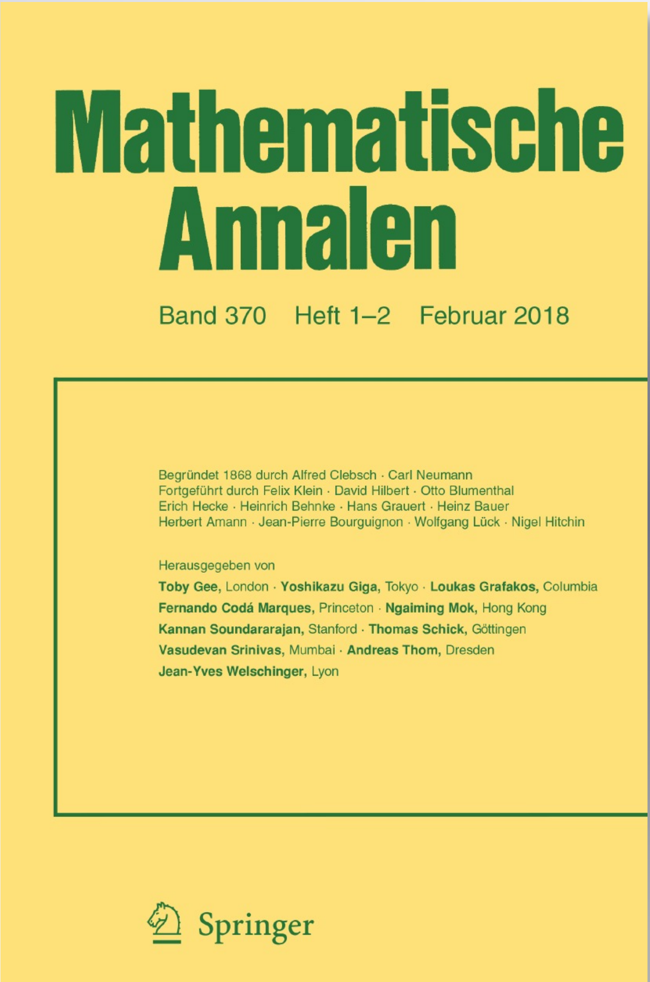
Effective equidistribution of shears and applications

Dubi Kelmer & Alex Kontorovich

Mathematische Annalen

ISSN 0025-5831
Volume 370
Combined 1-2

Math. Ann. (2018) 370:381-421
DOI 10.1007/s00208-017-1580-9



Math. Ann. (2018) 370:381–421
https://doi.org/10.1007/s00208-017-1580-9

Mathematische Annalen

CrossMark

Effective equidistribution of shears and applications

Dubi Kelmer[1] · **Alex Kontorovich**[2]

Received: 11 November 2015 / Revised: 1 August 2017 / Published online: 19 August 2017

Abstract A "shear" is a unipotent translate of a cuspidal geodesic ray in the quotient of the hyperbolic plane by a non-uniform discrete subgroup of $\mathrm{PSL}(2,\mathbb{R})$, possibly of infinite co-volume. We prove the regularized equidistribution of shears under large translates with effective (that is, power saving) rates. We also give applications to weighted second moments of GL(2) automorphic L-functions, and to counting lattice points on locally affine symmetric spaces.

Contents

Communicated by A. Venkatesh.

Kelmer is partially supported by NSF Grant DMS-1237412 and NSF Grant DMS-1401747.
Kontorovich is partially supported by an NSF CAREER Grant DMS-1254788/DMS-1455705, an NSF FRG Grant DMS-1463940, and an Alfred P. Sloan Research Fellowship.

✉ Alex Kontorovich
alex.kontorovich@rutgers.edu

Dubi Kelmer
kelmer@bc.edu

1 Boston College, Boston, MA, USA

2 Rutgers University, New Brunswick, NJ, USA

1 Introduction

In this paper, we prove the effective (meaning, with power savings rate) equidistribution of "shears" (see below for definitions) of "cuspidal" geodesic rays on hyperbolic surfaces. Our proofs are quite "soft," in that we only use mixing and standard properties of Eisenstein series, rather than explicit spectral decompositions, special functions, or any estimates on time spent near a cusp. This allows us to extend our methods to surfaces of infinite volume (in fact the proofs are surprisingly easier in this case). As a direct consequence, we complete the general problem of obtaining effective asymptotics for counting (in archimedean norm balls) discrete orbits on affine quadrics; as discribed in Sect. 1.3.1, exactly two lacunary settings remained unsolved, which are settled in this paper. Another application is to weighted second moments of GL(2) automorphic L-functions.

When the surface has infinite volume, we discover two new and completely unexpected phenomena: (1) the orbit count asymptotic can be proved with a uniform power savings error in congruence towers without inputting any information on the spectral gap.[1] And even more surprisingly, (2) orbits in such towers are not uniformly distributed among different cosets! The uniformity in cosets, were it true, would have allowed the application of an Affine Sieve in this archimedean ordering (see, e.g. [29]); our observation shows that the Affine Sieve procedure cannot be applied directly here, as the standard sieve axioms are not satisfied.[2]

1.1 Statements of the theorems

Our main equidistribution result is the following. Let Γ be a discrete, Zariski-dense,[3] geometrically finite[4] subgroup of $G := \mathrm{PSL}_2(\mathbb{R})$, and assume that the hyperbolic surface $\Gamma\backslash\mathbb{H}$, which may have finite or infinite volume, has at least one cusp. In particular, this forces the critical exponent[5] δ of Γ to exceed $1/2$; this will be our running assumption throughout.

The base point

$$\mathbf{x}_0 \in T^1(\Gamma\backslash\mathbb{H}) \cong \Gamma\backslash G$$

in the unit tangent bundle determines the visual (under the forward geodesic flow) limit point $\mathfrak{a}$ on the boundary $\Gamma\backslash\partial\mathbb{H}$. We call the point $\mathbf{x}_0$, as well as its forward geodesic ray, $\mathbf{x}_0 \cdot A^+$, **cuspidal** if $\mathfrak{a}$ is a cusp of Γ; here

$$A^+ := \{\begin{pmatrix} a & \\ & a^{-1} \end{pmatrix} : a > 1\}.$$

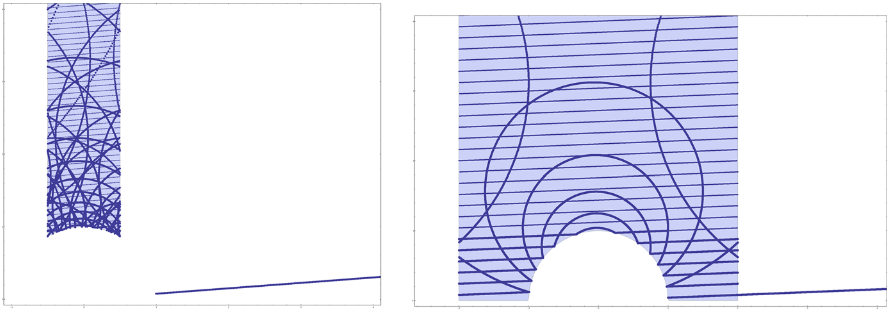
[1] By "spectral gap" we always mean the distance between the first eigenvalue λ_1 and the base eigenvalue λ_0 of the hyperbolic Laplacian; see Sect. 2.5.

[2] Of course one can instead order by wordlength in Γ, as is done in [3], to restore equidistribution and apply the Affine Sieve.

[3] Equivalently, non-elementary, that is, not virtually abelian.

[4] For surface groups, being geometrically finite is equivalent to being finitely generated.

[5] Roughly speaking, the critical exponent measures the asymptotic growth rate of Γ; see Sect. 2.1.



(a) **(b)**

Fig. 1 A shear of the cuspidal geodesic ray (**a**) Lattice case: $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$ (**b**) Thin case: $\Gamma = \left\langle \begin{pmatrix} 1 & 4 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \right\rangle$

(Note that we make no demands on the negative geodesic flow from $\mathbf{x}_0$.) Given such a ray, we define its **shear** (the ray is no longer geodesic), at time $T \in \mathbb{R}$, by:

$$\mathbf{x}_0 \cdot A^+ \cdot \mathfrak{s}_T \ \subset \ \Gamma\backslash G,$$

where

$$\mathfrak{s}_T \ := \ a_{\frac{1}{\sqrt{T^2+1}}} n_T, \quad a_y \ = \ \begin{pmatrix} \sqrt{y} & \\ & 1/\sqrt{y} \end{pmatrix}, \quad n_x \ = \ \begin{pmatrix} 1 & x \\ & 1 \end{pmatrix}. \tag{1.1}$$

For example, if $\Gamma = \mathrm{SL}_2(\mathbb{Z})$, then the base point $\mathbf{x}_0 = (i, \uparrow)$ has visual limit point $\mathfrak{a} = \infty$, and hence is cuspidal. The shear at time T of the forward ray from $\mathbf{x}_0$, projected to $\Gamma\backslash\mathbb{H}$, is then simply the Euclidean ray $\{re^{i\theta}\}_{r>1}$, where $\cot\theta = T$. See Figure 1a for an illustration of this ray and its projection mod $\mathrm{PSL}_2(\mathbb{Z})$. Similarly, Figure 1b gives the same picture but for a thin group Γ.

We are interested in the behavior of such shears as $T \to \infty$ (and similarly for $T \to -\infty$). To this end, define the measure μ_T on a smooth, compactly supported observable $\Psi \in C_0^\infty(\Gamma\backslash G)$ by

$$\mu_T(\Psi) \ := \ \int_{a\in A^+} \Psi(\mathbf{x}_0 \cdot a \cdot \mathfrak{s}_T) da \ = \ \int_1^\infty \Psi(\mathbf{x}_0 \cdot a_y \cdot \mathfrak{s}_T)\frac{dy}{y}. \tag{1.2}$$

A slight simplification of our main result (see Theorem 3.1) is the following

Theorem 1.1
Lattice case: *Assume that the quotient $\Gamma\backslash G$ has finite volume. Then there exists an $\eta > 0$, depending on the spectral gap for Γ, such that*

$$\mu_T(\Psi) \ = \ \log T \cdot \mu_{\Gamma\backslash G}(\Psi) + \mu_{\widetilde{Eis}}(\Psi) + O_\Psi\left(T^{-\eta}\right), \tag{1.3}$$

as $T \to \infty$. Here $\mu_{\Gamma\backslash G}$ is the probability Haar measure and $\mu_{\widetilde{Eis}}$ is a certain "regularized Eisenstein" distribution (see Remark 1.4 below).

Thin case: *Assume that Γ is thin, that is, the quotient $\Gamma\backslash G$ has infinite volume. Then there exists an $\eta > 0$, depending only on the critical exponent δ of Γ, and* not *on its spectral gap (!), such that*

$$\mu_T(\Psi) \;=\; \mu_{Eis}(\Psi) + O_\Psi\left(T^{-\eta}\right), \tag{1.4}$$

as $T \to \infty$. Here μ_{Eis} is an (un-regularized) Eisenstein distribution.

Some comments are in order.

Remark 1.2 For simplicity, we have stated Theorem 1.1 for compactly supported test functions Ψ, but our method applies just as well to a larger class of square-integrable functions with at least polynomial decay in the cusp $\mathfrak{a}$ (to ensure convergence of $\mu_T(\Psi)$); see Sect. 3.

Remark 1.3 Throughout we make no attempt to optimize the various error exponents η, as can surely be done with a modicum of effort; our point is to illustrate a soft method which is powerful enough to obtain new results with power savings errors.

Remark 1.4 Let us make the Eisenstein distributions arising in (1.3)–(1.4) less mysterious. These distributions are actually measures when Ψ is right K-invariant; we restrict attention to this case to simplify the discussion below. First assume that Γ is a lattice, that $\mathbf{x}_0 = (i, \uparrow)$ with $\mathfrak{a} = \infty$ a cusp of Γ of width 1, and let $\Gamma_\infty = \left(\begin{smallmatrix}1 & \mathbb{Z}\\ & 1\end{smallmatrix}\right)$ be the isotropy group of $\mathfrak{a}$ in Γ. Then one has the standard Eisenstein series

$$E(z,s) \;:=\; \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \mathfrak{Im}(\gamma z)^s, \quad (\mathfrak{Re}(s) > 1),$$

which is well known [46] to have meromorphic continuation and a simple pole at $s = 1$ with residue $\mathrm{vol}(\Gamma\backslash\mathbb{H})^{-1}$. Thus the function

$$\widetilde{E}(z,s) \;:=\; E(z,s) \;-\; \frac{1}{\mathrm{vol}(\Gamma\backslash\mathbb{H})\,(s-1)} \tag{1.5}$$

is regular at $s = 1$; for example, when $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$, we have (see, e.g., [25, (22.42), (22.63)–(22.69)])

$$\widetilde{E}(z,1) \;=\; \frac{3}{\pi}\left(2\gamma - 2\frac{\zeta'}{\zeta}(2) - \log\left(4y|\eta(z)|^4\right)\right), \tag{1.6}$$

where $\gamma = 0.577\cdots$ is Euler's constant, $\zeta(s)$ is the Riemann zeta function, and $\eta(z)$ is the Dedekind eta function. Then the measure $\mu_{\widetilde{Eis}}$ is simply given by:

$$\mu_{\widetilde{Eis}}(\Psi) \;=\; \left\langle \Psi, \widetilde{E}(\cdot,1)\right\rangle_{\Gamma\backslash\mathbb{H}}. \tag{1.7}$$

Note that $\log|\eta(z)|$ grows like y in the cusp, so $\mu_{\widetilde{Eis}}$ is also a non-finite measure. See (3.31) for the definition when Ψ is not K-finite.

In the thin case of (1.4), the Eisenstein series is itself regular at $s = 1$, that is, we can simply take $\widetilde{E}(z, s) = E(z, s)$; the spectral contribution is then *all* of lower order, so the power savings obtained in (1.4) is *independent* of any knowledge of a spectral gap for Γ.

Remark 1.5 The first factor $\log T$ on the right side of (1.3) is a manifestation of the logarithmic divergence of the measure μ_T. In the lattice case, a statement of the form

$$\mu_T(\Psi) \ = \ \textit{polynomial}(\log T) \cdot \mu_{\Gamma\backslash G}(\Psi) \cdot \Big(1 \ + \ o(1)\Big) \tag{1.8}$$

was suggested in work of Duke-Rudnick-Sarnak [15, see below (1.4)]. Recently, Oh-Shah [39] used a purely dynamical method to prove (a variant of) (1.8) with a log-savings rate, that is, with $o(1)$ replaced by $O(1/\log T)$. With such a rate it is of course impossible to see the second-order main term (that is, the regularized Eisenstein distribution), and this identification will be key to some of our applications below. Moreover, it is hard to imagine how a quantity like (1.6) can be captured using only dynamics; our approach is quite different.

Before discussing the proof of Theorem 1.1, we first give some applications.

1.2 Application 1: weighted second moments of GL(2) L-functions

Integrals like $\mu_T(\Psi)$ arise naturally in Sarnak's approach ("changing the test vector") [45] for second moments of L-functions (see also, e.g., [12,13,19,35,53]). We illustrate the method in the simplest case of a weight-k holomorphic Hecke cusp form f on $\mathrm{PSL}_2(\mathbb{Z})$, though the method works just as well for any GL(2) automorphic representation.

