J. Math. Pures Appl. 168 (2022) 1–64

Contents lists available at ScienceDirect

Journal de Mathématiques Pures et Appliquées

journal homepage: www.elsevier.com/locate/matpur

Beyond the spherical sup-norm problem

Valentin Blomer [a], Gergely Harcos [b,c,*], Péter Maga [b,c], Djordje Milićević [d]

[a] *Mathematisches Institut, Endenicher Allee 60, D-53115 Bonn, Germany*
[b] *Alfréd Rényi Institute of Mathematics, POB 127, Budapest H-1364, Hungary*
[c] *MTA Rényi Intézet Lendület Automorphic Research Group, POB 127, Budapest H-1364, Hungary*
[d] *Bryn Mawr College, Department of Mathematics, 101 North Merion Avenue, Bryn Mawr, PA 19010, USA*

ARTICLE INFO

Article history:
Received 4 April 2022
Available online 26 September 2022

MSC:
11F72
11F41
11F70
11J25
22E30
43A90

Keywords:
Sup-norm problem
Pre-trace formula
Paley–Wiener theorem
Spherical function

ABSTRACT

We open a new perspective on the sup-norm problem and propose a version for non-spherical Maaß forms when the maximal compact K is non-abelian and the dimension of the K-type gets large. We solve this problem for an arithmetic quotient of $G = \mathrm{SL}_2(\mathbb{C})$ with $K = \mathrm{SU}_2(\mathbb{C})$. Our results cover the case of vector-valued Maaß forms as well as all the individual scalar-valued Maaß forms of the Wigner basis, reaching sub-Weyl exponents in some cases. On the way, we develop analytic theory of independent interest, including uniform strong localization estimates for generalized spherical functions of high K-type and a Paley–Wiener theorem for the corresponding spherical transform acting on the space of rapidly decreasing functions. The new analytic properties of the generalized spherical functions lead to novel counting problems of matrices close to various manifolds that we solve optimally.

RÉSUMÉ

Nous ouvrons une nouvelle voie dans le problème de la norme sup et proposons un traitement de ce problème dans le cadre des formes de Maaß non sphériques lorsque le groupe compact maximal K n'est pas abélien et que la dimension des K-types croît. Nous resolvons ce problème pour un quotient arithmétique de $G = \mathrm{SL}_2(\mathbb{C})$ avec $K = \mathrm{SU}_2(\mathbb{C})$. Nos résultats contiennent le cas des formes de Maaß à valeurs vectorielles ainsi que des formes de Maaß à valeurs scalaires de la base de Wigner, obtenant des exposants allant au-delà de la borne de Weyl dans certains cas. Nous obtenons ce faisant des résultats analytiques ayant leur intérêt propre, incluant de puissantes estimations de localisation uniforme pour des fonctions sphériques généralisées de grands K-types, ainsi qu'un théorème de Paley–Wiener pour la transformation sphérique associée sur l'espace des fonctions rapidement décroissantes. Ces propriétés analytiques nouvelles des fonctions sphériques généralisées mènent à de nouveaux problèmes de comptage de matrices

* Corresponding author.
E-mail addresses: blomer@math.uni-bonn.de (V. Blomer), gharcos@renyi.hu (G. Harcos), magapeter@gmail.com (P. Maga), dmilicevic@brynmawr.edu (D. Milićević).

https://doi.org/10.1016/j.matpur.2022.09.009

proches de certaines variétés, que nous sommes en mesure de résoudre de manière optimale.

1. Introduction

1.1. The spherical sup-norm problem

The sup-norm problem on arithmetic Riemannian manifolds is a question at the interface of harmonic analysis and number theory that intrinsically combines techniques from both areas. Let $X = \Gamma\backslash G/K$ be a locally symmetric space of finite volume, where Γ is an arithmetic subgroup. Arithmetically and analytically, the most interesting functions in $L^2(X)$ are joint eigenfunctions ϕ of all invariant differential operators and the Hecke operators: these are precisely the functions that arise from (spherical) automorphic forms. The sup-norm problem asks for a quantitative comparison of the L^2-norm $\|\phi\|_2$ and the sup-norm $\|\phi\|_\infty$, most classically in terms of the Laplace eigenvalue λ_ϕ, but depending on the application also in terms of the volume of X or other relevant quantities. Upper bounds for the sup-norm in terms of the Laplace eigenvalue are a measure for the equidistribution of the mass of high energy eigenfunctions which sheds light on the question to what extent these eigenstates can localize ("scarring"). Besides the quantum mechanical interpretation, the sup-norm problem in its various incarnations has connections to the multiplicity problem, zero sets and nodal lines of automorphic functions, and bounds for Faltings' delta function, to name just a few. See [42,39,23,29].

If X is compact, the most general upper bound is due to Sarnak [42]:

$$\|\phi\|_\infty \ll_X \lambda_\phi^{(\dim X - \mathrm{rk}\, X)/4} \|\phi\|_2, \tag{1.1}$$

a bound which does not use the Hecke property and is in fact sharp (for general X) under these weaker assumptions. Sarnak derives this bound from asymptotics of spherical functions. A slightly different but ultimately related argument proceeds via a pre-trace inequality that bounds $\|\phi\|_\infty^2$ by a sum of an automorphic kernel over $\gamma \in \Gamma$. If the test function is an appropriate Paley–Wiener function, only the identity contributes to this sum, and one obtains as a ("trivial") upper bound for $\|\phi\|_\infty$ the square-root of the spectral density as given in terms of the Harish-Chandra **c**-function. If the Langlands parameters of ϕ are in generic position, this coincides with (1.1).

To go beyond (1.1), one uses a test function that localizes not only the archimedean Langlands parameters, but in addition the parameters at a large number of finite places (where "large" means a function tending to infinity as a small and carefully chosen power of λ_ϕ). This is called the amplification technique and leads, after estimating the automorphic kernel, to a problem in the geometry of numbers: count the elements of G which appear in Hecke correspondences and lie in regions of G according to the size of the kernel (such as counting rescaled integer matrices lying close to K). It has been implemented successfully in a variety of cases, see e.g. [27,26,7,5,35,48,41,3] and the references therein.

1.2. Automorphic forms with K-types

In this paper we open a new perspective on the sup-norm problem and propose a version of higher complexity. The sup-norm problem makes perfect sense not only on the level of symmetric spaces, but also on the level of groups, and a priori there is no reason why one should restrict to spherical, i.e. right K-invariant automorphic forms. Let τ be an irreducible unitary representation of K on some finite-dimensional complex vector space V^τ, and consider the homogeneous vector bundle over G/K defined by τ. A cross-

section may then be identified with a vector-valued function $f : G \to V^\tau$ which transforms on the right by K with respect to τ:

$$f(gk) = \tau(k^{-1})f(g), \qquad g \in G, \quad k \in K.$$

It is now an interesting question to bound the sup-norm of f or, more delicately, its components as the *dimension* of V^τ gets large. Such a situation cannot be realized in the classical case $G = \mathrm{SL}_2(\mathbb{R})$, since $K = \mathrm{SO}_2(\mathbb{R})$ is abelian, hence each V^τ is one-dimensional. In this paper, we offer a detailed investigation of the first nontrivial case $G = \mathrm{SL}_2(\mathbb{C})$. For concreteness, we choose the congruence lattice $\Gamma = \mathrm{SL}_2(\mathbb{Z}[i])$, although our results extend to more general arithmetic quotients of G using the techniques in [3].

Nontrivial irreducible unitary representations of G are principal series representations parametrized by certain pairs $(\nu, p) \in \mathfrak{a}_{\mathbb{C}}^* \times \frac{1}{2}\mathbb{Z}$, where as usual $\mathfrak{a}$ is the Lie algebra of the subgroup of positive diagonal matrices; see §2.2. (By a small abuse of notation we will later interpret ν simply as a complex number.) Each representation space V of G decomposes as a Hilbert space direct sum

$$V = \bigoplus_{\substack{\ell \geqslant |p| \\ \ell \equiv p \pmod 1}} V^\ell = \bigoplus_{\substack{\ell \geqslant |p| \\ \ell \equiv p \pmod 1}} \bigoplus_{\substack{|q| \leqslant \ell \\ q \equiv \ell \pmod 1}} V^{\ell,q}, \tag{1.2}$$

where $V^{\ell,q}$ is one-dimensional. Here and later, $\ell \in \frac{1}{2}\mathbb{Z}_{\geqslant 0}$ parametrizes the K-type, i.e. the $(2\ell+1)$-dimensional representation τ_ℓ of K, and the diagonal matrix $\mathrm{diag}(e^{i\varrho}, e^{-i\varrho}) \in K$ acts on $V^{\ell,q}$ by $e^{2qi\varrho}$. (The upper index ℓ in V^ℓ should not be mistaken for an ℓ-th power.)

Representations occurring in $L^2(\Gamma\backslash G)$ consist of even functions on G and have $p \in \mathbb{Z}$. A representation contains a spherical vector if and only if $p = 0$. In particular, the forms with $p \neq 0$ are untouched by any of the spherical sup-norm literature. For $p \neq 0$, no complementary series exists, so $\nu \in i\mathfrak{a}^*$.

1.3. Main results I: vector-valued forms

As explained above, we are interested in "big" K-types which occur for all representation parameters $|p| \leqslant \ell$, but arguably the most interesting case is when the K-type is "new" and no lower K-types appear in the same automorphic representation space. Hence from now on we restrict to $p = \ell$. The sup-norm problem for large ν was studied in detail in [3], so here we keep ν in a fixed compact subset $I \subset i\mathbb{R}$ and let ℓ vary. The spectral density is a constant multiple of $p^2 - \nu^2$. In particular, for a given K-type τ_ℓ, there are $\mathrm{O}_I(\ell^2)$ cuspidal automorphic representations $V \subset L^2(\Gamma\backslash G)$ with spectral parameter $\nu \in I$ and $p = \ell$ (see [13]), and in the light of the trace formula this bound is expected to be sharp. In each of these we consider the $(2\ell+1)$-dimensional subspace V^ℓ. Let us choose an orthonormal basis $\{\phi_q : |q| \leqslant \ell\}$ of V^ℓ, with $\phi_q \in V^{\ell,q}$ as in (1.2). The function $G \to \mathbb{C}^{2\ell+1}$ given by

$$g \mapsto (\phi_{-\ell}(g), \dots, \phi_\ell(g))^\top \tag{1.3}$$

is a vector-valued automorphic form for the group Γ with spectral parameter ν and K-type τ_ℓ. The Hermitian norm of this function,

$$\Phi(g) := \Big(\sum_{|q| \leqslant \ell} |\phi_q(g)|^2 \Big)^{1/2}, \qquad g \in G,$$

is independent of the choice of the orthonormal basis, and it satisfies $\|\Phi\|_2 = (2\ell+1)^{1/2}$. Let us fix a compact subset $\Omega \subset G$. Our remarks on spectral density and dimension suggest that

$$\|\Phi|_\Omega\|_\infty := \Big\| \sum_{|q|\leqslant \ell} |\phi_q|_\Omega|^2 \Big\|_\infty^{1/2} \ll_{I,\Omega} \ell^{3/2} \tag{1.4}$$

should be regarded as the "trivial" bound; this is made precise in Remark 2 below. Our first result is a power-saving improvement.

Theorem 1. *Let $\ell \geqslant 1$ be an integer, $I \subset i\mathbb{R}$ and $\Omega \subset G$ be compact sets. Let $V \subset L^2(\Gamma\backslash G)$ be a cuspidal automorphic representation with minimal K-type τ_ℓ and spectral parameter $\nu_V \in I$. Then for any $\varepsilon > 0$ we have*

$$\|\Phi|_\Omega\|_\infty \ll_{\varepsilon,I,\Omega} \ell^{4/3+\varepsilon}.$$

We will explain some ideas of the proof in a moment, but we remark already at this point that the exponent is the best possible, given that we sacrifice cancellation of the terms on the geometric side of the pre-trace formula and given our current knowledge on the construction of the most efficient amplifier. In other words, under these conditions we solve the arising matrix counting problem optimally. Since we trivially have $\|\Phi\|_\infty \gg \ell^{1/2}$, the above bound is one-sixth of the way from the trivial down to the best possible exponent (absent the possibility of some escape of mass into a cusp). This matches (after a renormalization) the original and still the best available subconvexity exponent 5/24 of Iwaniec–Sarnak [27] for the sup-norms of spherical Maaß forms of large Laplace eigenvalue on arithmetic hyperbolic surfaces.

1.4. Main results II: individual vectors

It is a much more subtle endeavor to investigate the sup-norm of the individual basis elements ϕ_q. Here one must contend with the inherent high multiplicity, a known serious barrier in the sup-norm problem. Indeed, a straightforward construction [42] shows that some scalar-valued L^2-normalized form $\phi \in V^\ell$ (essentially the projection of the vector-valued form (1.3) in the modulus-maximizing direction) has sup-norm on Ω as large as $\|\Phi|_\Omega\|_\infty$ in Theorem 1. However, our natural basis $\{\phi_q : |q| \leqslant \ell\}$ of V^ℓ is distinguished by consisting of eigenfunctions under the action of the group $\{\mathrm{diag}(e^{i\theta}, e^{-i\theta}) : \theta \in \mathbb{R}\}$ of diagonal matrices in K. This is the classical basis with respect to which the representation τ_ℓ is given by the Wigner D-matrix. By a similar heuristic reasoning as for (1.4), one might expect that the baseline bound should be $\|\phi_q|_\Omega\|_\infty \ll_{I,\Omega} \ell$. Indeed, we prove this bound in considerable generality up to a factor of ℓ^ε (cf. Remark 2 below), noting that it is not "trivial" in any sense other than that it does not require arithmeticity. Moreover, in the situation of Theorem 1, we are in fact able to break this barrier uniformly for all q, as shown by the next theorem.

Theorem 2. *Under the assumptions of Theorem 1, we have*

$$\max_{|q|\leqslant\ell} \|\phi_q|_\Omega\|_\infty \ll_{\varepsilon,I,\Omega} \ell^{26/27+\varepsilon}.$$

For special values of q we can improve on the exponent considerably. The central vector ϕ_0 is distinguished as the "archimedean newvector" [37] in the sense that its Whittaker function determines the archimedean L-factor of the underlying representation. Another interesting situation is the extreme case of the vector $\phi_{\pm\ell}$.

Theorem 3. *Keep the assumptions of Theorem 1.*

(a) For $q = 0$ we have

$$\|\phi_0|_\Omega\|_\infty \ll_{\varepsilon,I,\Omega} \ell^{7/8+\varepsilon}.$$

(b) Suppose that V lifts to an automorphic representation for $\mathrm{PGL}_2(\mathbb{Z}[i])\backslash\mathrm{PGL}_2(\mathbb{C})$. *For $q = \pm\ell$ we have*

$$\|\phi_{\pm\ell}|_\Omega\|_\infty \ll_{\varepsilon,I,\Omega} \ell^{1/2+\varepsilon}.$$

The strong numerical saving in the case $q = \pm\ell$, going far beyond the Weyl exponent, is quite remarkable, in particular in view of the seemingly weaker saving in Theorem 1 which might be regarded as an easier case. We will discuss this in §1.7. The assumption that V is associated to a representation of PGL_2 rather than SL_2 is only for technical simplicity and not essential to the method, cf. §2.7. This assumption holds if and only if the elements of V are fixed by the Hecke operator T_i (which is an involution on $L^2(\Gamma\backslash G)$).

Remark 1. In the case of the spherical sup-norm problem, Sarnak [42] put forward the purity conjecture that the accumulation points of the set

$$\left\{\frac{\log\|\psi\|_\infty}{\log\lambda_\psi} : \psi \text{ is a joint eigenfunction}\right\}$$

lie in $\frac{1}{4}\mathbb{Z}$. It would be very interesting to see if an analogous conjecture may be expected in the K-aspect, and even if there may be examples exhibiting different layers of power growth as in [36,4,6]. In particular, the savings in Theorem 3 produce already a considerable "exponent gap".

Remark 2. We record that our essentially best possible estimates on the spherical trace function in §1.5, which are of purely analytic nature, coupled with the formalism of the pre-trace inequality, yield what might be considered "trivial" geometric estimates: for any co-finite Kleinian subgroup $\Gamma \leqslant G$, without any arithmeticity assumption, we have

$$\|\Phi|_\Omega\|_\infty \ll_{I,\Omega,\Gamma} \ell^{3/2} \qquad \text{and} \qquad \max_{|q|\leqslant\ell} \|\phi_q|_\Omega\|_\infty \ll_{\varepsilon,I,\Omega,\Gamma} \ell^{1+\varepsilon}$$

for any L^2-normalized vector-valued Maaß eigenform $(\phi_{-\ell},\dots,\phi_\ell)^\top$ with spectral parameter $\nu \in I$ and K-type τ_ℓ (with $\phi_q \in V^{\ell,q}$ as before).

Our Theorems 1–3 above, and the non-spherical sup-norm problem in general, come with several novelties of representation theoretic, analytic and arithmetic nature that we discuss briefly in the following subsections.

1.5. Generalized spherical functions

The classical pre-trace formula features on the geometric side the Harish-Chandra transform $\check{h}$ of the test function h on the spectral side. This transform is a bi-K-invariant function obtained by integrating h against the elementary spherical functions (which themselves are bi-K-invariant, and hence in the case of $G = \mathrm{SL}_2(\mathbb{C})$ simply a function of one real variable). In typical applications there is no cancellation in this integral, so an asymptotic analysis of spherical functions is the first key step (see [7] for a general result in this direction). Our set-up requires a generalized version for homogeneous vector bundles over G/K. For $G = \mathrm{SL}_2(\mathbb{C})$, the corresponding *spherical trace function* equals (see §2.4 for details)

$$\varphi^\ell_{\nu,\ell}(g) = (2\ell+1)\int_K \psi_\ell(\kappa(k^{-1}gk))\, e^{(\nu-1)\rho(H(gk))}\,\mathrm{d}k, \tag{1.5}$$

where $\mathrm{d}k$ is the probability Haar measure on K, ρ is the unique positive root, κ (resp. H) is the KAN Iwasawa projection onto K (resp. $\mathfrak{a}$), and

$$\psi_\ell\left(\begin{pmatrix}\alpha & \beta\\ -\bar\beta & \bar\alpha\end{pmatrix}\right) := \bar\alpha^{2\ell}, \qquad \begin{pmatrix}\alpha & \beta\\ -\bar\beta & \bar\alpha\end{pmatrix} \in K. \tag{1.6}$$

The trivial bound is $|\varphi^\ell_{\nu,\ell}(g)| \leqslant 2\ell+1$, which is sharp for $g = \pm\mathrm{id}$, and the key question is how quickly $\varphi^\ell_{\nu,\ell}(g)$ decays, uniformly in ℓ, as $g \in G$ moves away from $\pm\mathrm{id}$. We observe that $\varphi^\ell_{\nu,\ell}(g)$ is invariant under conjugation by K, hence it suffices to investigate it for upper triangular matrices $g \in G$. We shall use the Frobenius norm $\|g\| := \sqrt{\mathrm{tr}(gg^*)}$, and we note that for $g \in G$ this is always at least $\sqrt{2}$. The following bound is new and most likely sharp for fixed $\nu \in i\mathbb{R}$ (up to factors ℓ^ε and powers of $\|g\|$, which we did not try to optimize).

Theorem 4. *Let $\ell \geqslant 1$ be an integer, and let $g = \left(\begin{smallmatrix} z & u \\ & z^{-1}\end{smallmatrix}\right) \in G$ be upper triangular. Then for any $\nu \in i\mathbb{R}$, $k \in K$, $\varepsilon > 0$, we have*

$$\varphi^\ell_{\nu,\ell}(k^{-1}gk) \ll_\varepsilon \min\left(\ell, \frac{\ell^\varepsilon\|g\|^6}{|z^2-1|^2}, \frac{\ell^{1/2+\varepsilon}\|g\|^3}{|u|}\right).$$

The proof shows that the factors ℓ^ε can be replaced with a suitable power of $\log 2\ell$. The same remark applies to Theorems 5 and 6 below.

The spherical trace function $\varphi^\ell_{\nu,\ell}$ can be used to analyze the vector-valued function (1.3). It is, unfortunately, unable to identify the individual components ϕ_q, and there does not seem to exist a general theory of spherical functions covering such cases. As the components are eigenfunctions of the action of the diagonal elements, we can single out ϕ_q by considering

$$\varphi^{\ell,q}_{\nu,\ell}(g) := \frac{1}{2\pi}\int_0^{2\pi} \varphi^\ell_{\nu,\ell}\left(g\,\mathrm{diag}(e^{i\varrho}, e^{-i\varrho})\right) e^{-2qi\varrho}\,\mathrm{d}\varrho. \tag{1.7}$$

The function $\varphi^{\ell,q}_{\nu,\ell}$ is an interesting object that does not seem to have been considered before. It is not conjugation invariant anymore, so it needs to be analyzed on the entire 6-dimensional group $G = \mathrm{SL}_2(\mathbb{C})$, and little preliminary reduction is possible. When restricted to K, it is not hard to see that $\varphi^{\ell,q}_{\nu,\ell}(k)$, for $k = k[u,v,w] \in K$ written in terms of Euler angles (cf. (2.1)), is essentially a Jacobi polynomial in $\cos 2v$. We refer to §5.4 for a more detailed discussion. In particular, $\varphi^{\ell,q}_{\nu,\ell}(\pm\mathrm{id}) = 1$. Therefore, at least heuristically, a safe baseline bound should be

$$\varphi^{\ell,q}_{\nu,\ell}(g) \ll_\varepsilon \ell^\varepsilon. \tag{1.8}$$

Unlike in the bi-K-invariant case, where the trivial bound is just an application of the triangle inequality and hence is indeed trivial, the expected baseline bound (1.8) turns out to be hard to prove. It requires very strong cancellation in the ϱ-integral, along with the decay properties of $\varphi^\ell_{\nu,\ell}$. Taking (1.8) for granted, we wish to investigate in what directions and with what speed we can identify decay as we move away from $\pm\mathrm{id} \in G$. Interestingly, this is extremely sensitive to the value of q.

Let $\mathcal{D} \subset G$ be the set of diagonal matrices, $\mathcal{S}$ the normalizer of A in K (which consists of the diagonal and the skew-diagonal matrices lying in K), and

$$\mathcal{N} := \left\{\begin{pmatrix}a & b\\ c & d\end{pmatrix} \in G : |a| = |d|,\ |b| = |c|\right\}. \tag{1.9}$$

It is clear that $\mathcal{S} \subset K \subset \mathcal{N} \subset G$. For $g \in G$ and non-empty $\mathcal{H} \subset G$, we shall write $\mathrm{dist}(g,\mathcal{H})$ for their distance $\inf_{h\in\mathcal{H}}\|g-h\|$. For later reference, we note that $\|g-h\| = \|g^{-1}-h^{-1}\|$, hence also

$$\mathrm{dist}(g,\mathcal{H}) = \mathrm{dist}(g^{-1},\mathcal{H}^{-1}). \tag{1.10}$$

As an alternative to $\operatorname{dist}(g,\mathcal{N})$, we shall also use

$$D(g) := \big||a|^2 - |d|^2\big| + \big||b|^2 - |c|^2\big|. \tag{1.11}$$

For orientation, we remark the elementary inequality

$$\operatorname{dist}(g,\mathcal{N})^2 \leqslant D(g) \leqslant 2\|g\| \operatorname{dist}(g,\mathcal{N}).$$

In the following theorem, we show that $\varphi_{\nu,\ell}^{\ell,q}(g)$ decays away from K and $\mathcal{D}$ in generic ranges, for all $|q| \leqslant \ell$, and with considerable uniformity.

Theorem 5. *Let $\ell, q \in \mathbb{Z}$ be such that $\ell \geqslant \max(1,|q|)$. Let $\nu \in i\mathbb{R}$ and $g \in G$. Then for any $\varepsilon > 0$ and $\Lambda > 0$, we have*

$$\varphi_{\nu,\ell}^{\ell,q}(g) \ll_{\varepsilon,\Lambda} \ell^{\varepsilon} \min\left(1, \frac{\|g\|}{\sqrt{\ell}\operatorname{dist}(g,K)^2 \operatorname{dist}(g,\mathcal{D})}\right) + \ell^{-\Lambda}. \tag{1.12}$$

The proof of Theorem 5 uses a soft argument that provides some decay for all $|q| \leqslant \ell$, despite the substantial dependence of $\varphi_{\nu,\ell}^{\ell,q}$ on this parameter. In the special case $q \in \{-\ell, 0, \ell\}$, we use more elaborate arguments for stronger bounds.

Theorem 6. *Let $\ell \geqslant 1$ be an integer, $\nu \in i\mathbb{R}$ and $g \in G$. Let $\varepsilon > 0$ and $\Lambda > 0$ be two parameters.*

(a) We have

$$\varphi_{\nu,\ell}^{\ell,0}(g) \ll_{\varepsilon,\Lambda} \ell^{\varepsilon} \min\left(1, \frac{1}{\sqrt{\ell}\operatorname{dist}(g,\mathcal{S})}\right) + \ell^{-\Lambda}. \tag{1.13}$$

Moreover, $\varphi_{\nu,\ell}^{\ell,0}(g) \ll_{\Lambda} \ell^{-\Lambda}$ holds unless $D(g) \ll_{\Lambda} \|g\|^2(\log \ell)/\sqrt{\ell}$.

(b) We have

$$\varphi_{\nu,\ell}^{\ell,\pm\ell}(g) \ll_{\varepsilon} \|g\|^{-2+\varepsilon}\ell^{\varepsilon}. \tag{1.14}$$

Moreover, $\varphi_{\nu,\ell}^{\ell,\pm\ell}(g) \ll_{\Lambda} \ell^{-\Lambda}$ holds unless $\operatorname{dist}(g,\mathcal{D}) \ll_{\Lambda} \|g\|\sqrt{\log \ell}/\sqrt{\ell}$.

We expect that the bounds in Theorem 6 are essentially best possible, possibly up to powers of ℓ^{ε} and $\|g\|$. The proof requires detailed analysis that could in principle be applied to all values of q and would detect, for instance, further Airy-type bumps in certain regions and for certain choices of parameters.

Remark 3. Less precise results but in a more general setting were obtained by Ramacher [38] using operator theoretical methods. Combined with an argument of Marshall [35], these were applied by Ramacher–Wakatsuki [40] to the sup-norm problem with K-types. For compact arithmetic quotients of $\mathrm{SL}_2(\mathbb{C})$, and for $\phi \in V^{\ell}$ as before, [40, Th. 7.12] yields $\|\phi\|_{\infty} \ll \ell^{5/2-\delta}$ with an unspecified constant $\delta > 0$; this does not even recover the baseline bound.

1.6. Paley–Wiener theory

For a reductive Lie group G, Paley–Wiener theory characterizes the image of $C_c^{\infty}(G)$ under the Harish-Chandra transform. For bi-K-invariant functions, this is a famous result of Gangolli [16]: the image consists of entire, Weyl group invariant functions satisfying certain growth conditions. For general K-finite functions,

the picture is much more complicated: any linear relation that holds for the matrix coefficients of generalized principal series also needs to hold for the matrix coefficients of the operator-valued Fourier transform (and hence for the τ-spherical transforms for $\tau \in \widehat{K}$). A complete list of these "Arthur–Campoli relations" requires a full knowledge of all the irreducible subquotients of the non-unitary principal series, which in general is not available. Arthur [1] describes them as a sequence of successive residues of certain meromorphic functions; see also [10]. Needless to say, a good knowledge of available functions on the spectral side is crucial for the quantitative analysis of the pre-trace formula in the sup-norm problem.

For the case of $G = \mathrm{SL}_2(\mathbb{C})$, in a somewhat neglected paper, Wang [50] devised an elegant argument to establish a completely explicit Paley–Wiener theorem for the τ_ℓ-spherical transform acting on $C_c^\infty(G)$: in addition to the Weyl group symmetry, we have the additional symmetry $(\nu, p) \leftrightarrow (p, \nu)$ whenever $\nu \equiv p \pmod 1$ and $|\nu|, |p| \leqslant \ell$; see Theorem 11 in §2.4. The additional symmetry is counter-intuitive at first (the pairs $(\nu, p) \neq (0, 0)$ satisfying $\nu \equiv p \pmod 1$ correspond to a discrete set of non-unitary representations), but it enters the picture as it fixes the eigenvalues $\nu^2 + p^2$ and νp of two generators of $Z(\mathcal{U}(\mathfrak{g}))$, and hence the infinitesimal character. See [50, Cor. 2] and its proof. A more conceptual explanation, along the lines of irreducible subquotients, can be found after (2.7). Wang's remarkable result is that these are *all* relations.

The extra symmetry makes the application of the pre-trace formula more delicate. For instance, it appears impossible to single out an individual value of p by a manageable test function on the spectral side. We circumvent this problem by employing a carefully chosen Gaussian (3.14) that at least asymptotically singles out our preferred value $p = \ell$. The price to pay for this maneuver is that we lose compact support. As a result of independent interest, we prove a new Paley–Wiener theorem for K-finite Schwartz class functions on $G = \mathrm{SL}_2(\mathbb{C})$. For the notation, see §2.4.

Theorem 7. *For $f \in \mathcal{H}(\tau_\ell)$, the following two conditions are equivalent (with implied constants depending on f).*

(a) The function $f(g)$ is smooth, and for any $m \in \mathbb{Z}_{\geqslant 0}$ and $A > 0$ we have

$$\frac{\partial^m}{\partial h^m} f(k_1 a_h k_2) \ll_{m,A} e^{-A|h|}, \qquad h \in \mathbb{R}, \quad k_1, k_2 \in K. \tag{1.15}$$

(b) The function $\widehat{f}(\nu, p)$ extends holomorphically to $\mathbb{C} \times \frac{1}{2}\mathbb{Z}$ such that

$$\widehat{f}(\nu, p) = \widehat{f}(p, \nu), \qquad \nu \equiv p \pmod 1, \quad |\nu|, |p| \leqslant \ell, \tag{1.16}$$

and for any $B, C > 0$ we have

$$\widehat{f}(\nu, p) \ll_{B,C} (1 + |\nu|)^{-C}, \qquad |\Re \nu| \leqslant B, \quad p \in \tfrac{1}{2}\mathbb{Z}. \tag{1.17}$$

The Schwartz space offers a lot more flexibility in applications. A less precise result for more general groups is given in [11, Th. 3], and we refer the reader to the introduction of that paper for additional discussion and motivation of Paley–Wiener type theorems for rapidly decaying functions.

1.7. Beyond the pre-trace formula: a fourth moment

We still owe an explanation for the sub-Weyl exponent in Theorem 3(b), where $q = \pm\ell$. The proof of this bound is different from the other results: it is inspired by a brilliant recent idea of Steiner and Khayutin–Steiner [45,32] in the *weight* aspect for the groups $\mathrm{SO}_3(\mathbb{R})$ and $\mathrm{SL}_2(\mathbb{R})$. The starting point is the desire to choose the amplifier so long that it works as self-amplification. In this way, the amplifier can be made

independent of the well-known but inefficient trick of using the Hecke relation $\lambda_p^2 - \lambda_{p^2} = 1$. A self-amplified second moment is in effect a fourth moment, and the key observation is that it can be realized as the diagonal term in a *double* pre-trace formula. This only has a chance to work if the corresponding geometric side can be analyzed sufficiently accurately, and to this end, two extra features are necessary: a special behavior of spherical functions with rapid decay conditions (such as, for instance, the Bergman kernel for $\mathrm{SL}_2(\mathbb{R})$) and the possibility for a *second moment* count on the geometric side, i.e. pairs of matrices, in a best possible way.

For the proof of Theorem 3(b), we implement this idea for the first time in the context of *principal series representations*. Our proof proceeds differently than both of [45] and [32]. We avoid the theta correspondence and instead detect the diagonal term in the double pre-trace formula by an argument that is reminiscent of the Voronoi formula for Rankin–Selberg L-functions over $\mathbb{Q}[i]$, cf. §2.8. As we lose positivity, we have to use the full power of the pre-trace formula, unlike our other results where the softer pre-trace inequality suffices. The argument is analytically subtle, since we also lose the possibility to choose the test function in the pre-trace formula freely: part of it is now given to us by the gamma kernel in the Voronoi summation formula (one of several new features compared to [45] and [32]). At this point we need a very precise understanding of the Harish-Chandra transform in Theorem 7 with complete uniformity in the auxiliary complex parameters, and the reader may observe that in the end only the strong g-dependence in (1.14) saves the final bound.

1.8. Matrix counting

Having discussed some of the analytic and representation theoretic novelties, we finally comment briefly on the arithmetic part. In all previous instances of the sup-norm problem, the analysis of the geometric side of the pre-trace formula amounts to counting matrices close to K, because the elementary spherical function is bi-K-invariant and decays away from K. Given the results on spherical trace functions in §1.5, it is clear that from an arithmetic point of view the sup-norm problem with big K-types is conceptually very different from the spherical sup-norm problem.

The localization behavior of generalized spherical functions has distinct features as reflected by Theorems 4 and 6. The spherical trace function $\varphi_{\nu,\ell}^{\ell}$ concentrates close to the identity. The functions $\varphi_{\nu,\ell}^{\ell,\pm\ell}$ localize sharply around diagonal matrices (but not necessarily within K). For $\varphi_{\nu,\ell}^{\ell,0}$, there is localization on diagonal and skew-diagonal matrices within K, then there is a gradual transition to a second layer in a neighborhood of the 4-dimensional manifold $\mathcal{N}$ defined by (1.9), and outside this neighborhood we see sharp decay. Theorem 5 is in some sense a combination of these two extreme cases. Correspondingly, the counting techniques in §§6–8 are still based on the geometry of numbers, but they differ conceptually and technically from the earlier treatment of the spherical sup-norm problem. In particular, as mentioned in §1.7, for the proof of Theorem 3(b) we have to achieve a best possible double matrix count, cf. Lemma 11.

1.9. Notation

The group $G = \mathrm{SL}_2(\mathbb{C})$ and its arithmetic subgroup $\Gamma = \mathrm{SL}_2(\mathbb{Z}[i])$ are fixed throughout the paper. We use the ε-convention in that $\varepsilon > 0$ denotes a number that may be different from line to line but may in each instance be taken to be as small as desired. As usual, we write $f \ll g$ or $f = \mathrm{O}(g)$ to denote that $|f| \leqslant Cg$, where the implied constant $C > 0$ may be different from line to line; it is absolute unless otherwise indicated by a subscript, except that we occasionally allow it to depend on the (fixed) quantities I and Ω as well as on ε. We also write $f \asymp g$ for $f \ll g \ll f$, and, when used as an asymptotic notation, $f \sim g$ for $\lim f/g = 1$, where the direction of the limit is clear from the context.

2. Preliminaries

2.1. Representations of $\mathrm{SU}_2(\mathbb{C})$

In this subsection, we review the representation theory of the maximal compact subgroup

$$K = \mathrm{SU}_2(\mathbb{C}) = \left\{ k[\alpha,\beta] := \begin{pmatrix} \alpha & \beta \\ -\bar\beta & \bar\alpha \end{pmatrix} : |\alpha|^2 + |\beta|^2 = 1 \right\}$$

of $G = \mathrm{SL}_2(\mathbb{C})$. We use [33, §2.1.1 & §2.2] as a convenient reference.

