LARGE VALUES OF EIGENFUNCTIONS ON ARITHMETIC HYPERBOLIC 3-MANIFOLDS

Djordje Milićević
University of Michigan
Department of Mathematics
530 Church Street, East Hall
Ann Arbor, Michigan 48109-1043
E-mail: djordje@umich.edu

Abstract. We prove that, on a distinguished class of arithmetic hyperolic 3-manifolds, a sequence of L^2-normalized high-energy Hecke-Maass eigenforms ϕ_j achieve values as large as $\lambda_j^{1/4+o(1)}$, where $(\Delta+\lambda_j)\phi_j=0$. Arithmetic hyperbolic 3-manifolds on which this exceptional behavior is exhibited are, up to commensurability, precisely those containing immersed totally geodesic surfaces. We adapt the method of resonators and connect values of eigenfunctions to global geometry of the manifold by employing the pre-trace formula and twists by Hecke correspondences. Automorphic representations corresponding to forms appearing with highest weights in the optimized spectral averages are characterized both in terms of base change lifts and in terms of theta lifts from GSp_2.

0. Introduction.

This paper and the companion paper [Mi] present results on extreme values of eigenfunctions of Laplacian on certain hyperbolic manifolds.

Suppose M is a compact Riemannian manifold. The Laplacian Δ_M is a self-adjoint operator on $L^2(M)$ and we have an orthonormal decomposition

$$L^2(M)=\overline{\bigoplus_{j\geqslant 0}\mathbb{C}\phi_j}.$$

Eigenfunctions ϕ_j are the building blocks for harmonic analysis on M, but their asymptotic behavior is also of importance for geometry and dynamics on M. In case of arithmetic manifolds M, these connections have been the basis of a growing body of number-theoretic results. A basic question about the harmonics ϕ_j is: How large can $\|\phi_j\|_\infty$ get?

In [Mi], we show that high-energy eigenfunctions ϕ_j on arithmetic hyperbolic surfaces exhibit much stronger fluctuations than what the so-called Random Wave Conjecture predicted. In the present work, we apply our method to the case of arithmetic hyperbolic 3-manifolds and show that there is a distinguished class of arithmetic 3-manifolds on which certain eigenfunctions actually exhibit *power growth*, namely

$$\|\phi_j\|_\infty = \Omega\left(\lambda_j^{1/4+o(1)}\right).$$

0.1. General setup.

We first give a brief overview of general results and conjectures for perspective on our results. A more detailed introductive review can be found in [Mi].

For eigenfunctions on a compact Riemannian manifold M of dimension n, one has in full generality the upper bound [Se-So]

$$\|\phi_j\|_\infty \ll \lambda_j^{(n-1)/4}.$$

This bound is *local* in that it is obtained by estimating $|\phi_j(z)|$ through analysis of ϕ_j in a neighborhood of z only, without taking into account the global geometry of M. We refer the reader to Burq–Gérard–Tzvetkov [Bu-Gé-Tz] for more general upper bounds of this type on L^p-norms ($2 \leqslant p \leqslant \infty$) of restrictions of eigenfunctions on Riemannian manifolds to certain embedded submanifolds.

If M is a locally symmetric space of dimension n and rank r, joint eigenfunctions ϕ_j of the commutative algebra of invariant differential operators span $L^2(M)$ and are known by local analysis employing spherical functions and stationary phase method ([Du-Ko-Va], [Va]) to satisfy (away from the calls of Weyl chamber)

$$\|\phi_j\|_\infty \ll \lambda_j^{(n-r)/4}. \tag{0.1}$$

Local bounds like the ones above are also referred to as convexity bounds and are often sharp for certain positively curved and flat manifolds.

Rate of growth of eigenvalues is also connected to the question of multiplicities of Laplacian eigenvalues $\mu(\lambda, M) = \dim V_\lambda$, $V_\lambda = \mathrm{Ker}(\Delta_M + \lambda)$, as it is not difficult to see that

$$\mu(\lambda, M) \leqslant \mu(M) \max_{\phi\in V_\lambda, \|\phi\|_2=1} \|\phi\|_\infty^2,$$

where $\mu(M)$ is the volume of a compact manifold M [Sa2]. Such a bound is in fact sharp for a globally symmetric space of compact type (positive curvature), so that in this case the growth of eigenfuncions and multiplicities of eigenvalues are one and the same question and can be answered by Weyl's character formula. For negatively curved locally symmetric spaces one believes that $\mu(\lambda, M)$ are uniformly bounded or that at most $\mu(\lambda, M) \ll \lambda^\epsilon$.

For M a compact Riemannian manifold of strictly negative curvature, understanding the behavior of ϕ_j goes under the name of *Quantum Chaos* [Sa1]. The geodesic flow on the unit tangent bundle of M is chaotic: it is ergodic with positive Lyapunov exponents and positive entropy. The semiclassical limit $\hbar \to 0$ of the quantized system corresponds precisely to the large eigenvalue limit. One such statement is the long-standing Quantum Unique Ergodicity conjecture, which states that $|\phi_j|^2$ become uniformly distributed as $\lambda_j \to \infty$. One a priori plausible way in which QUE may fail is if a weak limit of measures associated to $|\phi_j|^2$ is concentrated along some proper totally geodesic submanifold M' of M, meaning that high-energy eigenfunctions achieve unexpectedly large values close to M'; this phenomenon is referred to as scarring.

Given the strong mixing properties of the geodesic flow, one does not expect that local bounds (0.1) capture the whole truth. The conjecture that

$$\|\phi_j\|_\infty \ll \lambda_j^{(n-r)/4-\delta} \tag{0.2}$$

for some $\delta > 0$ is known as the *subconvexity problem* in this context. To our knowledge, there is currently no known way to effectively employ chaotic dynamics alone to bring in global geometry of M toward improving estimates on growth of eigenfunctions, and not much is known in the way of results.

However, more can be said about extremal behavior of eigenfunctions on certain arithmetic hyperbolic manifolds, where, as we will see below, one can employ the family of Hecke correspondences to bring some global geometry of M into the picture. In this work, we address the case of arithmetic hyperbolic 3-manifolds. A subconvexity result of type (0.2) for a specific arithmetic hyperbolic 3-manifold can be found in Koyama [Ko].

In the case of arithmetic hyperbolic surfaces, Iwaniec and Sarnak conjectured [Iw-Sa] that high-energy eigenfunctions satisfy $\|\phi_j\|_\infty \ll \lambda_j^\epsilon$. A naïve generalization of this statement to all arithmetic hyperbolic manifolds (of arbitrary dimension) is, however, false, as was first established by Rudnick and Sarnak [Ru-Sa]. In particular, the random wave model (see Berry [Be]), which predicts that, in a completely ergodic system, quantum states exhibit random patterns of interference extrema and thus more temperate intensity fluctuations compared to the integrable case (specifically [Sa1] that $\|\phi_j\|_\infty \asymp \sqrt{\log \lambda_j}$), does not apply universally. The true asymptotic order of magnitude of $\|\phi_j\|_\infty$ can be markedly different depending on the geometric or functorial properties of the arithmetic manifold in question. A more thorough discussion can be found in sections 0.5 and 0.6.

In our principal result, Theorem 1 (stated in section 0.4), we expand on the method of [Mi] and decribe a natural class of arithmetic hyperbolic 3-manifolds on which high-energy eigenfunctions actually exhibit power growth. We also discuss why we expect that, up to commensurability, our class is precisely the class of arithmetic hyperbolic 3-manifolds on which eigenfunctions exhibit power growth and address the question of identifying the exceptional eigenfunctions.

0.2. Hyperbolic 3-manifolds.

We now give a description of arithmetic hyperbolic 3-manifolds in concrete terms. This proceeds in many ways analogously to the more widely familiar case of surfaces; we refer the reader to [Ma-Re] and [El-Gr-Me1] for details. The universal cover of hyperbolic 3-manifolds is the upper half-space $\mathfrak{H} = \{v = (z,r) : z \in \mathbb{C}, r > 0\}$, equipped with measure $\mathrm{d}\mu v = |\mathrm{d}z|\mathrm{d}r/r^3$ and the hyperbolic Laplacian $\Delta = r^2(\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial r^2) - r\partial/\partial r$. The group $\mathrm{PSL}_2\mathbb{C} \cong \mathrm{Isom}\mathfrak{H}$ acts transitively on $\mathfrak{H}$ by

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

analogously to the case of the upper half-plane $\mathfrak{h}$, upon identification $\mathfrak{H} \cong \mathrm{PSL}_2\mathbb{C}/\mathrm{PSU}_2$, the above action corresponds to the regular action in $\mathrm{PSL}_2\mathbb{C}$.

Every discrete subgroup $\Gamma < \mathrm{PSL}_2\mathbb{C}$ (a Kleinian group) gives rise to a hyperbolic 3-manifold $M = \Gamma \setminus \mathfrak{H}$. If M has finite volume, a polyhedron with finitely many faces can be chosen as a (Dirichlet) fundamental domain for the action of Γ on $\hat{\mathfrak{H}}$, and its vertices on $\hat{\mathbb{C}}$ are called cusps. The spectral decomposition of $L^2(M)$ with respect to the positive semi-definite self adjoint operator $-\Delta$ proceeds similarly to the case of hyperbolic surfaces. One has a decomposition $L^2(M)=L^2_{\text{disc}}(M)\oplus L^2_{\text{Eis}}(M)$, where the discrete subspace corresponds to the pure point spectrum and is spanned by the Maass forms: $L^2_{\text{disc}}(M)=\oplus_{j=0}^{\infty}\mathbb{C}\phi_j$ with $(\Delta+\lambda_j)\phi_j=0$. For non-cocompact Γ, ϕ_j are Maass cusp forms for $j\geqslant j_0$, and there is also the continuous spectrum with L^2_{Eis} spanned by the Eisenstein series $E_{\mathfrak{a}}(ir)$, where $\mathfrak{a}$ runs through the set of cusps. Maass cusp forms satisfy the local bound $\|\phi_j\|_\infty\ll\lambda_j^{1/2}$. A subconvex bound of type (0.2) is expected to hold for all M of constant negative curvature, but, for non-arithmetic M, not much beyond the local bound is known (subconvexity and bounded multiplicities are expected).

Arithmetic hyperbolic 3-manifolds are associated to arithmetic Kleinian groups, which we describe next. As usual, $\mathrm{M}_2\mathbb{C}$ denotes the matrix quaternion algebra, $\mathrm{SL}_2\mathbb{C}$ is its group of units, and $P:\mathrm{SL}_2\mathbb{C}\to\mathrm{PSL}_2\mathbb{C}$ is the projection $\gamma\mapsto\{\pm\gamma\}$. Let L be a number field with exactly one complex embedding pair $(\rho,\bar\rho)$, and let $\mathfrak{O}=R_L$ be its ring of integers. Let further A be a quaternion algebra over L ramified at all real places. ρ can be extended to an L-embedding of quaternion algebras $A\hookrightarrow\mathrm{M}_2\mathbb{C}$; one way to do this is to fix a representation $A=\left(\frac{a,b}{L}\right)$, a fixed square root $\sqrt{\rho(a)}\in\mathbb{C}$, and $\omega,\omega'\in A$ such that $\omega^2=A$, $\omega'^2=b$ and $\omega\omega'+\omega'\omega=0$, and embed

$$\begin{aligned}x&=x_0+x_1\omega+x_2\omega'+x_3\omega\omega'\\&\mapsto\begin{pmatrix}\rho(x_0)+\rho(x_1)\sqrt{\rho(a)}&\rho(x_2)+\rho(x_3)\sqrt{\rho(a)}\\\rho(b)(\rho(x_2)-\rho(x_3)\sqrt{\rho(a)})&\rho(x_0)-\rho(x_1)\sqrt{\rho(a)}\end{pmatrix}.\end{aligned}\tag{0.3}$$

For every $\mathfrak{O}$-order $\mathcal{O}<A$ and the group $\mathcal{O}^1$ of its elements of norm one, the group $\Gamma=P\rho(\mathcal{O}^1)$ is a cofinite discrete subgroup of $\mathrm{PSL}_2\mathbb{C}$ [Ma-Re]. Kleinian groups commensurable with such Γ and the corresponding 3-manifolds $M=\Gamma\setminus\mathfrak{H}$ are called arithmetic. An arithmetic hyperbolic 3-manifold M is always of finite volume and is compact if and only if A is a division quaternion algebra, which is true of all A except the case $A=\mathrm{M}_2L$, $L=\mathbb{Q}(\sqrt{-d})$, when Γ is conjugate to a congruence subgroup of the Picard group $\mathrm{PSL}_2\mathfrak{O}$. We will be dealing with the case when $\mathcal{O}$ is an Eichler order and Γ is the unit group $P\rho(\mathcal{O}^1)$.

The action of $\mathrm{PSL}_2\mathbb{C}$ on $\mathfrak{H}$ lifts to action $\mathrm{GL}_2\mathbb{C}$ via composition with $\gamma\mapsto\{\pm\gamma/\sqrt{\det\gamma}\}$. On every arithmetic hyperbolic 3-manifold M, with notation as above, one can define a family of Hecke operators $T_{\mathfrak{n}}$ on $L^2(M)$, where $\mathfrak{n}$ runs through ideals of $\mathfrak{O}$, which we describe in more detail in section 1. In the case when L is of class number one, Hecke operator of order $\mathfrak{n}=(\eta)$ can be defined essentially by averaging the right-regular action over representatives of the finitely many Γ-orbits of elements of $\mathcal{O}$ of norm η. Hecke operators are self-adjoint, commute with each other and with the Laplacian Δ, so they can be simultaneously diagonalized with an orthonormal basis consisting of Hecke-Maass eigenforms ϕ_j: $(\Delta+\lambda_j)\phi_j=0$ (we also write $\lambda_j=1+r_j^2$), $T_{\mathfrak{n}}\phi_j=\lambda_j(\mathfrak{n})\phi_j$.

Arithmetic Kleinian groups form a very special class of finite-covolume Kleinian groups. One can associate to each finite-covolume Kleinian group Γ two inherent geometric invariants, its invariant trace field $k\Gamma$ and invariant trace algebra $A\Gamma$, a quaternion algebra over $k\Gamma$; commensurable Kleinian groups have the same invariant trace field and algebra. The Identification Theorem states that Γ is an arithmetic Kleinian group if and only if $k\Gamma$ has precisely one complex embedding pair, $A\Gamma$ is ramified at all real places of $k\Gamma$, and traces of all elements of Γ are algebraic integers. For an arithmetic Kleinian group Γ, invariant trace field and algebra completely determine its commensurability class and, for such a group Γ commensurable with $P\rho(\mathcal{O}^1)$ with notations as above, we have that $k\Gamma = L$ and $A\Gamma = \rho(A)$ precisely.

0.3. Our approach, QCM-points, and manifolds of Maclachlan–Reid type.

Soundararajan [So] pointed how one can use Dirichlet polynomials to give omega results for special values of L-functions in various families; this idea is often referred to as the method of resonators. The crucial observation in the context of asymptotic study of L^∞-behavior of Hecke-Maass eigenfunctions on arithmetic hyperbolic manifolds, employed in [Mi] and here, is that the appropriate analogue of Dirichlet polynomials is to weigh the twists of the pre-trace formula by Hecke operators with a parameter sequence. More precisely, we compare averages of the form

$$\sum_{j=0}^{\infty} h(r_j/T)\Big|\sum_{\mathfrak{N}\mathfrak{n}\leqslant M} a(\mathfrak{n})\lambda_j(\mathfrak{n})\Big|^2 |\phi_j(v)|^2 \quad\text{and}\quad \sum_{j=0}^{\infty} h(r_j/T)\Big|\sum_{\mathfrak{N}\mathfrak{n}\leqslant M} a(\mathfrak{n})\lambda_j(\mathfrak{n})\Big|^2, \tag{0.4}$$

where h is an appropriate fixed non-negative Schwarz function. As we will see in sections 3 and 4, these spectral averages $Q_1(a)$ and $Q(a)$ can be evaluated asymptotically if M is in an appropriate range compared to T, and the leading terms in these evaluations are certain quadratic forms in $a(\mathfrak{n})$'s. In section 5, the parameter sequence will be chosen to maximize the quotient $Q_1(a)/Q(a)$; this will prove that some $|\phi_j(v)|$ must assume values as large as those to be announced in the statement of Theorem 1. On one hand, this approach is reminiscent of the method of mollification for L-functions; on the other hand, thinking about $T_\mathfrak{p}$ as $\mathfrak{p}$-adic Laplacians, it brings all $\mathfrak{p}$-adic neighborhoods of v into the picture along with its archimedean neighborhood. We might call our approach *the spectral resonator method*.

In analyzing extreme values of eigenforms on hyperbolic 3-manifolds, one encounters features markedly different from the case of surfaces in several ways, and the first one is that the hyperbolic 3-space has no complex structure and no off-the-shelf concept of CM-points. The property of CM-points which is crucial in [Mi], after an application of twisted pre-trace formula to a spectral average of type (0.4), is that comparably many Hecke correspondents $\rho(\gamma)z$ return to z. In the upper half-space, the stabilizer $G_v = \mathrm{Stab}(v, \mathrm{SL}_2\mathbb{C})$ of any $v \in \mathfrak{H}$ can be identified with SU_2, but, more importantly for us, $\mathbb{H}_v := G_v\mathbb{R}_0^+$ naturally carries the structure of the Hamilton quaternions algebra. We have that $\mathrm{Stab}(v, \mathrm{GL}_2\mathbb{C}) = \mathbb{C}^\times\mathbb{H}_v^\times$, and every Hecke correspondence γ that is to have $\rho(\gamma)v = v$ is of the form $\lambda\gamma_0$ with $\gamma_0 \in \mathbb{H}_v$, $\lambda^2 \in L^\times$. The goal is to have as many such correspondences $\gamma \in A$ as possible. It is therefore natural to define v to be a *QCM-point* if $B_v := \rho^{-1}(\rho(A) \cap \mathbb{H}_v)$, an algebra over the maximal totally real subfield K of L, is of full dimension (four), that is, if B_v is a quaternion algebra over K, by necessity ramified at all archimedean places. Note that if v is to be of the sort we just described, we must have $A = B_v \otimes_K L$, and this is a serious restriction on A: not every arithmetic 3-manifold contains some QCM-points to begin with. We introduce the following definition.

Definition. *An arithmetic hyperbolic 3-manifold $M = \Gamma \setminus \mathfrak{H}$ is called a QCM-manifold if the invariant trace algebra A of the Kleinian group Γ is of the form $A = B \otimes_K L$ for some quaternion algebra B over the maximal totally real subfield K of the invariant trace field L of Γ which is ramified at all infinite places. If ρ denotes the underlying embedding $A \hookrightarrow \mathrm{GL}_2\mathbb{C}$, then we say that* *a point $v \in \mathfrak{H}$ is a QCM-point if it is stabilized by $P\rho(B)$ for some quaternion algebra B over K ramified at all infinite places such that $A = B \otimes_K L$.*

We remark that, on an arithmetic hyperbolic surface with an invariant trace field K and an invariant trace algebra B (so that K is a totally real field and B is a quaternion algebra over K ramified at all real places except one), one can analogously define a point z to be a CM-point if it is stabilized by $P\rho(F)$ for some totally imaginary quadratic extension $F \leqslant B$ of K which splits B; this is the characterization used in [Mi].