Write the Fourier expansion of f as

$$f(z) = \sum_{n\geq 1} a_f(n)e(nz),$$

where $a_f(1) = 1$ and the coefficients $a_f(n)$ are multiplicative, satisfying Hecke relations, and the Ramanujan bound $\left|a_f(p)\right| \leq 2p^{(k-1)/2}$ [11]. The standard L-function of f,

$$L(f, s) \ := \ \sum_{n\geq 1} \frac{a_f(n)}{n^{s+(k-1)/2}},$$

converges for $\mathfrak{Re}(s) > 1$, has analytic continuation to $\mathbb{C}$, and a functional equation sending $s \mapsto 1 - s$. The Rankin-Selberg L-function factors (see [26, (13.1)]) as

$$L(f \otimes \bar{f}, s) \ := \ \sum_{n\geq 1} \frac{\left|a_f(n)\right|^2}{n^{s+(k-1)}} \ = \ \frac{\zeta(s)}{\zeta(2s)} L(\mathrm{sym}^2 f, s),$$

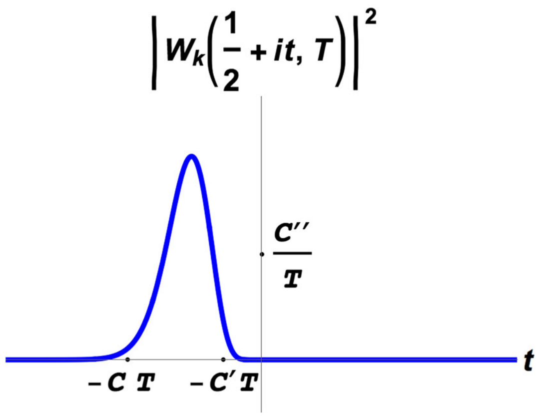


Fig. 2 A sample graph of the smoothed archimedean weight $|\mathcal{W}_k(\frac{1}{2}+it,T)|^2$

where $L(\mathrm{sym}^2 f, s)$ is the symmetric square L-function, known [18] to be the automorphic L-function of a cohomological form on GL(3).

A shear of the standard Hecke integral (already arising implicitly in classical work of Titchmarsch [51, Chap. VII]) is the following calculation:

$$\int_0^\infty f\,(Ty+iy)\,y^{s+(k-1)/2}\frac{dy}{y} = L(f,s)\mathcal{W}_k(s,T), \tag{1.9}$$

where

$$\mathcal{W}_k(s,T) := (2\pi)^{-\left(s+(k-1)/2\right)}\,\Gamma\left(s+\tfrac{k-1}{2}\right)\,(1-iT)^{-\left(s+(k-1)/2\right)}. \tag{1.10}$$

Applying Parseval to (1.9) gives (for $s=1/2+it$)

$$\int_0^\infty \left|f(Ty+iy)\right|^2 y^k\frac{dy}{y} = \frac{1}{2\pi}\int_{\mathbb{R}}\left|L(f,\tfrac12+it)\right|^2\left|\mathcal{W}_k(\tfrac12+it,T)\right|^2 dt. \tag{1.11}$$

A calculation with Stirling's formula (or see Fig. 2) shows that $|\mathcal{W}_k(\frac12+it,T)|^2$ has rapid decay as soon as $|t|>T^{1+\varepsilon}$, and is of size roughly $1/T$ in the bulk. Thus the quantity on the right side of (1.11) behaves like a smoothed second moment of $L(f,s)$ on the critical line. Applying Theorem 1.1 with $\Psi=|f|^2y^k$ (and a little more work, see §4) gives the following.

Theorem 1.6 *With notation as above, there is an $\eta>0$ such that*

$$\begin{aligned}
&\frac{1}{2\pi}\int_{\mathbb{R}}\left|L(f,\tfrac12+it)\right|^2\left|\mathcal{W}_k(\tfrac12+it,T)\right|^2 dt\\
&\quad = 2\frac{\|f\|^2}{\mathrm{vol}(\Gamma\backslash\mathbb{H})}\left(\log T \;+\; \frac{\Lambda'}{\Lambda}(\mathrm{sym}^2 f,1)+\gamma-2\frac{\zeta'}{\zeta}(2)\right)\\
&\qquad + O_f(T^{-\eta}),
\end{aligned} \tag{1.12}$$

as $T \to \infty$. Here γ is again Euler's constant, $\|f\|$ is the Petersson norm, and Λ'/Λ is the logarithmic derivative of the completed symmetric-square L-function,

$$\Lambda(\mathrm{sym}^2 f, s) = (4\pi)^{-(s+k-1)} \Gamma(s+k-1) L(\mathrm{sym}^2 f, s). \tag{1.13}$$

Remark 1.7 One can chase the various exponents in our proof to see that (1.12) holds with $\eta = 1/14 - \varepsilon$. Again, we are striving for simplicity of the method and not optimal exponents, see Remark 1.3. In fact, a straightforward refinement of the proof of Proposition 2.2 (using explicit spectral expansions in place of soft ergodic arguments) gives $\eta = 16/39 - \varepsilon$ on quoting the best-known bounds [30] towards the Ramanujan conjectures, and $\eta = 1/2 - \varepsilon$ conditionally. So in a sense, the proof of Theorem 1.6 is "sharp," as there is no "loss" in the rate from a best-possible one.

Remark 1.8 On comparing the lower order terms on the right hand side of (1.12) with the secondary term in (1.3), and using (1.6) and (1.7), one derives a Kronecker-type limit formula, in the form:

$$\frac{\left\langle \log\left(4y|\eta(z)|^4\right), |f|^2 y^k \right\rangle}{\|f\|^2} = \gamma - \frac{\Lambda'}{\Lambda}(\mathrm{sym}^2 f, 1). \tag{1.14}$$

This identity is surely classical, though we were not able to locate a precise reference.

1.3 Application 2: Archimedean counting for orbits on affine quadrics

Another standard context where integrals like $\mu_T(\Psi)$ arise naturally is in the execution of certain Margulis/Duke-Rudnick-Sarnak/Eskin-McMullen type arguments [15,16,33] for counting discrete orbits on quadrics in archimedean balls. The setting is as follows.

Let Q be a real ternary indefinite quadratic form (e.g., $Q(\mathbf{x}) = x^2 + y^2 - z^2$), fix $d \in \mathbb{R}$, and denote by $V = V_{Q,d}$ the affine quadric

$$V : Q = d. \tag{1.15}$$

The real points $V(\mathbb{R})$ enjoy an action by $G = \mathrm{SO}^\circ_Q(\mathbb{R})$, the connected component of the identity in the real special orthogonal group preserving Q. Let $\Gamma < G$ be a discrete, Zariski dense, geometrically finite subgroup of G, and assume, as throughout, that the critical exponent δ of Γ exceeds $1/2$. Fix a base point $\mathbf{x}_0 \in V(\mathbb{R})$, subject to the orbit

$$\mathcal{O} := \mathbf{x}_0 \cdot \Gamma \subset \mathbb{R}^3$$

being discrete.

For a fixed archimedean norm $\|\cdot\|$ on $\mathbb{R}^3$, let

$$B_T = \{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\| < T\}$$

Table 1 The new cases of Theorem 1.9, highlighted, are those with $C_1 > 0$

(Γ, Γ_H)	H		
	N	A	K
(Lattice, lattice)	[43,54]	[15,16]	[10,24,46]
(Lattice, thin)	Impossible by discreteness of $\mathcal{O}$	"Lacunary" case settled in (1.16)	Impossible by compactness of K
(Thin, lattice)	[28]	[6]	[31]
(Thin, thin)	[27]	$\begin{cases} \text{[6]}, & \text{if both } \xi_\pm \notin \Lambda, \\ \text{"Lacunary" case settled in (1.17)}, & \text{otherwise,} \end{cases}$	Impossible by compactness of K

be the norm ball of radius T. A very classical and well-studied problem is to give an effective (that is, with power savings error) estimate for

$$\mathcal{N}_{\mathcal{O}}(T) := |\mathcal{O} \cap B_T|.$$

Despite the vast attention this problem has received over the years, there remained exactly two lacunary cases in which hitherto resisted solution; see §1.3.1 and Table 1 below for a detailed taxonomy of the situation. Equipped with Theorem 1.1, we can now resolve the outstanding cases.

Theorem 1.9 *There exist constants C_1, C_2, and $\eta > 0$ such that the following holds:*

- *If Γ is a lattice in G, then*

$$\mathcal{N}_{\mathcal{O}}(T) = \big(C_1 T \log T + C_2 T\big)\big(1 + O(T^{-\eta})\big). \tag{1.16}$$

- *If Γ is thin, then*

$$\mathcal{N}_{\mathcal{O}}(T) = \big(C_1 T + C_2 T^{\delta}\big)\big(1 + O(T^{-\eta})\big). \tag{1.17}$$

Some comments are in order:

Remark 1.10 (1) All of the previously resolved cases of this problem (in the above generality) were such that the first term did not appear, that is, $C_1 = 0$ (whence $C_2 > 0$). Our new contributions are to the cases with $C_1 > 0$, which arise exactly when the stabilizer of $\mathbf{x}_0$ in G is diagonalizable but "cuspidal" in $\Gamma\backslash G$; see §1.3.1 below.

(2) If it happens that $\mathcal{O}$ is not just some arbitrary real discrete orbit but is actually the full integer quadric $V(\mathbb{Z})$ (assuming of course that the quadratic form Q is rational and that $V(\mathbb{Z})$ is non-empty), then one has many more tools available to approach the counting problem for $\mathcal{N}_{\mathcal{O}}(T)$. Specifically, one can use, e.g., classical methods of exponential sums (see [23]), or half-integral weight automorphic forms, Poincaré series and shifted convolutions [7,44,52], or multiple Dirichlet series [22]. These give, when Γ is a lattice and $C_1 > 0$, an estimate for $\mathcal{N}_{\mathcal{O}}(T)$ of the

same strength as (1.16), that is, with a secondary main term and a power savings error. Such tools do not seem to apply to the general orbit counting problem.

(3) In the thin case with $C_1 > 0$, the second term $C_2 T^\delta$ is swamped by the error, and should not be confused with a lower order "main" term. We would like to acknowledge here that Nimish Shah suggested to us that the main term in this setting is of order T rather than T^δ; see also [38, Remark 1.7].

(4) For the "new" cases with $C_1 > 0$, the exponent η depends on the same quantities as in Theorem 1.1; that is, η depends on the spectral gap in the lattice case, and only on the critical exponent in the thin case.

(5) The constants C_1, C_2 can be readily determined explicitly in terms of volumes, special values of (possibly regularized) Eisenstein series, and Patterson-Sullivan measures.

(6) A consequence of Oh-Shah's result discussed in Remark 1.5 gives, in the lattice case, the estimate (1.16), but with $O(T^{-\eta})$ replaced by the weaker error rate $O(1/(\log T)^\eta)$ for some small $\eta > 0$. This of course only identifies the first main term $C_1 T \log T$, as the secondary term $C_2 T$ is swallowed by the error.[6]

1.3.1 Taxonomy

To explain the lacunary cases settled by Theorem 1.9, we begin by passing from $\mathrm{SO}^\circ_Q(\mathbb{R})$ to its spin cover $\mathrm{PSL}_2(\mathbb{R}) \cong T^1(\mathbb{H})$. Abusing notation, we continue to write G and Γ for their pre-images in $\mathrm{PSL}_2(\mathbb{R})$.

Let H be the stabilizer of $\mathbf{x}_0$ in G,

$$H \ := \ \{h \in G : \mathbf{x}_0 \cdot h = \mathbf{x}_0\},$$

and let

$$\Gamma_H := \Gamma \cap H.$$

With $\mathbf{x}_0$ fixed, the norm $\|\cdot\|$ on $\mathbb{R}^3$ induces a left-H invariant norm $|||\cdot|||$ on G given by

$$|||g||| = \|\mathbf{x}_0 g\|. \tag{1.18}$$

We further abuse notation, writing B_T for the left-H-invariant norm-T ball in G, that is, $B_T \subset H\backslash G$. Then it is easy to see that

$$\mathcal{N}_{\mathcal{O}}(T) \ = \ |\Gamma_H\backslash\Gamma \ \cap \ B_T|.$$

To investigate this counting problem more precisely, we illustrate the geometry of B_T, which is determined by whether the stabilizer H is conjugate to the groups K, N, or A. That is, H is either a maximal compact, a unipotent subgroup, or diagonalizable

[6] Added in print: In private communication, Hee Oh has notified us that she and Nimish Shah have an unpublished manuscript in which they obtain a result similar to (1.16) in the lattice case.

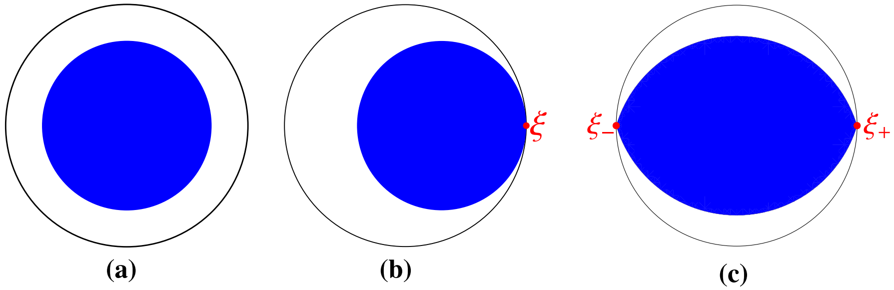


Fig. 3 The region B_T as a subset of the hyperbolic disk $\mathbb{D}$ (**a**) Case $H \cong K$ (**b**) Case $H \cong N$ (**c**) Case $H \cong A$

(over $\mathbb{R}$), and this corresponds to whether the real quadric $V(\mathbb{R})$ is a two-sheeted hyperboloid, a cone, or a one-sheeted hyperboloid, respectively. To visualize B_T as a left-H-invariant subset of G, we project to the base space $\mathbb{H}$ (or alternatively, assume that the norm $|||\cdot|||$ is right-K-invariant), so that B_T can be viewed as an H-invariant subset of the hyperbolic disk $\mathbb{D} \cong G/K$. Then B_T is illustrated in Fig. 3 in the three cases. Note that B_T has zero, one, or two limit points (denoted ξ or $\xi_\pm$) on the boundary $\partial\mathbb{D}$, corresponding to whether $H \cong K$, $H \cong N$, or $H \cong A$, respectively.

The asymptotic counting analysis depends in a fundamental way not only on whether Γ is a lattice in G, but also on whether

$$\Gamma_H \text{ is a lattice in } H. \tag{1.19}$$

Lemma 1.11 *If* (1.19) *does not hold, then the discreteness of* $\mathcal{O}$ *is equivalent to the endpoints* ξ *or* $\xi_\pm$ *of* H *not being radial limit points*[7] *for* Γ.

This follows from a simple topological argument; we omit the proof. We decompose the analysis according to whether Γ is a lattice or thin in G.

Case I: Γ is a lattice

Assuming that Γ is a lattice in G, and also demanding that (1.19) holds, Duke–Rudnick–Sarnak [15] and Eskin–McMullen [16] (see also [33]) showed (in much greater generality than considered here) that there is some $\eta > 0$ with

$$\mathcal{N}_{\mathcal{O}}(T) = \frac{\mathrm{vol}_H(\Gamma_H\backslash H)}{\mathrm{vol}_G(\Gamma\backslash G)} \mathrm{vol}_{H\backslash G}(B_T)\left(1 + O(T^{-\eta})\right), \tag{1.20}$$

as $T \to \infty$. Here the volumes are taken to be compatible with choices of Haar measure on G, H, and $H\backslash G$. Note that $\mathrm{vol}_{H\backslash G}(B_T)$ is of order T, so there is no logarithmic divergence in (1.20), that is, $C_1 = 0$ and $C_2 > 0$ in (1.16); see also Remark 1.10(1).