For $u, v, w \in \mathbb{R}$, we parametrize K using essentially Euler angles $(2u, 2v, 2w)$ as follows:

$$k[u,v,w] := \begin{pmatrix} e^{iu} & \\ & e^{-iu} \end{pmatrix} \begin{pmatrix} \cos v & i\sin v \\ i\sin v & \cos v \end{pmatrix} \begin{pmatrix} e^{iw} & \\ & e^{-iw} \end{pmatrix}. \tag{2.1}$$

Generating an equivalence relation $\sim$ on $\mathbb{R}^3$ by

$$(u,v,w) \ \sim \ (u+2\pi, v, w),\ (u, v, w+2\pi),\ (u+\pi, v+\pi, w),\ (u+\pi/2, -v, w-\pi/2) \tag{2.2}$$

we may parametrize $\mathrm{SU}_2(\mathbb{C})$ by $\mathbb{R}^3/\sim$, or by a specific fundamental domain such as $[0,\pi)\times[0,\pi/2]\times[-\pi,\pi)$, in which each point in $\mathrm{SU}_2(\mathbb{C})$ has exactly one pre-image other than those with $v \in \frac{\pi}{2}\mathbb{Z}$. The probability Haar measure on $\mathrm{SU}_2(\mathbb{C})$ is given by

$$\mathrm{d}k = (2\pi^2)^{-1} \sin 2v \,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w. \tag{2.3}$$

The irreducible representations of $K = \mathrm{SU}_2(\mathbb{C})$ are classified as $(2\ell+1)$-dimensional representations τ_ℓ, for $\ell \in \frac{1}{2}\mathbb{Z}_{\geqslant 0}$, described explicitly as the space $V_{2\ell}$ of polynomials of degree at most 2ℓ, with a basis given by $\{z^{\ell-q} : |q| \leqslant \ell,\ q \equiv \ell \pmod 1\}$ and $\mathrm{SU}_2(\mathbb{C})$ action given by

$$\tau_\ell(k[\alpha,\beta]) z^{\ell-q} = (\alpha z - \bar\beta)^{\ell-q} (\beta z + \bar\alpha)^{\ell+q} = \sum_{\substack{|p|\leqslant \ell \\ p \equiv \ell \pmod 1}} \Phi^\ell_{p,q}(k[\alpha,\beta]) z^{\ell-p}. \tag{2.4}$$

A K-invariant scalar product on $V_{2\ell}$ is given by $(z^{\ell-q}, z^{\ell-p}) = (\ell-q)!(\ell+q)!\delta_{q=p}$, so that $\Phi^\ell_{p,q}$ are (unnormalized) matrix coefficients of τ_ℓ. Moreover,

$$\left\{ \Phi^\ell_{p,q} \,:\, p,q,\ell \in \tfrac{1}{2}\mathbb{Z} \text{ and } |p|,|q| \leqslant \ell \text{ and } p,q \equiv \ell \pmod 1 \right\}$$

is an orthogonal basis of $L^2(K)$. In harmony with [51, §4.4.2], we denote by ξ_ℓ the character of τ_ℓ, by $d_\ell = 2\ell+1$ the dimension of τ_ℓ, and by $\chi_\ell = d_\ell \xi_\ell$ the normalized character of τ_ℓ. Finally, we denote by $\widehat{K} = \{\tau_\ell : \ell \in \frac{1}{2}\mathbb{Z}_{\geqslant 0}\}$ the unitary dual of K.

2.2. Representations of $\mathrm{SL}_2(\mathbb{C})$

For compatibility with the existing literature, we shall use the Iwasawa decomposition of $G = \mathrm{SL}_2(\mathbb{C})$ in two forms, $G = NAK$ and $G = KAN$, where N (resp. A) is the subgroup of unipotent upper-triangular (resp. positive diagonal) matrices, and $K = \mathrm{SU}_2(\mathbb{C})$ is the standard maximal compact subgroup.

We fix a Haar measure on G by setting

$$\mathrm{d}g = |\mathrm{d}z| \frac{\mathrm{d}r}{r^5} \mathrm{d}k \quad \text{for } g = \begin{pmatrix} 1 & z \\ & 1 \end{pmatrix} \begin{pmatrix} r & \\ & r^{-1} \end{pmatrix} k, \quad z \in \mathbb{C},\ r > 0,\ k \in K,$$

where $|\mathrm{d}z| = \mathrm{d}x\,\mathrm{d}y$ for $z = x + iy$, $x, y \in \mathbb{R}$, and $\mathrm{d}k$ is as in (2.3).

We write $\mathfrak{a} \simeq \mathbb{R}$ for the Lie algebra of A, ρ for the root on $\mathfrak{a}$ mapping $\left(\begin{smallmatrix} x & \\ & -x \end{smallmatrix}\right)$ to $2x$, $\exp : \mathfrak{a} \to A$ for the exponential map, and $\kappa : G \to K$ and $H : G \to \mathfrak{a}$ for the projection and height maps defined by $g \in \kappa(g)\exp(H(g))N$ for every $g \in G$. Thus explicitly, for $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in G$ we have

$$\kappa(g) = \begin{pmatrix} a/\sqrt{|a|^2+|c|^2} & * \\ c/\sqrt{|a|^2+|c|^2} & * \end{pmatrix}, \qquad \exp(H(g)) = \begin{pmatrix} \sqrt{|a|^2+|c|^2} & \\ & 1/\sqrt{|a|^2+|c|^2} \end{pmatrix}. \tag{2.5}$$

Finally, let $M \simeq S^1$ be the centralizer of A in K, which consists of diagonal matrices in K.

Following [17, Ch. III], we introduce for every pair $(\nu, p) \in \mathbb{C} \times \frac{1}{2}\mathbb{Z}$ the (generalized) principal series representation $\pi_{\nu,p}$. Let us denote by $C^\infty(\mathbb{C})$ the set of functions $\mathbb{C} \to \mathbb{C}$ that are smooth when regarded as functions $\mathbb{R}^2 \to \mathbb{C}$. The representation space $V_{\nu,p}$ consists of those functions $v \in C^\infty(\mathbb{C})$ for which the transformed functions

$$\pi_{\nu,p}\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) v(z) = |bz+d|^{2p+2\nu-2}(bz+d)^{-2p} v\left(\frac{az+c}{bz+d}\right), \qquad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in G, \tag{2.6}$$

extend to elements of $C^\infty(\mathbb{C})$. The above display then actually defines the representation $\pi_{\nu,p} : G \to \mathrm{GL}(V_{\nu,p})$. The space $V_{\nu,p}$ is complete with respect to the countable family of seminorms

$$\sup\big\{\big|v^{(a,b)}(x+yi)\big| + \big|\widehat{v}^{(a,b)}(x+yi)\big| : x^2+y^2 \leqslant c\big\}, \qquad (a,b,c) \in \mathbb{N}^3,$$

where we abbreviate $\widehat{v} := \pi_{\nu,p}\left(\left(\begin{smallmatrix} & -1 \\ 1 & \end{smallmatrix}\right)\right) v$ for $v \in V_{\nu,p}$. The action of G is continuous in the topology induced by these seminorms; thus, $\pi_{\nu,p}$ is a Fréchet space representation.

Using the action of $K = \mathrm{SU}_2(\mathbb{C})$ and its diagonal subgroup $\big\{\mathrm{diag}(e^{i\varrho}, e^{-i\varrho}) : \varrho \in \mathbb{R}\big\}$, we can decompose the K-finite part of $V_{\nu,p}$ into an *algebraic direct sum* of finite-dimensional subspaces and further into one-dimensional subspaces:

$$V_{\nu,p}^{K\text{-finite}} = \bigoplus_{\substack{\ell \geqslant |p| \\ \ell \equiv p \pmod 1}} V_{\nu,p}^{\ell} = \bigoplus_{\substack{\ell \geqslant |p| \\ \ell \equiv p \pmod 1}} \bigoplus_{\substack{|q| \leqslant \ell \\ q \equiv \ell \pmod 1}} V_{\nu,p}^{\ell,q}. \tag{2.7}$$

Precisely, $V_{\nu,p}^{\ell}$ is a $(2\ell+1)$-dimensional subspace on which $\pi_{\nu,p}|_K$ acts by $\tau_\ell \in \widehat{K}$.

If $\nu \not\equiv p \pmod 1$ or $|\nu| \leqslant |p|$, then $\pi_{\nu,p} \simeq \pi_{-\nu,-p}$ is irreducible, and these are all the equivalences among the representations $\pi_{\nu,p}$. If $\nu \equiv p \pmod 1$ and $|\nu| > |p|$, then $\pi_{\nu,p}$ and $\pi_{-\nu,-p}$ are reducible. Assume $\nu > 0$, say. Then the sum of $V_{\nu,p}^{\ell}$ with $|p| \leqslant \ell < \nu$ is a closed invariant subspace of $V_{\nu,p}$, and the representation induced on the quotient is irreducible. The closure of the sum of $V_{-\nu,-p}^{\ell}$ with $\ell \geqslant \nu$ is an invariant subspace of $V_{-\nu,-p}$, and the representation induced on it is irreducible. Both of these representations of G are isomorphic to $\pi_{p,\nu} \simeq \pi_{-p,-\nu}$. This observation will become relevant in (2.21) below.

The space $V_{\nu,p}$ has a G-invariant Hermitian inner product if and only if $\nu \in i\mathbb{R}$, or $p = 0$ and $\nu \in (-1,0) \cup (0,1)$. In the first case, we say that $\pi_{\nu,p}$ belongs to the (tempered) unitary principal series. In the second case, we say that $\pi_{\nu,p}$ belongs to the (non-tempered) complementary series. In either case, the Fréchet space representation $\pi_{\nu,p}$ induces an irreducible unitary representation on the Hilbert space completion $\widehat{V_{\nu,p}}$ that we shall still denote by $\pi_{\nu,p}$. The only equivalences among these unitary representations are $\pi_{\nu,p} \simeq \pi_{-\nu,-p}$. The equivalence classes, along with the trivial representation, form the unitary dual $\widehat{G}$ of G.

For $\pi \simeq \pi_{\nu,p} \in \widehat{G}$ we write

$$V_\pi := \widehat{V_{\nu,p}}, \qquad V_\pi^{\ell} := V_{\nu,p}^{\ell}, \qquad V_\pi^{\ell,q} := V_{\nu,p}^{\ell,q},$$

and then (2.7) is equivalent to the orthogonal Hilbert space decomposition (cf. (1.2)):

$$V_\pi = \bigoplus_{\substack{\ell \geqslant |p| \\ \ell \equiv p \pmod 1}} V_\pi^\ell = \bigoplus_{\substack{\ell \geqslant |p| \\ \ell \equiv p \pmod 1}} \bigoplus_{\substack{|q| \leqslant \ell \\ q \equiv \ell \pmod 1}} V_\pi^{\ell,q}.$$

The projection $V_\pi \to V_\pi^\ell$ is realized by the operator

$$\pi(\overline{\chi_\ell}) := \int_K \overline{\chi_\ell}(k)\pi(k)\,\mathrm{d}k \in \mathrm{End}(V_\pi), \tag{2.8}$$

where $\mathrm{End}(V_\pi)$ denotes the Hilbert space of Hilbert–Schmidt operators on V_π endowed with the Hilbert–Schmidt norm. This leads to the "block matrix decomposition"

$$\mathrm{End}(V_\pi) = \bigoplus_{\substack{m,n \geqslant |p| \\ m,n \equiv p \pmod 1}} \mathrm{Hom}(V_\pi^m, V_\pi^n), \tag{2.9}$$

where the direct sum is meant in the Hilbert space sense. Hence, for $f \in C_c(G)$, the (m,n)-component of the Hilbert–Schmidt operator (cf. [19, Th. 2])

$$\pi(f) := \int_G f(g)\pi(g)\,\mathrm{d}g \in \mathrm{End}(V_\pi) \tag{2.10}$$

equals

$$\pi(\overline{\chi_n})\pi(f)\pi(\overline{\chi_m}) = \pi(\overline{\chi_n} \star f \star \overline{\chi_m}) \in \mathrm{Hom}(V_\pi^m, V_\pi^n), \tag{2.11}$$

where the convolutions are meant over K.

2.3. Plancherel theorem

In this subsection, we review the Plancherel theorem for $G = \mathrm{SL}_2(\mathbb{C})$ pioneered by Gelfand and Naimark, following the original sources [19,20] and their translations [22,21]. We note that the list of unitary representations given in [20] is incomplete for higher rank groups (cf. [44,49,46]), but this does not affect the results we are quoting. In addition, we warn the reader that the translations contain some misprints not present in the originals, e.g. in the crucial formulae [22, (137)–(138)].

We identify once and for all (non-canonically) the *tempered unitary dual* $\widehat{G}_{\mathrm{temp}}$ with the set

$$\left\{\pi_{it,p} : (t,p) \in \left(\mathbb{R}_{>0} \times \tfrac{1}{2}\mathbb{Z}\right) \cup \left(\{0\} \times \tfrac{1}{2}\mathbb{Z}_{\geqslant 0}\right)\right\},$$

with topology inherited from the standard topology on $\mathbb{R}^2$. The *Plancherel measure* on $\widehat{G}$ is supported on $\widehat{G}_{\mathrm{temp}}$, and it is given explicitly as

$$\mathrm{d}\mu_{\mathrm{Pl}}(\pi_{it,p}) := \frac{1}{\pi^2}(t^2+p^2)\,\mathrm{d}t\,\mathrm{d}p, \tag{2.12}$$

with $\mathrm{d}t$ the Lebesgue measure on $\mathbb{R}_{\geqslant 0}$ and $\mathrm{d}p$ the counting measure on $\frac{1}{2}\mathbb{Z}$. For $\pi_{it,p} \in \widehat{G}_{\mathrm{temp}}$, the underlying Hilbert space $\widehat{V_{it,p}}$ is independent of the parameters: it equals $\mathcal{V} := L^2(\mathbb{C})$. On this common representation space, (2.6) defines the unitary action $\pi_{it,p} : G \to \mathrm{U}(\mathcal{V})$ that agrees with [19, (65)] for $(n,\rho) = (2p, 2t)$. The *operator-valued spherical transform* of $f \in C_c(G)$ is the map $\widehat{G}_{\mathrm{temp}} \to \mathrm{End}(\mathcal{V})$ given by $\pi \mapsto \pi(f)$ as in

(2.10). The Plancherel theorem for G concerns the extension of this transform to $L^2(G)$, and characterizes its image.

Theorem 8 *(Gelfand–Naimark). The map given by (2.10) extends (uniquely) to an L^2-isometry*

$$L^2(G) \longrightarrow L^2(\widehat{G}_{\mathrm{temp}} \to \mathrm{End}(\mathcal{V})),$$

where the operator-valued L^2-space on the right-hand side is meant with respect to the Hilbert–Schmidt norm $\|\cdot\|_{\mathrm{HS}}$ on $\mathrm{End}(\mathcal{V})$ and the Plancherel measure μ_{Pl} on $\widehat{G}_{\mathrm{temp}}$. In particular, for every $f \in L^2(G)$, the following Plancherel formula holds:

$$\int_G |f(g)|^2 \,\mathrm{d}g = \int_{\widehat{G}_{\mathrm{temp}}} \|\pi(f)\|_{\mathrm{HS}}^2 \,\mathrm{d}\mu_{\mathrm{Pl}}(\pi). \tag{2.13}$$

Proof. The theorem follows from [19, Th. 5]; we only need to check that our Plancherel measure corresponds to the one in [19, (137)]. We do this in four steps.

Step 1. We observe that the constant $(8\pi^4)^{-1}$ in [19, (137)] should be $(16\pi^4)^{-1}$ due to a small oversight in the derivation of [19, (130)] from [19, (129)]. The oversight is that the change of variables

$$(w_1, w_2, \lambda) \mapsto (\zeta_1, \zeta_2, \zeta_3) := (w_2, w_1\bar{\lambda} + w_2/\bar{\lambda}, w_1)$$

coming from [19, (123)] is not 1-to-1 but 2-to-1.

Step 2. We rewrite the corrected right-hand side of [19, (137)] as a sum over $p \in \frac{1}{2}\mathbb{Z}$ and an integral over $t > 0$, keeping in mind that (n, ρ) in [19] is $(2p, 2t)$ in our notation.

Step 3. We observe that the Haar measure $\mathrm{d}\mu(g)$ used by Gelfand–Naimark is $2\pi^2\mathrm{d}g$. Indeed, applying [19, (40)] to a right K-invariant test function $f \in C_c(G)$, we obtain by several changes of variables that

$$\begin{aligned}
\int_G f(g)\,\mathrm{d}\mu(g) &= \int_{\mathbb{C}\times\mathbb{C}^\times\times\mathbb{C}} f\left(\begin{pmatrix} w^{-1} & z \\ & w \end{pmatrix}\begin{pmatrix} 1 & \\ v & 1 \end{pmatrix}\right) |\mathrm{d}v|\,|\mathrm{d}w|\,|\mathrm{d}z| \\
&= \int_{\mathbb{C}\times\mathbb{C}^\times\times\mathbb{C}} f\left(\begin{pmatrix} w^{-1} & z \\ & w \end{pmatrix}\begin{pmatrix} 1/\sqrt{1+|v|^2} & \bar{v}/\sqrt{1+|v|^2} \\ & \sqrt{1+|v|^2} \end{pmatrix}\right) |\mathrm{d}v|\,|\mathrm{d}w|\,|\mathrm{d}z| \\
&= \int_{\mathbb{C}\times\mathbb{C}^\times\times\mathbb{C}} f\left(\begin{pmatrix} w^{-1} & z \\ & w \end{pmatrix}\right) \frac{|\mathrm{d}v|\,|\mathrm{d}w|\,|\mathrm{d}z|}{(1+|v|^2)^2} \\
&= \pi \int_{\mathbb{C}^\times\times\mathbb{C}} f\left(\begin{pmatrix} 1 & z \\ & 1 \end{pmatrix}\begin{pmatrix} w & \\ & w^{-1} \end{pmatrix}\right) \frac{|\mathrm{d}w|\,|\mathrm{d}z|}{|w|^6} = 2\pi^2 \int_G f(g)\,\mathrm{d}g.
\end{aligned}$$

Step 4. Putting everything together, the corrected version of [19, (137)] yields

$$\int_G |f(g)|^2\, 2\pi^2 \mathrm{d}g = \frac{1}{16\pi^4} \sum_p \int_0^\infty \|2\pi^2 \pi_{it,p}(f)\|_{\mathrm{HS}}^2 \,(4t^2 + 4p^2)\, 2\mathrm{d}t.$$

This formula is equivalent to (2.13), hence we are done. □

Remark 4. In the proof above, we claimed that the Plancherel measure in [19, Th. 5] is off by a factor of 2. For double checking this claim, we looked at [31, Th. 11.2], and we found (to our dismay) that the

Plancherel measure there is off by a factor of π. For example, for the test function $f(g) := 1/\operatorname{tr}(gg^*)^2$, the Fourier transform given by [31, (11.14)] equals $F_f^T(t) = \pi/\operatorname{tr}(tt^*)$, hence in [31, (11.17)] the left-hand side is π^2, while the right-hand side is π. For triple checking our claim, we verified that our Plancherel measure yields the correct inversion formula for the classical spherical transform (for bi-K-invariant functions), as in [15, §3.3].

Theorem 9 *(Gelfand–Naimark). Let $f \in C_c^\infty(G)$. For every $\pi \in \widehat{G}_{\mathrm{temp}}$, the operator $\pi(f) \in \mathrm{End}(\mathcal{V})$ is of trace class, and the following inversion formula holds:*

$$f(g) = \int_{\widehat{G}_{\mathrm{temp}}} \operatorname{tr}(\pi(f)\pi(g^{-1}))\, \mathrm{d}\mu_{\mathrm{Pl}}(\pi). \tag{2.14}$$

Proof. The theorem follows from [20, Th. 19] applied to $n = 2$ and $x = R(g)f$, or from [31, Th. 11.2], with appropriate correction of the Plancherel measure (cf. Remark 4). □

Remark 5. By a celebrated result of Dixmier–Malliavin [12], every $f \in C_c^\infty(G)$ can be written as a linear combination of convolutions $w \star w^*$, where $w \in C_c^\infty(G)$ and $w^*(g) := \overline{w(g^{-1})}$. Hence Theorem 9 also follows from Theorem 8 and [19, Th. 2]. In fact for this implication we only need that $w \in C_c(G)$, which is easier to achieve.

2.4. The τ_ℓ-spherical transform

For a given $\ell \in \frac{1}{2}\mathbb{Z}_{\geqslant 0}$, it is interesting to see what Theorems 8 and 9 yield for test functions $f \in L^2(G)$ with the following property: for almost every $\pi \in \widehat{G}_{\mathrm{temp}}$, the operator $\pi(f)$ acts by a scalar on V_π^ℓ and by zero on its orthocomplement $V_\pi^{\ell,\perp}$. In the light of (2.9), (2.11), (2.13), and Schur's lemma, these test functions form the Hilbert subspace $\mathcal{H}(\tau_\ell) \subset L^2(G)$ defined by the conditions

- $f(g) = f(kgk^{-1})$ for almost every $g \in G$ and $k \in K$;
- $f = \overline{\chi_\ell} \star f \star \overline{\chi_\ell}$.

Let $\widehat{G}_{\mathrm{temp}}(\tau_\ell)$ be the set of $\pi \in \widehat{G}_{\mathrm{temp}}$ whose restriction to K contains τ_ℓ. For $f \in \mathcal{H}(\tau_\ell)$, the operator-valued function $\pi \mapsto \pi(f)$ is supported on $\widehat{G}_{\mathrm{temp}}(\tau_\ell)$, and there it is simply determined by the scalar-valued function $\pi \mapsto \operatorname{tr}(\pi(f))$ via

$$\pi(f)|_{V_\pi^\ell} = \frac{\operatorname{tr}(\pi(f))}{2\ell+1} \cdot \mathrm{id}_{V_\pi^\ell} \qquad \text{and} \qquad \pi(f)|_{V_\pi^{\ell,\perp}} = 0. \tag{2.15}$$

In particular, for $\pi \in \widehat{G}_{\mathrm{temp}}(\tau_\ell)$ and $f \in \mathcal{H}(\tau_\ell)$,

$$\|\pi(f)\|_{\mathrm{HS}}^2 = \operatorname{tr}(\pi(f)\pi(f)^*) = \frac{|\operatorname{tr}(\pi(f))|^2}{2\ell+1}. \tag{2.16}$$

For $(\nu, p) \in i\mathbb{R} \times \frac{1}{2}\mathbb{Z}$, the condition $\pi_{\nu,p} \in \widehat{G}_{\mathrm{temp}}(\tau_\ell)$ is equivalent to $|p| \leqslant \ell$ and $p \equiv \ell \pmod 1$. Moreover, for $f \in L^1(G) \cap \mathcal{H}(\tau_\ell)$, the trace of $\pi_{\nu,p}(f)$ can be expressed in terms of the *τ_ℓ-spherical trace function*

$$\begin{aligned} \varphi_{\nu,p}^\ell(g) &:= \operatorname{tr}(\pi_{\nu,p}(\overline{\chi_\ell})\pi_{\nu,p}(g)\pi_{\nu,p}(\overline{\chi_\ell})) \\ &= \operatorname{tr}(\pi_{\nu,p}(\overline{\chi_\ell})\pi_{\nu,p}(g)) = \operatorname{tr}(\pi_{\nu,p}(g)\pi_{\nu,p}(\overline{\chi_\ell})) \end{aligned} \tag{2.17}$$

as (cf. (2.10) and (2.11))

$$\widehat{f}(\nu,p) := \mathrm{tr}(\pi_{\nu,p}(f)) = \int_G f(g)\,\varphi^\ell_{\nu,p}(g)\,\mathrm{d}g. \tag{2.18}$$

The function $\varphi^\ell_{\nu,p} : G \to \mathbb{C}$ vanishes unless $|p| \leqslant \ell$ and $p \equiv \ell \pmod 1$, for else τ_ℓ does not appear in $\pi_{\nu,p}$, and $\varphi^\ell_{\nu,p}(\mathrm{id}) = 2\ell+1$ in this latter case. Moreover, we have the integral representation of Harish-Chandra [52, Cor. 6.2.2.3]:

$$\varphi^\ell_{\nu,p}(g) = \int_K (\chi_\ell \star \eta_p)\,(\kappa(k^{-1}gk))\,e^{(\nu-1)\rho(H(gk))}\,\mathrm{d}k.$$

Here, $\eta_p : M \simeq S^1 \to \mathbb{C}^\times$ is the unitary character $\eta_p(z) = z^{-2p}$, the convolution is over M, and κ, ρ, and H are as in §2.2. For computational purposes, we spell out the $\chi_\ell \star \eta_p$ term explicitly, cf. (2.4), [50, (10) & Lemma 3.2], [25, Th. 29.18]:

$$\begin{aligned}(\chi_\ell \star \eta_p)\,(k[\alpha,\beta]) &= (2\ell+1)\Phi^\ell_{p,p}(k[\alpha,\beta]) \\ &= (2\ell+1)\sum_{r=0}^{\ell-|p|}(-1)^r\binom{\ell+p}{r}\binom{\ell-p}{r}\alpha^{\ell-p-r}\bar\alpha^{\ell+p-r}|\beta|^{2r}.\end{aligned}$$

We collect further useful properties of $\varphi^\ell_{\nu,p} : G \to \mathbb{C}$ in the next lemma, where we write

$$a_h := \mathrm{diag}(e^{h/2}, e^{-h/2}), \qquad h \in \mathbb{R}.$$

Lemma 1. *The τ_ℓ-spherical trace function $\varphi^\ell_{\nu,p}(g)$ extends holomorphically to $\nu \in \mathbb{C}$, and it satisfies the bound*

$$\left|\varphi^\ell_{\sigma+it,p}(k_1 a_h k_2)\right| \leqslant (2\ell+1)\frac{\sinh(\sigma h)}{\sigma\sinh(h)}, \qquad \sigma,t,h \in \mathbb{R}, \quad k_1,k_2 \in K. \tag{2.19}$$

(For $\sigma = 0$ or $h = 0$, the fraction on the right-hand side is understood as 1.) The extended function has the symmetries

$$\varphi^\ell_{\nu,p}(g) = \overline{\varphi^\ell_{-\overline{\nu},p}(g)} = \varphi^\ell_{\nu,p}(g^{-1}), \tag{2.20}$$

$$\varphi^\ell_{\nu,p}(g) = \varphi^\ell_{p,\nu}(g), \qquad \nu \equiv p \pmod 1, \quad |\nu|,|p| \leqslant \ell. \tag{2.21}$$

Proof. The holomorphic extension of $\varphi^\ell_{\nu,p}(g)$ and the bound (2.19) are a straightforward generalization of [50, Prop. 3.4] and its proof. The identity $\overline{\varphi^\ell_{-\overline{\nu},p}(g)} = \varphi^\ell_{\nu,p}(g^{-1})$ follows from (2.17) and $\pi(g)^* = \pi(g^{-1})$ for $\nu \in i\mathbb{R}$, and then also for $\nu \in \mathbb{C}$ by the uniqueness of analytic continuation. The identity $\varphi^\ell_{\nu,p}(g) = \varphi^\ell_{\nu,p}(g^{-1})$ is [50, Lemma 3.2], keeping in mind that $\pi_{\nu,p} \simeq \pi_{-\nu,-p}$ for $\nu \in i\mathbb{R}$ and again invoking analytic continuation. Finally, the remarkable symmetry (2.21) follows from [50, Cor. 2], or more conceptually from the discussion below (2.7). □

As we shall see in Theorem 10 below, the *τ_ℓ-spherical transform* defined by (2.18) is inverted by the following *inverse τ_ℓ-spherical transform*. For $h \in L^1(\widehat{G}_{\mathrm{temp}}(\tau_\ell)) \cap L^2(\widehat{G}_{\mathrm{temp}}(\tau_\ell))$ and $g \in G$, we define

$$\check{h}(g) := \frac{1}{(2\ell+1)\pi^2} \sum_{\substack{|p|\leqslant\ell \\ p\equiv\ell \pmod 1}} \int_0^\infty h(it,p)\,\varphi^\ell_{it,p}(g^{-1})\,(t^2+p^2)\,\mathrm{d}t. \tag{2.22}$$

Theorem 10. *The transforms defined by (2.18) and (2.22) extend (uniquely) to a pair of Hilbert space isometries inverse to each other:*

$$\mathcal{H}(\tau_\ell) \longleftrightarrow L^2(\widehat{G}_{\mathrm{temp}}(\tau_\ell)).$$

In particular, for $f \in \mathcal{H}(\tau_\ell)$, the following Plancherel formula holds:

$$\int_G |f(g)|^2 \,\mathrm{d}g = \frac{1}{(2\ell+1)\pi^2} \sum_{\substack{|p|\leqslant \ell \\ p\equiv \ell \pmod 1}} \int_0^\infty |\widehat{f}(it,p)|^2 \,(t^2+p^2)\,\mathrm{d}t. \tag{2.23}$$

Proof. The fact that $\widehat{\ }$ extends to a Hilbert space isomorphism $\mathcal{H}(\tau_\ell) \to L^2(\widehat{G}_{\mathrm{temp}}(\tau_\ell))$ follows from Theorem 8 and our discussion above. In particular, (2.23) is a special case of (2.13) in the light of (2.12), (2.16), (2.18). We are left with proving that $\check{\ }$ is the inverse of $\widehat{\ }$, and for this it suffices to verify that $\check{\ }$ applied after $\widehat{\ }$ is the identity on the dense subset $C_c^\infty(G) \cap \mathcal{H}(\tau_\ell)$ of the Hilbert space $\mathcal{H}(\tau_\ell)$. For $f \in C_c^\infty(G) \cap \mathcal{H}(\tau_\ell)$, (2.8), (2.10), (2.12), (2.14), (2.15), (2.17), (2.18) yield

$$\begin{aligned} f(g) &= \int_{\widehat{G}_{\mathrm{temp}}} \mathrm{tr}(\pi(f)\pi(g^{-1}))\,\mathrm{d}\mu_{\mathrm{Pl}} = \frac{1}{2\ell+1} \int_{\widehat{G}_{\mathrm{temp}}} \mathrm{tr}(\pi(f))\,\mathrm{tr}(\pi(\overline{\chi_\ell})\pi(g^{-1}))\,\mathrm{d}\mu_{\mathrm{Pl}} \\ &= \frac{1}{(2\ell+1)\pi^2} \sum_{\substack{|p|\leqslant \ell \\ p\equiv \ell \pmod 1}} \int_0^\infty \widehat{f}(it,p)\,\varphi_{it,p}^\ell(g^{-1})\,(t^2+p^2)\,\mathrm{d}t. \end{aligned}$$

The proof is complete. □

Wang [50] proved an analogue of the Paley–Wiener theorem for the τ_ℓ-spherical transform, and in particular characterized the image of $\mathcal{H}(\tau_\ell) \cap C_c^\infty(G)$ under the transform. The following is [50, Prop. 4.5] and should be compared to Theorem 7 in the introduction.

Theorem 11 *(Wang). Let $f \in \mathcal{H}(\tau_\ell)$ be a test function, and let $R > 0$. Then the following two conditions are equivalent.*

(a) The function $f(g)$ is smooth, and

$$f(k_1 a_h k_2) = 0, \qquad |h| > R, \quad k_1, k_2 \in K.$$

(b) The function $\widehat{f}(\nu,p)$ has a holomorphic extension to $\mathbb{C} \times \frac{1}{2}\mathbb{Z}$ such that

$$\widehat{f}(\nu,p) = \widehat{f}(p,\nu), \qquad \nu \equiv p \pmod 1, \quad |\nu|, |p| \leqslant \ell,$$

and for any $C > 0$ we have

$$\widehat{f}(\nu,p) \ll_C (1+|\nu|)^{-C} e^{R|\Re\nu|}, \qquad \nu \in \mathbb{C}, \quad p \in \tfrac{1}{2}\mathbb{Z}.$$

We now prove a Schwartz class version of this result as stated in Theorem 7.

Proof of Theorem 7. For harmony of notation with [50], in this proof we use $D_{p,q}^\ell(k)$ to denote the matrix coefficients of τ_ℓ relative to the basis obtained by normalizing the orthogonal basis $\{z^{\ell-q} : |q| \leqslant \ell, q \equiv \ell \pmod 1\}$ in the space $V_{2\ell}$ of §2.1. Thus we explicitly have the renormalization

$$D^{\ell}_{p,q}(k) = \left(\frac{(\ell-p)!(\ell+p)!}{(\ell-q)!(\ell+q)!}\right)^{1/2} \Phi^{\ell}_{p,q}(k).$$

Assume condition (a). The holomorphic extension of $\widehat{f}(\nu,p)$ follows from (2.19) coupled with (1.15) for $m=0$, and then (1.16) is immediate from (2.21). In order to derive (1.17), we use an alternate representation of $\widehat{f}(\nu,p)$. We shall assume that $|p| \leqslant \ell$ and $p \equiv \ell \pmod 1$, for else $\widehat{f}(\nu,p)=0$. By the third line of the second display on [50, p. 621] and [50, Lemma 3.2], we see that the (unique) holomorphic extension is also provided by

$$\widehat{f}(\nu,p) = \frac{2\ell+1}{2} \int_{-\infty}^{\infty} \breve{f}(h,p)\, e^{\nu h}\, \mathrm{d}h, \tag{2.24}$$

where

$$\breve{f}(h,p) := e^{h} \int_K \int_N f(ka_h n)\, D^{\ell}_{p,p}(k)\, \mathrm{d}k\, \mathrm{d}n. \tag{2.25}$$

We claim that, for any $m \in \mathbb{Z}_{\geqslant 0}$ and $A>0$, we have

$$\frac{\partial^m}{\partial h^m} \breve{f}(h,p) \ll_{m,A} e^{-A|h|}, \qquad h \in \mathbb{R}, \quad |p| \leqslant \ell, \quad p \equiv \ell \pmod 1. \tag{2.26}$$

For $|h|>1$ this follows by writing $a_h n = k_1 a_{h'} k_2$ in (2.25), and then combining (1.15) with some calculus to keep track of the dependence of $h' \in \mathbb{R}$ and $k_1, k_2 \in K$ on $h \in \mathbb{R}$. For $|h| \leqslant 1$ we proceed similarly for the part of the integral in (2.25) that corresponds to $n = \left(\begin{smallmatrix} 1 & z \\ & 1 \end{smallmatrix}\right)$ with $|z|>1$, while we estimate the (h-derivatives of the) remaining integral directly by the smoothness of $f(g)$. With (2.26) at hand, (1.17) follows from (2.24) via integration by parts. We proved that (a) implies (b).

Assume condition (b). By Theorem 10,

$$f(g) = \frac{1}{(2\ell+1)\pi^2} \sum_{\substack{|p|\leqslant \ell \\ p \equiv \ell \pmod 1}} \int_0^{\infty} \widehat{f}(it,p)\, \varphi^{\ell}_{it,p}(g^{-1})\, (t^2+p^2)\, \mathrm{d}t.$$

Let us restrict, without loss of generality, to $g = k_1 a_h k_2$ with $h>0$. Using the display below [50, (29)],[1] we infer

$$f(g) = \frac{1}{4\pi^2 \sinh(h)} \sum_{\substack{|p|,|j|\leqslant \ell \\ p,j \equiv \ell \pmod 1}} \int_{-h}^{h} \widetilde{f}(s,p)\, D^{\ell}_{-p,j}(v_\theta^{-1}) D^{\ell}_{j,j}(k_2 k_1) D^{\ell}_{j,-p}(v_{\theta'})\, \mathrm{d}s,$$

where $D^{\ell}_{-p,j}(v_\theta^{-1})$ and $D^{\ell}_{j,-p}(v_{\theta'})$ can be explicated using [50, (5) & (28)], and

$$\widetilde{f}(s,p) := \int_{-\infty}^{\infty} \widehat{f}(it,p)\, e^{-its}\, (t^2+p^2)\, \mathrm{d}t, \qquad s \in \mathbb{R}. \tag{2.27}$$

[1] We note that in [50, (29)] the product $k_2 k_1$ should be conjugated as $u_{\varphi_1}^{-1} k_2 k_1 u_{\varphi_1}$, and the integral over $0 \leqslant \varphi_1 \leqslant 2\pi$ with normalization factor $1/(2\pi)$ is missing. After this correction, the crucial next display follows as stated, by expanding the matrix coefficient $D^{\ell}_{-p,-p}$ (in our notation) via the entry-by-entry product of three matrices and executing the φ_1-integral.