Let now v be a QCM-point, so that $\mathrm{Stab}(v, A^\times) = L^\times B_v^\times$. For a large $\mathfrak{n} \subset \mathfrak{o} = R_K$ there are, roughly speaking, $\mathfrak{N}_K\mathfrak{n}$ correspondences $\gamma \in B_v$, according to theorems about representations of integers in totally real number fields by quaternary quadratic forms. We now explain how the optimization problem for quadratic forms Q_1 and Q naturally leads to an additional condition that $[L : K] = 2$. Optimal resonators $a(\mathfrak{n})$ are essentially concentrated on $\mathfrak{n} \subset \mathfrak{o}$, which is natural. It is less intuitive that to be able to produce *any* growth of $Q_1(a)/Q(a)$ (and hence of $|\phi_j(v)|$), it is not enough to merely have "many" $\gamma \in B_v \cap \Gamma(\mathfrak{n})$. It turns out that there is a critical exponent δ_0, such that if there are $\mathfrak{N}\mathfrak{n}^{\delta+o(1)}$ correspondences $\gamma \in B_v \cap \Gamma(\mathfrak{n})$, then the maximum value of $Q_1(a)/Q(a)$ exhibits (purely) *power growth* if $\delta > \delta_0$ and *no growth at all* if $\delta < \delta_0$. This is the result of Theorem 2, which we discuss later in section 0.6. In particular, if we did turn to (non-QCM) points v for which B_v is of less than full dimension, we could not deduce that $|\phi_j(v)|$ grow. Among those algebras A which allow a full quaternion algebra B_v (corresponding to v being a QCM-point on a QCM-manifold), only those with $[L : K] = 2$ allow good Diophantine control over Hecke actions (meaning that γv can return only so close to a QCM-point v without actually hitting it) — and it turns out that it is for arithmetic manifolds with these and only these underlying quaternion algebras that one can produce enough many Hecke correspondents $\gamma v = v$ in the meaning of Theorem 2. We propose to name these manifolds after Colin Maclachlan and Alan W. Reid, who first discovered their remarkable distinguishing geometric property which we describe next.

Definition. *We say that a QCM-manifold $M = \Gamma \setminus \mathfrak{H}$ is of Maclachlan–Reid type if the invariant trace field of Γ is a quadratic extension of its maximal totally real subfield.*

Maclachlan and Reid [Ma-Re] are concerned with classifying arithmetic hyperbolic 3-manifolds which contain immersed totally geodesic surfaces, or equivalently, the arithmetic Kleinian groups Γ which contain non-elementary Fuchsian subgroups. They show that such a Γ actually contains an *arithmetic* Fuchsian subgroup G and that their respective invariant trace fields $L = k\Gamma$ and $K = kG$ and invariant trace algebras $A = A\Gamma$ and $B = AG$ satisfy $[L : K] = 2$ and $A \cong B \otimes_K L$. As always, $L = K(\sqrt{D})$ has exactly one complex place $(\rho, \bar{\rho})$, and the quaternion algebra B is ramified at all archimedean places *except* $\rho|_K$. The converse is also true: given any two quaternion algebras A and B with these properties, every arithmetic group obtained from A contains arithmetic Fuchsian subgroups obtained from B, and correspondingly the 3-manifold arising from this group contains immersed arithmetic surfaces arising from subgroups of B. Moreover, there are then infinitely many other non-isomorphic quaternion algebras B (with prescribed ramification at infinite places) which yield as $B \otimes_K L$ this same L-algebra A — namely, one can prescribe at will the ramification of B at primes ramified or inert in $L|K$ at which A splits. This class of arithmetic hyperbolic 3-manifolds containing immersed totally geodesic surfaces coincides with the class of manifolds of

Maclachlan–Reid type in our definition. Namely, $B_1 = \left(\frac{a,b}{K}\right) \leftrightarrow \left(\frac{aD,bD}{K}\right) = B_2$ $(\rho(a), \rho(b) > 0)$ trivially corresponds (in a non-canonical way) to each quaternion algebra B_1 over K ramified at all infinite places except $\rho|_K$ a quaternion algebra B_2 over K ramified at all infinite places such that $B_1 \otimes_K L \cong B_2 \otimes_K L$, and vice versa.

The results of Maclachlan and Reid provide a sharp dichotomy in the class of arithmetic hyperbolic 3-manifolds: an arithmetic 3-manifold contains either contains no immersed totally geodesic surfaces, or it is of Maclachlan–Reid type, and then it contains infinitely many incommensurable immersed arithmetic surfaces, which correspond to incommensurable arithmetic Fuchsian subgroups of Γ. A manifold of Maclachlan–Reid type contains infinitely many QCM-points (corresponding to infinitely many non-isomorphic quaternion algebras B ramified at all infinite places such that $A = B \otimes_K L$), at which eigenfunctions are shown in Theorem 1 to achieve large values. It would appear extremely interesting to explain the analytic statements of Theorem 1 in terms of *geometry* of the immersed surfaces.

0.4. Main result.

For the manifolds M of Maclachlan–Reid type, on which $\|\phi_j\|_\infty$ can be shown to get large, they get handsomely large: this is the content of our principal result.

Theorem 1. *Let L be a number field with exactly one complex embedding pair, let K be its maximal totally real subfield, and suppose that $[L:K] = 2$ and that K and L are of narrow class number one. Let $\mathfrak{O} = R_L$ be the ring of integers of L, A be a quaternion algebra over L admitting QCM-points, $\mathcal{O}$ be an Eichler $\mathfrak{O}$-order in A and $\mathcal{O}^1$ be the group of elements of norm one in $\mathcal{O}$. Let ρ be an L-embedding of A in $\mathrm{M}_2\mathbb{C}$ and let $\Gamma = P\rho(\mathcal{O}^1) < \mathrm{PSL}_2\mathbb{C}$ be the corresponding arithmetic Kleinian group.*

The Hilbert space

$$L_0^2(\Gamma \setminus \mathfrak{H}) = \begin{cases} L^2(\Gamma \setminus \mathfrak{H}), & \text{if } \Gamma \text{ is cocompact,} \\ (L^2_{\mathrm{Eis}})^{\perp}, & \text{if } \Gamma \text{ is non-cocompact of finite volume} \end{cases}$$

has an orthonormal basis decomposition $L_0^2(\Gamma \setminus \mathfrak{H}) = \overline{\bigoplus_{j \geqslant 0} \mathbb{C}\phi_j}$, where ϕ_j are Hecke-Maass eigenforms (for $j \geqslant j_0$), with $(\Delta + \lambda_j)\phi_j = 0$. Then for every fixed QCM-point $v \in \mathfrak{H}$, we have, as $j \to \infty$,

$$|\phi_j(v)| = \Omega\left(\lambda_j^{\frac{1}{4}(1+\mathrm{O}(1/\log\log\lambda_j))}\right). \tag{0.5}$$

The version of Theorem 1 presented here includes the technical assumption that the number fields involved are of narrow class number one. This simplifies calculations with Hecke operators in our classical treatment, but there can be little doubt that results hold in the general case.

0.5. Exceptional eigenfunctions.

Statement of Theorem 1 does not point to some particular subset of eigenfunctions ϕ_j with large values at v, but the optimal weights

$$\sum_{\mathfrak{N}\mathfrak{n} \leqslant M, \mathfrak{n} \subset \mathfrak{o}} \mu_K^2(\mathfrak{n})\lambda_j(\mathfrak{n})$$

(in a variant of (0.4)) can be significantly larger at some distinguished ϕ_j than at others, strongly suggesting that it is in fact the corresponding forms ϕ_j which contribute to the power growth of the weighted average. In this section, we precisely classify forms for which this happens in the context of Theorem 1 and compare our results with the previously known particular results. In case of the Picard group $\Gamma = \mathrm{PSL}_2\mathfrak{O}_d$, the Dirichlet series

$$\sum_{n=1}^{\infty} \frac{\lambda_j(n)}{n^{s+1/2}} \tag{0.6}$$

is known as the Asai L-function of ϕ_j and can be analyzed using the Rankin convolution method. This was first performed by Asai [As1] for the analogous Hilbert modular case, where it was shown that the Asai L-function possesses an analytic continuation and has a pole at $s = 1$ if and only if $\lambda_j(\mathfrak{n}) = \lambda_j(\bar{\mathfrak{n}})$ for all $\mathfrak{n} \subset \mathfrak{O}_d$, which happens precisely when ϕ_j is a *base change lift* from some $\mathrm{PSL}_2\mathbb{Z}$-automorphic form. Proofs of these statements in [As1] are conditional on a mild, generally believed nonvanishing hypothesis, but they are now known unconditionally by Krishnamurthy [Kr] and Ramakrishnan [Ra]. We also refer the reader to Takase [Tak1] for results in the case of forms on quotients of the upper half-space.

Theorem 1 is not the first result in which power growth (0.5) is exhibited for some arithmetic 3-manifolds. Another way to approach values of eigenfunctions at special points are period formulae, going back to the important formula of Waldspurger [Wa]. Using his relative trace formula, Jacquet proved that, in the case of GL_n over a quadratic extension $E|F$, periods with respect to (adelic) unitary groups do not vanish precisely for representations which are base changes from F. Recently, Lapid and Offen [La-Of] proved a beautiful exact formula evaluating such periods of base change forms in terms of special values of L-functions at 1. Specialized to the case of congruence subgroups of $\mathrm{GL}_2\mathbb{Q}(\sqrt{-d})$, their formula reads as

$$\Big| \sum_{i \in \Lambda_\alpha} c_i \Phi(x_i) \Big|^2 \sim_\alpha |P_\alpha(\phi)|^2 \frac{\Lambda(1, \phi \times \tilde{\phi} \otimes \omega)}{\mathrm{Res}_{s=1} \Lambda(s, \phi \times \tilde{\phi})},$$

where Φ is the base change lift of a cusp form ϕ on $\mathrm{GL}_2\mathbb{Q}$, ω is the character attached to $E|F$ by class field theory, the weighted sum on the left-hand side is over the genus of the Hermitian form defined by α and P_α is a certain explicit product of local factors (in particular, $P_e = 1$). Λ-functions on the right-hand side are completed L-functions: by applying standard bounds for their finite parts and analyzing the Γ-factors with Stirling's formula one obtains $\|\Phi_j\|_\infty \gg \lambda_j^{1/4+o(1)}$. In the split case that it covers, this formula is extremely sharp in that it proves the lower bound for every base change form.

Another functorial approach involving theta lifts was used by Rudnick and Sarnak [Ru-Sa] to prove an omega result (of the strength of our Theorem 1) for eigenfunctions on a certain compact arithmetic hyperbolic 3-manifold. Suppose that Q is an integral quadratic form of signature (3,1) anisotropic over $\mathbb{Q}$. The group $\mathrm{PSL}_2\mathbb{C}$ can be identified with $G(\mathbb{R})$, the connected component of the identity in the orthogonal group of Q [El-Gr-Me1]. With $\Gamma = G(\mathbb{Z})$, K the orthogonal group of the majorant of Q and $D = \operatorname{disc} Q$, a classical theta function $\theta : \mathfrak{h} \times G(\mathbb{R}) \to \mathbb{C}$ can be constructed as in Shintani [Shin] so that integration against θ (against $\bar{\theta}$, respectively) corresponds to each eigenfunction on a compact arithmetic 3-manifold $X = \Gamma \setminus G(\mathbb{R})/K$ a cusp form of weight one on a congruence arithmetic surface $Y = \Gamma_0(4D) \setminus \mathfrak{h}$, and vice versa, with equivariance with respect to actions of Laplacians and Hecke operators. Images of cusp forms on Y in $L^2(X)$ are theta lifts, and one sees by counting eigenvalues that they span a proper subspace of $L^2(X)$ (in fact, of $\asymp \lambda^{3/2}$ eigenvalues on X with $\lambda_j \leqslant \lambda$, only $\asymp \lambda$ come from theta lifts); its orthogonal complement is the kernel of the theta correspondence. Now, if $x_1, x_2, \dots, x_h$ are representatives of the finitely many Γ-orbits on the quadric $V_m = \{\mathtt{x} : Q(\mathtt{x}) = m\}$ and $\Phi \in L^2(X)$, then a suitably weighted sum $\sum_{k=1}^h w_k^{-1}\Phi(g_k)$ ($g_k x_k = \pm x_0$, where $\operatorname{Stab} x_0 = K$) can be recognized as essentially m^{th} Fourier coefficient of the theta correspondent of Φ; in particular, it vanishes for all Φ orthogonal to theta lifts. As

$$\sum_{\lambda_j \leqslant \lambda} \left| \sum_{k=1}^h \Phi_j(g_k) \right|^2 \sim \frac{1}{(4\pi)^{3/2}} \sum_{k=1}^h \frac{1}{w_k^2} \lambda^{3/2},$$

by the pre-trace formula, and as all but $\asymp \lambda$ of the inner sums vanish, Rudnick and Sarnak conclude that $\|\Phi_j\|_\infty = \Omega(\lambda_j^{1/4})$.

We now compare our Theorem 1 with these particular results obtained through different approaches and present a conjectural unified context in which they can be viewed. We have not checked all details; precise statements and proofs will be found in our joint work with Takloo-Bighash [Mi-Ta].

Let B be a quaternion algebra over K ramified at all real places except ρ such that $A = B \otimes_K L$. (So this would be the B_1 from section 0.4.) Let $^\sigma$ denote the automorphism of A which acts trivially on B and extends the nontrivial Galois action on $L|K$, let $\bar{\ }$ denote the quaternion algebra conjugation on A, and let V_2 denote $\{a \in A \colon a = \bar{a}^\sigma\}$ considered as a K-quadratic space with respect to the quaternion algebra norm. Then V_2 is of signature $(3,1)$ and there is a natural injection $\epsilon_2 : a \mapsto (x \mapsto a x \bar{a}^\sigma)$ of $A^\times \hookrightarrow \mathrm{GO}(V_2)$. With analogous notation, consider $V_1 = \{g \in \mathrm{M}_2 L \colon g = {}^\top g^\sigma\}$ as a K-quadratic space with respect to the determinant norm, which is of signature $(3,1)$ and admits a natural injection $\epsilon_1 : \mathrm{GL}_2 L \hookrightarrow \mathrm{GO}(V_1)$ (which is a restriction of the classical isomoprhism $\mathrm{GL}_2 L \times K^\times / N_{L|K} L^\times \cong \mathrm{GO}(3,1)$). Finally, let ω be the classical isomorphism $\mathrm{GL}_2 K \cong \mathrm{GSp}_2 K$. For a reductive group G, let $\mathcal{A}(G)$ (initially) denote the set of cuspidal automorphic representations of G.

Consider the following diagram:

$$\begin{tikzcd}
\mathcal{A}(\mathrm{GSp}_2 F) \arrow[rr, "\theta"] \arrow[dr, equal, "\omega^*"] & & \mathcal{A}(\mathrm{GO}(V_1)) \arrow[dr, "\epsilon_1^*"] & \\
& \mathcal{A}(\mathrm{GL}_2 K) \arrow[rr, "\mathrm{BC}"] & & \mathcal{A}(\mathrm{GL}_2 L) \\
\omega^* \circ \mathrm{JL}(\mathcal{A}(B^\times)) \arrow[uu, hook, "\subset"] \arrow[rr, "\theta"'] \arrow[dr, equal, "\omega^* \circ \mathrm{JL}"'] & & \mathcal{A}(\mathrm{GO}(V_2)) \arrow[uu, hook, "\mathrm{GO-JL}"'] \arrow[dr, "\epsilon_2^*"] & \\
& \mathcal{A}(B^\times) \arrow[uu, hook, "\mathrm{JL}"'] \arrow[rr, "\mathrm{IF-BC}"'] & & \mathcal{A}(A^\times) \arrow[uu, hook, "\mathrm{JL}"']
\end{tikzcd}$$

Here, JL denote both classical Jacquet-Langlands liftings $\mathcal{A}(B^\times) \hookrightarrow \mathcal{A}(\mathrm{GL}_2K)$ and $\mathcal{A}(A^\times) \hookrightarrow \mathcal{A}(\mathrm{GL}_2L)$, θ denotes theta liftings $\mathcal{A}(\mathrm{GSp}_2K) \to \mathcal{A}(\mathrm{GO}(V_1))$ and $\mathcal{A}(\mathrm{GSp}_2K) \to \mathcal{A}(\mathrm{GO}(V_2))$, and ω^*, ϵ_1^*, and ϵ_2^* are the obvious morphisms arising from ω, ϵ_1, and ϵ_2. In [Mi-Ta], we construct liftings IF-BC : $\mathcal{A}(B^\times) \to \mathcal{A}(A^\times)$ and GO-JL : $\mathcal{A}(\mathrm{GO}(V_2)) \to \mathcal{A}(\mathrm{GO}(V_1))$ such that the diagram above commutes. We recall here that Strong Multiplicity One statements hold for all groups in question.

Recall that, for a given quaternion algebra A over L as in Theorem 1, $A = B \otimes_K L$ holds for infinitely many non-isomorphic quaternion algebras B; everything we have stated above is true for any one of them. Among these algebras, there is a unique one with minimal ramification; we assume from now on that B is such. Representations of other quaternion algebras B' simply inject naturally as proper subspaces of $\mathcal{A}(B^\times)$ via Jacquet-Langlands correspondences in the above diagram.

We have already noted that forms ϕ_j with the largest contribution to the power growth in (0.4) are precisely those whose Asai L-function $\sum_{\mathfrak{n}\subset\mathfrak{o}} \mu_K^2(\mathfrak{n})\lambda_j(\mathfrak{n})/(\mathfrak{N}_K\mathfrak{n})^{s+1/2}$ (analogous to (0.6)) has a pole at $s = 1$. We now describe what this means for the corresponding irreducible automorphic representation $\pi \in \mathcal{A}(A^\times)$. The Jacquet-Langlands lifting [Ha-Ta] associates to π a non-dihedral automorphic representation $\pi' = \mathrm{JL}(\pi) \in \mathcal{A}(\mathrm{GL}_2L)$. Asai L-function of π' also has a pole at $s = 1$, so that $\pi' \simeq BC(\pi_0') \otimes \eta$ for some $\pi_0' \in \mathcal{A}(\mathrm{GL}_2K)$ and an idele class character η on L trivial on ideles of K. It is checked in [Mi-Ta] that this π_0' is a Jacquet-Langlands lift $\pi_0' = \mathrm{JL}(\pi_0)$ of some automorphic respresentation $\pi_0 \in \mathcal{A}(B^\times)$, so that

$$\pi \otimes \bar{\eta} = \text{IF-BC}(\pi_0) = \epsilon_2^* \circ \theta(\pi_0''),$$

with $\pi_0'' = \omega^* \circ \mathrm{JL}(\pi_0)$. In this way, we see that representations corresponding to forms ϕ_j with the largest contribution in (0.4) can be characterized both in terms of base change lifts and in terms of theta lifts from GSp_2. In the non-compact and compact cases, our description of distinguished forms corresponds to the specific cases exhibited in Lapid–Offen and Rudnick–Sarnak, respectively.

To add perspective, we mention several results related to the above diagram of representations. It was first stated by Takase [Tak2] and proved by Krishnamurthy [Kr] that the Asai L-functions on general linear groups over an imaginary quadratic field L can be recognized as essentially the "standard" L-functions on the orthogonal group $\mathrm{O}(3,1)$ ($\mathrm{O}(V_2)$ in our case). Representations of $\mathrm{O}(V_2)$ whose L-functions have a pole at $s = 1$ are shown in Ginzburg–Jiang–Soudry [Gi-Ji-So] to be twists of theta lifts from GSp_2. Quadratic base change was realized as a theta lift by Cognet [Co] in the context of local fields as well as by Asai [As2] as a global lifting of holomorphic forms to imaginary quadratic fields. Flicker [Fl] showed that, more generally, a cuspidal representation of GL_nL whose twisted tensor L-function (as the Asai L-function is known in this context) has a pole at $s = 1$ is distinguished, subject to certain conditions on ramification (subsequent work was done on more precise versions). In case of quadratic extensions of GL_2, distinguished representations are base change lifts from GL_2K; this circle of questions is connected to the so-called Jacquet's conjecture.

Moreover, and this would be the principal result of [Mi-Ta], we expect the above diagram to commute even if $\mathcal{A}(G)$ denote spaces of individual automorphic *forms*. Here, Jacquet-Langlands lifts of forms are realized by Shimizu's lifting as in Watson [Wa] (see also [Shim]), and theta lifts are realized with specific Schwarz functions as in Shintani [Shin]. In view of Maclachlan-Reid's results, our hyperbolic 3-manifold $M = \Gamma\backslash\mathfrak{H}$ contains some immersed arithmetic surface M_0 corresponding to an arithmetic Fuchsian subgroup of B (of minimal level). Our construction would therefore describe a lifting of forms on M_0 (of appropriate weight) to forms on M, which would respect actions of Laplacian and Hecke operators. A form on M would have the property that its Asai L-function has a pole at $s = 1$ if and only if it lies in the image of this lifting IF-BC. We call such forms *exceptional*. Taking into account (0.4), where these forms are present with overwhelmingly large weights, we conjecture (see Section 0.6 below) that it is precisely exceptional forms which achieve power growth in Theorem 1.

