak{A} = \{a_N(n)\}_{n\geq 1}$ be as constructed in (3.12). Combining (3.23) and (4.5) gives the decomposition

$$|\mathfrak{A}_\mathfrak{q}| = \beta(\mathfrak{q})|\Pi| + r^{(1)}(\mathfrak{q}) + r^{(2)}(\mathfrak{q}) + r(\mathfrak{q}),$$

as in (3.18), with β given by (4.4). It is classical that (3.17) holds, so it remains to verify (3.19) with $\mathcal{Q}$ being the level of distribution. Write

$$X = N^x, \quad Y = N^y, \quad Z = N^z, \quad \mathcal{Q} = T^\alpha = N^{2\alpha}, \quad Q_0 = N^{\alpha_0},$$

with

$$x + y + z = 1. \tag{6.1}$$

The bounds (4.6) and (4.7) are sufficient as long as $y, \alpha_0 > 0$ and

$$\alpha_0 < y\Theta/C. \tag{6.2}$$

The three error terms in (5.3) are sufficiently controlled if

$$\alpha_0/4 > (x + z)(1 - \delta), \tag{6.3}$$

$$z/4 > (x + z)(1 - \delta), \tag{6.4}$$

$$x/4 > 8\alpha + (x + z)(1 - \delta). \tag{6.5}$$

Remark 6.6. If we take y and α_0 very small and x and δ very near 1, it is clear that (6.5) will not allow us to do better than $\alpha < 1/32$; this is what we achieve below.

Now, let $\eta > 0$ be given, sufficiently small, and set

$$\alpha = 1/32 - \eta,$$

as claimed in (3.16). We will assume further that

$$\delta > 1 - \eta$$

(more stringent restrictions on δ follow), and set

$$x = 1 - \eta,$$

so that (3.20) and (3.21) are satisfied. Then an elementary computation shows that (6.5) is satisfied.

After more elementary manipulations, we may set

$$z = \frac{\eta}{1 + C/\Theta}, \quad y = z \cdot \frac{C}{\Theta}, \quad \alpha_0 = \frac{5}{6}z,$$

and assume that

$$\delta > 1 - \frac{\eta}{5(1 + C/\Theta)}. \tag{6.7}$$

Then $y\Theta/C = \frac{6}{5}\alpha_0 > \alpha_0$, whence (6.2) is satisfied. Likewise,

$$z/4 > \alpha_0/4 = \tfrac{5}{24}z > \tfrac{1}{5}z > 1-\delta > (x+z)(1-\delta),$$

so that (6.3) and (6.4) hold. The condition (6.7) is guaranteed by taking $\mathcal{A}$ sufficiently large (see (1.22)). This completes the proof of Theorem 3.15.

Remark 6.8. We emphasize again that it is in the last step here that we need $\aleph$ to come from the fixed group $\Gamma_2 \subset \Gamma_{\mathcal{A}}$. Indeed, the constants Θ and C are then absolute (see Theorem 2.2), and do not depend on $\mathcal{A}$, so (6.7) can be ensured by taking $\mathcal{A}$ large.

7. Proof of Theorem 1.8

Having established Theorem 3.15 (and hence Theorem 3.10) in the last section, we are in a position to prove Theorem 1.23, from which we will deduce Theorem 1.8.

7.1. Proof of Theorem 1.23

The deduction from Theorem 3.10 is straightforward, but we give the details. We begin by first bounding the trace multiplicity.

Lemma 7.1. *For any $\mathcal{A} < \infty$, and any $t \geq 1$,*

$$\#\{\gamma \in \Gamma_{\mathcal{A}} : \mathrm{tr}(\gamma) = t\} \ll t^{1+\varepsilon}. \tag{7.2}$$

Proof. Let $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_{\mathcal{A}}$ have trace $a + d = t$. Since the entries of $\Gamma_{\mathcal{A}}$ are all positive, there are at most t choices of a, whence $d = t - a$ is determined. Then $bc = ad - 1 \leq t^2$ is determined, and there are $\ll t^{\varepsilon}$ choices for the divisors b and c. □

