ALMOST PRIME PYTHAGOREAN TRIPLES IN THIN ORBITS

ALEX KONTOROVICH AND HEE OH

Abstract. For the ternary quadratic form $Q(\mathbf{x}) = x^2+y^2-z^2$ and a non-zero Pythagorean triple $\mathbf{x}_0 \in \mathbb{Z}^3$ lying on the cone $Q(\mathbf{x}) = 0$, we consider an orbit $\mathcal{O} = \mathbf{x}_0\Gamma$ of a finitely generated subgroup $\Gamma < \mathrm{SO}_Q(\mathbb{Z})$ with critical exponent exceeding $1/2$.

We find infinitely many Pythagorean triples in $\mathcal{O}$ whose hypotenuse, area, and product of side lengths have few prime factors, where "few" is explicitly quantified. We also compute the asymptotic of the number of such Pythagorean triples of norm at most T, up to bounded constants.

Contents

1. Introduction 1
2. Background and More on Theorem 1.5 6
3. Equidistribution of Expanding Closed Horocycles 15
4. Proofs of the Counting Theorems 21
5. Proofs of the Sieving Theorems 26
Appendix A. Proof of Theorem 1.13 31
References 37

1. Introduction

1.1. **The Affine Linear Sieve.** In [BGS06], Bourgain, Gamburd, and Sarnak introduced the Affine Linear Sieve, which extends some classical sieve methods to thin orbits of non-abelian group actions. Its input is a pair $(\mathcal{O}, F)$, where

(1) $\mathcal{O}$ is a discrete orbit, $\mathcal{O} = \mathbf{x}_0 \cdot \Gamma$, generated by a discrete subgroup Γ of a linear group G. It is called "thin" if the volume of $\Gamma\backslash G$ is infinite; and
(2) F is a polynomial, taking integer values on $\mathcal{O}$.

Given the pair $(\mathcal{O}, F)$, the Affine Linear Sieve attempts to output a number $R = R(\mathcal{O}, F)$ as small as possible so that there are infinitely many integers $n \in F(\mathcal{O})$, with n having at most R prime factors (counted with multiplicities).

A special case of their main result is the following.

Kontorovich is partially supported by NSF grants DMS-0802998 and DMS-0635607, and the Ellentuck Fund at IAS.

Oh is partially supported by NSF grant DMS-0629322.

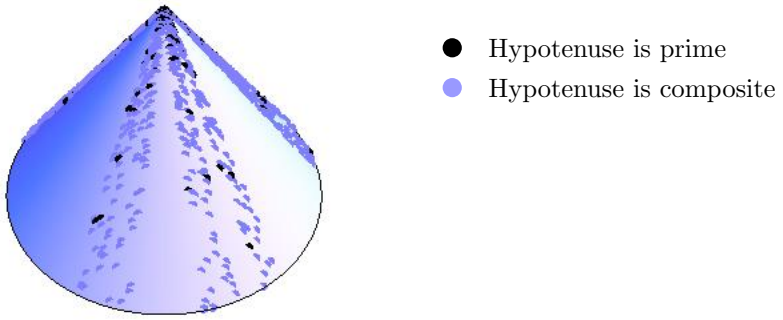

FIGURE 1. A thin orbit $\mathcal{O}$ of Pythagorean triples, sifted by hypotenuse, $F(x, y, z) = z$. The darker points denote those triples whose hypotenuse is prime.

Theorem 1.1 ([BGS06, BGS10])**.** *Let $G < \mathrm{GL}_n(\mathbb{R})$ be a $\mathbb{Q}$-form of SL_2, and let Γ be a non-elementary*[1] *subgroup of $G \cap \mathrm{GL}_n(\mathbb{Z})$. Let $\mathcal{O}$ be an orbit $\mathbf{x}_0\Gamma$ for some $\mathbf{x}_0 \in \mathbb{Z}^n \setminus \{0\}$ and F any polynomial which is integral on $\mathcal{O}$. Then there exists a number*

$$R = R(\mathcal{O}, F) < \infty$$

such that the set of $\mathbf{x} \in \mathcal{O}$ with $F(\mathbf{x})$ having at most R prime factors is Zariski dense in the Zariski closure of $\mathcal{O}$. In particular, this set is infinite.

Remark 1.2. As described in [BGS10, §2], Lagarias has given evidence that the result above may be false if one drops the condition that Γ is non-elementary.

For various special cases of $(\mathcal{O}, F)$, one can say more than just $R < \infty$; one can give explicit, "reasonable" values of $R(\mathcal{O}, F)$. This was achieved with some restrictions in [Kon07, Kon09], and it is our present goal to improve the results there in a more general setting.

In order to remove local obstructions which would increase R for trivial reasons, we will impose the strong primitivity condition on $(\mathcal{O}, F)$.

Definition 1.3. For a subset $\mathcal{O} \subset \mathbb{Z}^n$ and a polynomial $F(x_1, \cdots, x_n)$ taking integral values on $\mathcal{O}$, the pair $(\mathcal{O}, F)$ is called *strongly primitive* if for every integer $q \geq 2$ there is an $\mathbf{x} \in \mathcal{O}$ such that

$$F(\mathbf{x}) \neq 0 \pmod{q}.$$

Remark 1.4. The weaker condition of *primitivity* requires the above for q prime. See [BGS10, §2] for an example of $(\mathcal{O}, F)$ which is primitive but not strongly primitive.

To present a concrete number $R(\mathcal{O}, F)$, we will consider the quadratic form

$$Q(\mathbf{x}) = x^2 + y^2 - z^2.$$

Hence a non-zero vector $\mathbf{x} \in \mathbb{Z}^3$ is a Pythagorean triple whenever $Q(\mathbf{x}) = 0$. Let $G = \mathrm{SO}_Q(\mathbb{R})$ be the special orthogonal group preserving Q with real entries. For a

[1]Recall that a discrete subgroup $\Gamma < \mathrm{SL}(2, \mathbb{R})$ is elementary if and only if it has a cyclic subgroup of finite index.

discrete subgroup Γ of G, the critical exponent $\delta = \delta_\Gamma \in [0,1]$ of Γ is defined to be the abscissa of convergence of the Poincare series:

$$L_\Gamma(s) := \sum_{\gamma\in\Gamma} \|\gamma\|^{-s}$$

for any norm $\|\cdot\|$ on the vector space $\mathrm{M}_3(\mathbb{R})$ of 3×3 matrices. We remark that Γ is non-elementary if and only if $\delta > 0$. Moreover if Γ is finitely-generated, then Γ is of finite co-volume in G if and only if $\delta = 1$ [Pat76].

The detailed statement of our main result is given in Theorem 2.22. The following is a special case:

Theorem 1.5. *Let $\Gamma < \mathrm{SO}_Q(\mathbb{Z})$ be a finitely generated subgroup and set*

$$\mathcal{O} := \mathbf{x}_0 \cdot \Gamma,$$

for a primitive Pythagorean triple $\mathbf{x}_0$, e.g., $\mathbf{x}_0 = (3,4,5)$. Let the polynomial F be one of

$$\begin{cases} \textit{the hypotenuse:} & F_{\mathcal{H}}(\mathbf{x}) := z; \\ \textit{the ``area'':} & F_{\mathcal{A}}(\mathbf{x}) := \frac{1}{12}xy; \\ \textit{the product of coordinates :} & F_{\mathcal{C}}(\mathbf{x}) := \frac{1}{60}xyz. \end{cases}$$

We assume that the pair $(\mathcal{O}, F)$ is strongly primitive and that

$$\delta > \begin{cases} 0.9992 & \text{if } F = F_{\mathcal{H}}; \\ 0.99995 & \text{if } F = F_{\mathcal{A}}; \\ 0.99677 & \text{if } F = F_{\mathcal{C}}. \end{cases}$$

Then the following hold:

(1) *For infinitely many $\mathbf{x} \in \mathcal{O}$, the integer $F(\mathbf{x})$ has at most $R = R(\mathcal{O}, F)$ prime factors, where*

$$R = \begin{cases} 14 & \text{if } F = F_{\mathcal{H}}; \\ 25 & \text{if } F = F_{\mathcal{A}}; \\ 29 & \text{if } F = F_{\mathcal{C}}. \end{cases}$$

(2) *We have*[2]

$$\#\{\mathbf{x}\in\mathcal{O} : \|x\| < T,\ F(\mathbf{x}) \text{ has at most } R(\mathcal{O},F) \text{ prime factors}\} \asymp \frac{T^\delta}{(\log T)^\kappa},$$

where $\|\cdot\|$ is any norm on $\mathbb{R}^3$ and the sieve dimension κ is

$$\kappa = \begin{cases} 1 & \text{if } F = F_{\mathcal{H}}; \\ 4 & \text{if } F = F_{\mathcal{A}}; \\ 5 & \text{if } F = F_{\mathcal{C}}. \end{cases}$$

In particular, the set of $\mathbf{x} \in \mathcal{O}$ such that $F(\mathbf{x})$ has at most $R(\mathcal{O}, F)$ prime factors is Zariski dense in the cone $Q = 0$.

Remark 1.6. The functions $F_{\mathcal{A}}$ and $F_{\mathcal{C}}$ satisfy $F(3,4,5) = 1$; hence the pair $(\mathcal{O}, F)$ is strongly primitive for $\mathbf{x}_0 = (3,4,5)$, regardless of the choice of the group Γ. For the hypotenuse, $F_{\mathcal{H}}$, one must check, given Γ, that the pair $(\mathcal{O}, F)$ is strongly primitive.

[2] Recall that $f \asymp g$ means that $c^{-1}\cdot f(T) \le g(T) \le c\cdot f(T)$ for some $c > 1$ and for all $T > 1$.

Remark 1.7. The above theorem was proved in [Kon09] assuming that Γ contains a non-trivial (parabolic) stabilizer of $\mathbf{x}_0$. In this case, the orbit $\mathcal{O}$ contains an injection of affine space, and hence standard sieve methods [Iwa78] also produce integral points with few prime factors. Some of the most interesting cases which cannot be dealt with using standard methods and are now covered by our results are the so-called Schottky groups; these are groups generated by finitely many hyperbolic elements.

1.2. **A Counting theorem.** In order to sieve almost primes in a given orbit, one must know how to count points on such orbits, which we obtain without assuming the arithmetic condition on Γ.

Theorem 1.8. *Let Q be any ternary indefinite quadratic form, $G = \mathrm{SO}_Q(\mathbb{R})$, and $\Gamma < G$ a finitely generated discrete subgroup with $\delta > 1/2$. Let $\mathbf{x}_0 \in \mathbb{R}^3$ be a non-zero vector lying on the cone $Q = 0$ such that the orbit $\mathcal{O} = \mathbf{x}_0 \cdot \Gamma$ is discrete.*

Then there exist a constant $c_0 > 0$ and some $\zeta > 0$ such that as $T \to \infty$,

$$\#\{\mathbf{x} \in \mathcal{O} : \|\mathbf{x}\| < T\} = c_0 \cdot T^\delta + O(T^{\delta-\zeta}).$$

The norm $\|\cdot\|$ above is Euclidean.

Remark 1.9. Let N_0 denote the (unipotent) stabilizer of $\mathbf{x}_0$ in G. The Theorem 1.8 was proved in [Kon09] under the further assumption that $\Gamma \cap N_0$ is a lattice in N_0.

1.3. **Expanding Closed Horocycles.** The main difference between this paper and [Kon09] is the method used to establish counting theorems such as Theorem 1.8. While [Kon09] uses abstract operator theory, in the present work we prove the effective equidistribution of expanding closed horocycles on a hyperbolic surface X, allowing not only X to have infinite volume, but also allowing the closed horocycle to be infinite in length.

Let $G = \mathrm{SL}_2(\mathbb{R})$ and write the Iwasawa decomposition $G = NAK$ with

$$N = \left\{ n_x = \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} : x \in \mathbb{R} \right\}, \quad A = \left\{ a_y = \begin{pmatrix} \sqrt{y} & 0 \\ 0 & 1/\sqrt{y} \end{pmatrix} : y > 0 \right\}, \tag{1.10}$$

and $K = \mathrm{SO}_2(\mathbb{R})$.

We use the upper half plane $\mathbb{H} = \{z = x + iy : y > 0\}$ as a model for the hyperbolic plane with the metric $\frac{\sqrt{dx^2+dy^2}}{y}$. The group G acts on $\mathbb{H}$ by fractional linear transformations which give arise all orientation preserving isometries of $\mathbb{H}$:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{az+b}{cz+d}$$

for $\mathrm{Im}(z) > 0$. We compute:

$$n_x a_y(i) = x + iy.$$

Let $\Gamma < G$ be a finitely generated discrete subgroup with $\delta > 1/2$. Assume that the horocycle $(\Gamma \cap N)\backslash N$ is closed in $X := \Gamma\backslash G$, or equivalently the image of $N(i) = \{x + i : x \in \mathbb{R}\}$ is closed in $\Gamma\backslash\mathbb{H}$ under the canonical projection $\mathbb{H} \to \Gamma\backslash\mathbb{H}$. Geometrically, this is isomorphic to either a line $\mathbb{R}$ or to a circle $\mathbb{R}/\mathbb{Z}$, depending on whether or not $\Gamma \cap N$ is trivial. We push the closed horocycle $(N \cap \Gamma)\backslash N(i)$ in the orthogonal direction a_y, and are concerned with its asymptotic distribution near the boundary, corresponding to $y \to 0$.

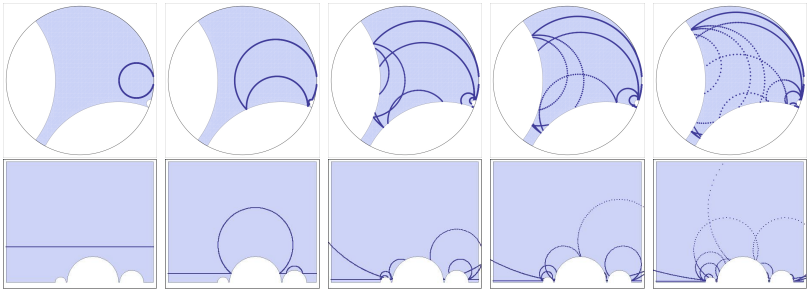

FIGURE 2. Equidistribution of expanding closed horocycles on a Schottky domain in the disk $\mathbb{D}$ and upper half plane $\mathbb{H}$ models.

Let $X = \Gamma\backslash\mathbb{H}$ and consider the Laplace operator $\Delta = -y^2(\partial_{xx} + \partial_{yy})$. By Patterson [Pat76] and Lax-Phillips [LP82], the spectral resolution of Δ acting on $L^2(X)$ consists of only finitely many eigenvalues in the interval $[0, 1/4)$, with the smallest given by $\lambda_0 = \delta(1-\delta)$. Denote the point spectrum below $1/4$ by

$$0 \leq \lambda_0 < \lambda_1 \leq \cdots \leq \lambda_k < 1/4.$$

Let $\phi_0, \ldots, \phi_k$ be the corresponding orthonormal eigenfunctions. Let $s_j > 1/2$ satisfy $\lambda_j = s_j(1-s_j)$, $j = 0, 1, 2, \ldots, k$, so that $s_0 = \delta$.

Theorem 1.11. *Fix notation as above and assume that $(\Gamma \cap N)\backslash N$ is closed. Then for any $\psi \in C_c^\infty(\Gamma\backslash\mathbb{H})$,*

$$\int\limits_{n_x \in (N\cap\Gamma)\backslash N} \psi(x+iy)\, dx = \sum_{j=0}^{k} \langle \psi, \phi_j \rangle \int\limits_{n_x \in (N\cap\Gamma)\backslash N} \phi_j(x+iy)\, dx + O_\varepsilon(y^{\frac{1}{2} - \frac{3}{5}(\delta - \frac{1}{2}) - \varepsilon}),$$

as $y \to 0$. Here the implied constant depends only on a Sobolev norm of ψ, and on $\varepsilon > 0$ which is arbitrary.

Moreover, the integrals above converge, and satisfy

$$\int\limits_{n_x \in (N\cap\Gamma)\backslash N} \phi_j(x+iy)\, dx \sim c_j \cdot y^{1-s_j}, \qquad \text{as } y \to 0,$$

where $c_0 > 0$, and $c_1, \cdots, c_k \in \mathbb{R}$.