Our setup is not entirely unlike Maass's original construction of special nonholomorphic wave forms on arithmetic quotients of the upper half-plane from "wave forms" (i.e. additive characters) on immersed geodesics. As in Maass's case, there is no obvious geometric way to effect the lifting. However, an important difference is the fact that in our situation forms on an immersed surface lift to a sequence of exceptional forms on a *fixed* 3-manifold. We also refer the reader to a series of papers by Kudla and Millson (see e.g. [Ku-Mi]), which relate geodesic cycles in locally symmetric spaces to specific dual harmonic forms, compare these duals to forms arising from the global Weil representation, and construct liftings of cohomologies of quotient spaces of orthogonal and unitary groups.

In addition to the classes of three-dimensional manifolds exhibiting power growth of eigenfunctions discussed above, we mention that Donnelly [Do] has generalized the method of [Ru-Sa] to produce, for every $n \geqslant 5$, arithmetic hyperbolic manifolds of dimension n on which a sequence of Laplacian eigenfunctions achieves $\|\phi_j\|_\infty = \Omega(\lambda_j^{(n-4)/4})$. Donnelly's construction involves arithmetic quotients $G_{\mathfrak{O}} \setminus G_{\mathbb{R}} \times \tilde{G}_{\mathbb{R}}$, where G $(\tilde{G})$ is the orthogonal group of the quadratic form $x_1^2 + x_2^2 + \cdots + x_n^2 + \sqrt{d}x_{n+1}^2$ $(x_1^2 + x_2^2 + \cdots + x_n^2 - \sqrt{d}x_{n+1}^2$, respectively), $\mathfrak{O}$ is the ring of integers of $\mathbb{Q}(\sqrt{d})$, with d a positive square-free integer, and $G_{\mathfrak{O}}$ is the set of $\mathfrak{O}$-points of G embedded into $G_{\mathbb{R}} \times \tilde{G}_{\mathbb{R}}$ as $g \mapsto (g, \tilde{g})$. Exceptional eigenfunctions are shown to be found among the theta lifts from certain eigenfunctions on quotients of $\mathfrak{h} \times \mathfrak{h}$ by appropriate congruence subgroups.

While constructions in specific cases can produce precise results, the method of this paper has, in our view, the advantage of being well suited to the general problem of *identifying* precise classes of manifolds on which eigenfunctions show power growth.

0.6. Remarks about discreteness.

The results of this work and [Mi] are part of a much more ambitious goal of general understanding the L^∞-growth of automorphic forms and, more generally, periods. As our results show, certain arithmetic manifolds show features not expected generically, of which the most striking is power growth along subsequences of eigenfunctions seen for specific families of quotients, and those arithmetic 3-manifolds that do are distinguished by both their geometric and functorial properties. In general, one wants to understand the structure that underlies this special behavior and learn how to distinguish quotients that exhibit it. Here we would like to point out a particular quantitative aspect of the connection between Hecke operators and large values of automorphic forms, for the case of 3-manifolds.

Theorem 2. *Let L be a number field with exactly one complex embedding pair, K be its maximal totally real subfield, $\mathfrak{O}$ and $\mathfrak{o}$ be their respective rings of integers, and $d = [L : K]$.*

Let further $f:\mathfrak{o}\to\mathbb{R}^+$ *be a multiplicative function and* $\Delta\geqslant 0$ *be such that*

$$f(\mathfrak{p})=\mathfrak{N}\mathfrak{p}^{\Delta/d}(1+O(\mathfrak{N}\mathfrak{p}^{-\delta})),\quad f(\mathfrak{p}^{u+v})\leqslant f(\mathfrak{p}^u)f(\mathfrak{p}^v).$$

For $\mathfrak{n}\subset\mathfrak{O}$*, define*

$$f_K(\mathfrak{n}_0)=\begin{cases} f(\mathfrak{n}_0), \textit{if } \mathfrak{n}=\mathfrak{n}_1^2\mathfrak{n}_0,\ \mathfrak{n}_0\subset\mathfrak{o},\ \mathfrak{n}_1 \textit{ minimal},\\ 0, \qquad\qquad\qquad\qquad\qquad\qquad \textit{else}.\end{cases}$$

Consider the quadratic forms

$$\begin{aligned} B(b)&=\sum_{\mathfrak{N}\mathfrak{d}\leqslant M}\mathfrak{N}\mathfrak{d}\,b(\mathfrak{d})^2\\ B_1(b)&=\sum_{\mathfrak{N}\mathfrak{d}\leqslant M}\mathfrak{N}\mathfrak{d}\sum_{\mathfrak{N}\mathfrak{m},\mathfrak{N}\mathfrak{n}\leqslant M/\mathfrak{N}\mathfrak{d}} b(\mathfrak{d}\mathfrak{m})b(\mathfrak{d}\mathfrak{n})\sum_{\mathfrak{u}^2|\mathfrak{m}}\sum_{\mathfrak{v}^2|\mathfrak{n}}\mu_L(\mathfrak{u})\mu_L(\mathfrak{v})f_K\left(\frac{\mathfrak{m}\mathfrak{n}}{\mathfrak{u}^2\mathfrak{v}^2}\right).\end{aligned}$$

Then

$$\begin{aligned} \max_b\frac{B_1(b)}{B(b)}&\asymp 1, && \textit{if } \Delta<(d-1)/2,\\ \max_b\frac{B_1(b)}{B(b)}&\asymp M^{(1+2\Delta)/d)-1}, && \textit{if } \Delta>(d-1)/2.\end{aligned}$$

Theorem 2 is a general optimization result about quadratic forms and is proved by a Peter-Paul procedure exactly like the one we employed in Lemma 2, which contains the case needed for the proof of Theorem 1. We omit the proof to avoid unnecessary repetition and instead focus on interpretation of the above statement. In the application to our problem, function f counts the number of Hecke correspondences fixing a chosen special point, and the conditions on f are natural and satisfied by all families of special points on 3-manifolds we have examined. The important message of the general optimization result of Theorem 2 for our method is that, to be able to showcase large values of eigenfunctions, it is not enough to merely produce points fixed under "any many" Hecke correspondences: there is a critical exponent of power growth that needs to be surpassed.

It is this property that distinguishes the Maclachlan-Reid type manifolds of our Theorem 1 and their QCM-points from all other 3-manifolds. For example, as noted in our proof, families of QCM-points fixed under at least $\mathfrak{N}_K\mathfrak{n}$ correspondences $\gamma\in\Gamma(\mathfrak{n})$ ($\mathfrak{n}\subset\mathfrak{o}$) exist on manifolds coming from unit groups of Eichler orders of $A=B\otimes_K L$ regardless of the degree $[L:K]$. While our proof is fine-tuned for the case $[L:K]=2$, remainder terms can be kept in check in all other cases as well. However, the above analysis shows that ultimately this growth ($\Delta=1$) is not enough to produce large values of $Q_1(a)/Q(a)$ unless $[L:K]=2$. Similar reasoning applies to families of special points whose stabilizers over K are not of full rank. For example, let $a,b_1\in\mathbb{Z}_{<0}$, $L=\mathbb{Q}(\sqrt{a})$, and $B_1=\left(\frac{a,b_1}{\mathbb{Q}}\right)$, so that L injects as a subalgebra $L_1<B_1$. Let further $A_1=B_1\otimes_{\mathbb{Q}}L$ and consider the (non-compact) arithmetic manifold corresponding to the unit group of some Eichler order in A_1. Motions in $P\rho(L_1)$ fix all points along a geodesic c on this manifold. In fact, if $M=\mathbb{Q}(b_2)$ is a real quadratic field such that $(\operatorname{disc}M,\operatorname{disc}L)=1$, $B_2=\left(\frac{a,b_2}{M}\right)$ and $A_2=B_2\otimes_M(LM)$, and if v is a QCM-point with respect to A_2 (so that $v\in c$), then one finds that, over A_1, v actually

has a Diophantine property that if γv is (in a certain precise sense) close enough to v for some $\gamma \in \Gamma(\mathfrak{n})$ with $\mathfrak{n} = (\eta)$ coprime to $\operatorname{disc} M \operatorname{disc} L$, then $\eta = \eta_1^2 n_0$ with $n_0 \in \mathbb{Z}$ and $\gamma/\eta_1 \in L_1$. This gives us a family of special points with the nice Diophantine separation property at which the number of modular correspondences fixing v is precisely counted as the number of ideals of L of prescribed norm; these points lie on a geodesic which as a whole is fixed by these correspondences. However, growth of this magnitude ($\Delta = 0$) once again cannot be successfully used to find large values of eigenfunctions (and by this we mean, not even the sub-power growth of the type exhibited in Theorem 1).

That there is such a strong discreteness condition goes well hand in hand with the recent purity conjecture of Sarnak [Sa2]. Let M be a compact hyperbolic locally symmetric space of dimension n, and let $E(M)$ be the set of accumulation points of

$$\frac{\log \|\phi_j\|_\infty}{\log \lambda_j}$$

as $j \to \infty$. (We discuss rank one case here; the full conjecture also covers higher rank cases, where more care is needed near the walls of Weyl chambers.) The general bounds from the introduction imply that $E(M) \subset [0, (n-1)/4]$, and subconvexity is the statement that $E(M) \subset [0, (n-1)/4)$. The purity conjecture states that, for an arithmetic manifold M,

$$E(M) \subset \mathbb{Z}/4.$$

For arithmetic 3-manifolds, this conjecture predicts $E(M) \subset \{0, 1/4\}$ and our Theorem 1 provides a family of manifolds achieving $\|\phi_j\|_\infty = \Omega(\lambda_j^{1/4+o(1)})$. In general, power growth along subsequences of eigenfunctions is expected to be a rare phenomenon and one would like to understand for which manifolds it occurs.

The discussion above strongly suggests that our manifolds of Maclachlan-Reid type are in fact precisely the class of arithmetic hyperbolic 3-manifolds showing this behavior. That this class is also distinguished by the geometric property of having infinitely many immersed arithmetic surfaces is intriguing. Moreover, on such distinguished manifolds, we have seen that power growth occurs at points with a large rational stabilizer and along a subsequence of forms which can be characterized either in terms of the associated representations being functorial lifts from the division algebras giving rise to immersed surfaces, or as theta lifts from appropriate congruence subgroups. We make the following

Conjecture. *On arithmetic hyperbolic 3-manifolds other than those of Maclachlan-Reid type, one has*

$$\|\phi_j\|_\infty \ll \lambda_j^\epsilon.$$

The same estimate holds for non-exceptional forms on Maclachlan-Reid type 3-manifolds.

Both statements of the above conjecture are very strong and appear to us to be out of reach of current methods.

In light of the description of the exceptional cases and of the purity conjecture stated above, one is tempted to speculate that, in general, it is plausible that one would see increasing discrete "layers of power growth" of periods of automorphic forms, with "scarring" (term used loosely) of varying severity along certain special points, geodesics, or, more generally, totally geodesic submanifolds. In that sense, understanding the general picture presents a threefold task: classifying, in geometric or functorial terms, the distinguished class of manifolds exhibiting this scarring phenomenon, describing special points or orbits along which large values are attained, and identifying the subsequence of eigenfunctions which achieve them. This paper and the companion paper [Mi] shed light on these questions in cases of arithmetic 2- and 3-manifolds and their connection to the geometry of Hecke correspondences. It will be very interesting to see how the interplay between the expected discreteness of $E(M)$ and the relevant quadratic forms plays out in cases of higher dimension and rank.

0.7. Acknowledgments.

I would like to thank Peter Sarnak, my doctoral advisor, for suggesting this problem as well as for his selfless support and encouragement. I have had very helpful conversations with Ramin Takloo-Bighash, especially with regard to paragraph 0.5. Wee Teck Gan was very patient and provided me with valuable references. Jeffrey Lagarias offered genuine help and encouragement in the final phases of work. Dihua Jiang kindly provided me with his preprint [Gi-Ji-So]. I have also benefited from discussions with Brian Conrad, Jayce Getz, Gopal Prasad, and Akshay Venkatesh.

Portions of this work were completed while the author was a Liftoff Fellow of the Clay Mathematics Institute.

1. Pre-trace formula.

The spectral expansion of the automorphic kernel for Kleinian groups proceeds analogously to the case of Fuchsian groups. We refer the reader to [Iw] and [Mi] for the two-dimensional case and to [El-Gr-Me1] for details of the three-dimensional case, which is of interest to us. For $v_i=(z_i,r_i)\in\mathfrak{H}$ $(i=1,2)$ write $u(v_1,v_2)=\dfrac{|z_1-z_2|^2+r_1^2+r_2^2}{2r_1r_2}$; u is a point-pair invariant as $u(v_1,v_2)=\cosh\rho(v_1,v_2)$, where ρ denotes the hyperbolic distance on $\mathfrak{H}$. Suppose h is an even function holomorphic in the strip $R=\{r\in\mathbb{C}\colon\ |\mathfrak{Im}\, r|<1/2+\epsilon\}$ such that $h(r)\ll(|r|+1)^{-3-\epsilon}$ in this strip. Such an $h:R\to\mathbb{C}$ has a Selberg transform $k:\mathbb{R}_{\geqslant1}\to\mathbb{C}$ given explicitly [Ko] as a composite

$$h(r)=\int_{-\infty}^{\infty}e^{iru}g(u)\,\mathrm{d}u,\quad g(u)=Q(\cosh u),\quad k(t)=-\frac{1}{2\pi}Q'(t).\tag{1.1}$$

From a point-pair invariant $k(v,w)=k(u(v,w))$ we can build the automorphic kernel $K(v,w)=\sum_{\gamma\in\Gamma}k(\gamma v,w)$, which in turn admits the spectral expansion

$$K(v,w)=\sum_{j\geqslant0}h(r_j)\phi_j(v)\overline{\phi_j(w)}+\delta_\Gamma\sum_{\mathfrak{a}}c_{\mathfrak{a}}\int_{-\infty}^{+\infty}h(r)E_{\mathfrak{a}}(v,ir)\overline{E_{\mathfrak{a}}(w,ir)}\,\mathrm{d}r,\tag{1.2}$$

where $\lambda_j=1+r_j^2$, $\delta_\Gamma=0$ or 1 according as Γ is cocompact or not, and, in the latter case, $\mathfrak{a}$ runs over the finite set of Γ-inequivalent cusps and $c_{\mathfrak{a}}$ are explicit positive constants defined in terms of the stabilizers of cusps. Among non-cocompact Γ, it suffices to consider the Picard groups $\mathrm{PSL}_2\mathfrak{O}$, where $\mathfrak{O}$ is the maximal order in some imaginary quadratic field $L=\mathbb{Q}(\sqrt{-d})$, as every non-cocompact arithmetic subgroup Γ with the invariant trace algebra $A=\mathrm{M}_2\mathbb{Q}(\sqrt{-d})$ is conjugate over A to some congruence subgroup Γ_1 of $\mathrm{PSL}_2\mathfrak{O}$, so that $L^2(\mathrm{PSL}_2\mathfrak{O}\setminus\mathfrak{H})\hookrightarrow L^2(\Gamma_1\setminus\mathfrak{H})\cong L^2(\Gamma\setminus\mathfrak{H})$. Under the technical assumptions of Theorem 1, there is only one equivalence class of cusps, represented by ∞, and $c_\infty=c=|\mathfrak{O}^*|/(4\pi\sqrt{\operatorname{disc}L})$.

We can choose g so that the kernel k is non-negative and compactly supported and its transform h is even and positive on the real and imaginary axes. As in [Mi], we can also arrange that, for some $\alpha>0$, h satisfies $h(x)\ll\exp(-\alpha\log|x|\log\log|x|)$ for large $|x|$; this innocent trick allows us to keep the constant in the exponent of the statement of Theorem 1 with an explicit error term instead of $1+o(1)$. To analyze values of $\phi_j(v)$, we will be using a "twisted" version of (1.2), eventually setting $v=w$. For effective asymptotic analysis, it is essential to have test functions k for which the left-hand size localizes quickly close to the diagonal, i.e. for u close to $u=1$ in $k(u)$. With this in mind, we fix a large $T>0$ and use (1.2) with $h_T(r)=h(r/T)$ and k_T corresponding to $g_T(u)=Tg(Tu)$; it is easy to see that $k_T(t)=0$ for $t-1\gg T^{-2}$, $|k_T(t)|\ll T^3$ for all t, and $k_T(1)\sim cT^3$ with $c=(-1/2\pi)g''(0)=(1/4\pi^2)\int_{-\infty}^{\infty}r^2h(r)\,\mathrm{d}r>0$.

Theorem 1 is concerned with eigenfunctions on arithmetic 3-manifolds. These manifolds come equipped with a family of modular correspondences, which we now proceed to describe for $M=\Gamma\backslash\mathfrak{H}$ with a group Γ as in the statement of Theorem 1. (We refer the reader to Eichler [Ei] for the original treatment.) We denote by $N(\alpha)$ the quaternion algebra norm of $\alpha\in A$ (also called the reduced norm of α). Let $\mathfrak{n}=\eta\mathfrak{O}$ be an ideal in $\mathfrak{O}$, where the generator η is chosen in the subgroup $L_+^\times$ of numbers positive in all real embeddings, and let

$$\mathcal{O}(\mathfrak{n})=\{\alpha\in\mathcal{O}\colon\ \eta^{-1}N(\alpha)\in\mathfrak{O}^*\cap L_+^\times\}.\tag{1.3}$$

$Z(\mathfrak{O})=\mathcal{O}(\mathfrak{O})\cap\rho^{-1}P^{-1}(\mathrm{id})$ is the center of $\mathcal{O}(\mathfrak{O})$, and $\mathcal{O}(\mathfrak{O})=Z(\mathfrak{O})\mathcal{O}^1$ acts on $\mathcal{O}(\mathfrak{n})$ by multiplication on the left with finitely many orbits $Z(\mathfrak{O})\mathcal{O}^1\setminus\mathcal{O}(\mathfrak{n})$. These give rise to the modular correspondences on M and in turn to the Hecke operators $T_{\mathfrak{n}}:L^2(M)\to L^2(M)$ defined by

$$T_{\mathfrak{n}}f(v)=\sum_{\alpha\in Z(\mathfrak{O})\mathcal{O}^1\backslash\mathcal{O}(\mathfrak{n})}f(\rho(\alpha)v).$$

(Hecke operators are typically normalized by scaling by a factor of $1/\sqrt{\mathfrak{N}\mathfrak{n}}$; the above is the normalization we will use.) There is an ideal $\mathfrak{q}_{\mathcal{O}}\subset\mathfrak{O}$ such that $T_{\mathfrak{n}}$'s with $(\mathfrak{n},\mathfrak{q}_{\mathcal{O}})=\mathfrak{O}$ satisfy the usual multiplicative property

$$T_{\mathfrak{n}}T_{\mathfrak{m}}=\sum_{\mathfrak{d}|(\mathfrak{m},\mathfrak{n})}\mathfrak{N}\mathfrak{d}T_{\mathfrak{mn}/\mathfrak{d}^2},\tag{1.4}$$

and, together with the Laplacian Δ, they form a commutative algebra of self-adjoint operators which may be simultaneously diagonalized to obtain an orthonormal Hecke eigenbasis $\{\phi_j\}_{j\geqslant0}$ of $L_0^2(M)$ with which we are concerned in Theorem 1:

$$(\Delta+\lambda_j)\phi_j=0,\quad T_{\mathfrak{n}}\phi_j=\lambda_j(\mathfrak{n})\phi_j.\tag{1.5}$$

In the non-cocompact case $\Gamma=\mathrm{PSL}_2\mathfrak{O}$, the operators $T_{\mathfrak{n}}$ may be applied to the Eisenstein series and one finds that they act on them by scalars, namely by divisor sums normalized by a factor of $\sqrt{\mathfrak{N}\mathfrak{n}}$ (see [El-Gr-Me1] for the explicit Fourier expansion of $E_{\mathfrak{a}}(v,s)$ and [Iw] for the analogous case of surfaces). Applying $T_{\mathfrak{n}}$ in variable v to (1.2) and unfolding, one obtains the explicit formula contained in the following lemma.