By (1.17) and Cauchy's theorem, it follows for any $n \in \mathbb{Z}_{\geqslant 0}$ and $D > 0$ that

$$\frac{\partial^n}{\partial s^n} \widetilde{f}(s,p) \ll_{n,D} e^{-D|s|}, \qquad s \in \mathbb{R}. \tag{2.28}$$

The smoothness of $f(g)$ is now straightforward, and this automatically verifies (1.15) for $|h| \leqslant 1$. From now on we can assume, without loss of generality, that $h > 1$. From (1.16), (2.28), and the calculation around [50, (38)–(41)], we see that

$$\sum_{\substack{|p|,|j| \leqslant \ell \\ p,j \equiv \ell \pmod 1}} \int_{-\infty}^{\infty} \widetilde{f}(s,p)\, D^{\ell}_{-p,j}(v_\theta^{-1}) D^{\ell}_{j,j}(k_2 k_1) D^{\ell}_{j,-p}(v_{\theta'})\, \mathrm{d}s = 0,$$

hence in fact

$$f(g) = \frac{-1}{4\pi^2 \sinh(h)} \sum_{\substack{|p|,|j| \leqslant \ell \\ p,j \equiv \ell \pmod 1}} \left(\int_{-\infty}^{-h} + \int_{h}^{\infty} \right) \widetilde{f}(s,p)\, D^{\ell}_{-p,j}(v_\theta^{-1}) D^{\ell}_{j,j}(k_2 k_1) D^{\ell}_{j,-p}(v_{\theta'})\, \mathrm{d}s.$$

From here it is straightforward to deduce (1.15) for $h > 1$, using (2.28) and the remarks above it. We proved that (b) implies (a). □

We shall denote by $\mathcal{H}(\tau_\ell)_\infty$ the set of functions satisfying the equivalent conditions (a) and (b) of Theorem 7. It is clear that $\mathcal{H}(\tau_\ell)_\infty$ is a convolution subalgebra of $L^1(G) \cap L^2(G)$.

Remark 6. In the previous display, we may estimate the product of the three matrix coefficients (recalling that each matrix $(D^{\ell}_{p,q}(k))_{p,q}$ is orthogonal) using the trivial bound $|D^{\ell}_{j,j}| \leqslant 1$ and the Cauchy–Schwarz inequality for the remaining two factors. Combining this with the observation $\widetilde{f}(s,p) = \widetilde{f}(-s,-p)$ yields the following refinement of (1.15) when $m = 0$:

$$\big|f(k_1 a_h k_2)\big| \leqslant \sum_{\substack{|p| \leqslant \ell \\ p \equiv \ell \pmod 1}} \int_{h}^{\infty} \big|\widetilde{f}(s,p)\big|\, \mathrm{d}s, \qquad h > 1, \quad k_1, k_2 \in K. \tag{2.29}$$

We end this subsection by stating a two-variable version of some of the previous definitions and results. Taking (topological) tensor products of Hilbert spaces, we can identify $\mathcal{H}(\tau_\ell) \hat{\otimes} \mathcal{H}(\tau_\ell)$ with the space of functions $f \in L^2(G \times G)$ satisfying

- $f(g_1, g_2) = f(k_1 g_1 k_1^{-1}, k_2 g_2 k_2^{-1})$ for almost every $g_1, g_2 \in G$ and $k_1, k_2 \in K$;
- $f = (\overline{\chi_\ell}, \overline{\chi_\ell}) \star f \star (\overline{\chi_\ell}, \overline{\chi_\ell})$ almost everywhere.

This can be seen by projecting the isomorphism between $L^2(G) \hat{\otimes} L^2(G)$ and $L^2(G \times G)$ (see e.g. [43, Cor. 4.11.9]) to $\mathcal{H}(\tau_\ell) \hat{\otimes} \mathcal{H}(\tau_\ell)$ and the (closed) subspace of functions in question. By Theorem 10, this space is isometrically isomorphic to $L^2(\widehat{G}_{\mathrm{temp}}(\tau_\ell)^2)$ via the obvious extension of the map (2.18):

$$\widehat{f}(\nu_1, p_1, \nu_2, p_2) := \int_{G_1 \times G_2} f(g_1, g_2)\, \varphi^{\ell}_{\nu_1, p_1}(g_1) \varphi^{\ell}_{\nu_2, p_2}(g_2)\, \mathrm{d}g_1\, \mathrm{d}g_2. \tag{2.30}$$

For $h \in L^1(\widehat{G}_{\mathrm{temp}}(\tau_\ell)^2) \cap L^2(\widehat{G}_{\mathrm{temp}}(\tau_\ell)^2)$, the inverse transform is given as in (2.22):

$$\breve{h}(g_1,g_2) := \frac{1}{(2\ell+1)^2\pi^4} \sum_{\substack{|p_1|,|p_2|\leqslant \ell \\ p_1\equiv p_2\equiv \ell \pmod 1}} \int_0^\infty\int_0^\infty h(it_1,p_1,it_2,p_2) \varphi^\ell_{it_1,p_1}(g_1^{-1})\varphi^\ell_{it_2,p_2}(g_2^{-1})\,(t_1^2+p_1^2)(t_2^2+p_2^2)\,\mathrm{d}t_1\,\mathrm{d}t_2. \tag{2.31}$$

It is straightforward to adapt the above presented proof of Theorem 7 to obtain the following variant for $\mathcal{H}(\tau_\ell)\hat{\otimes}\mathcal{H}(\tau_\ell)$:

Theorem 12. *For $f \in \mathcal{H}(\tau_\ell)\hat{\otimes}\mathcal{H}(\tau_\ell)$, the following two conditions are equivalent (with implied constants depending on f).*

(a) The function $f(g_1,g_2)$ is smooth, and for any $m_1,m_2\in\mathbb{Z}_{\geqslant 0}$ and $A>0$ we have

$$\frac{\partial^{m_1+m_2}}{\partial h_1^{m_1}\partial h_2^{m_2}} f(k_1a_{h_1}k_2,k_3a_{h_2}k_4) \ll_{m_1,m_2,A} e^{-A(|h_1|+|h_2|)}, \quad h_1,h_2\in\mathbb{R},\ k_1,k_2,k_3,k_4\in K.$$

(b) The function $\widehat{f}(\nu_1,p_1,\nu_2,p_2)$ has a holomorphic extension to $\mathbb{C}\times\frac12\mathbb{Z}\times\mathbb{C}\times\frac12\mathbb{Z}$ such that

$$\widehat{f}(\nu_1,p_1,\nu_2,p_2) = \widehat{f}(p_1,\nu_1,\nu_2,p_2), \qquad \nu_1\equiv p_1 \pmod 1, \quad |\nu_1|,|p_1|\leqslant\ell,$$
$$\widehat{f}(\nu_1,p_1,\nu_2,p_2) = \widehat{f}(\nu_1,p_1,p_2,\nu_2), \qquad \nu_2\equiv p_2 \pmod 1, \quad |\nu_2|,|p_2|\leqslant\ell,$$

and for any $B,C>0$ we have

$$\widehat{f}(\nu_1,p_1,\nu_2,p_2) \ll_{B,C} (1+|\nu_1|+|\nu_2|)^{-C}, \qquad |\Re\nu_1|,|\Re\nu_2|\leqslant B, \quad p_1,p_2\in\tfrac12\mathbb{Z}.$$

We shall denote by $\mathcal{H}(\tau_\ell,\tau_\ell)_\infty$ the set of functions satisfying the equivalent conditions (a) and (b) of Theorem 12; this is clearly a convolution subalgebra of $L^1(G\times G)\cap L^2(G\times G)$.

2.5. Hecke operators

The arithmetic quotient $\Gamma\backslash G$ comes equipped with a rich family of Hecke correspondences, which we now describe, referring to [2] for further details and references. For every $n\in\mathbb{Z}[i]\setminus\{0\}$, consider the set

$$\Gamma_n := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{M}_2(\mathbb{Z}[i]) : ad-bc = n \right\}.$$

In particular, $\Gamma_1=\Gamma$. Then we may define the Hecke operator T_n acting on functions $\phi:\Gamma\backslash G\to\mathbb{C}$ by

$$(T_n\phi)(g) := \frac{1}{|n|}\sum_{\gamma\in\Gamma\backslash\Gamma_n} \phi\left(\frac{1}{\sqrt{n}}\gamma g\right) = \frac{1}{4|n|}\sum_{ad=n}\sum_{b \bmod d} \phi\left(\frac{1}{\sqrt{n}}\begin{pmatrix} a & b \\ 0 & d\end{pmatrix} g\right), \tag{2.32}$$

where the result is independent of the choice of the square-root since $\pm\mathrm{id}\in\Gamma$. In particular, since $\Gamma_{-1}=\Gamma\cdot\left(\begin{smallmatrix}-1&\\&1\end{smallmatrix}\right)$ and $\frac1i\left(\begin{smallmatrix}-1&\\&1\end{smallmatrix}\right)=\left(\begin{smallmatrix}i&\\&-i\end{smallmatrix}\right)\in\Gamma$, we have $T_{-1}=T_1=\mathrm{id}$. We also observe that, as γ ranges through a set of representatives of $\Gamma\backslash\Gamma_n$, $n\gamma^{-1}$ ranges through a set of representatives of Γ_n/Γ.

These Hecke operators are self-adjoint on $L^2(\Gamma\backslash G)$, commute with each other and the Laplace operator; thus they act by constants $\lambda_n(V)$ on each irreducible component $V\subset L^2(\Gamma\backslash G)$, with non-zero vectors in each V being joint Hecke–Maaß eigenfunctions. They also satisfy the multiplicativity relation

$$T_mT_n = \sum_{(d)|(m,n)} T_{mn/d^2}, \qquad m,n\in\mathbb{Z}[i]\setminus\{0\}, \tag{2.33}$$

where it is clear that the right-hand side does not depend on the choice of the generator d. Finally we have the Rankin–Selberg bound

$$\sum_{|n|^2\leqslant x} |\lambda_n(V)|^2 \ll_V x. \tag{2.34}$$

2.6. Eisenstein series and spectral decomposition

In this subsection, we review the construction and properties of the (not necessarily spherical) Eisenstein series on $\Gamma\backslash G$. The quotient $\Gamma\backslash G$ has a unique cusp at ∞. For $\ell \in \mathbb{Z}_{\geqslant 0}$, $p, q \in \mathbb{Z}$ with $2 \mid p$ and $|p|, |q| \leqslant \ell$, and $\nu \in \mathbb{C}$ with $\Re\nu > 1$, we define the Eisenstein series of type (ℓ, q) at ∞ as in [33, Def. 3.3.1] by the absolutely and locally uniformly convergent series

$$E_{\ell,q}(\nu,p)(g) := \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \phi_{\ell,q}(\nu,p)(\gamma g), \tag{2.35}$$

where Γ_∞ is the subgroup of upper-triangular matrices in Γ (the stabilizer of ∞ in Γ), and

$$\phi_{\ell,q}(\nu,p)\left(\begin{pmatrix} r & * \\ & r^{-1}\end{pmatrix} k\right) := r^{2(1+\nu)}\Phi^\ell_{p,q}(k), \qquad r>0, \quad k\in K. \tag{2.36}$$

These Eisenstein series possess a meromorphic continuation to $\nu \in \mathbb{C}$, which is holomorphic along $i\mathbb{R}$ [33, §5.1]. An easy calculation with (2.32) and (2.4) shows that they are also eigenfunctions of the Hecke operators T_n with

$$T_n E_{\ell,q}(\nu,p) = \lambda_n(E(\nu,p))E_{\ell,q}(\nu,p), \qquad \lambda_n(E(\nu,p)) := \frac{1}{4}\sum_{n=ad} \chi_{\nu,p}(a)\chi_{-\nu,-p}(d), \tag{2.37}$$

where $\chi_{\nu,p}(z) := |z|^\nu (z/|z|)^{-p}$. In particular,

$$\lambda_{in}(E(\nu,p)) = (-1)^{p/2}\lambda_n(E(\nu,p)). \tag{2.38}$$

While $E_{\ell,q}(\nu,p)$ for individual $\nu \in i\mathbb{R}$ (barely) fail to lie in $L^2(\Gamma\backslash G)$, their averages against $C_c(i\mathbb{R})$ weights $f(\nu)$ comfortably do, and upon taking the Hilbert space closure of their span and orthocomplements one obtains the familiar orthogonal decomposition

$$L^2(\Gamma\backslash G) = \mathbb{C}\cdot 1 \oplus L^2(\Gamma\backslash G)_{\text{cusp}} \oplus L^2(\Gamma\backslash G)_{\text{Eis}}. \tag{2.39}$$

Let $H(\nu,p)$ be the linear span of all $\phi_{\ell,q}(\nu,p)$ with $|p|, |q| \leqslant \ell$. By (2.36), the functions $f \in H(\nu,p)$ satisfy

$$f\left(\begin{pmatrix} z & * \\ & z^{-1}\end{pmatrix} g\right) = |z|^2\chi_{\nu,p}(z^2)f(g), \qquad z\in\mathbb{C}^\times, \quad g\in G,$$

and they are determined by their restriction to K. In fact $H(\nu,p)$ as a $(\mathfrak{g}, K)$-module is isomorphic to the K-finite part of $V_{\nu,p}$ featured in (2.7). That is, the appropriate completion of $H(\nu,p)$ serves as a model of the Fréchet/Hilbert space representation $\pi_{\nu,p}$, and we shall denote by $H^\infty(\nu,p)$ the dense subspace of smooth vectors in this completion.

Denoting by $C^K(\Gamma\backslash G)$ the space of K-finite smooth functions on $\Gamma\backslash G$, an automorphic representation of type (ν,p) for $\Gamma\backslash G$ may be realized as a unitary $(\mathfrak{g}, K)$-module homomorphism $T : H(\nu,p) \to C^K(\Gamma\backslash G)$, with the corresponding $\pi_{\nu,p}$ an irreducible unitary representation on the Hilbert space $V_{\nu,p}$, cf. [33, §3.4 &

§8]. Such a T may arise as T_V for a cuspidal constituent $V \simeq V_{\nu,p}$ occurring discretely in $L^2(\Gamma\backslash G)_{\mathrm{cusp}}$, or from the Eisenstein series via

$$T_{E(\nu,p)}\phi_{\ell,q}(\nu,p) := E_{\ell,q}(\nu,p), \qquad |p|,|q| \leqslant \ell.$$

Indeed, by (2.35), the last display defines a $(\mathfrak{g},K)$-module homomorphism for $\Re\nu > 1$, hence by analytic continuation for all $\nu \in \mathbb{C}$ where the relevant Eisenstein series have no pole. Following custom, we lighten the notation by denoting a generic automorphic representation of type (ν,p), whether of type T_V or $T_{E(\nu,p)}$, as V, and its associated Hecke eigenvalues as $\lambda_n(V)$. Finally, we shall use that the above $(\mathfrak{g},K)$-module homomorphism extends uniquely to a G-module homomorphism $H^\infty(\nu,p) \to C^\infty(\Gamma\backslash G)$, and its image consists of functions of moderate growth.

Now (2.39) is explicated by the following two spectral identities. For f in the space $C_0^\infty(\Gamma\backslash G)$ of smooth complex-valued functions on $\Gamma\backslash G$ with all rapidly decaying derivatives, we have

$$\begin{aligned} f = \frac{\langle f,1\rangle}{\mathrm{vol}(\Gamma\backslash G)} &+ \sum_{V \text{ cuspidal}} \sum_{\substack{q,\ell\in\mathbb{Z}\\ |p_V|,|q|\leqslant\ell}} \frac{\langle f, T_V\phi_{\ell,q}(\nu_V,p_V)\rangle}{\|\Phi^\ell_{p_V,q}\|_K^2} T_V\phi_{\ell,q}(\nu_V,p_V) \\ &+ \frac{1}{\pi i}\int_{(0)} \sum_{p\in 2\mathbb{Z}} \sum_{\substack{q,\ell\in\mathbb{Z}\\ |p|,|q|\leqslant\ell}} \frac{\langle f, E_{\ell,q}(\nu,p)\rangle}{\|\Phi^\ell_{p,q}\|_K^2} E_{\ell,q}(\nu,p)\,\mathrm{d}\nu, \end{aligned} \tag{2.40}$$

with the obvious interpretation of $\langle f, E_{\ell,q}(\nu,p)\rangle$. For $f_1, f_2 \in C_0^\infty(\Gamma\backslash G)$, we have with the same interpretation

$$\begin{aligned} \langle f_1,f_2\rangle = \frac{\langle f_1,1\rangle\langle 1,f_2\rangle}{\mathrm{vol}(\Gamma\backslash G)} &+ \sum_{V \text{ cuspidal}} \sum_{\substack{q,\ell\in\mathbb{Z}\\ |p_V|,|q|\leqslant\ell}} \frac{\langle f_1, T_V\phi_{\ell,q}(\nu_V,p_V)\rangle\langle T_V\phi_{\ell,q}(\nu_V,p_V), f_2\rangle}{\|\Phi^\ell_{p_V,q}\|_K^2} \\ &+ \frac{1}{\pi i}\int_{(0)} \sum_{p\in 2\mathbb{Z}} \sum_{\substack{q,\ell\in\mathbb{Z}\\ |p|,|q|\leqslant\ell}} \frac{\langle f_1, E_{\ell,q}(\nu,p)\rangle\langle E_{\ell,q}(\nu,p), f_2\rangle}{\|\Phi^\ell_{p,q}\|_K^2}\,\mathrm{d}\nu. \end{aligned} \tag{2.41}$$

Compare with [14, Ch. 6, Th. 3.4] and [33, Th. 8.1].

We shorten the notation in two ways. First, for an automorphic representation V (cuspidal or Eisenstein) of type (ν,p) occurring in $L^2(\Gamma\backslash G)$, we write

$$\phi^V_{\ell,q} := \frac{T_V\phi_{\ell,q}(\nu,p)}{\|\Phi^\ell_{p,q}\|_K}, \qquad |p|,|q| \leqslant \ell.$$

In particular, when at least one of two such V and V' is cuspidal, $\langle \phi^V_{\ell,q}, \phi^{V'}_{\ell',q'}\rangle$ equals $\delta_{(\ell,q,V)=(\ell',q',V')}$. Second, while the decompositions in (2.40) and (2.41) are over all automorphic representations V (cuspidal or Eisenstein) occurring in $L^2(\Gamma\backslash G)$, keeping in mind the τ_ℓ-spherical transform of §2.4, it will be useful to introduce the shorthand notation $\int_{[\ell]} \mathrm{d}V$ for the sum-integral over those V of type (ν,p) such that $\pi_{\nu,p} \in \widehat{G}(\tau_\ell)$ (that is, with $|p| \leqslant \ell$ as well as $p \in 2\mathbb{Z}$ for V Eisenstein). Thus, for example, (2.40) may be rewritten in the more compact form

$$f = \frac{\langle f,1\rangle}{\mathrm{vol}(\Gamma\backslash G)} + \sum_{\ell\geqslant 0}\int_{[\ell]} \sum_{|q|\leqslant\ell} \langle f, \phi^V_{\ell,q}\rangle\phi^V_{\ell,q}\,\mathrm{d}V. \tag{2.42}$$

2.7. Rankin–Selberg convolutions

In this subsection, we review briefly the properties of Rankin–Selberg L-functions. We shall restrict to automorphic representations for $\Gamma\backslash G$ on which the Hecke operator T_i acts trivially, so that they lift to automorphic representations for $\mathrm{PGL}_2(\mathbb{Z}[i])\backslash\mathrm{PGL}_2(\mathbb{C})$. This allows us to refer to the theory of GL_2.

The Rankin–Selberg L-function of two automorphic representations V_j of type $(\nu_j, p_j) \in i\mathbb{R}\times\mathbb{Z}$ for $\Gamma\backslash G$ is defined by the absolutely convergent series (cf. (2.34))

$$L(s, V_1\times V_2) = \frac{1}{4}\zeta_{\mathbb{Q}(i)}(2s)\sum_{n\in\mathbb{Z}[i]\setminus\{0\}}\frac{\lambda_n(V_1)\lambda_n(V_2)}{(|n|^2)^s}, \qquad \Re s > 1. \tag{2.43}$$

This can be verified by matching the Euler factors on the two sides, using [28, Th. 15.1], [30, Prop. 3.5], [47, (3.1.3)], and [9, Lemma 1.6.1]. In particular,

$$L(s, V\times E(\nu,p)) = L(s-\tfrac{1}{2}\nu, V\otimes\chi_p)L(s+\tfrac{1}{2}\nu, V\otimes\chi_{-p})$$

for V cuspidal and $(\nu,p)\in i\mathbb{R}\times 4\mathbb{Z}$ according to (2.38), as well as

$$L\big(s, E(\nu_1,p_1)\times E(\nu_2,p_2)\big) = \prod_{\epsilon_1,\epsilon_2\in\{\pm1\}} L\big(s+\tfrac{1}{2}(\epsilon_1\nu_1+\epsilon_2\nu_2), \chi_{-\epsilon_1p_1-\epsilon_2p_2}\big),$$

with $(\nu_j,p_j)\in i\mathbb{R}\times 4\mathbb{Z}$ and $\chi_p(z) := (z/|z|)^{-p}$. All L-functions are meant over $\mathbb{Q}(i)$.

The Rankin–Selberg L-function $L(s, V_1\times V_2)$ possesses a meromorphic continuation to the entire complex plane with the exception of finitely many possible poles along the line $\Re s = 1$. It is in fact entire except as follows (cf. [18, Th. 2.2]):

- If $V_1 = V_2\,(=V)$ is cuspidal of type (ν,p) (that is, $(\nu_1,p_1) = \pm(\nu_2,p_2)$), there is a simple pole at $s=1$ with (strictly) positive residue

$$\operatorname*{res}_{s=1} L(s, V\times V) = \frac{\pi}{4}\cdot L(1, \mathrm{ad}^2 V) \gg_\varepsilon \big((1+|p|)(1+|\nu|)\big)^{-\varepsilon}, \tag{2.44}$$

  where the lower bound follows from [34, Prop. 3.2].
- If V_1 and V_2 are both Eisenstein series with $p_1 = \epsilon p_2$ for some $\epsilon\in\{\pm1\}$, there are simple poles at $s = 1+\eta(\nu_1-\epsilon\nu_2)/2$ for $\eta\in\{\pm1\}$ with residue

$$\mathcal{L}_\eta(V_1,V_2) := \frac{\pi}{4}\cdot\zeta_{\mathbb{Q}(i)}(1+\eta(\nu_1-\epsilon\nu_2))L(1+\eta\nu_1, \chi_{-2\eta p_1})L(1-\eta\epsilon\nu_2, \chi_{2\eta\epsilon p_2}), \tag{2.45}$$

  unless $\nu_1 = \pm\nu_2$ or $\nu_1 = 0$ or $\nu_2 = 0$, in which case, however, the definition still makes sense as a meromorphic function of ν_1 and ν_2.

Finally, the associated completed L-function satisfies the familiar functional equation

$$\Lambda(s, V_1\times V_2) := 16^s L(s, V_1\times V_2)L_\infty(s, V_1\times V_2) = \Lambda(1-s, V_1\times V_2), \tag{2.46}$$

where the exponential factor 16^s coming from the discriminant of $\mathbb{Q}(i)$ is included for convenience, and the factor at infinity is given by

$$L_\infty(s, V_1\times V_2) = \Gamma(s,\vec{\nu},\vec{p}) := \prod_{\epsilon_1,\epsilon_2\in\{\pm1\}} L_\infty(s, \chi_{\epsilon_1\nu_1,\epsilon_1p_1}\cdot\chi_{\epsilon_2\nu_2,\epsilon_2p_2})$$

$$= \prod_{\epsilon_1,\epsilon_2\in\{\pm 1\}} \Gamma_{\mathbb{C}}\left(s + \tfrac{1}{2}(\epsilon_1\nu_1 + \epsilon_2\nu_2) + \tfrac{1}{2}|\epsilon_1 p_1 + \epsilon_2 p_2|\right). \tag{2.47}$$

Here we used the abbreviations

$$\Gamma_{\mathbb{C}}(s) := 2(2\pi)^{-s}\Gamma(s), \qquad \vec{\nu} := (\nu_1,\nu_2), \qquad \vec{p} := (p_1,p_2).$$

Indeed, (2.46)–(2.47) follow from [28, Prop. 18.2], [47, §3], [53, Prop. 6 in §VII-2] and its proof, upon noting that V_j is isomorphic to the principal series representation induced from the pair of characters $(\chi_{-\nu_j,-p_j}, \chi_{\nu_j,p_j})$.

Lemma 2. *Let $f : i\mathbb{R} \to \mathbb{C}$ be a function decaying as $f(\nu) \ll (1+|\nu|)^{-3}$, and let $p \in \mathbb{Z}$. Then*

$$\int_{(0)}\int_{(0)} \sum_{\eta\in\{\pm 1\}} f(\nu_1)\overline{f(\nu_2)}\,\mathcal{L}_\eta((\nu_1,p),(\nu_2,p))\,\frac{\mathrm{d}\nu_1}{\pi i}\,\frac{\mathrm{d}\nu_2}{\pi i} \geqslant 0.$$

Proof. First we note that the η-sum cancels the individual poles of $\mathcal{L}_\eta((\nu_1,p),(\nu_2,p))$ at $\nu_1 = \nu_2$. For $\varepsilon > 0$ and $V_j = (\nu_j, p)$ with $j \in \{1,2\}$ define

$$\mathcal{L}_\eta(V_1,V_2,\varepsilon) := \frac{\pi}{4}\cdot \zeta_{\mathbb{Q}(i)}(1+\varepsilon+\eta(\nu_1-\nu_2))L(1+\eta\nu_1,\chi_{-2\eta p})L(1-\eta\nu_2,\chi_{2\eta p})$$

and

$$\mathcal{I}(\varepsilon) := \int_{(0)}\int_{(0)} \sum_{\eta\in\{\pm 1\}} f(\nu_1)\overline{f(\nu_2)}\,\mathcal{L}_\eta(V_1,V_2,\varepsilon)\,\frac{\mathrm{d}\nu_1}{\pi i}\,\frac{\mathrm{d}\nu_2}{\pi i}.$$

This function is continuous at $\varepsilon = 0$, so it suffices to show $\mathcal{I}(\varepsilon) \geqslant 0$ for $\varepsilon > 0$. Inserting the definition and opening the Dedekind zeta function, we see that

$$\mathcal{I}(\varepsilon) = \frac{\pi}{16}\sum_{\eta\in\{\pm 1\}}\sum_{n\in\mathbb{Z}[i]\setminus\{0\}} \frac{1}{|n|^{2+2\varepsilon}}\left|\int_{(0)} \frac{1}{|n|^{2\eta\nu}} L(1+\eta\nu,\chi_{-2\eta p})f(\nu)\frac{\mathrm{d}\nu}{\pi i}\right|^2 \geqslant 0$$

as desired. □

2.8. Diagonal detection of Voronoi type

In this subsection, we prove a Voronoi-type formula that allows us to detect equality of two automorphic representations occurring in $L^2(\Gamma\backslash G)$ in terms of a certain weighted orthogonality relation between their Hecke eigenvalues. We shall use that only tempered representations occur in $L^2(\Gamma\backslash G)$, e.g. by [14, Ch. 7, Prop. 6.2].

Lemma 3. *Let $P \geqslant 1$ be a parameter. There exists a function*

$$W_P : \mathbb{R}_{>0}\times\mathbb{C}^2\times\mathbb{Z}^2 \to \mathbb{C},$$

given explicitly by (2.50), with the following properties.

(a) $W_P(x,\vec{\nu},\vec{p})$ is an entire function of $\vec{\nu} = (\nu_1,\nu_2) \in \mathbb{C}^2$, and it is invariant under

$$(\nu_j,p_j) \mapsto (-\nu_j,-p_j) \qquad \text{as well as} \qquad (\nu_j,p_j) \mapsto (p_j,\nu_j) \quad (\nu_j \in \mathbb{Z}).$$

(b) Let us abbreviate $\tilde{P} := \big(1+|p_1+p_2|\big)\big(1+|p_1-p_2|\big)$. *Then for every* $A > |\Re\nu_1| + |\Re\nu_2|$ *we have*

$$W_P(x,\vec{\nu},\vec{p}) \ll_{A,\Re\nu_1,\Re\nu_2} \big(1+(\tilde{P}/P)^{2A-2}\big)\big(1+|\nu_1|+|\nu_2|\big)^{4A} x^{-A}. \tag{2.48}$$

(c) For every two automorphic representations V_j *of type* $(\nu_j, p_j) \in i\mathbb{R}\times\mathbb{Z}$ *for* $\Gamma\backslash G$ *we have*

$$\begin{aligned}&\sum_{n\in\mathbb{Z}[i]\setminus\{0\}} W_P\left(\frac{|n|}{P},\vec{\nu},\vec{p}\right)\lambda_n(V_1)\lambda_n(V_2)\\ &= \begin{cases} \frac{\pi}{4}L(1,\mathrm{ad}^2\, V_1)P^2, & V_1 = V_2 \text{ cuspidal};\\ \sum_{\eta\in\{\pm1\}}\mathcal{L}_\eta(V_1,V_2)P^{2+\eta(\nu_1-\epsilon\nu_2)}, & V_1, V_2 \text{ Eisenstein, } p_1=\epsilon p_2,\ \epsilon\in\{\pm1\};\\ 0, & \text{otherwise,}\end{cases}\end{aligned} \tag{2.49}$$

where $L(1,\mathrm{ad}^2\, V_1)$ *and* $\mathcal{L}_\eta(V_1,V_2)$ *are as in* (2.44) *and* (2.45).

Proof. Let $w:\mathbb{R}_{>0}\to\mathbb{C}$ be a smooth function supported inside $[1,2]$, and normalized so that its Mellin transform $\widehat{w}(s) = \int_0^\infty w(x)x^s\,\mathrm{d}x/x$ satisfies $\widehat{w}(1)=1$. We define

$$W_P(x,\vec{\nu},\vec{p}) := \frac{1}{8\pi i}\int_{(2)} \zeta_{\mathbb{Q}(i)}(2s)\left(\widehat{w}(s) - P^{2-4s}\frac{16^{2s-1}\Gamma(s,\vec{\nu},\vec{p})}{\Gamma(1-s,\vec{\nu},\vec{p})}\widehat{w}(1-s)\right)x^{-2s}\,\mathrm{d}s, \tag{2.50}$$

where $\Gamma(s,\vec{\nu},\vec{p})$ is as in (2.47).

Shifting the contour to the far right, we see that $W_P(x,\vec{\nu},\vec{p})$ is entire in $\vec{\nu}$. The symmetry with respect to $(\nu_j,p_j)\mapsto(-\nu_j,-p_j)$ is obvious from (2.47). For $r\in\frac12\mathbb{Z}$ we have the equality

$$\frac{\Gamma(z+r)}{\Gamma(1-z+r)} = \frac{\Gamma(z-r)}{\Gamma(1-z-r)}\cdot\frac{\sin(\pi(z-r))}{\sin(\pi(z+r))} = (-1)^{2r}\frac{\Gamma(z-r)}{\Gamma(1-z-r)}$$

of meromorphic functions in $z\in\mathbb{C}$. This shows that (cf. (2.47))

$$\begin{aligned}\frac{\Gamma(s,\vec{\nu},\vec{p})}{\Gamma(1-s,\vec{\nu},\vec{p})} &= \prod_{\epsilon_1,\epsilon_2\in\{\pm1\}}\frac{\Gamma_{\mathbb{C}}\left(s+\frac12(\epsilon_1\nu_1+\epsilon_2\nu_2)+\frac12|\epsilon_1p_1+\epsilon_2p_2|\right)}{\Gamma_{\mathbb{C}}\left(1-s-\frac12(\epsilon_1\nu_1+\epsilon_2\nu_2)+\frac12|\epsilon_1p_1+\epsilon_2p_2|\right)}\\ &= \prod_{\epsilon_1,\epsilon_2\in\{\pm1\}}\frac{\Gamma_{\mathbb{C}}\left(s+\frac12(\epsilon_1\nu_1+\epsilon_2\nu_2)+\frac12(\epsilon_1p_1+\epsilon_2p_2)\right)}{\Gamma_{\mathbb{C}}\left(1-s-\frac12(\epsilon_1\nu_1+\epsilon_2\nu_2)+\frac12(\epsilon_1p_1+\epsilon_2p_2)\right)}\end{aligned}$$

is symmetric with respect to $(\nu_j,p_j)\mapsto(p_j,\nu_j)$, completing the proof of (a).

Combining the first line of the previous display with [24, Lemma 3.2], we infer for $\Re(s) > \frac12|\Re\nu_1|+\frac12|\Re\nu_2|$ that

$$\begin{aligned}\left|\frac{\Gamma(s,\vec{\nu},\vec{p})}{\Gamma(1-s,\vec{\nu},\vec{p})}\right| &= \prod_{\epsilon_1,\epsilon_2\in\{\pm1\}}\left|\frac{\Gamma_{\mathbb{C}}\left(s+\frac12(\epsilon_1\nu_1+\epsilon_2\nu_2)+\frac12|\epsilon_1p_1+\epsilon_2p_2|\right)}{\Gamma_{\mathbb{C}}\left(1-\overline{s}-\frac12(\epsilon_1\overline{\nu_1}+\epsilon_2\overline{\nu_2})+\frac12|\epsilon_1p_1+\epsilon_2p_2|\right)}\right|\\ &\ll_{\Re s,\Re\nu_1,\Re\nu_2} \prod_{\epsilon_1,\epsilon_2\in\{\pm1\}}\left|s+\tfrac12(\epsilon_1\nu_1+\epsilon_2\nu_2)+\tfrac12|\epsilon_1p_1+\epsilon_2p_2|\right|^{\Re(2s+\epsilon_1\nu_1+\epsilon_2\nu_2)-1}\\ &\ll_{\Re s,\Re\nu_1,\Re\nu_2} \prod_{\epsilon_1,\epsilon_2\in\{\pm1\}}(1+|\epsilon_1p_1+\epsilon_2p_2|)^{\Re(2s+\epsilon_1\nu_1+\epsilon_2\nu_2)-1}(|s|+|\nu_1|+|\nu_2|)^{\Re(2s+\epsilon_1\nu_1+\epsilon_2\nu_2)}\\ &= (1+|p_1+p_2|)^{4\Re s-2}(1+|p_1-p_2|)^{4\Re s-2}(|s|+|\nu_1|+|\nu_2|)^{8\Re s}.\end{aligned}$$

Turning back to (2.50), the singularity of the integrand at $s = 1/2$ is removable, so we can shift the contour to $\Re s = A/2$. The bound (2.48) follows upon noting that

- $\widehat{w}(s) \ll_{C,\Re s} (1+|s|)^{-C}$ for all $C > 0$ and $s \in \mathbb{C}$;
- $\zeta_{\mathbb{Q}(i)}(2s) \ll (1+|s|)^2$ for $\Re s > 0$ and $|2s-1| > 1$.