Remark 1.12. If Γ is a lattice, then the closedness of $\Gamma \cap N\backslash N$ implies that $\Gamma \cap N\backslash N$ is compact. In this case, Sarnak [Sar81] proved the above result allowing $\psi \in C_c^\infty(\Gamma\backslash G)$ (that is, not requiring K-fixed), and with a best possible error term of

$$y^{\frac{1}{2}}$$

in place of our weaker bound

$$y^{\frac{1}{2} - \frac{3}{5}(\delta - \frac{1}{2})}.$$

1.4. **Bounds for Automorphic Eigenfunctions.** The proof of Theorem 1.11 requires control over the integrals of the eigenfunctions ϕ_j, which *a priori* are only square-integrable. For the base eigenfunction, one has extra structure coming from Patterson theory [Pat76] which makes this control possible. But for the other eigenfunctions, this analysis fails. Nevertheless, the problem of obtaining such

control was solved in the first-named author's thesis [Kon07]. The statement is the following (see also the Appendix).

Theorem 1.13 ([Kon07])**.** *Fix notation as in Theorem 1.11, and assume that the closed horocycle $(N \cap \Gamma)\backslash N$ is infinite. Let $\phi_j \in L^2(\Gamma\backslash\mathbb{H})$ be an eigenfunction of eigenvalue $\lambda_j = s_j(1-s_j) < 1/4$ with $s_j > 1/2$. Then*

$$\phi_j(n_x a_y) \ll_{\phi_j} \left(\frac{y}{x^2+y^2}\right)^{s_j},$$

as $|x| \to \infty$ and $y \to 0$.

1.5. **Organization of the Paper.** In §2 we give some background and elaborate further on Theorem 1.5. For the reader's convenience, in the Appendix we reproduce the proof of Theorem 1.13 from [Kon07], since this reference is not readily available. Equipped with such control, the proof of Theorem 1.11 follows with minor changes from the one given for one dimension higher in [KO08]. We sketch the argument in §3, and use it to prove Theorem 2.6 in §4. In §5, we verify the sieve axioms in Theorem 2.18 and conclude Theorem 2.22. At the end of §5, we derive the explicit values of R, in particular proving Theorem 1.5.

Acknowledgments. The authors wish to express their gratitude to Peter Sarnak for many helpful discussions, and the referee for many detailed comments and insightful suggestions.

2. Background and More on Theorem 1.5

In this section, we elaborate on Theorem 1.5. Let Q be a ternary rational quadratic form which is isotropic over $\mathbb{Q}$. Let $\Gamma < \mathrm{SO}_Q(\mathbb{Z})$ be a finitely generated subgroup with $\delta > 1/2$.

As Q is isotropic over $\mathbb{Q}$, we have a $\mathbb{Q}$-rational covering $\pi : \mathrm{SL}_2 \to \mathrm{SO}_Q$. Hence there exists a finitely generated subgroup, say Γ_0, of $\mathrm{SL}_2(\mathbb{Z})$ such that $\pi(\Gamma_0)$ is a subgroup of Γ with finite index. Since $\mathbf{x}_0\Gamma$ is a union of $\mathbf{x}_i\Gamma_0$ for finitely many primitive Pythagorean triples $\mathbf{x}_i$'s, we may assume without loss of generality that Γ is a finitely generated subgroup of $\mathrm{SL}_2(\mathbb{Z})$.

2.1. **Uniform Spectral Gaps.** For the application to sieving, Theorem 1.8 described in the introduction is insufficient. One requires uniformity along arithmetic progressions; hence we recall the notion of a spectral gap.

Let $\Gamma(q)$ denote the "congruence" subgroup of Γ of level q,

$$\Gamma(q) := \{\gamma \in \Gamma : \gamma \equiv I(q)\}.$$

The inclusion of vector spaces

$$L^2(\Gamma\backslash\mathbb{H}) \subset L^2(\Gamma(q)\backslash\mathbb{H})$$

induces the same inclusion on the spectral resolution of the Laplace operator:

$$\mathrm{Spec}(\Gamma\backslash\mathbb{H}) \subset \mathrm{Spec}(\Gamma(q)\backslash\mathbb{H}).$$

Definition 2.1. The *new spectrum*

$$\mathrm{Spec}_{new}(\Gamma(q)\backslash\mathbb{H})$$

at level q is defined to be the set of eigenvalues below $1/4$ which are in $\mathrm{Spec}(\Gamma(q)\backslash\mathbb{H})$ but not in $\mathrm{Spec}(\Gamma\backslash\mathbb{H})$.

Definition 2.2. A number θ in the interval $1/2 < \theta < \delta$ is called a *spectral gap* for Γ if there exists a *ramification number* $\mathfrak{B} \geq 1$ such that for any square-free

$$q = q'q'' \quad \text{with} \quad q' \mid \mathfrak{B} \text{ and } (q'', \mathfrak{B}) = 1,$$

we have

$$\mathrm{Spec}(\Gamma(q)\backslash\mathbb{H})_{new} \cap (0, \theta(1-\theta)) \quad \subset \quad \mathrm{Spec}(\Gamma(q')\backslash\mathbb{H})_{new}.$$

That is, the eigenvalues below $\theta(1-\theta)$ which are new for $\Gamma(q)$ are coming from the "bad" part q' of q. As $\mathfrak{B}$ is a fixed integer depending only on Γ, there are only finitely many possibilities for its divisors q'. Hence our definition of a gap agrees with the more standard one, since there is some other θ' for which no new eigenvalues appear below $\theta'(1-\theta')$. (The point is that our θ may sometimes be effectively known, whereas θ' is not.)

Collecting the results in [BG08, BGS10, BGS09] and the extension from prime to square-free of [Gam02], we have:

Theorem 2.3 ([Gam02, BG08, BGS10])**.**

(1) *For any finitely generated* $\Gamma < \mathrm{SO}_Q(\mathbb{Z})$ *with critical exponent* $\delta > 1/2$*, there exists a spectral gap*
$$1/2 < \theta < \delta.$$
(2) *If moreover* $\delta > 5/6$*, then there is a spectral gap with*
$$\theta = 5/6.$$

Remark 2.4. In [Gam02], part (2) above is proved under the further restriction that q is prime, but the proof (a major ingredient of which is a strong upper bound for the number of lattice points in a ball satisfying a congruence restriction) extends easily to the square-free case.

2.2. **Counting with Weights uniformly in Level.**

Allowing some "smoothing", one can count uniformly in cosets of orbits of level q with explicit error terms. We fix a non-zero vector $\mathbf{x}_0 \in \mathbb{Z}^3$ with $Q(\mathbf{x}_0) = 0$ and set

$$\mathcal{O} = \mathbf{x}_0\Gamma.$$

We denote by N_0 the stabilizer subgroup of $\mathbf{x}_0$ in $G := \mathrm{SL}_2(\mathbb{R})$. Then

$$N_0 = g_0 N g_0^{-1}$$

for some $g_0 \in \mathrm{SL}_2(\mathbb{Q})$, where N denotes the upper triangular subgroup of G.

Set $K_0 := g_0\,\mathrm{SO}_2(\mathbb{R})g_0^{-1}$ and let ψ be a non-negative, smooth, K_0-invariant function on G with $\int \psi dg = 1$ and with compact support which injects to $\Gamma\backslash G$.

Denote by B_T a K_0-invariant norm ball in $\mathbb{R}^3$ about the origin with radius T.

Definition 2.5. The weight $\xi_T : \mathbb{R}^3 \to \mathbb{R}_{\geq 0}$ is defined as follows:

$$\xi_T(\mathbf{x}) = \int_G \chi_T(\mathbf{x}g)\psi(g)dg$$

where χ_T denotes the characteristic function of B_T.

The sum of ξ_T over $\mathcal{O}$ is precisely a smoothed count for $\#\mathcal{O} \cap B_T$ satisfying:

$$\sum_{\mathbf{x}\in\mathcal{O}} \xi_T(\mathbf{x}) \asymp \#\mathcal{O} \cap B_T.$$

Theorem 2.6. *Let θ be a spectral gap for Γ.*

(1) *As $T \to \infty$,*

$$\Xi(T) := \sum_{\mathbf{x} \in \mathcal{O}} \xi_T(\mathbf{x}) \sim c \cdot T^{\delta},$$

for some $c > 0$.

(2) *For square-free q, write $q = q'q''$ with $q' \mid \mathfrak{B}$ and $(q'', \mathfrak{B}) = 1$. Let $\Gamma_1(q)$ be any group satisfying*

$$\Gamma(q) \subset \Gamma_1(q) \subset \Gamma.$$

Let N_0 be the stabilizer of $\mathbf{x}_0$ in G, and assume that

$$\Gamma_1(q) \cap N_0 = \Gamma \cap N_0.$$

Fix any $\gamma_0 \in \Gamma$ and $\varepsilon > 0$. Then as $T \to \infty$,

$$\begin{aligned} \sum_{\mathbf{x} \in \mathbf{x}_0 \Gamma_1(q)} \xi_T(\mathbf{x}\gamma_0) &= \frac{1}{[\Gamma : \Gamma_1(q)]} \cdot \Big(\Xi(T) + \mathcal{E}(T, q', \gamma_0) \Big) \\ &\quad + O_\varepsilon \Big(T^{\theta+\varepsilon} + T^{\frac{1}{2}+\frac{3}{5}(\delta-\frac{1}{2})+\varepsilon} \Big), \end{aligned}$$

where the implied constant does not depend on q or γ_0. Here the error term satisfies

$$\mathcal{E}(T, q', [\gamma_0]) \ll T^{\delta-\zeta}$$

for some fixed $\zeta > 0$, does not depend on q'', and depends only on the class $[\gamma_0]$ in $\Gamma_1(q')\backslash\Gamma$.

Remark 2.7. Assuming that $\Gamma \cap N_0$ is a lattice in N_0, [Kon09] gives the above uniform count with the last error term

$$T^{\frac{1}{2}+\frac{3}{5}(\delta-\frac{1}{2})+\varepsilon}$$

replaced by a best possible error of

$$T^{\frac{1}{2}} \log T.$$

2.3. **Zariski Density of Orbits of Pythagorean Triples.**

For simplicity, we will use the notation $\mathcal{P}(R)$ to denote the set of all integers having at most R prime divisors, counted with multiplicity.

Let $\mathbf{x}_0 \in \mathbb{Z}^3$ be a non-zero Pythagorean triple on the cone

$$Q(\mathbf{x}) = x^2 + y^2 - z^2 = 0$$

and $\Gamma < \mathrm{SO}_Q(\mathbb{Z})$ a non-elementary finitely generated subgroup. Set

$$\mathcal{O} := \mathbf{x}_0 \cdot \Gamma.$$

Given a polynomial F which is integral on $\mathcal{O}$, our goal is to find "small" values for $R = R(\mathcal{O}, F)$, for which F "often" has at most R prime factors.

In fact, when studying such thin orbits, a better notion of "often" is not "infinitely often", but instead one should require Zariski density. That is, the set of $\mathbf{x} \in \mathcal{O}$ for which $F(\mathbf{x}) \in \mathcal{P}(R)$ should not lie on a proper subvariety of the smallest variety containing $\mathcal{O}$. We illustrate this condition with the following examples.

2.3.1. *Example I: Area.*

Recall that given any integral Pythagorean triple $\mathbf{x} = (x, y, z)$ which is also primitive (that is, there is no common divisor of x, y and z), there exist coprime integers u and v of opposite parity (one even, one odd) such that, possibly after switching or negating x and y, we have the ancient parametrization

$$x = u^2 - v^2, \quad y = 2u\,v, \quad z = u^2 + v^2.$$

In fact, this is just a restatement of the group homomorphism $\mathrm{SL}_2(\mathbb{R}) \to \mathrm{SO}(2,1)$ given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \begin{pmatrix} \frac{a^2-b^2-c^2+d^2}{2} & ac-bd & \frac{a^2-b^2+c^2-d^2}{2} \\ ab-cd & bc+ad & ab+cd \\ \frac{a^2+b^2-c^2-d^2}{2} & ac+bd & \frac{a^2+b^2+c^2+d^2}{2} \end{pmatrix}$$

where $\mathrm{SL}(2,\mathbb{R})$ acts on (u, v) and $\mathrm{SO}(2,1)$ acts on (x, y, z).

Consider the "area" $\frac{1}{2}xy$ of the triple $\mathbf{x}$ (which may be negative). It is elementary that the area is always divisible by 6, so the function

$$F_{\mathcal{A}}(\mathbf{x}) := \frac{1}{12}xy = \frac{1}{6}(u+v)(u-v)u\,v \tag{2.8}$$

is integer-valued on $\mathcal{O}$.

Remark 2.9. As above, we insist that the polynomial F is integral on $\mathcal{O}$, but it need not necessarily have integer coefficients.

As (2.8) has four irreducible components, it is easy to show that there are only finitely many triples $\mathbf{x}$ for which $F_{\mathcal{A}}(\mathbf{x}) \in \mathcal{P}(2)$, that is, the product of at most two primes. Restricting to a subvariety such as

$$u = v + 1,$$

it follows conjecturally from the Hardy-Littlewood k-tuple conjectures [HL22] that

$$\frac{1}{12}xy = \frac{1}{6}(u+v)(u-v)u\,v = \frac{1}{6}(2v+1)\cdot 1\cdot(v+1)\cdot v$$

will be the product of three primes for infinitely many v.

Since the set of triples generated in this way lies on a subvariety, it is not Zariski dense. On the other hand, it was recognized in [BGS10] that the recent work of Green and Tao [GT10] proves the infinitude and Zariski density of the set of all primitive Pythagorean triples $\mathbf{x}$ for which $F_{\mathcal{A}}(\mathbf{x}) \in \mathcal{P}(4)$, that is, has at most four prime factors.

Remark 2.10. The results of Green-Tao do not apply to thin orbits, and neither do the conjectures of Hardy-Littlewood. Indeed, we conjecture that that if $\mathcal{O}$ is thin and Γ has no unipotent elements (which would furnish an affine injection into $\mathcal{O}$), then there are only *finitely-many* points $\mathbf{x}$ for which $F_{\mathcal{A}}(\mathbf{x}) \in \mathcal{P}(3)$! On the other hand, allowing 4 primes should lead to a Zariski dense set of triples $\mathbf{x}$. Below, we exhibit certain thin orbits for which there is a Zariski dense set of $\mathbf{x}$ with $F_{\mathcal{A}}(\mathbf{x}) \in \mathcal{P}(25)$.

Remark 2.11. The critical number, 4, of prime factors above is related to the *sieve dimension* for this pair $(\mathcal{O}, F)$. We return to this issue shortly, cf. Remark 2.13.

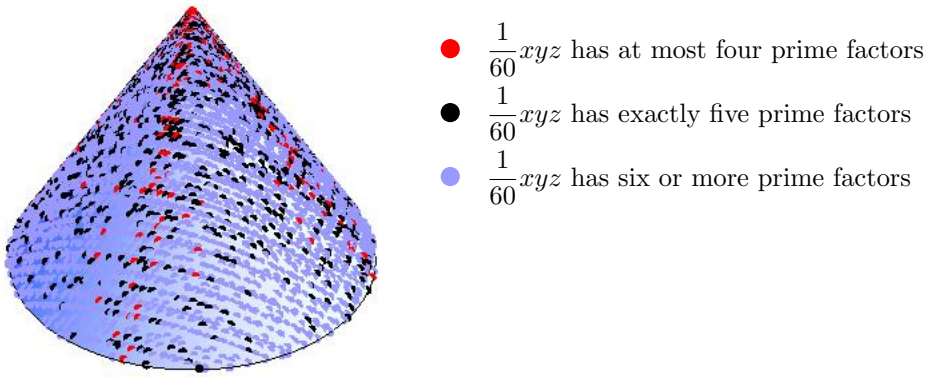

FIGURE 3. The full orbit of all primitive Pythagorean triples.

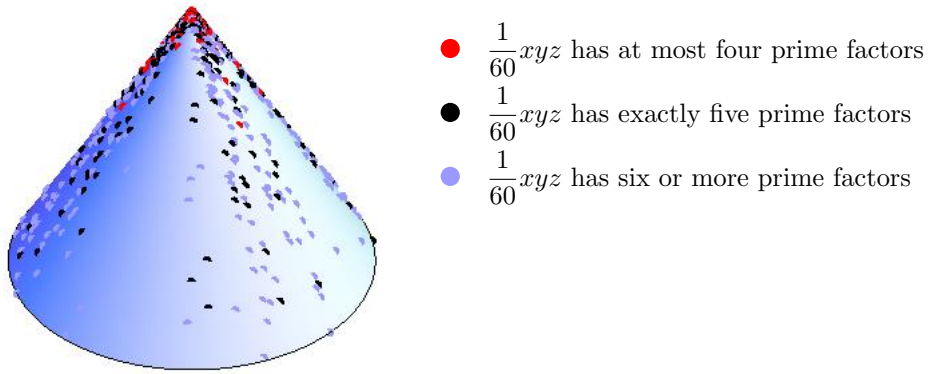

FIGURE 4. A thin orbit $\mathcal{O}$ of Pythagorean triples.