Lemma 1. (Twisted pre-trace formula.) *Let L be a number field with exactly one complex embedding pair of class number one and let $\mathfrak{O}=R_L$ be the ring of integers of L. Let A be a quaternion algebra over L, $\mathcal{O}$ be an Eichler $\mathfrak{O}$-order and $\mathcal{O}^1$ be the group of elements of norm one in $\mathcal{O}$.*

Let ρ be an L-embedding of A in $\mathrm{M}_2\mathbb{C}$, $\Gamma=P\rho(\mathcal{O}^1)$ be the corresponding arithmetic Kleinian group, and let $(\phi_j)_{j=0}^{\infty}$ be a basis of $L_0^2(\Gamma\setminus\mathfrak{H})$ consisting of Hecke-Maass eigenforms satisfying (1.5). In case when Γ is non-cocompact, assume $\Gamma=\mathrm{PSL}_2\mathfrak{O}$ and let the Eisenstein series $E(v,s)$ on $\Gamma\setminus\mathfrak{H}$ be defined as in (4.1)*. Let further $h_T:R\to\mathbb{C}$ and $k_T:\mathbb{R}_{\geqslant 1}\to\mathbb{C}$ be a Selberg transform pair, let $k_T(v,w)=k_T(u(v,w))$, and let $Z(\mathfrak{O})$ and $\mathcal{O}(\mathfrak{n})$ ($\mathfrak{n}\subset\mathfrak{O}$) be as in* (1.3)*. Then*

$$\sum_{\gamma\in Z(\mathfrak{O})\backslash\mathcal{O}(\mathfrak{n})} k_T(\rho(\gamma)v,w)=\sum_{j=0}^{\infty}h_T(r_j)\lambda_j(\mathfrak{n})\phi_j(v)\overline{\phi_j(w)} \tag{1.6}$$

$$+\,\delta_\Gamma c\sqrt{\mathfrak{N}\mathfrak{n}}\sum_{\mathfrak{n}=\mathfrak{a}\mathfrak{d}}\int_{-\infty}^{\infty}h_T(r)(\mathfrak{N}\mathfrak{a}/\mathfrak{N}\mathfrak{d})^{ir/2}E(v,ir)\overline{E(w,ir)}\,\mathrm{d}r. \qquad \square$$

This twisted pre-trace formula is the starting point for our asymptotic analysis. We will first deal with the cocompact case in sections 2 and 3. In section 4, we collect the additional terms present in the non-cocompact case; they will be less than the main terms by a power of T in the final asymptotic analysis.

2. The Diophantine lemma.

One of the key constructs of [Mi] is a Diophantine lemma which tells us that Hecke correspondents $\gamma(z)$ of a fixed CM-point can "return back" only so close to z before actually hitting it, which enables us to precisely count γ's with a nonzero contribution on the geometric side of the twisted pre-trace formula (1.5) (after proper localization of k_T to the diagonal as $T\to\infty$). QCM-points on arithmetic 3-manifolds do not have such nice separation properties *except* on manifolds described in our theorem. On the other hand, correspondences $\gamma\in\mathcal{O}(\mathfrak{n})$ fixing v can in any case be counted precisely. In what follows, for a positive definite quaternary integral quadratic form q over a totally real number field K with the ring of integers $\mathfrak{o}$ and for any $n\in\mathfrak{o}$, let $r(q,n)=\#\{\mathfrak{n}\in\mathfrak{o}^4:\ n=q(\mathfrak{n})\}$.

Lemma 2. *Let Γ be an arithmetic Kleinian group as in Theorem 1, and let $v\in\mathfrak{H}$ be a QCM-point fixed by $(B\otimes_\rho\mathbb{R})^\times\hookrightarrow\mathrm{GL}_2\mathbb{C}\twoheadrightarrow\mathrm{PSL}_2\mathbb{C}$, where $B\leqslant A$ is a quaternion algebra over K ramified at all infinite places.*

Denote by $\mathfrak{o}=R_K$ the ring of integers of K, by $J_{\mathfrak{O}}$ the group of fractional ideals of $\mathfrak{O}$, and let $J_{\mathfrak{o}}=\{a\mathfrak{O}\in J_{\mathfrak{O}}:\ a\in K^\times\}$. For $\mathfrak{n}\subset\mathfrak{O}$, let $\mathcal{O}(\mathfrak{n})$ be as in (1.3)*, and let $Z(\mathfrak{O})$ be the center of $\mathcal{O}(\mathfrak{O})$. For each $\mathfrak{b}\subset\mathfrak{O}$, fix an integral basis of the $\mathfrak{o}$-lattice $\mathcal{O}_K^{\mathfrak{b}}=\mathfrak{b}\mathcal{O}\cap B$ and write the quaternion algebra norm on this lattice as a quaternary quadratic form $q_{\mathfrak{b}}$ on $\mathfrak{o}^4$; for $\mathfrak{n}_0\in J_{\mathfrak{o}}$, define $\mathcal{O}_K^{\mathfrak{b}}(\mathfrak{n}_0)$ analogously to* (1.3)*. Finally, let $\mathfrak{c}_1,\dots,\mathfrak{c}_h$ be minimal integral representatives of classes in $\{\mathfrak{a}\in J_{\mathfrak{O}}:\ \mathfrak{a}^2\in J_{\mathfrak{o}}\}/J_{\mathfrak{o}}$, and write $\mathfrak{c}_i=c_i\mathfrak{O}$, $\mathfrak{c}_i^2\cap\mathfrak{o}=\tilde{c}_i\mathfrak{o}$ with $c_i\in L_+^\times$, $\tilde{c}_i\in K_+^\times$.*

a) There exists an ideal $\mathfrak{q}_{\mathcal{O}_K} \subset \mathfrak{O}$ such that Hecke correspondences $\gamma \in Z(\mathfrak{O}) \setminus \mathcal{O}(\mathfrak{n})$ $(\mathfrak{n} \subset \mathfrak{O})$ such that

$$\rho(\gamma)v = v$$

exist only if $\mathfrak{n} = \mathfrak{n}_1^2\mathfrak{n}_0$ for some $\mathfrak{n}_1 \in J_{\mathfrak{O}}$, $\mathfrak{n}_1 \subset \mathfrak{q}_{\mathcal{O}_K}^{-1}$ and $\mathfrak{n}_0 \subset \mathfrak{o}$. In that case, take the unique such representation with a minimal $\mathfrak{n}_1$ and write $\mathfrak{n}_1 = \eta_1\mathfrak{O}$ and $\mathfrak{n}_0 = \eta_0\mathfrak{o}$ with $\eta_1 \in L_+^\times$ and $\eta_0 = \eta/\eta_1^2 \in K_+^\times$. For $1 \leqslant i \leqslant h$, let $\mathfrak{b}_i = (\mathfrak{n}_1\mathfrak{c}_i)^{-1} \cap \mathfrak{O}$. Then, $\rho(\gamma)v = v$ if and only if

$$\gamma \in \left(\mathcal{O}(\mathfrak{n}) \cap \bigsqcup_{i=1}^{h} \eta_1 c_i Z(\mathfrak{O}) B\right) = \bigsqcup_{i=1}^{h} \eta_1 c_i Z(\mathfrak{O}) \mathcal{O}_K^{\mathfrak{b}_i}(\mathfrak{n}_0/\mathfrak{c}_i^2),$$

$$\textit{and} \quad \left| Z(\mathfrak{O}) \setminus \left(\mathcal{O}(\mathfrak{n}) \cap \bigsqcup_{i=1}^{j} \eta_1 c_i Z(\mathfrak{O}) B\right)\right| = \frac{1}{2}\sum_{i=1}^{h} r(q_{\mathfrak{b}_i}, \eta_0/\tilde{c}_i). \tag{2.1}$$

b) In the case $[L : K] = 2$, there exists a constant $C > 0$ depending on $B \leqslant A$ and $\mathcal{O}$ only, such that $\gamma \in \mathcal{O}(\mathfrak{n})$ satisfies

$$|\rho(\gamma)v - v| \leqslant C(\mathfrak{N}\mathfrak{n})^{-\Delta}$$

with $\Delta = 1/2$ if and only if actually $\rho(\gamma)v = v$.

Proof of Lemma 2. Let $\mathcal{O}_K = \mathcal{O} \cap B$. Every $\beta \in B \subset A$ has some multiple $\eta\beta \in \mathcal{O}$, $\eta \in \mathfrak{O}$, and therefore also a multiple $(N_{L|K}\eta)\beta \in \mathcal{O}_K$, $N_{L|K}\eta \in \mathfrak{o}$. Hence $\mathcal{O}_K$ is an $\mathfrak{o}$-order in B; let $(\xi_i)_{i=1}^4$ be an $\mathfrak{o}$-integral basis of $\mathcal{O}_K$. Call an element $\gamma_0 = \sum x_i\xi_i \in \mathcal{O}_K$ primitive if $\sum(x_i\mathfrak{o}) = \mathfrak{o}$. We fix $\mathfrak{q}_{\mathcal{O}_K} \subset \mathfrak{O}$ such that $\mathcal{O} \subset \mathfrak{q}_{\mathcal{O}_K}^{-1}\mathcal{O}_K$.

We prove that every $\gamma \in \mathcal{O}(\mathfrak{n})$ such that $\rho(\gamma)v = v$ can be written as

$$\gamma = \lambda\gamma_0, \tag{2.2}$$

where $\lambda = a/b \in \mathfrak{q}_{\mathcal{O}_K}^{-1}$, $a, b \in \mathfrak{O}$, $(a, b) = \mathfrak{O}$, and $\gamma_0 \in \mathcal{O}_K^{(b)}$ is primitive, and that this representation is unique up to multiplication by units in $\mathfrak{o}^*$. Indeed, write $\gamma = \sum y_i\xi_i$ with $y_i \in L^\times$; as $\rho(\gamma)v = v$, we also have that $\rho(\gamma) = \tilde{\lambda}(\sum \tilde{x}_i\rho(\xi_i))$ for some $\tilde{\lambda} \in \mathbb{C}^\times$, $\tilde{x}_i \in \mathbb{R}$. By changing $\tilde{\lambda}$ as necessary, we can ensure that $\tilde{x}_i = \rho(x_i)$ for some $x_i \in \mathfrak{o}$ such that $\gamma_0 = \sum x_i\xi_i$ is primitive, as well as $\tilde{\lambda} = \rho(\lambda)$ with $\lambda \in L^\times$; write $\lambda = a/b$ with $a, b \in \mathfrak{O}$, $(a, b) = \mathfrak{O}$. From $a\gamma_0 \in b\mathcal{O}$ it follows that $\gamma_0 \in b\mathcal{O} \cap B = \mathcal{O}_K^{(b)}$, while

$$(\lambda) = \textstyle\sum(\lambda x_i\mathfrak{O}) \subset \mathfrak{q}_{\mathcal{O}_K}^{-1}, \tag{2.3}$$

so that the decomposition $\gamma = \lambda\gamma_0$ has all the required properties. The uniqueness claim is immediate from (2.3).

In particular, we see from (2.2) that if there are Hecke correspondences $\gamma \in \mathcal{O}(\mathfrak{n})$ such that $\rho(\gamma)v = v$, then $\mathfrak{n} = \tilde{\mathfrak{n}}_1^2\tilde{\mathfrak{n}}_0$, with $\tilde{\mathfrak{n}}_1 = (\lambda) \subset \mathfrak{q}_{\mathcal{O}_K}^{-1}$, $\tilde{\mathfrak{n}}_0 = (N\gamma_0) \subset \mathfrak{o}$. It is easy to see that the set of all $\mathfrak{n}_1 \in J_{\mathfrak{O}}$, $\mathfrak{n}_1 \subset \mathfrak{q}_{\mathcal{O}_K}^{-1}$ such that $\mathfrak{n} = \mathfrak{n}_1^2\mathfrak{n}_0$ for some $\mathfrak{n}_0 \subset \mathfrak{o}$ is closed under addition; let $\mathfrak{n} = \mathfrak{n}_1^2\mathfrak{n}_0$ be the unique such representation with a minimal $\mathfrak{n}_1$ and write $\mathfrak{n}_1 = \eta_1\mathfrak{O}$ and $\mathfrak{n}_0 = \eta_0\mathfrak{o}$ with $\eta_1 \in L_+^\times$, $\eta_0 \in K_+^\times$. For every $\gamma = \lambda\gamma_0$ written as in (2.2) and the corresponding $\tilde{\mathfrak{n}}_1 = (\lambda)$, we have that $(\tilde{\mathfrak{n}}_1/\mathfrak{n}_1)^2 \in J_{\mathfrak{o}}$, so that there is a unique $1 \leqslant i \leqslant h$ such that $\tilde{\mathfrak{n}}_1 = c_i\tilde{n}_1\mathfrak{n}_1$ for some $\tilde{n}_1 \in \mathfrak{o}$. This allows us to write (2.2) in the form

$$\gamma = \lambda^1\gamma_0^1, \tag{2.4}$$

where $\lambda^1 = \lambda/\tilde{n}_1 = a/(b\tilde{n}_1)$, $(\lambda^1) = c_i\mathfrak{n}_1$, and $\gamma_0^1 = \tilde{n}_1\gamma_0^1 \in \mathcal{O}_K^{(b_i)}$ with $b_i\mathfrak{O} = (\lambda^1)^{-1} \cap \mathfrak{O} \supset \tilde{n}_1(b\mathfrak{O})$.

Summing up, we have shown that every $\gamma \in \mathcal{O}(\mathfrak{n})$ with $\rho(\gamma)v = v$ has exactly one representation in the form (2.2), and, grouping these representations according to the $\mathfrak{c}_i$ for which $\tilde{\mathfrak{n}}_1 \in \mathfrak{c}_i\mathfrak{n}_1 J_{\mathfrak{o}}$, each of them yields a unique representation of the form (2.4) with $(\lambda^1) = \mathfrak{c}_i\mathfrak{n}_1$ and $\gamma_0^1 \in \mathcal{O}_K^{\mathfrak{b}_i}(\mathfrak{n}_0/\mathfrak{c}_i^2)$, $\mathfrak{b}_i = (\mathfrak{c}_i\mathfrak{n}_1)^{-1} \cap \mathcal{O}$. Conversely, every product $\lambda^1\gamma_0^1$ of the form (2.4) with $(\lambda^1) = \mathfrak{c}_i\mathfrak{n}_1$ and $\gamma_0^1 \in \mathcal{O}_K^{\mathfrak{b}_i}(\mathfrak{n}_0/\mathfrak{c}_i^2)$ yields a $\gamma \in \mathcal{O}(\mathfrak{n})$. Part (a) follows.

We move on to the proof of part (b). Let $L = K(\delta)$, where $\delta^2 \in K$ is negative at ρ and positive at all other places of K (by conditions on archimedean embeddings of L). In particular, $A = B \oplus B\delta$ as a vector space. On each of $B_{\mathbb{R}} = B \otimes_\rho \mathbb{R}$ and $\mathrm{M}_2\mathbb{C} \cong A \otimes_\rho \mathbb{C}$, we have the quaternion algebra norm N, the (hermitian) coordinate L^2-norm $\|\cdot\|_2^2$ with respect to a fixed basis of B, and the entry-by-entry L^2-norm $\|\rho(\cdot)\|_2^2$ of the image inside the matrix algebra $M_2\mathbb{C}$ under the embedding ρ. On $B_{\mathbb{R}}$, all three are positive definite quadratic forms and so equivalent; on $\mathrm{M}_2\mathbb{C}$, the L^2-norms are equivalent (so we can always use them interchangeably) and the algebra norm is bounded by either. As in Lemma 2 of [Mi], the inequality $\|AB\|_2 \leqslant \|A\|_2\|B\|_2$ for $A, B \in \mathrm{M}_2\mathbb{C}$ is a simple consequence of Cauchy-Schwarz inequality, and $B_{\mathbb{R}}$ and $\delta B_{\mathbb{R}}$ are literally perpendicular to each other in the coordinate L^2-norm.

Suppose now that $\gamma \in \mathcal{O}(\mathfrak{n})$ (that is, $\gamma \in \mathcal{O}$ with $\eta^{-1}N\gamma \in (\mathfrak{O}^* \cap L_+^\times)$) satisfies $|\rho(\gamma)v - v| \ll (\mathfrak{N}\mathfrak{n})^{-\Delta}$. The action $\tilde{\gamma} \mapsto P(\tilde{\gamma})v$ induces the continuous Iwasawa homeomorphism $\mathrm{GL}_2\mathbb{C}/\mathbb{C}^\times B_{\mathbb{R}}^\times \cong \mathfrak{H}$, so that $|\rho(\gamma)v - v| \ll (\mathfrak{N}\mathfrak{n})^{-\Delta}$ is equivalent with

$$\|\rho(\gamma)\lambda^{-1}\gamma_{\mathrm{f}}^{-1} - I\|_2 \ll (\mathfrak{N}\mathfrak{n})^{-\Delta}$$

for some $\gamma_{\mathrm{f}} \in B_{\mathbb{R}}^\times$ and some $\lambda \in \mathbb{C}^\times$, $|\lambda| = 1$. If we write $\lambda = \lambda_0 + \rho(\delta)\lambda_1$ ($\lambda_i \in \mathbb{R}$) and $\gamma = \gamma_0 + \delta\gamma_1$ ($\gamma_i \in B$), it follows that

$$\begin{aligned}\frac{\|\rho(\gamma_i) - \lambda_i\gamma_{\mathrm{f}}\|_2}{\|\gamma_{\mathrm{f}}\|_2} &\ll \frac{\|\rho(\gamma_0) - \lambda_0\gamma_{\mathrm{f}} + \rho(\delta\gamma_1) - \rho(\delta)\lambda_1\gamma_{\mathrm{f}}\|_2}{\|\gamma_{\mathrm{f}}\|_2} \\ &\leqslant \|\left(\rho(\gamma) - \lambda\gamma_{\mathrm{f}}\right)\gamma_{\mathrm{f}}^{-1}\|_2 \ll (\mathfrak{N}\mathfrak{n})^{-\Delta}.\end{aligned}$$

For any $\gamma \in B$ and any $\tilde{\gamma} \in B_{\mathbb{R}}$, let γ_μ and $\tilde{\gamma}_\mu$ ($1 \leqslant \mu \leqslant 4$) denote the coordinates of γ and $\tilde{\gamma}$ with respect to a fixed basis of B. From $\rho(\gamma_i)_\mu = \lambda_i\gamma_{\mathrm{f}\mu} + \mathrm{O}((\mathfrak{N}\mathfrak{n})^{-\Delta}\|\gamma_{\mathrm{f}}\|_2)$ it follows that $\rho(N\gamma) = \lambda^2 N\gamma_{\mathrm{f}} + \mathrm{O}((\mathfrak{N}\mathfrak{n})^{-\Delta}\|\gamma_{\mathrm{f}}\|_2^2)$ and so in particular $\|\gamma_{\mathrm{f}}\|_2 \sim |\rho(\eta)|^{1/2}$, and further

$$|\rho(\gamma_{0\mu}\gamma_{1\nu} - \gamma_{1\mu}\gamma_{0\nu})| \ll |\rho(\eta)|(\mathfrak{N}\mathfrak{n})^{-\Delta}.$$

On the other hand, by pairing the real places of L so that (ρ_i, ρ_i') corresponds to a single place of K, we see from the simple inequality ($\epsilon = \rho_i(\delta) > 0$)

$$(ad - bc)^2 \ll \left((a + b\epsilon)^2 + (c + d\epsilon)^2\right)\left((a - b\epsilon)^2 + (c - d\epsilon)^2\right)$$

that we are justified in estimating

$$\begin{aligned}|\mathfrak{N}(\gamma_{0\mu}\gamma_{1\nu} - \gamma_{1\mu}\gamma_{0\nu})| &= |\rho(\gamma_{0\mu}\gamma_{1\nu} - \gamma_{1\mu}\gamma_{0\nu})|^2 \prod |\rho_i\rho_i'(\gamma_{0\mu}\gamma_{1\nu} - \gamma_{1\mu}\gamma_{0\nu})| \\ &\ll |\rho(\eta)|^2(\mathfrak{N}\mathfrak{n})^{-2\Delta} \prod \rho_i(\eta)\rho_i'(\eta) = (\mathfrak{N}\mathfrak{n})^{1-2\Delta}.\end{aligned}$$

As all $\gamma_{i\mu}$ belong to some fixed ideal in $\mathfrak{o}$, this implies that $\gamma_{0\mu}\gamma_{1\nu} = \gamma_{1\mu}\gamma_{0\nu}$ if the implied constant is small enough, but this means that γ_0 and γ_1 are scalar multiples of each other. This is precisely the condition for $\rho(\gamma)v = v$. □

Lemma 2 provides a lower bound for the geometric side of (1.2) in terms of the representation numbers of certain positive definite quadratic forms. This lower bound is obtained by precisely counting Hecke correspondences with $\rho(\gamma)v = v$ regardless of degree $[L : K]$, but it is only for $[L : K] = 2$, a situation we might term the Galois case, that the bound should be expected to be precise. In other cases, there are presumably Hecke correspondents $\rho(\gamma)v$ which come very close to v in an unpredictable fashion and the full picture touches upon very delicate Diophantine questions.

