HYPERBOLIC LATTICE POINT COUNTING IN UNBOUNDED RANK

VALENTIN BLOMER AND CHRISTOPHER LUTSKO

Abstract. We use spectral analysis to give an asymptotic formula for the number of matrices in $\mathrm{SL}(n,\mathbb{Z})$ of height at most T with strong error terms, far beyond the previous known, both for small and large rank.

1. Introduction

1.1. **The main result.** Counting lattice points in specified regions dates back at least to Gauß who gave an asymptotic formula for the number of integer points inside a circle of large radius R:

$$\#\{(x_1,x_2)\in\mathbb{Z}^2 \mid x_1^2+x_2^2\le R^2\} = \pi R^2 + O(R).$$

The error term can be interpreted as the length of the circumference. Following Davenport, this method is now called the Lipschitz principle. By basic harmonic analysis, the error can be improved to $O(R^{2/3})$ (see [Sz], or [IK, Corollary 4.9] for a modern treatment).

This question is equally interesting in hyperbolic geometry. The prototypical result goes back to Selberg. Let $\|.\|$ denote the Frobenius norm on $\mathrm{SL}_n(\mathbb{R})$, i.e. $\|g\|^2=\mathrm{tr}(g^\top g)$. Then

$$\#\{\gamma\in\mathrm{SL}_2(\mathbb{Z}) \mid \|\gamma\|\le T\} = 6T^2 + O(T^{4/3}). \tag{1}$$

While considered classical nowadays, this is nevertheless a difficult result which has never been improved. Selberg's proof was never published. A modern version can be found in [Iw, Section 12][1], where the highly non-trivial estimation of the spherical transform is left as an exercise. A long and very different proof of a more general (and only marginally weaker) result can be found in [LP].

The aim of this paper is a hyperbolic lattice point count in arbitrary rank. For $z,w\in\mathrm{SL}_n(\mathbb{R})$ let

$$\mathcal{N}_n(T;z,w) := \#\{\gamma\in\mathrm{SL}_n(\mathbb{Z}) \mid \|z^{-1}\gamma w\|\le T\}.$$

This counts the number of lattice points in the orbit $\mathrm{SL}_n(\mathbb{Z})w$ in a ball of radius T about z. Define

$$c_n = \frac{\pi^{n^2/2}}{\Gamma(\frac{n^2-n+2}{2})\Gamma(\frac{n}{2})\zeta(2)\cdots\zeta(n)}. \tag{2}$$

Theorem 1. *For $T\ge 1$, $\varepsilon>0$, $n\ge 3$ and $z,w\in\mathrm{SL}_n(\mathbb{R})$ we have*

$$\mathcal{N}_n(T;z,w) = c_n T^{n(n-1)} + O_{z,w,\varepsilon,n}(T^{n(n-1)-\delta_n+\varepsilon})$$

2010 *Mathematics Subject Classification.* Primary 11P21, 11N45, 11F72, 22E30.

Key words and phrases. lattice points, spherical functions, pretrace formula, Eisenstein series.

First author supported by DFG through SFB-TRR 358 and EXC-2047/1 - 390685813 and by ERC Advanced Grant 101054336. The second author would like to thank the University of Bonn for hosting him in the summer 2023.

[1] whose proof is written for $\mathrm{PSL}_2(\mathbb{Z})$ and hence differs by a factor 2 in the main term

where

$$\delta_3 = 1, \quad \delta_4 = 6/5, \quad \delta_n = 1 + \frac{1}{\sqrt{2}} + O\Big(\frac{1}{n}\Big).$$

The proof works without substantial modifications also for congruence subgroups of $\mathrm{SL}_n(\mathbb{Z})$. As we will argue below, the numerical values for δ_3 and δ_4 are most likely not improveable by spectral techniques and should be seen as the appropriate generalization of Selberg's bound (1).

The first result in this direction was proved by Duke-Rudnick-Sarnak [DRS, Theorem 1.10] and reads

$$\delta_n^{\mathrm{DRS}} = \frac{1}{n+1}.$$

This was improved only about 25 year later [GNY, Theorem 2] to

$$\delta_n^{\mathrm{GNY}} = \frac{2(n-1)}{(n+1)(n+\eta)}, \quad \eta = \begin{cases} 0, & n \text{ even}, \\ 1, & n \text{ odd}. \end{cases} \tag{3}$$

which is asymptotically roughly twice as good. Neither of these bounds recovers (1) for $n = 2$. For comparison with Theorem 1, the corresponding savings in (3) are

$$\delta_3^{\mathrm{GNY}} = 1/4, \quad \delta_4^{\mathrm{GNY}} = 3/10, \quad \delta_n^{\mathrm{GNY}} = O(1/n).$$

In contrast, our exponent in Theorem 1 is uniformly bounded from below. We can slightly improve the asymptotic performance on average over z.

Theorem 2. *Let $\Omega \subseteq \mathrm{SL}_n(\mathbb{R})$ be a compact set. For $T \geq 1$, $\varepsilon > 0$ and $n \geq 3$ we have*

$$\int \mathcal{N}_n(T; z, z)dz = c_n \mathrm{vol}(\Omega) T^{n(n-1)} + O_{\Omega,\varepsilon,n}(T^{n(n-1)-\delta_n+\varepsilon})$$

where $\delta_n = 2 + O(1/n)$.

As we will see below, except for the $O(1/n)$ term this is probably very hard to improve by spectral techniques.