2.3.2. *Example II: Product of Coordinates.* Consider now the product of coordinates xyz for triples $\mathbf{x} \in \mathcal{O}$. It is elementary that xyz is divisible by 60, so the function

$$F_{\mathcal{C}}(\mathbf{x}) := \frac{1}{60}xyz = \frac{1}{30}(u+v)(u-v)u\,v(u^2+v^2) \tag{2.12}$$

is integer-valued.

Now we note that as (2.12) has five irreducible components (and the *sieve dimension* is five). Therefore there are only finitely many triples $\mathbf{x}$ for which $F_{\mathcal{C}}(\mathbf{x}) \in \mathcal{P}(3)$. Restricting to a subvariety such as

$$u = v + 3,$$

it follows conjecturally from Schinzel's Hypothesis H [SS58] that

$$\frac{1}{30}(u+v)(u-v)u\,v(u^2+v^2) = \frac{1}{30}(2v+3)\cdot(3)\cdot(v+3)\cdot v\cdot(2v^2+6v+9)$$

will be the product of four primes for infinitely many v. Again, this set is not Zariski dense in the cone. See Figure 3, where it is clear that such points frolic near the x or y axes.

On the other hand, it is a folklore conjecture that the set of triples $\mathbf{x}$ for which $F_{\mathcal{C}}(\mathbf{x}) \in \mathcal{P}(5)$ spreads out in every direction. For the full orbit of all primitive Pythagorean triples (rather than a thin one), the best known bound for the number of prime factors in $F_{\mathcal{C}}(\mathbf{x})$ follows from the Diamond-Halberstam-Richert sieve [DHR88, DH97]. Their work shows that $F_{\mathcal{C}}(\mathbf{x}) \in \mathcal{P}(17)$ infinitely often.[3]

Again, when the orbit $\mathcal{O}$ is thin without affine injections, we conjecture that there will be only finitely many $\mathbf{x}$ for which $F_{\mathcal{C}}(\mathbf{x}) \in \mathcal{P}(4)$ whereas five factors will be Zariski dense. Compare Figure 3 to Figure 4. We will exhibit certain thin orbits for which there is a Zariski dense set of $\mathbf{x}$ with $F_{\mathcal{C}}(\mathbf{x}) \in \mathcal{P}(29)$.

Remark 2.13. Sieve dimension is not merely a function of the polynomial F but really depends on the pair $(\mathcal{O}, F)$. In the related recent work [LS07], Liu and Sarnak consider $F_{\mathcal{C}}(\mathbf{x}) = xyz$, where the orbit $\mathcal{O} = \mathbf{x}_0 \cdot \Gamma$ is generated from a point $\mathbf{x}_0 \in \mathbb{Z}^3$ on a one- or two-sheeted hyperboloid $Q(\mathbf{x}) = t$, where $t \neq 0$ and Q is an indefinite integral ternary quadratic form which is anisotropic. Then the spin group of $G = \mathrm{SO}_Q(\mathbb{R})$ consists of the elements of norm one in a quaternion division algebra over $\mathbb{Q}$, and Γ is the set of *all* such integral elements.

In particular, their orbit is full, whereas the focus of this paper is on thin orbits. A common feature, though, is that there do not exist non-constant polynomial parametrizations of points in $\mathcal{O}$ (in our case this corresponds precisely to $N_0 \cap \Gamma$ being trivial).

The sieve dimension for Liu-Sarnak's pair $(\mathcal{O}, F_{\mathcal{C}})$ is 3 (whereas in our case, the same function $F = F_{\mathcal{C}}$ has sieve dimension 5), and they prove the Zariski density of the set of points $\mathbf{x} \in \mathcal{O}$ for which $F_{\mathcal{C}}(\mathbf{x})$ is in $\mathcal{P}(26)$. The precise definition of sieve dimension is given in Definition 2.15.

2.4. **The Diamond-Halberstam-Richert Weighted Sieve.**

Let $\mathcal{A} = \{a_n\}$ be a sequence of non-negative real numbers, all but finitely many of which are zero. Recall that $\mathcal{P}(R)$ is the set of R-almost primes.

Sieve theory allows one to extract estimates for

$$\sum_{n \in \mathcal{P}(R)} a_n,$$

that is, the number of R-almost primes in the sequence $\mathcal{A}$ (counted with multiplicity), from knowledge of the distribution of $\mathcal{A}$ along certain arithmetic progressions. For $q \geq 1$ a square-free integer, let

$$\mathcal{A}_q := \{a_n \in \mathcal{A} : n \equiv 0(q)\} \quad \text{and}$$

and

$$|\mathcal{A}_q| := \sum_{n \equiv 0(q)} a_n.$$

[3] In fact they restrict to a subvariety in deriving the numer 17.

Assume there exists an approximation $\mathcal{X}$ to $|\mathcal{A}| := \sum_n a_n$ and a non-negative multiplicative function $g(q)$ so that $g(q)\mathcal{X}$ is an approximation to $|\mathcal{A}_q|$. Assume that $g(1) = 1$, $g(q) \in [0, 1)$ for $q > 1$, and that for constants $K \geq 2$, $\kappa \geq 1$ we have the local density bound

$$\prod_{z_1 \leq p \leq z} (1 - g(p))^{-1} \leq \left(\frac{\log z}{\log z_1}\right)^{\kappa} \left(1 + \frac{K}{\log z_1}\right) \tag{2.14}$$

for any $2 \leq z_1 < z$.

Definition 2.15. The number κ appearing in (2.14) is called the *sieve dimension*. (Note that it is not unique, as any larger value also satisfies (2.14); in practice one typically takes the least allowable value.)

We require that the remainder terms

$$r_q := |\mathcal{A}_q| - g(q)\mathcal{X}$$

be small an average, that is, for some constants $\tau \in (0, 1)$ and $A \geq 1$,[4]

$$\sum_{\substack{q < \mathcal{X}^{\tau}(\log \mathcal{X})^{-A} \\ q \text{ squarefree}}} 4^{\nu(q)} |r_q| \ll \frac{\mathcal{X}}{\log^{\kappa+1} \mathcal{X}}. \tag{2.16}$$

Finally, we introduce a parameter μ which controls the number of terms in $\mathcal{A}$ which are non-zero. Precisely, we require that

$$\max\{n \geq 1 : a_n \neq 0\} \leq \mathcal{X}^{\tau\mu}. \tag{2.17}$$

We now state

Theorem 2.18 ([DHR88, DH97], see Thm 11.1 of [DH08])**.** *Let $\mathcal{A}$, z, $\mathcal{X}$, g, κ, μ and τ be as described above.*

(1) *Let $\sigma_\kappa(u)$ be the continuous solution of the differential-difference problem:*

$$\begin{cases} u^{-\kappa}\sigma(u) = A_\kappa^{-1}, & \text{for } 0 < u \leq 2,\ A_\kappa = (2e^\gamma)^\kappa \Gamma(\kappa + 1), \\ (u^{-\kappa}\sigma(u))' = -\kappa u^{-\kappa-1}\sigma(u-2), & \text{for } u > 2, \end{cases} \tag{2.19}$$

where γ is the Euler constant. Then there exist two numbers α_κ and β_κ satisfying $\alpha_\kappa \geq \beta_\kappa \geq 2$ such that the following simultaneous differential-difference system has continuous solutions $F_\kappa(u)$ and $f_\kappa(u)$ which satisfy

$$F_\kappa(u) = 1 + O(e^{-u}), \quad f_\kappa(u) = 1 + O(e^{-u}),$$

and F_κ (resp. f_κ) decreases (resp. increases) monotonically towards 1 as $u \to \infty$:

$$\begin{cases} F(u) = 1/\sigma_\kappa(u), & \text{for } 0 < u \leq \alpha_\kappa, \\ f(u) = 0, & \text{for } 0 < u \leq \beta_\kappa, \\ (u^\kappa F(u))' = \kappa u^{\kappa-1} f(u-1), & \text{for } u > \alpha_\kappa, \\ (u^\kappa f(u))' = \kappa u^{\kappa-1} F(u-1), & \text{for } u > \beta_\kappa. \end{cases} \tag{2.20}$$

[4] Recall that $\nu(q)$ denotes the number of prime factors of q.

(2) *We have*

$$\sum_{n\in\mathcal{P}(R)} a_n \gg \mathcal{X} \prod_{p<\mathcal{X}^{1/v}} (1-g(p)),$$

provided that

$$\tau^{-1} < u \leq v, \beta_\kappa < \tau v,$$

and

$$R > \tau\mu u - 1 + \frac{\kappa}{f_\kappa(\tau v)} \int_1^{v/u} F_\kappa(\tau v - s)\left(1 - \frac{u}{v}s\right)\frac{ds}{s}. \tag{2.21}$$

2.5. **Statement of the Main Theorem.**

The following is the main result of this paper.

Theorem 2.22. *Let $Q(\mathbf{x}) = x^2+y^2-z^2$, $\mathbf{x}_0 \in \mathbb{Z}^3$ a non-zero vector with $Q(\mathbf{x}_0) = 0$, and $\Gamma < \mathrm{SO}_Q(\mathbb{Z})$ a finitely generated subgroup with $\delta > 1/2$. Let θ denote a spectral gap for Γ. Let F be a polynomial which is integral on $\mathcal{O}$, such that the pair $(\mathcal{O}, F)$ is strongly primitive.*

Then the following hold:

(1) *There are infinitely many $\mathbf{x} \in \mathcal{O}$ such that $F(\mathbf{x})$ has at most R prime factors, where $R(\mathcal{O}, F)$ is given by* (2.21) *with*

$$\tau < \min\left(\frac{\delta-\theta}{2\delta}, \frac{\delta-\frac{1}{2}}{5\delta}\right), \qquad \text{and} \qquad \mu > \max\left(\frac{2}{\delta-\theta}, \frac{5}{\delta-1/2}\right). \tag{2.23}$$

(2) *Under the further assumption that $N_0 \cap \Gamma$ is a lattice in N_0 for $N_0 = \mathrm{Stab}_G(\mathbf{x}_0)$, the bounds in* (2.23) *improve to*

$$\tau < \frac{\delta-\theta}{2\delta}, \qquad \text{and} \qquad \mu > \frac{2}{\delta-\theta}. \tag{2.24}$$

(3) *Denoting by κ be the sieve dimension of $(\mathcal{O}, F)$,*

$$\#\{\mathbf{x} \in \mathcal{O} : \|\mathbf{x}\| < T, F(\mathbf{x}) \in \mathcal{P}(R)\} \asymp \frac{T^\delta}{(\log T)^\kappa}. \tag{2.25}$$

In particular, this set is Zariski dense in the cone $Q = 0$.

Remark 2.26. Zariski-density follows from the count of the number of points produced in (2.25). See the proof of Prop. 3.2 in [BGS10].

2.6. **Explicit Values of $R(\mathcal{O}, F)$.**

One must still work somewhat to obtain actual values of R from Theorem 2.22. We now state the smallest values of $R(\mathcal{O}, F)$ which are achieved from Theorem 2.22, at the expense of requiring the critical exponent δ to be close to 1.

The first statement we give is unconditional. Assuming $\delta > 5/6$ and using Gamburd's spectral gap $\theta = 5/6$ from Theorem 2.3, the bounds (2.23) and (2.24) are equivalent. Hence assuming $N_0 \cap \Gamma$ is a lattice in N_0 and using the optimal error terms in [Kon09] does not improve the final values of R. Said another way, the fact that our counting theorem is not optimal does not hurt the final values of R, unconditionally.

In the next three theorems, we keep the notation $Q, \Gamma, \mathcal{O}$ from Theorem 2.22. Let

$$F_{\mathcal{H}}(\mathbf{x}) = z, \qquad F_{\mathcal{A}}(\mathbf{x}) = \frac{1}{12}xy, \qquad F_{\mathcal{C}}(\mathbf{x}) = \frac{1}{60}xyz,$$

and assume the pair $(\mathcal{O}, F)$ is strongly primitive.

The following is the same as Theorem 1.5:

Theorem 2.27. *Let Γ have critical exponent*

$$\delta_{\mathcal{H}} > 0.9992, \qquad \delta_{\mathcal{A}} > 0.99995, \qquad \delta_{\mathcal{C}} > 0.99677.$$

Then the proportion of $\mathbf{x} \in \mathcal{O}$ with $F(\mathbf{x}) \in \mathcal{P}(R)$ with $\|\mathbf{x}\| < T$ is

$$\asymp \frac{1}{(\log T)^{\kappa}},$$

where

$$\kappa_{\mathcal{H}} = 1, \qquad \kappa_{\mathcal{A}} = 4, \qquad \kappa_{\mathcal{C}} = 5,$$

and

$$\boxed{R_{\mathcal{H}} = 14, \qquad R_{\mathcal{A}} = 25, \qquad R_{\mathcal{C}} = 29.}$$

We now observe the effect of the worse error term on the quality of $R(\mathcal{O}, F)$, conditioned on an improved spectral gap. Assuming $N_0 \cap \Gamma$ is a lattice in N_0, the counting theorem in [Kon09] gives an optimal error term, which together with our sieve analysis gives the following.

Theorem 2.28. *Assume that $N_0 \cap \Gamma$ is a lattice in N_0. Then if*

$$\delta_{\mathcal{H}} > 0.9265, \qquad \delta_{\mathcal{A}} > 0.98805, \qquad \delta_{\mathcal{C}} > 0.981675,$$

the conclusion of Theorem 2.27 holds with

$$\boxed{R_{\mathcal{H}} = 6, \qquad R_{\mathcal{A}} = 14, \qquad R_{\mathcal{C}} = 17.}$$

Lastly, we demonstrate the values of R which we can obtain without assuming that $\Gamma \cap N_0$ is a lattice in N_0.

Theorem 2.29. *We make no assumptions on $N_0 \cap \Gamma$. Then if*

$$\delta_{\mathcal{H}} > 0.991, \qquad \delta_{\mathcal{A}} > 0.97895, \qquad \delta_{\mathcal{C}} > 0.99905,$$

the conclusion of Theorem 2.27 holds with

$$\boxed{R_{\mathcal{H}} = 12, \qquad R_{\mathcal{A}} = 23, \qquad R_{\mathcal{C}} = 26.}$$

Theorems 2.27, 2.28, and 2.29 follow from Theorem 2.22 and Table 1, the computation of which is discussed in §5.3.

Remark 2.30. Gamburd [Gam09] has informed us that in general it is not true that a spectral gap can be arbitrarily close to 1/2, in that he has observed counterexamples to what would be an analogue of the "Selberg 1/4 conjecture." So the above Theorems 2.28 and 2.29 may only serve to illustrate the relative dependence of R on the our present error term in Theorem 2.6 (2), versus the optimal error term obtained in [Kon07, Kon09].

3. Equidistribution of Expanding Closed Horocycles

Let $G = \mathrm{SL}(2,\mathbb{R})$. We keep the notation for N, A, K, n_x, a_y, etc., from (1.10). We have the Cartan decomposition

$$G = KA^+K$$

where $A^+ := \{a_y : 0 < y \leq 1\}$, as well as the Iwasawa decomposition $G = NAK$.

Let $\Gamma < G$ be a discrete finitely generated subgroup with critical exponent $\delta > 1/2$. Assume the horocycle $(N \cap \Gamma)\backslash N$ is closed in $\Gamma\backslash G$. In this section, we prove Theorem 1.11.

3.1. **Automorphic Representations and Spectral Bounds.**

Let $1/2 < s < 1$, and consider the character χ_s on the upper-triangular subgroup $B := NA$ of G defined by

$$\chi_s(na_y) = y^s$$

where $a_y = \mathrm{diag}(\sqrt{y}, \sqrt{y}^{-1})$ is as before, and $n \in N$.

The unitarily induced representation $(\pi_s := \mathrm{Ind}_B^G \chi_s, V_s)$ admits a K-invariant unit vector, say v_s, unique up to a scalar multiplication.

By the theory of spherical functions,

$$f_s(g) := \langle \pi_s(g)v_s, v_s\rangle = \int_K v_s(kg)dk$$

is the unique bi K-invariant function of G with $f_s(e) = 1$ and with $\mathcal{C}f_s = s(1-s)f_s$ where $\mathcal{C}$ is the Casimir operator of G. Moreover, there exist some $c_s > 0$ and $\alpha > 0$ such that for all y small

$$f_s(a_y) = c_s \cdot y^{1-s}(1 + O(y^\alpha))$$

by [GV88, 4.6].