With Fig. 3 and Lemma 1.11 in mind, we analyze separately the possible conjugacy classes of H.

[7] Recall that the limit set, Λ, of Γ decomposes disjointly into cusps (i.e., parabolic fixed points) and radial limit points (also called "points of approximation"); the complement $\partial\mathbb{H}\backslash\Lambda$ is called the free boundary (which is empty if Γ is a lattice). See §2.1.

- First if $H \cong K$ is compact, then (1.19) clearly holds automatically. In this case, the counting result (1.20) corresponds to counting in norm balls of G, which dates back to Delsarte [10], Huber [24], and Selberg [46].
- If $H \cong N$ is unipotent, then by Lemma 1.11, the boundary point ξ of H must be a cusp. That is, $\Gamma_H \backslash H$ is a closed horocycle, so Γ_H is a lattice in H, and (1.19) is again automatically satisfied. In this case, the count takes place in a strip $\Gamma_H \backslash G$, and the equidistribution of low-lying closed horocycles [33,43,54] can be used to establish (1.20).
- Lastly, if $H \cong A$ is diagonalizable (over $\mathbb{R}$), then Lemma 1.11 forces one of two settings. Either
  - (i): Γ_H is a lattice in H, whence $\Gamma_H \backslash H$ corresponds to a closed geodesic on $\Gamma \backslash G$. Then (1.19) holds, so (1.20) follows from [15]. Or
  - (ii): Γ_H is finite, but both limit points ξ_+ and ξ_- of H (see Figure 3c) are cusps of Γ. Here we are in the diagonalizable but "cuspidal" setting referred to in Remark 1.10(1). fig:BT This case is the only one (when Γ is a lattice in G) not satisfying (1.19) despite the discreteness of the orbit $\mathcal{O}$; it is precisely the new case settled by Theorem 1.9.

Case II: Γ is thin

In this setting, we again decompose the problem of estimating $\mathcal{N}_{\mathcal{O}}(T)$ into separate cases, depending on the conjugacy class of H, and on whether condition (1.19) holds (there are now more situations in which $\mathcal{O}$ is discrete but (1.19) can fail).

- When $H \cong K$ is compact, the condition (1.19) is again automatically satisfied, and in this case Lax-Phillips [31] showed that

$$\mathcal{N}_{\mathcal{O}}(T) = C_2\, T^{\delta}\left(1 + O(T^{-\eta})\right), \tag{1.21}$$

  where $\eta > 0$ depends on the spectral gap for Γ. This corresponds to the case $C_1 = 0$ in (1.17); again, see Remark 1.10(1).
- If $H \cong N$ is unipotent, the discreteness of $\mathcal{O}$ forces one of two cases. Either
  - (i) Γ_H is a lattice in H, so (1.19) holds, and $\Gamma_H \backslash H$ corresponds to a closed horocycle. In this case, the asymptotic formula is (1.21) was shown in the second-named author's thesis [28]. Or
  - (ii) Γ_H is trivial, whence Lemma 1.11 forces the limit point ξ of H to be in the free boundary, that is, it is not in the limit set of Γ. The asymptotic here is also of the form (1.21); see [27].
- Finally, when $H \cong A$ is diagonalizable, there are three separate cases to consider. The discreteness of $\mathcal{O}$ now implies either
  - (i) Γ_H is a lattice in H, again corresponding to a closed geodesic on $\Gamma \backslash G$. Then the same asymptotic (1.21) follows from now-standard methods using the equidistribution result of Bourgain-Kontorovich-Sarnak [6]. Or
  - (ii) Γ_H is thin in H, in which case each of the two endpoints $\xi_\pm$ of H is either a cusp of Γ or in the free boundary. If
    - (a) both endpoints $\xi_\pm$ are in the free boundary, then the methods of [6] can again be used to show the same asymptotic (1.21). (The key is that only a

finite portion of the sheared geodesic ray interacts with the limit set — see [27] where a similar phenomenon was studied in the case of a unipotent stabilizer.) Otherwise,

(b) at least one of $\xi_\pm$ is a cusp of Γ. This is the other new lacunary case of Theorem 1.9, and is the only thin case for which (1.17) has $C_1 > 0$. Note that if one boundary point is a cusp and the other is in the free boundary, the former has contribution of order T, while the latter's contribution, of order T^δ, is dominated by the former's error term.

This concludes our taxonomy. To summarize, the following table serves to illustrate that Theorem 1.9 above completes the effective solution to the general counting orbital problem in our context:

Remark 1.12 When the critical exponent $\delta \leq 1/2$, work of Naud [36], extending Dolgopyat's methods [14], allows one to conclude, in the cases not excluded by Lemma 1.11, an effective asymptotic of the form (1.21). So the lacunary cases do not occur here, as there are no cusps (and hence no cuspidal geodesic rays) when $\delta \leq 1/2$.

Remark 1.13 As pointed out in [39, p. 917] (at least for Γ a lattice), the only lacunary cases in the more general setting of Q having signature (n, m) are precisely those of signature $(2, 1)$, that is, those considered here; so we have lost no generality in restricting to $\mathrm{PSL}_2(\mathbb{R})$. This is because the stabilizer H is either unipotent, compact, or fixes a form of signature $(n-1, m)$ or $(n, m-1)$. The only non-compact such not generated by unipotents has signature $(1, 1)$, whence Q has signature $(2, 1)$.

1.4 Surprise: Non-equidistribution in congruence cosets!

The most unexpected consequence of Theorem 1.9 comes from studying the thin case in cosets of congruence towers, as we now describe. Assume that Q is not just a real quadratic form but an integral one, and that Γ is a subgroup of the *integral* special orthogonal group

$$\Gamma \ < \ \mathrm{SO}_Q(\mathbb{Z}).$$

Given an integer $q \geq 1$, we can then form the level-q "congruence" subgroups $\Gamma(q) < \Gamma$, defined as

$$\Gamma(q) \ := \ \ker(\Gamma \to \mathrm{SO}_Q(\mathbb{Z}/q\mathbb{Z})).$$

For many applications, one wishes to study the same counting problem as above, with the orbit $\mathcal{O} = \mathbf{x}_0 \cdot \Gamma$ replaced by the congruence orbit $\mathbf{x}_0 \cdot \Gamma(q)$, or better yet, by some "congruence coset" orbit,

$$\mathcal{O}_{q,\varpi} \ := \ \mathbf{x}_0 \cdot \varpi \cdot \Gamma(q),$$

for a given $\varpi \in \Gamma/\Gamma(q)$. Let

$$\mathcal{N}_{q,\varpi}(T) \ := \ |\mathcal{O}_{q,\varpi} \ \cap \ B_T|$$

be the corresponding counting function, which we wish to estimate *uniformly* with q and T (and ϖ) varying in some allowable range.

Theorem 1.9 applies just as well to estimate $\mathcal{N}_{q,\varpi}(T)$, and in all previously studied examples, the asymptotic analysis showed that

$$\mathcal{N}_{q,\varpi}(T) \sim \frac{1}{[\Gamma : \Gamma(q)]}\mathcal{N}_{\mathcal{O}}(T), \quad (T \to \infty), \tag{1.22}$$

that is, the asymptotic is independent of ϖ, so the orbits are *equidistributed* among congruence cosets. (Moreover, (1.22) even holds with q allowed to grow sufficiently slowly with T.) This equidistribution is a key input, for example, in executing an Affine Sieve in an archimedean ordering (see, e.g., [4,5,21,27,28,32,34,37]). An analysis of Theorem 1.9 shows that, for thin orbits, there are cosets which are *not* uniformly distributed in archimedean balls!

Proposition 1.14 *Assume that Γ is thin, and that the orbit $\mathcal{O}$ is has diagonalizable and cuspidal stabilizer, that is, $C_1 > 0$ in (1.17). Then the equidistribution (1.22) in congruence cosets is* **false**. *For example, for each fixed q, there is some $\varpi \in \Gamma/\Gamma(q)$ such that*

$$\mathcal{N}_{q,\varpi}(T) \gg \frac{1}{q}\, T, \tag{1.23}$$

while

$$\frac{1}{[\Gamma : \Gamma(q)]}\mathcal{N}_{\mathcal{O}}(T) \ll \frac{1}{q^3}\, T,$$

as $T \to \infty$. (The implied constants above may depend on Γ and $\mathbf{x}_0$, but obviously not on q or T.)

This means that the standard Affine Sieve procedure *cannot* be executed in this ordering. (Note that the case considered in Proposition 1.14 is precisely the one omitted in the sieving statement [34, Cor 1.19].)

1.5 Outline of the proofs and paper

Our proof of Theorem 1.1 is surprisingly simple, and proceeds in two stages. The first is to show that $\mu_T(\Psi)$ in some sense equidistributes in the "strip $\Gamma_\infty\backslash\mathbb{H}$", but with respect to $dx\,dy/y$, as opposed to Haar measure $dx\,dy/y^2$ (for a hint of this, look again at Figure 1); this fact uses only the decay of the Fourier coefficients of Ψ (itself a simple consequence of mixing via low-lying horocycles; see Proposition 2.2). Then stage two is to relate this equidistribution to Eisenstein series, where we mimic Sarnak's approach in [43,54] to conclude the proof.

The rest of the paper proceeds as follows. In Sect. 2, we set notation and recall basic facts needed through the paper. Then in Sect. 3, we prove the main equidistribution result, Theorem 1.1, and its generalization (Theorem 3.1). Then Theorem 1.6 is proved in Sect. 4 using the Rankin-Selberg "unfolding" technique. Finally, Theorem 1.9 and Proposition 1.14 are proved in Sect. 5.

1.6 Notation

Constants $0 < C < \infty$ and $0 < \eta < 1$ can change from line to line, and $\varepsilon > 0$ represents an arbitrarily small quantity. The transpose of a matrix g is written ${}^\top g$. Unless otherwise specified, implied constants depend at most on Γ, which is treated as fixed. The symbol $\mathbf{1}_{\{\cdot\}}$ represents the indicator function of the event $\{\cdot\}$.

2 Preliminaries

In this section, we set all notation and basic facts used throughout.

2.1 Hyperbolic Geometry

Let $\mathbb{H} := \{z \in \mathbb{C} : \mathfrak{Im} z > 0\}$ denote the hyperbolic upper half-plane. At each point $z \in \mathbb{H}$, and tangent vector $\zeta \in T_z\mathbb{H} \cong \mathbb{C}$, the Riemannian structure is $\|\zeta\|_z := |\zeta|/\mathfrak{Im} z$. The unit tangent bundle $T^1\mathbb{H}$ is then

$$T^1\mathbb{H} := \{(z, \zeta) \in \mathbb{H} \times \mathbb{C} : \|\zeta\|_z = 1\}.$$

The group $G = \mathrm{PSL}_2(\mathbb{R})$ acts on $T^1\mathbb{H}$ via

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} : (z, \zeta) \mapsto \left(\frac{az + b}{cz + d}, \frac{\zeta}{(cz + d)^2} \right),$$

and moreover we can identify $G \cong T^1\mathbb{H}$ under $g \mapsto g(i, \uparrow)$. We also use the disk model $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$, identified with $\mathbb{H}$ under the map

$$\mathbb{H} \ni z \mapsto (z - i)/(z + i) \in \mathbb{D}.$$

Let Γ be a finitely generated, Zariski dense, discrete subgroup of G. As above, we identify $T^1(\Gamma\backslash\mathbb{H}) \cong \Gamma\backslash G$. For a fixed base point $\mathfrak{o} \in \mathbb{H}$, the **critical exponent**

$$\delta = \delta(\Gamma) \in [0, 1]$$

of Γ is the abscissa of convergence of the Poincaré series

$$\sum_{\gamma \in \Gamma} \exp(-s d(\gamma\mathfrak{o}, \mathfrak{o})), \quad (\mathfrak{Re}(s) > \delta).$$

Here $d(\cdot, \cdot)$ is hyperbolic distance, and δ does not depend on the choice of $\mathfrak{o}$. Let dg be a choice of Haar measure on G; we call Γ a **lattice** if $\Gamma\backslash G$ has finite measure, and **thin** otherwise. This is measured by the critical exponent δ, as $\delta = 1$ or $\delta < 1$ exactly when Γ is a lattice or thin, respectively [41]; the Zariski-density of Γ implies that $\delta > 0$. The **limit set**

$$\Lambda = \Lambda(\Gamma) \subset \partial\mathbb{H} \cong S^1 \cong \mathbb{R} \sqcup \{\infty\}$$

of Γ is the set of limit points of $\gamma\mathfrak{o}$, $\gamma \in \Gamma$; it also does not depend on the choice of $\mathfrak{o}$. The Hausdorff dimension of Λ is exactly equal to the critical exponent δ [41,50]. A boundary point $\xi \in \partial\mathbb{H}$ is a **cusp** of Γ if it is the fixed point of a parabolic element in Γ; these all lie in the limit set Λ, and we let Λ_{cusp} denote the subset of cusps. A limit point $\xi \in \Lambda$ is called **radial** (or a "point of approximation") if there is a sequence $\{\gamma_n\mathfrak{o}\}$, $\gamma_n \in \Gamma$, which stays a bounded distance away from a geodesic ray ending at ξ. Let Λ_{rad} denote the set of radial limit points; it is a basic fact [2] that the limit set decomposes disjointly into radial and cuspidal points,

$$\Lambda = \Lambda_{cusp} \sqcup \Lambda_{rad}.$$

The complement of the limit set in $\partial\mathbb{H}$ is called the **free boundary** of Γ,

$$\mathcal{F} = \mathcal{F}(\Gamma) := \partial\mathbb{H}\backslash\Lambda,$$

and $\mathcal{F} = \varnothing$ if and only if Γ is a lattice. We record here the decomposition

$$\partial\mathbb{H} = \mathcal{F} \sqcup \Lambda_{cusp} \sqcup \Lambda_{rad}. \tag{2.1}$$

We assume henceforth that Γ has at least one cusp, whence its critical exponent exceeds $1/2$,

$$\delta > 1/2. \tag{2.2}$$

2.2 Spectral theory

The hyperbolic Laplace operator $\Delta := -y^2(\partial_{xx} + \partial_{yy})$ acts (after unique extension) on the space $L^2(\Gamma\backslash\mathbb{H})$ of square-integrable automorphic functions, and is self-adjoint and positive semi-definite. Let $\Omega = \Omega(\Gamma) \subset [0, \infty)$ denote the spectrum of Δ. The assumption that Γ has at least one cusp implies the existence of continuous spectrum above $1/4$, that is, $[1/4, \infty) \subset \Omega$ (there may also be embedded discrete spectrum in this range, which only occurs when Γ is a lattice [40,46]). Below $1/4$ the spectrum is finite [31] and nonempty (by (2.2)); we denote these eigenvalues, often referred to as the "exceptional spectrum," by

$$0 \le \lambda_0 < \lambda_1 \le \cdots \le \lambda_{max} < \frac{1}{4},$$

and introduce spectral parameters $1/2 < s_j \leq 1$ defined by

$$\lambda_j = s_j(1 - s_j),$$

so that

$$1 \geq s_0 > s_1 \geq \cdots \geq s_{max} > \frac{1}{2}. \tag{2.3}$$

The bottom eigenvalue λ_0 is simple, and is related to the geometry of Λ via the Patterson-Sullivan formula [41,50]

$$\lambda_0 = \delta(1 - \delta),$$

that is, $s_0 = \delta$.