In the proof of the omega result of Theorem 1, we will for simplicity of notation restrict ourselves to just the contribution of $\mathfrak{b} = \mathfrak{O}$ (i.e. $\mathfrak{n}_1 \subset \mathfrak{O}$) and write $q = q_{\mathfrak{O}}$; the cost of doing so is at any rate at most a constant multiple. For the following discussion involving quadratic forms over totally real fields, we refer the reader to [Ki] and [SP] for standard facts. Starting from a positive definite quaternary integral quadratic form q over the totally real number field K as above, we can form its theta series

$$\theta(q,z) := \sum_{\mathfrak{n}\in\mathfrak{o}^4} e^{\pi i \mathrm{Tr}(\lambda^{-1}q(\mathfrak{n})z)} = \sum_{n\in\mathfrak{o}} r(q,n)e^{\pi i\mathrm{Tr}(\lambda^{-1}nz)},$$

where $z \in \mathfrak{h}^m$, where $m = [K : \mathbb{Q}]$, (λ) is the absolute different of K, and Tr is the linear extension of the field trace $K \to \mathbb{Q}$ to $\mathbb{C}^m \to \mathbb{C}$. The theta series can be defined in the obvious analogous way for a quadratic form over any $\mathfrak{o}$-lattice $\Lambda < K^4$ of full rank and is known to be a Hilbert modular form of weight two for a certain congruence subgroup of $\mathrm{PSL}_2\mathfrak{o}$ that is locally everywhere conjugate to a subgroup of type $\Gamma_0(N)$. Using the transformation formula for θ, it is seen that the values at cusps of such a theta series are expressed as Gauss sums over certain quotients $\Lambda/c\Lambda$, where c depends only on the congruence properties of (Λ, q), and so in particular are same for all forms in the genus of q.

Let us denote by $q_1 = q, q_2, \ldots, q_h$ the full set of representatives of isometry classes of forms (with their implicit underlying $\mathfrak{o}$-lattices) in the genus $\operatorname{gen} q$, by $\mathrm{O}(q_i)$ the finite group of isometries of q_i,

$$r(\operatorname{gen} q, n) = \left(\sum_{i=1}^{h}\frac{1}{|\mathrm{O}(q_i)|}\right)^{-1}\sum_{i=1}^{h}\frac{r(q_i,n)}{|\mathrm{O}(q_i)|},$$

and $\theta(\operatorname{gen} q, z) = \sum_{n\in\mathfrak{o}} r(\operatorname{gen} q, n)e^{\pi i\mathrm{Tr}(\lambda^{-1}nz)}$. Then our previous remarks show that $r_{\mathrm{cusp}}(q,n) := r(q,n) - r(\operatorname{gen} q, n)$ are Fourier coefficients of a *cusp* form $\theta(q,z) - \theta(\operatorname{gen} q, z)$. On the other hand, it is a classical theorem of Siegel that

$$r(\operatorname{gen} q, n) = c\mathfrak{N}_K(n)\prod_{\mathfrak{p}}\alpha_{\mathfrak{p}}(q,n),$$

with a constant $c > 0$ depending on K and q only. (We write $\mathfrak{N}_K$ for the absolute norm of (n) as an ideal of K to avoid confusion.) The "local densities" $\alpha_{\mathfrak{p}}(q,n)$ for places $\mathfrak{p}$ at which q is not unimodular are only partially understood, but a lower bound is known on their product which is uniform in n as long as n is not highly divisible by such $\mathfrak{p}$'s (of which there are finitely many) and

n is representable by q, which will be the case if we restrict n by congruence conditions modulo a certain $\mathfrak{q}_{\mathrm{gen}}$. In fact, noting that classes in $G=(\mathfrak{o}/\mathfrak{q}_{\mathrm{gen}})^{*2}<(\mathfrak{o}/\mathfrak{q}_{\mathrm{gen}})^{*}$ are representable by forms in the genus of q and that, under our assumptions, $\mathfrak{o}^{*+}=\mathfrak{o}^{*2}$, we see that there exists a finite set Ξ of ray class characters of $\mathfrak{o}$ modulo $\mathfrak{q}_{\mathrm{gen}}$ trivial on $\mathfrak{o}^{\times 2}$ (namely, the set of all ray class characters modulo $\mathfrak{q}_{\mathrm{gen}}$ trivial on G) such that, for any $(n,\mathfrak{q}_{\mathrm{gen}})=\mathfrak{o}$, the class $n+\mathfrak{q}_{\mathrm{gen}}$ is representable by forms in the genus of q if $\sum_{\chi\in\Xi}\chi((n))>0$. The product of the remaining local densities can be shown to be

$$\gg\prod_{\mathfrak{p}\mid n}(1+\chi_{\mathfrak{p}}(q)\mathfrak{N}_K\mathfrak{p}^{-1})$$

where $\chi_{\mathfrak{p}}(q)$ is ±1 according to whether the localization of the quadratic space of q at $\mathfrak{p}$ is hyperbolic or not. For our q, which is the norm form of a quaternion algebra A, it is clear that $\chi_{\mathfrak{p}}(q)=1$ whenever A is unramified at $\mathfrak{p}$. Summing up, this discussion proves the following

Lemma 3. *Let K be a totally real number field of narrow class number one with the ring of integers $\mathfrak{o}$. Let q be a positive definite quaternary integral quadratic form over K which is locally almost everywhere equivalent to the norm form of a certain quaternion algebra B over K, and, for every $n\in\mathfrak{o}$, let $r(q,n)=\#\{\mathfrak{n}\in\mathfrak{o}^4:\ q(\mathfrak{n})=n\}$. Then there is an ideal $\mathfrak{q}_q\subset\mathfrak{o}$ and a finite set Ξ of ray class of characters of $\mathfrak{o}$ such that for all $((n),\mathfrak{q}_q)=\mathfrak{o}$,*

$$r(q,n)\gg\sum_{\chi\in\Xi}\sigma_{K,\mu^2,\chi}((n))+\mathrm{O}(|r_{\mathrm{cusp}}(q,n)|). \tag{2.5}$$

Here, $\sigma_{K,\mu^2,\chi}(\mathfrak{n}_0)=\chi(\mathfrak{n}_0)\sum_{\mathfrak{n}_0=\mathfrak{e}\mathfrak{f},\,\mathfrak{e},\mathfrak{f}\subset\mathfrak{o}}\mathfrak{N}_K\mathfrak{e}\,\mu_K^2(\mathfrak{f})$, and $r_{\mathrm{cusp}}(q,n)$ are Fourier coefficients of a certain cusp Hilbert modular form of weight two for a certain congruence subgroup of $\mathrm{PSL}_2\mathfrak{o}$. □

Returning to the notation of Theorem 1, for $\mathfrak{n}\subset\mathfrak{O}$, $(\mathfrak{n},\mathfrak{q}_q)=\mathfrak{O}$, we set

$$\sigma_{K,\mu^2,\chi}(\mathfrak{n})=\sigma_{K,\mu^2,\chi}(\mathfrak{n}_0)\quad\text{and}\quad r^K_{\mathrm{cusp}}(\mathfrak{n})=r^K_{\mathrm{cusp}}(q,\eta_0) \tag{2.6}$$

if $\mathfrak{n}=\mathfrak{n}_1^2\mathfrak{n}_0$ with $\mathfrak{n}_0=\eta_0\mathfrak{o}$ ($\eta_0\in\mathfrak{o}^+$) and $\mathfrak{n}_1$ minimal, and $\sigma_{K,\mu^2,\chi}(\mathfrak{n})=r^K_{\mathrm{cusp}}(\mathfrak{n})=0$ otherwise. Lemmas 2 and 3 together give an effective lower bound on the number of Hecke correspondences $\gamma\in Z(\mathfrak{O})\setminus\mathcal{O}(\mathfrak{n})$ such that $\rho(\gamma)v=v$ in terms of numbers $\sigma_{K,\mu^2,\chi}(\mathfrak{n})$. A remarkable feature of the estimate of Lemma 3 is its near-universality: the specific quaternion algebra A and QCM-point $v=(z,r)$ are reflected in the implied constants, $\mathfrak{q}_q$, collection of characters Ξ, and the cuspidal term only.

3. Resonator and the asymptotics of $Q_1(a)$ and $Q(a)$.

We now fix a large $M>0$, introduce a "resonator" — a sequence of non-negative real numbers $a(\mathfrak{n})$ ($(\mathfrak{n},\mathfrak{q}_{\mathcal{O}}\mathfrak{q}_L\mathfrak{q}_q)=\mathfrak{O}$, $\mathfrak{N}\mathfrak{n}\leqslant M$), to be chosen later, and consider the following two spectral averages, where v is a QCM-point:

$$Q_1(a)=\sum_{j=0}^{\infty}h_T(r_j)\left|\sum_{\mathfrak{N}\mathfrak{n}\leqslant M}a(\mathfrak{n})\lambda_j(\mathfrak{n})\right|^2|\phi_j(v)|^2, \tag{3.1}$$

$$Q(a)=\sum_{j=0}^{\infty}h_T(r_j)\left|\sum_{\mathfrak{N}\mathfrak{n}\leqslant M}a(\mathfrak{n})\lambda_j(\mathfrak{n})\right|^2. \tag{3.2}$$

Sections 3 and 4 are devoted to the proof of the following Lemma 4.

Lemma 4. *Let Γ be an arithmetic Kleinian group as in Theorem 1, and let $v \in \mathfrak{H}$ be a QCM-point fixed by $(B \otimes_\rho \mathbb{R})^\times \hookrightarrow \mathrm{GL}_2\mathbb{C} \twoheadrightarrow \mathrm{PSL}_2\mathbb{C}$, where $B \leqslant A$ is a quaternion algebra over K ramified at all infinite places. Let $(\phi_j)_{j=0}^\infty$ be a basis of $L_0^2(\Gamma \setminus \mathfrak{H})$ consisting of Hecke-Maass eigenforms satisfying (1.5), let $h_T : R \to \mathbb{C}$ and $k_T : \mathbb{R}_{\geqslant 1} \to \mathbb{C}$ be a Selberg transform pair, let $a(\mathfrak{n})$ be a non-negative resonator as above, and define $Q_1(a)$ and $Q(a)$ as in* (3.1) *and* (3.2). *Then, for $M \ll T^2$ with a sufficiently small implied constant,*

$$Q_1(a) \gg k_T(1) \sum_{\chi \in \Xi} \sum_{\mathfrak{N}\mathfrak{d} \leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\mathfrak{N}\mathfrak{m},\, \mathfrak{N}\mathfrak{n} \leqslant M/\mathfrak{N}\mathfrak{d}} a(\mathfrak{d}\mathfrak{m})a(\mathfrak{d}\mathfrak{n})\sigma_{K,\mu^2,\chi}(\mathfrak{m}\mathfrak{n})$$
$$+ \mathrm{O}\left(k_T(1) \sum_{\mathfrak{N}\mathfrak{d} \leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\mathfrak{N}\mathfrak{m},\, \mathfrak{N}\mathfrak{n} \leqslant M/\mathfrak{N}\mathfrak{d}} a(\mathfrak{d}\mathfrak{m})a(\mathfrak{d}\mathfrak{n})|r_{\mathrm{cusp}}^K(\mathfrak{m}\mathfrak{n})| \right) + \delta_\Gamma Q_1^\infty(a),$$

$$Q(a) = k_T(1)\mu(\Gamma \setminus \mathfrak{H}) \sum_{\mathfrak{N}\mathfrak{d} \leqslant M} \left(\sum_{\mathfrak{d}=\mathfrak{e}\mathfrak{f}} \mathfrak{N}\mathfrak{e}\mu_L^2(\mathfrak{f}) \right) \left(\sum_{\mathfrak{N}\mathfrak{n} \leqslant \sqrt{M/\mathfrak{N}\mathfrak{d}}} a(\mathfrak{d}\mathfrak{n}^2) \right)^2$$
$$+ \mathrm{O}\left(\sum_{\mathfrak{N}\mathfrak{d} \leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\mathfrak{N}\mathfrak{m}, \mathfrak{N}\mathfrak{n} \leqslant M/\mathfrak{N}\mathfrak{d}} |a(\mathfrak{d}\mathfrak{m})a(\mathfrak{d}\mathfrak{n})|\ell_T(\mathfrak{m}\mathfrak{n}) \cdot e^{2C\frac{\log T}{\log\log T}} \right) + \delta_\Gamma Q^\infty(a),$$

where $\delta_\Gamma = 1$ if Γ is non-cocompact (in which case we assume $\Gamma = \mathrm{PSL}_2\mathfrak{O}$), and

$$Q_1^\infty(a) \ll T^{2+\epsilon} \sum_{\mathfrak{N}\mathfrak{d} \leqslant M} \mathfrak{N}\mathfrak{d} \left(\sum_{\mathfrak{N}\mathfrak{m} \leqslant M/\mathfrak{N}\mathfrak{d}} a(\mathfrak{d}\mathfrak{m})\sqrt{\mathfrak{N}\mathfrak{m}} \right)^2, \quad Q^\infty(a) \ll T^{1+\epsilon} \left(\sum_{\mathfrak{N}\mathfrak{m} \leqslant M} a(\mathfrak{m})\sqrt{\mathfrak{N}\mathfrak{m}} \right)^2.$$

Here, Ξ is a certain finite collection of ray class characters of K described in the proof of Lemma 3, $\sigma_{K,\mu^2,\chi}$ and r_{cusp}^K are functions on ideals of L defined in (2.6), *and $\ell_T(\mathfrak{n}) = \mathfrak{N}\mathfrak{n} + T^2$.*

Proof of Lemma 4. In $Q_1(a)$, we expand the square and use the multiplicative relation (1.4) for Hecke operators and the spectral expansion (1.5) to obtain

$$Q_1(a) = \sum_{\mathfrak{N}\mathfrak{m},\, \mathfrak{N}\mathfrak{n} \leqslant M} a(\mathfrak{m})a(\mathfrak{n}) \sum_{\mathfrak{d} | (\mathfrak{m},\mathfrak{n})} \mathfrak{N}\mathfrak{d} \sum_{\gamma \in Z(\mathfrak{o}) \setminus \mathcal{O}(\mathfrak{m}\mathfrak{n}/\mathfrak{d}^2)} k_T(\rho(\gamma)v, v).$$

For v a QCM-point, the resulting innermost sum can be estimated by Lemmas 2 and 3 as

$$Q_1(a) \gg k_T(1) \sum_{\mathfrak{N}\mathfrak{m},\, \mathfrak{N}\mathfrak{n} \leqslant M} a(\mathfrak{m})a(\mathfrak{n}) \sum_{\mathfrak{d} | (\mathfrak{m},\mathfrak{n})} \mathfrak{N}\mathfrak{d} \left(\sum_{\chi \in \Xi} \sigma_{K,\mu^2,\chi}(\mathfrak{m}\mathfrak{n}/\mathfrak{d}^2) + \mathrm{O}(|r_{\mathrm{cusp}}^K(\mathfrak{m}\mathfrak{n}/\mathfrak{d}^2)|) \right). \quad (3.3)$$

On the other hand, integrating (1.5) along the diagonal $v = w \in \Gamma \setminus \mathfrak{H}$ gives

$$Q(a) = \sum_{\mathfrak{N}\mathfrak{m}, \mathfrak{N}\mathfrak{n} \leqslant M} a(\mathfrak{m})a(\mathfrak{n}) \sum_{\mathfrak{d} | (\mathfrak{m},\mathfrak{n})} \mathfrak{N}\mathfrak{d} \sum_{\gamma \in Z(\mathfrak{O}) \setminus \mathcal{O}(\mathfrak{m}\mathfrak{n}/\mathfrak{d}^2)} \int_{\Gamma \setminus \mathfrak{H}} k_T(\rho(\gamma)v, v)\, \mathrm{d}\mu v. \quad (3.4)$$

We collect the contributions from individual $\gamma \in \mathcal{O}(\mathfrak{n})$ to $Q(a)$ according to the geometric type of $P\rho(\gamma)$. This proceeds somewhat parallel to the case of surfaces dealt with in [Mi]. By multiplying by a suitable totally positive unit we may assume that the representative $\eta \in L_+^\times$ of $\mathfrak{n} = \eta\mathfrak{O}$ is chosen so that $|\rho(\eta)| \asymp \sqrt{\mathfrak{N}\mathfrak{n}}$, $\rho_i(\eta) \asymp 1$. We note for reference that

$$\mathrm{Tr}^2 P\rho(\gamma) - 4 = \frac{\mathrm{Tr}^2\rho(\gamma) - 4\rho(\eta)}{\rho(\eta)} = -\frac{4\rho(N(\gamma-\gamma_0))}{\rho(\eta)}.$$

In particular, we note that $\mathrm{Tr}\, P\rho(\gamma) = \pm 2$, the condition that $P\rho(\gamma)$ is a parabolic or identity element, holds only if $\gamma = \gamma_0$. So, these elements contribute only when $\mathfrak{n}$ is a square, in which case their contribution is $k_T(1)\mu(\Gamma \setminus \mathfrak{H})$.