Since the Casimir operator is equal to the Laplace operator Δ on K-invariant functions, this immediately implies the following.

Theorem 3.1. *Let $\phi_s \in L^2(\Gamma\backslash G)^K \cap C^\infty(\Gamma\backslash G)$ satisfy $\Delta\phi_s = s(1-s)\phi_s$ and $\|\phi_s\|_2 = 1$. Then there exist $c_s > 0$ and $\alpha > 0$ such that for all $y \ll 1$,*

$$\langle a_y\phi_s, \phi_s\rangle_{L^2(\Gamma\backslash G)} = c_s \cdot y^{1-s}(1 + O(y^\alpha)).$$

The irreducible unitary representations of G with K-fixed vector consist of principal series and the complementary series. We use the parametrization of $s \in \{1/2 + i\mathbb{R}\} \cup [1/2, 1]$ so that the vertical line $1/2 + i\mathbb{R}$ corresponds to the principal series and the complementary series is parametrized by $1/2 < s \leq 1$ with $s = 1$ corresponding to the trivial representation.

Let $\{X_1, X_2, X_3\}$ denote an orthonormal basis of the (real) Lie algebra of G with respect to an Ad-invariant scalar product. For $f \in C^\infty(\Gamma\backslash G) \cap L^2(\Gamma\backslash G)$, we consider the following Sobolev norm $\mathcal{S}_m(f)$:

$$\mathcal{S}_m(f) = \max\{\|X_{i_1} \cdots X_{i_m}(f)\|_2 : 1 \leq i_j \leq 3\}.$$

Proposition 3.2. *Let (π, V) be a representation of G which does not weakly contain any complementary series representation V_s. Then for any $\varepsilon > 0$, and any smooth vectors $v_1, v_2 \in V$,*

$$|\langle \pi(a_y)v_1, v_2\rangle| \ll_\varepsilon y^{1/2-\varepsilon} \cdot \|\mathcal{S}_1(v_1)\| \cdot \|\mathcal{S}_1(v_2)\|, \qquad \text{as} \qquad y \to 0.$$

Here the implied constant depends only on $\epsilon > 0$ (and is independent of π).

Proof. The assumption on π implies that π is a tempered representation. Hence by [CHH88, Thm.2], for any K-finite vectors v_1 and v_2,

$$|\langle \pi(a_y)v_1, v_2\rangle| \leq d_{v_1} d_{v_2} \|v_1\| \, \|v_2\|, \Xi(a_y)$$

where d_{v_i} is the dimension of the K-span of v_i and Ξ is the Harish-Chadra function of $\mathrm{SL}_2(\mathbb{R})$: $\Xi(a_y) = \int_K \beta^{-1/2}(a_y k)dk$ where β is the left-modular function of B. To obtain our claim on smooth vectors, we proceed as in [War72, Ch. 4]. This step is well-known but we add the details for the sake of completeness. Write $\pi = \oplus_{n\in\mathbb{Z}}\pi_n$ where π_n is the one dimensional space where K acts as scalar $k_\theta.\pi_n = e^{in\theta}\pi_n$ for $k_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$. We write v_i as the sum $\sum_{n\in\mathbb{Z}} v_{in}$ with $v_{in} \in \pi_n$. Choosing an element X of the Lie algebra of K, we have $\|X.v_{in}\| = c_0|n|\|v_{in}\|$ for some uniform constant c_0 depending only on the choice of X. Then

$$\begin{aligned}
&\langle v_1, a_y v_1\rangle = \sum_{n,m} \langle v_{1n}, a_y v_{2m}\rangle \\
&\leq \Xi(a_y) \sum_{n,m} \|v_{1n}\| \cdot \|v_{2m}\| \\
&= \Xi(a_y)(\sum_n \|v_{1n}\|) \cdot (\sum_m \cdot \|v_{2m}\|) \\
&= \Xi(a_y) c_0^2 4 (\sum_n (|n|+1)^{-1}\|X.v_{1n}\|)(\sum_m (|m|+1)^{-1}\|X.v_{2m}\|) \\
&= \Xi(a_y) c_0^2 4 (\sum_n (|n|+1)^{-2})^{1/2}\|X.v_1\|)(\sum_m (|m|+1)^{-2})^{1/2}\|X.v_2\|) \ll y^{1/2-\epsilon}\mathcal{S}_1(v_1)\mathcal{S}_2(v_2)
\end{aligned}$$

since $\Xi(a_y) \ll_\epsilon y^{1/2-\epsilon}$. □

3.2. **Approximations to Integrals over closed horocycles.**

As $\delta > 1/2$, there exists a unique positive L^2-eigenfunction ϕ_0 of the Laplace operator $\Delta = -y^2(\partial_{xx} + \partial_{yy})$ on $\Gamma\backslash\mathbb{H}$ with smallest eigenvalue $\delta(1-\delta)$ and of unit norm [Pat76]. The spectrum of Δ acting on $L^2(\Gamma\backslash\mathbb{H})$ has finitely many discrete points below $1/4$ and is purely continuous above $1/4$ [LP82].

Order the other discrete eigenvalues $\lambda_j = s_j(1-s_j)$, with $s_j > 1/2$, $j = 1, \ldots, k$ and let ϕ_j denote the corresponding eigenfunction with $\|\phi_j\| = 1$. For uniformity of notation, set $s_0 := \delta$.

The map $g \mapsto g(i)$ gives an identification of G/K with $\mathbb{H}$. Below a K-invariant function ϕ of G may be considered as a function on the upper half plane $\mathbb{H}$ by $\phi(x+iy) := \phi(n_x a_y)$ and vice versa.

Let dg denote the Haar measure of G given by

$$d(n_x a_y k) = y^{-2}dxdydk$$

where dk is the probability Haar measure on K, and dx and dy are Lebesque measures. Hence for $\psi_1, \psi_2 \in L^2(\Gamma\backslash G)^K$,

$$\langle \psi_1, \psi_2\rangle = \int_{\Gamma\backslash G} \psi_1(g)\overline{\psi_2(g)}\, dg = \int_{\Gamma\backslash\mathbb{H}} \psi_1(x+iy)\overline{\psi_2(x+iy)}\, dx\frac{dy}{y^2}.$$

Proposition 3.3. *For any smooth $\psi_1 \in L^2(\Gamma\backslash G)^K$ and $\psi_2 \in C_c^\infty(\Gamma\backslash G)$,*

$$\langle a_y\psi_1, \psi_2\rangle = \sum_{j=0}^{k}\langle\psi_1,\phi_j\rangle\langle a_y\phi_j,\psi_2\rangle + O_\varepsilon(y^{1/2-\varepsilon}\mathcal{S}_1(\psi_1)\cdot\mathcal{S}_1(\psi_2)).$$

as $y \to 0$. Here the implied constant is independent of Γ.

Proof. Write

$$L^2(\Gamma\backslash G) = W_{\lambda_0}\oplus\cdots\oplus W_{\lambda_k}\oplus V$$

where W_{λ_j} is isomorphic as a G-representation to the complementary series representation V_{s_j}, $\lambda_j = s_j(1-s_j)$, and V is tempered. Write

$$\psi_1 = \langle\psi_1,\phi_0\rangle\phi_0 + \cdots + \langle\psi_1,\phi_k\rangle\phi_k + \psi_1^\perp.$$

Since the ϕ_j's are K-invariant, we have $\psi_1^\perp \in V^K$. Hence by Proposition 3.2, for any $\varepsilon > 0$ and $y \ll 1$,

$$\begin{aligned}\langle a_y\psi_1,\psi_2\rangle &= \sum_{j=0}^{k}\langle\psi_1,\phi_j\rangle\langle a_y\phi_j,\psi_2\rangle + \langle a_y\psi_1^\perp,\psi_2\rangle\\ &= \sum_{j=0}^{k}\langle\psi_1,\phi_j\rangle\langle a_y\phi_j,\psi_2\rangle + O_\varepsilon(y^{1/2-\varepsilon}\mathcal{S}_1(\psi_1)\cdot\mathcal{S}_1(\psi_2))\end{aligned}$$

since $\mathcal{S}_1(\psi_1^\perp) \ll \mathcal{S}_1(\psi_1)$. □

The main goal of this subsection is to study the averages of the ϕ_j's along the translates $(N\cap\Gamma)\backslash Na_y$.

We first need to recall some geometric facts. The limit set $\Lambda(\Gamma)$ of Γ is the set of all accumulation points of an orbit $\Gamma(z_0)$ for some $z_0 \in \mathbb{H}$. As Γ is discrete, it easily follows that $\Lambda(\Gamma)$ is a subset of $\partial_\infty(\mathbb{H}) = \mathbb{R}\cup\{\infty\}$.

Geometrically, $N(i) \subset \mathbb{H}$ is a horocycle based at ∞. Since Γ is finitely generated, the closedness of its projection to $\Gamma\backslash\mathbb{H}$ implies either that ∞ is a parabolic fixed point, that is, $N\cap\Gamma$ is non-trivial, or that ∞ lies outside the limit set $\Lambda(\Gamma)$ [Dal00].

We define for each $0 \le j \le k$

$$\phi_j^N(a_y) := \int_{\Gamma\cap N\backslash N}\phi_j(na_y)dn. \tag{3.4}$$

Theorem 3.5. *For any $j = 0,\dots,k$, the integrals in (3.4) converge absolutely. Moreover there exist constants c_j and d_j, depending on ϕ_j, such that*

$$\phi_j^N(a_y) = c_j y^{1-s_j} + d_j y^{s_j}.$$

Furthermore, $c_0 > 0$.

Proof. We first establish the convergence of the integral in (3.4). If ∞ is a parabolic fixed point, that is, if $N\cap\Gamma$ is a lattice in N, then the domain of this integral is compact, and of course the eigenfunctions of the Laplace operator are bounded, so we are done.

On the other hand, if $\infty \notin \Lambda(\Gamma)$, then Theorem A.1 applies, that for y fixed,

$$\phi_j(n_xa_y) \ll_y (1+x^2)^{-s_j}.$$

Then the integral $\int_{-\infty}^{\infty}\phi_j(n_xa_y)dx$ clearly converges, since $s_j > 1/2$.

From

$$\Delta\phi_j = s_j(1-s_j)\phi_j,$$

it follows that

$$-y^2\frac{\partial^2}{\partial y^2}\phi_j^N = s_j(1-s_j)\phi_j^N.$$

The two independent solutions to this equation are y^{s_j} and y^{1-s_j}.

Lastly, we must demonstrate that $c_0 > 0$. If $\infty \notin \Lambda(\Gamma)$, this is done as in [Kon07] (see also [KO08, Equation (4.11)]), by proving explicitly that

$$\phi_0^N(a_y) = c_0\ y^{1-\delta},$$

i.e. $d_0 = 0$. As ϕ_0 is a positive function, this implies $c_0 > 0$.

If on the other hand ∞ is a parabolic fixed point for Γ, then as the Dirichlet domain for Γ is a finite sided polygon with ∞ as a vertex [Bea83], it follows that for some $Y_0 \gg 1$, $(N\cap\Gamma)\backslash N \times [Y_0,\infty)$ injects to $\Gamma\backslash\mathbb{H}$. Therefore, if we had $c_0 = 0$ and hence $\phi_0(n_x a_y) = d_0 y^\delta$, then

$$\begin{aligned} 1 = \|\phi_0\|^2 \geq \int_{Y_0}^\infty \int_{n_x\in(N\cap\Gamma)\backslash N} |\phi_0(n_x a_y)|^2 dx\frac{dy}{y^2} \\ \geq \int_{(N\cap\Gamma)\backslash N} 1\ dn \cdot \int_{Y_0}^\infty d_0^2 y^{2\delta-2} dy = \infty, \end{aligned}$$

since $\delta > 1/2$.

This contradiction gives the desired result. □

The following Proposition shows that to compute ϕ_j^N, it suffices to integrate over a bounded set; the error term, if any, is small. We first define an appropriate bounded set J as follows.

Definition 3.6. If $N\cap\Gamma$ is a lattice in N, set $J := (N\cap\Gamma)\backslash N$. Otherwise as $\infty \notin \Lambda(\Gamma)$, $\Lambda(\Gamma)$ is a compact subset of $\mathbb{R}$, and we let $J \subset (N\cap\Gamma)\backslash N = \mathbb{R}$ be an open bounded interval which contains $\Lambda(\Gamma)$. In either case J is a bounded interval.

Proposition 3.7. *We have*

$$\phi_j^N(a_y) = \int_J \phi_j(n_x a_y)\ dx + O(y^{s_j}), \tag{3.8}$$

as $y \to 0$.

Proof. If $N\cap\Gamma$ is a lattice in N, then ϕ_j^N and $\int_J \phi_j$ are identical; there is nothing to prove. Otherwise, denoting by J^c the complement of J, we have

$$\int_{n_x\in J^c} \phi_j(n_x a_y) dn_x \ll y^{s_j}.$$

by Theorem A.1, which completes the proof. □

Next, we approximate ϕ_j^N by smoothing further in a transverse direction. Denote by U_ϵ the ball of radius ϵ about e in G. Let J be as in Proposition 3.7.

Denoting by N^- the lower triangular subgroup of G, we note that NAN^- forms an open dense neighborhood of e in G.

Definition 3.9.

- We fix a non-negative function $\eta \in C_c^\infty(\Gamma \cap N\backslash N)$ with $\eta = 1$ on J.
- Let U_{c_0} denote the c_0-neighborhood of the identity e, and fix $c_0 > 0$ so that the multiplication map
$$\mathrm{supp}(\eta) \times (U_{c_0} \cap AN^-) \to \mathrm{supp}(\eta)(U_{c_0} \cap AN^-) \subset \Gamma\backslash G$$
is a bijection onto its image.
- For each $\epsilon < c_0$, let r_ϵ be a non-negative smooth function in AN^- whose support is contained in
$$W_\epsilon := (U_\epsilon \cap A)(U_{c_0} \cap N^-)$$
and $\int_{W_\epsilon} r_\epsilon \, d\nu = 1$.
- We define the following function $\rho_{\eta,\epsilon}$ on $\Gamma\backslash G$ which is 0 outside $\mathrm{supp}(\eta)U_{c_0}$, and for $g = n_x a n^- \in \mathrm{supp}(\eta)(U_{c_0} \cap AN^-)$,
$$\rho_{\eta,\epsilon}(g) := \eta(n_x) \cdot r_\epsilon(an^-).$$

Proposition 3.10. *We have for all small $0 < \epsilon \ll \epsilon_0$ and for all $0 < y < 1$,*
$$|\phi_j^N(a_y) - \langle a_y\phi_j, \rho_{\eta,\epsilon}\rangle_{L^2(\Gamma\backslash G)}| \ll \epsilon \cdot y^{1-s_j}.$$

Proof. This follows in the same way as Proposition 6.4 in [KO08]. □

The next corollary follows immediately from Proposition 3.10 and Theorem 3.5.

Corollary 3.11. *For $\epsilon < 1$, and $j = 0, 1, \ldots, k$,*
$$\langle a_y\phi_j, \rho_{\eta,\epsilon}\rangle_{L^2(\Gamma\backslash G)} \ll y^{1-s_j},$$
as $y \to 0$.

3.3. **Proof of Theorem 1.11.**

The proof is almost identical to the proof of Theorem 6.1 in [KO08]. We sketch the main steps.

Recall that we wish to evaluate
$$\int_{N\cap\Gamma\backslash N} \psi(na_y)dn, \tag{3.12}$$
as $y \to 0$. For y sufficiently small, we can replace $N \cap \Gamma\backslash N$ with J from Def. 3.6.

Definition 3.13. For a given $\psi \in C^\infty(\Gamma\backslash G)^K$ and $\eta \in C_c(\Gamma \cap N\backslash N)$, define the function $\mathcal{I}_\eta(\psi)$ on G by
$$\mathcal{I}_\eta(\psi)(a_y) := \int_{n\in(\Gamma\cap N)\backslash N} \psi(na_y)\eta(n)\, dn.$$

Hence (3.12) is the same as $\mathcal{I}_\eta(\psi)(a_y)$, where η is as in Def 3.9. It remains to evaluate $\mathcal{I}_\eta(\psi)(a_y)$.

Lemma 3.14. *For $\psi \in C_c^\infty(\Gamma\backslash G)^K$, there exists $\psi^\sharp \in C_c^\infty(\Gamma\backslash G)^K$ such that*

(1) *for any $\epsilon < \epsilon_0$ and $h \in U_\epsilon$,*
$$|\psi(g) - \psi(gh)| \le \epsilon \cdot \psi^\sharp(g) \quad \text{for all } g \in \Gamma\backslash G.$$
(2) *$\mathcal{S}_1(\psi^\sharp) \ll \mathcal{S}_3(\psi)$, where the implied constant depends only on* $\mathrm{supp}(\psi)$.