2.3 Algebra

We will use standard notation for the subgroups N, A, and K of G, given by:

$$N := \begin{pmatrix} 1 & \mathbb{R} \\ 0 & 1 \end{pmatrix}, \quad A := \{\mathrm{diag}(a, 1/a) : a > 0\}, \quad K := \mathrm{SO}(2), \tag{2.4}$$

and containing typical elements

$$n_x := \begin{pmatrix} 1 & x \\ & 1 \end{pmatrix}, \quad a_y := \begin{pmatrix} \sqrt{y} & \\ & 1/\sqrt{y} \end{pmatrix}, \quad k_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

As right actions, they correspond, respectively, to the unipotent flow, geodesic flow, and rotation of the tangent vector, keeping the base point fixed. On the other hand, as left actions, the correspond, respectively, to horizontal translation, scaling, and motion around a hyperbolic circle centered at i. Haar measure dg in Iwasawa coordinates $g = n_x a_y k_\theta$ is then given by $dg = dx\, y^{-2} dy\, d\theta$. The right-action by the semigroup $A^+ := \{a_y : y > 1\}$ corresponds to the positive geodesic flow, so that a given point $\mathbf{x}_0 \in G \cong T^1\mathbb{H}$ gives rise to the geodesic ray $\mathbf{x}_0 A^+$.

2.4 Representation theory

By the Duality Theorem [17], the spectral decomposition (2.3) corresponds to the decomposition of the right regular representation of G on $L^2(\Gamma\backslash G)$ as

$$L^2(\Gamma\backslash G) = V_0 \oplus V_1 \oplus \cdots \oplus V_{max} \oplus V_{temp}.$$

Here V_{temp} consists of the tempered spectrum (a reducible subspace); each V_j, $j = 1, \ldots, max$ is an irreducible complementary series representation of parameter

s_j; and V_0 is either the trivial representation (if Γ is a lattice), or a complementary series representation of parameter $s_0 = \delta$ (if Γ is thin).

We record here a Sobolev-norm version [8] of the exponential decay of matrix coefficients. Fix a basis $\mathscr{B} = \{X_1, X_2, X_3\}$ for the Lie algebra $\mathfrak{g}$ of G, and given a smooth test function $\Psi \in C^\infty(\Gamma\backslash G)$, define the "$L^p$, order-$d$" Sobolev norm $\mathcal{S}_{p,d}(\Psi)$ as

$$\mathcal{S}_{p,d}(\Psi) := \sum_{\mathrm{ord}(\mathscr{D}) \le d} \|\mathscr{D}\Psi\|_{L^p(\Gamma\backslash G)}.$$

Here $\mathscr{D}$ ranges over monomials in $\mathscr{B}$ of order at most d.

Theorem 2.1 ([9,47,53]) *Let (π, V) be a unitary G-representation, and assume there is a number $\Theta > 1/2$ so that V does not weakly contain any complementary series with parameter $s > \Theta$. Then for all smooth $v, w \in V^\infty$, we have*

$$|\langle \pi(g).v, w\rangle| \ll \|g\|^{2(1-\Theta)} \mathcal{S}_{2,1}(v)\mathcal{S}_{2,1}(w), \tag{2.5}$$

as $\|g\|^2 := \mathrm{tr}(g^\top g) \to \infty$. The implied constant is absolute.

Later we will encounter other Sobolev norms which are convex combinations of those above.

2.5 Uniform spectral gaps

Recalling the spectral decomposition (2.3), we call a number $\Theta \in (1/2, \delta)$ a **spectral gap** for Γ if $\Theta > s_1$. To make sense of a *uniform* such gap, we assume integrality. As in §1.4, if Γ consists of integer matrices, $\Gamma < \mathrm{PSL}_2(\mathbb{Z})$, we may, given an integer $q \ge 1$, define the level-q "congruence" subgroup

$$\Gamma(q) := \ker(\Gamma \to \mathrm{PSL}_2(\mathbb{Z}/q\mathbb{Z})).$$

Let $\Omega(q)$ be the spectrum of Δ on $L^2(\Gamma(q)\backslash\mathbb{H})$; clearly $\Omega \subset \Omega(q)$, and in general this inclusion is strict. We will call a number $\Theta \in (1/2, \delta)$ a **uniform spectral gap** for Γ if, for all $q \ge 1$,

$$\Omega(q) \cap (\delta(1-\delta), \Theta(1-\Theta)] = \varnothing, \tag{2.6}$$

that is, there are no eigenvalues at any level q in a neighborhood of the base eigenvalue $\lambda_0 = \delta(1-\delta)$. (Note that this definition is different from other related definitions in the literature.) In a number of statements below, several quantities depend on the *spectral gap* in the lattice case, but only on the *critical exponent* in the thin case; to unify the two notions, we will say that such quantities depend on the **first non-zero eigenvalue** of Γ.

2.6 Eisenstein series

In this section we recall some basic facts from the theory of Eisenstein series. We will assume here that the Eisenstein series is with respect to a cusp at ∞ and note that Eisenstein series corresponding to other cusps are defined similarly, after conjugating that cusp to ∞. In our applications, we will not have the flexibility to demand the cusp have width 1, so deal below with arbitrary width.

The Eisenstein series corresponding to a cusp at ∞ of width $\omega > 0$ is defined[8] in the half-plane $\mathfrak{Re}(s) > 1$ by the convergent series

$$E(z,s) := \frac{1}{\omega} \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \mathfrak{Im}(\gamma z)^s, \tag{2.7}$$

where $\Gamma_\infty = \left(\begin{smallmatrix} 1 & \omega\mathbb{Z} \\ 0 & 1 \end{smallmatrix}\right)$.

Assume first that Γ is a lattice. Then $E(z,s)$ has a meromorphic continuation to $\mathbb{C}$ with a functional equation sending $s \mapsto 1-s$. In fact, it is analytic in the half plane $\mathfrak{Re}(s) > 1/2$ except for a simple pole at $s = 1$ and perhaps finitely many poles

$$1 > \sigma_1 \geq \cdots \geq \sigma_h > 1/2. \tag{2.8}$$

These poles comprise the "residual spectrum," which is a subset of the "exceptional spectrum" in (2.3); the remaining spectrum in this range, if any, is cuspidal. The residue at $s = 1$ is

$$\mathrm{Res}_{s=1}\, E(s,z) = \frac{1}{\mathrm{vol}(\Gamma\backslash\mathbb{H})}$$

and

$$\varphi_{\sigma_j}(z) = \mathrm{Res}_{s=\sigma_j}\, E(s,z)$$

are the residual forms.

For any integer $n \in \mathbb{Z}$ we also define the weight $2n$ Eisenstein series by

$$E_n(z,s) := \frac{1}{\omega} \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \mathfrak{Im}(\gamma z)^s \epsilon_\gamma(z)^{2n}, \tag{2.9}$$

where

$$\epsilon_g(z) = \frac{cz+d}{|cz+d|}, \quad g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

[8] We should note that we are using a non-standard definition for the Eisenstein series, where we multiply the series by $\frac{1}{\omega}$ instead of by $\frac{1}{\omega^s}$. Using the standard definition instead will result in minor changes to the regularized Eisenstein series in the lattice case.

Unless the weight $n = 0$, the E_n's are all regular at $s = 1$, that is, $E_0 = E$ is the only Eisenstein series with a pole at $s = 1$. In the range $\frac{1}{2} < \mathfrak{Re}(s) < 1$, the poles of E_n are the same $s = \sigma_1, \ldots, \sigma_h$ as those of E. For each such pole $\sigma = \sigma_j$ we denote by

$$\varphi_{\sigma,n} = \mathrm{Res}_{s=\sigma}\, E_n(z, s), \tag{2.10}$$

the (un-normalized) residual form of weight $2n$.

We note for future reference that the weight-$2n$ Eisenstein series and the weight-$2n$ residual forms all lie in the space $(\Gamma, 2n)$ of functions on $\mathbb{H}$ transforming by

$$f(\gamma z) = \epsilon_\gamma(z)^{2n} f(z). \tag{2.11}$$

Still assuming that Γ is a lattice, we also have the following bounds coming from the spectral decomposition of $L^2(\Gamma\backslash G)$, see, e.g. [26]. For any square-integrable $f \in (\Gamma, 2n)$, we have the bound

$$\frac{1}{2\pi}\int_{\mathbb{R}} \left|\left\langle f, E_n\left(\cdot, \tfrac{1}{2} + ir\right)\right\rangle\right|^2 dr \leq \|f\|_2^2. \tag{2.12}$$

When Γ is thin, we will only use the fact that the defining series in (2.7) and (2.9) converge absolutely in the range $\mathfrak{Re}(s) > \delta$.

2.7 Decay of Fourier coefficients

In this subsection we wish to record the basic fact that the (parabolic) Fourier coefficients of an automorphic function decay in the cusp, in a uniform sense. The method we use to establish this is completely standard, though the requisite uniformity does not seem to be in the literature; hence we give sketches of proofs for the reader's convenience. We again assume that Γ has a cusp at ∞ of width $\omega > 0$, that is, the isotropy group Γ_∞ of ∞ is generated by the translation $z \mapsto z + \omega$.

Then a smooth, square-integrable, Γ-automorphic function $\Psi \in L^2 \cap C^\infty(T^1(\Gamma\backslash\mathbb{H}))$ has a Fourier expansion:

$$\Psi(x + iy, \zeta) = \sum_{m\in\mathbb{Z}} a_\Psi(m; y, \zeta)\, e_\omega(mx), \tag{2.13}$$

where $e_\omega(x) := e^{2\pi i x/\omega}$, and the Fourier coefficients are given by

$$a_\Psi(m; y, \zeta) := \frac{1}{\omega}\int_0^\omega \Psi(x + iy, \zeta) e_\omega(-mx)dx. \tag{2.14}$$

The next proposition records the decay of such as $y \to 0$ (in a uniform statement); the subscripts F below stand for "Fourier coefficients."

Proposition 2.2 *There is a "Sobolev" norm $\mathcal{S}_F(\Psi)$ and constants $0 < C_F < \infty$ and $0 < \alpha_F < 1$, such that, uniformly over all $y > 0$ and $m \in \mathbb{Z}\backslash\{0\}$, we have*

$$|a_\Psi(m; y, \zeta)| \ll \mathcal{S}_F(\Psi)\, |m|^{C_F}\, y^{\alpha_F}. \tag{2.15}$$

The constants α_F and C_F depend on the first non-zero eigenvalue of Γ.

The Sobolev norm and exact values of the constants C_F and α_F are given below in (2.20), (2.21), and (2.22), respectively; the last claim of the proposition is then clear, namely, that the constants depend on the spectral gap when Γ is a lattice, and only on the critical exponent when Γ is thin. As we are not striving for optimal exponents (recall Remark 1.3), we have chosen to suppress their precise values so as not to clutter the presentation. Recall also our convention that implied constants may depend at most on Γ, unless stated otherwise.

The Proposition is an easy consequence of another standard fact, namely the equidistribution of (pieces of) "low-lying" horocycles; the subscripts H below stand for "Horocycle pieces."

Proposition 2.3 *There is a "Sobolev" norm $\mathcal{S}_H(\Psi)$ and constants $C_H < \infty$ and $0 < \alpha_H < 1$, such that, uniformly over all $y > 0$ and open intervals $\mathcal{I} \subset (0, \omega)$, we have*

$$\frac{1}{|\mathcal{I}|}\int_{\mathcal{I}} \Psi(x + iy, \zeta)dx = \frac{1}{\mathrm{vol}(\Gamma\backslash G)}\int_{\Gamma\backslash G} \Psi\, dg + O\Big(\mathcal{S}_H(\Psi)\, |\mathcal{I}|^{-C_H}\, y^{\alpha_H}\Big). \tag{2.16}$$

This statement holds whether Γ is a lattice or not, with the interpretation that the first term on the right-hand-side of (2.16) vanishes in the thin case. The constants C_H and α_H depend on the first non-zero eigenvalue of Γ.

Again, the norm and constants are detailed in (2.17), (2.18), and (2.19), which we have suppressed in the interest of exposition. Much stronger versions of (2.16) exist in the literature (at least in the lattice case, for which see, e.g., [48,49]), but for the reader's convenience, we provide a quick

Proof of Proposition 2.3 (Sketch) We may assume that $y < 1$ (for the statement is obviously true otherwise), and we may moreover assume that Ψ is right-K-invariant (or replace Ψ by $\tilde{\Psi} := \pi(k_\theta^{-1})\Psi$, where 2θ is the angle of ζ measured counterclockwise from the vertical). Then the left hand side of (2.16) is

$$\mathscr{M} := \frac{1}{|\mathcal{I}|}\int_{x\in\mathcal{I}} \Psi(n_x a_y)dx.$$

Let ρ be a smooth, non-negative function on $\mathbb{R}$ with support in $[-1, 1]$ and $\int_{\mathbb{R}} \rho = 1$. For $\eta > 0$ to be chosen later, concentrate ρ to $\rho_\eta(x) := \eta^{-1}\rho(x/\eta)$. Write $\overline{N} := {}^\top N$, $\overline{n}_x := {}^\top n_x$ for the opposite horocyclic group and element. Then the multiplication map

$$N \times A \times \overline{N} \;\to\; G \;:\; (n, a, \overline{n}) \mapsto na\overline{n}$$

is bijective on an open neighborhood of the origin. Define $\xi : G \to \mathbb{R}_{\geq 0}$, supported in such a neighborhood, via:

$$\xi(n_x a_t \overline{n}_s) \;:=\; c_\xi \cdot \left(\rho_\eta \star \frac{1}{|\mathcal{I}|}\mathbf{1}_{\mathcal{I}}\right)(x) \cdot \rho_\eta(\log t) \cdot \rho_\eta(s),$$

where $\star$ denotes convolution, and $c_\xi \asymp 1$ is a constant (independent of η) chosen so that $\int_G \xi = 1$. Automorphize ξ to

$$\Xi(g) \;:=\; \sum_{\gamma \in \Gamma} \xi(\gamma g),$$

which is a function on $\Gamma\backslash G$ with $\int_{\Gamma\backslash G} \Xi = 1$. Finally, consider the matrix coefficient:

$$\mathscr{C} \;:=\; \left\langle \pi(a_y).\Psi,\, \Xi \right\rangle,$$

which we evaluate in two ways. Using the decay of matrix coefficients (2.5), we see immediately that

$$\mathscr{C} \;=\; \frac{1}{\mathrm{vol}(\Gamma\backslash G)} \int_{\Gamma\backslash G} \Psi + O\left(y^{1-\Theta} \mathcal{S}_{2,1}(\Psi)\mathcal{S}_{2,1}(\Xi)\right),$$

where $\Theta = s_1 + \varepsilon$ is a spectral gap in the lattice case, and $\Theta = \delta + \varepsilon$ in the thin case (in which case the "main" term vanishes). It is easy to estimate, crudely, that

$$\mathcal{S}_{2,1}(\Xi) \ll |\mathcal{I}|^{-1}\eta^{-3}.$$

For a second evaluation of $\mathscr{C}$, unfold the inner product to obtain

$$\mathscr{C} \;=\; \int_{NA\overline{N}} \Psi(n_x a_t \overline{n}_s a_y)\xi(n_x a_t \overline{n}_s) d\overline{n}_s da_t dn_x$$