1.2. **The methods.** In the situation of the classical Gauß circle problem, the strategy is well-known: after a bit of smoothing one applies the Poisson summation formula. The zero frequency yields the main term, the remaining terms are estimated sharply in absolute value. One then optimizes the smoothing parameter to obtain the desired asymptotic formula. Any improvement requires cancellation between the non-zero frequencies (which is possible to some extent in this particular situation using exponential sum techniques).

In the hyperbolic case, one proceeds similarly. After a bit of smoothing, one applies harmonic analysis in the form of the pretrace formula. The main difference in the non-commutative set-up is that the spectral side looks very different than the geometric side. A sharp estimate of the spherical transform yields (1). Any improvement would require cancellation in the spectral sum over eigenvalues of the hyperbolic Laplacian which is not available with present methods. Consequently, Selberg's error term has never been improved. Note that a Lipschitz principle is not applicable here since the circumference of a large hyperbolic circle has the same order of magnitude as its area.

In order to put Theorem 1 into perspective, let us first get a feeling for what we can possibly hope for and what might us prevent from obtaining this. The higher rank case offers two major difficulties:

a) a sharp estimate for the spherical transform of the characteristic function on a ball or a smoothed version thereof. This is a problem in the analysis on Lie groups which we solve in a best possible way in this paper; cf. Section 3.

b) an understanding of the non-tempered spectrum. This is deeper than it might sound. First of all, in higher rank there could be infinitely many linearly independent cusp forms violating the Ramanujan conjecture. More seriously, residual Eisenstein series are known to violate the Ramanujan conjecture, some of them quite drastically, and will therefore contribute a large error term. One could hope that such Eisenstein series make up only a small portion of the spectrum which compensates their degree of non-temperedness. However, an analysis of the pretrace formula also asks for pointwise bound of all appearing spectral components, and in higher rank we have very little control on the sup-norm of automorphic forms, regardless of whether they are cuspidal or Eisenstein. The best we can do in general is to use the pretrace formula backwards and estimate the sup-norm by the square-root of the spectral density. In this way, however, we sacrifice a large portion of our knowledge on the sparsity of such Eisenstein series. The reader may notice from the proof that on the other hand two features are working in our favor. On the one hand residual Eisenstein series lie on many Weyl chamber walls which reduces the spectral density (somewhat). On the other hand, the spherical transform of the characteristic function of the set of matrices with norm $\leq T$ behaves (slightly) better on the non-tempered spectrum.

Let us make the preceding remarks a bit more quantitative. We smooth out the characteristic function on $[0,T]$ to a function with support in $[0,T(1+\delta)]$ where δ will be chosen later as a function of T. This introduces an error in the main term of $O(T^{n(n-1)}\delta)$. We will show in Section 3 below that the spherical transform of such a function decays roughly like

$$\frac{T^{\frac{1}{2}n(n-1)+n\|\Re\mu\|}}{\|\mu\|^{\frac{1}{4}n(n+1)}}(1+\delta\|\mu\|)^{-A}$$

for any $A>0$ and a spectral parameter $\mu=(\mu_1,\dots,\mu_n)$ which we normalize so that it is tempered if all entries are purely imaginary. Here $\|.\|$ denotes the maximum norm of a vector. This bound is an oversimplification and holds only if μ is in generic position, i.e. away from the Weyl chamber walls, but let us proceed anyway. Recall that by Weyl's law [Mu] there are $O(\delta^{1-n(n+1)/2})$ linearly independent cusp forms with $\mu \ll \delta^{-1}$, the point after which the spherical transform becomes negligible.

If we ignore the non-tempered spectrum as well as the problem of pointwise bounds for automorphic functions, we obtain a total error term

$$\ll T^{n(n-1)}\delta + T^{\frac{1}{2}n(n-1)}\delta^{1-\frac{1}{2}n(n+1)+\frac{1}{4}n(n+1)} \ll T^{n(n-1)-2\frac{n-1}{n+1}} \tag{4}$$

upon choosing $\delta = T^{-2(n-1)/(n+1)}$. This is the most optimistic generalization of Selberg's argument to higher rank and recovers (1) for $n=2$. Since even for $n=2$ this has been the state of the art for more than half a century, it seems essentially impossible to go beyond such a bound. We emphasize that unlike (3), this suggests a saving of *asymptotically constant* size.

This is precisely what we achieve in Theorem 1 with a constant $1+1/\sqrt{2}\approx 1.7$. The cases $n=3$ and $n=4$ match our (rather optimistic) heuristic (4). In Theorem 2 we reach asymptotically the "best-possible" constant 2 on average over z. (For $n=2$, the paper [PR] shows that a rather sophisticated additional application of the Kuznetsov formula can

exploit the average a bit more strongly, but this seems very hard to implement in higher rank.)

We remark that our main results are somewhat similar in spirit and quality (relative to previous results) as the recent paper [JK2], however with the major difference that Theorem 1 is a pointwise result, valid for all (fixed) z, w, whereas [JK2] holds for almost all points.

1.3. **Acknowledgments.** We thank Peter Sarnak for the discussion which led to this project, and Alex Kontorovich for insightful discussions.