Proof. Let $f_0 \in C_c^\infty(\Gamma\backslash G)^K$ such that $f_0(g) = 1$ for all $g \in \mathrm{supp}(\psi)U_{\epsilon_0}^{-1}K$ and $f_0(g) = 0$ for all $g \in \Gamma\backslash G - \mathrm{supp}(\psi)U_{2\epsilon_0}^{-1}K$.

Set $C_\psi := \sup_{g\in\mathrm{supp}(\psi)} \sum_{i=1}^3 |X_i(\psi)(g)|$. Then there exists a constant $c_0 \geq 1$ such that for all $g \in \Gamma\backslash G$, $h \in U_\epsilon$, and $\epsilon < \epsilon_0$,

$$|\psi(g) - \psi(gh)| \leq \epsilon \cdot c_0 C_\psi.$$

Hence if we define $\psi^\sharp \in C_c^\infty(\Gamma\backslash G)^K$ by $\psi^\sharp(g) = c_0 C_\psi f_0(g)$ for $g \in \Gamma\backslash G$, then (1) holds.

Now by the Sobolev imbedding theorem (cf. [Aub82, Thm. 2.30]), we have

$$C_\psi \leq \mathcal{S}_3(\psi).$$

Since $\mathcal{S}_1(\psi^\sharp) \ll C_\psi$, this proves (2). □

Proposition 3.15. *Let $\psi \in C^\infty(\Gamma\backslash G)^K$. Then for any $0 < y < 1$ and any small $\epsilon > 0$,*

$$|\mathcal{I}_\eta(\psi)(a_y) - \langle a_y\psi, \rho_{\eta,\epsilon}\rangle| \ll (\epsilon + y) \cdot \mathcal{I}_\eta(\psi^\sharp)(a_y),$$

where $\psi^\sharp$ is given by Lemma 3.14.

Proof. This is the same as Proposition 6.6 in [KO08]. □

By Proposition 3.7, we have that

$$\phi_j^N(a_y) = \mathcal{I}_\eta(\phi_j)(a_y) + O(y^{s_j}).$$

For simplicity, we set $\rho_\epsilon = \rho_{\eta,\epsilon}$ where $\rho_{\eta,\epsilon} = \eta \otimes r_\epsilon$ is defined as in Def. 3.9. Noting that r_ϵ is essentially an ϵ-approximation only in the A-direction, we obtain that $\mathcal{S}_1(\rho_\epsilon) = O(\epsilon^{-3/2})$.

Set $p = 3/2$. Fix ℓ, a parameter to be chosen later. Setting $\psi_0(g) := \psi(g)$, we define for $1 \leq i \leq \ell$, inductively

$$\psi_i(g) := \psi_{i-1}^\sharp(g)$$

where $\psi_{i-1}^\sharp$ is given by Lemma 3.14.

Applying Proposition 3.15 to each ψ_i, we obtain for $0 \leq i \leq \ell - 1$

$$\mathcal{I}_\eta(\psi_i)(a_y) = \langle a_y\psi_i, \rho_\epsilon\rangle + O((\epsilon + y)\mathcal{I}_\eta(\psi_{i+1})(a_y))$$

and

$$\mathcal{I}_\eta(\psi_\ell)(a_y) = \langle a_y\psi_\ell, \rho_\epsilon\rangle + O((\epsilon + y)\mathcal{S}_1(\psi_\ell)).$$

Note that by Corollary 3.3, we have for each $1 \leq i \leq \ell$

$$\begin{aligned}\langle a_y\psi_i, \rho_\epsilon\rangle &= \sum_{j=0}^{k} \langle \psi_i, \phi_j\rangle\langle a_y\phi_j, \rho_\epsilon\rangle + O(y^{1/2}\mathcal{S}_1(\rho_\epsilon)) \\ &\ll_\psi y^{1-\delta}.\end{aligned}$$

Combining the above with Proposition 3.10 and Corollary 3.11, we get that for any $y < \epsilon$,

$$\begin{aligned}\mathcal{I}_\eta(\psi)(a_y) &= \langle a_y\psi, \rho_\epsilon\rangle + O\left(\sum_{i=1}^{\ell-1}\langle a_y\psi_i, \rho_\epsilon\rangle(\epsilon+y)^k\right) + O(\mathcal{S}_1(\psi_\ell)(\epsilon+y)^\ell)\\ &= \sum_{j=0}^{k}\langle\psi,\phi_j\rangle\langle a_y\phi_j,\rho_\epsilon\rangle + O(y^{1/2}\mathcal{S}_1(\rho_\epsilon)) + O\left(\sum_{i=1}^{\ell-1} y^{1-\delta}\epsilon^k\right) + O(\epsilon^\ell)\\ &= \sum_{j=0}^{k}\langle\psi,\phi_j\rangle\phi_j^N(a_y) + O\left(\epsilon\cdot y^{1-\delta}\right) + O_\varepsilon(y^{1/2-\varepsilon}\epsilon^{-p}) + O(\epsilon^\ell),\end{aligned}$$

where the implied constants depend on the Sobolev norms of ψ.

Balancing the first two error terms and recalling $p = 3/2$, one arrives at the optimal choice

$$\epsilon = y^{(\delta-1/2)/(1+p)}.$$

With this choice of ϵ, the first two error terms are

$$\ll_\alpha y^{\frac{4-3\delta}{5}-\alpha},$$

for any $\alpha > 0$. Lastly, we choose ℓ to make the final error term of the same quality, namely

$$\ell = \frac{4-3\delta}{2\delta-1}.$$

Therefore

$$\int_{(N\cap\Gamma)\backslash N}\psi(na_y)\,dn = \sum_{j=0}^{k}\langle\psi,\phi_j\rangle\phi_j^N(a_y) + O_{\psi,\alpha}(y^{\frac{4-3\delta}{5}-\alpha}),$$

for any $\alpha > 0$.

This completes the proof of Theorem 1.11.

4. Proofs of the Counting Theorems

With Theorem 1.11 at hand, we now count establish the counting theorems; these are Theorems 1.8 and 2.6.

4.1. **Proof of Theorem 1.8.** Recall that Q is a ternary indefinite quadratic form, $\mathrm{SO}_Q(\mathbb{R})$, and Γ a finitely generated discrete subgroup with $\delta > 1/2$. Fix a non-zero vector $\mathbf{x}_0 \in \mathbb{R}^3$, lying on the cone $Q = 0$ such that the orbit $\mathcal{O} = \mathbf{x}_0\Gamma$ is discrete.

Let

$$\mathcal{N}(T) := \#\{\mathbf{x}\in\mathcal{O} : \|\mathbf{x}\| < T\}$$

where $\|\cdot\|$ denotes a Euclidean norm on $\mathbb{R}^3$.

Using the spin cover $\iota : \mathrm{SL}_2 \to \mathrm{SO}_Q$, we may assume without loss of generality that Γ is a finitely generated subgroup of $G := \mathrm{SL}_2(\mathbb{R})$ with $\delta > 1/2$. We use the notation N, A, K, a_y, n_x, etc from the introduction. Let dg denote the Haar measure given by

$$d(n_x a_y k) = y^{-2}dxdydk$$

where dk is the probability Haar measure on K, and dx and dy are Lebesque measures.

As $Q(\mathbf{x}_0)=0$, the stabilizer of $\mathbf{x}_0$ in G is conjugate to N and hence by replacing Γ with a conjugate if necessary, we may assume without loss of generality that N is precisely the stabilizer subgroup of $\mathbf{x}_0$ in G. It follows that $\mathbf{x}_0 a_y = y^{-1}\mathbf{x}_0$. Let B_T be a K-invariant ball of radius T in $\mathbb{R}^3$, and let χ_T be the characteristic function of this ball. Note that χ_T is right K-invariant, that is, $\chi_T(\mathbf{x}gk)=\chi_T(\mathbf{x}g)$, for any $g\in G$ and $k\in K$. Also, as N is the stabilizer of $\mathbf{x}_0$, we have $\chi_T(\mathbf{x}_0 ng)=\chi_T(\mathbf{x}_0 g)$, for any $n\in N$, that is, χ_T is left N-invariant.

We define the following counting function on $\Gamma\backslash G$:

$$F_T(g) := \sum_{\gamma\in N\cap\Gamma\backslash\Gamma} \chi_T(\mathbf{x}_0\gamma g).$$

Lemma 4.1. *For any $\Psi\in C_c(\Gamma\backslash G)^K$,*

$$\langle F_T,\Psi\rangle_{\Gamma\backslash G} = \int_{y>T^{-1}\|\mathbf{x}_0\|}\int_{n_x\in\Gamma\cap N\backslash N}\Psi(n_x a_y)y^{-2}dxdy.$$

Proof. We observe that by unfolding and using the K-invariance of Ψ and χ_{B_T},

$$\begin{aligned}\langle F_T,\Psi\rangle_{\Gamma\backslash G} &= \int_{\Gamma\backslash G}\sum_{\Gamma\cap N\backslash\Gamma}\chi_{B_T}(\mathbf{x}_0\gamma g)\Psi(g)dg\\ &= \int_{\Gamma\cap N\backslash G}\chi_{B_T}(\mathbf{x}_0 g)\Psi(g)dg\\ &= \int_{\|\mathbf{x}_0 a_y\|<T}\int_{\Gamma\cap N\backslash N}\Psi(na_y)y^{-2}dndy.\end{aligned}$$

As $\mathbf{x}_0 a_y = y^{-1}\mathbf{x}_0$, the claim follows. □

As before, we order discrete eigenvalues $0\le\lambda_0<\lambda_1\le\lambda_2\cdots\le\lambda_k<1/4$ of Δ on $L^2(\Gamma\backslash\mathbb{H})$ with $\lambda_j=s_j(1-s_j)$, with $s_j>1/2$, $j=1,\dots,k$ and let $\phi_j\in L^2(\Gamma\backslash\mathbb{H})$ denote the corresponding eigenfunction with $\|\phi_j\|=1$.

Recall by Theorem 3.5, there exist constants c_j and d_j, depending on ϕ_j, such that

$$\phi_j^N(a_y) = c_j y^{1-s_j} + d_j y^{s_j}.$$

Furthermore, $c_0>0$.

By inserting the asymptotic formula for $\int_{\Gamma\cap N\backslash N}\Psi(na_y)dn$ from Theorem 1.11, we deduce:

Proposition 4.2. *For any $\Psi\in C_c^\infty(\Gamma\backslash G)^K$ and $\epsilon>0$,*

$$\begin{aligned}\langle F_T,\Psi\rangle_{\Gamma\backslash G} &= \sum_{j=0}^{k}\langle F_T,\phi_j\rangle_{\Gamma\backslash G}\langle\Psi,\phi_j\rangle_{\Gamma\backslash G} + O_\varepsilon\left(T^{\frac12+\frac35(\delta-\frac12)+\varepsilon}\right)\\ &= \sum_{j=0}^{k}\langle\Psi,\phi_j\rangle_{\Gamma\backslash G}\left(\frac{c_j T^{s_j}}{s_j\|\mathbf{x}_0\|^{s_j}} + \frac{d_j T^{1-s_j}}{(1-s_j)\|\mathbf{x}_0\|^{1-s_j}}\right) + O_\epsilon(T^{\frac12+\frac35(\delta-\frac12)+\epsilon})\end{aligned}$$

where the implied constant depends only on a Sobolev norm of Ψ and ϵ.

For all small $\eta>0$, consider an η-neighborhood U_η of e in G, which is K-invariant, such that for all $T\gg 1$,

$$B_T U_\eta \subset B_{(1+\eta)T} \quad\text{and}\quad B_{(1-\eta)T}\subset\cap_{u\in U_\eta}B_T u.$$

Let $\psi_\eta \in C^\infty(\mathbb{H})$ denote a non-negative function supported on U_η with $\int_G \psi dg = 1$. We lift ψ_η to $\Gamma\backslash G$ by

$$\Psi_\eta(g) := \sum_{\gamma\in\Gamma} \psi_\eta(\gamma g).$$

Then

$$\langle F_{(1-\eta)T}, \Psi_\eta\rangle \le F_T(e) \le \langle F_{(1+\eta)T}, \Psi_\eta\rangle. \tag{4.3}$$

On the other hand, recalling that $\{X_1, X_2, X_3\}$ is an orthonormal basis for the Lie algebra $\mathfrak{g} = \mathfrak{s}l(2,\mathbb{R})$ of G, we have

$$\langle \Psi_\eta, \phi_j\rangle = \phi_j(e) + O(\eta \sup_{g\in U_\eta} \sup_{i=1,2,3} X_i\phi_j(g))$$

where the implied constant is absolute.

Therefore Proposition 4.2 yields for $C_0 = \frac{\phi_0(e)c_0}{\delta\|\mathbf{x}_0\|^\delta} > 0$,

$$\langle F_{(1-\eta)T}, \Psi_\eta\rangle = C_0T^\delta + O_\epsilon(T^{s_1} + \eta T^\delta + \eta^{-A}T^{\frac{1}{2}+\frac{3}{5}(\delta-\frac{1}{2})+\epsilon})$$

for some $A > 0$ and any $\epsilon > 0$, where we used that the Sobolev norms of Ψ_η grow at most polynomially in η; the implied constant now depends only on ϵ. Therefore by equating the last two terms, we can choose $\eta = T^{-r}$ for some $r > 0$ and obtain from (4.3) that

$$F_T(e) = C_0T^\delta + O(T^{\delta-\zeta})$$

for some $\zeta > 0$. This proves Theorem 1.8.

4.2. **Proof of Theorem 2.6:** As in the discussion preceding Theorem 2.6, we may assume that Γ is a finitely generated subgroup of $\mathrm{SL}_2(\mathbb{Z})$. We may also assume without loss of generality that $g_0 = e$ by replacing Γ by $g_0\Gamma g_0^{-1}$ and hence the stabilizer subgroup of $\mathbf{x}_0$ in G is precisely the upper triangular subgroup N. Recall the weight ξ_T, depending on ψ, as given in Definition 2.5.

Let $q \ge 1$ be squarefree, and let $\Gamma_1(q)$ be any group satisfying

$$\Gamma(q) \subset \Gamma_1(q) \subset \Gamma$$

and

$$N \cap \Gamma_1(q) = N \cap \Gamma, \tag{4.4}$$

where $\Gamma(q) = \{\gamma \in \Gamma : \gamma \equiv I(q)\}$ is the "congruence" subgroup of Γ of level q.

We define the following K-invariant functions on $\Gamma_1(q)\backslash G$:

$$F_T^q(g) := \sum_{\gamma\in(N\cap\Gamma)\backslash\Gamma_1(q)} \chi_T(\mathbf{x}_0\gamma g),$$

and for any fixed $\gamma_0 \in \Gamma$,

$$\Psi_{\gamma_0}^q(g) := \sum_{\gamma\in\Gamma_1(q)} \psi(\gamma_0^{-1}\gamma g).$$

Proposition 4.5. *We have*

$$\sum_{\mathbf{x}\in\mathbf{x}_0\Gamma_1(q)} \xi_T(\mathbf{x}\gamma_0) = \left\langle F_T^q, \Psi_{\gamma_0}^q\right\rangle_{\Gamma_1(q)\backslash G}. \tag{4.6}$$

Proof. Starting with the left hand side of (4.6), we insert Definition 2.5, use that N stabilizes $\mathbf{x}_0$, and that $N \cap \Gamma = N \cap \Gamma_1(q)$:

$$\begin{aligned}\sum_{\mathbf{x}\in\mathbf{x}_0\Gamma_1(q)} \xi_T(\mathbf{x}\gamma_0) &= \sum_{\mathbf{x}\in\mathbf{x}_0\Gamma_1(q)} \int_G \chi_T(\mathbf{x}\gamma_0 g)\psi(g)dg \\ &= \sum_{\gamma\in(N\cap\Gamma)\backslash\Gamma_1(q)} \int_G \chi_T(\mathbf{x}_0\gamma g)\psi(\gamma_0^{-1}g)dg \\ &= \int_G F_T^q(g)\psi(\gamma_0^{-1}g)dg \\ &= \sum_{\gamma\in\Gamma_1(q)} \int_{\Gamma_1(q)\backslash G} F_T^q(g)\psi(\gamma_0^{-1}\gamma g)dg \\ &= \int_{\Gamma_1(q)\backslash G} F_T^q(g)\Psi_{\gamma_0}^q(g)dg,\end{aligned}$$

where we used the definition of F_T^q, "refolded" (the reverse of the "unfolding trick"), and used the definition of $\Psi_{\gamma_0}^q$. □

Note that the Sobolev norms of $\Psi_{\gamma_0}^q$ are same as those of Ψ, i.e., independent of γ_0 and q (since the functions are built out of the same ψ). Applying Proposition 4.2 to each $\Gamma_1(q)$, we then obtain:

Proposition 4.7. *Let*

$$0 < \delta(1-\delta) = \lambda_0^{(q)} < \lambda_1^{(q)} \leq \cdots \leq \lambda_{k_q}^{(q)} < 1/4$$

denote the point spectrum in $L^2(\Gamma_1(q)\backslash G)$, and let $\phi_0^{(q)}, \ldots, \phi_{k_q}^{(q)}$ be the corresponding orthonormal eigenfunctions. Then for any $\varepsilon > 0$,

$$\begin{aligned}\left\langle F_T^q, \Psi_{\gamma_0}^q \right\rangle_{\Gamma_1(q)\backslash G} &= \sum_{j=0}^{k_q} \left\langle F_T^q, \phi_j^{(q)} \right\rangle_{\Gamma_1(q)\backslash G} \left\langle \phi_j^{(q)}, \Psi_{\gamma_0}^q \right\rangle_{\Gamma_1(q)\backslash G} \\ &\quad + O_\varepsilon\left(T^{\frac{1}{2}+\frac{3}{5}(\delta-\frac{1}{2})+\varepsilon}\right),\end{aligned}$$

as $T \to \infty$. The implied constant depends on ε and a Sobolev norm of ψ, but not on q or γ_0.