The "wavefront lemma" (in this case, trivial) states that $\overline{n}_s a_y = a_y \overline{n}_{sy}$, and we estimate

$$\Psi(n_x a_y a_t \overline{n}_{sy}) \;=\; \Psi(n_x a_y) + O\left(\eta\, \mathcal{S}_{\infty,1}(\Psi)\right).$$

Hence

$$\begin{aligned}\mathscr{C} &= \int_{\mathbb{R}} \Psi(n_x a_t) \left(\rho_\eta \star \frac{1}{|\mathcal{I}|}\mathbf{1}_{\mathcal{I}}\right)(x)\, dx + O\left(\eta\, \mathcal{S}_{\infty,1}(\Psi)\right)\\ &= \mathscr{M} + O\left(\eta\, \mathcal{S}_{\infty,1}(\Psi)\right).\end{aligned}$$

Combining the errors and choosing $\eta = y^{(1-\Theta)/4}\mathcal{S}_{2,1}(\Psi)^{1/4}\,\mathcal{S}_{\infty,1}(\Psi)^{-1/4}|\mathcal{I}|^{-1/4}$ gives (2.16) with

$$\mathcal{S}_H(\Psi) := \mathcal{S}_{2,1}(\Psi)^{1/4}\,\mathcal{S}_{\infty,1}(\Psi)^{3/4}, \tag{2.17}$$

$$C_H := 1/4, \tag{2.18}$$

and

$$\alpha_H := \frac{1-\Theta}{4}, \tag{2.19}$$

as claimed. □

Equipped with Proposition 2.3, we may now give a quick

Proof of Proposition 2.2 Again we may assume that Ψ is right-K-invariant and $y < 1$. Let $J \geq 1$ be an integer parameter to be chosen later, and write

$$a_\Psi(m; y) = \frac{1}{\omega}\sum_{j=0}^{J-1}\int_{\omega j/J}^{\omega(j+1)/J}\Psi(x+iy)e_\omega(-mx)dx.$$

On each short interval, we estimate $e_\omega(-mx) = e(-mj/J) + O(|m|/J)$, whence

$$a_\Psi(m; y) = \sum_{j=0}^{J-1} e(-mj/J)\frac{1}{\omega}\int_{\omega j/J}^{\omega(j+1)/J}\Psi(x+iy)dx + O(\|\Psi\|_\infty|m|/J).$$

Now on each little integral, we apply the equidistribution of pieces of "low-lying" horocycles in the form (2.16), that is,

$$\frac{1}{\omega}\int_{\omega j/J}^{\omega(j+1)/J}\Psi(x+iy)dx = \frac{1}{J}\frac{1}{\mathrm{vol}(\Gamma\backslash G)}\int_{\Gamma\backslash G}\Psi\,dg + O\Big(\mathcal{S}_H(\Psi)\,y^{\alpha_H}J^{C_H-1}\Big).$$

Inserting this expression into $a_\Psi(m; y)$ and using $m \neq 0$, the roots of unity cancel out, leaving only error terms:

$$|a_\Psi(m; y)| \ll \mathcal{S}_H(\Psi)\,y^{\alpha_H}J^{C_H} + \|\Psi\|_\infty|m|\,J^{-1}.$$

Setting

$$J \asymp \Big(\mathcal{S}_H(\Psi)^{-1}\,y^{-\alpha_H}\,\|\Psi\|_\infty\,|m|\Big)^{1/(C_H+1)},$$

we arrive at (2.15) with

$$\mathcal{S}_F(\Psi) := \mathcal{S}_H(\Psi)^{1/(C_H+1)}\,\|\Psi\|_\infty^{C_H/(C_H+1)}, \tag{2.20}$$

$$C_F := C_H/(C_H+1), \tag{2.21}$$

and

$$\alpha_F := \alpha_H/(C_H + 1). \tag{2.22}$$

This completes the proof. □

Remark 2.4 It should be noted that actually Propositions 2.3 and 2.2 are *equivalent*, in the sense that one can also use the uniform decay of Fourier coefficients to prove a version of (2.3) (though with possibly worse exponents).

Remark 2.5 In the thin case, the proof of Proposition 2.2 can be made *much* simpler. Namely, one can first trivially bound the mth coefficient by the constant one, $|a_\Psi(m; y, \zeta)| \le |a_\Psi(0; y, \zeta)|$, and then use (2.16) with $\mathcal{I} = (0, \omega)$ to estimate the constant coefficient. (Note though that if Γ is a lattice, then of course $a_\Psi(0; y, \zeta)$ need not decay!)

3 Equidistribution of shears

Recall our running assumption that $\Gamma < G = \mathrm{PSL}_2(\mathbb{R})$ is a geometrically finite, Zariski dense, discrete group with at least one cusp, and hence critical exponent δ exceeding $1/2$. As in (1.2), we will study the limit as $|T| \to \infty$ of the measures

$$\mu_T(\Psi) := \int_{a \in A^+} \Psi(\mathbf{x}_0 \cdot a \cdot \mathfrak{s}_T)da.$$

To study the equidistribution of such, we need an appropriate space of test functions; in particular, we will require smoothness and at least polynomial decay at the cusp. Toward this end, for any cusp $\mathfrak{a}$ of Γ and integer $m \ge 1$, we introduce the space

$$\mathscr{P}^m_{\mathfrak{a}}(\Gamma\backslash G) \subset L^2 \cap C^\infty(\Gamma\backslash G)$$

of smooth, square-integrable, automorphic functions with the following added property. We will state it in the case $\mathfrak{a} = \infty$; for a general cusp $\mathfrak{a}$, conjugate $\mathfrak{a}$ to ∞ in the standard way.

We require that, for each $\Psi \in \mathscr{P}^m_\infty(\Gamma\backslash G)$, there are constants $1 \le C_\Psi < \infty$ and $0 < \alpha_\Psi$, such that

$$\left| \frac{\partial^j}{\partial\theta^j} \Psi(n a_y k_\theta) \right| < C_\Psi \, y^{-\alpha_\Psi}, \tag{3.1}$$

holds uniformly for all $j \le m$, $y > C_\Psi$, and all $n \in N$, $k \in K$. That is, after a certain point high up in the specified cusp, we have completely uniform polynomial decay in Ψ its first m derivatives in $\mathfrak{k} = \mathrm{Lie}(K)$. Note that we make no demands on decay properties (beyond square-integrability) in any other non-compact regions (cusps or possibly flares) of $\Gamma\backslash G$ besides the specified cusp $\mathfrak{a}$. Also note that the space $\mathscr{P}^m_{\mathfrak{a}}$

is non-empty, since, e.g., it contains the subspace of smooth, compactly supported functions, or better yet, cusp forms.

Our main theorem, from which Theorem 1.1 follows immediately, is the following.

Theorem 3.1 *Let* $\mathbf{x}_0 \cdot A^+$ *be a cuspidal geodesic ray ending in a cusp* $\mathfrak{a}$ *of* Γ*, and let* $\Psi \in \mathscr{P}^2_{\mathfrak{a}}(\Gamma\backslash G)$ *be a test function (i.e., assume* (3.1) *is satisfied for all* $j \leq 2$*). Then there is a finite-order "Sobolev" norm* $\mathcal{S}(\Psi)$ *(which depends on the constants* C_Ψ *and* α_Ψ *in* (3.1)*), and an* $\eta > 0$ *depending only on the first non-zero eigenvalue of* Γ*, such that: if* Γ *is a lattice,*

$$\mu_T(\Psi) \;=\; \log|T|\, \mu_{\Gamma\backslash G}(\Psi) + \mu_{\widetilde{Eis}}(\Psi) + O(\mathcal{S}(\Psi)T^{-\eta}),$$

and if Γ *is thin, then*

$$\mu_T(\Psi) \;=\; \mu_{Eis}(\Psi) + O(\mathcal{S}(\Psi)T^{-\eta}),$$

as $|T| \to \infty$*. Here* $\mu_{\Gamma\backslash G}(\Psi) := \mathrm{vol}(\Gamma\backslash G)^{-1}\int_{\Gamma\backslash G}\Psi$ *is the Haar probability measure,* $\mu_{\widetilde{Eis}}$ *is the "regularized Eisenstein" distribution given in* (3.31)*, and* μ_{Eis} *is the distribution given in* (3.25)*.*

As a first simplification, we can immediately apply an auxiliary conjugation to move $\mathbf{x}_0$ to the origin $e \cong (i, \uparrow)$, whence the cusp $\mathfrak{a}$ moves to ∞. Unfortunately, we have thus exhausted our free parameters, and cannot control the *width* of the resulting cusp, which we denote ω; that is, the isotropy group Γ_∞ is generated by the translation $z \mapsto z + \omega$.

As outlined in the introduction, the proof of Theorem 3.1 now proceeds in two stages, as encapsulated in the following two theorems.

Theorem 3.2 (Equidistribution in the "strip" $\mathfrak{S} = \Gamma_\infty\backslash G$) *For a test function* $\Psi \in \mathscr{P}^2_\infty(\Gamma\backslash G)$*, define the measure:*

$$\begin{aligned}\mu_{T,\mathfrak{S}}(\Psi) &:= \frac{1}{\mathrm{vol}(\Gamma_\infty\backslash N)}\int_{n\in\Gamma_\infty\backslash N}\int_{a\in A^+}\Psi(n\,a\,a_{1/T})da\,dn\\ &= \frac{1}{\omega}\int_0^\omega\int_{1/T}^\infty \Psi(n_x\,a_y)\frac{dy}{y}\,dx.\end{aligned} \tag{3.2}$$

(Recall that in this context, da_y *is* dy/y*,* not dy/y^2*.) Then there is a "Sobolev" norm* $\mathcal{S}_{\mathfrak{S}}(\Psi)$ *and a constant* $\alpha_{\mathfrak{S}} > 0$*, defined in* (3.18) *and* (3.19)*, respectively, such that*

$$\mu_T(\Psi) \;=\; \mu_{T,\mathfrak{S}}(\Psi) \;+\; O\left(\mathcal{S}_{\mathfrak{S}}(\Psi)T^{-\alpha_{\mathfrak{S}}}\right), \tag{3.3}$$

as $|T| \to \infty$*. Here* $\alpha_{\mathfrak{S}}$ *only depends on the first non-zero eigenvalue of* Γ*.*

Note that Theorem 3.2 makes no distinction between whether Γ is a lattice or thin. This dichotomy is only evident in the second stage:

Theorem 3.3 (Eisenstein distributions) *Let* $\Psi \in \mathscr{P}^2_\infty(\Gamma\backslash G)$ *as above.* **Lattice case:** *If* Γ *is a lattice in* G*, then there is a distribution* $\mu_{\widetilde{Eis}}$ *defined in* (3.31)*, and "residual" distributions* μ_{σ_j} *corresponding to* (2.8) *and defined in* (3.33)*, such that:*

$$\mu_{T,\mathfrak{S}}(\Psi) = \mu(\Psi)\log(T) + \mu_{\widetilde{Eis}}(\Psi) + \sum_{j=1}^{h} \frac{T^{\sigma_j-1}}{\sigma_j - 1}\mu_{\sigma_j}(\Psi) + O(\mathcal{S}_{2,1}(\Psi)T^{-1/2}).$$

Thin case: *If* Γ *is thin in* G*, then there is a distribution* μ_{Eis} *defined in* (3.25) *such that:*

$$\mu_{T,\mathfrak{S}}(\Psi) \;=\; \mu_{Eis}(\Psi) \;+\; O\Big(\mathcal{S}_H(\Psi)T^{-\alpha_H}\Big). \tag{3.4}$$

Here $\mathcal{S}_H$ *and* α_H *are as in Proposition* 2.3*.*

It is clear that Theorem 3.1 follows immediately from Theorems 3.2 and 3.3.

3.1 Stage 1: Proof of Theorem 3.2

We proceed with a series of elementary lemmata. Beginning with the definition (1.2), we express μ_T in terms of coordinates in $T^1(\Gamma\backslash\mathbb{H})$:

$$\mu_T(\Psi) \;:=\; \int_1^\infty \Psi\left(\frac{yT}{\sqrt{T^2+1}} + i\frac{y}{\sqrt{T^2+1}}, \uparrow\right)\frac{dy}{y}. \tag{3.5}$$

All of our manipulations below will not affect the direction of the tangent vector, so we drop the $\uparrow$. (Alternatively, pretend Ψ is right-K-invariant.)

Lemma 3.4 *With* C_Ψ *and* α_Ψ *from* (3.1)*, we let*

$$U \;>\; C_\Psi T \tag{3.6}$$

be a parameter to be chosen later in (3.17)*. Then*

$$\mu_T(\Psi) \;=\; \mathscr{M}_1 \;+\; O\left(\|\Psi\|_\infty T^{-2} + C_\Psi\left(\frac{T}{U}\right)^{\alpha_\Psi}\right), \tag{3.7}$$

where

$$\mathscr{M}_1 \;:=\; \int_1^U \Psi\left(y + i\frac{y}{T}\right)\frac{dy}{y}. \tag{3.8}$$

Proof From (3.5), make a change of variables $y \mapsto y\sqrt{T^2+1}/T$, and simplify to

$$\mu_T(\Psi) = \int_{T/\sqrt{T^2+1}}^{\infty} \Psi\left(y + i\frac{y}{T}\right)\frac{dy}{y} = \int_1^{\infty} \Psi\left(y + i\frac{y}{T}\right)\frac{dy}{y} + O\left(\|\Psi\|_\infty T^{-2}\right).$$

With U as in (3.6), break the range of integration $[1,\infty) = [1,U] \cup (U,\infty)$. On the latter range, apply (3.1), whence (3.7) follows. □

Now we invoke the Fourier expansion (2.13). Define

$$\Psi^{\perp}(x+iy) := \sum_{m\in\mathbb{Z}\setminus\{0\}} a_\Psi(m;y)\, e_\omega(mx),$$

so that

$$\Psi(x+iy) = a_\Psi(0;y) + \Psi^{\perp}(x+iy). \tag{3.9}$$

Inserting (3.9) into (3.8) splits $\mathscr{M}_1$ into a "main term" and "error":

$$\mathscr{M}_1 = \mathscr{M}_2 + \mathscr{E}_1,$$

where

$$\mathscr{M}_2 := \int_1^U a_\Psi\left(0; \frac{y}{T}\right)\frac{dy}{y}, \tag{3.10}$$

and

$$\mathscr{E}_1 := \int_1^U \Psi^{\perp}\left(y + i\frac{y}{T}\right)\frac{dy}{y}. \tag{3.11}$$

We first analyze $\mathscr{M}_2$.