2. Preliminaries

2.1. **Some notation.** We write $G = \mathrm{SL}_n(\mathbb{R})$, $\Gamma = \mathrm{SL}_n(\mathbb{Z})$, $K = \mathrm{SO}(n)$, $A \subseteq \mathrm{SL}_n(\mathbb{R})$ for the group of diagonal matrices with positive entries and determinant 1, N for the group unipotent upper triangular matrices, W for the Weyl group. We decompose $G = KAN = NAK$. Accordingly, we have the Iwasawa projections[2] $A, H : G \to \mathfrak{a}$ such that $g \in K\exp(H(g))N = N\exp(A(g))K$. We denote by dn, da, dk, dg the usual Haar measures, normalized as in [DRS, Appendix A][3], in particular $\mathrm{vol}(K) = 1$. As usual, we write $\rho = (\frac{n-1}{2}, \frac{n-3}{2}, \dots, \frac{3-n}{2}, \frac{1-n}{2}) \in \mathfrak{a}^*$ and C_ρ for the convex hull of the points $\{w\rho \mid w \in W\}$. We often identify $\mathfrak{a}_{\mathbb{C}}^* \cong \{\mu \in \mathbb{C}^n \mid \sum \mu_j = 0\}$ and equip vectors in $\mathfrak{a}_{\mathbb{C}}^*$ with the max-norm $\|.\|$. We recall that all relevant spectral parameters $\mu \in \mathfrak{a}_{\mathbb{C}}^*$ satisfy

$$\sum_{j=1}^{n} \mu_j = 0, \quad \{\mu_1, \dots, \mu_n\} = \{-\bar{\mu}_1, \dots, -\bar{\mu}_n\}, \quad \mu \in i\mathfrak{a}^* + C_\rho. \tag{5}$$

In the following we regard $n \geq 2$ as fixed, and all implied constants may depend on n.

2.2. **Spherical transform and pretrace formula.** For $\mu \in \mathfrak{a}_{\mathbb{C}}^*$ we define the spherical function (cf. [He, p. 418 & p. 435] with μ in place of $i\lambda$)

$$\phi_\mu(g) = \int_K e^{(-\rho+\mu)H(gk)}dk = \int_K e^{(\rho+\mu)A(kg)}dk.$$

We have

$$\phi_\rho(g) = 1, \quad |\phi_\mu(g)| \leq 1 \tag{6}$$

for all $g \in G$ and μ satisfying (5). For a (measurable) compactly supported, bi-K-invariant function $f : G \to \mathbb{C}$ we define the spherical transform (cf. [He, p. 449])

$$\tilde{f}(\mu) = \int_G f(g)\phi_{-\mu}(g)dg.$$

It is well-known ([He, p. 450]) that this is the composition of the Abel transform

$$\mathcal{A}_f(a) = e^{\rho(\log a)} \int_N f(an)dn, \quad a \in A,$$

and the Fourier transform

$$\tilde{f}(\mu) = \int_A \mathcal{A}_f(a)e^{-\mu \log a}da. \tag{7}$$

[2]The double use of A will not lead to confusion.

[3]but the particular normalization plays no role for the purpose of this paper

If f_1, f_2 are two (measurable) compactly supported, bi-K-invariant functions, the convolution

$$(f_1 * f_2)(x) = \int_G f_1(xg^{-1}) f_2(g) dg$$

satisfies [He, p. 454]

$$\widetilde{f_1 * f_2} = \tilde{f}_1 \tilde{f}_2. \tag{8}$$

The spherical transform comes up in the pretrace formula [Se]: for a smooth, compactly supported, bi-K-invariant function $f : G \to \mathbb{C}$ and $z, w \in \Gamma\backslash G/K$ the spectral expansion of the automorphic kernel on the left hand side (cf. [Se, (2.5)]) of the following display together with the uniqueness principle (cf. [Se, (1.8)]) is

$$\sum_{\gamma \in \Gamma} f(z^{-1}\gamma w) = \int \tilde{f}(\mu_\varpi) \varpi(z) \overline{\varpi(w)} d\varpi. \tag{9}$$

The notation should be interpreted as follows: the right hand side runs over cusp forms and Eisenstein series (including residual Eisenstein series) for the group Γ, and $d\varpi$ is the counting measure on the discrete spectrum. Each spectral parameter $\mu_\varpi \in \mathfrak{a}_{\mathbb{C}}^*$ of ϖ is only defined modulo the action of the Weyl group W. We parameterize the spectrum in detail in Subsection 2.5.

2.3. **Sup-norm bounds.** We apply the pretrace formula in the other direction to obtain the following (generic) bounds for automorphic forms appearing on the right hand side of (9).

Lemma 1. *Let $\mu \in i\mathfrak{a}^*$, $B(\mu) \subseteq \mathfrak{a}_{\mathbb{C}}^*$ a ball of size $O(1)$ about μ and $z \in G$. Then we have*

$$\int_{B(\mu)} |\varpi(z)|^2 d\varpi \ll_z \prod_{1 \le i < j \le n} (1 + |\mu_i - \mu_j|).$$

Proof. This is well-known and implicit for instance in [BM]. For convenience we recall the proof. Let f be a fixed function on $\mathfrak{a}_{\mathbb{C}}^*$ with compactly supported Fourier transform such that f is real on $i\mathfrak{a}^*$, $\Re f$ is non-negative in the strip $\{\lambda \in \mathfrak{a}_{\mathbb{C}}^* : |\Re \lambda_j| \le \|\rho\|\}$, and $\Re f \ge 1$ on a ball in $\mathfrak{a}_{\mathbb{C}}^*$ about 0 of radius $\|\rho\|_2$. Such a function was constructed explicitly in [BM, Lemma 1] and the subsequent display. Then we choose

$$\tilde{f}_\mu(\lambda) := \Big(\sum_{w \in W} f(\mu - w \cdot \lambda) \Big)^2.$$