Finally, to show (c), we start from the following identity, a consequence of (2.43):

$$\frac{1}{16} \sum_{m,n \in \mathbb{Z}[i]\setminus\{0\}} w\left(\frac{|m|^4|n|^2}{P^2}\right) \lambda_n(V_1)\lambda_n(V_2) = \frac{1}{2\pi i} \int_{(2)} L(s, V_1 \times V_2)\widehat{w}(s)P^{2s}\,\mathrm{d}s.$$

We shift the contour to $\Re s = -1$; the contribution of the possible poles (on the line $\Re s = 1$) is recorded on the right-hand side of (2.49). In the remaining integral we apply the functional equation (2.46) and change variables $s \mapsto 1-s$ getting

$$\frac{1}{2\pi i} \int_{(2)} L(s, V_1 \times V_2)P^{2-4s} \frac{16^s \Gamma(s, \vec{\nu}, \vec{p})}{16^{1-s}\Gamma(1-s, \vec{\nu}, \vec{p})} \widehat{w}(1-s)P^{2s}\,\mathrm{d}s.$$

Moving this term to the other side, we obtain the desired formula (2.49), first for $(\nu_1, p_1) \neq \pm(\nu_2, p_2)$, but then by analytic continuation everywhere. This completes the proof of (c). □

3. Pre-trace formula and amplification

In this section, we first implement a pre-trace setup, using integral kernels that are (by necessity) not bi-K-invariant, first in §3.1 as the full pre-trace formula based on the theory of Eisenstein series and then as a streamlined pre-trace inequality in §3.2. In §§3.3–3.5, we couple the pre-trace setup with either amplification by Hecke operators or self-amplification via diagonal detection of Voronoi type in §2.8 to derive estimates on pointwise values of automorphic forms in terms of estimates on generalized spherical trace functions and Diophantine counts.

3.1. Amplified pre-trace formula

In this subsection, we prove an amplified pre-trace formula based on the theory of Eisenstein series and the spectral decomposition of $L^2(\Gamma\backslash G)$ (see §2.6). This is a familiar identity between spectral and geometric data, and its full force will be needed in the proof of Theorem 3(b); in fact, as an even more general version, we shall use a double pre-trace formula (see §3.4) in two variables.

Let A be a bounded operator on $L^2(\Gamma\backslash G)$ preserving the subspace $C_0^\infty(\Gamma\backslash G)$ of smooth functions with all rapidly decreasing derivatives. Assume that for the basis forms $\phi_{\ell,q}^V$, indexed as in (2.42) by V occurring in $L^2(\Gamma\backslash G)$ (cuspidal or Eisenstein) and $\ell, q \in \mathbb{Z}$ satisfying $\ell \geqslant \max(|p_V|, |q|)$, there are constants $c_{\ell,q}^V(A) \in \mathbb{C}$ such that

$$\langle A\psi, \phi_{\ell,q}^V\rangle = c_{\ell,q}^V(A)\langle \psi, \phi_{\ell,q}^V\rangle, \qquad \psi \in C_0^\infty(\Gamma\backslash G). \tag{3.1}$$

Then (2.41) yields, for every $\psi \in C_0^\infty(\Gamma\backslash G)$,

$$\langle A\psi, \psi\rangle = \frac{\langle A\psi, 1\rangle\langle 1, \psi\rangle}{\mathrm{vol}(\Gamma\backslash G)} + \sum_{\ell \geqslant 0} \int_{[\ell]} \sum_{|q| \leqslant \ell} c_{\ell,q}^V(A)|\langle \psi, \phi_{\ell,q}^V\rangle|^2\,\mathrm{d}V. \tag{3.2}$$

For $f \in C_0(G)$ a rapidly decaying continuous function on G, and $\psi \in L^2(\Gamma\backslash G)$, we may consider the function $R(f)\psi \in L^2(\Gamma\backslash G)$ defined by

$$\begin{aligned}(R(f)\psi)(g) &:= \int_G f(h)\psi(gh)\,\mathrm{d}h = \int_G f(g^{-1}h)\psi(h)\,\mathrm{d}h \\ &= \int_{\Gamma\backslash G} k_f(g,h)\psi(h)\,\mathrm{d}h, \qquad k_f(g,h) := \sum_{\gamma\in\Gamma} f(g^{-1}\gamma h).\end{aligned}$$

Thus $R(f)$ is a bounded integral operator on $L^2(\Gamma\backslash G)$ with kernel k_f. It is clear that $R(f)$ preserves $C_0^\infty(\Gamma\backslash G)$, and its adjoint equals $R(f)^* = R(f^*)$ with

$$f^*(g) := \overline{f(g^{-1})}, \qquad g \in G.$$

Further, for a finitely supported sequence of complex coefficients $x = (x_n)_{n\in\mathbb{Z}[i]\setminus\{0\}}$, let $R_{\mathrm{fin}}(x)$ be the operator on $L^2(\Gamma\backslash G)$ given by

$$R_{\mathrm{fin}}(x) := \sum_{n\in\mathbb{Z}[i]\setminus\{0\}} x_n T_n. \tag{3.3}$$

The adjoint of this operator equals $R_{\mathrm{fin}}(x)^* = R_{\mathrm{fin}}(\overline{x})$.

Let us now fix an integer $\ell \geqslant 1$. Let $f \in \mathcal{H}(\tau_\ell)_\infty$ be such that $f = f^*$, and let $x = (x_n)$ be as above such that $x = \overline{x}$, the self-adjointness conditions serving only to lighten the notation below. Further, let V be a non-identity (cuspidal or Eisenstein) automorphic representation of arbitrary type (ν_V, p_V) occurring in $L^2(\Gamma\backslash G)$, and let $\ell', q \in \mathbb{Z}$ be such that $\ell' \geqslant \max(|p_V|, |q|)$. For V cuspidal, (2.15) and (2.18) show that

$$\begin{aligned} R(f)\phi^V_{\ell',q} &= \delta_{\ell'=\ell}\frac{\widehat{f}(V)}{2\ell+1}\phi^V_{\ell',q}, & \widehat{f}(V) &:= \widehat{f}(\nu_V, p_V); \\ R_{\mathrm{fin}}(x)\phi^V_{\ell',q} &= \widehat{x}(V)\phi^V_{\ell',q}, & \widehat{x}(V) &:= \sum_{n\in\mathbb{Z}[i]\setminus\{0\}} x_n\lambda_n(V).\end{aligned} \tag{3.4}$$

For V Eisenstein, these equations are still valid with the obvious extension of $R(f)$ and $R_{\mathrm{fin}}(x)$ to functions in $C^\infty(\Gamma\backslash G)$ of moderate growth, as follows from (2.37) and the discussion between (2.39) and (2.40). Therefore, following the usual argument that $R(f)$ and $R_{\mathrm{fin}}(x)$ are self-adjoint, we obtain that $A := R(f)R_{\mathrm{fin}}(x)$ satisfies (3.1) with

$$c^V_{\ell',q}(A) = \delta_{\ell'=\ell}\frac{\widehat{f}(V)\widehat{x}(V)}{2\ell+1}.$$

Hence (3.2) holds with these coefficients and ℓ-summation replaced by ℓ'-summation. We note that the coefficients decay rapidly in ν by Theorem 7. Moreover, $A(1) = R(f)(1)$ vanishes by $f = f \star \overline{\chi_\ell}$ and the orthogonality of characters (recalling that $\ell \geqslant 1$).

Applying (3.2) and recalling our observation below (2.32) about $n\gamma^{-1}$ as $\gamma \in \Gamma \setminus \Gamma_n$, we obtain for every $\psi \in C_0^\infty(\Gamma\backslash G)$ that

$$\begin{aligned}\int_{[\ell]} \sum_{|q|\leqslant\ell} c^V_{\ell,q}(A)|\langle\psi, \phi^V_{\ell,q}\rangle|^2\,\mathrm{d}V &= \iint_{(\Gamma\backslash G)^2} k_f(g,h) \sum_{n\in\mathbb{Z}[i]\setminus\{0\}} x_n T_n\psi(h)\overline{\psi(g)}\,\mathrm{d}g\,\mathrm{d}h \\ &= \iint_{(\Gamma\backslash G)^2} \sum_{n\in\mathbb{Z}[i]\setminus\{0\}} \frac{x_n}{|n|} \sum_{\gamma\in\Gamma_n} f(g^{-1}\tilde{\gamma}h)\overline{\psi(g)}\psi(h)\,\mathrm{d}g\,\mathrm{d}h,\end{aligned}$$

where $\tilde\gamma$ abbreviates $\gamma/\sqrt{\det\gamma}$. Letting ψ range through smooth, nonnegative, L^1-normalized functions supported in increasingly small open neighborhoods of a fixed point $\Gamma g \in \Gamma\backslash G$, and taking limits using the rapid decay of $c^V_{\ell,q}(A)$, we obtain the desired *amplified pre-trace formula*

$$\int_{[\ell]} \frac{\widehat f(V)\widehat x(V)}{2\ell+1} \sum_{|q|\leqslant \ell} |\phi^V_{\ell,q}(g)|^2 \,\mathrm{d}V = \sum_{n\in\mathbb{Z}[i]\setminus\{0\}} \frac{x_n}{|n|} \sum_{\gamma\in\Gamma_n} f(g^{-1}\tilde\gamma g). \tag{3.5}$$

The pre-trace formula (3.5) isolates forms $\phi^V_{\ell,q}$ with a specific value of ℓ (thus, forms in the chosen constituent V^ℓ in the decomposition (1.2) for various V's), a starting point for a proof of Theorem 1. To further isolate eigenforms in the specific constituent $V^{\ell,q}$ (for a fixed $|q| \leqslant \ell$), starting from our earlier $f \in \mathcal{H}(\tau_\ell)_\infty$ satisfying $f = f^*$, we define a smooth function $f_q \in C_0(G)$ by

$$f_q(g) := \frac{1}{2\pi}\int_0^{2\pi} f\big(g\,\mathrm{diag}(e^{i\varrho}, e^{-i\varrho})\big)\, e^{2qi\varrho}\,\mathrm{d}\varrho = \frac{1}{2\pi}\int_0^{2\pi} f\big(\,\mathrm{diag}(e^{i\varrho}, e^{-i\varrho})g\big)\, e^{2qi\varrho}\,\mathrm{d}\varrho. \tag{3.6}$$

We note that $f_q = f_q^*$, but f_q need not lie in $\mathcal{H}(\tau_\ell)_\infty$. By the orthogonality of characters on $\mathbb{R}/\mathbb{Z}$, we have

$$R(f_q) = R(f)\Pi_q = \Pi_q R(f), \tag{3.7}$$

where Π_q is the projection onto the closed subspace consisting of $\psi \in L^2(\Gamma\backslash G)$ such that $\psi(g\,\mathrm{diag}(e^{i\varrho}, e^{-i\varrho})) = e^{2qi\varrho}\psi(g)$. In particular, $R(f_q)$ is a bounded, self-adjoint operator, which preserves $C_0^\infty(\Gamma\backslash G)$. Moreover, by (3.4) and the surrounding discussion,

$$R(f_q)\phi^V_{\ell',q'} = \delta_{(\ell',q')=(\ell,q)}\frac{\widehat f(V)}{2\ell+1}\phi^V_{\ell',q'}$$

holds for V cuspidal, and also for V Eisenstein with the obvious extension of $R(f_q)$ to functions in $C^\infty(\Gamma\backslash G)$ of moderate growth. Thus, applying as above (3.2) with $A = R(f_q)R_{\mathrm{fin}}(x)$, we obtain the following amplified pre-trace formula for individual forms:

$$\int_{[\ell]} \frac{\widehat f(V)\widehat x(V)}{2\ell+1} |\phi^V_{\ell,q}(g)|^2 \,\mathrm{d}V = \sum_{n\in\mathbb{Z}[i]\setminus\{0\}} \frac{x_n}{|n|} \sum_{\gamma\in\Gamma_n} f_q(g^{-1}\tilde\gamma g). \tag{3.8}$$

We proved (3.5) and (3.8) for every $f \in \mathcal{H}(\tau_\ell)_\infty$ and finitely supported $x = (x_n)$ under the assumption that $f = f^*$ and $x = \overline{x}$. In fact (3.5) and (3.8) hold without this assumption, because both sides are $\mathbb{C}$-linear in f and x. Alternatively, one can modify the above proof to work without the self-adjointness assumption, starting with the analogue of (3.4) for $R(f^*)\phi^V_{\ell',q}$ and $R_{\mathrm{fin}}(\overline{x})\phi^V_{\ell',q}$.

3.2. Positivity and amplified pre-trace inequality

In many situations, the coefficients on the left-hand (spectral) side of (3.5) and (3.8) are nonnegative, and the pre-trace formula is simply used as an inequality, by dropping all but the terms of interest. This is the case for the proofs of Theorems 1, 2 and 3(a). In this subsection, we derive such amplified pre-trace inequalities in a streamlined way with substantially less heavy machinery, drawing inspiration from [3, §3]. For example, here we do not even need to mention Eisenstein series.

Let A be a positive operator on $L^2(\Gamma\backslash G)$, and let $\mathfrak{B}$ be a finite orthonormal system of eigenfunctions ϕ of A with (not necessarily distinct) eigenvalues $(c_\phi(A))_{\phi\in\mathfrak{B}}$. Then, A preserves the orthodecomposition

$L^2(\Gamma\backslash G) = \mathrm{Span}(\mathfrak{B}) \oplus \mathrm{Span}(\mathfrak{B})^\perp$, and for any $\psi \in L^2(\Gamma\backslash G)$ the corresponding decomposition $\psi = \psi_1 + \psi_2$ with

$$\psi_1 := \sum_{\phi\in\mathfrak{B}} \langle \psi, \phi\rangle \phi \qquad \text{and} \qquad \psi_2 := \psi - \psi_1$$

gives

$$\langle A\psi, \psi\rangle = \langle A\psi_1, \psi_1\rangle + \langle A\psi_2, \psi_2\rangle \geqslant \langle A\psi_1, \psi_1\rangle = \sum_{\phi\in\mathfrak{B}} c_\phi(A) |\langle \psi, \phi\rangle|^2. \tag{3.9}$$

We will apply this positivity argument to the operators $A = R(f)R_{\mathrm{fin}}(x)$ and $A = R(f_q)R_{\mathrm{fin}}(x)$, where $f \in \mathcal{H}(\tau_\ell)_\infty$ and $x = (x_n)$ are as in the previous subsection. Positivity is achieved by making the operators $R(f)$ and $R_{\mathrm{fin}}(x)$ individually positive, because Hecke operators commute with integral operators, and Π_q in (3.7) is a positive operator commuting with $R(f)$. For the positivity of $R(f)$, it suffices that

$$f = u \star u \qquad \text{for some } u \in \mathcal{H}(\tau_\ell)_\infty \text{ satisfying } u = u^*. \tag{3.10}$$

For the positivity of $R_{\mathrm{fin}}(x)$, it suffices that

$$\begin{aligned} R_{\mathrm{fin}}(x) &= \Big(\sum_{l\in P} y_l T_l\Big) \star \Big(\sum_{m\in P} \overline{y_m} T_m\Big) + \Big(\sum_{l\in P} z_l T_{l^2}\Big) \star \Big(\sum_{m\in P} \overline{z_m} T_{m^2}\Big), \\ x_n &:= \sum_{\substack{l,m\in P\\ (d)|(l,m)\\ lm/d^2=n}} y_l \overline{y_m} + \sum_{\substack{l,m\in P\\ (d)|(l^2,m^2)\\ l^2m^2/d^2=n}} z_l \overline{z_m}, \end{aligned} \tag{3.11}$$

where $(y_l)_{l\in P}$ and $(z_l)_{l\in P}$ are arbitrary complex coefficients supported on a finite set $P \subset \mathbb{Z}[i] \setminus \{0\}$. Here we used that each Hecke operator T_n is self-adjoint.

Now, let V be a cuspidal automorphic representation that occurs in $L^2(\Gamma\backslash G)$ and contains τ_ℓ-type vectors. Let $\mathfrak{B} = \{\phi_q : |q| \leqslant \ell\}$ be an orthonormal basis of V^ℓ, with $\phi_q \in V^{\ell,q}$. As in the previous subsection, we evaluate the left-hand side of (3.9) geometrically, and then apply a limit in ψ to both sides. This way we obtain the following *amplified pre-trace inequalities* in place of (3.5) and (3.8):

$$\frac{\widehat{f}(V)\widehat{x}(V)}{2\ell+1} \sum_{\phi\in\mathfrak{B}} |\phi(g)|^2 \leqslant \sum_{n\in\mathbb{Z}[i]\setminus\{0\}} \frac{x_n}{|n|} \sum_{\gamma\in\Gamma_n} f(g^{-1}\tilde{\gamma}g), \tag{3.12}$$

$$\frac{\widehat{f}(V)\widehat{x}(V)}{2\ell+1} |\phi_q(g)|^2 \leqslant \sum_{n\in\mathbb{Z}[i]\setminus\{0\}} \frac{x_n}{|n|} \sum_{\gamma\in\Gamma_n} f_q(g^{-1}\tilde{\gamma}g). \tag{3.13}$$

3.3. Test functions and amplifier

The main idea of the amplified pre-trace inequality (3.12) is that it can provide a good upper bound for $\sum_{\phi\in\mathfrak{B}} |\phi(g)|^2$ as long as the test function $f \in \mathcal{H}(\tau_\ell)_\infty$ and the amplifier $x = (x_n)$ in §3.2 are chosen so that $\widehat{f}(V)$ and $\widehat{x}(V)$ are sizeable while the right-hand side is not too large. In this subsection, we make these choices.

As in Theorems 1, 2 and 3, let $\ell \geqslant 1$ be an integer, $I \subset i\mathbb{R}$ and $\Omega \subset G$ be compact sets. Let $V \subset L^2(\Gamma\backslash G)$ be a cuspidal automorphic representation with minimal K-type τ_ℓ and spectral parameter $\nu_V \in I$. Let us introduce the spectral weights

$$h(\nu,p) := \begin{cases} e^{(p^2-\ell^2+\nu^2)/2}, & \nu\in\mathbb{C},\quad p\in\tfrac12\mathbb{Z},\quad |p|\leqslant \ell,\\ 0, & \nu\in\mathbb{C},\quad p\in\tfrac12\mathbb{Z},\quad |p|>\ell. \end{cases} \tag{3.14}$$

According to Theorems 10 and 7, the inverse τ_ℓ-spherical transform $f := \check{h}$ given by (2.22) belongs to $\mathcal{H}(\tau_\ell)_\infty$, and it satisfies $\widehat{f} = h$. Moreover, if we set $u := \check{v}$ with

$$v(\nu,p) := \begin{cases} (2\ell+1)^{1/2}e^{(p^2-\ell^2+\nu^2)/4}, & \nu\in\mathbb{C},\quad p\in\tfrac12\mathbb{Z},\quad |p|\leqslant \ell,\\ 0, & \nu\in\mathbb{C},\quad p\in\tfrac12\mathbb{Z},\quad |p|>\ell, \end{cases}$$

then $u\in\mathcal{H}(\tau_\ell)_\infty$, $u = u^*$ by (2.20) and (2.22), and $\widehat{f} = \widehat{u}^2/(2\ell+1) = \widehat{u\star u}$. This shows that (3.10) is satisfied. Hence $R(f)$ is the kind of positive operator considered in §3.2, and by (3.4) we have

$$\widehat{f}(V) = h(\nu_V,\ell) \gg_I 1. \tag{3.15}$$

With the notation (2.27), we have

$$\widetilde{f}(s,p) = \sqrt{2\pi}(p^2+1-s^2)e^{(p^2-\ell^2-s^2)/2},$$

whence by (2.29), (2.22), and the trivial bound $\left|\varphi^\ell_{\nu,p}(g^{-1})\right|\leqslant 2\ell+1$, we have

$$f(g) \ll \ell^2 e^{-\log^2\|g\|}. \tag{3.16}$$

We shall also use the following supplement, a consequence of (2.20) and (2.22):

$$f(g) \ll \ell \sup_{\nu\in i\mathbb{R}} \left|\varphi^\ell_{\nu,\ell}(g)\right| + \ell^{-50}. \tag{3.17}$$

We now choose our amplifier, which we do as in [2, §5]. Let $L\geqslant 7$ be a parameter, to be chosen at the very end of the proof of Theorems 1, 2 and 3, and set

$$P(L) := \left\{l\in\mathbb{Z}[i] \text{ prime} : 0<\arg(l)<\tfrac{\pi}{4} \text{ and } L\leqslant |l|^2\leqslant 2L\right\};$$
$$y_l := \operatorname{sgn}(\lambda_l(V)),\qquad z_l := \operatorname{sgn}(\lambda_{l^2}(V)),\qquad l\in P(L).$$

It follows from the result of Breusch [8, Teil II] (or from the prime number theorem for arithmetic progressions, for sufficiently large L) that $P(L)\neq\emptyset$, while in (3.3) and (3.11) we have

$$x_n = \begin{cases} \sum_{l\in P(L)}(y_l^2+{z_l}^2) \ll L/\log L, & n=1;\\ (1+\delta_{l_1\neq l_2})y_{l_1}y_{l_2}+\delta_{l_1=l_2}z_{l_1}z_{l_2}\ll 1, & n=l_1l_2 \text{ for some } l_1,l_2\in P(L);\\ (1+\delta_{l_1\neq l_2})z_{l_1}z_{l_2}\ll 1, & n=l_1^2l_2^2 \text{ for some } l_1,l_2\in P(L);\\ 0, & \text{otherwise.} \end{cases} \tag{3.18}$$

This formula is the analogue of [3, (9.16)], except that we forgot to insert the factors $1+\delta_{l_1\neq l_2}$ there. In particular, by the inequality $|\lambda_l(V)|+|\lambda_{l^2}(V)|>1/2$ that follows from (2.33), we have

$$\widehat{x}(V) = \Big(\sum_{l\in P(L)}|\lambda_l(V)|\Big)^2 + \Big(\sum_{l\in P(L)}|\lambda_{l^2}(V)|\Big)^2 \gg \frac{L^2}{\log^2 L}. \tag{3.19}$$

Let $\mathfrak{B}$ be an orthonormal basis of V^ℓ. Entering the lower bounds (3.15) and (3.19) into the amplified pre-trace inequality (3.12), we obtain

$$\frac{L^{2-\varepsilon}}{\ell}\sum_{\phi\in\mathfrak{B}}|\phi(g)|^2 \ll_{\varepsilon,I} \sum_{n\in\mathbb{Z}[i]\setminus\{0\}}\frac{|x_n|}{|n|}\sum_{\gamma\in\Gamma_n}|f(g^{-1}\tilde{\gamma}g)|. \tag{3.20}$$

Let us assume that $g\in\Omega$. A straightforward counting combined with the divisor bound shows that

$$\#\left\{\gamma\in\Gamma_n : \|g^{-1}\tilde{\gamma}g\|\leqslant R\right\} \ll_{\varepsilon,\Omega} R^{4+\varepsilon}|n|^{2+\varepsilon}, \tag{3.21}$$

so that, splitting into dyadic ranges for $\|g^{-1}\tilde{\gamma}g\|$ and using (3.16), we obtain

$$\sum_{\substack{\gamma\in\Gamma_n\\ \log\|g^{-1}\tilde{\gamma}g\|>8\sqrt{\log\ell}}}|f(g^{-1}\tilde{\gamma}g)| \ll_{\varepsilon,\Omega} \ell^{-50}|n|^{2+\varepsilon}.$$

Thus from (3.17) and (3.20) we conclude that

$$\sum_{\phi\in\mathfrak{B}}|\phi(g)|^2 \ll_{\varepsilon,I,\Omega} L^{-2+\varepsilon}\ell^2 \sum_{\substack{n\in\mathbb{Z}[i]\setminus\{0\}\\ \gamma\in\Gamma_n\\ \log\|g^{-1}\tilde{\gamma}g\|\leqslant 8\sqrt{\log\ell}}}\frac{|x_n|}{|n|}\sup_{\nu\in i\mathbb{R}}|\varphi^{\ell}_{\nu,\ell}(g^{-1}\tilde{\gamma}g)| + L^{2+\varepsilon}\ell^{-48}. \tag{3.22}$$

The bound (3.22) explicitly reduces the non-spherical sup-norm problem of estimating $\sum_{\phi\in\mathfrak{B}}|\phi(g)|^2$ via the amplification method to two ingredients:

- estimates on $\varphi^{\ell}_{\nu,\ell}(g^{-1}\tilde{\gamma}g)$ for $g^{-1}\tilde{\gamma}g\in G$ of moderate size;
- counting $\gamma\in\Gamma_n$ according to the size of $\varphi^{\ell}_{\nu,\ell}(g^{-1}\tilde{\gamma}g)$.

We now also derive a version of (3.22) adapted to estimating a single form $|\phi_q(g)|^2$ for some $|q|\leqslant\ell$. With the specific $f\in\mathcal{H}(\tau_\ell)_\infty$ provided by (2.22) and (3.14), we obtain by averaging as in (3.6) the test function

$$f_q(g) := \frac{1}{(2\ell+1)\pi^2}\sum_{|p|\leqslant\ell}\int_0^\infty e^{(p^2-\ell^2-t^2)/2}\,\varphi^{\ell,q}_{it,p}(g^{-1})\,(t^2+p^2)\,\mathrm{d}t,$$

where

$$\varphi^{\ell,q}_{\nu,p}(g) := \frac{1}{2\pi}\int_0^{2\pi}\varphi^{\ell}_{\nu,p}\big(g\,\mathrm{diag}(e^{i\varrho},e^{-i\varrho})\big)\,e^{-2qi\varrho}\,\mathrm{d}\varrho.$$

In particular, this definition generalizes (1.7), and by (2.20) we have the symmetry

$$\varphi^{\ell,-q}_{\nu,p}(g) = \overline{\varphi^{\ell,q}_{-\overline{\nu},p}(g)} = \varphi^{\ell,q}_{\nu,p}(g^{-1}). \tag{3.23}$$

The analogues of (3.16)–(3.17) clearly hold for the $\mathbb{R}/\mathbb{Z}$-average f_q, hence by (3.13) the following analogue of (3.22) holds as well:

$$|\phi_q(g)|^2 \ll_{\varepsilon,I,\Omega} L^{-2+\varepsilon}\ell^2 \sum_{\substack{n\in\mathbb{Z}[i]\setminus\{0\}\\ \gamma\in\Gamma_n\\ \log\|g^{-1}\tilde{\gamma}g\|\leqslant 8\sqrt{\log\ell}}}\frac{|x_n|}{|n|}\sup_{\nu\in i\mathbb{R}}|\varphi^{\ell,q}_{\nu,\ell}(g^{-1}\tilde{\gamma}g)| + L^{2+\varepsilon}\ell^{-48}. \tag{3.24}$$

3.4. A double pre-trace formula and a fourth moment

In this subsection, we use a different argument, outlined in §1.7, to estimate values $|\phi_q(g)|$ in terms of Diophantine counts of *pairs* of Hecke correspondences and estimates on generalized spherical functions; see (3.29) and (3.30) below. The argument, reminiscent of self-amplification, relies on using diagonal detection of Voronoi type of §2.8 in a double pre-trace formula (see (3.25) below) to get a handle on the fourth spectral moment of $|\phi_q(g)|$.

Let us fix two integers $\ell, q \in \mathbb{Z}$ with $\ell \geqslant \max(1, |q|)$. Let $n \in \mathbb{Z}[i] \setminus \{0\}$ and $g \in G$. By (3.8) and the remarks below it, for any $f \in \mathcal{H}(\tau_\ell)_\infty$ we have

$$\int_{[\ell]} \widehat{f}(V)\lambda_n(V)|\phi^V_{\ell,q}(g)|^2 \,\mathrm{d}V = \frac{2\ell+1}{|n|}\sum_{\gamma\in\Gamma_n} f_q(g^{-1}\tilde{\gamma}g).$$

It is straightforward to adapt, first the two-variable versions of (2.15), (2.18), and (2.41), and then the proof of the above pre-trace formula to yield the following two-variable version. Let $n_1, n_2 \in \mathbb{Z}[i] \setminus \{0\}$ and $g_1, g_2 \in G$. Then for any $f \in \mathcal{H}(\tau_\ell, \tau_\ell)_\infty$ (recalling the notation introduced after Theorem 12) we have

$$\begin{aligned}\int_{[\ell]}\int_{[\ell]} \widehat{f}(V_1,V_2)\lambda_{n_1}(V_1)\lambda_{n_2}(V_2)|\phi^{V_1}_{\ell,q}(g_1)|^2|\phi^{V_2}_{\ell,q}(g_2)|^2 \,\mathrm{d}V_1\,\mathrm{d}V_2 \\ = \frac{(2\ell+1)^2}{|n_1n_2|}\sum_{\gamma_1\in\Gamma_{n_1}}\sum_{\gamma_2\in\Gamma_{n_2}} f_q(g_1^{-1}\tilde{\gamma}_1 g_1, g_2^{-1}\tilde{\gamma}_2 g_2),\end{aligned} \tag{3.25}$$

where $\widehat{f}(V_1, V_2)$ is given by (2.30) when V_j is of type $(\nu_j, p_j) \in i\mathbb{R}\times\mathbb{Z}$, and

$$f_q(g_1,g_2) := \frac{1}{(2\pi)^2}\int_0^{2\pi}\int_0^{2\pi} f\big(g_1\,\mathrm{diag}(e^{i\varrho_1}, e^{-i\varrho_1}), g_2\,\mathrm{diag}(e^{i\varrho_2}, e^{-i\varrho_2})\big)\, e^{2qi(\varrho_1+\varrho_2)}\,\mathrm{d}\varrho_1\,\mathrm{d}\varrho_2.$$

In (3.25), we can restrict to pairs (V_1, V_2) satisfying $\lambda_i(V_j) = 1$ by introducing an averaging over $\{n_1, in_1\}\times\{n_2, in_2\}$:

$$\begin{aligned}\int_{[\ell]'}\int_{[\ell]'} \widehat{f}(V_1,V_2)\lambda_{n_1}(V_1)\lambda_{n_2}(V_2)|\phi^{V_1}_{\ell,q}(g_1)|^2|\phi^{V_2}_{\ell,q}(g_2)|^2 \,\mathrm{d}V_1\,\mathrm{d}V_2 \\ = \frac{(2\ell+1)^2}{4|n_1n_2|}\sum_{\gamma_1\in\Gamma_{n_1}\cup\Gamma_{in_1}}\sum_{\gamma_2\in\Gamma_{n_2}\cup\Gamma_{in_2}} f_q(g_1^{-1}\tilde{\gamma}_1 g_1, g_2^{-1}\tilde{\gamma}_2 g_2).\end{aligned} \tag{3.26}$$

The prime symbol in $[\ell]'$ indicates that we sum-integrate over automorphic representations with a lift to $\mathrm{PGL}_2(\mathbb{Z}[i])\backslash\mathrm{PGL}_2(\mathbb{C})$, so that the results of §2.7 and §2.8 are applicable.

Now we consider, for any $n \in \mathbb{Z}[i] \setminus \{0\}$, the spectral weights

$$H(V_1, V_2; n) := h(\nu_1, p_1)h(\nu_2, p_2)W_\ell\left(\frac{|n|}{\ell}, \vec{\nu}, \vec{p}\right),$$

where h is as in (3.14) and W_ℓ is as in Lemma 3. Combining the Hilbert space isomorphism

$$\mathcal{H}(\tau_\ell)\hat{\otimes}\mathcal{H}(\tau_\ell) \longleftrightarrow L^2(\widehat{G}_{\mathrm{temp}}(\tau_\ell)\times\widehat{G}_{\mathrm{temp}}(\tau_\ell))$$

induced by Theorem 10 with Theorem 12 and parts (a)–(b) of Lemma 3, we see that the function $(g_1, g_2) \mapsto \breve{H}(g_1, g_2; n)$ given by (2.31) belongs to $\mathcal{H}(\tau_\ell, \tau_\ell)_\infty$, and its double τ_ℓ-spherical transform equals $H(V_1, V_2; n)$. Therefore, applying (3.26) with $f = \breve{H}(\cdot, \cdot; n)$, $n_1 = n_2 = n$, and $g_1 = g_2 = g$, and then summing up over n, we arrive at

$$\begin{aligned}\sum_{n\in\mathbb{Z}[i]\setminus\{0\}} \int_{[\ell]'}\int_{[\ell]'} H(V_1,V_2;n)\lambda_n(V_1)\lambda_n(V_2)|\phi_{\ell,q}^{V_1}(g)|^2|\phi_{\ell,q}^{V_2}(g)|^2\,\mathrm{d}V_1\,\mathrm{d}V_2 \\ = \sum_{n\in\mathbb{Z}[i]\setminus\{0\}} \frac{(2\ell+1)^2}{4|n|^2} \sum_{\gamma_1,\gamma_2\in\Gamma_n\cup\Gamma_{in}} \breve{H}_q(g^{-1}\tilde{\gamma}_1 g, g^{-1}\tilde{\gamma}_2 g; n).\end{aligned} \tag{3.27}$$

By Lemma 3(c), the left-hand side of (3.27) equals

$$\frac{\pi}{4}\ell^2 \sum_{\substack{V \text{ cuspidal}\\ T_i(V)=1,\ |p_V|\leqslant\ell}} h(\nu_V,p_V)^2\, L(1,\mathrm{ad}^2\, V)\,|\phi_{\ell,q}^V(g)|^4 \ +\ \text{Eis}, \tag{3.28}$$

where the term Eis is the contribution of Eisenstein representations:

$$\begin{aligned}\text{Eis} = \ell^2 \sum_{\epsilon,\eta\in\{\pm1\}} \sum_{\substack{p\in 4\mathbb{Z}\\ |p|\leqslant\ell}} \int_{(0)}\int_{(0)} \ell^{\eta(\nu_1-\epsilon\nu_2)}\, h(\nu_1,\epsilon p)h(\nu_2,p)\,\mathcal{L}_\eta((\nu_1,\epsilon p),(\nu_2,p)) \\ \times |\phi_{\ell,q}^{E(\nu_1,\epsilon p)}(g)|^2|\phi_{\ell,q}^{E(\nu_2,p)}(g)|^2\,\frac{\mathrm{d}\nu_1}{\pi i}\,\frac{\mathrm{d}\nu_2}{\pi i}.\end{aligned}$$

We make a change of variable $(\nu_1, \nu_2, p) \mapsto (\eta\nu_1, \eta\epsilon\nu_2, \eta\epsilon p)$. By invariance, we can replace the resulting pairs $(\eta\nu_1, \eta p)$ and $(\eta\epsilon\nu_2, \eta\epsilon p)$ by (ν_1, p) and (ν_2, p), respectively. In this way we see that

$$\begin{aligned}\text{Eis} = 4\ell^2 \sum_{\substack{p\in 4\mathbb{Z}\\ |p|\leqslant\ell}} \int_{(0)}\int_{(0)} \ell^{\nu_1-\nu_2}\, h(\nu_1,p)h(\nu_2,p)\,\mathcal{L}_\eta((\nu_1,p),(\nu_2,p)) \\ \times |\phi_{\ell,q}^{E(\nu_1,p)}(g)|^2|\phi_{\ell,q}^{E(\nu_2,p)}(g)|^2\,\frac{\mathrm{d}\nu_1}{\pi i}\,\frac{\mathrm{d}\nu_2}{\pi i}.\end{aligned}$$

By Lemma 2, we conclude that Eis $\geqslant 0$. In particular, the right-hand side of (3.27) is real, and it provides an upper bound for the contribution of each cuspidal V in (3.28):

$$h(\nu_V,p_V)^2\, L(1,\mathrm{ad}^2\, V)\,|\phi_{\ell,q}^V(g)|^4 \ll \sum_{n\in\mathbb{Z}[i]\setminus\{0\}} \frac{1}{|n|^2} \sum_{\gamma_1,\gamma_2\in\Gamma_n\cup\Gamma_{in}} \breve{H}_q(g^{-1}\tilde{\gamma}_1 g, g^{-1}\tilde{\gamma}_2 g; n).$$

Here we can restrict the n-sum to $|n| \leqslant \ell^{1+\varepsilon}$ at the cost of an error of $\mathrm{O}_\varepsilon(\ell^{-50})$. Indeed, the contributions of $|n| > \ell^{1+\varepsilon}$ on the two sides of (3.27) are equal, and this contribution is $\mathrm{O}_\varepsilon(\ell^{-50})$ thanks to the bound $H(V_1, V_2; n) \ll_A (|n|/\ell)^{-A}$ for any $A > 0$ that follows from Lemma 3(b) and the exponential decay in (3.14).