Returning to Theorem 1.23, let $\eta > 0$ be given. Applying Theorem 3.10 gives a sufficiently large $\mathcal{A} = \mathcal{A}(\eta)$ and a set $\Pi \subset \Gamma_{\mathcal{A}} \cap B_N$ such that (3.11) holds. To illustrate more clearly the mechanism below, write

$$\alpha = 1/350,$$

so that

$$\Pi_{\mathrm{AP}} = \{\varpi \in \Pi : p \,|\, (\mathrm{tr}(\varpi)^2 - 4) \Rightarrow p > N^{\alpha}\}.$$

Now, we have

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

where we have used (3.11) and defined

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

Now, for each $\gamma \in \Pi_{\mathrm{AP}}^{\square}$, there is a prime p with $p^2 \mid (\mathrm{tr}(\gamma)^2 - 4)$. Since $\gamma \in \Pi_{\mathrm{AP}}$, we thus have $p > N^{\alpha}$, and moreover p^2 divides either $\mathrm{tr}(\gamma) + 2$ or $\mathrm{tr}(\gamma) - 2$; in particular, $p \ll N^{1/2}$. Therefore, reversing orders and applying (7.2), we have

$$\begin{aligned}\#\Pi_{\mathrm{AP}}^{\square} &\leq \sum_{N^{\alpha} < p \ll N^{1/2}} \sum_{\substack{t < N \\ t^2 - 4 \equiv 0 \,(p^2)}} \#\{\gamma \in \Gamma_{\mathcal{A}} \cap B_N : \mathrm{tr}(\gamma) = t\} \\ &\ll \sum_{N^{\alpha} < p \ll N^{1/2}} \frac{N}{p^2} N^{1+\varepsilon} \ll N^{2-\alpha+\varepsilon}.\end{aligned}$$

Inserting this estimate into (7.3) gives

$$\#\{\gamma \in \Gamma_{\mathcal{A}} \cap B_N : \mathrm{tr}(\gamma)^2 - 4 \text{ is square-free}\} > N^{2\delta-\eta} - O(N^{2-\alpha+\varepsilon}).$$

The above estimate is sufficient to establish (1.24), as long as, roughly,

$$2\delta > 2 - \alpha. \tag{7.4}$$

This explains (up to constants) the discussion on p. 1339 that the exponent of distribution needs to be strong enough to overcome the thinness of $\Gamma_{\mathcal{A}}$. Of course, since we have proved the above with the absolute quantity $\alpha = 1/350$, it follows that as long as $\delta - \eta/2 > 1 - 1/700$ (equivalently, $\mathcal{A}$ sufficiently large), we ensure that (7.4) holds. This completes the proof of Theorem 1.23.

7.2. *Proof of Theorem 1.8*

Again, this will be an easy consequence of Theorem 1.23. Let

$$\mathcal{T} := \{t \geq 1 : t^2 - 4 \text{ is square-free}\},$$

and for an integer t and $\mathcal{A} < \infty$, let the trace multiplicity be

$$\mathcal{M}_{\mathcal{A}}(t) := \#\{\gamma \in \Gamma_{\mathcal{A}} : \mathrm{tr}(\gamma) = t\}.$$

Our main claim is that, for any $\eta > 0$, there is an $\mathcal{A}$ sufficiently large such that

$$\sum_{t \in \mathcal{T} \cap [1,N]} \mathbf{1}_{\{\mathcal{M}_{\mathcal{A}}(t) \geq t^{2\delta-1-\eta}\}} > N^{2\delta-1-\eta}. \tag{7.5}$$

Indeed, from Theorem 1.23, we have

$$N^{2\delta-\eta} < \sum_{t \in \mathcal{T} \cap [1,N]} \mathcal{M}_{\mathcal{A},N}(t) = \sum_{t \in \mathcal{T} \cap [1,N]} \mathcal{M}_{\mathcal{A},N}(t)(\mathbf{1}_{\{\mathcal{M}_{\mathcal{A},N}(t) \geq W\}} + \mathbf{1}_{\{\mathcal{M}_{\mathcal{A},N}(t) < W\}}),$$

where we have introduced a parameter W to be chosen later. Using (7.2) gives

$$N^{2\delta-\eta} \ll N^{1+\varepsilon} \sum_{t \in \mathcal{T} \cap [1,N]} \mathbf{1}_{\{\mathcal{M}_{\mathcal{A},N}(t) \geq W\}} + NW,$$

from which (7.5) follows on setting $W = N^{2\delta-1-2\eta}$, say, and renaming constants.