This again has compactly supported Fourier transform, and the support is independent of μ. One verifies quickly that

$$\tilde{f}_\mu(\lambda) \ge 0$$

for all λ satisfying (5) and

$$\tilde{f}_\mu(\mu) \ge 1.$$

Moreover, the rapid decay along the real axis shows

$$\tilde{f}_\mu(\lambda) \ll_A \max_{w \in W} (1 + \|\Im \mu - w \cdot \lambda\|)^{-A}$$

for $\lambda \in i\mathfrak{a}^*$ and any $A > 0$. By the Harish-Chandra inversion formula featuring the Harish-Chandra $\mathbf{c}$-function together with the trivial bound (6) for elementary spherical functions,

we see that the inverse spherical transform f_μ of $\tilde{f}_\mu$ has compact support and satisfies the bound

$$f_\mu(g) \ll \prod_{1\le i<j\le n}(1+|\mu_i-\mu_j|).$$

We now conclude from (9) that

$$\int_{B(\mu)}|\varpi(z)|^2 d\varpi \ll \int_{B(\mu)}|\varpi(z)|^2\tilde{f}_\mu(\mu_\varpi)d\varpi = \sum_{\gamma\in\Gamma} f_\mu(z^{-1}\gamma z) \ll \prod_{1\le i<j\le n}(1+|\mu_i-\mu_j|)$$

as desired.

In Subsection 2.5 we give a better bound on average over z, taken from [JK1].

2.4. **Smoothing.** For $T>1$ we write $\chi_T : K\backslash G/K \to \mathbb{R}_{\ge 0}$ for the characteristic function on $\|g\|\le T$. For $0<\delta<1$ and let $\psi_\delta : K\backslash G/K\to\mathbb{R}_{\ge 0}$ be a smooth $L^1(AN)$-normalized function supported in $B_\delta := \{g\in G \mid \max(\|g\|_2, \|g^{-1}\|_2)\le 1+\delta\}$ where $\|.\|_2$ is the matrix norm induced from the Euclidean vector norm. Let

$$\chi_{T,\delta} := \chi_T * \psi_\delta. \tag{10}$$

By [DRS, Lemma 3.3] we have for $hg^{-1}\in\mathrm{supp}(\chi_T)$ and $g\in\mathrm{supp}(\psi_\delta)$ the inequalities

$$T\ge \|hg^{-1}\| \ge \frac{\|hg^{-1}g\|}{\|g\|_2}\ge \frac{\|h\|}{1+\delta}.$$

On the other hand we have for $\|h\|\le T(1+\delta)^{-1}$ and $g\in\mathrm{supp}(\psi_\delta)$ that

$$\|hg^{-1}\|\le \|h\|\|g^{-1}\|_2\le T.$$

These two inequalities imply $\chi_{T(1+\delta)^{-1}}\le\chi_{T,\delta}\le\chi_{T(1+\delta)}$ and hence

$$\chi_{T(1+\delta)^{-1},\delta}\le\chi_T\le\chi_{T(1+\delta),\delta}. \tag{11}$$

2.5. **Parametrization of the spectrum.** We describe the various types of Eisenstein series in classical language. See [MW1, MW2] for a detailed description in representation theoretic terms with full proofs and [GH, Chapter 10] for a concise summary. We start with a partition

$$n = d_1+d_2+\ldots+d_r$$

of GL_n into blocks of dimension $d_j\ge 1$. For each j we choose a divisor $f_j\mid d_j$ and in the case $f_j\ge 2$ a cusp form u_j for the group $\mathrm{SL}_{f_j}(\mathbb{Z})$ with spectral parameter $\mu_j\in\mathbb{C}^{f_j}$ (satisfying (5) with f_j in place of n). In addition we choose r imaginary numbers $s_1,\ldots s_r\in i\mathbb{R}$ satisfying $\sum_j d_j s_j = 0$. We call the set of such Eisenstein series $E(d,f)$; it consists of all (discrete) choices of such $u_1,\ldots,u_r$ (for $f_j\ge 2$) and such numbers $s_1,\ldots,s_r$ (that vary continuously). The case of cusp forms corresponds to $r=1$, $d_1=f_1=n$. In representation theoretic terms, if π is a cuspidal automorphic representation on $\mathrm{GL}_f(\mathbb{A})$ corresponding to the cusp form u, this corresponds to the Speh representation $\mathrm{Speh}(\pi, d/f)$ of $\mathrm{GL}_d(\mathbb{A})$ which is the unique irreducible subrepresentation of

$$\mathrm{Ind}_{P_{d/f}(\mathbb{A})}^{\mathrm{GL}_d(\mathbb{A})}\big(|\cdot|_{\mathbb{A}}^{\frac{d/f-1}{2}}\pi\otimes\ldots\otimes|\cdot|_{\mathbb{A}}^{-\frac{d/f-1}{2}}\pi\big) \tag{12}$$

where $P_{d/f}$ is the standard parabolic subgroup associated to the partition $f+\ldots+f=d$.