Lemma 4.8. *For $q'|q$ and $\phi^{(q')} \in L^2(\Gamma_1(q')\backslash G)$ of norm one, consider the normalized old form $\phi^{(q)}$ in $L^2(\Gamma_1(q)\backslash G)$ of level q':*

$$\phi^{(q)} = \frac{1}{\sqrt{[\Gamma_1(q') : \Gamma_1(q)]}} \phi^{(q')}. \tag{4.9}$$

Then for any $\gamma_0 \in \Gamma$,

$$\frac{\left\langle \phi^{(q)}, \Psi_{\gamma_0}^q \right\rangle_{\Gamma_1(q)\backslash G}}{\left\langle \phi^{(q')}, \Psi_{\gamma_0}^{q'} \right\rangle_{\Gamma_1(q')\backslash G}} = \frac{1}{\sqrt{[\Gamma_1(q') : \Gamma_1(q)]}} = \frac{\left\langle \phi^{(q)}, F_T^q \right\rangle_{\Gamma_1(q)\backslash G}}{\left\langle \phi^{(q')}, F_T^{q'} \right\rangle_{\Gamma_1(q')\backslash G}}. \tag{4.10}$$

Proof. Consider the inner product

$$\begin{aligned}\left\langle \phi^{(q)}, \Psi^q_{\gamma_0} \right\rangle_{\Gamma_1(q)\backslash G} &= \int_{\Gamma_1(q)\backslash G} \phi^{(q)}(g)\Psi^q_{\gamma_0}(g)dg \\ &= \int_G \phi^{(q)}(g)\psi(\gamma_0^{-1}g)dg \\ &= \frac{1}{\sqrt{[\Gamma_1(q') : \Gamma_1(q)]}} \int_G \phi^{(q')}(\gamma_0 g)\psi(g)dg \\ &= \frac{1}{\sqrt{[\Gamma_1(q') : \Gamma_1(q)]}} \left\langle \phi^{(q')}, \Psi^{q'}_{\gamma_0} \right\rangle_{\Gamma\backslash G},\end{aligned}$$

where we unfolded, used (4.4) and the $\Gamma_1(q')$-invariance of $\phi^{(q')}$, and refolded.

The second equality in (4.10) is proved in the same way. □

Lemma 4.11. *Let $q = q'q''$ and assume that*

$$\phi_j^{(q)} = \frac{1}{\sqrt{[\Gamma_1(q') : \Gamma_1(q)]}} \phi_j^{(q')}.$$

Then

$$\left\langle F_T^q, \phi_j^{(q)} \right\rangle = \frac{1}{\sqrt{[\Gamma : \Gamma_1(q)]}} \left(c_j^{(q')}\ T^{s_j} + d_j^{(q')}\ T^{1-s_j} \right),$$

where $c_j^{(q')}$ and $d_j^{(q')}$ are independent of q''.

Proof. The inner product $\left\langle F_T^q, \phi_j^{(q)} \right\rangle$ can be unfolded again, giving

$$\begin{aligned}\left\langle F_T^q, \phi_j^{(q)} \right\rangle &= \int_{y>\|\mathbf{x}_0\|/T} \int_{(N\cap\Gamma)\backslash N} \phi_j^{(q)}(na_y)\ dn\ \frac{dy}{y^2} \\ &= \frac{1}{\sqrt{[\Gamma_1(q') : \Gamma_1(q)]}} \int_{y>\|\mathbf{x}_0\|/T} \int_{(N\cap\Gamma)\backslash N} \phi_j^{(q')}(na_y)\ dn\ \frac{dy}{y^2} \\ &= \frac{1}{\sqrt{[\Gamma_1(q') : \Gamma_1(q)]}} \int_{y>\|\mathbf{x}_0\|/T} \left(c_j^{(q')} y^{1-s_j} + d_j^{(q')} y^{s_j} \right) \frac{dy}{y^2} \\ &= \frac{\sqrt{[\Gamma : \Gamma_1(q')]}}{\sqrt{[\Gamma : \Gamma_1(q)]}} \int_{y>\|\mathbf{x}_0\|/T} \left(c_j^{(q')} y^{1-s_j} + d_j^{(q')} y^{s_j} \right) \frac{dy}{y^2}\end{aligned}$$

where we used Theorem 3.5 as well as the identity

$$[\Gamma : \Gamma_1(q)] = [\Gamma : \Gamma_1(q')][\Gamma_1(q') : \Gamma_1(q)].$$

The claim follows from a simple computation and renaming the constants. □

Recall from Definition 2.2 that square-free q are to be decomposed as $q = q'q''$ with $q' \mid \mathfrak{B}$ and $(q'', \mathfrak{B}) = 1$. Let

$$0 < \delta(1-\delta) = \lambda_0^{(q)} < \lambda_1^{(q)} \leq \cdots \leq \lambda_{k_q}^{(q)} < 1/4$$

be the eigenvalues of the Laplacian acting on $L^2(\Gamma(q)\backslash\mathbb{H})$. The eigenvalues below $\theta(1-\theta)$ are all oldforms coming from level 1, with the possible exception of finitely many eigenvalues coming from level $q' \mid \mathfrak{B}$.

For ease of exposition, assume the spectrum below $\theta(1-\theta)$ consists of only the base eigenvalue $\lambda_0 = \delta(1-\delta)$ corresponding to $\phi^{(q)}$, and one newform $\tilde{\phi}^{(q)}$ from the "bad" level $q' \mid \mathfrak{B}$. The general case is a finite sum of such terms.

Combining (4.6) and Proposition 4.7 with Lemmata 4.8 and 4.11 gives

$$\begin{aligned}\sum_{\mathbf{x}\in\mathbf{x}_0\Gamma_1(q)}\xi_T(\mathbf{x}\gamma_0) &= \frac{1}{[\Gamma:\Gamma_1(q)]}\left\langle F_T,\phi^{(1)}\right\rangle_{\Gamma\backslash G}\left\langle\phi^{(1)},\Psi^1\right\rangle_{\Gamma\backslash G}\\ &+\frac{[\Gamma:\Gamma_1(q')]}{[\Gamma:\Gamma_1(q)]}\left\langle F_T,\tilde{\phi}^{(q')}\right\rangle_{\Gamma_1(q')\backslash G}\left\langle\tilde{\phi}^{(q')},\Psi^{q'}_{\gamma_0}\right\rangle_{\Gamma_1(q')\backslash G}\\ &+O_\varepsilon(T^{\theta+\varepsilon}+T^{\frac{1}{2}+\frac{3}{5}(\delta-\frac{1}{2})+\varepsilon}).\end{aligned}$$

Setting

$$\mathcal{E}(T,q',\gamma_0):=[\Gamma:\Gamma_1(q')]\left\langle F_T,\tilde{\phi}^{(q')}\right\rangle_{\Gamma_1(q')\backslash G}\left\langle\tilde{\phi}^{(q')},\Psi^{q'}_{\gamma_0}\right\rangle_{\Gamma_1(q')\backslash G},$$

the proposition follows by recognizing the main term as the main contribution to

$$\Xi(T)=\sum_{\mathbf{x}\in\mathbf{x}_0\Gamma}\xi_T(\mathbf{x})=\langle F_T,\Psi\rangle_{\Gamma\backslash G}.$$

This completes the proof of Theorem 2.6, part (2). Part (1) follows immediately from Proposition 4.2.

5. Proofs of the Sieving Theorems

We now consider

$$Q=x^2+y^2-z^2$$

and fix a finitely generated subgroup $\Gamma<\mathrm{SO}_Q(\mathbb{Z})$ with $\delta>1/2$. Let $\mathbf{x}_0\in\mathbb{Z}^3\setminus\{0\}$ with $Q(\mathbf{x}_0)=0$.

Again by considering the spin cover $G:=\mathrm{SL}_2\to\mathrm{SO}_Q$ over $\mathbb{Q}$, we may assume without loss of generality that Γ is a finitely generated subgroup of $\mathrm{SL}_2(\mathbb{Z})$.

Let F be a polynomial which is integral on the orbit $\mathcal{O}=\mathbf{x}_0\Gamma$.

5.1. **Strong Approximation.**

We first pass to a finite index subgroup of our original Γ which is chosen so that its projection to $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ is either the identity or all of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$.

Lemma 5.1. *There exists an integer $\mathfrak{B}\geq 1$ with $\mathfrak{B}\equiv 0(\mathrm{mod}\,6)$ so that*

$$\Gamma(\mathfrak{B})=\{\gamma\in\Gamma:\gamma\equiv I(\mathrm{mod}\,\mathfrak{B})\}$$

projects onto $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ for $p\nmid\mathfrak{B}$. Obviously the projection of $\Gamma(\mathfrak{B})$ in $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ for $p\mid\mathfrak{B}$ is the identity.

This follows from Strong Approximation; see, e.g. [Gam02, §2]. From now on we replace Γ by $\Gamma(\mathfrak{B})$. We use Goursat's Lemma (e.g. [Lan02], p. 75) to have a similar statement for the reduction modulo a square-free parameter:

Theorem 5.2 (Thm 2.1 of [BGS10])**.** *There exists a number $\mathfrak{B}$ such that if $q=q'q''$ is square-free with $q'|\mathfrak{B}$ and $(q'',\mathfrak{B})=1$ then the projection of Γ in $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ is isomorphic to $\mathrm{SL}_2(\mathbb{Z}/q''\mathbb{Z})$.*

5.2. **Executing the sieve.**

Choose a weight ξ_T as in Definition 2.5. Let $\mathcal{A} = \mathcal{A}(T) = \{a_n(T)\}$ where

$$a_n(T) := \sum_{\substack{\mathbf{x}\in\mathcal{O}\\ F(\mathbf{x})=n}} \xi_T(\mathbf{x}).$$

Then

$$|\mathcal{A}(T)| = \sum_n a_n(T) = \sum_{\mathbf{x}\in\mathcal{O}} \xi_T(\mathbf{x}) = \Xi(T) =: \mathcal{X},$$

where the leading term $\mathcal{X} \asymp T^\delta$.

For q square-free

$$|\mathcal{A}_q(T)| = \sum_{n\equiv 0(q)} a_n(T) = \sum_{\substack{\mathbf{x}\in\mathcal{O}\\ F(\mathbf{x})\equiv 0(n)}} \xi_T(\mathbf{x}).$$

Let $\Gamma_{\mathbf{x}_0}(q)$ be the subgroup of Γ which stabilizes $\mathbf{x}_0 \bmod q$, i.e.

$$\Gamma_{\mathbf{x}_0}(q) := \{\gamma \in \Gamma : \mathbf{x}_0\gamma \equiv \gamma(q)\}.$$

Clearly

$$\Gamma(q) \subset \Gamma_{\mathbf{x}_0}(q) \subset \Gamma,$$

so Theorem 2.6 (2) applies.

Then

$$|\mathcal{A}_q(T)| = \sum_{\substack{\gamma'\in\Gamma_{\mathbf{x}_0}(q)\backslash\Gamma\\ F(\mathbf{x}_0\gamma)\equiv 0(q)}} \left(\sum_{\gamma\in \mathrm{Stab}_\Gamma(\mathbf{x}_0)\backslash\Gamma_{\mathbf{x}_0}(q)} \xi_T(\mathbf{x}_0\gamma\gamma') \right).$$

Let

$$|\mathcal{O}^F(q)| := \sum_{\substack{\gamma\in\Gamma_{\mathbf{x}_0}(q)\backslash\Gamma\\ F(\mathbf{x}_0\gamma)\equiv 0(q)}} 1.$$

Theorem 2.6 then gives

$$\begin{aligned} |\mathcal{A}_q(T)| &= \frac{|\mathcal{O}^F(q)|}{[\Gamma : \Gamma_{\mathbf{x}_0}(q)]}(\mathcal{X} + \mathcal{X}_{q'}) \\ &\quad + O_\varepsilon\left(T^\varepsilon |\mathcal{O}^F(q)| \left(T^\theta + T^{\frac{1}{2}+\frac{3}{5}(\delta-\frac{1}{2})} \right) \right), \end{aligned}$$

with

$$\mathcal{X}_{q'} \ll \mathcal{X}^{1-\zeta}, \tag{5.3}$$

for some $\zeta > 0$, uniformly in q'. Define

$$g^F(q) := \frac{|\mathcal{O}^F(q)|}{[\Gamma : \Gamma_{\mathbf{x}_0}(q)]}.$$

Lemma 5.4. *Assume the pair $(\mathcal{O}, F)$ is strongly primitive. Then for q square-free, $g^F(q)$ is completely multiplicative, and $g^F(p) < 1$. If $q = q'q''$ with $q' \mid \mathfrak{B}$ and $q' \geq 2$, then $g(q) = 0$.*

Furthermore,

(1) *For $F_{\mathcal{H}}(\mathbf{x}) = z$, we have*

$$g^{F_{\mathcal{H}}}(p) = \begin{cases} 2/(p+1), & \text{if } p \equiv 1(4), \\ 0, & \text{otherwise.} \end{cases} \tag{5.5}$$

(2) *For* $F = F_{\mathcal{A}} = \frac{1}{12}xy$, *we have* $g^{F_{\mathcal{A}}}(p) = 4/(p+1)$.
(3) *For* $F = F_{\mathcal{C}} = \frac{1}{60}xyz$, *we have*

$$g^{F_{\mathcal{C}}}(p) = \begin{cases} 4/(p+1), & \text{if } p \equiv 1(4), \\ 6/(p+1), & \text{if } p \equiv 3(4). \end{cases} \tag{5.6}$$

Proof. For unramified $q = q_1 q_2$ with q_1 and q_2 relatively prime and both prime to $\mathfrak{B}$, then $\mathcal{O}(q)$, the orbit of $\mathbf{x}_0$ mod q is equal to $\mathcal{O}(q_1) \times \mathcal{O}(q_2)$ in $(\mathbb{Z}/q_1\mathbb{Z})^3 \times (\mathbb{Z}/q_2\mathbb{Z})^3 = (\mathbb{Z}/q\mathbb{Z})^3$.

For ramified $q = q'q''$, with $q' \mid \mathfrak{B}$, then Γ projects onto the identity mod q', so $\mathcal{O}(q')$ is just one point, i.e. $\mathcal{O}(q') = \{\mathbf{x}_0\}$. It also follows in this case that $\mathcal{O}^F(q'q'')$ is isomorphic to $\mathcal{O}^F(q') \times \mathcal{O}^F(q'')$.

Since $[\Gamma : \Gamma_{\mathbf{x}_0}(q)] = |\mathcal{O}(q)|$, we have shown that

$$g^F(q) = \frac{|\mathcal{O}^F(q)|}{|\mathcal{O}(q)|}$$

is multiplicative, and thus is determined by its values on the primes (only square-free q are ever used).

From the assumption that the pair $(\mathcal{O}, F)$ is strongly primitive, it immediately follows that $|\mathcal{O}^F(p)| < |\mathcal{O}(p)|$, i.e. $g^F(p) < 1$. Notice that if $p \mid \mathfrak{B}$ then $|\mathcal{O}(p)| = 1$ and $|\mathcal{O}^F(p)| = 0$, so $g^F(p) = 0$.