Lemma 3.5 *Recalling the measure $\mu_{T,\mathfrak{S}}$ in (3.2), we have*

$$\mathscr{M}_2 = \mu_{T,\mathfrak{S}}(\Psi) + O\left(C_\Psi\left(\frac{T}{U}\right)^{\alpha_\Psi}\right) \tag{3.12}$$

Proof Inserting (2.14) into (3.10) gives

$$\mathscr{M}_2 = \int_1^U \frac{1}{\omega}\int_0^{\omega} \Psi\left(x + i\frac{y}{T}\right)dx\frac{dy}{y} = \frac{1}{\omega}\int_0^{\omega}\int_{1/T}^{U/T} \Psi(x+iy)\frac{dy}{y}dx.$$

Extending the y integral from U/T to ∞ and applying (3.1) again gives the claimed main and error terms in (3.12). □

Returning to $\mathscr{E}_1$ in (3.11), our next goal is to incorporate the Fourier expansion, via the following

Lemma 3.6 *Let*

$$\mathscr{E}_2 := \sum_{u=1}^{U} \frac{1}{u^2} \sum_{m \in \mathbb{Z}\setminus\{0\}} \frac{\left|a_\Psi\left(m; \frac{u}{T}\right)\right|}{|m|}.$$

Then

$$|\mathscr{E}_1| \ll \mathscr{E}_2 + \mathcal{S}_{\infty,1}(\Psi)\, \frac{\log U}{T}. \tag{3.13}$$

Proof We first straighten out the sheared integral in (3.11) by breaking it into sums:

$$\mathscr{E}_1 = \sum_{u=1}^{U-1} \int_u^{u+1} \Psi^\perp\left(y + i\frac{y}{T}\right) \frac{dy}{y}.$$

On each interval, estimate

$$\Psi^\perp\left(y + i\frac{y}{T}\right) = \Psi^\perp\left(y + i\frac{u}{T}\right) + O\left(\mathcal{S}_{\infty,1}(\Psi)\, \frac{1}{T}\right),$$

and Fourier expand

$$\Psi^\perp\left(y + i\frac{u}{T}\right) = \sum_{m \neq 0} a_\Psi\left(m; \frac{u}{T}\right) e_\omega(my).$$

Thus

$$\mathscr{E}_1 = \sum_{u=1}^{U-1} \sum_{m\neq 0} a_\Psi\left(m; \frac{u}{T}\right) \left[\int_u^{u+1} e_\omega(my)\frac{dy}{y}\right] + O\left(\mathcal{S}_{\infty,1}(\Psi)\, \frac{\log U}{T}\right).$$

Inserting absolute values and estimating the bracketed term by partial integration gives (3.13), as claimed. □

Our final task is to estimate $\mathscr{E}_2$; we cannot directly use the decay of Fourier coefficients (2.15) in the full range of m, so introduce a parameter M to be chosen later, and decompose

$$\mathscr{E}_2 = \mathscr{E}_\geq + \mathscr{E}_<,$$

where for $\square \in \{\geq, <\}$,

$$\mathscr{E}_\square := \sum_{u=1}^{U} \frac{1}{u^2} \sum_{0 \neq |m| \,\square\, M} \frac{\left|a_\Psi\left(m; \frac{u}{T}\right)\right|}{|m|}.$$

We first estimate the large range trivially.

Lemma 3.7

$$\mathscr{E}_{\geq} \ll \|\Psi\|_{\infty}\, M^{-1/2}. \tag{3.14}$$

Proof Cauchy-Schwarz and Parseval give:

$$\mathscr{E}_{\geq} \ll \sum_{u=1}^{U} \frac{1}{u^2} \left(\sum_{|m|\geq M} \left| a_{\Psi}\left(m; \frac{u}{T}\right)\right|^2 \right)^{1/2} \left(\sum_{|m|\geq M} \frac{1}{|m|^2} \right)^{1/2}$$
$$\ll \sum_{u=1}^{U} \frac{1}{u^2} \left(\frac{1}{\omega} \int_0^{\omega} \left| \Psi\left(x + i\frac{u}{T}\right)\right|^2 dx \right)^{1/2} M^{-1/2},$$

which can be estimated by (3.14), as claimed. □

Finally, we estimate the range of small m using decay of Fourier coefficients. Note that this is the *only* part of the argument involving any spectral theory; nevertheless, thanks to the uniformity of Proposition 2.2, we do not at this stage perceive any difference between the lattice and thin cases.

Lemma 3.8 *Recalling the Sobolev norm $\mathcal{S}_F$ and constants C_F and α_F from Proposition 2.2, we have*

$$\mathscr{E}_{<} \ll \mathcal{S}_F(\Psi)\, T^{-\alpha_F}\, M^{C_F}. \tag{3.15}$$

Proof Applying (2.15) gives

$$\mathscr{E}_{<} = \sum_{u=1}^{U} \frac{1}{u^2} \sum_{1\leq |m| < M} \frac{\left|a_{\Psi}\left(m; \frac{u}{T}\right)\right|}{|m|} \ll \sum_{u=1}^{U} \frac{1}{u^2} \sum_{1\leq |m|<M} \mathcal{S}_F(\Psi)|m|^{C_F-1}\left|\frac{u}{T}\right|^{\alpha_F},$$

which is bounded as claimed in (3.15). □

Proof of Theorem 3.2 This is now a simple matter of combining the above lemmata. To balance (3.14) and (3.15), set

$$M = \left(\|\Psi\|_{\infty}\, \mathcal{S}_F(\Psi)^{-1}\, T^{\alpha_F} \right)^{1/(C_F+1/2)},$$

for a net error, crudely, of

$$\mathscr{E}_2 = \mathscr{E}_{\geq} + \mathscr{E}_{<} = O\left(\max(\mathcal{S}_F(\Psi), \mathcal{S}_{\infty,1}(\Psi)) \cdot T^{-\alpha_F/(2C_F+1)} \right). \tag{3.16}$$

To balance the error in (3.7) and (3.12) with that of (3.13), we take U to be some power of T, say,

$$U = T^{1+1/\alpha_{\Psi}}, \tag{3.17}$$

assuming that T is large enough for (3.6) to be satisfied. Then the errors in (3.12) and (3.7) are $O(C_\Psi/T)$, and the second error term in (3.13) is $O((1+\frac{1}{\alpha_\Psi})\mathcal{S}_{\infty,1}\log T/T)$, which subsumes $O(\|\Psi\|_\infty T^{-2})$ in (3.7). On (again, crudely) setting

$$\mathcal{S}_{\mathfrak{S}}(\Psi) := \left(C_\Psi + \frac{1}{\alpha_\Psi}\right) \cdot \max(\mathcal{S}_F(\Psi), \mathcal{S}_{\infty,1}(\Psi)), \tag{3.18}$$

and

$$\alpha_{\mathfrak{S}} := \alpha_F/(2C_F+1), \tag{3.19}$$

as in (3.16), one can verify directly that the net error is as claimed in (3.3).

This completes the proof. □

3.2 Stage 2: Proof of Theorem 3.3

We first give the proof in the thin case, as it is significantly easier.

3.2.1 Assume Γ is thin in G

Returning to (3.2), write $\mu_{T,\mathfrak{S}}(\Psi)$ as:

$$\mu_{T,\mathfrak{S}}(\Psi) = \frac{1}{\omega}\left(\int_0^\infty - \int_0^{1/T}\right)\int_0^\omega \Psi(z,\uparrow)\, y\, dz =: \mathscr{T}_1 - \mathscr{T}_2,$$

say. Here we have set $dz := dx\,dy/y^2$. We bound $\mathscr{T}_2$ by

$$\begin{aligned} |\mathscr{T}_2| &\le \int_0^{1/T} |a_\Psi(0; y, \uparrow)|\; y\, \frac{dy}{y^2} \\ &\ll \int_0^{1/T} \mathcal{S}_H(\Psi) y^{\alpha_H} y\, \frac{dy}{y^2} = \mathcal{S}_H(\Psi)\, T^{-\alpha_H}, \end{aligned} \tag{3.20}$$

where we applied (2.16) (with $\mathcal{I} = (0, \omega)$).

Recalling that $\Gamma_\infty = \begin{pmatrix} 1 & \omega\mathbb{Z} \\ & 1 \end{pmatrix} < \Gamma$, we next deal with

$$\mathscr{T}_1 := \frac{1}{\omega}\int_{\Gamma_\infty\backslash\mathbb{H}} \Psi(z,\uparrow)\, y\, dz. \tag{3.21}$$

Note that the integral converges absolutely; for $y \to \infty$, this is due to (3.1), while for $y \to 0$, we can again use (2.16).

For ease of exposition, it is convenient at this point to first assume that Ψ is right-K-invariant, that is,

$$\Psi(z,\zeta) = \Psi(z). \tag{3.22}$$

Below we detail the modifications needed to handle the general case.

Recall from (2.7) that

$$E(z,s) := \frac{1}{\omega}\sum_{\gamma\in\Gamma_\infty\backslash\Gamma}\mathfrak{Im}(\gamma z)^s$$

is the Eisenstein series at ∞ of a cusp of width ω. Note that the defining sum converges absolutely and uniformly on compacta in the range $\mathfrak{Re}(s) > \delta$, since Γ is assumed to be a thin subgroup of G. In particular, $E(z,s)$ is regular at $s = 1$.

Then, letting $\mathscr{F}$ be a fixed fundamental domain for $\Gamma\backslash\mathbb{H}$, we can "re-fold" and write (3.21) as

$$\begin{aligned}\mathscr{T}_1 &= \frac{1}{\omega}\sum_{\gamma\in\Gamma_\infty\backslash\Gamma}\int_{\gamma\mathscr{F}}\Psi(z)\, y\, dz\\ &= \frac{1}{\omega}\sum_{\gamma\in\Gamma_\infty\backslash\Gamma}\int_{\mathscr{F}}\Psi(z)\,\mathfrak{Im}(\gamma z)\, dz = \langle\Psi, E(\cdot,1)\rangle\,.\end{aligned}$$

Setting

$$\mu_{Eis}(\Psi) := \langle\Psi, E(\cdot,1)\rangle\,, \tag{3.23}$$

we immediately see that $\mathscr{T}_1 = \mu_{Eis}(\Psi)$, which combined with (3.20) gives:

$$\mu_{T,\mathfrak{S}}(\Psi) = \mu_{Eis}(\Psi) + O\left(\mathcal{S}_H(\Psi)\, T^{-\alpha_H}\right),$$

as claimed.

Finally, we remove the assumption (3.22) and extend the proof to the general case. For a unit tangent vector ζ at z, write $\theta\in[-\pi,\pi)$ for the "angle" of $\zeta = \zeta_\theta$, measured from the vertical $\uparrow$ counterclockwise. We first decompose $\Psi(z,\zeta)$ in a Fourier series in ζ, writing:

$$\Psi(z,\zeta) = \sum_{n\in\mathbb{Z}}\widehat{\Psi}_n(z)\,\chi_n(\zeta), \tag{3.24}$$

where $\chi_n(\zeta_\theta) = e^{in\theta}$ in the above notation, and

$$\widehat{\Psi}_n(z) := \frac{1}{2\pi}\int_{-\pi}^{\pi}\Psi(z,\zeta_\theta)\,\overline{\chi}_n(\zeta_\theta)\, d\theta.$$

Note that each $\widehat{\Psi}_n$ lives in the space $(\Gamma, 2n)$ of functions on $\mathbb{H}$ given in (2.11).

Returning to $\mathscr{T}_1$ in (3.21), we insert (3.24) (with $\zeta = \uparrow$), and "re-fold " again, obtaining:

$$\begin{aligned}\mathscr{T}_1 &= \frac{1}{\omega} \sum_{\gamma \in \Gamma_\infty \backslash \Gamma} \int_{\gamma \mathscr{F}} \sum_{n \in \mathbb{Z}} \widehat{\Psi}_n(z)\, \mathfrak{Im}(z)\, dz \\ &= \frac{1}{\omega} \sum_n \sum_{\gamma \in \Gamma_\infty \backslash \Gamma} \int_{\mathscr{F}} \widehat{\Psi}_n(z)\, \epsilon_\gamma(z)^{2n} \mathfrak{Im}(\gamma z)\, dz \\ &= \sum_n \left\langle \widehat{\Psi}_n, E_n(\cdot, 1) \right\rangle.\end{aligned}$$

Here, $E_{2n}(z, s)$ are the "wight-$2n$" Eisenstein series given by the series (2.9); these all converge absolutely for $\mathfrak{Re}(s) > \delta$. The absolute convergence of the sum over n is guaranteed by (3.1) after taking two derivatives in θ and noting that $\frac{\partial^2 \Psi}{\partial \theta^2} \in \mathscr{P}^0_{\mathfrak{a}}(\Gamma \backslash G)$. Then, on defining

$$\mu_{Eis}(\Psi) := \sum_n \left\langle \widehat{\Psi}_n, E_n(\cdot, 1) \right\rangle, \tag{3.25}$$

(3.4) follows immediately. Note that if Ψ is K-invariant, the two definitions (3.23) and (3.25) agree, and moreover μ_{Eis} is actually a measure. In general, μ_{Eis} is a distribution, as we need some derivatives of $\widehat{\Psi}_n$ to ensure the convergence of (3.25). This completes the proof in the thin case.

3.2.2 Case Γ is a lattice in G

In this case, our analysis precedes in a similar fashion to that in [43]. We begin with the following

Lemma 3.9 *For $1 < \sigma < 1 + \alpha_\Psi$, we have*

$$\mu_{T,\mathfrak{S}}(\Psi) = \sum_n \frac{1}{2\pi i} \int_{(\sigma)} \frac{T^{s-1}}{s-1} \left\langle \widehat{\Psi}_n, E_n(\cdot, \bar{s}) \right\rangle ds. \tag{3.26}$$

Proof Starting with (3.2), write

$$\mu_{T,\mathfrak{S}}(\Psi) = \frac{1}{\omega} \int_0^\infty \int_0^\omega \Psi(z, \uparrow)\, h_T(y)\, dz, \tag{3.27}$$

where we have set

$$h_T(y) := y \cdot \mathbf{1}_{\{y > 1/T\}}.$$

Note the Mellin transform/inverse pair:

$$\widetilde{h_T}(s) := \int_0^\infty h_T(y)\, y^{-s} \frac{dy}{y} = \frac{T^{s-1}}{s-1},$$

and

$$h_T(y) \;=\; \frac{1}{2\pi i}\int_{(\sigma)} \frac{T^{s-1}}{s-1}\, y^s\, ds.$$

The first integral converges absolutely for $\mathfrak{Re}(s) = \sigma > 1$; the second is henceforth interpreted (after partial integration) as the absolutely convergent integral

$$h_T(y) \;=\; \frac{1}{2\pi i}\int_{(\sigma)} \frac{T^{s-1}}{\log(Ty)(s-1)^2}\, y^s\, ds. \tag{3.28}$$

Inserting (3.28) into (3.27) with the above convention gives

$$\mu_{T,\mathfrak{S}}(\Psi) \;=\; \frac{1}{\omega}\frac{1}{2\pi i}\int_{(\sigma)} \frac{T^{s-1}}{s-1}\int_0^\infty\int_0^\omega \Psi(z,\uparrow)\, y^s\, dz\, ds, \tag{3.29}$$

which is absolutely convergent in the range $1 < \sigma < 1 + \alpha_\Psi$ using (3.1).