We now further explicate this bound within the context of Theorems 1–3 (in particular, in preparation for use in Theorem 3(b)). Let $I \subset i\mathbb{R}$ and $\Omega \subset G$ be compact subsets. We fix a cuspidal automorphic representation $V \subset L^2(\Gamma\backslash G)$ with $\nu_V \in I$, $p_V = \ell$, $\lambda_i(V) = 1$, and we pick a cusp form $\phi_q \in V^{\ell,q}$ with $\|\phi_q\|_2 = 1$. We shall also assume that $g \in \Omega$. By (2.44) and our findings above,

$$|\phi_q(g)|^4 \ll_{\varepsilon,I} \ell^\varepsilon \sum_{\substack{n\in\mathbb{Z}[i]\setminus\{0\}\\ |n|\leqslant\ell^{1+\varepsilon}}} \frac{1}{|n|^2} \sum_{\gamma_1,\gamma_2\in\Gamma_n\cup\Gamma_{in}} \breve{H}_q(g^{-1}\tilde{\gamma}_1 g, g^{-1}\tilde{\gamma}_2 g; n) + \ell^{-50}. \tag{3.29}$$

We will analyze the right-hand side of (3.29) to localize γ_1, γ_2 which contribute non-negligibly, and to bound these contributions in terms of generalized spherical functions $\varphi_{\nu,\ell}^{\ell,q}$.

We estimate $\breve{H}$ (hence also $\breve{H}_q$) in terms of Cartan coordinates using the two-dimensional analogue of (2.29):

$$\begin{aligned}\big|\breve{H}(k_1 a_{h_1} k_2, k_3 a_{h_2} k_4; n)\big| \leqslant \sum_{|p_1|,|p_2|\leqslant \ell} \iint\limits_{\substack{s_1>h_1\\ s_2>h_2}} \Bigg| \iint\limits_{\substack{t_1\in\mathbb{R}\\ t_2\in\mathbb{R}}} W_\ell\Big(\frac{|n|}{\ell}, (it_1, it_2), (p_1,p_2)\Big)\\ \times e^{-\ell^2+(p_1^2+p_2^2)/2}\, e^{-(t_1^2+t_2^2)/2}\, e^{-it_1 s_1 - it_2 s_2}\, (t_1^2+p_1^2)(t_2^2+p_2^2)\, \mathrm{d}t_1\, \mathrm{d}t_2 \Bigg| \mathrm{d}s_1\, \mathrm{d}s_2.\end{aligned}$$

This estimate holds for $k_j \in K$ and $h_j > 1$. Assuming without loss of generality that $s_1 \geqslant s_2 > 0$ and shifting the t_1-contour, we conclude from Lemma 3(b) that for any $\varepsilon, B > 0$ the inner double integral is

$$\begin{aligned}&\ll_{\varepsilon,B} \left(1 + \frac{(1+|p_1+p_2|)(1+|p_1-p_2|)}{\ell}\right)^{2B-2+2\varepsilon} e^{-\ell^2+(p_1^2+p_2^2)/2} \ell^4 e^{-Bs_1} \left(\frac{|n|}{\ell}\right)^{-B-\varepsilon}\\ &\ll_{\varepsilon,B} \frac{\ell^{4+\varepsilon} e^{-B\max(s_1,s_2)}}{\big(1+\ell-|p_1|\big)^2\big(1+\ell-|p_2|\big)^2} \left(\frac{|n|}{\ell}\right)^{-B}.\end{aligned}$$

It follows that

$$\breve{H}(k_1 a_{h_1} k_2, k_3 a_{h_2} k_4; n) \ll_{\varepsilon,B} \ell^{4+\varepsilon} e^{-B\max(h_1,h_2)} \big(\ell/|n|\big)^B$$

for any $\varepsilon, B > 0$ and $h_1, h_2 > 1$. This estimate remains true for general $h_1, h_2 \geqslant 0$, as can be seen by using (2.22) and the trivial bound $|\varphi_{it,p}^{\ell}| \leqslant 2\ell + 1$ instead of (2.29) for the respective variable if one or both of h_1, h_2 are at most 1. The same bound applies for $\breve{H}_q$, that is,

$$\breve{H}_q(g_1, g_2; n) \ll_{\varepsilon,B} \ell^{4+\varepsilon} \left(\frac{\ell/|n|}{\|g_1\|^2 + \|g_2\|^2}\right)^B$$

for any $\varepsilon, B > 0$ and $g_1, g_2 \in G$. So we can refine (3.29) to

$$|\phi_q(g)|^4 \ll_{\varepsilon,I,\Omega} \ell^{\varepsilon} \sum_{\substack{n\in\mathbb{Z}[i]\setminus\{0\}\\ |n|\leqslant \ell^{1+\varepsilon}}} \frac{1}{|n|^2} \sum_{\substack{\gamma_1,\gamma_2\in\Gamma_n\cup\Gamma_{in}\\ \|g^{-1}\tilde{\gamma}_j g\|\leqslant \ell^\varepsilon\sqrt{\ell/|n|}}} \breve{H}_q(g^{-1}\tilde{\gamma}_1 g, g^{-1}\tilde{\gamma}_2 g; n) + \ell^{-50}.$$

In the last sum, we estimate the terms more directly by (2.31), (3.23), and Lemma 3(b):

$$\breve{H}_q(g^{-1}\tilde{\gamma}_1 g, g^{-1}\tilde{\gamma}_2 g; n) \ll_{\varepsilon} \ell^{2+\varepsilon} F(\gamma_1) F(\gamma_2) + \ell^{-80},$$

where we temporarily abbreviate (suppressing g and q from the notation)

$$F(\gamma) := \sup_{\nu\in i\mathbb{R}} |\varphi_{\nu,\ell}^{\ell,q}(g^{-1}\tilde{\gamma} g)|, \qquad \gamma \in \mathrm{GL}_2(\mathbb{C}).$$

Recalling also (3.21), we obtain an inequality of bilinear type:

$$|\phi_q(g)|^4 \ll_{\varepsilon,I,\Omega} \ell^{2+\varepsilon} \sum_{\substack{n\in\mathbb{Z}[i]\setminus\{0\}\\ |n|\leqslant \ell^{1+\varepsilon}}} \frac{1}{|n|^2} \sum_{\substack{\gamma_1,\gamma_2\in\Gamma_n\cup\Gamma_{in}\\ \|g^{-1}\tilde{\gamma}_j g\|\leqslant \ell^\varepsilon\sqrt{\ell/|n|}}} F(\gamma_1) F(\gamma_2) + \ell^{-50}.$$

With the shorthand notation

$$S(n) := \sum_{\substack{\gamma\in\Gamma_n \\ \|g^{-1}\tilde\gamma g\|\leqslant \ell^\varepsilon\sqrt{\ell/|n|}}} F(\gamma),$$

we observe that the innermost sum in the previous display equals $(S(n)+S(in))^2$, hence it does not exceed $2(S(n)^2+S(in)^2)$. In the end, we conclude

$$|\phi_q(g)|^4 \ll_{\varepsilon,I,\Omega} \ell^{2+\varepsilon} \sum_{\substack{n\in\mathbb{Z}[i]\setminus\{0\} \\ |n|\leqslant \ell^{1+\varepsilon}}} \frac{1}{|n|^2} \sum_{\substack{\gamma_1,\gamma_2\in\Gamma_n \\ \|g^{-1}\tilde\gamma_j g\|\leqslant \ell^\varepsilon\sqrt{\ell/|n|}}} \prod_{j=1}^{2} \sup_{\nu\in i\mathbb{R}} |\varphi_{\nu,\ell}^{\ell,q}(g^{-1}\tilde\gamma_j g)| + \ell^{-50}, \tag{3.30}$$

which serves as an analogue of (3.24).

3.5. Reduction to Diophantine counting

In this subsection, we input into the preliminary estimates (3.22), (3.24) and (3.30) the results of Theorems 4, 5 and 6, which provide the desired estimates on spherical trace functions. We shall assume (as we can) that ℓ is sufficiently large in terms of ε.

We begin by explicating the estimate (3.22) using (3.18) and Theorem 4. For $\mathcal{L}\geqslant 1$ and $\vec\delta=(\delta_1,\delta_2)\in\mathbb{R}^2_{>0}$, let

$$D(L,\mathcal{L}) := \left\{ n\in\mathbb{Z}[i] : \mathcal{L}\leqslant |n|^2\leqslant 16\mathcal{L},\ \begin{array}{c} n=1 \text{ or } n=l_1l_2 \text{ or } n=l_1^2l_2^2 \\ \text{for some } l_1,l_2\in P(L)\end{array}\right\},$$

$$M(g,L,\mathcal{L},\vec\delta) := \sum_{n\in D(L,\mathcal{L})} \#\left\{\gamma\in\Gamma_n : g^{-1}\tilde\gamma g = k\begin{pmatrix} z & u \\ & z^{-1}\end{pmatrix}k^{-1} \ \begin{array}{l}\text{for some } k\in K,\ |z|\geqslant 1, \\ \min|z\pm 1|\leqslant\delta_1,\ |u|\leqslant\delta_2\end{array}\right\}.$$

Note that every element of G is of the form $k\left(\begin{smallmatrix} z & u \\ & z^{-1}\end{smallmatrix}\right)k^{-1}$ for some $k\in K$, $|z|\geqslant 1$, and $u\in\mathbb{C}$. Indeed, such a decomposition is immediate with $k\in G$, $z\in\mathbb{C}^\times$, and $u=0$ unless $z=\pm1$, after which the claim follows by replacing k by $k\left(\begin{smallmatrix} & -1 \\ 1 & \end{smallmatrix}\right)$ if needed and using the Iwasawa decomposition of k.

Thus to each γ occurring in (3.22) we may associate a dyadic vector $\vec\delta=(\delta_1,\delta_2)$ (that is, $\log_2\delta_j\in\mathbb{Z}$) such that $1/\sqrt{\ell}\leqslant\delta_j\leqslant\ell^\varepsilon$ and δ_j are minimal such that γ is counted in the corresponding $M(g,L,\mathcal{L},\vec\delta)$. Therefore, applying (3.18) and the estimates of Theorem 4 in (3.22) leads to the following result.

Lemma 4. *Let $\ell\geqslant 1$ be an integer, $I\subset i\mathbb{R}$ and $\Omega\subset G$ be compact sets. Let $V\subset L^2(\Gamma\backslash G)$ be a cuspidal automorphic representation with minimal K-type τ_ℓ and spectral parameter $\nu_V\in I$. Let $\mathfrak{B}$ be an orthonormal basis of V^ℓ, and let $g\in\Omega$. Then for any $L\geqslant 7$ and $\varepsilon>0$ we have*

$$\sum_{\phi\in\mathfrak{B}} |\phi(g)|^2 \ll_{\varepsilon,I,\Omega} \ell^{3+\varepsilon}L^\varepsilon \sum_{\substack{\vec\delta \text{ dyadic} \\ 1/\sqrt{\ell}\leqslant\delta_j\leqslant\ell^\varepsilon}} \min\left(\frac{1}{\ell\delta_1^2},\frac{1}{\sqrt{\ell}\delta_2}\right) \times\left(\frac{M(g,L,1,\vec\delta)}{L}+\frac{M(g,L,L^2,\vec\delta)}{L^3}+\frac{M(g,L,L^4,\vec\delta)}{L^4}\right)+L^{2+\varepsilon}\ell^{-48}.$$

Lemma 4 is free of any choices of the test function, amplifier, and spherical trace function. It reduces the estimation of $\sum_{\phi\in\mathfrak{B}}|\phi(g)|^2$ to the Diophantine counting problem of estimating $M(g,L,\mathcal{L},\vec\delta)$ uniformly in L, $\mathcal{L}$, and $\vec\delta$.

Now, we similarly explicate the estimate (3.24) using (3.18) and Theorems 5–6(a). Recall the sets $\mathcal{D} \subset G$ and $\mathcal{S} \subset K \subset \mathcal{N} \subset G$ introduced before Theorem 6. With $D(L, \mathcal{L})$ as above, we define for $|q| \leqslant \ell$, $\mathcal{L} \geqslant 1$, $\delta > 0$, and $\vec{\delta} = (\delta_1, \delta_2) \in \mathbb{R}^2_{>0}$, the matrix counts

$$M_0^*(g, L, \mathcal{L}, \delta) := \sum_{n \in D(L,\mathcal{L})} \# \left\{ \gamma \in \Gamma_n : \operatorname{dist}(g^{-1}\tilde{\gamma} g, \mathcal{S}) \leqslant \delta,\ \frac{D(g^{-1}\tilde{\gamma} g)}{\|g^{-1}\tilde{\gamma} g\|^2} \ll \frac{\log \ell}{\sqrt{\ell}} \right\},$$

$$M^*(g, L, \mathcal{L}, \vec{\delta}) := \sum_{n \in D(L,\mathcal{L})} \# \left\{ \gamma \in \Gamma_n : \operatorname{dist}\left(g^{-1}\tilde{\gamma} g, K\right) \leqslant \delta_1,\ \operatorname{dist}(g^{-1}\tilde{\gamma} g, \mathcal{D}) \leqslant \delta_2 \right\},$$

with a sufficiently large implied constant in the definition of $M_0^*(g, L, \mathcal{L}, \delta)$.

For $q = 0$, we estimate the size of $\varphi_{\nu,\ell}^{\ell,q}(g^{-1}\tilde{\gamma} g)$ in (3.24) using Theorem 6(a). Since there are at most $\mathrm{O}_{\varepsilon,\Omega}(\ell^\varepsilon |n|^{2+\varepsilon})$ elements $\gamma \in \Gamma_n$ contributing to the right-hand side of (3.24), the total contribution of those elements which fail to satisfy $D(g^{-1}\tilde{\gamma} g) \ll \|g^{-1}\tilde{\gamma} g\|^2 (\log \ell)/\sqrt{\ell}$ with a sufficiently large implied constant may be absorbed into the existing $\mathrm{O}_{\varepsilon,I,\Omega}(L^{2+\varepsilon}\ell^{-48})$ error term. We may thus restrict to $\gamma \in \Gamma_n$ satisfying these conditions. We associate to each remaining γ in (3.24) the smallest dyadic $1/\sqrt{\ell} \leqslant \delta \leqslant \ell^\varepsilon$ such that γ is counted in the corresponding $M_0^*(g, L, \mathcal{L}, \delta)$. For a general $|q| \leqslant \ell$, we associate to each γ in (3.24) the lexicographically smallest dyadic vector $\vec{\delta} = (\delta_1, \delta_2)$ such that $\delta_j \leqslant \ell^\varepsilon$ and $\delta_1^2 \delta_2 \geqslant 1/\sqrt{\ell}$ and γ is counted in the corresponding $M_q(g, L, \mathcal{L}, \vec{\delta})$. Applying (3.18) and the estimates of Theorems 5–6(a) in (3.24) leads to the following result.

Lemma 5. *Let $\ell \geqslant 1$ be an integer, $I \subset i\mathbb{R}$ and $\Omega \subset G$ be compact sets. Let $V \subset L^2(\Gamma\backslash G)$ be a cuspidal automorphic representation with minimal K-type τ_ℓ and spectral parameter $\nu_V \in I$. Let $\phi_q \in V^{\ell,q}$ such that $\|\phi_q\|_2 = 1$ and let $g \in \Omega$. Then for any $L \geqslant 7$ and $\varepsilon > 0$ we have*

$$|\phi_0(g)|^2 \ll_{\varepsilon,I,\Omega} \ell^{2+\varepsilon} L^\varepsilon \sum_{\substack{\delta \text{ dyadic} \\ 1/\sqrt{\ell} \leqslant \delta \leqslant \ell^\varepsilon}} \frac{1}{\sqrt{\ell}\delta} \times \left(\frac{M_0^*(g, L, 1, \delta)}{L} + \frac{M_0^*(g, L, L^2, \delta)}{L^3} + \frac{M_0^*(g, L, L^4, \delta)}{L^4} \right) + L^{2+\varepsilon}\ell^{-48}.$$

Moreover, for $|q| \leqslant \ell$ we have

$$|\phi_q(g)|^2 \ll_{\varepsilon,I,\Omega} \ell^{2+\varepsilon} L^\varepsilon \sum_{\substack{\vec{\delta} \text{ dyadic},\ \delta_j \leqslant \ell^\varepsilon \\ \delta_1^2\delta_2 \geqslant 1/\sqrt{\ell}}} \frac{1}{\sqrt{\ell}\delta_1^2\delta_2} \times \left(\frac{M^*(g, L, 1, \vec{\delta})}{L} + \frac{M^*(g, L, L^2, \vec{\delta})}{L^3} + \frac{M^*(g, L, L^4, \vec{\delta})}{L^4} \right) + L^{2+\varepsilon}\ell^{-48}.$$

Similarly, we explicate (3.30) using Theorem 6(b). Here we introduce the double matrix count

$$Q(g, L, H_1, H_2) := \sum_{L \leqslant |n| \leqslant 2L} \# \left\{ (\gamma_1, \gamma_2) \in \Gamma_n^2 :\ \|g^{-1}\tilde{\gamma}_j g\| \leqslant \sqrt{\frac{H_j}{L}},\ \operatorname{dist}(g^{-1}\tilde{\gamma}_j g, \mathcal{D}) \ll \sqrt{\frac{H_j \log \ell}{L\ell}} \right\},$$

with a sufficiently large implied constant in the distance condition.

Lemma 6. *Let $\ell \geqslant 1$ be an integer, $I \subset i\mathbb{R}$ and $\Omega \subset G$ be compact sets. Let $V \subset L^2(\Gamma\backslash G)$ be a cuspidal automorphic representation with minimal K-type τ_ℓ and spectral parameter $\nu_V \in I$. Suppose that V lifts to*

an automorphic representation for $\mathrm{PGL}_2(\mathbb{Z}[i])\backslash\mathrm{PGL}_2(\mathbb{C})$. *Let* $\phi_{\pm\ell} \in V^{\ell,\pm\ell}$ *such that* $\|\phi_{\pm\ell}\|_2 = 1$ *and let* $g \in \Omega$. *Then for any* $\varepsilon > 0$ *we have*

$$|\phi_{\pm\ell}(g)|^4 \ll_{\varepsilon,I,\Omega} \ell^{2+\varepsilon} \max_{1\leqslant L,H_1,H_2\leqslant \ell^{1+\varepsilon}} \frac{Q(g,L,H_1,H_2)}{H_1H_2} + \ell^{-50}.$$

4. Proof of Theorem 4

In this section, we prove Theorem 4. It is clear from the definition (2.17) that we can restrict to $k = 1$ without loss of generality, and the first bound holds in the stronger form $|\varphi^{\ell}_{\nu,\ell}(g)| \leqslant 2\ell + 1$. In particular, Theorem 4 is trivial for $\ell = 1$, hence we shall assume (for notational simplicity) that $\ell \geqslant 2$. In addition, the exponential factor in (1.5) has absolute value less than $\|g\|^2$ thanks to (2.5) and the identity

$$|ad - bc|^2 + |a\bar{b} + c\bar{d}|^2 = (|a|^2 + |c|^2)(|b|^2 + |d|^2),$$

hence it suffices to prove that

$$\int_K |\psi_\ell(\kappa(k^{-1}gk))|\,\mathrm{d}k \ll_\varepsilon \ell^\varepsilon \min\left(\frac{\|g\|^4}{|z^2-1|^2\ell}, \frac{\|g\|}{|u|\sqrt{\ell}}\right). \tag{4.1}$$

Finally, we shall use the obvious fact that

$$|u|, |z|, |z^{-1}| \leqslant \|g\|. \tag{4.2}$$

Writing $k = k[\phi,\theta,\psi]$ in Euler angles as in (2.1), and setting

$$x := (z^2-1)\cos\theta + ie^{-2i\phi}uz\sin\theta,$$

one computes

$$k[\phi,\theta,\psi]^{-1}gk[\phi,\theta,\psi] = \begin{pmatrix} (1+x\cos\theta)/z & * \\ -ie^{2i\psi}x\sin\theta/z & * \end{pmatrix}.$$

Our goal is to estimate then

$$\int_0^\pi \int_0^{\pi/2} \int_{-\pi}^{\pi} \left|\psi_\ell\left(\kappa\left(\begin{pmatrix} (1+x\cos\theta)/z & * \\ -ie^{2i\psi}x\sin\theta/z & * \end{pmatrix}\right)\right)\right| \sin 2\theta\,\mathrm{d}\psi\,\mathrm{d}\theta\,\mathrm{d}\phi. \tag{4.3}$$

We introduce the notation $\lambda := \sqrt{\log \ell}$.

4.1. Small values of the integrand

First we identify a region where $|\psi_\ell|$ in the integral (4.3) is small. Assume that

$$\min\big(\tan\theta, |x|\sin\theta\big) > \frac{4\lambda}{\sqrt{\ell}}. \tag{4.4}$$

Then in

$$\kappa\left(\begin{pmatrix} (1+x\cos\theta)/z & * \\ -ie^{2i\psi}x\sin\theta/z & * \end{pmatrix}\right) = \begin{pmatrix} \frac{1+x\cos\theta}{\sqrt{|1+x\cos\theta|^2+|x\sin\theta|^2}} & * \\ * & * \end{pmatrix} \in K$$

the upper left entry has absolute square less than $1-\lambda^2/\ell$, hence

$$\left|\psi_\ell\left(\kappa\left(\begin{pmatrix}(1+x\cos\theta)/z & *\\ -ie^{2i\psi}x\sin\theta/z & *\end{pmatrix}\right)\right)\right| < \left(1-\frac{\log\ell}{\ell}\right)^\ell < \frac{1}{\ell}.$$

In view of (4.2), this is admissible for (4.1). In the next subsection, we consider the case when (4.4) fails.

4.2. Large values of the integrand

Assume first that $\tan\theta \leqslant 4\lambda/\sqrt{\ell}$. Then $\theta \leqslant 4\lambda/\sqrt{\ell}$, hence the corresponding contribution to (4.3) is $\ll \lambda^2/\ell$. This is admissible for (4.1) in the light of (4.2).

Now assume that $|x|\sin\theta \leqslant 4\lambda/\sqrt{\ell}$, and decompose the relevant integration domain for θ as follows. For any $m,n\in\mathbb{Z}_{\geqslant 0}$ and $\phi\in[0,\pi]$, let

$$I(m,n,\phi) := \left\{\theta\in\left(0,\frac{\pi}{2}\right) \,:\, |x|\sin\theta\leqslant\frac{4\lambda}{\sqrt{\ell}},\ \frac{1}{2}<2^m\sin\theta\leqslant 1,\ \frac{1}{2}<2^n\cos\theta\leqslant 1\right\}.$$

If $\theta\notin I(m,n,\phi)$ holds for every $0\leqslant m,n\leqslant 2\log\ell$, then $\sin 2\theta = 2\sin\theta\cos\theta\leqslant 1/\ell$, which is admissible for (4.1). Therefore, by (4.2) and (4.3), it suffices to prove the bound

$$\int_0^\pi \int_{I(m,n,\phi)} \sin 2\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \ll \min\left(\frac{\lambda^2}{\ell|z^2-1|^2}, \frac{\lambda}{\ell^{1/2}|uz|}\right) \tag{4.5}$$

for every $0\leqslant m,n\leqslant 2\log\ell$. We shall assume that $\min(m,n)=0$, for otherwise $I(m,n,\phi)=\emptyset$. We record also that the Lebesgue measure of $I(m,n,\phi)$ is $\mathrm{O}(2^{-m-n})$, because if $n=0$, then $\sin\theta\asymp\theta$, while if $m=0$, then $\cos\theta\asymp\pi/2-\theta$. Hence, for any $\phi\in[0,\pi]$, we have

$$\int_{I(m,n,\phi)} \sin 2\theta\,\mathrm{d}\theta = \int_{I(m,n,\phi)} 2\sin\theta\cos\theta\,\mathrm{d}\theta \ll 2^{-2m-2n}.$$

First consider the case when in $x=(z^2-1)\cos\theta+ie^{-2i\phi}uz\sin\theta$, whose absolute value does not exceed $2^{m+3}\lambda/\sqrt{\ell}$, neither of the two summands is large:

$$|z^2-1|2^{-n}\leqslant 2^{m+6}\frac{\lambda}{\sqrt{\ell}}, \qquad |uz|2^{-m}\leqslant 2^{m+6}\frac{\lambda}{\sqrt{\ell}}.$$

Recalling $\min(m,n)=0$, the previous two displays imply for any $\phi\in[0,\pi]$ that

$$\int_{I(m,n,\phi)} \sin 2\theta\,\mathrm{d}\theta \ll \min\left(\frac{\lambda^2}{\ell|z^2-1|^2}, \frac{\lambda}{\ell^{1/2}|uz|}\right).$$

So in this case (4.5) is clear.

Now consider the case when in $x=(z^2-1)\cos\theta+ie^{-2i\phi}uz\sin\theta$, whose absolute value does not exceed $2^{m+3}\lambda/\sqrt{\ell}$, the two summands are individually large:

$$|z^2-1|2^{-n}>2^{m+4}\frac{\lambda}{\sqrt{\ell}}, \qquad |uz|2^{-m}>2^{m+4}\frac{\lambda}{\sqrt{\ell}}, \qquad |z^2-1|2^{-n}\asymp|uz|2^{-m}. \tag{4.6}$$

We claim that this localizes ϕ. Indeed, setting

$$2\phi_0 = \arg(iuz) - \arg(z^2-1),$$

we see that

$$|z^2-1|\cos\theta + e^{2i(\phi_0-\phi)}|uz|\sin\theta \ll 2^m\frac{\lambda}{\sqrt{\ell}},$$

and comparing the imaginary parts, we have that

$$\sin(2\phi-2\phi_0) \ll \frac{2^{2m}\lambda}{|uz|\sqrt{\ell}}, \quad \text{and so} \quad \phi \equiv \phi_0 + \mathrm{O}\left(\frac{2^{2m}\lambda}{|uz|\sqrt{\ell}}\right) \pmod{\pi/2}.$$

Also, θ is localized, since

$$|z^2-1|\cos\theta - |uz|\sin\theta \ll 2^m\frac{\lambda}{\sqrt{\ell}},$$

and the first term here is monotone decreasing, the second one is monotone increasing in θ. We see that θ is localized to an interval of length $\mathrm{O}(2^m\lambda/|uz|\sqrt{\ell})$ for $\sin\theta \leqslant \cos\theta$ (in which case $n=0$), and to an interval of length $\mathrm{O}(\lambda/|z^2-1|\sqrt{\ell})$ for $\cos\theta \leqslant \sin\theta$ (in which case $m=0$).

We estimate the left-hand side of (4.5) by exploiting the above localizations and all three parts of (4.6). If $\sin\theta \leqslant \cos\theta$, then $n=0$ and $\sin 2\theta \leqslant 2^{1-m}$, so altogether we obtain a contribution to (4.5) of size

$$\ll 2^{-m}\cdot\frac{2^m\lambda}{|uz|\sqrt{\ell}}\cdot\frac{2^{2m}\lambda}{|uz|\sqrt{\ell}} \ll \min\left(\frac{\lambda^2}{|z^2-1|^2\ell}, \frac{\lambda}{|uz|\sqrt{\ell}}\right).$$

Similarly, if $\cos\theta \leqslant \sin\theta$, then $m=0$ and $\sin 2\theta \leqslant 2^{1-n}$, so altogether we obtain a contribution to (4.5) of size

$$\ll 2^{-n}\cdot\frac{\lambda}{|z^2-1|\sqrt{\ell}}\cdot\frac{\lambda}{|uz|\sqrt{\ell}} \ll \min\left(\frac{\lambda^2}{|z^2-1|^2\ell}, \frac{\lambda}{|uz|\sqrt{\ell}}\right).$$

The proof of Theorem 4 is complete.

5. Proof of Theorems 5 and 6

In this section, we prove Theorems 5 and 6. We recall that the key player is the function

$$\varphi^{\ell,q}_{\nu,\ell}(g) := \frac{1}{2\pi}\int_0^{2\pi}\varphi^{\ell}_{\nu,\ell}\big(gk[0,0,\varrho]\big)\,e^{-2qi\varrho}\,\mathrm{d}\varrho, \tag{5.1}$$

where

$$\varphi^{\ell}_{\nu,\ell}(g) := (2\ell+1)\int_K \psi_\ell(\kappa(k^{-1}gk))\,e^{(\nu-1)\rho(H(gk))}\,\mathrm{d}k.$$

The function $\psi_\ell : K\to\mathbb{C}$ was defined in (1.6), but for calculational purposes we extend it now to $\mathrm{GL}_2(\mathbb{C})$:

$$\psi_\ell\left(\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}\right) := \bar\alpha^{2\ell}, \qquad \begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}\in\mathrm{GL}_2(\mathbb{C}). \tag{5.2}$$

5.1. Preliminary computations

We write g in Cartan form

$$g = k[u_1, v_1, w_1] \begin{pmatrix} r & \\ & r^{-1} \end{pmatrix} k[u_2, v_2, w_2], \tag{5.3}$$

where $r \geqslant 1$, and we allow $u_j, v_j, w_j \in \mathbb{R}$ to be arbitrary for convenience. Spelling out the definitions, and using that the height in the Iwasawa decomposition is left K-invariant, we see that $\varphi_{\nu,\ell}^{\ell,q}(g)$ equals

$$\frac{d_\ell}{4\pi^3} \int\limits_{\substack{0\leqslant u\leqslant \pi \\ 0\leqslant v\leqslant \pi/2 \\ 0\leqslant w\leqslant 2\pi \\ 0\leqslant \varrho\leqslant 2\pi}} \psi_\ell\left(k[-w,-v,-u]k[u_1,v_1,w_1]\kappa\left(\begin{pmatrix} r & \\ & r^{-1}\end{pmatrix} k[u_2,v_2,w_2]k[0,0,\varrho]k[u,v,w]\right)\right)$$

$$\cdot\, e^{-2iq\varrho}\, e^{(\nu-1)\rho\left(H\left(\left(\begin{smallmatrix} r & \\ & r^{-1}\end{smallmatrix}\right)k[u_2,v_2,w_2]k[0,0,\varrho]k[u,v,w]\right)\right)} \sin 2v \,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w\,\mathrm{d}\varrho.$$

With a change of variables $k[u_2, v_2, w_2]k[0, 0, \rho]k[u, v, w] \mapsto k[u, v, w]$ and dropping the normalized w-integration (which is legitimate since the conjugation by $k[0, 0, w]$ does not alter the ψ_ℓ-value, and the height in the Iwasawa decomposition is also unaffected by right-multiplication by $k[0, 0, w]$), we arrive at

$$\frac{d_\ell}{2\pi^2} \int\limits_{\substack{0\leqslant u\leqslant \pi \\ 0\leqslant v\leqslant \pi/2 \\ 0\leqslant \varrho\leqslant 2\pi}} \psi_\ell\left(k[0,-v,-u]k[u_2,v_2,w_2]k[0,0,\varrho]k[u_1,v_1,w_1]\kappa\left(\begin{pmatrix} r & \\ & r^{-1}\end{pmatrix} k[u,v,0]\right)\right)$$

$$\cdot\, e^{-2iq\varrho}\, e^{(\nu-1)\rho\left(H\left(\left(\begin{smallmatrix} r & \\ & r^{-1}\end{smallmatrix}\right)k[u,v,0]\right)\right)} \sin 2v \,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}\varrho.$$

The sum of absolute squares in the first column of $\mathrm{diag}(r, r^{-1})k[u, v, 0]$ equals

$$h(r, v) := r^2 \cos^2 v + r^{-2} \sin^2 v,$$

hence recalling the definitions (2.5) and (5.2), we can rewrite the integral as

$$\frac{d_\ell}{2\pi^2} \int\limits_{\substack{0\leqslant u\leqslant \pi \\ 0\leqslant v\leqslant \pi/2 \\ 0\leqslant \varrho\leqslant 2\pi}} \psi_\ell\left(k[0,-v,-u]k[u_2,v_2,w_2]k[0,0,\varrho]k[u_1,v_1,w_1]\begin{pmatrix} r & \\ & r^{-1}\end{pmatrix} k[u,v,0]\right)$$

$$\cdot\, e^{-2iq\varrho}\, h(r,v)^{\nu-1-\ell} \sin 2v \,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}\varrho.$$

Replacing ϱ by $\varrho - u_1 - w_2$, the integral further simplifies to

$$\frac{d_\ell e^{2iq(u_1+w_2)}}{2\pi^2} \int\limits_{\substack{0\leqslant u\leqslant \pi \\ 0\leqslant v\leqslant \pi/2 \\ 0\leqslant \varrho\leqslant 2\pi}} \psi_\ell\left(\begin{pmatrix} e^{-i\varrho}I + e^{i\varrho}J & * \\ * & * \end{pmatrix}\right) e^{-2iq\varrho}\, h(r,v)^{\nu-1-\ell} \sin 2v \,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}\varrho,$$

where

$$I := \left(r^{-1}e^{-2iu-iw_1}\sin v\cos v_1 + re^{iw_1}\cos v\sin v_1\right)\left(e^{2iu-iu_2}\sin v\cos v_2 - e^{iu_2}\cos v\sin v_2\right),$$
$$J := \left(-r^{-1}e^{-2iu-iw_1}\sin v\sin v_1 + re^{iw_1}\cos v\cos v_1\right)\left(e^{2iu-iu_2}\sin v\sin v_2 + e^{iu_2}\cos v\cos v_2\right).$$

Evaluating the ϱ-integral, we obtain

$$\varphi_{\nu,\ell}^{\ell,q}(g) = \frac{d_\ell e^{2iq(u_1+w_2)}}{\pi}\binom{2\ell}{\ell+q} \int\limits_{\substack{0\leqslant u\leqslant \pi\\ 0\leqslant v\leqslant \pi/2}} \frac{\sin 2v}{h(r,v)^{\ell+1-\nu}}\, \bar{I}^{\ell+q}\bar{J}^{\ell-q}\,\mathrm{d}u\,\mathrm{d}v. \tag{5.4}$$

Taking the complex conjugate of the right-hand side, and introducing the new variables $t := r^{-1}\tan v$ and $\phi := 2u$, we get

$$\begin{aligned}\left|\varphi_{\nu,\ell}^{\ell,q}(g)\right| = \frac{d_\ell}{\pi}\binom{2\ell}{\ell+q}\Bigg| \int_0^\infty\int_0^{2\pi} \frac{t}{(1+(t/r)^2)^{\ell+1+\nu}(1+(tr)^2)^{\ell+1-\nu}} \\ \left(e^{-i\phi-2iw_1}(t/r)\cos v_1+\sin v_1\right)^{\ell+q}\left(e^{i\phi-2iu_2}(tr)\cos v_2-\sin v_2\right)^{\ell+q} \\ \left(e^{-i\phi-2iw_1}(t/r)\sin v_1-\cos v_1\right)^{\ell-q}\left(e^{i\phi-2iu_2}(tr)\sin v_2+\cos v_2\right)^{\ell-q}\,\mathrm{d}\phi\,\mathrm{d}t\Bigg|.\end{aligned}$$

Now comes the last key step: in the inner ϕ-integral, we can remove the r's. This is so because $e^{-i\phi}$ must be chosen equally many times as $e^{i\phi}$, and the r's will cancel out in all terms surviving the integration. Another way to see the same thing is to shift the contour as in $\phi \mapsto \phi + i\log r$ where the boundary terms cancel out by 2π-periodicity. Either way, using also the opportunity to replace $\phi \mapsto \phi + u_2 - w_1$, and writing $\Delta := u_2 + w_1$, we finally obtain

$$\begin{aligned}\left|\varphi_{\nu,\ell}^{\ell,q}(g)\right| \leqslant \frac{d_\ell}{\pi}\binom{2\ell}{\ell+q}\int_0^\infty \frac{t}{((1+(t/r)^2)(1+(tr)^2))^{\ell+1}} \\ \times\int_0^{2\pi}\left|e^{i\phi+i\Delta}t\cos v_1+\sin v_1\right|^{\ell+q}\left|e^{i\phi-i\Delta}t\cos v_2-\sin v_2\right|^{\ell+q} \\ \left|e^{i\phi+i\Delta}t\sin v_1-\cos v_1\right|^{\ell-q}\left|e^{i\phi-i\Delta}t\sin v_2+\cos v_2\right|^{\ell-q}\mathrm{d}\phi\,\mathrm{d}t.\end{aligned}$$

We estimate the inner integrand using the following lemma, which is purely about inequalities. We state it formally so as to clearly separate issues. (In the case $q = \pm\ell$, all expressions raised to exponent 0 should simply be omitted.) As in the previous section, we introduce the notation $\lambda := \sqrt{\log \ell}$.