Now let $\epsilon > 0$ be given, and take $\eta > 0$ small enough and $\mathcal{A}$ large enough that

$$2\delta - 1 - \eta > 1 - \epsilon. \tag{7.6}$$

This choice of $\mathcal{A} = \mathcal{A}(\epsilon)$ corresponds to a compact region

$$\mathcal{Y} = \mathcal{Y}(\epsilon) \subset \mathcal{X} \quad (= T^1(\mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H})),$$

as in §1.2.1. Define the set $\mathscr{D} = \mathscr{D}(\epsilon)$ to be

$$\mathscr{D} := \{D = t^2 - 4 : t \in \mathcal{T},\ \mathcal{M}_{\mathcal{A}}(t) > t^{2\delta-1-\eta}\}.$$

All $D \in \mathscr{D}$ are square-free, and hence fundamental, as are the corresponding geodesics by Lemma 1.14. Moreover,

$$\#(\mathscr{D} \cap [1, T]) \geq \#\{t \in \mathcal{T} \cap [1, \sqrt{T}] : \mathcal{M}_{\mathcal{A}}(t) > t^{2\delta-1-\eta}\} > T^{1/2-\epsilon}$$

by (7.5) and (7.6). This confirms (1.10).

For each $D = t^2 - 4 \in \mathscr{D}$, the corresponding trace multiplicity satisfies

$$\mathcal{M}_{\mathcal{A}}(t) > t^{1-\epsilon} > (\sqrt{D})^{1-\epsilon} \gg |\mathscr{C}_D|^{1-\epsilon}. \tag{7.7}$$

Now, it is *not* the case that each $\gamma \in \Gamma_{\mathcal{A}}$ corresponds uniquely to a closed geodesic on $\mathcal{X}$, but since the corresponding visual points (1.13) of the geodesic are all reduced, any two differ by a cyclic permutation of their partial quotients. Recalling from (3.5) that the wordlength metric is commensurable with the log-norm, there can be at most $C \log t$ such permutations. Together with (7.7), this gives (1.9), and completes the proof of Theorem 1.8.

Appendix. Proof of Theorem 1.11

In this appendix, we prove Theorem 1.11; it is a pleasure to thank Elon Lindenstrauss for suggesting the argument given here. Again, the method is more-or-less standard in the ergodic-theory community, so we only give a sketch.

Let $\mathcal{Y} \subset \mathcal{X}$ be a given compact region, and let D be a large non-square number (we do not require that D be fundamental in this section), with corresponding class group $\mathscr{C}_D$ and class number h_D. Recall again that Duke's theorem (now in effective form) states that, for a smooth function ψ on $\mathcal{X} = T^1(\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H})$, we have

$$\int_{\mathcal{X}} \psi \, d\mu_D = \int_{\mathcal{X}} \psi \, d\mu_{\mathcal{X}} + O(D^{-c}\mathcal{S}\psi) \quad (D \to \infty), \tag{A.1}$$

where, as in (1.2), $\mu_{\mathcal{X}}$ is Haar measure on $\mathcal{X}$, μ_D is the measure associated to $\mathscr{C}_D$, namely,

$$\mu_D = \frac{1}{h_D} \sum_{\gamma \in \mathscr{C}_D} \mu_\gamma,$$

and $\mathcal{S}\psi$ is a finite-order Sobolev norm of ψ (see, e.g., [CU04]). The constant $c > 0$ in the error rate of (A.1) could be made precise in terms of subconvexity bounds for certain L-functions, but we prefer to keep the exponent qualitative for ease of exposition.

Let $0 \le F \le 1$ be a fixed function on $\mathcal{X}$ which smoothly approximates the indicator function of the complement $\mathcal{X} \setminus \mathcal{Y}$; in particular, we assume the support of F is outside of $\mathcal{Y}$. Now suppose that $\gamma \in \mathscr{C}_D$ is a low-lying geodesic, $\gamma \subset \mathcal{Y}$. Then, writing T for the time-1 shift under the geodesic flow, we see for any $x \in \gamma$ that $T^\ell.x \in \gamma$, and hence

$$\{x, T.x, \dots, T^{k-1}.x\} \cap \operatorname{supp} F = \emptyset.$$

Let

$$M := \int_{\mathcal{X}} F \, d\mu_{\mathcal{X}}$$

be the mean of F, so that $F_0 := F - M$ has mean zero. Furthermore, for a parameter k to be chosen later (of size roughly $\log D$), define