Every element $\beta \in \mathrm{PSL}_2\mathbb{C}$ other than identity or parabolic has a pair of fixed points on the boundary $\hat{\mathbb{C}}$ of $\mathfrak{H}$, and the unique geodesic joining these two points is called the axis of β, A_β. Such a β is conjugate to a matrix of the form $h_{t,\varphi} = \begin{pmatrix} te^{i\varphi} & \\ & 1/(te^{i\varphi}) \end{pmatrix}$ and geometrically acts as a rotation around A_β with angle 2φ followed by a hyperbolic translation along the same axis by distance $2\log t$. Only those β that have $t = 1$ have fixed points in $\mathfrak{H}$: these are the elliptic elements and are distinguished by the condition $\mathrm{Tr}\,\beta \in \mathbb{R}$, $|\mathrm{Tr}\,\beta| < 2$. The remaining elements are loxodromic, with pure translations ($\varphi \in \pi\mathbb{Z}$, i.e. $\mathrm{Tr}\,\beta \in \mathbb{R}$, $|\mathrm{Tr}\,\beta| > 2$) usually termed hyperbolic. In any case it follows from $u(h_{t,\varphi}v, v) = \dfrac{|t^2e^{2i\varphi} - 1|^2|z|^2 + (1+t^4)r^2}{2t^2r^2}$ that

$$\inf_{v\in\mathfrak{H}} u(\beta v, v) = 1 + \frac{(t^2-1)^2}{2t^2}.$$

We now estimate the contribution in (3.4) of $\gamma \in \mathcal{O}(\mathfrak{n})$ for which $\beta = P\rho(\gamma)$ is neither identity nor parabolic. In particular, if the contribution of some γ is to be nonzero, the parameter t corresponding to $\beta = P\rho(\gamma)$ must satisfy $t = 1 + \mathrm{O}(T^{-1})$. Further, as $\mathrm{Tr}\, P\rho(\gamma) = te^{i\varphi} + (1/t)e^{-i\varphi} = 2\cos\varphi + \mathrm{O}(T^{-1})$, we have that $\rho(N\gamma_0)/\rho(\eta) = \cos^2\varphi + \mathrm{O}(T^{-1})$. In fact, as

$$\frac{|\rho(N(\gamma-\gamma_0))|^2}{|\rho(\eta)|^2}\mathfrak{N}\mathfrak{n} \geqslant |\rho(N(\gamma-\gamma_0))|^2 \prod_i \rho_i\eta \geqslant |\mathfrak{N}N(\gamma-\gamma_0)| \gg 1,$$

we have that

$$|2\sin\varphi + \mathrm{O}(T^{-1})|^2 = \left|4 - \left(te^{i\varphi} + \frac{1}{t}e^{-i\varphi}\right)^2\right| = \frac{4|\rho(N(\gamma-\gamma_0))|}{|\rho(\eta)|} \gg \frac{1}{\sqrt{\mathfrak{N}\mathfrak{n}}}.$$

Recall that $\mathfrak{N}\mathfrak{n} \leqslant M^2 \ll T^4$ in our ranges for application in (3.4), so that, by assuming that $M \ll T^2$ with a sufficiently small implied constant, we can ensure that $\sin\varphi > cT^{-1}$ for a suitably large $c > 0$. In particular, if $u(\rho(\gamma)v, v) - 1 \ll T^{-2}$ for some v in a fixed, compact fundamental domain for Γ, the axis $A_{\rho(\gamma)}$ will actually intersect some slightly larger fixed compact domain $\mathcal{F}$. The fixed points of $\rho(\gamma)$ on $\hat{\mathbb{C}}$ must also lie in some fixed domain $\mathcal{F}'$.

For each such γ, fix the positive integer $c^2 < r_\gamma \leqslant T^2$ such that

$$\frac{r_\gamma - 1}{T^2} < \sin^2\varphi \leqslant \frac{r_\gamma}{T^2},$$

and note that $|\rho(N(\gamma-\gamma_0))| \asymp |\rho(\eta)|r_\gamma T^{-2}$ as well. Consider the quadratic extension $M = L(\omega) < A$, all of whose embeddings are complex, and write $\gamma = \gamma_M + \gamma_M'\omega'$ with $\gamma_M = \gamma_0 + \gamma_1\omega$,

$\gamma_M' = \gamma_2 + \gamma_3\omega \in M$. For a fixed $r = r_\gamma$, let x_r, y_r, z_r and I_r be, respectively, the number of possible choices for γ_0, the number of choices for γ_1, the maximum possible number of choices for γ_2 and γ_3 once γ_0 and γ_1 have been chosen, and the maximum possible contribution of such a γ to (3.4).

In light of $\rho(N\gamma_0) = \rho(\eta)\cos^2\varphi + \mathrm{O}(|\rho(\eta)|T^{-1})$ and $\rho_i(N\gamma_0) \ll \rho_i(\eta)$, we have that

$$x_r \ll \left(\frac{\sqrt{|\rho(\eta)|}}{\max(T\sqrt{1-r^2/T^2}, T^{1/2})} + 1\right)^2 \ll \frac{|\rho(\eta)|}{T} + 1.$$

From the quadratic formula, the fixed points of $\rho(\gamma)$ on $\hat{\mathbb{C}}$ are

$$z_{1,2} = \frac{1}{\rho(\bar{\gamma}_M'\omega')}\left(\rho(\gamma_1\omega) \pm \sqrt{-\rho(N(\gamma-\gamma_0))}\right);$$

the condition that these lie in $\mathcal{F}'$ translates into

$$|\rho(\bar{\gamma}_M')| \ll \frac{\sqrt{|\rho(\eta)|}}{T}\sqrt{r}, \quad |\rho(\gamma_1)| \ll \frac{\sqrt{|\rho(\eta)|}}{T}\sqrt{r},$$
$$|\rho(N\gamma_M')| \ll \frac{|\rho(\eta)|}{T^2}r, \quad |\rho(\gamma_M')| \ll \frac{\sqrt{|\rho(\eta)|}}{T}\sqrt{r}.$$

(Note that all of these also hold if one of the fixed points on $\hat{\mathbb{C}}$ is ∞.) The second estimate above coupled with $|\rho_i(\gamma_1)| \ll \sqrt{\rho_i(\eta)}$ shows that

$$\sum_{s\leqslant r} y_s \ll \left(\frac{\sqrt{\mathfrak{N}\mathfrak{n}}}{T^2} + 1\right) r.$$

Finally, once γ_0 and γ_1 are fixed, γ is subject to the quadratic condition $N\gamma_M' = x := (N\gamma_M - \eta)/\omega'^2$, which, up to multiplication by units, is a question of representing a fixed ideal (x) of L as a norm of a principal ideal (γ_M') with generator in a fixed order. Taking into account the estimates on $|\rho(\gamma_M')|$, $|\rho(\bar{\gamma}_M')|$ above as well as $|\rho_i(\gamma_M')|, |\rho_i(\bar{\gamma}_M')| \ll \sqrt{\rho_i(\eta)}$, and noting that units $\epsilon \in \mathfrak{O}_M^*$ such that $N_{M|L}\epsilon = 1$ form a free subgroup of rank one, we see that this can be done in no more than $\ll d_L(x)\log T$ ways, so that in any case

$$z_r \ll e^{C\frac{\log T}{\log\log T}}.$$

Finally, the contribution to (3.4) of each γ is an integral along a compact interval on $A_{\rho(\gamma)}$ (of length uniformly bounded by $\operatorname{diam}\mathcal{F} + 1$) of surface integrals, each of which is at most (recalling that $P\rho(\gamma)$ is conjugate to $h_{t,\varphi}$)

$$\int_{\mathbb{C}} k_T\left(1 + \frac{|t^2e^{2i\varphi}-1|^2|z|^2}{2y^2}\right)\frac{|\mathrm{d}z|}{y^2} \leqslant \frac{2\pi}{|t^2e^{2i\varphi}-1|^2}\int_1^\infty k_T(w)\,\mathrm{d}w$$
$$= \frac{g_T(0)}{|t^2e^{2i\varphi}-1|^2} \ll \frac{T^3}{r},$$

so that $I_r \ll T^3/r$.

Summing up, the total contribution from elements $\gamma \in Z(\mathfrak{O}) \setminus \mathcal{O}(\mathfrak{n})$ such that $P\rho(\gamma)$ is not identity or parabolic is hence (e.g. by splitting into dyadic intervals)

$$\sum_{r=\lfloor c^2 \rfloor+1}^{\lceil T/c \rceil} x_r y_r z_r I_r \ll \left(\frac{\sqrt{\mathfrak{N}\mathfrak{n}}}{T} + 1 \right)^2 T^2 e^{C \frac{\log T}{\log\log T}} \log T \ll (\mathfrak{N}\mathfrak{n} + T^2) e^{2C \frac{\log T}{\log\log T}}.$$

This proves that

$$\begin{aligned} Q(a) =& k_T(1)\mu(\Gamma \setminus \mathfrak{H}) \sum_{\substack{\mathfrak{N}\mathfrak{m},\, \mathfrak{N}\mathfrak{n} \leqslant M \\ \mathfrak{m}\mathfrak{n} \text{ square}}} a(\mathfrak{m})a(\mathfrak{n}) \sum_{\mathfrak{d} | (\mathfrak{m},\mathfrak{n})} \mathfrak{N}\mathfrak{d} \\ &+ \mathrm{O}\Bigg(\sum_{\mathfrak{N}\mathfrak{d} \leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\mathfrak{N}\mathfrak{m}, \mathfrak{N}\mathfrak{n} \leqslant M/\mathfrak{N}\mathfrak{d}} |a(\mathfrak{d}\mathfrak{m})a(\mathfrak{d}\mathfrak{n})| \ell_T(\mathfrak{m}\mathfrak{n}) \cdot e^{2C \frac{\log T}{\log\log T}} \Bigg) \\ =& k_T(1)\mu(\Gamma \setminus \mathfrak{H}) \sum_{\mathfrak{N}\mathfrak{d} \leqslant M} \Bigg(\sum_{\mathfrak{d}=\mathfrak{e}\mathfrak{f}} \mathfrak{N}\mathfrak{e} \mu_L^2(\mathfrak{f}) \Bigg) \Bigg(\sum_{\mathfrak{N}\mathfrak{n} \leqslant \sqrt{M/\mathfrak{N}\mathfrak{d}}} a(\mathfrak{d}\mathfrak{n}^2) \Bigg)^2 \\ &+ \mathrm{O}\Bigg(\sum_{\mathfrak{N}\mathfrak{d} \leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\mathfrak{N}\mathfrak{m}, \mathfrak{N}\mathfrak{n} \leqslant M/\mathfrak{N}\mathfrak{d}} |a(\mathfrak{d}\mathfrak{m})a(\mathfrak{d}\mathfrak{n})| \ell_T(\mathfrak{m}\mathfrak{n}) \cdot e^{2C \frac{\log T}{\log\log T}} \Bigg), \end{aligned} \tag{3.5}$$

where $\ell_T(\mathfrak{n}) = \mathfrak{N}\mathfrak{n} + T^2$ as in the statement of Lemma 4.

4. Non-compact case and the Eisenstein series contribution.

In this section, we deal with the additional terms that occur in the case when Γ is the Picard group $\mathrm{PSL}_2\mathfrak{O}$, with $\mathfrak{O}$ the maximal order of the imaginary quadratic field $L = \mathbb{Q}(\sqrt{-d})$. We first turn our attention to the contribution of Eisenstein series to (1.5) for the QCM-point v of Theorem 1. Under the technical assumptions made in Theorem 1, there is only one cusp at ∞ and the corresponding Eisenstein series is defined for $\mathfrak{Re}\, s > 1$ as

$$E(v,s) = \sum_{(c,d)=\mathfrak{O}} \left(\frac{r}{|cz+d|^2 + |c|^2 r^2} \right)^{1+s}. \tag{4.1}$$

Here, $\tilde{q}^H(c,d) = |cz+d|^2 + |c|^2 r^2$ is a binary Hermitian form over L. In fact, it is not difficult to check that, as a definite quaternary quadratic form over $\mathbb{Q}$, it is equivalent to the form q considered in section 2. We can write $\tilde{q}^H(c,d) = q_0 q^H(c,d)$ for some $q_0 \in L^\times$ and a primitive binary Hermitian form q^H on $\mathfrak{O}$, so that

$$E(v,s) = \left(\frac{r}{q_0^2} \right)^{1+s} \frac{Z(q^H, 1+s)}{\zeta_L(1+s)}$$

with $Z(q^H,s) = \sum_{n=1}^\infty r(q^H,n)/n^s$ and $r(q^H,n) = \#\{(c,d) \in \mathfrak{O}^2 :\ n = q^H(c,d)\}$.

While this is not strictly necessary for our purposes, we now refer to the work of Elstrodt, Grünewald, and Mennicke [El-Gr-Me2], in which the local densities for primitive binary Hermitian forms over imaginary quadratic fields are explicitly computed. (We note that, to our knowledge, this result is not available over a quadratic extension of an arbitrary totally real ground field, which is why we could not appeal to it in section 2.) To state their result, let $q_1^H = q^H, q_2^H, \dots, q_\ell^H$ be representatives of $\mathrm{GL}_2\mathfrak{O}$-equivalence classes of Herimitian forms in the genus of q^H, let $E(q_i^H)$ denote the group of $\mathrm{GL}_2\mathfrak{O}$-units of q_i^H, and let

$$r(\operatorname{gen} q^H, n) = \left(\sum_{i=1}^{\ell} \frac{1}{|E(q_i^H)|}\right)^{-1} \sum_{i=1}^{\ell} \frac{r(q_i^H, n)}{|E(q_i^H)|}.$$

Denote correspondingly $Z(\operatorname{gen} q^H, s) = \sum_{n=1}^{\infty} r(\operatorname{gen} q^H, n)/n^s$, as well as $r_{\mathrm{cusp}}(q^H, n) = r(q^H, n) - r(\operatorname{gen} q^H, n)$, $Z_{\mathrm{cusp}}(q^H, s) = \sum_{n=1}^{\infty} r_{\mathrm{cusp}}(q^H, n)/n^s$. Then [El-Gr-Me2] explicitly determine a finite collection X of quadratic Dirichlet characters χ to conductors dividing $(\operatorname{disc} L, \operatorname{disc} q^H)$, constants c_χ, and finite Euler products $P_\chi(s)$ of polynomials in p^{-s}, such that

$$Z(\operatorname{gen} q^H, s+1) = \sum_{\chi \in X} c_\chi P_\chi(s) L(s, \chi) L(s+1, \chi).$$

As all forms q_i^H are everywhere locally equivalent also as quaternary quadratic forms over K, we have, as in section 2, that their values at cusps are all equal, and so $r_{\mathrm{cusp}}(q^H, n)$ are Fourier coefficients of a cusp Hilbert modular form f for a certain congruence subgroup of $\mathrm{PSL}_2\mathbb{Z}$; $Z_{\mathrm{cusp}}(q^H, s)$ is the L-function of this cusp form.

The total contribution of Eisenstein series to the spectral average in (3.1), after expanding the square and applying (1.5), is

$$Q_1^\infty(a) = \frac{r^2}{q_0^4} \sum_{\mathfrak{N}\mathfrak{m}, \mathfrak{N}\mathfrak{n} \leqslant M} a(\mathfrak{m}) a(\mathfrak{n}) \sqrt{\mathfrak{N}\mathfrak{m}\mathfrak{N}\mathfrak{n}} \sum_{\mathfrak{u} | (\mathfrak{m}, \mathfrak{n})} \sum_{\mathfrak{m}\mathfrak{n}/\mathfrak{u}^2 = \mathfrak{a}\mathfrak{d}}$$
$$\int_{-\infty}^{\infty} \frac{h_T(r)}{|\zeta_L(1+ir)|^2} \left| \sum_{\chi \in X} c_\chi P_\chi(ir) L(ir, \chi) L(1+ir, \chi) + L(1+ir, f) \right|^2 \left(\frac{\mathfrak{N}\mathfrak{a}}{\mathfrak{N}\mathfrak{d}}\right)^{ir/2} \mathrm{d}r.$$

Estimating all integrals trivially, using convexity bounds at the edge of the critical strip for the L-functions as well as the lower bound $|\zeta_L(1+ir)| \gg (1+|r|)^{-\epsilon}$, we find that

$$Q_1^\infty(a) \ll T^{2+\epsilon} \sum_{\mathfrak{N}\mathfrak{d} \leqslant M} \mathfrak{N}\mathfrak{d} \left(\sum_{\mathfrak{N}\mathfrak{m} \leqslant M/\mathfrak{N}\mathfrak{d}} a(\mathfrak{d}\mathfrak{m}) \sqrt{\mathfrak{N}\mathfrak{m}} \right)^2. \tag{4.2}$$

It is quite possible that this bound can be substantially improved by exercising some care in the above estimation. In [Mi], such improvement was achieved by using Gallagher's form of large sieve inequality. However, the above suffices for our purposes.

We next account for the additional terms which, in the non-compact case, occur in the asymptotic analysis of (1.5) after integration over v in a fixed fundamental domain $\mathcal{F}_0$ for $\Gamma \backslash \mathfrak{H}$. This proceeds analogously to the proof of the trace formula for Picard groups [Tan]. We refer to [El-Gr-Me1] for specific evaluations needed in our treatment. Integrals of both sides are seen to be actually mildly divergent in the cusp; therefore, we first integrate over $\mathcal{F}_Y = \{v = (z, r) \in \mathcal{F}_0 \colon\ r \leqslant Y\}$ and then let $Y \to \infty$. On the spectral side, the integral present is the same as in the trace formula with $h_T(r)(\mathfrak{N}\mathfrak{a}/\mathfrak{N}\mathfrak{d})^{ir/2}$ in place of $h(r)$ and evaluates as

$$\begin{aligned} c\int_{\mathcal{F}_Y} \sqrt{\mathfrak{N}\mathfrak{n}} \sum_{\mathfrak{n}=\mathfrak{a}\mathfrak{d}} \int_{-\infty}^{\infty} h_T(r) \left(\frac{\mathfrak{N}\mathfrak{a}}{\mathfrak{N}\mathfrak{d}}\right)^{ir/2} |E(v,ir)|^2 \,\mathrm{d}r \\ = \sqrt{\mathfrak{N}\mathfrak{n}} \sum_{\mathfrak{n}=\mathfrak{a}\mathfrak{d}} \Bigg[g_T\left(-\frac{1}{2}\log\frac{\mathfrak{N}\mathfrak{a}}{\mathfrak{N}\mathfrak{d}}\right)\log Y + \frac{\phi(0)h_T(0)}{4} \\ - \frac{1}{4\pi}\int_{-\infty}^{\infty} h_T(r)\frac{\phi'}{\phi}(ir)\left(\frac{\mathfrak{N}\mathfrak{a}}{\mathfrak{N}\mathfrak{d}}\right)^{ir/2} \mathrm{d}r\Bigg] + o(1), \quad (Y\to\infty), \end{aligned} \tag{4.3}$$

where, for L as in Theorem 1, $\phi(s) = \dfrac{2\pi}{\sqrt{|d_L|}} \dfrac{\zeta_L(s)}{\zeta_L(1+s)}$, $\phi(0) = -1$.

On the geometric side, we estimate the contribution of a $\gamma \in Z(\mathfrak{O}) \setminus \mathcal{O}(\mathfrak{n})$ after integration over $\mathcal{F}_Y$ separately according to the geometric type of $P\rho(\gamma)$. The analysis of $P\rho(\gamma)$ which are neither parabolic nor cusp-elliptic or cusp-loxodromic fixing ∞, runs verbatim as in the compact case, with the enlarged domain $\mathcal{F}$ being replaced by a domain of the shape $\tilde{\mathcal{F}} = \mathcal{F} \cup \{v = (z,r)\colon |r| \geqslant c|z|\}$ for some suitable $c > 0$. One sees that the, if $u(\rho(\gamma)v,v) - 1 \ll T^{-2}$ for some $v \in \mathcal{F}_0$, the axis $A_{\rho(\gamma)}$ must actually intersect $\tilde{\mathcal{F}}$, and this suffices to arrive at the same bounds for $|\rho(\bar{\gamma}'_M)|$, $\rho(\gamma_1)|$, and, consequently, x_r, y_r and z_r. The estimate on I_r is the same except for the length of the interval along $A_{\rho(\gamma)}$ whose length is $\ll \log T$, so that the total contribution of these elements is still included in $Q(a)$.