Lemma 7. *Let $\ell, q \in \mathbb{Z}$ be such that $\ell \geqslant \max(1,|q|)$. Let $X > 0$ and $\Lambda > 0$.*

(a) If $A, B \geqslant 0$ satisfy $A^2 + B^2 = X^2$, then

$$\left(\frac{2\ell}{\ell+q}\right)^{(\ell+q)/2}\left(\frac{2\ell}{\ell-q}\right)^{(\ell-q)/2} A^{\ell+q}B^{\ell-q} \leqslant X^{2\ell}. \tag{5.5}$$

Moreover, the left-hand side is $\mathrm{O}_\Lambda(X^{2\ell}\ell^{-\Lambda})$ unless

$$A^2 = \frac{\ell+q}{2\ell}X^2 + \mathrm{O}_\Lambda\left(X^2\frac{\lambda^2+\lambda\sqrt{\ell-|q|}}{\ell}\right), \qquad B^2 = \frac{\ell-q}{2\ell}X^2 + \mathrm{O}_\Lambda\left(X^2\frac{\lambda^2+\lambda\sqrt{\ell-|q|}}{\ell}\right). \tag{5.6}$$

(b) If $A, B, C, D \geqslant 0$ satisfy $A^2 + B^2 = C^2 + D^2 = X^2$, then

$$\binom{2\ell}{\ell+q} A^{\ell+q} B^{\ell-q} C^{\ell+q} D^{\ell-q} \ll \frac{X^{4\ell}}{1+\sqrt{\ell-|q|}}.$$

Moreover, the left-hand side is $\mathrm{O}_\Lambda(X^{4\ell}\ell^{-\Lambda})$ unless (5.6) and the analogous estimates for C, D are satisfied.

Proof. Let us first assume $|q| < \ell$. We use Young's inequality

$$xy \leqslant \frac{x^a}{a} + \frac{y^b}{b}, \qquad \frac{1}{a} + \frac{1}{b} = 1,$$

to conclude with

$$x := \left(\sqrt{\frac{2\ell}{\ell+q}}\frac{A}{X}\right)^{\frac{\ell+q}{\ell}}, \qquad y := \left(\sqrt{\frac{2\ell}{\ell-q}}\frac{B}{X}\right)^{\frac{\ell-q}{\ell}}, \qquad a := \frac{2\ell}{\ell+q}, \qquad b := \frac{2\ell}{\ell-q} \tag{5.7}$$

that

$$\left(\sqrt{\frac{2\ell}{\ell+q}}\frac{A}{X}\right)^{\frac{\ell+q}{\ell}} \left(\sqrt{\frac{2\ell}{\ell-q}}\frac{B}{X}\right)^{\frac{\ell-q}{\ell}} \leqslant \frac{A^2+B^2}{X^2} = 1.$$

This is equivalent to (5.5). We also conclude (still using the notation (5.7)) that the left-hand side of (5.5) is $\mathrm{O}_\Lambda(X^{2\ell}\ell^{-\Lambda})$ unless

$$xy > 1/2, \qquad xy = 1 + \mathrm{O}_\Lambda(\delta), \qquad \delta := \lambda^2/\ell. \tag{5.8}$$

Let us explore the consequences of (5.8). First, by $x^a/a + y^b/b = 1$ we have

$$1/3 < x, y < 3/2.$$

Without loss of generality, $q \geqslant 0$ (i.e. $a \leqslant b$), and then $x^a < a \leqslant 2$. Moreover,

$$b \log x < (b/a) \log a < (b/a)(a-1) = 1,$$

hence also $-b \log y < 1 + \mathrm{O}_\Lambda(b\delta)$. In particular, $y^b \gg_\Lambda 1$ whenever $b\delta < 1$. Now let us consider the function

$$F(t) := \frac{x^a}{a} + \frac{t^b}{b} - xt.$$

Note that $F(y) = 1 - xy$, and $F(y_0) = F'(y_0) = 0$ for $y_0 := x^{a-1}$. Hence, using Lagrange's form for the remainder term in Taylor's theorem, we see that

$$\delta \gg_\Lambda F(y) \geqslant \frac{(b-1)}{2} \min(y_0^{b-2}, y^{b-2})\,(y-y_0)^2.$$

Here $y_0^{b-2} = x^{2-a} \gg 1$. Now let us assume that $y^b > 1$ or $b\delta < 1$. Then $y^b \gg_\Lambda 1$, whence $y - y_0 \ll_\Lambda \sqrt{\delta/b}$ by the previous display. From here and (5.8) we get the following two approximations for bxy:

$$bxy = bxy_0 + \mathrm{O}_\Lambda(\sqrt{b\delta}) = bx^a + \mathrm{O}_\Lambda(\sqrt{b\delta}),$$
$$bxy = b + \mathrm{O}_\Lambda(b\delta) = (b-1)x^a + y^b + \mathrm{O}_\Lambda(b\delta).$$

Comparing the right-hand sides, we conclude that

$$x^a - y^b \ll_\Lambda b\delta + \sqrt{b\delta}. \tag{5.9}$$

In the remaining case when $y^b \leqslant 1$ and $b\delta \geqslant 1$, the inequality (5.9) holds automatically in the stronger form $|x^a - y^b| < 2 \leqslant 2b\delta$.

We proved that (5.8) implies (5.9) in all ranges. For our specific set-up (5.7), the inequality (5.9) says that

$$aA^2 - bB^2 \ll_\Lambda X^2(b\delta + \sqrt{b\delta}),$$

and this is equivalent to (5.6) in the light of $A^2 + B^2 = X^2$. This shows (a) under the assumption $0 \leqslant q < \ell$, but it is easily seen to continue to hold also for $q = \ell$ in which case (5.8) simply reads $A^2 = X^2 + \mathrm{O}_\Lambda(X^2\delta)$. The argument for $-\ell \leqslant q < 0$ is identical.

Turning to (b), we conclude from (a) that

$$\left(\frac{2\ell}{\ell+q}\right)^{\ell+q}\left(\frac{2\ell}{\ell-q}\right)^{\ell-q} A^{\ell+q}B^{\ell-q}C^{\ell+q}D^{\ell-q} \leqslant X^{4\ell}.$$

On the other hand, using Stirling's formula $n! \sim (n/e)^n\sqrt{2\pi n}$, we have for $|q| < \ell$ that

$$\binom{2\ell}{\ell+q} \asymp \frac{(2\ell)^{2\ell}}{(\ell+q)^{\ell+q}(\ell-q)^{\ell-q}}\sqrt{\frac{2\ell}{(\ell+q)(\ell-q)}},$$

and so combining the two most recent displays we have the announced bound

$$\binom{2\ell}{\ell+q} A^{\ell+q}B^{\ell-q}C^{\ell+q}D^{\ell-q} \ll \frac{X^{4\ell}}{1+\sqrt{\ell-|q|}}.$$

We added artificially the 1+ term in the denominator, so that the inequality also holds for the previously excluded case $|q| = \ell$ in view of $AC, BD \leqslant X^2$ (which follows directly from $A^2 + C^2 = B^2 + D^2 = X^2$). The claim that the left-hand side is negligible unless (5.6) holds for (A, B) and (C, D) is immediate from (a). □

We now return to the double integral in the upper bound for $\varphi^{\ell,q}_{\nu,\ell}(g)$. We estimate the inner integral by writing the integrand as $A^{\ell+q}B^{\ell-q}C^{\ell+q}D^{\ell-q}$ in the obvious way and applying Lemma 7, where

$$A^2 + B^2 = C^2 + D^2 = X^2 = 1 + t^2,$$

and

$$A^2 = \frac{1+t^2}{2} + \frac{t^2-1}{2}\cos 2v_1 + t\sin 2v_1\cos(\phi+\Delta),$$

with analogous expressions for B^2, C^2, and D^2. Since

$$\frac{(1+(t/r)^2)(1+(tr)^2)}{(1+t^2)^2} = 1 + \left(\frac{r-r^{-1}}{t+t^{-1}}\right)^2,$$

we conclude that the contribution of the inner integral is $\mathrm{O}_\Lambda(\ell^{-\Lambda})$ unless

$$\min(t, t^{-1}) \ll_\Lambda \frac{\lambda}{(r-1)\sqrt{\ell}}. \tag{5.10}$$

For $r = 1$ we treat the right-hand side as infinity. We may then summarize our findings as follows.

Lemma 8. *Let $\Lambda \in \mathbb{N}$. Let $\ell, q \in \mathbb{Z}$ be such that $\ell \geqslant \max(1, |q|)$, and let $\nu \in i\mathbb{R}$. Assume that $g \in \mathrm{SL}_2(\mathbb{C})$ is given by (5.3). Let us abbreviate $\Delta := u_2 + w_1$ and $\lambda := \sqrt{\log \ell}$. Let $\mathcal{M} = \mathcal{M}(v_1, v_2, \Delta, r, \Lambda)$ be the set of $(\phi, t) \in [0, 2\pi] \times [0, \infty)$ satisfying (5.10) as well as*

$$\begin{aligned} 2t \sin 2v_1 \cos(\phi + \Delta) &= (1 - t^2) \cos 2v_1 + \frac{q}{\ell}(1 + t^2) + \mathrm{O}_\Lambda\left((1 + t^2)\frac{\lambda^2 + \lambda\sqrt{\ell - |q|}}{\ell}\right), \\ 2t \sin 2v_2 \cos(\phi - \Delta) &= (t^2 - 1) \cos 2v_2 - \frac{q}{\ell}(1 + t^2) + \mathrm{O}_\Lambda\left((1 + t^2)\frac{\lambda^2 + \lambda\sqrt{\ell - |q|}}{\ell}\right), \end{aligned} \tag{5.11}$$

with a sufficiently large (but fixed) implied constant depending on Λ. Then

$$\varphi^{\ell,q}_{\nu,\ell}(g) \ll_\Lambda \frac{\ell}{1 + \sqrt{\ell - |q|}} \int\limits_{\mathcal{M}} \frac{t}{(1 + t^2)^2} \, \mathrm{d}\phi \, \mathrm{d}t + \ell^{-\Lambda}. \tag{5.12}$$

5.2. Simplifying assumptions

For the proof of Theorems 5 and 6, we can and we shall assume that $|\Delta| \leqslant \pi/4$. Indeed, using the last relation in (2.2) multiple times, we can choose the coordinates in (5.3) so that this bound is satisfied. Moreover, we can replace g by

$$g^{-1} = k\left[\frac{\pi}{2} - w_2, v_2 - \frac{\pi}{2}, u_2 + \frac{\pi}{2}\right] \begin{pmatrix} r & \\ & r^{-1} \end{pmatrix} k\left[w_1 - \frac{\pi}{2}, v_1 - \frac{\pi}{2}, \frac{\pi}{2} - u_1\right]$$

if needed, because the quantities Δ, $\|g\|$, $D(g)$ do not change under this replacement, $\left|\varphi^{\ell,q}_{\nu,\ell}(g)\right| = \left|\varphi^{\ell,q}_{\nu,\ell}(g^{-1})\right|$ holds by (3.23), and

$$\operatorname{dist}(g, \mathcal{H}) = \operatorname{dist}(g^{-1}, \mathcal{H}), \qquad \mathcal{H} \in \{K, \mathcal{D}, \mathcal{S}\}$$

holds by (1.10).

We shall derive (most of) the bounds in Theorems 5 and 6 from (5.12). In Lemma 8, the pair (Δ, r) does not change under the above discussed replacement $g \mapsto g^{-1}$, while the corresponding integration domains $\mathcal{M}$ are related by

$$(\phi, t) \in \mathcal{M}\left(v_2 - \frac{\pi}{2}, v_1 - \frac{\pi}{2}, \Delta, r, \Lambda\right) \quad \Longleftrightarrow \quad (\phi, t^{-1}) \in \mathcal{M}(v_1, v_2, \Delta, r, \Lambda).$$

Moreover, the integrand in (5.12) is invariant under $t \mapsto t^{-1}$, hence we can assume that the contribution of $t \leqslant 1$ is not smaller than the contribution of $t > 1$. So from now on we restrict $\mathcal{M}$ in (5.12) to the corresponding subset of $[0, 2\pi] \times [0, 1]$. On this subset we have, by (5.10),

$$t \in [0, 1] \qquad \text{and} \qquad t \ll_\Lambda \frac{\lambda}{(r-1)\sqrt{\ell}}. \tag{5.13}$$

5.3. Proof of Theorem 5

The bound (1.12) is trivial for $\ell \ll_\Lambda 1$, hence we shall assume that ℓ is sufficiently large in terms of Λ. With the notation

$$\alpha := \operatorname{dist}(g, K) \asymp r - 1 \qquad \text{and} \qquad \beta := \operatorname{dist}(g, \mathcal{D}),$$

it follows from (5.12) and the previous subsection that it suffices to show

$$\frac{\ell}{1 + \sqrt{\ell - |q|}} \int_{\mathcal{M}} t \, \mathrm{d}\phi \, \mathrm{d}t \ll_{\varepsilon,\Lambda} \ell^\varepsilon \min\left(1, \frac{\|g\|}{\sqrt{\ell}\alpha^2\beta}\right), \tag{5.14}$$

where $\mathcal{M}$ is now restricted by (5.13). In fact our arguments below will show that ℓ^ε can be replaced by $(\log \ell)^3$.

We start with the first bound of (5.14). With the notation

$$\sigma := \lambda^2 + \lambda\sqrt{\ell - |q|}, \qquad \mu := \frac{q}{\ell} - \cos 2v_1, \qquad \rho := \sin 2v_1,$$

the first equation in (5.11) becomes

$$\mu t^2 - 2t\rho\cos(\phi + \Delta) + \frac{2q}{\ell} - \mu + \mathrm{O}_\Lambda\left(\frac{\sigma}{\ell}\right) = 0. \tag{5.15}$$

Without loss of generality, $\mu \neq 0$, and then we can view (5.15) as a quadratic equation for t. Multiplying by μ and completing the square, we obtain the alternative form

$$\big(\mu t - \rho\cos(\phi + \Delta)\big)^2 + \big(\rho\sin(\phi + \Delta)\big)^2 = 1 - \frac{q^2}{\ell^2} + \mathrm{O}_\Lambda\left(\frac{|\mu|\sigma}{\ell}\right). \tag{5.16}$$

In particular, the discriminant of (5.15) equals $4D(\phi) + \mathrm{O}_\Lambda(|\mu|\sigma/\ell)$, where

$$D(\phi) := 1 - \frac{q^2}{\ell^2} - \big(\rho\sin(\phi + \Delta)\big)^2 \tag{5.17}$$

We assume first that $|q|/\ell \leqslant 5/6$, and decompose $\mathcal{M}$ into two parts $\mathcal{M}^\pm$ according as $|\rho\sin(\phi+\Delta)|$ exceeds $1/2$ or not. On $\mathcal{M}^+$, the equation (5.15) localizes ϕ within $\ll_\Lambda \sigma/(\ell t)$ for each given $t \in [0,1]$. On $\mathcal{M}^-$, we have $D(\phi) \geqslant 1/18$, hence the equation (5.16) localizes t within $\ll_\Lambda \sigma/\ell$ for each given $\phi \in [0, 2\pi]$. This shows that

$$\int_{\mathcal{M}} t \, \mathrm{d}\phi \, \mathrm{d}t \ll_\Lambda \int_0^1 t \, \frac{\sigma}{\ell t} \, \mathrm{d}t + \int_0^{2\pi} \frac{\sigma}{\ell} \, \mathrm{d}\phi \ll \frac{\sigma}{\ell},$$

hence the first bound of (5.14) follows in stronger form. From now on we assume that $|q|/\ell > 5/6$. We decompose $\mathcal{M}$ into two parts $\mathcal{M}^\pm$ according as $D(\phi)$ is positive or not, and we make two initial observations. First, $\mathcal{M}^+$ is clearly empty when $|q| = \ell$. Second, $|\mu| > 1/6$ holds for large ℓ, because (5.15) coupled with $t \in [0,1]$ yields

$$\frac{2|q|}{\ell} - |\mu| - \mathrm{O}_\Lambda\left(\frac{\sigma}{\ell}\right) \leqslant 2t|\rho| \leqslant 2\sqrt{1 - \left(\frac{q}{\ell} - \mu\right)^2}.$$

In order to estimate the contribution of $\mathcal{M}^+$ in (5.14), we decompose $\mathcal{M}^+$ into pieces

$$\mathcal{M}^+(\mathtt{D},\eta) := \big\{(\phi,t)\in\mathcal{M}^+ : D(\phi)\asymp \mathtt{D} \text{ and } |\cos(\phi+\Delta)|\asymp\eta\big\}.$$

If $\eta\leqslant \ell^{-10}$, we can estimate trivially, so there are only $\mathrm{O}(\log\ell)$ relevant values for η. If $\rho\geqslant\ell^{-10}$, then by the same argument there are only $\mathrm{O}(\log\ell)$ relevant values for $\mathtt{D}$. If $\rho<\ell^{-10}$, then $\mathtt{D}>\ell^{-1}$ by $|q|<\ell$, hence again there are only $\mathrm{O}(\log\ell)$ relevant values for $\mathtt{D}$. So in all cases it suffices to restrict to $\mathrm{O}((\log\ell)^2)$ pairs $(\mathtt{D},\eta)$. Our current assumptions localize $\sin(\phi+\Delta)$ within $\ll\sqrt{\mathtt{D}}/|\rho|$, and hence ϕ within $\ll\min(1,\sqrt{\mathtt{D}}/|\rho\eta|)$, independently of t. On the other hand, given ϕ, the equation (5.16) localizes t within $\ll_\Lambda \min((\sigma/\ell)\mathtt{D}^{-1/2},\sqrt{\sigma/\ell})$. Such t are of size $\ll_\Lambda |\rho\eta|+\sqrt{\mathtt{D}}+\sqrt{\sigma/\ell}$, so that

$$\int\limits_{\mathcal{M}^+(\mathtt{D},\eta)} t\,\mathrm{d}\phi\,\mathrm{d}t \ll_\Lambda \left(|\rho\eta|+\sqrt{\mathtt{D}}+\sqrt{\frac{\sigma}{\ell}}\right)\min\left(\frac{\sigma}{\ell\sqrt{\mathtt{D}}},\sqrt{\frac{\sigma}{\ell}}\right)\min\left(1,\frac{\sqrt{\mathtt{D}}}{|\rho\eta|}\right)\ll\frac{\sigma}{\ell}.$$

This contribution is admissible for the first bound of (5.14). It remains to estimate the contribution of $\mathcal{M}^-$ in (5.14). On this set we have

$$0\leqslant -D(\phi)\ll_\Lambda \frac{\sigma}{\ell}$$

by (5.16). The argument is similar as for $\mathcal{M}^+$, in fact simpler as we only need $\mathrm{O}(\log\ell)$ pieces $\mathcal{M}^-(\eta)$ defined by $|\cos(\phi+\Delta)|\asymp\eta$. Initially we localize ϕ within $\ll\min(1,|\rho\eta|^{-1}\sqrt{\sigma/\ell})$, independently of t. The equation (5.16) localizes t within $\ll_\Lambda\sqrt{\sigma/\ell}$, and such t are of size $\ll_\Lambda|\rho\eta|+\sqrt{\sigma/\ell}$. We obtain altogether

$$\int\limits_{\mathcal{M}^-(\eta)} t\,\mathrm{d}\phi\,\mathrm{d}t \ll_\Lambda \left(|\rho\eta|+\sqrt{\frac{\sigma}{\ell}}\right)\sqrt{\frac{\sigma}{\ell}}\min\left(1,\frac{1}{|\rho\eta|}\sqrt{\frac{\sigma}{\ell}}\right)\ll\frac{\sigma}{\ell},$$

which is again admissible for the first bound of (5.14).

We now turn to the second bound of (5.14). We shall assume (as we can) that $\mathcal{M}\neq\emptyset$ and $\sqrt{\ell}\alpha^2\beta>\|g\|$. We pick an arbitrary point $(\phi,t)\in\mathcal{M}$. Combining (5.11) and (5.13), we get

$$\cos 2v_j = -\mathrm{sgn}(q)+\mathrm{O}(t)\sin 2v_j+\mathrm{O}_\Lambda\left(t^2+\frac{\lambda^2+\ell-|q|}{\ell}\right),$$

where for $q=0$ we can replace $\mathrm{sgn}(q)$ by 1. After squaring and solving for $\sin 2v_j$, then feeding back the result into the previous display, we get

$$\sin 2v_j=\mathrm{O}_\Lambda\left(t+\frac{\lambda+\sqrt{\ell-|q|}}{\sqrt{\ell}}\right),\qquad \cos 2v_j=-\mathrm{sgn}(q)+\mathrm{O}_\Lambda\left(t^2+\frac{\lambda^2+\ell-|q|}{\ell}\right).$$

Recalling also (5.3), and using (5.13) again, we infer that

$$\beta\ll_\Lambda\|g\|\left(\frac{\lambda}{\alpha\sqrt{\ell}}+\frac{\lambda+\sqrt{\ell-|q|}}{\sqrt{\ell}}\right).$$

Hence we always have

$$1\ll_\Lambda\frac{\|g\|\lambda}{\alpha\beta\sqrt{\ell}}\qquad\text{or}\qquad 1\ll_\Lambda\|g\|\lambda\frac{1+\sqrt{\ell-|q|}}{\beta\sqrt{\ell}}.$$

In either case, for any $c>0$, the previous display combined with (5.13) yields that

$$\frac{\ell}{1+\sqrt{\ell-|q|}}\int_{\mathcal{M}} t\,\mathrm{d}\phi\,\mathrm{d}t \ll_{\Lambda,c} \frac{\lambda^2}{(1+\sqrt{\ell-|q|})\alpha^2}\left(\left(\frac{\|g\|\lambda}{\alpha\beta\sqrt{\ell}}\right)^c + \|g\|\lambda\frac{1+\sqrt{\ell-|q|}}{\beta\sqrt{\ell}}\right)$$

$$\ll_{\varepsilon,\Lambda,c} \ell^{\varepsilon}\left(\frac{\|g\|^c}{\ell^{c/2}\alpha^{2+c}\beta^c} + \frac{\|g\|}{\sqrt{\ell}\alpha^2\beta}\right).$$

Choosing $c=2$, and recalling our initial assumption $\sqrt{\ell}\alpha^2\beta > \|g\|$, we obtain the second bound of (5.14) in stronger form.

The proof of Theorem 5 is complete.

5.4. *Proof of Theorem 6(a)*

The averaged spherical trace function $\varphi^{\ell,q}_{\nu,\ell}(g)$ exhibits starkly different behavior depending on the value of $-\ell \leqslant q \leqslant \ell$. Some of these features are already visible along $K = \mathrm{SU}_2(\mathbb{C})$. From (5.1) and (5.4) we can see that, in the notation of (2.1) and (2.4),

$$\varphi^{\ell,q}_{\nu,\ell}(k[u,v,w]) = \Phi^{\ell}_{q,q}(k[u,v,w]) = e^{2\pi i q(u+w)}(\cos v)^{2q}P^{(0,2q)}_{\ell-q}(\cos 2v).$$

The absolute value of the right-hand side exhibits a primary peak at $v \in \pi\mathbb{Z}$ of size 1. For $q = \pm\ell$, this is followed by a sharp drop to $\mathrm{O}_N(\ell^{-N})$ after a range of length about $\ell^{-1/2}$. For a generic q, the drop becomes soft through a highly oscillatory range of magnitude $\ell^{-1/2}$ (faster and more oscillatory for smaller q) and a secondary, Airy-type peak of size about $\ell^{-1/3}$ before the delayed sharp drop. For $q=0$, the secondary peak grows to a full peak of size 1 at $v \in \frac{1}{2}\pi + \pi\mathbb{Z}$ (corresponding to skew-diagonal matrices in K) and the sharp drop disappears. These varying features, which are illustrated in Fig. 1, become vastly more complicated off K, where the hard work in Theorems 5 and 6 lies. Nevertheless, their traces are visible in the hard localization to $\mathcal{D}$ (but none to K!) for $q = \pm\ell$ and the hard localization to $\mathcal{N}$ with soft localization to $\mathcal{S} \subset K \subset \mathcal{N}$ for $q=0$.

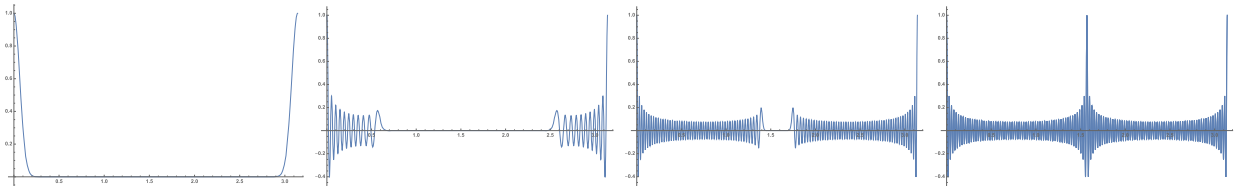

Fig. 1. Plots of $(\cos v)^{2q}P^{(0,2q)}_{\ell-q}(\cos 2v)$ for $0 \leqslant v \leqslant \pi$, $\ell = 120$, $q = 120$, $q = 100$, $q = 20$, and $q = 0$.

In this subsection, we consider in more detail the case $q=0$. Then (5.12) simplifies to

$$\varphi^{\ell,0}_{\nu,\ell}(g) \ll_{\Lambda} \sqrt{\ell}\int_{\mathcal{M}} \frac{t}{(1+t^2)^2}\,\mathrm{d}\phi\,\mathrm{d}t + \ell^{-\Lambda}, \tag{5.18}$$

where by (5.11) and the last paragraph of §5.1, the set $\mathcal{M}$ can be described by the constraints given in (5.13) and

$$\begin{aligned} 2t\sin 2v_1\cos(\phi+\Delta) &= (1-t^2)\cos 2v_1 + \mathrm{O}_{\Lambda}(\lambda/\sqrt{\ell}),\\ 2t\sin 2v_2\cos(\phi-\Delta) &= (t^2-1)\cos 2v_2 + \mathrm{O}_{\Lambda}(\lambda/\sqrt{\ell}). \end{aligned} \tag{5.19}$$

We shall use the notations

$$P(\phi) := \max(|\sin 2v_1 \cos(\phi+\Delta)|, |\sin 2v_2 \cos(\phi-\Delta)|),$$
$$R := \max(|\cos 2v_1|, |\cos 2v_2|),$$
$$N := \max(|\sin(2v_1+2v_2)\cos\Delta|, |\sin(2v_1-2v_2)\sin\Delta|).$$

Recall also the earlier notations (1.9) and (1.11). As

$$|a|^2-|d|^2 = \frac{r^2-r^{-2}}{2}(\cos 2v_1+\cos 2v_2), \qquad |b|^2-|c|^2 = \frac{r^2-r^{-2}}{2}(\cos 2v_1-\cos 2v_2),$$

we can identify $\mathcal{N}$ as the set of matrices with $r=1$ or $\cos 2v_1 = \cos 2v_2 = 0$. More precisely, by (5.19) and (5.13) we have

$$D(g) \ll r(r-1)R \ll_\Lambda r(r-1)\left(t+\frac{\lambda}{\sqrt{\ell}}\right) \ll_\Lambda \frac{r^2\lambda}{\sqrt{\ell}},$$

so that unless $D(g) \ll_\Lambda \|g\|^2\lambda\ell^{-1/2}$, we have $\mathcal{M}=\emptyset$, yielding $\varphi^{\ell,0}_{\nu,\ell}(g) \ll_\Lambda \ell^{-\Lambda}$. Hence we are left with proving (1.13).

In (5.18), the contribution of the t-integral over the interval $[0,\ell^{-\Lambda/2-1/4}]$ is negligible, and we split the rest of $\mathcal{M}$ in dyadic ranges $\mathcal{M}(\delta)$ according to $\ell^{-\Lambda/2-1/4} < t \asymp \delta \leqslant 1$. The number of such ranges is $\mathrm{O}_\Lambda(\log\ell)$. Assume $(\phi,t)\in\mathcal{M}(\delta)$. The discriminants of the two quadratic equations (5.19) are $4D_j(\phi)+\mathrm{O}_\Lambda(\lambda/\sqrt{\ell})$, where

$$D_1(\phi) := 1-\sin^2(2v_1)\sin^2(\phi+\Delta), \qquad D_2(\phi) := 1-\sin^2(2v_2)\sin^2(\phi-\Delta).$$

A simple calculation gives that

$$D_1(\phi)+D_2(\phi) \geqslant P(\phi)^2+R^2. \tag{5.20}$$

If $|\sin 2v_1 \sin(\phi+\Delta)| > 1/2$, then for any fixed t, (5.19) localizes ϕ to a set of measure $\mathrm{O}_\Lambda(\lambda/\sqrt{\ell})$. Otherwise, for any fixed ϕ, (5.19) localizes t to a set of measure $\mathrm{O}_\Lambda(\lambda/\sqrt{\ell})$. We conclude that

$$\mathrm{meas}(\mathcal{M}(\delta)) \ll_\Lambda \lambda/\sqrt{\ell}. \tag{5.21}$$

Now we prove the alternative bound

$$\mathrm{meas}(\mathcal{M}(\delta)) \ll_\Lambda \frac{\lambda^4}{N\ell}. \tag{5.22}$$

We shall assume that $N\ell > 1$, for otherwise (5.22) follows from (5.21). Under this assumption, we have $\max(|\sin 2v_1|, |\sin 2v_2|) \gg \ell^{-1}$, which implies that

$$\mathrm{meas}(\{(\phi,t)\in\mathcal{M}(\delta) : P(\phi)\leqslant\ell^{-3}\}) \ll \ell^{-1}.$$

Indeed, if ϕ changes by at least ℓ^{-1} and at most $\pi/4$, then $\cos(\phi\pm\Delta)$ both change by $\Omega(\ell^{-2})$, hence $P(\phi)$ changes by $\Omega(\ell^{-3})$. This implies that $P(\phi)\leqslant\ell^{-3}$ localizes ϕ to a set of measure $\mathrm{O}(\ell^{-1})$. Therefore, the contribution of $\{(\phi,t)\in\mathcal{M}(\delta) : P(\phi)\leqslant\ell^{-3}\}$ to the left-hand side of (5.22) is $\mathrm{O}_\Lambda(\delta/\sqrt{\ell})$, which is admissible by $N\leqslant 1$. We decompose the rest of $\mathcal{M}(\delta)$ into dyadic ranges $\mathcal{M}(\delta,\mathtt{P})$ according to $\ell^{-3}\leqslant P(\phi)\asymp\mathtt{P}\leqslant 1$. The number of such ranges is $\mathrm{O}(\log\ell)$, hence in order to verify (5.22), it suffices to prove

$$\mathrm{meas}(\mathcal{M}(\delta,\mathtt{P})) \ll_\Lambda \frac{\lambda^2}{N\ell}.$$

The proof of this estimate immediately reduces to the following two localizations:

$$\mathrm{meas}(\{\phi \in [0,2\pi] : (\phi,t) \in \mathcal{M}(\delta,\mathtt{P}) \text{ for some } t \asymp \delta\}) \ll_\Lambda \frac{(\mathtt{P}+R)\lambda}{N\sqrt{\ell}}, \tag{5.23}$$

and for any $\phi \in [0,2\pi]$,

$$\mathrm{meas}(\{t \in [0,1] : (\phi,t) \in \mathcal{M}(\delta,\mathtt{P})\}) \ll_\Lambda \frac{\lambda}{(\mathtt{P}+R)\sqrt{\ell}}. \tag{5.24}$$

Now we prove these localizations.

Starting out from (5.19), we execute two eliminations: one to eliminate the main terms of the right-hand sides, and the other one to eliminate the left-hand sides. Introducing

$$F(\phi) := \cos\phi\cos\Delta\sin(2v_1+2v_2) - \sin\phi\sin\Delta\sin(2v_1-2v_2),$$

these give

$$tF(\phi) \ll_\Lambda R\lambda/\sqrt{\ell} \qquad \text{and} \qquad (1-t^2)F(\phi) \ll_\Lambda \mathtt{P}\lambda/\sqrt{\ell}.$$

In particular, we obtain both for $t > 1/2$ and $t \leqslant 1/2$ that

$$F(\phi) \ll_\Lambda (\mathtt{P}+R)\lambda/\sqrt{\ell}. \tag{5.25}$$

Letting

$$N' := \sqrt{\sin^2(2v_1+2v_2)\cos^2(\Delta) + \sin^2(2v_1-2v_2)\sin^2(\Delta)} \asymp N,$$

and choosing $\psi \in [0,2\pi)$ such that

$$\cos\psi = \frac{\sin(2v_1+2v_2)\cos\Delta}{N'}, \qquad \sin\psi = -\frac{\sin(2v_1-2v_2)\sin\Delta}{N'},$$

(5.25) gives rise to

$$\cos(\phi-\psi) \ll_\Lambda \frac{(\mathtt{P}+R)\lambda}{N\sqrt{\ell}}.$$

This localizes ϕ to a set of measure $\mathrm{O}_\Lambda((\mathtt{P}+R)\lambda/N\sqrt{\ell})$. Indeed, if the right-hand side is very small in terms of the implied constant, then $\phi-\psi$ is bounded away from $\pi\mathbb{Z}$, hence the derivative $\cos'(\phi-\psi)$ is bounded away from zero, while otherwise the claimed localization is trivial. This gives (5.23). Fixing $\phi \in [0,2\pi]$, and solving under (5.19) the quadratic equation in t of the larger discriminant, we see by (5.20) that t is localized to a set of measure $\mathrm{O}_\Lambda(\lambda/(\mathtt{P}+R)\sqrt{\ell})$. This gives (5.24). Altogether, the proof of (5.22) is complete.