Example: Let $n = 3$. If $r = 1$ and $d_1 = 3$, we have either $f_1 = 3$ in which case we get cusp forms for $\mathrm{SL}_3(\mathbb{Z})$, or $f_1 = 1$ in which case we get the constant function. If $r = 2$ with $d_1 = 2$ and $d_2 = 1$, we have either $f_1 = 2$ in which case we get maximal Eisenstein series with a cusp form u_1 for $\mathrm{SL}_2(\mathbb{Z})$, or $f_1 = 1$ in which case we get the maximal degenerate Eisenstein series (Epstein zeta function). If $r = 3$ with $d_j = f_j = 1$, we get minimal Eisenstein series.

From (12) we see that the spectral parameter of an element $\varpi \in E(d, f)$ is

$$\Big(\underbrace{\mu_1 + s_1 + \tfrac{1}{2}(\tfrac{d_1}{f_1} - 1)}_{\in \mathbb{C}^{f_1}}, \underbrace{\mu_1 + s_1 + \tfrac{1}{2}(\tfrac{d_1}{f_1} - 3)}_{\in \mathbb{C}^{f_1}}, \dots, \underbrace{\mu_1 + s_1 + \tfrac{1}{2}(1 - \tfrac{d_1}{f_1})}_{\in \mathbb{C}^{f_1}},$$
$$\dots,$$
$$\underbrace{\mu_r + s_r + \tfrac{1}{2}(\tfrac{d_r}{f_r} - 1)}_{\in \mathbb{C}^{f_r}}, \underbrace{\mu_r + s_r + \tfrac{1}{2}(\tfrac{d_r}{f_r} - 3)}_{\in \mathbb{C}^{f_r}}, \dots, \underbrace{\mu_r + s_r + \tfrac{1}{2}(1 - \tfrac{d_r}{f_r})}_{\in \mathbb{C}^{f_r}}\Big)$$

with $\mu_j = 0$ if $f_j = 1$. For each cusp form u_j we use the Jacquet-Shalika bounds to bound $\|\Re \mu_j\| \le 1/2$, so that for $\varpi \in E(d, f)$ we have

$$\|\Re \mu_\varpi\| \le \max_{1 \le j \le r} \Big(\frac{1}{2}\Big(\frac{d_j}{f_j} - 1\Big) + \delta_{f_j \ge 2}\frac{1}{2}\Big). \tag{13}$$

Next, a cusp form in each block of dimension d_j has at most f_j different entries in its spectral parameter, so out of the differences $\mu_{j,i} - \mu_{j,k}$ with $1 \le i < k \le d_j$ at least

$$f_j \cdot \frac{1}{2}\frac{d_j}{f_j}\Big(\frac{d_j}{f_j} - 1\Big) = \frac{d_j}{2}\Big(\frac{d_j}{f_j} - 1\Big)$$

unordered pairs coincide. Using Lemma 1, we conclude that

$$\int_{B(\mu) \cap E(d,f)} |\varpi(z)|^2 d\varpi \ll_z (1 + \|\mu\|)^{\frac{1}{2}n(n-1) - \frac{1}{2}\sum_{j=1}^r d_j(d_j/f_j - 1)} \tag{14}$$

for $\mu \in i\mathfrak{a}^*$ and $z \in G$. This is a reflection of the fact that degenerate Eisenstein series lie on many Weyl chamber walls and therefore have a (somewhat) smaller sup-norm. The bound is in general probably far from optimal; see [Bl] for sup-norm bounds for Eisenstein series in a very special case. Using [JK1, Theorem 1] in combination with a local (upper bound) Weyl law [Mu] for each of the blocks we can do better on average over z, namely

$$\int_\Omega \int_{B(\mu) \cap E(d,f)} |\varpi(z)|^2 d\varpi \, dz \ll_{\Omega,\varepsilon} (1 + \|\mu\|)^{\frac{1}{2}\sum_j f_j(f_j - 1) + \varepsilon} \tag{15}$$

for each compact $\Omega \subseteq \mathrm{SL}_n(\mathbb{R})$.

Finally, the set $\{\varpi \in E(d, f) : \|\mu_\varpi\| \le R\}$ can be covered by

$$\ll R^{\sum_{j=1}^r (f_j - 1) + (r-1)} = R^{-1 + \sum_{j=1}^r f_j} \tag{16}$$

balls $B(\mu)$.

3. Spherical transforms

Our goal is to understand the spherical transforms $\tilde{\chi}_T$ and $\tilde{\psi}_\delta$. By (6) we have

$$\tilde{\psi}_\delta(\rho) = 1, \quad \tilde{\chi}_T(\rho) = \int_{\|g\| \le T} dg = c_n \mathrm{vol}(\Gamma\backslash G) T^{n(n-1)} \tag{17}$$

where the constant c_n, defined in (2), is computed in [DRS, Appendix 1].

Lemma 2. *For $0 < \delta < 1$, $A > 0$ and $\Re\mu \ll 1$ we have*

$$\tilde{\psi}_\delta(\mu) \ll_A (1 + \delta\|\mu\|)^{-A}.$$

Proof. By (7) we have

$$\tilde{\psi}_\delta(\mu) = \int_A a^\rho \int_N \psi_\delta(an)\, dn\, a^{-\mu} da$$

where we recall that ψ_δ is L^1 supported in a ball of radius δ about the identity of the $(\frac{1}{2}n(n+1)-1)$–dimensional space AN, so $\|\psi_\delta\|_\infty \ll \delta^{1-n(n+1)/2}$. The N-integral vanishes unless $n \ll \delta$ and $a - \mathrm{id} \ll \delta$. This gives immediately the trivial bound $\tilde{\psi}_\delta(\mu) \ll 1$ for $\Re\mu \ll 1$. On the other hand, we can use an $(n-1)$-dimensional local coordinate system about the identity in A, so that the A-integral looks like