$$f := \frac{1}{k}(F_0 + T.F_0 + \cdots + T^{k-1}.F_0).$$

Note that for such x, we have $f(x) = -M$, and hence

$$\begin{aligned}\frac{1}{h_D} \sum_{\gamma \in \mathscr{C}_D} \mathbf{1}_{\{\gamma \subset \mathcal{Y}\}} &\le \mu_D(\{x : \{x, T.x, \dots, T^{k-1}.x\} \cap \operatorname{supp} F = \emptyset\}) \\ &\le \mu_D(\{x : |f(x)| \ge M\}) \le \frac{1}{M^{2\ell}} \int_{\mathcal{X}} f^{2\ell} \, d\mu_D,\end{aligned}$$

where we have introduced another parameter $1 \le \ell \le k$ to be chosen later (of size a small constant times k).

Apply Duke's theorem (A.1) to the last term, getting

$$\frac{1}{h_D} \sum_{\gamma \in \mathscr{C}_D} \mathbf{1}_{\{\gamma \subset \mathcal{Y}\}} \le \frac{1}{M^{2\ell}} \int_{\mathcal{X}} f^{2\ell} \, d\mu_{\mathcal{X}} + O(D^{-c} C^k), \tag{A.2}$$

where we have estimated $\mathcal{S}(f^{2\ell}) < C^k$, since F is fixed. Now, the geodesic flow is a Bernoulli flow, and hence mixing of all orders. It follows that

$$\int_{\mathcal{X}} f^{2\ell} \, d\mu_{\mathcal{X}} \ll_{F_0} \left(\frac{2\ell}{k}\right)^{\ell}. \tag{A.3}$$

Inserting (A.3) into (A.2) gives

$$\frac{1}{h_D} \sum_{\gamma \in \mathscr{C}_D} \mathbf{1}_{\{\gamma \subset \mathcal{Y}\}} \ll_{\mathcal{Y}} \left(\frac{2\ell}{kM^2}\right)^{\ell} + D^{-c} C^k. \tag{A.4}$$

Choosing

$$k = \frac{c}{2 \log C} \cdot \log D,$$

say, makes the second error in (A.4) of size $D^{-c}C^k = D^{-c/2}$. Choosing

$$\ell = \frac{M^2}{4} \cdot k,$$

say, makes the first term in (A.4) of size

$$\left(\frac{2\ell}{kM^2}\right)^{\ell} = \left(\frac{1}{2}\right)^{\ell} = D^{-cM^2 \log 2/(8 \log C)}.$$

This last exponent determines $\epsilon = \epsilon(\mathcal{Y})$, completing the proof.

Acknowledgments. It is our pleasure to thank Tim Browning, Curt McMullen, Michael Rubinstein, Zeev Rudnick, and Peter Sarnak for illuminating conversations. The second-named author would like to thank for the hospitality of the IAS, where much of this work was carried out.

JB is partially supported by NSF grant DMS-0808042.

AK is partially supported by an NSF CAREER grant DMS-1254788, an Alfred P. Sloan Research Fellowship, a Yale Junior Faculty Fellowship, and support at IAS from The Fund for Math and The Simonyi Fund.

References

[Art24] Artin, E.: Ein mechanisches System mit quasiergodischen Bahnen. Abh. Math. Sem. Univ. Hamburg **3**, 170–175 (1924) JFM 50.0677.11 MR 3069425

[BGS06] Bourgain, J., Gamburd, A., Sarnak, P.: Sieving and expanders. C. R. Math. Acad. Sci. Paris **343**, 155–159 (2006) Zbl 1217.11081 MR 2246331

[BGS10] Bourgain, J., Gamburd, A., Sarnak, P.: Affine linear sieve, expanders, and sum-product. Invent. Math. **179**, 559–644 (2010) Zbl 1239.11103 MR 2587341

[BGS11] Bourgain, J., Gamburd, A., Sarnak, P.: Generalization of Selberg's 3/16th theorem and affine sieve. Acta Math. **207**, 255–290 (2011) Zbl 1276.11081 MR 2892611