Now we proceed as in the thin case, decomposing

$$\Psi(z,\uparrow) \;=\; \sum_n \widehat{\Psi}_n(z)$$

and "unfolding"; for each $n \in \mathbb{Z}$ this gives

$$\int_{\Gamma_\infty\backslash\mathbb{H}} \widehat{\Psi}_n(z)\, y^s\, dz \;=\; \left\langle \widehat{\Psi}_n, E_n(\cdot,\bar{s})\right\rangle.$$

Summing over n and inserting into (3.29) gives (3.26), as claimed. □

To finish the proof of Theorem 3.3, we make the following definition:

$$\widetilde{E}_n(z,s) = \begin{cases} E_n(z,s) & n \neq 0 \\ E(z,s) - \frac{1}{\mathrm{vol}(\Gamma\backslash\mathbb{H})(s-1)} & n = 0 \end{cases},$$

which, again, is regular at $s = 1$ for all n. Then (3.26) can be rewritten as

$$\mu_{T,\mathfrak{S}}(\Psi) = \mu_{\Gamma\backslash G}(\Psi)\log(T) \;+\; \sum_{n\in\mathbb{Z}} \frac{1}{2\pi i}\int_{(\sigma)} \frac{T^{s-1}}{s-1}\left\langle \widehat{\Psi}_n, \widetilde{E}_n(\cdot,\bar{s})\right\rangle ds, \tag{3.30}$$

where we used that

$$\left\langle \widehat{\Psi}_0, \frac{1}{\mathrm{vol}(\Gamma\backslash\mathbb{H})}\right\rangle \;=\; \mu_{\Gamma\backslash G}(\Psi), \quad \text{and} \quad \frac{1}{2\pi i}\int_{(\sigma)} \frac{T^{s-1}}{(s-1)^2} ds \;=\; \log T.$$

Now shifting the contour of integration to $\mathfrak{Re}(s) = \frac{1}{2}$, we pick up residues from the simple pole at $s = 1$ and from the residual spectrum at $s = \sigma_j$ as in (2.8). The residue at $s = 1$ is

$$\sum_{n\in\mathbb{Z}}\left\langle\widehat{\Psi}_n, \widetilde{E}_n(\cdot,1)\right\rangle =: \mu_{\widetilde{Eis}}(\Psi), \tag{3.31}$$

that is, this is our "second-order" contribution, and is a distribution (as opposed to a measure) since Ψ is not assumed to be K-finite. Note that if Ψ is K-fixed, then (3.31) simplifies to just

$$\mu_{\widetilde{Eis}}(\Psi) = \left\langle\Psi, \widetilde{E}_0(\cdot,1)\right\rangle, \tag{3.32}$$

as claimed in (1.7).

Each pole $s=\sigma_j$ contributes a residue $\frac{T^{\sigma_j-1}}{\sigma_j-1}\mu_{\sigma_j}(\Psi)$, where

$$\mu_{\sigma_j}(\Psi) := \sum_{n\in\mathbb{Z}}\left\langle\widehat{\Psi}_n, \varphi_{\sigma_j,n}\right\rangle, \tag{3.33}$$

with $\varphi_{\sigma_j,n}$ the "weight-$2n$" residual form given in (2.10). Note that these distributions are exactly the same as those arising in Sarnak's analysis [43, p. 737].

We thus obtain

$$\mu_{T,\mathfrak{S}}(\Psi) = \mu(\Psi)\log(T) + \mu_{\widetilde{Eis}}(\Psi) + \sum_{j=1}^{h}\frac{T^{\sigma_j-1}}{\sigma_j-1}\mu_{\sigma_j}(\Psi)$$
$$+\sum_n\frac{1}{2\pi i}\int_{(1/2)}\frac{T^{s-1}}{s-1}\left\langle\widehat{\Psi}_n, E_n(\cdot,\bar{s})\right\rangle ds.$$

Finally, taking absolute values and combining Cauchy Schwartz with (2.12), we can bound each of the terms in the last sum by

$$\left|\frac{1}{2\pi i}\int_{(1/2)}\frac{T^{s-1}}{s-1}\left\langle\widehat{\Psi}_n, E_n(\cdot,\bar{s})\right\rangle ds\right| \ll T^{-1/2}\|\widehat{\Psi}_n\|_2.$$

On estimating $\sum_n\|\widehat{\Psi}_n\|_2 \ll \mathcal{S}_{2,1}(\Psi)$, we finally conclude the proof of Theorem 3.3.

4 Application 1: moments of *L*-functions

Theorem 1.6 now follows readily from Theorem 1.1, as we explain below. Recall that we will illustrate the method on the simplest case of f being a weight-k holomorphic Hecke cusp form on $\mathrm{PSL}_2(\mathbb{Z})$; the calculation for general cuspidal GL(2) automorphic representations is similar.

Let $\Psi(x+iy) = |f(x+iy)|^2y^k$, and use (1.11) to write the left hand side of (1.12) as

$$\frac{1}{2\pi}\int_{\mathbb{R}}\left|L\left(f,\tfrac{1}{2}+it\right)\right|^2\left|\mathcal{W}_k\left(\tfrac{1}{2}+it,T\right)\right|^2 dt = \left(\int_0^{1/\widetilde{T}} + \int_{1/\widetilde{T}}^{\infty}\right)\Psi\left(Ty+iy\right)\frac{dy}{y},$$

where we have set

$$\widetilde{T} := \sqrt{T^2+1}$$

for convenience. Theorem 1.1 can be applied directly to the range $[1/\widetilde{T}, \infty)$, but the range $(0, 1/\widetilde{T})$ must be manipulated. Changing variables $y \mapsto 1/y$, using the automorphy of Ψ that $\Psi(-1/z) = \Psi(z)$, and changing $y \mapsto \widetilde{T}^2 y$ gives:

$$\int_0^{1/\widetilde{T}} \Psi\left(Ty+iy\right)\frac{dy}{y} = \int_{1/\widetilde{T}}^{\infty} \Psi\left(-yT+iy\right)\frac{dy}{y}.$$

Now we can apply Theorem 1.1 to both contributions, giving

$$\begin{aligned}\frac{1}{2\pi}\int_{\mathbb{R}}\left|L\left(f,\tfrac{1}{2}+it\right)\right|^2\left|\mathcal{W}_k\left(\tfrac{1}{2}+it,T\right)\right|^2 dt\\ = 2\mu_{\Gamma\backslash G}(\Psi)\log(T) + 2\mu_{\widetilde{Eis}}(\Psi) + O_\Psi(T^{-\eta}).\end{aligned} \tag{4.1}$$

The first term is of course

$$\mu_{\Gamma\backslash G}(\Psi) = \frac{\|f\|^2}{\mathrm{vol}(\Gamma\backslash\mathbb{H})},$$

where the norm is with respect to the Petersson inner product. It remains to show that the second term, that is, the Eisenstein measure $\mu_{\widetilde{Eis}}(\Psi)$, can be expressed as special value (at the edge of the critical strip) of a symmetric square L-function. Note that Ψ is a function on $\mathbb{H}$, that is, as a function on G it is right-K-invariant; therefore $\mu_{\widetilde{Eis}}(\Psi)$ is determined by the simpler expression (3.32) (or (1.7)).

Proposition 4.1 *With the above notation, we have*

$$\mu_{\widetilde{Eis}}(\Psi) = \frac{\|f\|^2}{\mathrm{vol}(\Gamma\backslash\mathbb{H})}\left(\frac{\Lambda'}{\Lambda}(\mathrm{sym}^2 f, 1) + \gamma - 2\frac{\zeta'}{\zeta}(2)\right).$$

Clearly Proposition 4.1 inserted into (4.1) gives the right hand side of (1.12), completing the proof of Theorem 1.6.

Proof of Proposition 4.1 To evaluate

$$\mathcal{I} := \mu_{\widetilde{Eis}}(\Psi) = \left\langle |f|^2 y^k, \widetilde{E}(\cdot,1)\right\rangle,$$

use (1.5) to write

$$\mathcal{I} = \lim_{s\to 1}\left(\left\langle |f|^2 y^k, \widetilde{E}(\cdot,\bar{s})\right\rangle\right) = \lim_{s\to 1}\left(\left\langle |f|^2 y^k, E(z,\bar{s})\right\rangle - \frac{1}{(s-1)V}\|f\|^2\right), \tag{4.2}$$

where

$$V := \mathrm{vol}(\Gamma\backslash\mathbb{H}).$$

By standard Rankin–Selberg theory, we have

$$\left\langle f\,\bar{g}\,y^k, E(\cdot,\bar{s})\right\rangle = (4\pi)^{-(s+k-1)}\Gamma(s+k-1)L(f\otimes\bar{g},s) = \Lambda(f\otimes\bar{g},s).$$

When $f = g$, the Rankin–Selberg L-function factors (see, e.g., [26, p. 232]) as

$$L(f\otimes\bar{f},s) = \frac{\zeta(s)}{\zeta(2s)}L(\mathrm{sym}^2 f,s).$$

Hence

$$\left\langle |f|^2 y^k, E(\cdot,\bar{s})\right\rangle = \Lambda(f\otimes\bar{f},s) = \frac{\zeta(s)}{\zeta(2s)}\Lambda(\mathrm{sym}^2 f,s), \tag{4.3}$$

where $\Lambda(\mathrm{sym}^2 f,s)$ is as in (1.13). Taking residues at $s = 1$ on both sides of (4.3) gives

$$\frac{\|f\|^2}{V} = \frac{1}{\zeta(2)}\Lambda(\mathrm{sym}^2 f,1). \tag{4.4}$$

Inserting (4.4) and (4.3) into (4.2) gives

$$\mathcal{I} = \lim_{s\to 1}\left(\frac{\zeta(s)}{\zeta(2s)}\Lambda(\mathrm{sym}^2 f,s) - \frac{1}{(s-1)}\frac{1}{\zeta(2)}\Lambda(\mathrm{sym}^2 f,1)\right).$$

Using $\zeta(s) - \frac{1}{s-1} \to \gamma$ as $s \to 1$ (Euler's constant), and elementary calculus, we have that

$$\mathcal{I} = \frac{\|f\|^2}{V}\left(\frac{\Lambda'}{\Lambda}(\mathrm{sym}^2 f,1) + \gamma - 2\frac{\zeta'}{\zeta}(2)\right),$$

on using (4.4) again. This completes the proof. □

Note that one could also extend our method to Eisenstein series, and then evaluate the (weighted) fourth moment of the Riemann zeta function using Theorem 1.1.

4.1 Subconvexity?

We leave open the problem of extracting from the effective second moment (1.12) a subconvex bound

$$\left|L\left(f,\tfrac{1}{2}+it\right)\right| \ll_f |t|^{1/2-\eta}$$

in the t-aspect. Such is already known [19,42] in the Maass case via trace formulae, explicit expansions, and shifted convolutions, but it would be interesting to give a new proof using only equidistribution. (Of course the general GL(2) subconvexity problem has been resolved [35]; the interest here would be in the method used.)

The key issue is that the archimedean factor $|\mathcal{W}_k|^2$ in (1.12) is a smooth weight, which does not allow truncation; if the weight could be replaced by a sharp cutoff while still having a power savings rate, then the subconvexity bound would follow immediately. This could be accomplished by finding a function $\Psi_X(T)$ so that

$$\int_{\mathbb{R}} \Psi_X(T) \left|\mathcal{W}_k\left(\tfrac{1}{2}+it, T\right)\right|^2 dT \overset{?}{=} \mathbf{1}_{|t|<X}; \tag{4.5}$$

indeed, then one would multiply both sides of (1.12) by $\Psi_X(T)$ and integrate in T, obtaining

$$\begin{aligned}\frac{1}{2\pi}\int_{|t|<X} \left|L\left(f, \tfrac{1}{2}+it\right)\right|^2 dt &= \int_{\mathbb{R}} \Psi_X(T)\left(C_1 \log T + C_2 + O(T^{-\eta})\right) dT\\ &\overset{?}{=} C_1' X \log X + C_2' X + O(X^{1-\eta'}).\end{aligned}$$

Another approach is "shorten the interval," that is, to replace the right hand side of (4.5) by $\mathbf{1}_{|t-X|<Y}$, with $Y < X^{1-\eta}$.

Either way, one would need to invert the "$\mathcal{W}$-transform":

$$\Psi(T) \mapsto \widetilde{\Psi}(t) := \int_{\mathbb{R}} \Psi(T) \left|\mathcal{W}_k\left(\tfrac{1}{2}+it, T\right)\right|^2 dT.$$

Unfortunately, there are basic difficulties with said inversion, namely a Paley-Weiner (or Heisenberg uncertainty) analysis shows that the transform has insufficient harmonics to be invertible and functions Ψ_X as above do not exist, even in this simple holomorphic case! (Cf. the related discussion in, e.g., [20, Appendix].) The case of non-holomorphic Hecke-Maass forms is seemingly even more complicated as the weights (1.10) will involve Bessel functions.

A potential method to circumvent this issue (since our equidistribution theorem is proved in the generality of the unit tangent bundle) is to use all the harmonics afforded us by f, that is, by applying Maass raising and lowering operators. This does not change the L-function, but results in effective second moments with a large span of weight functions $\mathcal{W}$. One can hope that enough combinations of these can recover the desired sharp cutoff functions Ψ_X, and we plan to return to this question later.

5 Application 2: Counting and non-equidistribution

5.1 Proof of Theorem 1.9

As the method of counting from equidistribution is by now completely standard, we give a brief sketch (only the setup is not completely obvious). Let $G = \mathrm{SL}_2(\mathbb{R})$ be the spin double-cover $G \xrightarrow{\iota} \mathrm{SO}^\circ_Q(\mathbb{R})$ of the (identity component of the) special orthogonal group preserving an indefinite ternary quadratic form Q. Let $\Gamma < G$ be discrete, Zariski-dense, geometrically finite, and have at least one cusp, and given $\mathbf{x}_0 \in \mathbb{R}^3$, let $\mathcal{O} = \mathbf{x}_0\,\iota(\Gamma)$ be a discrete orbit. Let $H = \mathrm{Stab}_G\,\mathbf{x}_0$ be the stabilizer of $\mathbf{x}_0$ in G, and let $\Gamma_H := \Gamma \cap H$ be the stabilizer in Γ. Given an archimedean norm $\|\cdot\|$ on $\mathbb{R}^3$, we obtain a norm-T ball B_T in $H\backslash G$ as in (1.18). Our goal is to estimate $\mathcal{N}_{\mathcal{O}}(T) = |\mathcal{O} \cap B_T|$, that is,

$$\mathcal{N}_{\mathcal{O}}(T) \;=\; \{\gamma \in \Gamma_H\backslash\Gamma \;:\; \|\mathbf{x}_0\,\iota(\gamma)\| < T\}.$$

Thanks to the discussion in §1.3.1 (see Table 1), there are only two new cases to prove, both occurring only when $H = \mathrm{Stab}_G\,\mathbf{x}_0$ is diagonalizable. We can choose the spin cover ι up to conjugation, and hence can assume that $H = A$. Having made such a choice, we will henceforth drop ι from the notation. To handle the two lacunary cases, we may assume that Γ_H is trivial. (In principle, Γ_H could be finite.)

We then break $\mathcal{N}_{\mathcal{O}}(T)$ into two contributions as follows. Recalling the shear $\mathfrak{s}_t$ in (1.1), we decompose each $g \in G = ANK$ uniquely as $g = a\mathfrak{s}_t k$, and write

$$G^{\pm} := \{g = a\mathfrak{s}_t k \in G : a \in A^{\pm}\}.$$

Hence we can write

$$\mathcal{N}_{\mathcal{O}}(T) = \mathcal{N}^+_{\mathcal{O}}(T) + \mathcal{N}^-_{\mathcal{O}}(T),$$

say, where

$$\mathcal{N}^{\pm}_{\mathcal{O}}(T) \;:=\; \{\gamma \in \Gamma \cap G^{\pm} \;:\; \|\mathbf{x}_0\,\gamma\| < T\},$$

and treat only $\mathcal{N}^+_{\mathcal{O}}(T)$, the other contribution being the same (after conjugation).