$$\int_{\mathbb{R}_{>0}^{n-1}} \Psi_\delta(y_1, \dots, y_{n-1}) y_1^{\mu_1} \cdots y_{n-1}^{\mu_{n-1}} (y_1 \cdots y_{n-1})^{-\mu_n} dy$$

where Ψ_δ is supported in $y_j = 1 + O(\delta)$ and $\mathcal{D}\Psi_\delta(y) \ll \delta^{-(n-1)-k}$ for any differential operator of degree k with constant coefficients. Integrating by parts k times with respect to y_j we obtain the bound $\tilde{\psi}_\delta(\mu) \ll (\delta|\mu_j - \mu_n|)^{-k}$. Choosing a different local coordinate system, we can replace n with any other index. This completes the proof.

For $a = (a_1 \dots, a_n) \in \mathbb{R}^n$ we define

$$G(a) = \prod_{j=1}^n (1 + (\max_i a_i) - a_j)^{-1/2} + \prod_{j=1}^n (1 + |(\min_i a_i) - a_j|)^{-1/2}.$$

Lemma 3. *Let $T \geq 1$, $\kappa, \varepsilon > 0$ and μ satisfying (5).*
There exists a constant $B \in \mathbb{R}$ depending only on n such that

$$\tilde{\chi}_T(\mu) \ll_\varepsilon T^{\frac{1}{2}n(n-1)+n\|\Re\mu\|+\varepsilon}(1 + \|\mu\|)^B.$$

If $\|\mu\| \ll T^{2-\kappa}$, then

$$\tilde{\chi}_T(\mu) \ll_{\kappa,\varepsilon} T^{\frac{1}{2}n(n-1)+n\|\Re\mu\|+\varepsilon} \frac{G(\Im\mu)}{(1+\|\mu\|)^{\frac{1}{4}n(n-1)+\frac{1}{2}(1+\|\Re\mu\|)}}.$$

Proof. Again we start with (7) and compute first

$$\int_N \chi_T(an) dn = \mathrm{meas}\Big(\Big\{ \sum_{j=1}^n a_i^2 + \sum_{1 \leq i < j \leq n} n_{ij}^2 a_i^2 \leq T^2 \mid n_{ij} \in \mathbb{R} \Big\}\Big).$$

We integrate successively each n_{ij} at a time using the formula [GR, 3.191.1]

$$\int_{-\sqrt{Z}}^{\sqrt{Z}} (Z - x^2 y^2)^\alpha dx = \frac{Z^{\frac{1}{2}+\alpha}}{y} \frac{\pi^{1/2}\Gamma(1+\alpha)}{\Gamma(\frac{3}{2}+\alpha)}$$

for $\alpha, Z, y \geq 0$. In this way we obtain

$$\int_N \chi_T(an) dn = \gamma_n \delta_{\|a\|_2 \leq T} \frac{(T^2 - \|a\|_2^2)^{n(n-1)/4}}{a_1^{n-1} a_2^{n-2} \cdots a_{n-1}}, \quad \gamma_n = \frac{\pi^{n(n-1)/4}}{\Gamma(1 + \frac{1}{4}n(n-1))},$$

so that

$$\tilde{\chi}_T(\mu) = \gamma_n T^{\frac{1}{2}n(n-1)} \int_{\|a\|_2 \le T} \Big(1 - \frac{\|a\|_2}{T}\Big)^{n(n-1)/4} \prod_j a_j^{-\mu_j - \frac{n-1}{2}} da = \gamma_n \int_A f_\mu\Big(\frac{a}{T}\Big) da$$

with

$$f_\mu(a) = \delta_{\|a\| \le 1} (1 - \|a\|_2^2)^{\frac{1}{4}n(n-1)} \prod_j a_j^{-\mu_j - \frac{n-1}{2}}$$

(using that $\sum_j \mu_j = 0$). Following [DRS], it is most convenient to estimate this integral asymptotically by passing to the torus in GL_n^+. Define

$$F_\mu(s) = \gamma_n \int_{\mathbb{R}^n_{>0}} f_\mu(y) (\det y)^s \frac{dy_1}{y_1} \cdots \frac{dy_n}{y_n}.$$

Then by Mellin inversion we have

$$\tilde{\chi}_T(\mu) = \gamma_n \int_{(c)} F_\mu(s) T^{ns} \frac{ds}{2\pi i}$$

for some sufficiently large $c > 0$. We compute

$$F_\mu(s) = \gamma_n \frac{\Gamma(1 + \frac{n(n-1)}{4})}{2^n \Gamma(\frac{n}{2}s + 1)} \prod_{j=1}^n \Gamma\Big(\frac{s}{2} - \frac{\mu_j}{2} - \frac{n-1}{4}\Big)$$

using again [GR, 3.191.1] and the fact that $\sum_j \mu_j = 0$ (this condition on μ will be used frequently in following arguments). We conclude that

$$\tilde{\chi}_T(\mu) = \frac{\pi^{n(n-1)/4}}{2^n} T^{n(n-1)/2} \int_{(c)} \frac{T^{ns}}{\Gamma(\frac{n}{2}s + \frac{n(n-1)}{4} + 1)} \prod_{j=1}^n \Gamma\Big(\frac{s - \mu_j}{2}\Big) \frac{ds}{2\pi i}$$