We next cosider the contribution of non-identity parabolic elements, which are present only if $\mathfrak{n}$ is a square. We first claim that parabolic elements $P\rho(\gamma) = \left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right)$ fixing a $z_0 \in \hat{\mathbb{C}}$ other than ∞ do not contribute in our ranges. Indeed, it is easily verified that, under conjugation by $\gamma_0 = \left(\begin{smallmatrix} 0 & 1 \\ -1 & z_0\end{smallmatrix}\right)$, $\gamma_0\rho(\gamma)\gamma_0^{-1} = \left(\begin{smallmatrix} (a+d)/2 & -c \\ 0 & (a+d)/2\end{smallmatrix}\right)$, while the domain $\gamma_0\mathcal{F}_0$ lies strictly below the plane $r = r_0$ (where r_0 may be taken to be the reciprocal of the smallest r-coordinate of all points in $\mathcal{F}_0$). As $(a+d)/2 = \operatorname{Tr} P\rho(\gamma)/2 = \pm\sqrt{\rho(\eta)}$, this shows that $k_T(\rho(\gamma)v,v) = 0$ for all $v \in \mathcal{F}_0$ if $M \ll T^2$ with a small enough implied constant. Moving on to parabolic elements fixing ∞, we find that their contribution is

$$\begin{aligned} \int_{\mathcal{F}_Y} \sideset{}{^*}\sum_{\mu\in\mathfrak{O}} k_T\left(\frac{|\mu|^2}{2|d|^2r^2}+1\right)\frac{|\mathrm{d}z|\,\mathrm{d}r}{r^3} &= \frac{\sqrt{\operatorname{disc} L}}{|\mathfrak{O}^*|} \sideset{}{^*}\sum_{\mu\in\mathfrak{O}} \int_0^Y k_T\left(\frac{|\mu|^2}{2|d|^2r^2}+1\right)\frac{\mathrm{d}r}{r^3} \\ &= \sqrt{\operatorname{disc} L}\,\mathfrak{N}\mathfrak{d} \sum_{\mathfrak{m}\subset\mathfrak{O}} \frac{1}{\mathfrak{N}\mathfrak{m}} \int_{\frac{\mathfrak{N}\mathfrak{m}}{2\mathfrak{N}\mathfrak{d}Y^2}}^{\infty} k_T(u+1)\,\mathrm{d}u. \end{aligned}$$

Here, $d \in \mathfrak{O}$ is such that $\mathfrak{d} = (d)$ satisfies $\mathfrak{d}^2 = \mathfrak{n}$, and the first equality is justified because we can arrange, by assuming $M \ll T^2$ with a sufficiently small implied constant, that the integrand is zero unless r is greater than a large positive number of our choosing. By contour shifting and the class number formula, we have that, uniformly,

$$\sum_{\mathfrak{m}\subset\mathfrak{O},\,\mathfrak{N}\mathfrak{m}\leqslant x} \frac{1}{\mathfrak{N}\mathfrak{m}} = \frac{2\pi}{|\mathfrak{O}^*|\sqrt{\operatorname{disc} L}}\left(\log x + C_L + \Phi(x)\right),$$

where $C_L = \gamma + (L'/L)(1, \chi_L)$, γ is the Euler constant, χ_L is the quadratic character associated to the complex conjugation in L by the class field theory, and $\Phi(x) = \mathrm{O}(\min(1/x, |\log x|))$. Hence the above contribution is

$$\begin{aligned}\frac{2\pi}{|\mathfrak{O}^*|}\sqrt{\mathfrak{N}\mathfrak{n}}\int_0^\infty & k_T(u+1)\left(\log(2\sqrt{\mathfrak{N}\mathfrak{n}}Y^2u)+C_L+\Phi\left(2\sqrt{\mathfrak{N}\mathfrak{n}}Y^2u\right)\right)\mathrm{d}u\\ &=\frac{\sqrt{\mathfrak{N}\mathfrak{n}}}{|\mathfrak{O}^*|}\bigg[\left(2\log Y+\log(\sqrt{\mathfrak{N}\mathfrak{n}})+C_L-2\gamma\right)g_T(0)\\ &\qquad+\frac{1}{2}h_T(0)-\frac{1}{\pi}\int_{-\infty}^{\infty}h_T(r)\frac{\Gamma'}{\Gamma}(1-ir)\,\mathrm{d}r\bigg]+\mathrm{O}_{T,\mathfrak{n}}\left(\frac{\log Y}{Y^2}\right),\end{aligned}\tag{4.4}$$

by using the expicit evaluation of $\int_0^\infty k_T(u+1)\log u\,\mathrm{d}u$ in [El-Gr-Me1], section 6.5.

We now pass to the contribution of the cusp-elliptic and cusp-loxodromic elements fixing ∞. Note that the contribution of the integral over a compact part of $\mathcal{F}_0$ for these elements is already included in the estimation of $Q(a)$, so that we only need to consider integrals over domain of the shape $\Lambda \times [r_0, +\infty)$, where $\Lambda \subset \mathbb{C}$ is a fundamental domain under the action of the stabilizer of ∞ in $\mathrm{PSL}_2\mathfrak{O}$. The elements under consideration split into classes of the form $\left\{\left(\begin{smallmatrix} a\epsilon & b\epsilon^{-1}\\ 0 & d\epsilon^{-1}\end{smallmatrix}\right) : b \in \mathfrak{O},\ \epsilon \in \mathfrak{O}^*/\{\pm 1\}\right\}$, where a representative η of $(\eta) = \mathfrak{n}$ is picked and, for every decomposition $\mathfrak{n} = \mathfrak{a}\mathfrak{d}$, a and d with $\mathfrak{a} = (a)$, $\mathfrak{d} = (d)$ are picked arbitrarily subject only to the condition that $a = d$ if $\mathfrak{n}$ is a square. In cusp-loxodromic classes, elements in one class contribute

$$\begin{aligned}\int_{r_0}^{Y}\sum_{\epsilon\in\mathfrak{O}^*/\{\pm1\}}\sum_{b\in\mathfrak{O}}\int_\Lambda & k_T\left(\frac{|(a-d)\epsilon^2z+b|^2+(|a|^2+|d|^2)r^2}{2|ad|r^2}\right)|\mathrm{d}z|\,\frac{\mathrm{d}r}{r^3}\\ &=\int_{r_0}^{Y}\mathfrak{N}(a-d)\int_{\mathbb{C}}k_T\left(\frac{|(a-d)z|^2+(|a|^2+|d|^2)r^2}{2|ad|r^2}\right)|\mathrm{d}z|\,\frac{\mathrm{d}r}{r^3}\\ &=\sqrt{\mathfrak{N}\mathfrak{n}}\,g_T\left(\frac{1}{2}\log\frac{\mathfrak{N}\mathfrak{a}}{\mathfrak{N}\mathfrak{d}}\right)(\log Y-\log r_0).\end{aligned}\tag{4.5}$$

Finally, we consider classes of cusp-elliptic elements $\left(\begin{smallmatrix} d\epsilon & b\epsilon^{-1}\\ 0 & d\epsilon^{-1}\end{smallmatrix}\right)$, $\epsilon \in \mathfrak{O}^*/\pm1$, $\epsilon \neq \pm 1$, which are present only when $L = \mathbb{Q}(\sqrt{-1})$ or $L = \mathbb{Q}(\sqrt{-3})$ and $\mathfrak{n}$ is square. Their contribution is

$$\begin{aligned}\int_{r_0}^{Y}\sum_{\substack{\epsilon\in\mathfrak{O}^*/\{\pm1\}\\ \epsilon\neq\pm1}}\sum_{b\in\mathfrak{O}}\int_\Lambda & k_T\left(\frac{|d(\epsilon^2-1)z+b|^2}{2|d|^2r^2}+1\right)|\mathrm{d}z|\,\frac{\mathrm{d}r}{r^3}\\ &=\int_{r_0}^{\infty}\left(1-\frac{2}{|\mathfrak{O}^*|}\right)\mathfrak{N}\mathfrak{d}\int_{\mathbb{C}}k_T\left(\frac{|z|^2}{2r^2}+1\right)|\mathrm{d}z|\,\frac{\mathrm{d}r}{r^3}\\ &=\sqrt{\mathfrak{N}\mathfrak{n}}\left(1-\frac{2}{|\mathfrak{O}^*|}\right)g_T(0)(\log Y-\log r_0),\end{aligned}\tag{4.6}$$

where the second line is justified simply by comparing a fundamental domain for $\mathbb{C}/\mathfrak{O}$ with $(\epsilon^2-1)\Lambda$ in each of the two cases for L separately.

Comparing the total contribution of $P\rho(\gamma)$ of all geometric types in (4.4)–(4.6) with that present on the spectral side in (4.3), we see that, as $Y \to \infty$, the leading terms of order $\log Y$ cancel out and one obtains a finite limit. Combining these contributions for various $\mathfrak{m}$ and $\mathfrak{n}$, we see that the additional contribution to the spectral average in (3.2) present in the case $\Gamma = \mathrm{PSL}_2\mathfrak{O}$ is given by

$$\begin{aligned} Q^{\infty}(a) = \sum_{\mathfrak{N}\mathfrak{m},\mathfrak{N}\mathfrak{n}\leqslant M} & a(\mathfrak{m})a(\mathfrak{n})\sqrt{\mathfrak{N}\mathfrak{m}\mathfrak{N}\mathfrak{n}} \sum_{\mathfrak{u}\mid(\mathfrak{m},\mathfrak{n})} \Bigg[\frac{d(\mathfrak{m}\mathfrak{n}/\mathfrak{u}^2)}{4}h(0) \\ & + \frac{1}{4\pi}\sum_{\mathfrak{m}\mathfrak{n}/\mathfrak{u}^2=\mathfrak{a}\mathfrak{d}} \int_{-\infty}^{\infty} h_T(r)\frac{\phi'}{\phi}(ir)\left(\frac{\mathfrak{N}\mathfrak{a}}{\mathfrak{N}\mathfrak{d}}\right)^{ir/2}\mathrm{d}r - \log r_0 \sum_{\mathfrak{m}n/\mathfrak{u}^2=\mathfrak{a}\mathfrak{d}} g_T\left(\frac{1}{2}\log\frac{\mathfrak{N}\mathfrak{a}}{\mathfrak{N}\mathfrak{d}}\right)\Bigg] \\ + \sum_{\substack{\mathfrak{N}\mathfrak{m},\mathfrak{N}\mathfrak{n}\leqslant M\\ \mathfrak{m}\mathfrak{n}\,\text{square}}} & a(\mathfrak{m})a(\mathfrak{n})\sqrt{\mathfrak{N}\mathfrak{m}\mathfrak{N}\mathfrak{n}}\frac{d((\mathfrak{m},\mathfrak{n}))}{|\mathfrak{O}^*|}\Bigg[\frac{1}{2}h(0) - \frac{1}{\pi}\int_{-\infty}^{\infty} h_T(r)\frac{\Gamma'}{\Gamma}(1-ir)\,\mathrm{d}r \\ & + \left(\log\sqrt{\mathfrak{N}[\mathfrak{m},\mathfrak{n}]} + C_L - 2\gamma + (2-|\mathfrak{O}^*|)\log r_0\right)g_T(0)\Bigg]. \end{aligned}$$

Estimating all terms trivially, we conclude that

$$Q^{\infty}(a) \ll T^{1+\epsilon}\left(\sum_{\mathfrak{N}\mathfrak{m}\leqslant M} a(\mathfrak{m})\sqrt{\mathfrak{N}\mathfrak{m}}\right)^2. \tag{4.7}$$

Collecting the estimates (3.3), (4.2), (3.5), and (4.7), we get the full statement of Lemma 4. □

5. Optimal resonators and conclusion.

In this section, we optimize the resonator sequence $a(\mathfrak{n})$ to make the quotient $Q_1(a)/Q(a)$ as large as possible and conclude the proof of Theorem 1. We introduce one final substitution $b(\mathfrak{n}) = \sum_{\mathfrak{N}\mathfrak{m}\leqslant\sqrt{M/\mathfrak{N}\mathfrak{n}}} a(\mathfrak{n}\mathfrak{m}^2)$, which inverts as $a(\mathfrak{n}) = \sum_{\mathfrak{N}\mathfrak{m}\leqslant\sqrt{M/\mathfrak{N}\mathfrak{n}}} \mu_L(\mathfrak{m})b(\mathfrak{n}\mathfrak{m}^2)$. The coprimality condition $b(\mathfrak{n}) = 0$ unless $(\mathfrak{n}, \mathfrak{q}_{\mathcal{O}}\mathfrak{q}_L\mathfrak{q}_q) = \mathfrak{O}$ is equivalent to the same condition for $a(\mathfrak{n})$. We can rewrite the conclusions of Lemma 4 in terms of $b(\mathfrak{n})$ as

$$\begin{aligned} Q_1(a) &\gg k_T(1)B_1(b) + \mathrm{O}(k_T(1)R_{\text{cusp}}(b)) + \mathrm{O}(\delta_\Gamma Q_1^{\infty}(a)) \\ Q(a) &= k_T(1)\mu(\Gamma\setminus\mathfrak{H})B(b) + \mathrm{O}\left(R_T(b)e^{2C\frac{\log T}{\log\log T}}\right) + \mathrm{O}(\delta_\Gamma Q^{\infty}(a)), \end{aligned} \tag{5.1}$$

where B, B_1, R_T and R_{cusp} are quadratic forms in $b(\mathfrak{n})$ given by

$$\begin{aligned} B(b) &= \sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \sigma_{\mu^2}(\mathfrak{d})b(\mathfrak{d})^2 \\ B_1(b) &= \sum_{\chi\in\Xi}\sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\mathfrak{N}\mathfrak{m},\mathfrak{N}\mathfrak{n}\leqslant M/\mathfrak{N}\mathfrak{d}} b(\mathfrak{d}\mathfrak{m})b(\mathfrak{d}\mathfrak{n})\sum_{\mathfrak{u}^2\mid\mathfrak{m}}\sum_{\mathfrak{v}^2\mid\mathfrak{n}} \mu_L(\mathfrak{u})\mu_L(\mathfrak{v})\sigma_{K,\mu^2,\chi}\left(\frac{\mathfrak{m}\mathfrak{n}}{\mathfrak{u}^2\mathfrak{v}^2}\right) \\ R_T(b) &= \sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\mathfrak{N}\mathfrak{m},\mathfrak{N}\mathfrak{n}\leqslant M/\mathfrak{N}\mathfrak{d}} |b(\mathfrak{d}\mathfrak{m})b(\mathfrak{d}\mathfrak{n})| \sum_{\mathfrak{u}^2\mid\mathfrak{m}}\sum_{\mathfrak{v}^2\mid\mathfrak{n}} |\mu_L(\mathfrak{u})\mu_L(\mathfrak{v})|\ell_T\left(\frac{\mathfrak{m}\mathfrak{n}}{\mathfrak{u}^2\mathfrak{v}^2}\right) \\ R_{\text{cusp}}(b) &= \sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\mathfrak{N}\mathfrak{m},\mathfrak{N}\mathfrak{n}\leqslant M/\mathfrak{N}\mathfrak{d}} |b(\mathfrak{d}\mathfrak{m})b(\mathfrak{d}\mathfrak{n})| \sum_{\mathfrak{u}^2\mid\mathfrak{m}}\sum_{\mathfrak{v}^2\mid\mathfrak{n}} |\mu_L(\mathfrak{u})\mu_L(\mathfrak{v})|\left|r_{\text{cusp}}^K\left(\frac{\mathfrak{m}\mathfrak{n}}{\mathfrak{u}^2\mathfrak{v}^2}\right)\right|. \end{aligned}$$

The following lemma shows that the problem of maximizing $B_1(b)/B(b)$ (and hence $Q_1(a)/Q(a)$ after we choose $M \ll T^2$ small enough to keep the remainders in check) is substantially different from the corresponding problem for surfaces [Mi]. In place of analytical subtlety required of optimal resonators, one encounters interesting combinatorics of the contribution from $\sigma_{K,\mu^2,\chi}$.

Lemma 5. *For large M, the maximum value attained by $B_1(b)/B(b)$ is $\asymp M^{1/2}$.*

Proof of Lemma 2. Consider the auxiliary totally multiplicative function g defined by $g(\mathfrak{p}) = 1/(\sqrt{\mathfrak{N}\mathfrak{p}}-1)$, and let $\sigma_*(\mathfrak{n}) = \sum_{\mathfrak{d}|\mathfrak{n}} g(\mathfrak{d})$. These functions are constructed so that $1+\sigma_*(\mathfrak{p})/\sqrt{\mathfrak{N}\mathfrak{p}} = \sigma_*(\mathfrak{p})$, and in fact it is easily checked that σ_* satisfies

$$\sum_{\mathfrak{d}|\mathfrak{n}} \frac{\sigma_*(\mathfrak{d})}{\sqrt{\mathfrak{N}\mathfrak{d}}} \asymp \sigma_*(\mathfrak{n}).$$

Fix a $1<\beta<3/2$. We can write each $\mathfrak{n} \subset \mathfrak{O}$ as $\mathfrak{n} = \mathfrak{n}_0\mathfrak{n}_1$ with $\mathfrak{n}_0 \subset \mathfrak{o}$ and $\mathfrak{n}_1$ minimal. With this, let us further denote $\sigma^+(\mathfrak{n}) = \sum_{\mathfrak{d}|(\mathfrak{n}_1\cap\mathfrak{o}),\mathfrak{d}\subset\mathfrak{o}} \sigma_{\mu^2}(\mathfrak{d})\mathfrak{N}\mathfrak{d}^{-\beta/2}$ and $\sigma^*(\mathfrak{n}) = \sigma^+(\mathfrak{n})\sigma_*(\mathfrak{n})$. We also note for future reference that

$$\sum_{\mathfrak{N}\mathfrak{n}\leqslant N,\, \mathfrak{n}\subset\mathfrak{o}} \sigma_{\mu^2}^2(\mathfrak{n})\sigma_*(\mathfrak{n})\mathfrak{N}\mathfrak{n}^{-\beta} \ll N^{3/2-\beta},$$

because $\sum_{\mathfrak{n}\subset\mathfrak{o}} \sigma_{\mu^2}^2(\mathfrak{n})\sigma_*(\mathfrak{n})\mathfrak{N}\mathfrak{n}^{-s} = \zeta_K(2s-2)G(s)$ with a certain Dirichlet series $G(s)$ absolutely convergent in $\sigma>1$.