It remains to compute the values of g^F on primes $p \nmid \mathfrak{B}$. Denote by V the cone defined by $Q = x^2 + y^2 - z^2 = 0$, minus the origin. I.e.

$$V = \{(x, y, z) \neq (0, 0, 0) : x^2 + y^2 - z^2 = 0\}.$$

For $F = F_{\mathcal{H}} = z$, let

$$W_1 = \{\mathbf{x} \in V : F_{\mathcal{H}}(\mathbf{x}) = 0\} = \{(x, y, 0) \neq (0, 0, 0) : x^2 + y^2 = 0\}.$$

As V is a homogeneous space of G with a connected stabilizer, we have

$$\mathcal{O}(p) = V(\mathbb{F}_p), \quad \text{and hence } \mathcal{O}^{F_{\mathcal{H}}}(p) = W_1(\mathbb{F}_p).$$

We can easily calculate $|V(\mathbb{F}_p)| = p^2 - 1$. If $p \equiv 3(4)$, then $W_1(\mathbb{F}_p)$ is empty. If $p \equiv 1(4)$, then W_1 is the disjoint union of the two lines $\{(x, y) \neq 0 : x = \pm\sqrt{-1}y\}$, each of cardinality $p - 1$. This proves claim (1).

For $F = F_{\mathcal{A}} = \frac{1}{12}xy$, we set

$$\begin{aligned} W_2 &:= \{\mathbf{x} \in V : F_{\mathcal{A}}(\mathbf{x}) = 0\} \\ &= \{(0, y, z) \neq (0, 0, 0) : y^2 - z^2 = 0\} \sqcup \{(x, 0, z) \neq (0, 0, 0) : x^2 - z^2 = 0\} \end{aligned}$$

Thus W_2 is the disjoint union of four lines, $x = \pm z$, $y = \pm z$, proving claim (2).

For $F = F_{\mathcal{C}} = \frac{1}{60}xyz$, we see immediately that

$$W_3 := \{\mathbf{x} \in V : F_{\mathcal{C}}(\mathbf{x}) = 0\} = W_1 \sqcup W_2,$$

proving claim (3). □

From Lemma 5.4, the computation leading to (2.14) is a classical exercise (see e.g. [Lan53]), with sieve dimensions

$$\kappa = 1, 4 \text{ and } 5 \text{ for } F = F_{\mathcal{H}},\ F_{\mathcal{A}} \text{ and } F_{\mathcal{C}}, \text{ respectively.} \tag{5.7}$$

Define $r_q := |\mathcal{O}^F(q)|T^{\theta+\varepsilon}$. From the proof of Lemma 5.4, we have $|\mathcal{O}^F(p)| \ll p$, so

$$\sum_{\substack{q<\mathcal{X}^\tau \\ q \text{ squarefree}}} 4^{\nu(q)}|r_q| \ll_\varepsilon \mathcal{X}^{2\tau+\varepsilon}T^\theta.$$

As $\mathcal{X} \sim cT^\delta$, this error term is admissible, that is, satisfies (2.16), for any

$$\tau < \frac{\delta-\theta}{2\delta}. \tag{5.8}$$

The elements $a_n(T)$ are zero for $n \gg T \gg \mathcal{X}^{1/\delta}$, so (2.17) is satisfied for

$$\mu > \frac{2}{\delta-\theta}. \tag{5.9}$$

We have thus established that our sequence $\mathcal{A}$ satisfies

$$|\mathcal{A}_q| = g(q)\mathcal{X} + g(q)\mathcal{X}_{q'} + r_q.$$

But note in fact that (we thank the referee for observing this simplification of a previous argument) by Lemma 5.4, $g(q) = 0$ if $q' \geq 2$, and so we always have just

$$|\mathcal{A}_q| = g(q)\mathcal{X} + r_q.$$

We now apply Theorem 2.18. For any R satisfying (2.21), have

$$\sum_{n\in\mathcal{P}(R)} a_n \gg \frac{\mathcal{X}}{\log^\kappa \mathcal{X}},$$

where κ is determined in (5.7), according to the choice of $F \in \{F_{\mathcal{H}}, F_{\mathcal{A}}, F_{\mathcal{C}}\}$.

Having verified the sieve axioms, the upper bound of the same order of magnitude follows from a standard application of a combinatorial sieve, see e.g. [Kon09, Theorem 2.5] where the details are carried out.

5.3. **Explicit values of R.** It remains to determine values of R for which the above discussion holds.

The values of α_κ and β_κ in Theorem 2.18 can be tabulated, see for instance Appendix III on p. 345 of [DHR88]. The sets $\mathcal{A}$ appearing in this paper have sieve dimensions $\kappa = 1, 4$ and 5; for these values, we have

$$\alpha_1 = \beta_1 = 2, \alpha_4 = 11.5317.., \beta_4 = 9.0722.., \alpha_5 = 14.7735.., \beta_5 = 11.5347... \tag{5.10}$$

We will also need precise estimates on the functions which appear in (2.21). Although these are difficult to extract by hand, the following procedure is quite effective in practice. Usually, u is chosen so that τu is near 1, and v so that τv exceeds α_κ. Precisely, for any $\zeta \in (0, \beta_\kappa)$, set

$$\tau u = 1 + \zeta - \zeta/\beta_\kappa, \quad \tau v = \beta_\kappa/\zeta + \beta_\kappa - 1.$$

Then by Halberstam-Richert [HR74], equations (10.1.10), (10.2.4) and (10.2.7), we obtain

$$\frac{\kappa}{f_\kappa(\tau v)} \int_1^{v/u} F_\kappa(\tau v - s)\left(1 - \frac{u}{v}s\right)\frac{ds}{s} \leq (\kappa+\zeta)\log\frac{\beta_\kappa}{\zeta} - \kappa + \zeta\frac{\kappa}{\beta_\kappa}.$$

Thus Theorem 2.18 holds with

$$R > \mu(1+\zeta-\zeta/\beta_\kappa) - 1 + (\kappa+\zeta)\log\frac{\beta_\kappa}{\zeta} - \kappa + \zeta\frac{\kappa}{\beta_\kappa} =: m(\zeta), \tag{5.11}$$

for any $0 < \zeta < \beta_\kappa$. After inputting the values of μ, κ, α_κ and β_κ, the minimum of $m(\zeta)$ is easily determined by hand or with computer assistance.

F	$(\Gamma \cap N)\backslash N$	δ	θ	μ	$m(\zeta)$	R
$F_{\mathcal{H}}$	Any	1	5/6	12	13.93..	14
$F_{\mathcal{H}}$	Any	0.9992	5/6	12.05	13.99..	14
$F_{\mathcal{H}}$	Any	1	39/64	5.12	6.48..	7
$F_{\mathcal{H}}$	Finite	1	1/2	4	5.22..	6
$F_{\mathcal{H}}$	Finite	0.9265	1/2	4.69	5.99..	6
$F_{\mathcal{H}}$	Infinite	1	1/2	10	11.8..	12
$F_{\mathcal{H}}$	Infinite	0.991	1/2	10.2	11.9..	12
$F_{\mathcal{A}}$	Any	1	5/6	12	24.9..	25
$F_{\mathcal{A}}$	Any	0.99995	5/6	12.0	24.9..	25
$F_{\mathcal{A}}$	Any	1	39/64	5.12	15.6..	16
$F_{\mathcal{A}}$	Finite	1	1/2	4	13.8..	14
$F_{\mathcal{A}}$	Finite	0.98805	1/2	4.1	13.9..	14
$F_{\mathcal{A}}$	Infinite	1	1/2	10	22.4..	23
$F_{\mathcal{A}}$	Infinite	0.97895	1/2	10.4	22.9..	23
$F_{\mathcal{C}}$	Any	1	5/6	12	28.7..	29
$F_{\mathcal{C}}$	Any	0.99677	5/6	12.2	28.99..	29
$F_{\mathcal{C}}$	Any	1	39/64	5.12	18.7..	19
$F_{\mathcal{C}}$	Finite	1	1/2	4	16.7..	17
$F_{\mathcal{C}}$	Finite	0.981675	1/2	4.2	16.99..	17
$F_{\mathcal{C}}$	Infinite	1	1/2	10	25.9..	26
$F_{\mathcal{C}}$	Infinite	0.99905	1/2	10.02	25.9..	26

TABLE 1. Values of R depending on δ, θ, and whether $N \cap \Gamma$ is assumed to be a lattice in N.

For the function $F_{\mathcal{H}} = z$, we have $\kappa = 1$, and $\alpha_1 = 2 = \beta_1$. The best value of R is obtained for $\delta \to 1$, where we can take $\theta = 5/6$. Then (5.9) gives $\mu > 12$, and we have collected everything required to compute the minimum of $m(\zeta)$ defined in (5.11). We find that the minimum value is attained at $\zeta = 0.1203..$ with $m(\zeta) = 13.931...$ Thus $R = 14$ is the limit of our method. For $\delta > 1 - 1/1250$ and $\theta = 5/6$, we find the minimum value $m(0.1198..) = 13.992$, which still allows $R = 14$. For comparison, consider instead a finite co-volume congruence subgroup; then $\delta = 1$ and can take $\theta = 39/64$ by [KS03]. This gives $m(0.238..) = 6.48..$, allowing $R = 7$. If one could take θ arbitrarily close to $1/2$, the above calculation gives $m(0.292..) = 5.216..$, or $R = 6$.

Table 1 summarizes the discussion above and extends it to the other choices $F_{\mathcal{A}} = \frac{1}{12}xy$ and $F_{\mathcal{C}} = \frac{1}{60}xyz$, with various possibilities for δ, θ, and whether $\Gamma \cap N$ is a lattice in N. For comparison, we also show a spectral gap $\theta = 39/64$ [KS03] for congruence subgroups of $\mathrm{SL}(2, \mathbb{Z})$. These values of R are precisely those quoted in Theorems 2.27, 2.28, and 2.29, in particular proving Theorem 1.5.

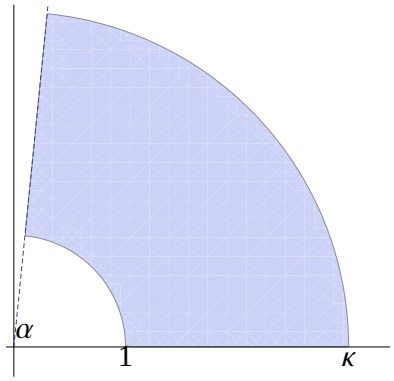

FIGURE 5. A flare domain.

APPENDIX A. PROOF OF THEOREM 1.13

Recall the notation (1.10). Let $\Gamma < G$ be a discrete, finitely generated subgroup with critical exponent $\delta > 1/2$. Let $\phi \in L^2(\Gamma\backslash\mathbb{H})$ be an eigenfunction of the hyperbolic Laplace-Beltrami operator Δ with eigenvalue $\lambda = s(1-s) < 1/4$ and $s > 1/2$.

The aim of this section is to reproduce the proof of Theorem 1.13, which was demonstrated in [Kon07]. We give the statement again.

Theorem A.1 ([Kon07])**.** *Assume that the volume of $\Gamma\backslash\mathbb{H}$ is infinite, and that the horocycle $(N \cap \Gamma)\backslash N$ is closed and infinite. There exist $x_0 > 0$ and $y_0 < 1$ such that if $|x| > x_0$ and $y < y_0$, then*

$$\phi(n_x a_y) \ll \left(\frac{y}{x^2+y^2}\right)^s,$$

as $|x| \to \infty$ and $y \to 0$.

The proof of this fact is reminiscent of the arguments given in Patterson [Pat75] and Lax-Phillips [LP82] showing that a square-integrable eigenfunction of the Laplacian acting on an infinite volume surface must have eigenvalue $\lambda < 1/4$, i.e. the spectrum above $1/4$ is purely continuous. The key ingredient is that being L^2 forces an asymptotic formula for the rate of decay of the eigenfunction as it approaches the free boundary in the flare.

A.1. **Fourier Expansion in the Flare.** Recall that a "flare" in the fundamental domain is a region bounded by two geodesics, containing a free boundary. Concretely, after conjugation in $\mathrm{SL}_2(\mathbb{R})$, we can assume that our group Γ contains the fixed hyperbolic cyclic subgroup generated by the element $\begin{pmatrix} \sqrt{k} & 0 \\ 0 & 1/\sqrt{k} \end{pmatrix} : z \mapsto kz$. So a flare domain is isometric to a domain of the form $\{z : 1 < |z| < k; 0 < \arg z < \alpha\}$, where $\alpha < \pi/2$. See Fig. 5.

Such a conjugation sends the point at infinity to some point $\xi \in [1, \kappa]$, and a horocycle at infinity to a circle h tangent to ξ. See Fig. 6.

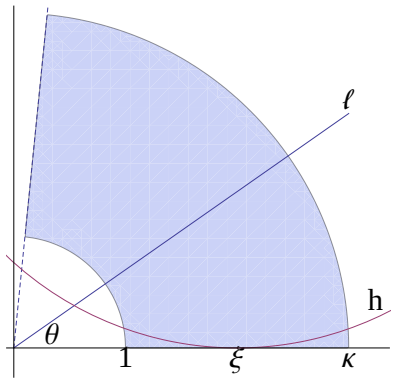

FIGURE 6. The horocycle h is based at the point ξ which is in the free boundary. The line ℓ intersects the real line at angle θ.

Let $\phi \in L^2(\Gamma\backslash\mathbb{H})$, $||\phi||_2 = 1$, with $\Delta\phi = s(1-s)\phi$. We proceed with the Fourier development of ϕ using polar coordinates in this domain. As we are in the flare, $\phi(\kappa z) = \phi(z)$, $\kappa > 1$. Write $z = re^{i\theta}$ in polar coordinates and separate variables:

$$\phi(z) = \phi(r,\theta) = f(r)g(\theta),$$

with $f(\kappa r) = f(r)$. Expand f in a (logarithmic) Fourier series:

$$\phi(r,\theta) = \sum_{n\in\mathbb{Z}} g_n(\theta) e^{2\pi i n \log r/\log\kappa}.$$

Then the solution to the differential equation induced on g_n is (see e.g. [Gam02], pages 180–181):

$$g_n(\theta) = c_n\sqrt{\sin\theta}P_\nu^\mu(\cos\theta),$$

where $c_n \in \mathbb{C}$ are some coefficients and P_ν^μ is the associated Legendre function of the first kind with

$$\begin{aligned} \mu &= \frac{1}{2} - s \\ \nu &= -\frac{1}{2} + \frac{2\pi i n}{\log\kappa}. \end{aligned} \tag{A.2}$$

We have proved

Proposition A.3. *There are some coefficients $c_n \in \mathbb{C}$ such that*

$$\phi(r,\theta) = \sum_n c_n\ e^{2\pi i n \log r/\log\kappa}\sqrt{\sin\theta}P_\nu^\mu(\cos\theta) \tag{A.4}$$

in the flare, with $1 \le r \le \kappa$ and $0 \le \theta \le \alpha < \pi/2$.

Next we need bounds on the coefficients c_n.

A.2. **Bounds on the Fourier Coefficients.**

Proposition A.5. *The coefficients c_n in* (A.4) *satisfy*

$$c_n \ll n^s e^{-\pi n\alpha/\log\kappa},$$

as $n \to \infty$. The implied constant depends on α and κ.

Proof. In Cartesian coordinates $z = x + iy$, the Haar measure is $dz = \frac{dxdy}{y^2}$. In polar coordinates this becomes

$$dz = \frac{drd\theta}{r\sin^2\theta}.$$

Input the expansion (A.4) into $||\phi|| = 1$, and consider only the contribution from the flare domain. This gives

$$\begin{aligned} 1 &\geq \int_1^\kappa \int_0^\alpha |\phi(r,\theta)|^2 \frac{drd\theta}{r\sin^2\theta} \\ &= \log\kappa \sum_n |c_n|^2 \int_0^\alpha |P_\nu^\mu(\cos\theta)|^2 \frac{d\theta}{\sin\theta}. \\ &\geq \log\kappa \sum_n |c_n|^2 \int_{\alpha/2}^\alpha |P_\nu^\mu(\cos\theta)|^2 \frac{d\theta}{\sin\theta}, \end{aligned} \tag{A.6}$$

where we have decreased the range of integration by positivity.

By Stirling's formula and the values of μ and ν in (A.2),

$$\frac{\Gamma(\nu+\mu+1)}{\Gamma(\nu+\frac{3}{2})} \gg n^{-s}, \tag{A.7}$$

for $n \gg 1$. The implied constants depend only on κ.