Combining (5.21) and (5.22), we obtain

$$\mathrm{meas}(\mathcal{M}(\delta)) \ll_{\varepsilon,\Lambda} \ell^{\varepsilon-1}\mu, \qquad \mu := \min(\sqrt{\ell}, N^{-1}).$$

We claim that

$$\mathrm{dist}(g,\mathcal{S}) \ll_\Lambda \lambda\delta^{-1}\mu^{-1}, \qquad \text{if } \mathcal{M}(\delta) \neq \emptyset. \tag{5.26}$$

This implies the inequality

$$\sqrt{\ell} \int\limits_{\mathcal{M}(\delta)} t \, \mathrm{d}\phi \, \mathrm{d}t \ll_{\varepsilon,\Lambda} \ell^{\varepsilon-1/2}\delta\mu \ll_{\varepsilon,\Lambda} \frac{\ell^{\varepsilon}}{\sqrt{\ell}\,\mathrm{dist}(g,\mathcal{S})},$$

which, summed over the $\mathrm{O}(\log \ell)$ dyadic ranges for δ, suffices for the proof of (1.13). Note that the bound $\varphi_{\nu,\ell}^{\ell,q}(g) \ll_{\varepsilon} \ell^{\varepsilon}$ is already covered by Theorem 5.

To complete the proof of Theorem 6(a), it remains to show (5.26). For this final argument, we can and we shall assume that $-\pi/8 \leqslant v_1, v_2 \leqslant 3\pi/8$, because replacing (u_1, v_1) by $(-u_1, v_1 + \pi/2)$, or (v_2, w_2) by $(v_2+\pi/2, -w_2)$, has the effect of multiplying g by $\left(\begin{smallmatrix} & i \\ i & \end{smallmatrix}\right)$ from either side without altering Δ or the statement (5.26). We fix a pair $(\phi, t) \in \mathcal{M}(\delta)$.

Now, $N \leqslant \mu^{-1}$ implies that

$$v_1 + v_2 \in \frac{\pi}{2}\mathbb{Z} + \mathrm{O}\left(\frac{1}{\mu}\right) \qquad \text{and} \qquad v_1 - v_2 \in \frac{\pi}{2}\mathbb{Z} + \mathrm{O}\left(\frac{1}{\mu|\Delta|}\right). \tag{5.27}$$

Let us introduce the short-hand notation

$$m[v] := \begin{pmatrix} \cos v & i\sin v \\ i\sin v & \cos v \end{pmatrix}, \qquad v \in \mathbb{R}.$$

Keeping (2.1) and (5.3) in mind, we observe initially that

$$m[v_1]\,\mathrm{diag}\left(re^{i\Delta}, r^{-1}e^{-i\Delta}\right) m[v_2] = m[v_1+v_2] + \mathrm{O}\big(r - 1 + |\Delta|\big). \tag{5.28}$$

On the right-hand side, we have $\mathrm{dist}(m[v_1+v_2], \mathcal{S}) \ll \mu^{-1}$ by (5.27), and also

$$r - 1 \ll_{\Lambda} \frac{\lambda}{t\sqrt{\ell}} \ll \frac{\lambda}{\delta\mu} \tag{5.29}$$

by (5.13) and $\mu \leqslant \sqrt{\ell}$. Hence (5.26) follows from (5.28) as long as $\Delta \ll_{\Lambda} \lambda\delta^{-1}\mu^{-1}$. In other words, we can and we shall assume that $|\Delta| \gg_{\Lambda} \lambda\delta^{-1}\mu^{-1}$ holds with a sufficiently large implied constant depending on Λ. In particular, we shall assume that the error terms in (5.27), and similar error terms for angles in the rest of this subsection, are less than $\pi/8$ in size. Under this assumption, (5.27) breaks into two cases.

Case 1: $v_1, v_2 \ll \mu^{-1}|\Delta|^{-1}$ and $v_1 + v_2 \ll \mu^{-1}$. In this case, we refine (5.28) to

$$\begin{aligned} &m[v_1]\,\mathrm{diag}\left(re^{i\Delta}, r^{-1}e^{-\Delta}\right) m[v_2] \\ &= m[v_1+v_2] + m[v_1]\,\mathrm{diag}\left(re^{i\Delta} - 1, r^{-1}e^{-i\Delta} - 1\right) m[v_2] \\ &= m[v_1+v_2] + \mathrm{diag}\left(re^{i\Delta} - 1, r^{-1}e^{-i\Delta} - 1\right) + \mathrm{O}\big(r - 1 + \mu^{-1}\big) \\ &= \mathrm{diag}\left(e^{i\Delta}, e^{-i\Delta}\right) + \mathrm{O}\big(r - 1 + \mu^{-1}\big). \end{aligned}$$

The main term $\mathrm{diag}\left(e^{i\Delta}, e^{-i\Delta}\right)$ lies in $\mathcal{S}$, hence (5.26) follows by (5.29).

Case 2: $v_1, v_2 = \pi/4 + \mathrm{O}(\mu^{-1}|\Delta|^{-1})$ and $v_1 + v_2 = \pi/2 + \mathrm{O}(\mu^{-1})$. As we shall see, this case does not occur. The assumptions imply that $\sin 2v_1$ and $\sin 2v_2$ exceed $1/2$. We multiply the second equation in (5.19) by $\sin 2v_1$, and the first equation in (5.19) by $\sin 2v_2$. Adding and subtracting the resulting two equations, we obtain

$$\begin{aligned} 4t \sin 2v_1 \sin 2v_2 \cos\phi \cos\Delta &= (t^2-1)\sin(2v_1 - 2v_2) + \mathrm{O}_{\Lambda}(\lambda/\sqrt{\ell}), \\ 4t \sin 2v_1 \sin 2v_2 \sin\phi \sin\Delta &= (t^2-1)\sin(2v_1 + 2v_2) + \mathrm{O}_{\Lambda}(\lambda/\sqrt{\ell}). \end{aligned}$$

We infer that

$$\delta \ll |t\cos\phi| + |t\sin\phi| \ll_\Lambda |\sin(2v_1-2v_2)| + \frac{|\sin(2v_1+2v_2)|}{|\Delta|} + \frac{\lambda}{\sqrt{\ell}|\Delta|} \ll_\Lambda \frac{\lambda}{\mu|\Delta|}.$$

This contradicts our earlier assumption that $\Delta \gg_\Lambda \lambda\delta^{-1}\mu^{-1}$ holds with a sufficiently large implied constant depending on Λ.

The proof of Theorem 6(a) is complete.

5.5. *Proof of Theorem 6(b)*

We finally consider the case $q = \pm\ell$. By the symmetries (1.10) and (3.23), we can restrict to $q = \ell$. We have already shown the bound $\varphi^{\ell,\ell}_{\nu,\ell}(g) \ll_\varepsilon \ell^\varepsilon$ in greater generality in Theorem 5. As a first step, we complement this with a stronger bound for $r \geqslant 2$. To this end, we return to (5.4). As $q = \ell$, the binomial coefficient and the J-factor disappear. When $\overline{I}^{2\ell}$ is expanded, we see a Laurent polynomial of e^{2iu}. When we integrate in u from 0 to π, all the terms but the constant one vanish. We calculate the constant term using the binomial theorem and the original product definition of I. This way we see that

$$\varphi^{\ell,\ell}_{\nu,\ell}(g) = d_\ell e^{2i\ell(u_1-u_2-w_1+w_2)} r^{2\ell} \sum_{m=0}^{2\ell} \binom{2\ell}{m}^2 (r^{-2}e^{2iu_2+2iw_1}\cos v_1\cos v_2)^m$$
$$(-\sin v_1\sin v_2)^{2\ell-m} \int_0^{\pi/2} (\sin^2 v)^m (\cos^2 v)^{2\ell-m} \frac{\sin 2v}{h(r,v)^{\ell+1-\nu}}\,\mathrm{d}v.$$

Using the variable $x := \sin^2 v$, we rewrite this as

$$\varphi^{\ell,\ell}_{\nu,\ell}(g) = d_\ell e^{2i\ell(u_1-u_2-w_1+w_2)} r^{2\nu-2} \sum_{m=0}^{2\ell} \binom{2\ell}{m}^2 (r^{-2}e^{2iu_2+2iw_1}\cos v_1\cos v_2)^m$$
$$(-\sin v_1\sin v_2)^{2\ell-m} \int_0^1 \frac{x^m(1-x)^{2\ell-m}}{(1-x+r^{-4}x)^{\ell+1-\nu}}\,\mathrm{d}x.$$

With the short-hand notation

$$U := r^{-1}e^{iu_2+iw_1}\sqrt{x\cos v_1\cos v_2} \qquad \text{and} \qquad V := i\sqrt{(1-x)\sin v_1\sin v_2},$$

we obtain finally

$$\begin{aligned} |\varphi^{\ell,\ell}_{\nu,\ell}(g)| &\leqslant \frac{2\ell+1}{r^2}\left|\int_0^1 \sum_{m=0}^{2\ell}\binom{2\ell}{m}^2 \frac{U^{2m}V^{4\ell-2m}}{(1-x+r^{-4}x)^{\ell+1-\nu}}\,\mathrm{d}x\right| \\ &= \frac{2\ell+1}{r^2}\left|\int_0^1 \frac{1}{2\pi}\int_0^{2\pi} \frac{(Ue^{i\phi}+V)^{2\ell}(Ue^{-i\phi}+V)^{2\ell}}{(1-x+r^{-4}x)^{\ell+1-\nu}}\,\mathrm{d}\phi\,\mathrm{d}x\right|. \end{aligned} \tag{5.30}$$

Using that $(Ue^{i\phi}+V)(Ue^{-i\phi}+V) = U^2+V^2+2UV\cos\phi$ is on the line segment connecting $(U+V)^2$ and $(U-V)^2$, we observe that

$$\frac{|(Ue^{i\phi}+V)(Ue^{-i\phi}+V)|^2}{1-x+r^{-4}x} \leqslant \max_{\pm} \frac{|U\pm V|^4}{1-x+r^{-4}x},$$

which by the Cauchy–Schwarz inequality can be further upper bounded by

$$\leqslant \frac{(1-x+r^{-2}x)^2}{1-x+r^{-4}x} = 1 - \frac{x}{1+\frac{2}{r^2-1}+\frac{1}{(r^2-1)^2(1-x)}}.$$

Hence the contribution to the rightmost expression in (5.30) of $x \in [0,1]$ satisfying

$$x > \delta\left(1+\frac{2}{r^2-1}+\frac{1}{(r^2-1)^2(1-x)}\right), \qquad \delta := \frac{\log \ell}{\ell},$$

is admissible for (1.14). By $r \geqslant 2$, the remaining values $x \in [0,1]$ satisfy

$$x < 3\delta \qquad \text{or} \qquad x(1-x) < \frac{3\delta}{(r^2-1)^2},$$

hence also $x < 3\delta$ or $1-x < 8\delta/r^4$. So the remaining contribution is

$$\leqslant \frac{2\ell+1}{r^2} \int\limits_{[0,3\delta)\cup(1-8\delta/r^4,1]} \frac{\mathrm{d}x}{1-x+r^{-4}x} \ll \frac{\log \ell}{r^2},$$

which is again admissible for (1.14).

By (5.12), it remains to show that

$$\operatorname{dist}(g,\mathcal{D}) \ll_\Lambda \|g\|\lambda/\sqrt{\ell}, \qquad \text{if } \mathcal{M} \neq \emptyset. \tag{5.31}$$

In the present case $q=\ell$, the condition (5.11) simplifies to

$$\begin{aligned} 2t\sin 2v_1\cos(\phi+\Delta) &= (1-t^2)\cos 2v_1 + (1+t^2) + \mathrm{O}_\Lambda(\lambda^2/\ell),\\ 2t\sin 2v_2\cos(\phi-\Delta) &= (t^2-1)\cos 2v_2 - (1+t^2) + \mathrm{O}_\Lambda(\lambda^2/\ell), \end{aligned} \tag{5.32}$$

hence for the proof of (5.31) we can and we shall assume that $|v_1+v_2| \leqslant \pi/2$. Indeed, replacing v_1 by $v_1+\pi$ has the effect of replacing g by $-g$ without altering Δ or the statement (5.31). We fix a pair $(\phi,t) \in \mathcal{M}$.

The two equations in (5.32) yield readily that

$$(\sin 2v_j)^2 \leqslant 2+2\cos 2v_j \ll_\Lambda t^2 + t|\sin 2v_j| + \lambda^2/\ell.$$

Hence $\sin 2v_j \ll_\Lambda t + \lambda/\sqrt{\ell}$, that is,

$$v_1, v_2 \in \frac{\pi}{2}\mathbb{Z} + \mathrm{O}_\Lambda\left(t+\frac{\lambda}{\sqrt{\ell}}\right). \tag{5.33}$$

Combining (5.32) with the Cauchy–Schwarz inequality, we also get

$$(1+t^2)^2 + \mathrm{O}_\Lambda(\lambda^2/\ell) \leqslant (1-t^2)^2 + 4t^2\cos^2(\phi\pm\Delta).$$

Equivalently,

$$\sin(\phi\pm\Delta) \ll_\Lambda \frac{\lambda}{t\sqrt{\ell}}.$$

Using also our initial assumption $|\Delta| \leqslant \pi/4$, we conclude that

$$\Delta \ll_\Lambda \frac{\lambda}{t\sqrt{\ell}} \qquad \text{and} \qquad \phi \in \pi\mathbb{Z} + \mathrm{O}_\Lambda\left(\frac{\lambda}{t\sqrt{\ell}}\right). \tag{5.34}$$

In particular, $\cos(\phi \pm \Delta) = \epsilon + \mathrm{O}_\Lambda(\lambda^2/t^2\ell)$ for some $\epsilon \in \{\pm 1\}$. Plugging this back to (5.32), and using also (5.33) along with

$$t\left(t + \frac{\lambda}{\sqrt{\ell}}\right) \min\left(1, \frac{\lambda^2}{t^2\ell}\right) \ll \frac{\lambda^2}{\ell},$$

we obtain

$$\begin{aligned} 2t\epsilon \sin 2v_1 &= (1-t^2)\cos 2v_1 + (1+t^2) + \mathrm{O}_\Lambda(\lambda^2/\ell), \\ 2t\epsilon \sin 2v_2 &= (t^2-1)\cos 2v_2 - (1+t^2) + \mathrm{O}_\Lambda(\lambda^2/\ell). \end{aligned} \tag{5.35}$$

Now consider the following three unit vectors in $\mathbb{R}^2$:

$$\mathbf{v}_1 := (\cos 2v_1, \sin 2v_1), \qquad \mathbf{v}_2 := (\cos 2v_2, -\sin 2v_2), \qquad \mathbf{t} := \left(\frac{t^2-1}{t^2+1}, \frac{2t\epsilon}{t^2+1}\right).$$

By (5.35), the scalar products $\mathbf{v}_j\mathbf{t}$ are $1 + \mathrm{O}_\Lambda(\lambda^2/\ell)$, hence the directed angles $\arg(\mathbf{v}_j) - \arg(\mathbf{t})$ lie in $2\pi\mathbb{Z} + \mathrm{O}_\Lambda(\lambda/\sqrt{\ell})$. It follows that

$$\arg(\mathbf{v}_1) - \arg(\mathbf{v}_2) \in 2\pi\mathbb{Z} + \mathrm{O}_\Lambda(\lambda/\sqrt{\ell}),$$

and then the assumption $|v_1 + v_2| \leqslant \pi/2$ forces that

$$v_1 + v_2 \ll_\Lambda \lambda/\sqrt{\ell}. \tag{5.36}$$

We are now ready to complete the proof of Theorem 6(b). By (5.33) and (5.36), there exists a multiple v of $\pi/2$ such that

$$\begin{aligned} m[v_1] &= m[v] + \mathrm{O}_\Lambda\big(t + \lambda/\sqrt{\ell}\big), \\ m[v_2] &= m[-v] + \mathrm{O}_\Lambda\big(t + \lambda/\sqrt{\ell}\big), \\ m[v_1 + v_2] &= \mathrm{id} + \mathrm{O}_\Lambda\big(\lambda/\sqrt{\ell}\big). \end{aligned}$$

Therefore, using also (5.13) and (5.34), we conclude that

$$\begin{aligned} &m[v_1]\,\mathrm{diag}\left(re^{i\Delta}, r^{-1}e^{-\Delta}\right) m[v_2] \\ &= m[v_1+v_2] + m[v_1]\,\mathrm{diag}\left(re^{i\Delta} - 1, r^{-1}e^{-i\Delta} - 1\right) m[v_2] \\ &= m[v_1+v_2] + m[v]\,\mathrm{diag}\left(re^{i\Delta} - 1, r^{-1}e^{-i\Delta} - 1\right) m[-v] + \mathrm{O}_\Lambda\big(r\lambda/\sqrt{\ell}\big) \\ &= m[v]\,\mathrm{diag}\left(re^{i\Delta}, r^{-1}e^{-i\Delta}\right) m[-v] + \mathrm{O}_\Lambda\big(r\lambda/\sqrt{\ell}\big). \end{aligned}$$

The main term $m[v]\,\mathrm{diag}\left(re^{i\Delta}, r^{-1}e^{-i\Delta}\right) m[-v]$ lies in $\mathcal{D}$, hence (5.31) follows.

The proof of Theorem 6(b) is complete.

6. Proof of Theorem 1

In this section, we prove Theorem 1. Lemma 4, which results from the amplified pre-trace inequality and estimates on the spherical trace function, proves an estimate on $|\Phi(g)|^2$ for $g \in \Omega$ in terms of the Diophantine counts $M(g, L, \mathcal{L}, \vec{\delta})$. We begin with the key remaining step of estimating these counts.

We allow all implied constants within this section to depend on Ω, and we drop the subscript from notation. Moreover, we adopt the notation $A \preccurlyeq B$ to mean that $|A| \ll_\varepsilon (\ell L)^\varepsilon B$, where $\varepsilon > 0$ is fixed but may be taken as small as desired at each step, and the implied constant is allowed to depend on ε.

For each $\mathcal{L} \in \{1, L^2, L^4\}$ and $\vec{\delta} = (\delta_1, \delta_2)$ with $0 < \delta_1, \delta_2 \leqslant \ell^\varepsilon$, we will estimate the count $M(g, L, \mathcal{L}, \vec{\delta})$ of matrices

$$\begin{aligned} &\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{M}_2(\mathbb{Z}[i]), && \det \gamma = n \in D(L, \mathcal{L}), && |n| \asymp \mathcal{L}^{1/2}, \\ &g^{-1}\tilde{\gamma} g = k \begin{pmatrix} z & u \\ & z^{-1} \end{pmatrix} k^{-1} && \text{for some } k \in K \text{ such that} \\ & && |z| \geqslant 1,\ \min|z \pm 1| \leqslant \delta_1,\ |u| \leqslant \delta_2, \end{aligned} \tag{6.1}$$

where as before $\tilde{\gamma} = \gamma/\sqrt{n}$. By the symmetry $\gamma \leftrightarrow -\gamma$, we can and we shall assume that $|z - 1| \leqslant |z + 1|$. Then the conditions imply that both $|z - 1|$ and $|z^{-1} - 1|$ are at most δ_1, hence

$$\left| \frac{a+d}{\sqrt{n}} - 2 \right| = |\operatorname{tr} \tilde{\gamma} - 2| = |z + z^{-1} - 2| = |z - 1||z^{-1} - 1| \leqslant \delta_1^2.$$

On the other hand, since $\|g\| \asymp_\Omega 1$, we also have that

$$\|\tilde{\gamma} - \mathrm{id}\| = \left\| gk \begin{pmatrix} z - 1 & u \\ & z^{-1} - 1 \end{pmatrix} k^{-1} g^{-1} \right\| \ll \delta_1 + \delta_2.$$

Summarizing, we need to estimate the number of matrices γ as in (6.1) such that

$$\left| a + d - 2\sqrt{n} \right| \leqslant \delta_1^2 \sqrt{|n|}, \qquad |a - d|, |b|, |c| \ll (\delta_1 + \delta_2)\sqrt{|n|}. \tag{6.2}$$

In particular, we have $|a + d| \preccurlyeq \sqrt{|n|}$ and

$$(a - d)^2 + 4bc = (a + d)^2 - 4n \preccurlyeq \delta_1^2 |n|. \tag{6.3}$$

As is often the case, parabolic matrices γ (those with trace $\pm 2\sqrt{n}$) play a distinctive role in this counting problem, and we split the count accordingly into the parabolic and non-parabolic subcounts as

$$M(g, L, \mathcal{L}, \vec{\delta}) = M^{\mathrm{p}}(g, L, \mathcal{L}, \vec{\delta}) + M^{\mathrm{np}}(g, L, \mathcal{L}, \vec{\delta}).$$

We shall prove the following result using (6.1), (6.2), and (6.3).

Lemma 9. *Let $\Omega \subset G$ be a compact subset, $L \geqslant 1$, and $\mathcal{L} \in \{1, L^2, L^4\}$. For $g \in \Omega$ and $\vec{\delta} = (\delta_1, \delta_2)$ with $0 < \delta_1, \delta_2 \preccurlyeq 1$, we have the following bounds:*

$$M(g, L, 1, \vec{\delta}) \preccurlyeq_\Omega 1, \tag{6.4}$$

$$M^{\mathrm{p}}(g, L, \mathcal{L}, \vec{\delta}) \preccurlyeq_\Omega \mathcal{L}^{1/2} + \mathcal{L}\delta_2^2, \tag{6.5}$$

$$M^{\mathrm{np}}(g, L, L^2, \vec{\delta}) \preccurlyeq_\Omega L^4 \delta_1^4 (\delta_1^2 + \delta_2^2), \tag{6.6}$$

$$M^{\mathrm{np}}(g, L, L^4, \vec{\delta}) \preccurlyeq_\Omega L^6 \delta_1^4 (\delta_1^2 + \delta_2^2). \tag{6.7}$$

Moreover,

$$M^{\mathrm{np}}(g, L, \mathcal{L}, \vec{\delta}) = 0 \qquad \textit{unless} \qquad \delta_1 \succcurlyeq \mathcal{L}^{-1/4}. \tag{6.8}$$

Proof. The bound (6.4) is immediate from (6.2). We turn to the bound (6.5), which counts parabolic matrices γ. In this case, we have $(a-d)^2+4bc=0$ and $z=1$, hence in particular (6.2) holds with 0 in place of δ_1. If $bc \neq 0$, then there are $\ll \mathcal{L}^{1/2}$ choices for $a+d=2\sqrt{n}$, and $\ll \mathcal{L}^{1/2}\delta_2^2$ choices for $a-d \neq 0$. The difference $a-d$ determines the product bc uniquely, hence by the divisor bound, there are $\preccurlyeq 1$ choices for (b,c). This is admissible for (6.5). If $bc=0$, then there are $\ll \mathcal{L}^{1/2}$ choices for $a=d=\sqrt{n}$, and $\ll 1+\mathcal{L}^{1/2}\delta_2^2$ choices for (b,c). This is again admissible for (6.5).

From now on we count non-parabolic matrices γ, in which case $(a-d)^2+4bc \neq 0$. The statement (6.8) is immediate from (6.3), so we are left with proving (6.6) and (6.7), where we may assume $\delta_1 \succcurlyeq \mathcal{L}^{-1/4}$. If $bc \neq 0$, then there are $\preccurlyeq \mathcal{L}^{1/2}(\delta_1^2+\delta_2^2)$ choices for $a-d$, and, for given $a-d$, there are $\preccurlyeq \mathcal{L}\delta_1^4$ choices for (b,c) by (6.3) and the divisor bound. If $bc=0$, then there are $\preccurlyeq \mathcal{L}^{1/2}\delta_1^2$ choices for $a-d$ by (6.3), and $\preccurlyeq \mathcal{L}^{1/2}(\delta_1^2+\delta_2^2)$ choices for (b,c). Altogether, there are $\preccurlyeq \mathcal{L}^{3/2}\delta_1^4(\delta_1^2+\delta_2^2)$ choices for the triple $(a-d,b,c)$. In the middle range $\mathcal{L}=L^2$, we additionally use that there are $\preccurlyeq \mathcal{L}^{1/2}$ choices for $a+d$, whence (6.6) follows. In the high range $\mathcal{L}=L^4$, $n=l_1^2l_2^2$ is a square, and $(a-d)^2+4bc \neq 0$ factors as $(a+d+2l_1l_2)(a+d-2l_1l_2)$. Hence the triple $(a-d,b,c)$ in fact determines $a+d$ up to $\preccurlyeq 1$ possibilities by the divisor bound, and (6.7) follows. □

Combining Lemmata 4 and 9, we obtain that

$$\sum_{\phi\in\mathfrak{B}} |\phi(g)|^2 \preccurlyeq_{I,\Omega} \ell^3 \left(\frac{1}{L} + S^{\mathrm{p}}(L) + S^{\mathrm{np}}(L,L^2) + S^{\mathrm{np}}(L,L^4)\right) + L^2\ell^{-48},$$

where

$$S^{\mathrm{p}}(L) := \sum_{\substack{\vec{\delta}\ \text{dyadic} \\ 1/\sqrt{\ell}\leqslant\delta_j\preccurlyeq 1}} \frac{1}{\sqrt{\ell}\delta_2}\left(\frac{L+L^2\delta_2^2}{L^3} + \frac{L^2+L^4\delta_2^2}{L^4}\right) \preccurlyeq \frac{1}{L^2} + \frac{1}{\sqrt{\ell}},$$

$$S^{\mathrm{np}}(L,L^2) := \sum_{\substack{\vec{\delta}\ \text{dyadic} \\ 1/\sqrt{\ell}\leqslant\delta_j\preccurlyeq 1}} \frac{1}{\ell\delta_1^2}\cdot\frac{L^4\delta_1^4(\delta_1^2+\delta_2^2)}{L^3} \preccurlyeq \frac{L}{\ell},$$

$$S^{\mathrm{np}}(L,L^4) := \sum_{\substack{\vec{\delta}\ \text{dyadic} \\ 1/\sqrt{\ell}\leqslant\delta_j\preccurlyeq 1}} \frac{1}{\ell\delta_1^2}\cdot\frac{L^6\delta_1^4(\delta_1^2+\delta_2^2)}{L^4} \preccurlyeq \frac{L^2}{\ell}.$$

Putting everything together, we conclude that

$$\sum_{\phi\in\mathfrak{B}} |\phi(g)|^2 \preccurlyeq_{I,\Omega} \ell^3\left(\frac{1}{L} + \frac{1}{\sqrt{\ell}} + \frac{L^2}{\ell}\right) + L^2\ell^{-48} \ll \ell^{8/3},$$

by making the essentially optimal choice $L := 7\ell^{1/3}$ (which satisfies our earlier condition $L \geqslant 7$).

The proof of Theorem 1 is complete.

7. Proof of Theorem 3

In this section, we prove Theorem 3. For $q = 0$, Lemma 5 provides an estimate on $|\phi_q(g)|^2$ for $g \in \Omega$ in terms of the Diophantine count $M_0^*(g, L, \mathcal{L}, \delta)$, while for $q = \pm\ell$ we need to analyze $Q(g, L, H_1, H_2)$ as follows from Lemma 6. We begin by estimating these counts. We keep the notational conventions from §6.

7.1. A comparison lemma

The Diophantine counts in Lemmata 5 and 6 involve the positioning relative to certain special sets of the matrix $g^{-1}\tilde{\gamma}g$, which we now explicate in preparation for a counting argument. Using $g \in \Omega$, we may write explicitly

$$g = \begin{pmatrix} g_1 & g_2 \\ g_3 & g_4 \end{pmatrix}, \qquad g_j \ll 1.$$

An explicit calculation shows that

$$g^{-1}\begin{pmatrix} a & b \\ c & d \end{pmatrix} g = \begin{pmatrix} \frac{a+d}{2} + L_1 & L_2 \\ L_3 & \frac{a+d}{2} - L_1 \end{pmatrix},$$

where

$$\begin{aligned} L_1 = &\quad (a-d)\big(\tfrac{1}{2} + g_2g_3\big) + bg_3g_4 - cg_1g_2, \\ L_2 = &\quad (a-d)g_2g_4 \qquad\quad + bg_4^2 \quad - cg_2^2, \\ L_3 = &-(a-d)g_1g_3 \qquad\quad - bg_3^2 \quad + cg_1^2. \end{aligned} \tag{7.1}$$

We record the following simple but effective result, which will be used in both parts of Theorem 3.

Lemma 10. *Let $\Omega \subset G$ be a compact subset, and $g \in \Omega$. Let $a, b, c, d \in \mathbb{C}$ and $\Delta > 0$ be such that $L_2, L_3 \ll \Delta$.*

(a) For at least one $s \in \{a-d, b, c\}$, we have

$$[a-d \quad b \quad c]^\top = [\lambda_1 \quad \lambda_2 \quad \lambda_3]^\top s + \mathrm{O}(\Delta)$$

with $\lambda_1, \lambda_2, \lambda_3 \ll 1$ depending only on g.

(b) For the same choice of $s \in \{a-d, b, c\}$, we have

$$(a-d)^2 + 4bc = \mu s^2 + \mathrm{O}(\Delta|s| + \Delta^2),$$

with $\mu = \lambda_1^2 + 4\lambda_2\lambda_3 \gg 1$. If additionally $(a-d)^2 + 4bc = 0$, then $a-d, b, c \ll \Delta$.

Proof. We may write the defining equations for L_2 and L_3 as $[L_2 \quad L_3]^\top = M\,[a-d \quad b \quad c]^\top$ for a 2×3 matrix M whose 2×2 minors we compute to be

$$\begin{vmatrix} g_4^2 & -g_2^2 \\ -g_3^2 & g_1^2 \end{vmatrix} = g_1g_4 + g_2g_3, \qquad \begin{vmatrix} g_2g_4 & -g_2^2 \\ -g_1g_3 & g_1^2 \end{vmatrix} = g_1g_2, \qquad \begin{vmatrix} g_2g_4 & g_4^2 \\ -g_1g_3 & -g_3^2 \end{vmatrix} = g_3g_4.$$

At least one of these minors exceeds $1/3$ in absolute value, since

$$(g_1g_4 + g_2g_3)^2 - 4g_1g_2g_3g_4 = 1.$$

Consider the case when $|g_1g_4 + g_2g_3| > 1/3$. Then we may solve the latter two equations in (7.1) for b, c, which yields

$$\begin{bmatrix} b \\ c \end{bmatrix} = \begin{bmatrix} -g_1g_2 \\ g_3g_4 \end{bmatrix} \frac{a-d}{g_1g_4+g_2g_3} + \mathrm{O}(\Delta).$$

This settles the first claim in the lemma with $s = a - d$. The second claim follows from

$$(a-d)^2 + 4bc = \frac{(a-d)^2}{(g_1g_4+g_2g_3)^2} + \mathrm{O}\big(\Delta|a-d| + \Delta^2\big).$$

The other cases (of which it suffices to consider one) are similar. For example, under $|g_1g_2| > 1/3$ we have

$$\begin{bmatrix} a-d \\ c \end{bmatrix} = \begin{bmatrix} -g_1g_4 - g_2g_3 \\ -g_3g_4 \end{bmatrix} \frac{b}{g_1g_2} + \mathrm{O}(\Delta), \qquad (a-d)^2 + 4bc = \frac{b^2}{(g_1g_2)^2} + \mathrm{O}\big(\Delta|b| + \Delta^2\big),$$

from which the lemma follows. □

7.2. Second moment count for $q = \pm\ell$

We will now establish an upper bound for the quantity $Q(g, L, H_1, H_2)$ counting pairs of matrices (γ_1, γ_2) such that

$$\begin{gathered} \gamma_j = \begin{pmatrix} a_j & b_j \\ c_j & d_j \end{pmatrix} \in \mathrm{M}_2(\mathbb{Z}[i]), \qquad \det\gamma_1 = \det\gamma_2 = n, \qquad L \leqslant |n| \leqslant 2L, \\ \|g^{-1}\tilde{\gamma}_j g\| \leqslant \sqrt{\frac{H_j}{L}}, \qquad \operatorname{dist}(g^{-1}\tilde{\gamma}_j g, \mathcal{D}) \ll \sqrt{\frac{H_j \log \ell}{L\ell}}. \end{gathered} \tag{7.2}$$

We denote the quantities in (7.1) corresponding to γ_j as L_{1j}, L_{2j}, L_{3j}. From (7.1) and (7.2) we deduce that

$$\|\gamma_j\| \ll \sqrt{H_j}, \qquad L_{2j}, L_{3j} \preccurlyeq \sqrt{H_j/\ell}, \tag{7.3}$$

and

$$(a_1+d_1)^2 - (a_2+d_2)^2 = (a_1-d_1)^2 + 4b_1c_1 - (a_2-d_2)^2 - 4b_2c_2. \tag{7.4}$$

We shall prove the following result using (7.2), (7.3), and (7.4).

Lemma 11. *Let $\Omega \subset G$ be a compact subset and $L \geqslant 1$. For $g \in \Omega$ and $1 \leqslant H_1, H_2 \preccurlyeq \ell$, we have*

$$Q(g, L, H_1, H_2) \preccurlyeq_\Omega H_1H_2. \tag{7.5}$$

Proof. We shall use that the entries $a_j, b_j, c_j, d_j \in \mathbb{Z}[i]$ of each participating γ_j satisfy the conditions of Lemma 10 with $\Delta_j \preccurlyeq 1$ in the role of Δ. Indeed, this follows from (7.3) and $H_1, H_2 \preccurlyeq \ell$.

Let $s_j \in \{a_j - d_j, b_j, c_j\}$ be as in Lemma 10(a). By Lemma 10(a) and (7.3), for a given pair (s_1, s_2), there are $\preccurlyeq 1$ choices for the two triples $(a_j - d_j, b_j, c_j)$, which then determine both sides of (7.4). Using this preliminary observation, we do the counting in two steps.

First we count (γ_1, γ_2) satisfying (7.2) and $(a_1+d_1)^2 \neq (a_2+d_2)^2$. By (7.3), there are $\ll H_1H_2$ choices for the pair (s_1, s_2), hence $\preccurlyeq H_1H_2$ choices for the two triples $(a_j - d_j, b_j, c_j)$. Given the triples, by (7.3)–(7.4) and the divisor bound, there are $\preccurlyeq 1$ choices for $(a_1 + d_1, a_2 + d_2)$. This is admissible for (7.5).