As for

$$\Sigma_\chi(\mathfrak{m},\mathfrak{n}) = \sum_{\mathfrak{u}^2|\mathfrak{m}}\sum_{\mathfrak{v}^2|\mathfrak{n}} \mu_L(\mathfrak{u})\mu_L(\mathfrak{v})\sigma_{K,\mu^2,\chi}(\mathfrak{m}\mathfrak{n}/\mathfrak{u}^2\mathfrak{v}^2),$$

we note that $\Sigma_\chi(\mathfrak{m},\mathfrak{n}) = 0$ unless $\mathfrak{m}\mathfrak{n} = \mathfrak{s}_1^2\mathfrak{s}_0$ for some $\mathfrak{s}_0 \subset \mathfrak{o}$, in which case we assume $\mathfrak{s}_1$ to be minimal as usual and write $\mathfrak{s}_0 = \mathfrak{m}_0\mathfrak{n}_0\mathfrak{s}^+$ with $\mathfrak{s}^+ = \mathfrak{m}^+\mathfrak{n}^+$, $\mathfrak{m}^+ \mid \mathfrak{m}_1^2$, $\mathfrak{n}^+ \mid \mathfrak{n}_1^2$, and note that then $\sigma_{K,\mu^2,\chi}(\mathfrak{m}\mathfrak{n}/\mathfrak{u}^2\mathfrak{v}^2) = \chi(\mathfrak{s}_0)\sigma_{K,\mu^2,\chi_0}(\mathfrak{m}\mathfrak{n}/\mathfrak{u}^2\mathfrak{v}^2)$ with the principal character χ_0. Further, $\Sigma_\chi(\mathfrak{m},\mathfrak{n}) = 0$ if $\mathfrak{p} \mid (\mathfrak{n}_1/\mathfrak{n}^+)$ and $\mathfrak{p}^2 \nmid \mathfrak{m}$ for some $\mathfrak{p}$, and a moment's reflection shows that in any case

$$|\Sigma_\chi(\mathfrak{m},\mathfrak{n})| \leqslant \sigma_{\mu^2}(\mathfrak{s}_0) \leqslant \sigma_{\mu^2}(\mathfrak{m}_0)\sigma_{\mu^2}(\mathfrak{n}_0)\sigma_{\mu^2}(\mathfrak{s}^+).$$

To bound $B_1(b)$ from above, we use Peter-Paul inequality as follows:

$$\begin{aligned}
B_1(b) &\leqslant |\Xi| \sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\mathfrak{N}\mathfrak{m},\mathfrak{N}\mathfrak{n}\leqslant M/\mathfrak{N}\mathfrak{d}} \frac{1}{2}\Bigg(b(\mathfrak{d}\mathfrak{m})^2 \frac{\mathfrak{N}\mathfrak{m}^{\beta}}{\mathfrak{N}\mathfrak{n}^{\beta}} \frac{\sigma_{\mu^2}(\mathfrak{n}_0)}{\sigma_{\mu^2}(\mathfrak{m}_0)} \frac{\sigma^*(\mathfrak{n})}{\sigma^*(\mathfrak{m})} \\
&\qquad\qquad + b(\mathfrak{d}\mathfrak{n})^2 \frac{\mathfrak{N}\mathfrak{n}^{\beta}}{\mathfrak{N}\mathfrak{m}^{\beta}} \frac{\sigma_{\mu^2}(\mathfrak{m}_0)}{\sigma_{\mu^2}(\mathfrak{n}_0)} \frac{\sigma^*(\mathfrak{m})}{\sigma^*(\mathfrak{n})}\Bigg)\Sigma(\mathfrak{m},\mathfrak{n}) \\
&\leqslant |\Xi| \sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\mathfrak{N}\mathfrak{m}\leqslant M/\mathfrak{N}\mathfrak{d}} \frac{\mathfrak{N}\mathfrak{m}^{\beta}\, b(\mathfrak{d}\mathfrak{m})^2}{\sigma_*(\mathfrak{m})\sigma^+(\mathfrak{m})} \sum_{\mathfrak{N}\mathfrak{n}\leqslant M/\mathfrak{N}\mathfrak{d}} \sigma_{\mu^2}(\mathfrak{s}^+)\sigma_{\mu^2}^2(\mathfrak{n}_0)\sigma^*(\mathfrak{n})\mathfrak{N}\mathfrak{n}^{-\beta} \\
&\leqslant |\Xi| \sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \sigma_*(\mathfrak{d})\mathfrak{N}\mathfrak{d} \sum_{\mathfrak{N}\mathfrak{m}\leqslant M/\mathfrak{N}\mathfrak{d}} \frac{\mathfrak{N}\mathfrak{m}^{\beta}}{\sigma_*(\mathfrak{d}\mathfrak{m})} b(\mathfrak{d}\mathfrak{m})^2 \\
&\qquad\qquad \sum_{\mathfrak{N}\mathfrak{n}_0\leqslant M/\mathfrak{N}\mathfrak{d}} \sigma_{\mu^2}^2(\mathfrak{n}_0)\sigma_*(\mathfrak{n}_0)\mathfrak{N}\mathfrak{n}_0^{-\beta} \sum_{\mathfrak{n}} \sigma^*(\mathfrak{n}^2)\mathfrak{N}\mathfrak{n}^{-2\beta} \\
&\ll \sum_{\mathfrak{N}\mathfrak{m}\leqslant M} \frac{\mathfrak{N}\mathfrak{m}}{\sigma_*(\mathfrak{m})} b(\mathfrak{m})^2 \sum_{\mathfrak{d}|\mathfrak{m}} \sigma_*(\mathfrak{d}) \left(\frac{M}{\mathfrak{N}\mathfrak{d}}\right)^{(\beta-1)+(3/2-\beta)} \\
&\ll M^{1/2}B(b).
\end{aligned}$$

That this bound is actually tight can be seen by taking e.g. $b(\mathfrak{n})=1$ for square-free $\mathfrak{n}\subset\mathfrak{o}$ and 0 for all other $\mathfrak{n}$. With this particular choice,

$$\begin{aligned}
B(b) &= \sum_{\mathfrak{N}_K\mathfrak{d}\leqslant M^{1/2},\mathfrak{d}\subset\mathfrak{o}} \mu_K^2(\mathfrak{d})(\mathfrak{N}_K\mathfrak{d})^2 = c_K M^{3/2} + \mathrm{O}(M^{5/4}), \\
B_1(b) &\geqslant \sum_{\mathfrak{N}_K\mathfrak{d}\leqslant M^{1/2},\mathfrak{d}\subset\mathfrak{o}} (\mathfrak{N}_K\mathfrak{d})^2 \sum_{\substack{\mathfrak{N}_K\mathfrak{m},\mathfrak{N}_K\mathfrak{n}\leqslant M^{1/2}/\mathfrak{N}_K\mathfrak{d}\\ \mathfrak{m},\mathfrak{n}\subset\mathfrak{o},\ \mathfrak{m}\mathfrak{d},\mathfrak{n}\mathfrak{d}\text{ square-free}}} \sum_{\chi\in\Xi} \chi(\mathfrak{m}\mathfrak{n})\mathfrak{N}_K(\mathfrak{m}\mathfrak{n}) \qquad (5.2) \\
&\geqslant \sum_{\mathfrak{N}_K\mathfrak{d}\leqslant M^{1/2-o(1)},\mathfrak{d}\subset\mathfrak{o}} (\varphi_K(\mathfrak{d}))^2 \left(\frac{M^{1/2}}{\mathfrak{N}_K\mathfrak{d}} + \mathrm{O}(1)\right)^4 \asymp M^2. \qquad \square
\end{aligned}$$

We proceed to the proof of Theorem 1. For this, we employ Lemma 4 with the choice of $b(\mathfrak{n})$ from Lemma 5. The proof of Lemma 5 (specifically (5.2)) gives the order of magnitude of the leading terms; we now also account for the remainder terms. For any $\mathfrak{d}\subset\mathfrak{O}$, denote $\mathfrak{d}'=(\mathfrak{d}\cap\mathfrak{o})\mathfrak{d}^{-1}$. Then

$$\begin{aligned}
R(b) &\leqslant \sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\substack{\mathfrak{N}\mathfrak{m},\mathfrak{N}\mathfrak{n}\leqslant M/\mathfrak{N}(\mathfrak{d}\mathfrak{d}')\\ \mathfrak{m},\mathfrak{n}\subset\mathfrak{o}}} \left(\mathfrak{N}(\mathfrak{m}\mathfrak{n}\mathfrak{d}'^2) + T^2\right) \\
&= \sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \mathfrak{N}\mathfrak{d}\mathfrak{N}(\mathfrak{d}'^2) \left(\sum_{\mathfrak{N}_K\mathfrak{m}\leqslant (M/\mathfrak{N}(\mathfrak{d}\mathfrak{d}'))^{1/2}} (\mathfrak{N}_K\mathfrak{m})^2\right)^2 + T^2 \sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \mathfrak{N}\mathfrak{d}\frac{M}{\mathfrak{N}(\mathfrak{d}\mathfrak{d}')} \qquad (5.3) \\
&\ll M^3 \sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \frac{1}{(\mathfrak{N}\mathfrak{d})^2\mathfrak{N}\mathfrak{d}'} + T^2 M \sum_{\mathfrak{N}\mathfrak{d}'\leqslant M} \frac{1}{\mathfrak{N}\mathfrak{d}'} \left(\frac{M}{\mathfrak{N}\mathfrak{d}'^2}\right)^{1/2} \\
&\ll M^3 + T^2 M^{3/2}.
\end{aligned}$$

As for R_{cusp}, we can write the underlying cusp form of weight two as a linear combination (depending on v only) of Hecke cusp forms $\sum_{n\in\mathfrak{o}} r_i(n)e^{\pi i\mathrm{Tr}(\lambda^{-1}nz)}$ normalized to have $r_i(1)=1$; using the Rankin-Selberg bound for $r_i(n)$'s and their multiplicative properties we can estimate

$$\begin{aligned}
R_{\mathrm{cusp}}(b) &\ll \sum_i \sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\substack{\mathfrak{N}\mathfrak{m},\mathfrak{N}\mathfrak{n}\leqslant M/\mathfrak{N}(\mathfrak{d}\mathfrak{d}')\\ \mathfrak{m},\mathfrak{n}\subset\mathfrak{o},\,(\mathfrak{m},\mathfrak{d}\mathfrak{d}')=(\mathfrak{n},\mathfrak{d}\mathfrak{d}')=\mathfrak{o}}} |\mu_K(\mathfrak{m})\mu_K(\mathfrak{n})||r_i(\mathfrak{m}\mathfrak{n})|\\
&\ll \sum_i \sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\substack{\mathfrak{N}_K\mathfrak{e}\leqslant(M/\mathfrak{N}(\mathfrak{d}\mathfrak{d}'))^{1/2}\\ \mathfrak{e}\subset\mathfrak{o}}}\\
&\qquad\qquad \Bigg(\sum_{\substack{\mathfrak{N}_K\mathfrak{m}\leqslant(M/\mathfrak{N}(\mathfrak{d}\mathfrak{d}'\mathfrak{e}))^{1/2}\\ \mathfrak{m}\subset\mathfrak{o}}} |r_i(\mathfrak{m})|\Bigg)^2 \sum_{\substack{\mathfrak{e}=\mathfrak{e}_1\mathfrak{e}_2\\ \mathfrak{e}_i\subset\mathfrak{o}}} r_i(\mathfrak{e}_1)^2\mathfrak{N}_K\mathfrak{e}_2 \qquad (5.4)\\
&\ll \sum_i \sum_{\mathfrak{N}\mathfrak{d}\leqslant M} \mathfrak{N}\mathfrak{d} \sum_{\substack{\mathfrak{N}_K(\mathfrak{e}_1\mathfrak{e}_2)\leqslant(M/\mathfrak{N}(\mathfrak{d}\mathfrak{d}'))^{1/2}\\ \mathfrak{e}_1,\mathfrak{e}_2\subset\mathfrak{o}}} r_i(\mathfrak{e}_1)^2\mathfrak{N}_K\mathfrak{e}_2\left(\frac{M}{\mathfrak{N}(\mathfrak{d}\mathfrak{d}'\mathfrak{e}_1\mathfrak{e}_2)}\right)^{3/2}\\
&\ll M^{3/2}\sum_{\mathfrak{N}\mathfrak{d}\leqslant M}\frac{1}{\mathfrak{N}\mathfrak{d}^{1/2}\mathfrak{N}\mathfrak{d}'^{3/2}} \ll M^{3/2}\log M.
\end{aligned}$$

Finally, in case when Γ is not cocompact, we can estimate by Lemma 4

$$\begin{aligned}
Q_1^{\infty}(a) &\ll T^{2+\epsilon}\sum_{\mathfrak{N}\mathfrak{d}\leqslant M}\mathfrak{N}\mathfrak{d}\Bigg(\sum_{\mathfrak{N}\mathfrak{m}\leqslant M/\mathfrak{N}\mathfrak{d},\,\mathfrak{d}\mathfrak{m}\subset\mathfrak{o}}\sqrt{\mathfrak{N}\mathfrak{m}}\Bigg)^2\\
&\ll T^{2+\epsilon}\sum_{\mathfrak{N}\mathfrak{d}\leqslant M}\mathfrak{N}\mathfrak{d}\mathfrak{N}\mathfrak{d}'\Bigg(\sum_{\mathfrak{N}_K\mathfrak{m}\leqslant(M/\mathfrak{N}(\mathfrak{d}\mathfrak{d}'))^{1/2},\,\mathfrak{m}\subset\mathfrak{o}}\mathfrak{N}_K\mathfrak{m}\Bigg)^2 \qquad (5.5)\\
&\ll T^{2+\epsilon}M^2\sum_{\mathfrak{N}\mathfrak{d}\leqslant M}\frac{1}{\mathfrak{N}\mathfrak{d}\mathfrak{N}\mathfrak{d}'} \ll T^{2+\epsilon}M^2,\\
Q^{\infty}(a) &\ll T^{1+\epsilon}\Bigg(\sum_{\mathfrak{N}_K\mathfrak{m}\leqslant M^{1/2},\,\mathfrak{m}\subset\mathfrak{o}}\mathfrak{N}_K\mathfrak{m}\Bigg)^2 \ll T^{1+\epsilon}M^2.
\end{aligned}$$

Collecting all terms from (5.2)–(5.5) into (5.1), we find that

$$\begin{aligned}
Q_1(a) &\gg k_T(1)M^2 + \mathrm{O}(k_T(1)M^{3/2}\log M + \delta_\Gamma T^{2+\epsilon}M^2),\\
Q(a) &= c_K k_T(1)M^{3/2} + \mathrm{O}\left((M^3+T^2M^{3/2})e^{2C\frac{\log T}{\log\log T}} + k_T(1)M^{5/4} + \delta_\Gamma T^{1+\epsilon}M^2\right).
\end{aligned}$$

Recalling that $k_T(1)\sim cT^3$, we see that, with the choice $M=T^2\exp(-A\log T/\log\log T)$ for any $A>2C$, remainder terms are smaller than the leading terms.

Returning now to (3.1) and (3.2), we can truncate the spectral sums beyond $T\exp(N\log T/\log\log T)$ with a negligible remainder by our choice of h. We obtain that

$$\max_{r_j\leqslant T\exp\left(\frac{N\log T}{\log\log T}\right)}|\phi_j(v)|^2 \gg \frac{Q_1(a)}{Q(a)} \gg Te^{-\frac{A}{2}\frac{\log T}{\log\log T}}.$$

As $T \gg \lambda_j^{1/2}\exp(-N\log\lambda_j/2\log\log\lambda_j)$, Theorem 1 is thus proved. □

References

[As1] Asai, T.: *On certain Dirichlet series associated with Hilbert modular forms and Rankin's method.* Math. Ann., Vol. 226 (1977), No. 1, 81–94.

[As2] Asai, T.: *On the Doi–Naganuma lifting associated with imaginary quadratic fields.* Nagoya Math. J., Vol. 71 (1978), 149–167.

[Be] Berry, M.V.: *Regular and irregular semiclassical wavefunctions.* J. Phys. A: Math. Gen., Vol. 10 (1977), No. 12, 2083–2091.

[Bu-Gé-Tz] Burq, N., Gérard, P., Tzvetkov, N.: *Restrictions of the Laplace-Beltrami eigenfunctions to submanifolds.* Duke Math. J., Vol. 138 (2007), No. 3, 445–486.

[Do] Donnelly, H.: *Exceptional sequences of eigenfunctions for hyperbolic manifolds.* Proc. Amer. Math. Soc., Vol. 135 (2007), No. 5, 1551–1555.

[Du-Ko-Va] Duistermaat, J.J., Kolk, J.A.C., Varadarajan, V.S.: *Functions, flows and oscillatory integrals on flag manifolds and conjugacy classes in real semisimple Lie groups.* Compositio Math., Vol. 49 (1983), No. 3, 309–398.

[Ei] Eichler, M.: *Lectures on modular correspondences.* Tata Inst. Fund. Res., Lectures Math. Phys. 9, 1955.

[El-Gr-Me1] Elstrodt, J., Grünewald, F., Mennicke, J.: *Groups acting on hyperbolic space: harmonic analysis and number theory.* Springer, 1997.

[El-Gr-Me2] Elstrodt, J., Grünewald, F., Mennicke, J.: *Zeta functions of binary Hermitian forms and special values of Eisenstein series on three-dimensional hyperbolic space.* Math. Ann., Vol. 277 (1987), No. 4, 655–708.

[Fl] Flicker, Y.: *Twisted tensors and Euler products.* Bull. Soc. Math. France, Vol. 116 (1988), No. 3, 295–313.

[Gi-Ji-So] Ginzburg, D., Jiang, D., Soudry, D.: *Poles of L-functions and theta liftings for orthogonal groups.* J. Inst. Math. Jussieu, Vol. 8 (2009), No. 4, 693–741.

[Gr-Pr] Gross, B.H., Prasad, D.: *On the decomposition of a representation of* SO_n *when restricted to* SO_{n-1}. Canad. J. Math., Vol. 44 (1992), No. 5, 974–1002.

[He-Ra] Hejhal, D.A., Rackner, B.N.: *On the topography of Maass waveforms for* $\mathrm{PSL}(2,\mathbf{Z})$. Experiment. Math. 1 (1992), No. 4, 275–305.

[He] Helgason, S.: *Differential geometry and symmetric spaces.* Pure Appl. Math., Vol. XII, Acad. Press, 1962.

[Iw] Iwaniec, H.: *Introduction to the Spectral Theory of Automorphic Forms.* Revista Matemática Iberoamericana, 1995.

[Iw-Sa] Iwaniec, H., Sarnak, P.: L^∞ *norms of eigenfunctions on arithmetic surfaces.* Ann. of Math., 2nd Ser., Vol. 141 (1995), No.2, 301–320.

[Ka-Sa] Katok, S., Sarnak, P.: *Heegner points, cycles and Maass forms.* Israel J. Math., Vol. 84 (1993), No. 1–2, 193–227.

[Ki] Kitaoka, Y.: *Arithmetic of quadratic forms.* Cambridge Univ. Press, 1993.

[Ko] Koyama, S.: L^∞*-norms of eigenfunctions for arithmetic hyperbolic 3-manifolds.* Duke Math. J., Vol. 77 (1995), No. 3, 799–817.

[Kr] Krishnamurthy, M.: *The Asai transfer to* GL_4 *via the Langlands–Shahidi method.* Int. Math. Res. Not., 2003, No. 41, 2221–2254.

[Ku-Mi] *Intersection numbers of cycles on locally symmetric spaces and Fourier coefficients of holomorphic modular forms in several complex variables.* Inst. Hautes Études Sci. Publ. Math., Vol. 71 (1990), 121–172.

[La-Of] Lapid, E., Offen, O.: *Compact unitary periods.* Compos. Math., Vol. 143 (2007), No. 2, 323–338.

[Ma-Re] Maclachlan, C., Reid, A.W.: *The Arithmetic of Hyperbolic 3-manifolds.* Springer-Verlag, 2003.

[Mi] Milićević, D.: *Large values of eigenfunctions on arithmetic hyperbolic surfaces.* Duke Math. J., accepted for publication.

[Mi-Ta] Milićević, D., Takloo-Bighash, R.: *Base change, theta lifting, and arithmetic hyperbolic 3-manifolds of Maclachlan-Reid type.* In preparation.

[Ra] Ramakrishnan, D.: *Modularity of solvable Artin representations of* $\mathrm{GO}(4)$*-type.* Int. Math. Res. Not., 2002, No.1, 1–54.

[Ru-Sa] Rudnick, Z., Sarnak, P.: *The behaviour of eigenstates of arithmetic hyperbolic manifolds.* Comm. Math. Phys., Vol. 161 (1994), No. 1, 195–213.

[Sa1] Sarnak, P.: *Arithmetic quantum chaos.* The R.A.Blyth lectures, Univ. Toronto, 1993.

[Sa2] Sarnak, P.: A letter to Cathleen Morawetz, 2004.

[Shim] Shimizu, H.: *Theta series and automorphic forms on* GL_2. J. Math. Soc. Japan, Vol. 24 (1972), No. 4, 638–683.

[Shin] Shintani, T.:: *On construction of holomorphic cusp forms of half integral weight.* Nagoya Math J., Vol. 58 (1975), 83–126.

[So] Soundararajan, K.: *Extreme values of zeta and L-functions.* Math.. Ann.., Vol. 342 (2008), No. 2, 467–486.

[SP] Schulze-Pillot, R.: *Representation by integral quadratic forms—a survey.* Algebraic and arithmetic theory of quadratic forms, Contemp. Math. 344, Amer. Math. Soc., 2004, 323–337.

[Tak1] Takase, K.: *On certain Dirichlet series associated with automorphic forms on* $\mathrm{SL}(2, \mathbf{C})$. Man. Math. 56, 1986, 293–312.

[Tak2] Takase, K.: *Wave forms on* $\mathrm{O}(1, q+1)$ *and associated Dirichlet series.* Proc. Japan Acad., Vol. 62, Ser. A, 1986, 112–115.

[Tan] Tanigawa, Y.: *Selberg trace formula for Picard groups.* Algebraic Number Theory, Kyoto Int. Symp. 1976, S. Iyanaga (Ed.), Tokyo 1977.

[To-Ze] Toth, J.A., Zelditch, S.: *Norms of modes and quasi-modes revisited.* Harmonic analysis at Mount Holyoke, Contemp. Math. 320, Amer. Math. Soc., 2003, 435–458.

[Va] Varadarajan, V.S.: *The method of stationary phase and applications to geometry and analysis on Lie groups.* Algebraic and analytic methods in representation theory, European School of Group Theory 1994, B. Ørsted and H. Schlichtkrull (Eds.), Persp. Math. 17, Acad. Press., 1997, 167–242.

[Wa] Watson, T.C.: *Rankin triple products and quantum chaos.* To appear in Ann.Math., accepted for publication in 2009.