for sufficiently large $c > 0$. Estimating this integral is an elaborate exercise in Stirling's formula. Let us write $\mu_j = m_j + i\tau_j$, $s = \sigma + it$ and assume without loss of generality $\tau_1 \ge \tau_2 \ge \ldots \ge \tau_n$. Since $\sum \tau_j = 0$, we have $\tau_1 \asymp -\tau_n \asymp \|\tau\|$. The exponential behavior of the integrand is given by

$$\exp\Big(-\frac{\pi}{4}\sum_{j=1}^n |t - \tau_j| + \frac{\pi n}{4}|t|\Big) \le \begin{cases} \exp(-\frac{\pi}{2}\min(|\tau_1 - t|, |t - \tau_n|)), & \tau_n \le t \le \tau_1, \\ 1, & \text{else.} \end{cases}$$

In particular, there is no exponential increase, but exponential decrease for $t \in [\tau_n, \tau_1]$. The polynomial behavior of the gamma quotient is

$$\ll_\sigma (1 + |t|)^{-\frac{1}{2}n\sigma - \frac{1}{4}n(n-1) - \frac{1}{2}} \prod_{j=1}^n (1 + |t - \tau_j|)^{(\sigma - m_j - 1)/2}$$

away from poles. For $|t| \ge 1 + 2\|\tau\|$ this is $\ll_\sigma |t|^{-\frac{1}{4}(n^2+n+2)}$, in particular, the integrand is absolutely integrable on every vertical line (not crossing poles). Moreover, for $\sigma \le -n$, $t \ll \|\tau\|$, the gamma quotient is very coarsely bounded by $(1 + \|\tau\|)^{\frac{1}{2}n|\sigma|}$. In particular, for $\|\mu\| \le T^{2-\kappa}$, the integral over a line sufficiently far to the left becomes negligible. We use these considerations in the following estimations.

The right-most pole appears[4] at $s = \|\Re\mu\|$. Shifting the contour to $c = \|\Re\mu\| + \varepsilon$ and estimating trivially, we obtain immediately the first part of the lemma.

Suppose from now on $\|\mu\| \ll T^{2-\kappa}$. We the shift the contour to the far left and pick up the residues. For notational simplicity let us first assume that the μ_j are pairwise distinct. The general case follows by a straightforward limit procedure. The residues in $\Re s \geq -K$ are given by

$$2\sum_{j=1}^{n}\sum_{0\leq k\leq (K+\Re\mu_j)/2}\frac{(-1)^k}{k!}\frac{T^{n\mu_j-2nk}}{\Gamma(\frac{n}{2}\mu_j+\frac{n(n-1)}{4}+1-nk)}\prod_{i\neq j}\Gamma\Big(\frac{\mu_j-\mu_i}{2}-k\Big).$$

By the same computation as above, each summand is

(18)
$$\ll \frac{T^{nm_j-2nk}}{(1+|\tau_j|)^{\frac{n}{2}m_j+\frac{1}{4}n(n-1)+\frac{1}{2}-nk}}\exp\Big(-\frac{\pi}{2}\min(|\tau_1-\tau_j|,|\tau_j-\tau_n|)\Big)\prod_{i=1}^{n}(1+|\tau_j-\tau_i|)^{\frac{1}{2}(m_j-m_i-1-2k)}.$$

For $\|\tau\| \ll T^{2-\kappa}$ this is decreasing in k, so it suffices to consider the term $k = 0$. The expression is also increasing in m_j, so we can and will assume $m_j = \|m\|$. From the exponential term we can assume that $|\tau_j| \asymp \|\tau\|$ in which case

$$\frac{\prod_{i=1}^{n}(1+|\tau_j-\tau_i|)^{\frac{1}{2}(m_j-m_i)}}{(1+|\tau_j|)^{\frac{n}{2}m_j}} \ll \frac{1}{(1+|\tau_j|)^{\frac{1}{2}m_j}}\frac{(1+|\tau_j|)^{\frac{1}{2}(n-1)m_j-\frac{1}{2}\sum m_i}}{(1+|\tau_j|)^{\frac{n-1}{2}m_j}} = \frac{1}{(1+|\tau_j|)^{\frac{1}{2}m_j}}.$$

We therefore bound (18) by

$$\ll \frac{T^{n\|m\|}}{(1+|\tau_j|)^{\frac{1}{4}n(n-1)+\frac{1}{2}(1+\|m\|)}}\exp\Big(-\frac{\pi}{2}\min(|\tau_1-\tau_j|,|\tau_j-\tau_n|)\Big)\prod_{i=1}^{n}(1+|\tau_j-\tau_i|)^{-1/2}$$

which gives the desired bound. This completes the proof.