Now we count (γ_1, γ_2) satisfying (7.2) and $(a_1 + d_1)^2 = (a_2 + d_2)^2$. In this case, Lemma 10(b) coupled with (7.3)–(7.4) shows that $s_1^2 - s_2^2 \preccurlyeq \sqrt{H_1} + \sqrt{H_2}$. Hence, by the divisor bound (separating the case when $s_1^2 = s_2^2$), there are $\preccurlyeq \max(H_1, H_2)$ choices for the pair (s_1, s_2) and same for the two triples $(a_j - d_j, b_j, c_j)$. Independently of the triples, by (7.3), there are $\ll \min(H_1, H_2)$ choices for $(a_1 + d_1, a_2 + d_2)$. This is again admissible for (7.5). □

7.3. Interlude: a first moment count

For the proof of Theorem 2 in §8 below, we need a variation of the previous Diophantine argument that is most conveniently stated and proved at this point. For $\mathcal{L} \in \{1, L^2, L^4\}$ and every $0 < \delta \preccurlyeq 1$, we will establish an upper bound on the quantity

$$M_{\mathcal{D}}(g, L, \mathcal{L}, \varepsilon, \delta) := \sum_{n \in D(L,\mathcal{L})} \#\left\{\gamma \in \Gamma_n : \|g^{-1}\tilde{\gamma}g\| \ll \ell^{\varepsilon},\ \operatorname{dist}(g^{-1}\tilde{\gamma}g, \mathcal{D}) \leqslant \delta\right\}, \tag{7.6}$$

where the implied constant is absolute. As before, we conclude from the conditions in (7.6) and the explicit description in (7.1) that

$$\|\gamma\| \preccurlyeq \mathcal{L}^{1/4} \qquad \text{and} \qquad L_2, L_3 \ll \mathcal{L}^{1/4}\delta. \tag{7.7}$$

We shall prove the following result using (7.7) and the identity

$$(a - d)^2 + 4bc = (a + d)^2 - 4n. \tag{7.8}$$

Lemma 12. *Let $\Omega \subset G$ be a compact subset, $L \geqslant 1$, and $\varepsilon > 0$. For $g \in \Omega$ and $0 < \delta \preccurlyeq 1$, we have the following bounds:*

$$M_{\mathcal{D}}(g, L, 1, \varepsilon, \delta) \preccurlyeq_{\Omega} 1, \tag{7.9}$$

$$M_{\mathcal{D}}(g, L, L^2, \varepsilon, \delta) \preccurlyeq_{\Omega} L^2 + L^4\delta^4, \tag{7.10}$$

$$M_{\mathcal{D}}(g, L, L^4, \varepsilon, \delta) \preccurlyeq_{\Omega} L^2 + L^6\delta^4. \tag{7.11}$$

Proof. The bound (7.9) corresponds to $\mathcal{L} = 1$, and it is immediate from (7.7). Hence we focus on the bounds (7.10)–(7.11) that correspond to $\mathcal{L} \in \{L^2, L^4\}$. We shall use that the entries $a, b, c, d \in \mathbb{Z}[i]$ of each participating γ satisfy the conditions of Lemma 10 with $\Delta = \mathcal{L}^{1/4}\delta$, as follows from (7.7).

First we count parabolic matrices γ. In this case, we have $(a-d)^2+4bc = 0$, hence also $a-d, b, c \ll \mathcal{L}^{1/4}\delta$ by Lemma 10(b). If $bc \neq 0$, then there are $\ll \mathcal{L}^{1/2}$ choices for $a + d = \pm 2\sqrt{n}$, and $\ll \mathcal{L}^{1/2}\delta^2$ choices for $a - d \neq 0$. The difference $a - d$ determines the product bc uniquely, hence by the divisor bound, there are $\preccurlyeq 1$ choices for (b, c). This is admissible for (7.10)–(7.11). If $bc = 0$, then there are $\ll \mathcal{L}^{1/2}$ choices for $a = d = \pm\sqrt{n}$, and $\ll 1 + \mathcal{L}^{1/2}\delta^2$ choices for (b, c). This is again admissible for (7.10)–(7.11).

Now we count non-parabolic matrices γ, in which case $(a - d)^2 + 4bc \neq 0$. Let $s \in \{a - d, b, c\}$ be as in Lemma 10(a). There are $\preccurlyeq \mathcal{L}^{1/2}$ choices for s, and for a given s, there are $\ll 1 + \mathcal{L}\delta^4$ choices for the triple $(a-d, b, c)$ by Lemma 10(a). Altogether, there are $\preccurlyeq \mathcal{L}^{1/2} + \mathcal{L}^{3/2}\delta^4$ choices for the triple $(a-d, b, c)$. In the middle range $\mathcal{L} = L^2$, we additionally use that there are $\preccurlyeq \mathcal{L}^{1/2}$ choices for $a + d$, whence (7.10) follows. In the high range $\mathcal{L} = L^4$, $n = l_1^2 l_2^2$ is a square, and $(a - d)^2 + 4bc \neq 0$ factors as $(a + d + 2l_1l_2)(a + d - 2l_1l_2)$. Hence the triple $(a - d, b, c)$ in fact determines $a + d$ up to $\preccurlyeq 1$ possibilities by the divisor bound, and (7.11) follows. □

7.4. Counting setup for $q = 0$

For each $\mathcal{L} \in \{1, L^2, L^4\}$ and $0 < \delta \preccurlyeq 1$, we will establish an upper bound on the quantity $M_0^*(g, L, \mathcal{L}, \delta)$ consisting of matrices

$$\begin{gathered}\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{M}_2(\mathbb{Z}[i]), \qquad \det\gamma = n \in D(L, \mathcal{L}), \qquad |n| \asymp \mathcal{L}^{1/2} \\ \operatorname{dist}(g^{-1}\tilde{\gamma}g, \mathcal{S}) \leqslant \delta, \qquad \frac{D(g^{-1}\tilde{\gamma}g)}{\|g^{-1}\tilde{\gamma}g\|^2} \ll \frac{\log \ell}{\sqrt{\ell}}.\end{gathered} \tag{7.12}$$

From the first distance condition in (7.12) we conclude that

$$a, b, c, d \preccurlyeq \mathcal{L}^{1/4}. \tag{7.13}$$

Using the description in (7.1), the distance conditions in (7.12) imply that

$$\begin{cases} L_2, L_3 \ll \delta\sqrt{|n|} \\ \left|\frac{a+d}{2} \pm L_1\right| = (1 + \mathrm{O}(\delta))\sqrt{|n|} \end{cases} \quad \text{or} \quad \begin{cases} a + d, L_1 \ll \delta\sqrt{|n|} \\ |L_2|, |L_3| = (1 + \mathrm{O}(\delta))\sqrt{|n|}; \end{cases} \tag{7.14}$$

$$\left|\tfrac{a+d}{2} + L_1\right|^2 - \left|\tfrac{a+d}{2} - L_1\right|^2 \preccurlyeq \sqrt{\mathcal{L}/\ell} \qquad \text{and} \qquad |L_2|^2 - |L_3|^2 \preccurlyeq \sqrt{\mathcal{L}/\ell}. \tag{7.15}$$

As in §6, we split the count into the parabolic and non-parabolic subcounts as

$$M_0^*(g, L, \mathcal{L}, \delta) = M_0^{*\mathrm{p}}(g, L, \mathcal{L}, \delta) + M_0^{*\mathrm{np}}(g, L, \mathcal{L}, \delta).$$

We shall prove the following result using (7.12)–(7.15) and (7.8).

Lemma 13. *Let $\Omega \subset G$ be a compact subset, $L \geqslant 1$, and $\mathcal{L} \in \{L^2, L^4\}$. For $g \in \Omega$ and $0 < \delta \preccurlyeq 1$, we have the following bounds:*

$$M_0^*(g, L, 1, \delta) \preccurlyeq_\Omega 1, \tag{7.16}$$

$$M_0^{*\mathrm{p}}(g, L, \mathcal{L}, \delta) \preccurlyeq_\Omega \mathcal{L}^{1/2} + \mathcal{L}\delta^2, \tag{7.17}$$

$$M_0^*(g, L, L^2, \delta) \preccurlyeq_\Omega L^{3/2} + L^3\delta^3 + \frac{L^2 + L^{7/2}\delta^2}{\sqrt{\ell}} + \frac{L^4\delta^2}{\ell}, \tag{7.18}$$

$$M_0^{*\mathrm{np}}(g, L, L^4, \delta) \preccurlyeq_\Omega L^3 + L^5\delta^2 + \frac{L^4 + L^6\delta^2}{\sqrt{\ell}}. \tag{7.19}$$

Proof. The bound (7.16) is immediate from (7.13). For the proof of (7.17), we observe that, in the parabolic case, (7.14) implies $L_2, L_3 \ll \mathcal{L}^{1/4}\delta$. Indeed, this is clear when the first half of (7.14) holds. Otherwise, the conditions $a + d = \pm 2\sqrt{n}$ and $a + d \ll \delta\sqrt{|n|}$ force $\delta \gg 1$, so the claimed bound is clear again. Applying Lemma 10(b), we infer that $a - d, b, c \ll \mathcal{L}^{1/4}\delta$ holds in the parabolic case. From here (7.17) follows readily, as in the second paragraph of the proof of Lemma 12. Finally, we shall prove (7.18) and (7.19) in the next two subsections. □

7.5. Volume argument

Here, we present a volume argument that we will use repeatedly to estimate the number of lattice points satisfying (7.12)–(7.15). The symbol vol will refer to the Lebesgue measure in $\mathbb{C}^m \simeq \mathbb{R}^{2m}$, with m being clear from the context.

The explicit expressions for the linear forms in (7.1) may be rewritten as

$$[L_1 \quad L_2 \quad L_3]^\top = A_0(g)\,[a-d \quad b \quad c]^\top, \tag{7.20}$$

where $A_0 : \Omega \to \mathrm{GL}_3(\mathbb{C})$ is a continuous function. It is straightforward to verify that $\det A_0(g) = 1/2$ holds identically. We shall also use the 4-dimensional variant

$$[a+d \quad L_1 \quad L_2 \quad L_3]^\top = \mathrm{diag}(1, A_0(g))\,[a+d \quad a-d \quad b \quad c]^\top. \tag{7.21}$$

Now, let $m \geqslant 1$ be a fixed integer ($m \in \{2,3,4\}$ in our applications), and let $A : \Omega \to \mathrm{GL}_m(\mathbb{C})$ be a fixed continuous function. As Ω is compact, there exists a fixed compact subset $K = K(A,\Omega) \subset \mathbb{C}^m$ such that each $2m$-dimensional lattice $A(g)\mathbb{Z}[i]^m \subset \mathbb{C}^m$ ($g \in \Omega$) has a fundamental parallelepiped lying in K and of volume $\asymp 1$. It follows by a standard volume argument that for any compact subset $V \subset \mathbb{C}^m$ and $g \in \Omega$ we have

$$\#\big(V \cap A(g)\mathbb{Z}[i]^m\big) \ll \mathrm{vol}\, V^\bullet \qquad \text{where} \qquad V^\bullet := V + K. \tag{7.22}$$

We also record for repeated reference a simple volume computation. For $r, \Delta > 0$, we define the sets

$$W_1(r,\Delta) := \big\{(z_1,z_2) \in \mathbb{C}^2 : |z_1|, |z_2| \leqslant r,\ \Re(z_1\overline{z_2}) \leqslant \Delta\big\},$$
$$W_2(r,\Delta) := \big\{(z_1,z_2) \in \mathbb{C}^2 : |z_1|, |z_2| \leqslant r,\ \big||z_1|^2 - |z_2|^2\big| \leqslant \Delta\big\}.$$

Cutting these into two parts according to whether $|z_2| \leqslant |z_1|$ or $|z_2| > |z_1|$, we obtain readily by Fubini's theorem that

$$\mathrm{vol}\, W_j(r,\Delta) \ll \min(r^4, r^2\Delta).$$

On the other hand, we have

$$W_j(r,\Delta)^\bullet \subset W_j\big(r + \mathrm{O}(1), \Delta + \mathrm{O}(r+1)\big)$$

with implied constants depending only on A and Ω, hence

$$\mathrm{vol}\, W_j(r,\Delta)^\bullet \ll \min\big((r+1)^4, (r+1)^2(\Delta + r + 1)\big) \ll 1 + r^2\Delta + r^3. \tag{7.23}$$

7.6. Middle and high range for $q = 0$

We now estimate the count $M_0^*(g, L, \mathcal{L}, \delta)$ in the "middle range" $\mathcal{L} = L^2$ and the "high range" $\mathcal{L} = L^4$. In the high range, we shall focus on the non-parabolic contribution $M_0^{*\mathrm{np}}(g, L, L^4, \delta)$, since we have already proved (7.17), and here we shall profit substantially from the fact that $\det \gamma$ is a square.

7.6.1. Middle range

In the middle range $\mathcal{L} = L^2$, we estimate the number of choices in $M_0^*(g, L, L^2, \delta)$ as follows.

For the case when the first half of (7.14) holds, we introduce the set

$$V_1(\delta) := \big\{(z_0,z_1,z_2,z_3) \in \mathbb{C}^4 : z_0, z_1 \preccurlyeq \mathcal{L}^{1/4},\ \Re(z_0\overline{z_1}) \preccurlyeq \sqrt{\mathcal{L}/\ell},$$
$$z_2, z_3 \ll \mathcal{L}^{1/4}\delta,\ |z_2|^2 - |z_3|^2 \preccurlyeq \sqrt{\mathcal{L}/\ell}\big\},$$

suppressing from notation the dependence implicit in $\preccurlyeq$. Then we have by (7.23)

$$\begin{aligned}\operatorname{vol} V_1(\delta)^\bullet &\preccurlyeq \operatorname{vol} W_1(\mathcal{L}^{1/4}, \sqrt{\mathcal{L}/\ell})^\bullet \cdot \operatorname{vol} W_2(\mathcal{L}^{1/4}\delta, \sqrt{\mathcal{L}/\ell})^\bullet \\ &\ll (\mathcal{L}^{3/4} + \mathcal{L}/\sqrt{\ell})(1 + \mathcal{L}^{3/4}\delta^3 + \mathcal{L}\delta^2/\sqrt{\ell}). \end{aligned} \tag{7.24}$$

For the case when the second half of (7.14) holds, we introduce the set

$$\begin{aligned}V_2(\delta) = \big\{(z_0, z_1, z_2, z_3) \in \mathbb{C}^4 : z_0, z_1 \ll \mathcal{L}^{1/4}\delta,\ \Re(z_0\overline{z_1}) \preccurlyeq \sqrt{\mathcal{L}/\ell},\\ z_2, z_3 \preccurlyeq \mathcal{L}^{1/4},\ |z_2|^2 - |z_3|^2 \preccurlyeq \sqrt{\mathcal{L}/\ell}\big\},\end{aligned}$$

suppressing from notation the dependence implicit in $\preccurlyeq$. Then we have by (7.23)

$$\begin{aligned}\operatorname{vol} V_2(\delta)^\bullet &\preccurlyeq \operatorname{vol} W_1(\mathcal{L}^{1/4}\delta, \sqrt{\mathcal{L}/\ell})^\bullet \cdot \operatorname{vol} W_2(\mathcal{L}^{1/4}, \sqrt{\mathcal{L}/\ell})^\bullet \\ &\ll (\mathcal{L}^{3/4} + \mathcal{L}/\sqrt{\ell})(1 + \mathcal{L}^{3/4}\delta^3 + \mathcal{L}\delta^2/\sqrt{\ell}). \end{aligned} \tag{7.25}$$

Using (7.12)–(7.15), (7.21)–(7.22), and (7.24)–(7.25), we conclude (7.18) in the form

$$M_0^*(g, L, L^2, \delta) \preccurlyeq (L^{3/2} + L^2/\sqrt{\ell})(1 + L^{3/2}\delta^3 + L^2\delta^2/\sqrt{\ell}).$$

7.6.2. High range

As in the proof of Lemmata 9 and 12, in the high range $\mathcal{L} = L^4$, once the triple $(a-d, b, c)$ is determined for a non-parabolic matrix γ (so that (7.8) holds), $a+d$ and along with it γ is determined up to $\preccurlyeq 1$ choices by the divisor bound, using that $n = l_1^2 l_2^2$ is a square. We now estimate the number of choices in $M_0^{*\mathrm{np}}(g, L, L^4, \delta)$ as follows.

For the case when the first half of (7.14) holds, we introduce the set

$$V_3(\delta) := \big\{(z_1, z_2, z_3) \in \mathbb{C}^3 : z_1 \preccurlyeq \mathcal{L}^{1/4},\ z_2, z_3 \ll \mathcal{L}^{1/4}\delta,\ |z_2|^2 - |z_3|^2 \preccurlyeq \sqrt{\mathcal{L}/\ell}\big\},$$

suppressing from notation the dependence implicit in $\preccurlyeq$. Then we have by (7.23)

$$\operatorname{vol} V_3(\delta)^\bullet \preccurlyeq \sqrt{\mathcal{L}} \cdot \operatorname{vol} W_2(\mathcal{L}^{1/4}\delta, \sqrt{\mathcal{L}/\ell})^\bullet \ll \sqrt{\mathcal{L}}(1 + \mathcal{L}\delta^2/\sqrt{\ell} + \mathcal{L}^{3/4}\delta^3). \tag{7.26}$$

For the case when the second half of (7.14) holds, we introduce the set

$$V_4(\delta) := \big\{(z_1, z_2, z_3) \in \mathbb{C}^3 : z_1 \ll \mathcal{L}^{1/4}\delta,\ z_2, z_3 \preccurlyeq \mathcal{L}^{1/4},\ |z_2|^2 - |z_3|^2 \preccurlyeq \sqrt{\mathcal{L}/\ell}\big\},$$

suppressing from notation the dependence implicit in $\preccurlyeq$. Then we have by (7.23)

$$\operatorname{vol} V_4(\delta)^\bullet \preccurlyeq (1 + \mathcal{L}^{1/4}\delta)^2 \cdot \operatorname{vol} W_2(\mathcal{L}^{1/4}, \sqrt{\mathcal{L}/\ell})^\bullet \ll (1 + \sqrt{\mathcal{L}}\delta^2)(\mathcal{L}^{3/4} + \mathcal{L}/\sqrt{\ell}). \tag{7.27}$$

Using (7.8), (7.12)–(7.15), (7.20), (7.22), and (7.26)–(7.27), we conclude (7.19) in the form

$$M_0^{*\mathrm{np}}(g, L, L^4, \delta) \preccurlyeq L^2(1 + L^4\delta^2/\sqrt{\ell} + L^3\delta^3) + (1 + L^2\delta^2)(L^3 + L^4/\sqrt{\ell}).$$

The proof of Lemma 13 is complete.

7.7. Proof of Theorem 3

In the case $q = 0$, we combine Lemmata 5 and 13 to see that

$$|\phi_0(g)|^2 \preccurlyeq_{I,\Omega} \ell^2 \left(\frac{1}{L} + S_0^*(L, L^2) + S_0^*(L, L^4)\right) + L^2\ell^{-48},$$

where

$$S_0^*(L, L^2) := \sum_{\substack{\delta \text{ dyadic} \\ 1/\sqrt{\ell} \leqslant \delta \preccurlyeq 1}} \frac{1}{\sqrt{\ell}\delta L^3}\left(L^{3/2} + L^3\delta^3 + \frac{L^2 + L^{7/2}\delta^2}{\sqrt{\ell}} + \frac{L^4\delta^2}{\ell}\right) \preccurlyeq \frac{1}{L^{3/2}} + \frac{1}{\sqrt{\ell}} + \frac{L}{\ell^{3/2}},$$

$$S_0^*(L, L^4) := \sum_{\substack{\delta \text{ dyadic} \\ 1/\sqrt{\ell} \leqslant \delta \preccurlyeq 1}} \frac{1}{\sqrt{\ell}\delta L^4}\left(L^3 + L^5\delta^2 + \frac{L^4 + L^6\delta^2}{\sqrt{\ell}}\right) \preccurlyeq \frac{1}{L} + \frac{L}{\sqrt{\ell}} + \frac{L^2}{\ell}.$$

Putting everything together, we conclude that

$$|\phi_0(g)|^2 \preccurlyeq_{I,\Omega} \ell^2 \left(\frac{1}{L} + \frac{L}{\sqrt{\ell}} + \frac{L^2}{\ell}\right) + L^2\ell^{-48} \ll \ell^{7/4},$$

by making the essentially optimal choice $L := 7\ell^{1/4}$ (which satisfies our earlier condition $L \geqslant 7$).

The case $q = \pm\ell$ is immediate from Lemmata 6 and 11, hence the proof of Theorem 3 is complete.

8. Proof of Theorem 2

In this section, we prove Theorem 2. Here we take the aim of the softest possible proof based on the localization properties of the averaged spherical trace function (proved in Theorem 5 and then encoded in the form of the amplified pre-trace inequality in Lemma 5) and the already available ingredients for the counting problem.

For each $\mathcal{L} \in \{1, L^2, L^4\}$ and $\vec{\delta} = (\delta_1, \delta_2)$ with $0 < \delta_1, \delta_2 \leqslant \ell^{\varepsilon}$, the count $M^*(g, L, \mathcal{L}, \vec{\delta})$ in Lemma 5 may be estimated in a split fashion as

$$M^*(g, L, \mathcal{L}, \vec{\delta}) \leqslant \min\left(M_K(g, L, \mathcal{L}, \delta_1), M_{\mathcal{D}}(g, L, \mathcal{L}, \varepsilon, \delta_2)\right),$$

where

$$M_K(g, L, \mathcal{L}, \delta) := \sum_{n \in D(L,\mathcal{L})} \#\left\{\gamma \in \Gamma_n : \operatorname{dist}\left(g^{-1}\tilde{\gamma}g, K\right) \leqslant \delta\right\},$$

and $M_{\mathcal{D}}(g, L, \mathcal{L}, \varepsilon, \delta)$ is as in (7.6). The quantity $M_K(g, L, \mathcal{L}, \delta)$ is the classical Diophantine count in the spherical sup-norm problem in the eigenvalue aspect, which in the present context was treated in detail in [2]. In the notation of that paper, we have:

- $u(\tilde{\gamma}gK, gK) \asymp \operatorname{dist}(g^{-1}\tilde{\gamma}g, K)^2$ in [2, (5.3)];
- $N = 1$, and $r \asymp_{\Omega} 1$ for $g \in \Omega$, in [2, (6.2)].

Thus the count $M_K(g, L, \mathcal{L}, \delta_1)$ agrees with $M(gK, L, \mathcal{L}, \mathrm{O}(\delta_1^2))$ in [2, (5.17)–(5.18)]. Importing estimates [2, (7.1), (7.2), (7.5), (11.1), (11.6)], we conclude that

$$M_K(g,L,1,\delta_1) \preccurlyeq_\Omega 1, \quad M_K(g,L,L^2,\delta_1) \preccurlyeq_\Omega L^2 + L^4\delta_1, \quad M_K(g,L,L^4,\delta_1) \preccurlyeq_\Omega L^3 + L^6\delta_1.$$

The count $M_{\mathcal{D}}(g,L,\mathcal{L},\varepsilon,\delta)$ was estimated in Lemma 12. Combining everything, we obtain the following lemma.

Lemma 14. *For $g \in \Omega$, $L > 0$, and arbitrary $\varepsilon > 0$ and $\vec{\delta} = (\delta_1,\delta_2)$ with $0 < \delta_j \preccurlyeq 1$, the quantity $M^*(g,L,\mathcal{L},\vec{\delta})$ in Lemma 5 satisfies*

$$\begin{aligned} M^*(g,L,1,\vec{\delta}) &\preccurlyeq_\Omega 1,\\ M^*(g,L,L^2,\vec{\delta}) &\preccurlyeq_\Omega \min\left(L^2 + L^4\delta_1, L^2 + L^4\delta_2^4\right),\\ M^*(g,L,L^4,\vec{\delta}) &\preccurlyeq_\Omega \min\left(L^3 + L^6\delta_1, L^2 + L^6\delta_2^4\right). \end{aligned}$$

We are now ready for the proof of Theorem 2. From Lemma 14, we have for every pair $\vec{\delta} = (\delta_1,\delta_2)$ with $0 < \delta_1,\delta_2 \leqslant \ell^\varepsilon$ that

$$\frac{M^*(g,L,1,\vec{\delta})}{L} + \frac{M^*(g,L,L^2,\vec{\delta})}{L^3} + \frac{M^*(g,L,L^4,\vec{\delta})}{L^4} \preccurlyeq_\Omega \left(\frac{1}{L} + L^2\min\left(\delta_1,\delta_2^4\right)\right).$$

Inserting this into Lemma 5, we find that

$$\begin{aligned} |\phi_q(g)|^2 &\preccurlyeq_{I,\Omega} \ell^2 \sum_{\substack{\vec{\delta}\text{ dyadic, }\delta_j\preccurlyeq 1\\ \delta_1^2\delta_2\geqslant 1/\sqrt{\ell}}} \frac{1}{\sqrt{\ell}\delta_1^2\delta_2}\left(\frac{1}{L} + L^2\min\left(\delta_1,\delta_2^4\right)\right) + L^2\ell^{-48}\\ &\preccurlyeq \ell^2\left(\frac{1}{L} + \sum_{\substack{\vec{\delta}\text{ dyadic, }\delta_j\preccurlyeq 1\\ \delta_1^2\delta_2\geqslant 1/\sqrt{\ell}}} L^2\min\left(\frac{1}{\sqrt{\ell}\delta_1\delta_2},\delta_1,\delta_2^4\right)\right) \preccurlyeq \ell^2\left(\frac{1}{L} + \frac{L^2}{\ell^{2/9}}\right), \end{aligned}$$

where we used $\min(A,B,C) \leqslant A^{4/9}B^{4/9}C^{1/9}$ in the last step. The choice $L := 7\ell^{2/27}$ is optimal up to a constant, and it satisfies our earlier condition $L \geqslant 7$, hence we obtain Theorem 2 in the form

$$\|\phi_q|_\Omega\|_\infty \preccurlyeq_{I,\Omega} \ell^{26/27}.$$

The proof of Theorem 2 is complete.

Acknowledgements

This work began during D.M.'s term as Director's Mathematician in Residence at the Budapest Semesters of Mathematics program in the summer of 2018; D.M. would like to thank BSM, the Alfréd Rényi Institute of Mathematics, as well as the Max Planck Institute for Mathematics for their hospitality and excellent working conditions.

This work was supported in part by the DFG-SNF Lead Agency Program grant BL 915/2-2 (V.B.), Germany's Excellence Strategy grant EXC-2047/1 - 390685813 (V.B.), National Science Foundation Grant DMS-1903301 (D.M.), European Research Council grant CoG-648017 (G.H.), the MTA Rényi Intézet Lendület Automorphic Research Group (G.H. & P.M.), NKFIH (National Research, Development and Innovation Office) grants K 119528 (G.H. & P.M.), KKP 133819 (P.M.), FK 135218 (P.M.), and the Premium Postdoctoral Fellowship of the Hungarian Academy of Sciences (P.M.).

References

[1] J. Arthur, A Paley-Wiener theorem for real reductive groups, Acta Math. 150 (1–2) (1983) 1–89.
[2] V. Blomer, G. Harcos, D. Milićević, Bounds for eigenforms on arithmetic hyperbolic 3-manifolds, Duke Math. J. 165 (4) (2016) 625–659.
[3] V. Blomer, G. Harcos, P. Maga, D. Milićević, The sup-norm problem for GL(2) over number fields, J. Eur. Math. Soc. 22 (1) (2020) 1–53.
[4] V. Blomer, Epstein zeta-functions, subconvexity, and the purity conjecture, J. Inst. Math. Jussieu 19 (2) (2020) 581–596.
[5] V. Blomer, P. Maga, Subconvexity for sup-norms of cusp forms on PGL(n), Sel. Math. New Ser. 22 (3) (2016) 1269–1287.
[6] F. Brumley, S. Marshall, Lower bounds for Maass forms on semisimple groups, Compos. Math. 156 (5) (2020) 959–1003.
[7] V. Blomer, A. Pohl, The sup-norm problem on the Siegel modular space of rank two, Am. J. Math. 138 (4) (2016) 999–1027.
[8] R. Breusch, Zur Verallgemeinerung des Bertrandschen Postulates, daß zwischen x und $2x$ stets Primzahlen liegen, Math. Z. 34 (1) (1932) 505–526.
[9] D. Bump, Automorphic Forms and Representations, Cambridge Studies in Advanced Mathematics, vol. 55, Cambridge University Press, Cambridge, 1997.
[10] O.A. Campoli, Paley-Wiener type theorems for rank-1 semisimple Lie groups, Rev. Unión Mat. Argent. 29 (3) (1979/80) 197–221.
[11] P. Delorme, M. Flensted-Jensen, Towards a Paley-Wiener theorem for semisimple symmetric spaces, Acta Math. 167 (1–2) (1991) 127–151.
[12] J. Dixmier, P. Malliavin, Factorisations de fonctions et de vecteurs indéfiniment différentiables, Bull. Sci. Math. (2) 102 (4) (1978) 307–330.
[13] L. Dever, D. Milićević, Ambient prime geodesic theorems on hyperbolic 3-manifolds, Int. Math. Res. Not. (2021) rnab048.
[14] J. Elstrodt, F. Grunewald, J. Mennicke, Groups Acting on Hyperbolic Space: Harmonic Analysis and Number Theory, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 1998.
[15] M. Frączyk, G. Harcos, P. Maga, D. Milićević, The density hypothesis for horizontal families of lattices, arXiv e-prints, arXiv:2007.13961, 2020.
[16] R. Gangolli, On the Plancherel formula and the Paley-Wiener theorem for spherical functions on semisimple Lie groups, Ann. Math. (2) 93 (1971) 150–165.
[17] I.M. Gelfand, M.I. Graev, N.Ya. Vilenkin, Generalized Functions. Vol. 5: Integral Geometry and Representation Theory, Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1966 [1977], translated from the Russian by Eugene Saletan.
[18] S. Gelbart, H. Jacquet, A relation between automorphic representations of GL(2) and GL(3), Ann. Sci. Éc. Norm. Supér. (4) 11 (4) (1978) 471–542.
[19] I.M. Gelfand, M.A. Naimark, Unitary representations of the Lorentz group, Izv. Akad. Nauk SSSR, Ser. Mat. 11 (1947) 411–504.
[20] I.M. Gelfand, M.A. Naimark, Unitary Representations of the Classical Groups, Trudy Mat. Inst. Steklova, vol. 36, Izdat. Nauk SSSR, Moscow-Leningrad, 1950.
[21] I.M. Gelfand, M.A. Naimark, Unitäre Darstellungen der klassischen Gruppen, Akademie-Verlag, Berlin, 1957.
[22] I.M. Gelfand, M.A. Naimark, Unitary representations of the Lorentz group, in: Collected Papers of I.M. Gelfand, Vol. II, Springer-Verlag, Berlin, 1988, pp. 41–123.
[23] A. Ghosh, A. Reznikov, P. Sarnak, Nodal domains of Maass forms I, Geom. Funct. Anal. 23 (5) (2013) 1515–1568.
[24] G. Harcos, Uniform approximate functional equation for principal L-functions, Int. Math. Res. Not. (18) (2002) 923–932.
[25] E. Hewitt, K.A. Ross, Abstract Harmonic Analysis. Vol. II: Structure and Analysis for Compact Groups. Analysis on Locally Compact Abelian Groups, Die Grundlehren der mathematischen Wissenschaften, Band 152, Springer-Verlag, New York-Berlin, 1970.
[26] G. Harcos, N. Templier, On the sup-norm of Maass cusp forms of large level. III, Math. Ann. 356 (1) (2013) 209–216.
[27] H. Iwaniec, P. Sarnak, L^∞ norms of eigenfunctions of arithmetic surfaces, Ann. Math. (2) 141 (2) (1995) 301–320.
[28] H. Jacquet, Automorphic Forms on GL(2). Part II, Lecture Notes in Mathematics, vol. 278, Springer-Verlag, Berlin-New York, 1972.
[29] J. Jorgenson, J. Kramer, Bounding the sup-norm of automorphic forms, Geom. Funct. Anal. 14 (6) (2004) 1267–1277.
[30] H. Jacquet, R.P. Langlands, Automorphic Forms on GL(2), Lecture Notes in Mathematics, vol. 114, Springer-Verlag, Berlin-New York, 1970.
[31] A.W. Knapp, Representation Theory of Semisimple Groups, Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, 2001, reprint of the 1986 original.
[32] I. Khayutin, R.S. Steiner, Theta functions, fourth moments of eigenforms, and the sup-norm problem I, arXiv e-prints, arXiv:2009.07194, 2020.
[33] H. Lokvenec-Guleska, Sum formula for SL_2 over imaginary quadratic number fields, Ph.D. thesis, Utrecht University, 2004.
[34] P. Maga, Subconvexity and shifted convolution sums over number fields, Ph.D. thesis, Central European University, Budapest, 2013.
[35] S. Marshall, Upper bounds for Maass forms on semisimple groups, arXiv e-prints, arXiv:1405.7033, 2014.
[36] D. Milićević, Large values of eigenfunctions on arithmetic hyperbolic 3-manifolds, Geom. Funct. Anal. 21 (6) (2011) 1375–1418.
[37] A. Popa, Whittaker newforms for Archimedean representations, J. Number Theory 128 (6) (2008) 1637–1645.
[38] P. Ramacher, The equivariant spectral function of an invariant elliptic operator. L^p-bounds, caustics, and concentration of eigenfunctions, J. Math. Pures Appl. (9) 120 (2018) 33–90.
[39] Z. Rudnick, On the asymptotic distribution of zeros of modular forms, Int. Math. Res. Not. (34) (2005) 2059–2074.

[40] P. Ramacher, S. Wakatsuki, Subconvex bounds for Hecke-Maass forms on compact arithmetic quotients of semisimple Lie groups, Math. Z. 298 (3–4) (2021) 1383–1424.
[41] A. Saha, Hybrid sup-norm bounds for Maass newforms of powerful level, Algebra Number Theory 11 (5) (2017) 1009–1045.
[42] P. Sarnak, Letter to Morawetz, 2004, https://publications.ias.edu/sarnak/.
[43] B. Simon, Real Analysis: A Comprehensive Course in Analysis, Part 1, Amer. Math. Soc., Providence, RI, 2015.
[44] E.M. Stein, Analysis in matrix spaces and some new representations of SL(N, C), Ann. Math. (2) 86 (1967) 461–490.
[45] R.S. Steiner, Sup-norm of Hecke-Laplace eigenforms on S^3, Math. Ann. 377 (1–2) (2020) 543–553.
[46] M. Tadić, An external approach to unitary representations, Bull. Am. Math. Soc. (N.S.) 28 (2) (1993) 215–252.
[47] J. Tate, Number theoretic background, in: Automorphic Forms, Representations and L-Functions, Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977, Part 2, in: Proc. Sympos. Pure Math., vol. XXXIII, Amer. Math. Soc., Providence, RI, 1979, pp. 3–26.
[48] N. Templier, Hybrid sup-norm bounds for Hecke-Maass cusp forms, J. Eur. Math. Soc. 17 (8) (2015) 2069–2082.
[49] D.A. Vogan Jr., The unitary dual of GL(n) over an Archimedean field, Invent. Math. 83 (3) (1986) 449–505.
[50] A.B. Wang, An analogue of the Paley-Wiener theorem for certain function spaces on SL(2, **C**), Pac. J. Math. 52 (1974) 617–629.
[51] G. Warner, Harmonic Analysis on Semi-Simple Lie Groups. I, Die Grundlehren der mathematischen Wissenschaften, Band 188, Springer-Verlag, New York-Heidelberg, 1972.
[52] G. Warner, Harmonic Analysis on Semi-Simple Lie Groups. II, Die Grundlehren der mathematischen Wissenschaften, Band 189, Springer-Verlag, New York-Heidelberg, 1972.
[53] A. Weil, Basic Number Theory, third ed., Die Grundlehren der mathematischen Wissenschaften, Band 144, Springer-Verlag, New York-Berlin, 1974.