Next we record the elementary bound

$$\cos\left[(\nu+\frac{1}{2})\theta + (\mu-\frac{1}{2})\frac{\pi}{2}\right] \gg e^{2\pi n\theta/\log\kappa}. \tag{A.8}$$

Finally we have the formula (see [GR07] p.1003 #3),

$$P_\nu^\mu(\cos\theta) = \frac{2}{\sqrt{\pi}} \frac{\Gamma(\nu+\mu+1)}{\Gamma\left(\nu+\frac{3}{2}\right)} \frac{\cos\left[(\nu+\frac{1}{2})\theta + (\mu-\frac{1}{2})\frac{\pi}{2}\right]}{\sqrt{2\sin\theta}} \left(1 + O\left(\frac{1}{\nu}\right)\right), \tag{A.9}$$

valid whenever

(P1) $\mu \in \mathbb{R}$, $|\nu| \gg 1$,
(P2) $|\nu| \gg |\mu|$,
(P3) $|\arg\nu| < \pi$,
(P4) $0 < \varepsilon < \theta < \pi - \varepsilon$, and
(P5) $|\nu| \gg \frac{1}{\varepsilon}$.

The big-Oh constant in (A.9) depends only on the implied constants in (P1)–(P5).

The conditions (P1) and (P2) are immediately satisfied from the values of μ and ν in (A.2). The argument of ν approaches $\frac{\pi}{2}$ for n large, so (P3) is easily satisfied. We will use this formula for θ in a fixed range away from zero, $\theta \in [\alpha/2, \alpha]$. Thus (P4) is satisfied, and (P5) is equivalent to (P1).

Since θ is bounded away from zero, so is the factor $\sin\theta$ in the denominator of (A.9). Putting together (A.7), (A.8) and (A.9) gives

$$|P_\nu^\mu(\cos\theta)| \gg n^{-s} e^{2\pi n\theta/\log\kappa} \gg n^{-s} e^{\pi n\alpha/\log\kappa}. \tag{A.10}$$

Returning to (A.6), consider the contribution from just the Nth coefficient and use (A.10)

$$\begin{aligned} 1 &\geq \log\kappa \sum_n |c_n|^2 \int_{\alpha/2}^{\alpha} |P_\nu^\mu(\cos\theta)|^2 \frac{d\theta}{\sin\theta} \\ &\geq |c_N|^2 \int_{\alpha/2}^{\alpha} |P^{1/2-s}_{-1/2+2\pi i N/\log\kappa}(\cos\theta)|^2 d\theta \\ &\gg |c_N|^2 N^{-2s} e^{2\pi N\alpha/\log\kappa}, \end{aligned}$$

as $N \to \infty$.

This completes the proof of Proposition A.5. □

A.3. **Radial Bounds for the Eigenfunction.** Next we get bounds on the eigenfunction ϕ as the angle θ decreases to zero.

Proposition A.11. *Let ϕ be as above with eigenvalue $\lambda = s(1-s)$, $s > 1/2$. Then*

$$\phi(r,\theta) \ll \theta^s,$$

as $\theta \to 0$.

Proof. For θ small,

$$\sin\theta \asymp \theta,$$

and

$$1 - \cos\theta \asymp \theta^2.$$

We require some more estimates. First we use the following standard bound on the Gauss hypergeometric series:

$$F(a,b,c;x) = 1 + O(\left|\frac{abx}{c}\right|),$$

valid for

$$|x| \max_{\ell\in\mathbb{Z}} \left|\frac{(a+\ell)(b+\ell)}{(c+\ell)(1+|\ell|)}\right| \leq \frac{1}{2}.$$

In particular, with

$$\begin{aligned} a &= -\nu = \frac{1}{2} - \frac{2\pi\imath n}{\log\kappa}, \\ b &= 1+\nu = \bar{a}, \\ c &= 1-\mu = \frac{1}{2} + s, \text{ and} \\ x &= \frac{1-\cos\theta}{2} \ll \theta^2, \end{aligned}$$

the above gives:

$$F(-\nu, 1+\nu, 1-\mu; \frac{1-\cos\theta}{2}) \ll 1, \tag{A.12}$$

whenever

$$n \ll \frac{1}{\theta}.$$

We require [GR07] p. 999 formula 8.702:

$$P_\nu^\mu(z) = \frac{1}{\Gamma(1-\mu)} \left(\frac{1+z}{1-z}\right)^{\mu/2} F(-\nu, \nu+1; 1-\mu; \frac{1-z}{2}).$$

For $z = \cos\theta$, we have

$$\left(\frac{1+z}{1-z}\right)^{\mu/2} \asymp \theta^{-\mu} = \theta^{s-1/2},$$

so together with (A.12) we arrive at

$$P_\nu^\mu(\cos\theta) \ll \theta^{-\mu} = \theta^{s-1/2} \tag{A.13}$$

for $n \ll \frac{1}{\theta}$.

The analysis leading to (A.9) also gives

$$|P_\nu^\mu(\cos\theta)| \ll n^{-s} e^{2\pi n\theta/\log\kappa}\theta^{-1/2} \tag{A.14}$$

in the range $n \gg \frac{1}{\theta}$.

Thus we split the Fourier series as follows:

$$\begin{aligned}\phi(r,\theta) &= \sum_n c_n e^{2\pi\imath n\log r/\log\kappa}\sqrt{\sin\theta}P_\nu^\mu(\cos\theta)\\ &= \sum_{n\le X} + \sum_{n>X}\\ &= S_1 + S_2,\end{aligned}$$

with $X \asymp \frac{1}{\theta}$.

On S_1 we use the bound (A.13):

$$\begin{aligned}|S_1| &\le \sum_{n\le X} |c_n|\sqrt{\sin\theta}|P_\nu^\mu(\cos\theta)|\\ &\ll \theta^{1/2}\theta^{-\mu}\sum_n |c_n|\\ &\ll \theta^s,\end{aligned}$$

by the exponential decay of c_n (clearly the series converges).

On S_2, we use the bound (A.14):

$$\begin{aligned}|S_2| &\le \sum_{n>X} |c_n|\sqrt{\sin\theta}|P_\nu^\mu(\cos\theta)|.\\ &\ll \theta^{1/2}\sum_{n>X} n^s e^{-\pi n\alpha/\log\kappa} n^{-s} e^{2\pi n\theta/\log\kappa}\theta^{-1/2}\\ &\ll \sum_{n>X} e^{-\pi n(\alpha-2\theta)/\log\kappa}\\ &\ll \exp(-\pi X(\alpha-2\theta)/\log\kappa)\\ &\ll \exp(-\frac{1}{\theta}\pi\alpha/\log\kappa),\end{aligned}$$

since $X \asymp \frac{1}{\theta}$.

Combining the exponential decay of S_2 with the polynomial decay of S_1, we arrive at

$$\phi(r,\theta) \ll \theta^s,$$

as $\theta \to 0$.

This completes the proof of Proposition A.11. □

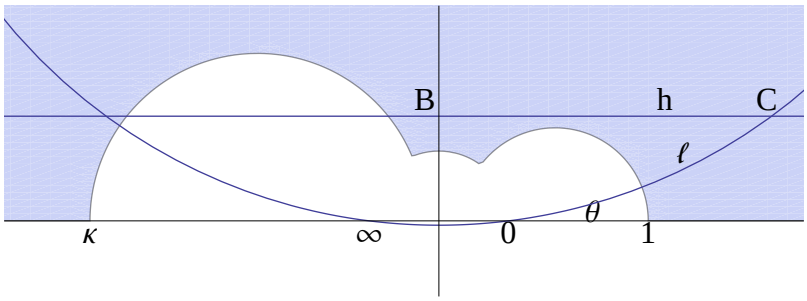

FIGURE 7. The image of Fig. 6 after a conformal transformation sending $\xi \mapsto \infty$, $0 \mapsto 1$, and $\infty \mapsto -1$. The point C is a point of intersection of h and ℓ, and B is the intersection of h with the y-axis.

A.4. **Proof of Theorem A.1.** Finally, we return to cartesian coordinates to convert the bound above into Theorem A.1. We require the following geometric analysis. Recall Fig. 6, where ξ is a point on the free boundary of the fundamental domain, h is a horocycle tangent to ξ, and θ is the angle between the real line and the line ℓ from zero to infinity intersecting the horocycle, h at a point C.

To return to cartesian coordinates, we redraw our picture after the conformal mapping

$$z \to \frac{z+\xi}{-z+\xi},$$

which sends the triple $(0, \infty, \xi) \mapsto (1, -1, \infty)$. See Fig. 7.

Lines and circles are mapped to lines and circles, and angles of incidence are preserved. The horocycle tangent to ξ is now the horizontal line, h (tangent to ξ, which has been mapped to infinity). Similarly, the line ℓ from zero to infinity passing through the horocycle is now a circle passing through the same points, having the same angle of incidence, θ, with the real line.

We reconstruct this configuration yet again in Fig. 8. Let A be the center of the circle ℓ, having radius $R = R(\theta)$, let B be the intersection of the horocycle h with the y-axis, C the intersection of the horocycle with the circle, and let D denote the origin.

It is easy to see through elementary geometry (since $\overline{A0}$ is tangent to the circle) that angle $0AD = \theta$. Looking at triangle $0AD$, we see that

$$R \sin\theta = \overline{0D} = 1. \tag{A.15}$$

Let $x = x(\theta)$ represent the length of $\overline{BC}$, and $y = y(\theta)$ be the distance from B to D. We aim to compute the precise dependence of x and y on θ.

We collect two identities for $\overline{AB}$:

$$\begin{aligned} \overline{AB} &= \overline{AD} - \overline{BD} = R\cos\theta - y, \text{ and} \\ \overline{AB}^2 &= \overline{AC}^2 - \overline{BC}^2 = R^2 - x^2. \end{aligned}$$

This implies

$$x^2 = R^2 - (R\cos\theta - y)^2 = R^2 \sin^2\theta + 2Ry\cos\theta - y^2,$$

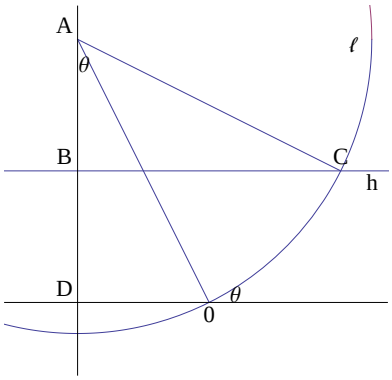

FIGURE 8. A cleaner version of Fig. 7. The point A is the center of the circle ℓ, and D denotes the origin.

which together with (A.15) gives

$$\frac{x^2+y^2-1}{2y} = \frac{\cos\theta}{\sin\theta} \asymp \frac{1}{\theta}, \tag{A.16}$$

as $\theta \to 0$.

Thus (A.16), together with Proposition A.11, concludes the proof of Theorem A.1.

REFERENCES

[Aub82] Thierry Aubin. *Nonlinear analysis on manifolds. Monge-Ampère equations*, volume 252 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York, 1982.

[Bea83] Alan F. Beardon. *The Geometry of Discrete Groups*, volume 91 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1983.

[BG08] Jean Bourgain and Alex Gamburd. Uniform expansion bounds for Cayley graphs of $\mathrm{SL}_2(\mathbb{F}_p)$. *Ann. of Math. (2)*, 167(2):625–642, 2008.

[BGS06] Jean Bourgain, Alex Gamburd, and Peter Sarnak. Sieving and expanders. *C. R. Math. Acad. Sci. Paris*, 343(3):155–159, 2006.

[BGS09] Jean Bourgain, Alex Gamburd, and Peter Sarnak. Generalization of Selberg's theorem and Selberg's sieve, 2009. Preprint.

[BGS10] Jean Bourgain, Alex Gamburd, and Peter Sarnak. Affine linear sieve, expanders, and sum-product. *Invent. Math.*, 179(3):559–644, 2010.

[CHH88] M. Cowling, U. Haagerup, and R. Howe. Almost L^2 matrix coefficients. *J. Reine Angew. Math.*, 387:97–110, 1988.

[Dal00] F. Dal'bo. Topologie du feuilletage fortement stable. *Ann. Inst. Fourier (Grenoble)*, 50(3):981–993, 2000.

[DH97] H. Diamond and H. Halberstam. Some applications of sieves of dimension exceeding 1. In *Sieve methods, exponential sums, and their applications in number theory (Cardiff, 1995)*, volume 237 of *London Math. Soc. Lecture Note Ser.*, pages 101–107. Cambridge Univ. Press, Cambridge, 1997.

[DH08] Harold G. Diamond and H. Halberstam. *A higher-dimensional sieve method*, volume 177 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2008. With an appendix ("Procedures for computing sieve functions") by William F. Galway.

[DHR88] H. Diamond, H. Halberstam, and H.-E. Richert. Combinatorial sieves of dimension exceeding one. *J. Number Theory*, 28(3):306–346, 1988.

[Gam02] Alex Gamburd. On the spectral gap for infinite index "congruence" subgroups of $\mathrm{SL}_2(\mathbf{Z})$. *Israel J. Math.*, 127:157–200, 2002.

[Gam09] A. Gamburd. Private communication, 2009.
[GR07] I.S. Gradshteyn and I.M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, 2007.
[GT10] Benjamin Green and Terence Tao. Linear equations in primes. *Ann. of Math. (2)*, 171(3):1753–1850, 2010.
[GV88] Ramesh Gangolli and V. S. Varadarajan. *Harmonic analysis of spherical functions on real reductive groups*, volume 101 of *Ergebnisse der Mathematik und ihrer Grenzgebiete [Results in Mathematics and Related Areas]*. Springer-Verlag, Berlin, 1988.
[HL22] G. H. Hardy and J. E. Littlewood. Some problems of 'Partitio Numerorum': III. on the expression of a number as a sum of primes. *Acta Math.*, 44:1–70, 1922.
[HR74] H. Halberstam and H.-E. Richert. *Sieve methods*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], London-New York, 1974. London Mathematical Society Monographs, No. 4.
[Iwa78] Henryk Iwaniec. Almost-primes represented by quadratic polynomials. *Invent. Math.*, 47:171–188, 1978.
[KO08] A. Kontorovich and H. Oh. Apollonian circle packings and closed horospheres on hyperbolic 3-manifolds, 2008. With an appendix by H. Oh and N. Shah. To appear, JAMS. *http://arxiv.org/abs/0811.2236*.
[Kon07] A. V. Kontorovich. *The Hyperbolic Lattice Point Count in Infinite Volume with Applications to Sieves*. Columbia University Thesis, 2007.
[Kon09] A. V. Kontorovich. The hyperbolic lattice point count in infinite volume with applications to sieves. *Duke J. Math.*, 149(1):1–36, 2009. *http://arxiv.org/abs/0712.1391*.
[KS03] H. Kim and P. Sarnak. Refined estimates towards the Ramanujan and Selberg conjectures. *J. Mar. Math. Soc.*, 16:175–181, 2003.
[Lan53] Edmund Landau. *Handbuch der Lehre von der Verteilung der Primzahlen. 2 Bände*. Chelsea Publishing Co., New York, 1953. 2d ed, With an appendix by Paul T. Bateman.
[Lan02] Serge Lang. *Algebra*, volume 211 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, third edition, 2002.
[LP82] P.D. Lax and R.S. Phillips. The asymptotic distribution of lattice points in Euclidean and non-Euclidean space. *Journal of Functional Analysis*, 46:280–350, 1982.
[LS07] Jianya Liu and Peter Sarnak. Integral points on quadrics in three variables whose coordinates have few prime factors, 2007. Preprint.
[Pat75] S. J. Patterson. The Laplacian operator on a Riemann surface. *Compositio Math.*, 31(1):83–107, 1975.
[Pat76] S.J. Patterson. The limit set of a Fuchsian group. *Acta Mathematica*, 136:241–273, 1976.
[Sar81] Peter Sarnak. Asymptotic behavior of periodic orbits of the horocycle flow and Eisenstein series. *Comm. Pure Appl. Math.*, 34(6):719–739, 1981.
[SS58] A. Schinzel and W. Sierpiński. Sur certaines hypothèses concernant les nombres premiers. *Acta Arith. 4 (1958), 185–208; erratum*, 5:259, 1958.
[War72] Garth Warner. *Harmonic analysis on semi-simple Lie groups. I*. Springer-Verlag, New York, 1972. Die Grundlehren der mathematischen Wissenschaften, Band 188.

E-mail address: alexk@math.brown.edu

Mathematics department, Brown University, Providence, RI and Institute for Advanced Study, Princeton, NJ

E-mail address: heeoh@math.brown.edu

Mathematics department, Brown University, Providence, RI and Korea Institute for Advanced Study, Seoul, Korea