[BK10] Bourgain, J., Kontorovich, A.: On representations of integers in thin subgroups of SL(2, **Z**). Geom. Funct. Anal. **20**, 1144–1174 (2010) Zbl 1230.11050 MR 2746949

[BK11] Bourgain, J., Kontorovich, A.: On Zaremba's conjecture. C. R. Math. Acad. Sci. Paris **349**, 493–495 (2011) Zbl 1215.11005 MR 2802911

[BK14a] Bourgain, J., Kontorovich, A.: On Zaremba's conjecture. Ann. of Math. **180**, 137–196 (2014) Zbl 06316068 MR 3194813

[BK14b] Bourgain, J., Kontorovich, A.: On the local-global conjecture for integral Apollonian gaskets. Invent. Math. **196**, 589–650 (2014) Zbl 1301.11046 MR 3211042

[BK15] Bourgain, J., Kontorovich, A.: The affine sieve beyond expansion I: thin hypotenuses. Int. Math. Res. Notices **2015**, 9175–9205 Zbl 1347.11068 MR 3431590

[CU04] Clozel, L., Ullmo, E.: Équidistribution des pointes de Hecke. In: Contribution to Automorphic Forms, Geometry and Number Theory, Johns Hopkins Univ. Press, 193–254 (2004) Zbl 1068.11042 MR 2058609

[Duk88] Duke, W.: Hyperbolic distribution problems and half-integral weight Maass forms. Invent. Math. **92**, 73–90 (1988) Zbl 0628.10029 MR 0931205

[ELMV09] Einsiedler, M., Lindenstrauss, E., Michel, P., Venkatesh, A.: Distribution of periodic torus orbits on homogeneous spaces. Duke Math. J. **148**, 119–174 (2009) Zbl 1172.37003 MR 2515103

[FI10] Friedlander, J., Iwaniec, H.: Opera de Cribro, Amer. Math. Soc. Colloq. Publ. 57, Amer. Math. Soc., Providence, RI (2010) Zbl 1226.11099 MR 2647984

[HM06] Harcos, G., Michel, P.: The subconvexity problem for Rankin–Selberg L-functions and equidistribution of Heegner points. II. Invent. Math. **163**, 581–655 (2006) Zbl 1111.11027 MR 2207235

[Hen89] Hensley, D.: The distribution of badly approximable numbers and continuants with bounded digits. In: Théorie des nombres (Québec, PQ, 1987), de Gruyter, Berlin, 371–385 (1989) Zbl 0689.10042 MR 1024576

[Hen92] Hensley, D.: Continued fraction Cantor sets, Hausdorff dimension, and functional analysis. J. Number Theory **40**, 336–358 (1992) Zbl 0745.28005 MR 1154044

[Hum16] Humbert, G.: Sur les fractions continues ordinaires et les formes quadratiques binaires indéfinies. J. Math. Pures Appl. (7) **2**, 104–154 (1916) JFM 46.0272.03

[KS03] Kim, H., Sarnak, P.: Refined estimates towards the Ramanujan and Selberg conjectures (appendix in a paper of H. Kim). J. Amer. Math. Soc. **16**, 175–181 (2003) Zbl 1018.11024 MR 1937203

[Pop06] Popa, A.: Central values of Rankin L-series over real quadratic fields. Compos. Math. **142**, 811–866 (2006) Zbl 1144.11041 MR 2249532

[SGS13] Salehi Golsefidy, A., Sarnak, P.: The affine sieve. J. Amer. Math. Soc. **26**, 1085–1105 (2013) Zbl 1283.20055 MR 3073885

[Sar07] Sarnak, P.: Reciprocal geodesics. In: Analytic Number Theory, Clay Math. Proc. 7, Amer. Math. Soc., Providence, RI, 217–237 (2007) Zbl 1198.11039 MR 2362203

[Sel65] Selberg, A.: On the estimation of Fourier coefficients of modular forms. In: Proc. Sympos. Pure Math. 7, Amer. Math. Soc., Providence, RI, 1–15 (1965) Zbl 0142.33903 MR 0182610

[Ser85] Series, C.: The modular surface and continued fractions. J. London Math. Soc. (2) **31**, 69–80 (1985) Zbl 0545.30001 MR 0810563

[Zag82] Zagier, D.: On the number of Markoff numbers below a given bound. Math. Comp. **39**, 709–723 (1982) Zbl 0501.10015 MR 0669663