If Γ is a lattice, then the “lacunary” case occurs only when both 0 and ∞ (that is, the two endpoints of A) are cusps of Γ. When Γ is thin, the “lacunary” cases occur when at least one of $0, \infty$ is a cusp; Lemma 1.11 forces the other endpoint to be either a cusp or in the free boundary. If ∞, say, is in the free boundary, then $\mathcal{N}^+_{\mathcal{O}}(T)$ gives a contribution of order N^{δ}, as described below (1.21). So to restrict attention to the lacunary case, we assume that ∞ is a cusp of Γ. Now we continue with the standard smoothing/unsmoothing argument applied to the equidistribution theorem. For ease of exposition, assume that the norm $\|\cdot\|$ is right-K-invariant. (This assumption is standard to relax.)

Let $\psi : G \to \mathbb{R}_{\geq 0}$ be a right-K-invariant bump function supported in an $\varepsilon > 0$ ball about the origin in G/K with $\int_G \psi = 1$. Set $\Psi(g) := \sum_{\gamma\in\Gamma} \psi(\gamma g)$, so that $\int_{\Gamma\backslash G} \Psi = 1$. Let

$$f^+_T(g) := \mathbf{1}_{\{\|\mathbf{x}_0 g\| < T,\ g \in G^+\}},$$

and $\mathcal{F}^+_T(g) \;:=\; \sum_{\gamma\in\Gamma} f^+_T(\gamma g)$. Then

$$\mathcal{F}^+_T(e) \;=\; \mathcal{N}^+_{\mathcal{O}}(T)$$

and

$$\langle \mathcal{F}_T^+, \Psi \rangle = \mathcal{N}_{\mathcal{O}}^+(T)(1 + O(\varepsilon)),$$

since Ψ is a bump function about the origin. Unfolding the inner product gives

$$\langle \mathcal{F}_T^+, \Psi \rangle = \int_G f_T^+(g)\Psi(g)dg = \int_{\mathfrak{s}_t} f_T^+(\mathfrak{s}_t)\left[\int_{a\in A^+} \Psi(a\mathfrak{s}_t)da\right] d\mathfrak{s}_t.$$

Applying Theorem 1.1 to the bracketed term and integrating in t completes the sketch of the two remaining cases of Theorem 1.9.

5.2 Proof of Proposition 1.14

As above, let the stabilizer $H = A$ be diagonalizable, let ∞ be a cusp of Γ, and assume for ease of exposition that the norm $\|\cdot\|$ is right-K-invariant. The statement of Proposition 1.14 assumes that $\Gamma < \mathrm{SL}_2(\mathbb{Z})$ is integral and thin. For an integer $q \geq 1$, let $\Gamma(q) < \Gamma$ be its level-q principal congruence subgroup, and for a fixed $\varpi \in \Gamma/\Gamma(q)$, let

$$\mathcal{O}_{q,\varpi} := \mathbf{x}_0\varpi\Gamma(q)$$

be the congruence coset orbit. The corresponding counting function is then

$$\mathcal{N}_{q,\varpi}(T) := |\mathcal{O}_{q,\varpi} \cap B_T|.$$

We claim that this count depends on ϖ, that is, is not distributed uniformly among the cosets.

One way to explain this is to unravel the formalism of the previous proof, and note that C_1 in (1.17) is essentially the evaluation at $s = 1$ of an (unregularized) Eisenstein series corresponding to one of the cusps of $\Gamma(q)$; these values do not coincide for all cosets, which one can see as follows.

Assume for simplicity that $\begin{pmatrix} 1 & \mathbb{Z} \\ & 1 \end{pmatrix} < \Gamma$; then the isotropy group of ∞ in $\Gamma(q)$ is

$$\Gamma_{\infty,q} := \begin{pmatrix} 1 & q\mathbb{Z} \\ & 1 \end{pmatrix} = \Gamma(q) \cap \Gamma_\infty.$$

Certainly the orbit $\mathcal{O}_{q,\varpi}$ contains the points

$$\mathbf{x}_0\varpi\Gamma_{\infty,q} \subset \mathcal{O}_{q,\varpi},$$

so

$$\mathcal{N}_{q,\varpi}(T) \geq |\mathbf{x}_0\varpi\Gamma_{\infty,q} \cap B_T|.$$

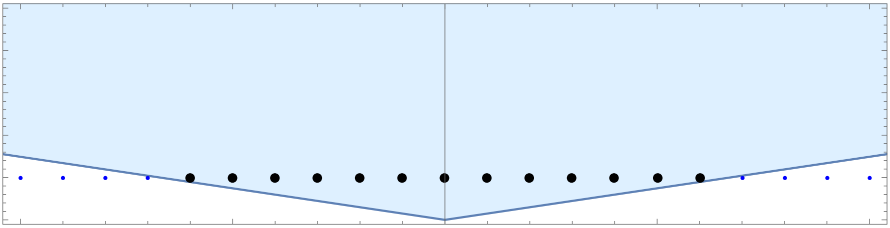


Fig. 4 The orbit $\mathbf{x}_0 \varpi \Gamma_{\infty,q}$ for $\varpi = e$ inside $B_T \subset \mathbb{H}$

Converting Fig. 3c from the disk $\mathbb{D}$ to the hyperbolic plane $\mathbb{H}$, we show in Fig. 4 how the shaded region B_T contains the orbit points $\varpi \Gamma_{\infty,q}$ for $\varpi = e$. A moment's reflection (or rather, translation) shows that (1.23) holds for this orbit.

Something similar holds when one takes congruence cosets with the subgroups

$$\Gamma_0(q) := \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma : c \equiv 0(q)\},$$

say, instead of $\Gamma(q)$. The isotropy group Γ_∞ now remains unchanged, but the same picture shows that, for the identity coset (with $\varpi = e$), the number of points in an orbit is $\gg T$, whereas is average count is of order T/q. We leave it as an interesting challenge to develop sieve methods which apply to this non-uniformly distributed (in the archimedean ordering) setting.

Acknowledgements The authors would like to express their gratitude to Jens Marklof, Curt McMullen, and Peter Sarnak for many enlightening comments and suggestions. The second author would especially like to thank Valentin Blomer, Farrell Brumley, and Nicolas Templier for many hours of discussion about this problem; see also the related work in [1].

References

1. Blomer, V., Brumley, F., Kontorovich, A., Templier, N.: Bounding hyperbolic and spherical coefficients of Maass forms. J. Théor. Nombres Bordeaux **26**(3), 559–579 (2014)
2. Beardon, Alan F.: The Geometry of Discrete Groups, volume 91 of Graduate Texts in Mathematics. Springer, New York (1983)
3. Bourgain, J., Gamburd, A., Sarnak, P.: Affine linear sieve, expanders, and sum-product. Invent. Math. **179**(3), 559–644 (2010)
4. Bourgain, J., Gamburd, A., Sarnak, P.: Generalization of Selberg's 3/16th theorem and affine sieve. Acta Math **207**, 255–290 (2011)
5. Bourgain, J., Kontorovich, A.: The affine sieve beyond expansion I: Thin hypotenuses. Int. Math. Res. Not. IMRN **19**, 9175–9205 (2015)
6. Bourgain, J., Kontorovich, A., Sarnak, P.: Sector estimates for hyperbolic isometries. GAFA **20**(5), 1175–1200 (2010)
7. Blomer, Valentin: Sums of Hecke eigenvalues over values of quadratic polynomials. *Int. Math. Res. Not. IMRN*, (16):Art. ID rnn059. 29, (2008)
8. Bernstein, Joseph, Reznikov, André: Sobolev norms of automorphic functionals and Fourier coefficients of cusp forms. C. R. Acad. Sci. Paris Sér. I Math. **327**(2), 111–116 (1998)
9. Cowling, M., Haagerup, U., Howe, R.: Almost L^2 matrix coefficients. J. Reine Angew. Math. **387**, 97–110 (1988)
10. Delsarte, J.: Sur le gitter fuchsien. C. R. Acad. Sci. Paris **214**, 147–179 (1942)
11. Deligne, P.: La conjecture de Weil. I. Inst. Hautes Études Sci. Publ. Math. **43**, 273–307 (1974)

12. Diaconu, A., Garrett, P.: Subconvexity bounds for automorphic L-functions. J. Inst. Math. Jussieu **9**(1), 95–124 (2010)
13. Diaconu, A., Garrett, P., Goldfeld, D.: Moments for L-functions for $GL_r \times GL_{r-1}$. In: *Contributions in analytic and algebraic number theory*, volume 9 of *Springer Proc. Math.*, pp. 197–227. Springer, New York (2012)
14. Dolgopyat, Dmitry: On decay of correlations in Anosov flows. Ann. of Math. (2) **147**(2), 357–390 (1998)
15. Duke, W., Rudnick, Z., Sarnak, P.: Density of integer points on affine homogeneous varieties. Duke Math. J. **71**(1), 143–179 (1993)
16. Eskin, A., McMullen, C.: Mixing, counting and equidistribution in lie groups. Duke Math. J. **71**, 143–180 (1993)
17. Gelfand, I.M., Graev, M.I., Pjateckii-Shapiro, I.I.: *Teoriya predstavlenii i avtomorfnye funktsii*. Generalized functions, No. 6. Izdat. Nauka, Moscow, (1966)
18. Gelbart, S., Jacquet, H.: A relation between automorphic representations of GL(2) and GL(3). Ann. Sci. École Norm. Sup. (4) **11**(4), 471–542 (1978)
19. Good, A.: The convolution method for Dirichlet series. In: *The Selberg trace formula and related topics (Brunswick, Maine, 1984)*, volume 53 of *Contemp. Math.*, pages 207–214. Amer. Math. Soc., Providence, RI, (1986)
20. Ghosh, A., Reznikov, A., Sarnak, P.: Nodal domains of Maass forms I. Geom. Funct. Anal. **23**(5), 1515–1568 (2013)
21. Hong, J., Kontorovich, A.: Almost prime coordinates for anisotropic and thin pythagorean orbits. Israel J. Math. **209**(1), 397–420 (2015)
22. Hulse, Thomas A., Kiral, E. Mehmet, Kuan, Chan Ieong, Lim, Li-Mei: Counting square discriminants, 2013. Preprint, arXiv:1307.6606
23. Hooley, C.: On the number of divisors of a quadratic polynomial. Acta Math. **110**, 97–114 (1963)
24. Huber, H.: über eine neue Klasse automorpher Funktionen und ein Gitterpunktproblem in der hyperbolischen Ebene. i. Comment. Math. Helv. **30**, 20–62 (1956)
25. Iwaniec, H., Kowalski, E.: Analytic number theory, volume 53 of American Mathematical Society Colloquium Publications. American Mathematical Society, Providence, RI (2004)
26. Iwaniec, H.: Topics in classical automorphic forms. American Mathematical Society, Providence, RI (1997)
27. Kontorovich, A., Oh, H.: Almost prime Pythagorean triples in thin orbits. J. reine angew. Math. **667**, 89–131 (2012). arXiv:1001.0370
28. Kontorovich, A.: The hyperbolic lattice point count in infinite volume with applications to sieves. Duke J. Math. **149**(1), 1–36 (2009). arXiv:0712.1391
29. Kontorovich, A.: Levels of distribution and the affine sieve. Ann. Fac. Sci. Toulouse Math. (6) **23**(5), 933–966 (2014)
30. Kim, H., Sarnak, P.: Refined estimates towards the Ramanujan and Selberg conjectures. J. Amer. Math. Soc. **16**(1), 175–181 (2003)
31. Lax, P.D., Phillips, R.S.: The asymptotic distribution of lattice points in Euclidean and non-Euclidean space. Journal of Functional Analysis **46**, 280–350 (1982)
32. Liu, J., Sarnak, P.: Integral points on quadrics in three variables whose coordinates have few prime factors. Israel J. Math **178**, 393–426 (2010)
33. Margulis, G.A.: *On some aspects of the theory of Anosov systems*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2004. With a survey by Richard Sharp: Periodic orbits of hyperbolic flows, Translated from the Russian by Valentina Vladimirovna Szulikowska
34. Mohammadi, A., Oh, H.: Matrix coefficients, counting and primes for orbits of geometrically finite groups, (2013). To appear, *JEMS*
35. Michel, P., Venkatesh, A.: The subconvexity problem for GL_2. Publ. Math. Inst. Hautes Études Sci. **111**, 171–271 (2010)
36. Naud, F.: Expanding maps on Cantor sets and analytic continuation of zeta functions. Ann. Sci. École Norm. Sup. (4) **38**(1), 116–153 (2005)
37. Nevo, A., Sarnak, P.: Prime and almost prime integral points on principal homogeneous spaces. Acta Math. **205**(2), 361–402 (2010)
38. Oh, H., Shah, N.A.: Equidistribution and counting for orbits of geometrically finite hyperbolic groups. J. Amer. Math. Soc. **26**(2), 511–562 (2013)

39. Oh, H., Shah, N.A.: Limits of translates of divergent geodesics and integral points on one-sheeted hyperboloids. Israel J. Math. **199**(2), 915–931 (2014)
40. Patterson, S.J.: The Laplacian operator on a Riemann surface. Compositio Math. **31**(1), 83–107 (1975)
41. Patterson, S.J.: The limit set of a Fuchsian group. Acta Mathematica **136**, 241–273 (1976)
42. Petridis, Y.N., Sarnak, P.: Quantum unique ergodicity for $\mathrm{SL}_2(O)\backslash\mathbf{H}^3$ and estimates for L-functions. J. Evol. Equ. **1**(3), 277–290 (2001)
43. Sarnak, P.: Asymptotic behavior of periodic orbits of the horocycle flow and Eisenstein series. Comm. Pure Appl. Math. **34**(6), 719–739 (1981)
44. Sarnak, P.: Additive number theory and Maass forms. In: *Number theory (New York, 1982)*, vol. 1052 of *Lecture Notes in Math.*, pp. 286–309. Springer, Berlin (1984)
45. Sarnak, P.C.: Fourth moments of grossencharakteren zeta functions. Comm. Pure Appl. Math. **38**, 167–178 (1985)
46. Selberg, A.: Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series. J. Indian Math. Soc. (N.S.) **20**, 47–87 (1956)
47. Shalom, Y.: Rigidity, unitary representations of semisimple groups, and fundamental groups of manifolds with rank one transformation group. Ann. of Math. (2) **152**(1), 113–182 (2000)
48. Strombergsson, A.: On the uniform equidistribution of long closed horocycles. Duke Math. J. **123**, 507–547 (2004)
49. Sarnak, Peter, Ubis, Adrián: The horocycle flow at prime times. J. Math. Pures Appl. (9) **103**(2), 575–618 (2015)
50. Sullivan, D.: Entropy, Hausdorff measures old and new, and limit sets of geometrically finite Kleinian groups. Acta Math. **153**(3–4), 259–277 (1984)
51. Titchmarsh, E.C.: The Theory of the Riemann Zeta-Function. Oxford Univ. Press, London/New York, Oxford (1951)
52. Templier, N., Tsimerman, J.: Non-split sums of coefficients of $GL(2)$-automorphic forms. Israel J. Math. **195**(2), 677–723 (2013)
53. Venkatesh, A.: Sparse equidistribution problems, period bounds and subconvexity. Ann. of Math. (2) **172**(2), 989–1094 (2010)
54. Zagier, D.: Eisenstein series and the Riemann zeta function. In: *Automorphic forms, representation theory and arithmetic (Bombay, 1979)*, volume 10 of *Tata Inst. Fund. Res. Studies in Math.*, pages 275–301. Tata Inst. Fundamental Res., Bombay, (1981)