Combining (8), Lemma 2 and Lemma 3, we conclude

Corollary 4. *Define $\chi_{T,\delta}$ as in (10). For $n \geq 2$, $\varepsilon, \kappa, A > 0$,*

$$T^{-2+\kappa} \ll \delta < 1 \leq T$$

and μ satisfying (5) we have

$$\tilde{\chi}_{T,\delta}(\mu) \ll_{\varepsilon,n} T^{\frac{1}{2}n(n-1)+n\|\Re\mu\|+\varepsilon}\frac{G(\Im\mu)(1+\delta\|\mu\|)^{-A}}{(1+\|\mu\|)^{\frac{1}{4}n(n-1)+\frac{1}{2}(1+\|\Re\mu\|)}}.$$

4. Lattice point count

For $0 < \delta < 1 \leq T$ we define

$$\mathcal{N}_{n,\delta}(T; z, w) := \sum_{\gamma\in\Gamma}\chi_{T,\delta}(z^{-1}\gamma w).$$

From (11) and (9) we conclude

$$\mathcal{N}_n(T; z, w) \leq \mathcal{N}_{n,\delta}(T(1+\delta); z, w) = \int \tilde{\chi}_{T(1+\delta),\delta}(\mu_\varpi)\varpi(z)\overline{\varpi(w)}d\varpi.$$

[4]Note that for $\mu = \rho$ the residue of the right-most pole at $s = (n-1)/2$ yields exactly the asymptotic [DRS, (A1.15)] as it should

From the right hand side we extract the L^2-normalized constant function corresponding to $\mu_\varpi = \rho$. By (17) this contributes

$$c_n(T(1+\delta))^{n(n-1)} = c_n T^{n(n-1)} + O(T^{n(n-1)}\delta).$$

Similarly we obtain a lower bound and hence conclude the basic asymptotic

$$\mathcal{N}_n(T; z, w) = c_n T^{n(n-1)} + O\Big(T^{n(n-1)}\delta + \int_{\mu_\varpi \neq \rho} |\tilde{\chi}_{T(1\pm\delta),\delta}(\mu_\varpi)|(|\varpi(z)|^2 + |\varpi(w)|^2)d\varpi\Big).$$

We need to estimate the second term.

4.1. **The general argument for $n \geq 5$.** We partition the spectrum into parameters (d, f) as in Section 2.5, excluding the case $r = 1$, $d_1 = n$, $f_1 = 1$, which corresponds to the constant function. We assume that $|\log \delta| \asymp \log T$ and specifically $\delta = T^{-\alpha}$ for $0 < \alpha < 2$ (in order to apply Corollary 4).

Combining Corollary 4 (where we drop the factor $G(\mu)$ for simplicity and also simplify the denominator a bit), (13), (14) and (16), we have

$$\begin{aligned}\int_{E(d,f)} (...) &\ll_{z,w} T^{\frac{1}{2}n(n-1)+n\max_j(\frac{1}{2}(\frac{d_j}{f_j}-1)+\delta_{f_j\geq 2}\frac{1}{2})+\varepsilon} \\ &\times \Big(1 + \delta^{\frac{1}{4}n(n-1)+\frac{1}{2}-\frac{1}{2}n(n-1)+\frac{1}{2}\sum_j d_j(d_j/f_j-1)+1-\sum_j f_j}\Big) \\ &= T^{\frac{1}{2}n(n-1)+\frac{n}{2}\max_j(\frac{d_j}{f_j}-\delta_{f_j=1})+\varepsilon}\Big(1 + \delta^{-\frac{1}{4}n(n+1)+\frac{3}{2}+\frac{1}{2}\sum_j(d_j^2/f_j-2f_j)}\Big).\end{aligned} \tag{19}$$

Suppose without loss of generality that $j = 1$ is the index at which the maximum in the exponent is attained. Clearly for all other indices the worst case is $f_j = d_j$, so we are left with analyzing

$$T^{\frac{1}{2}n(n-1)+\frac{n}{2}(\frac{d_1}{f_1}-\delta_{f_1=1})+\varepsilon}\Big(1 + \delta^{-\frac{1}{4}n(n+1)+\frac{3}{2}+\frac{1}{2}(\frac{d_1^2}{f_1}-2f_1-\sum_{j\geq 2} d_j)}\Big). \tag{20}$$

Before we optimize f_1, we treat by hand the case $d_1 = n-1$, $f_1 = 1$ (so that $r = 2$, $d_2 = 1$), in which case the preceding display becomes

$$T^{n(n-1)-\frac{n}{2}+\varepsilon}\Big(1 + \delta^{\frac{1}{4}(n^2-5n+2)}\Big) \ll T^{n(n-1)-\frac{n}{2}+\varepsilon}$$

for $n \geq 5$. This error term is certainly acceptable.

From now on we weaken (20) a bit and consider

$$\begin{aligned}&T^{\frac{1}{2}n(n-1)+\frac{nd_1}{2f_1}+\varepsilon}\Big(1 + \delta^{-\frac{1}{4}n(n+1)+\frac{3}{2}+\frac{1}{2}(\frac{d_1^2}{f_1}-2f_1-\sum_{j\geq 2} d_j)}\Big) \\ =&T^{\frac{1}{2}n(n-1)+\frac{nd_1}{2f_1}+\varepsilon}\Big(1 + \delta^{-\frac{1}{4}n(n+1)+\frac{3}{2}+\frac{1}{2}(\frac{d_1^2}{f_1}-2f_1-n+d_1)}\Big).\end{aligned} \tag{21}$$

Since the cases $d_1 \in \{n-1, n\}$, $f_1 = 1$ have been ruled out, we always have $d_1/f_1 \leq \max(n-2, n/2) = n-2$ for $n \geq 5$ and so

$$T^{\frac{1}{2}n(n-1)+\frac{nd_1}{2f_1}+\varepsilon} \leq T^{n(n-1)-n/2+\varepsilon}$$

which is clearly acceptable. For the second term we need to analyze

$$\phi(\alpha, n, d_1, f_1) = \frac{1}{2}n(n-1) + \frac{nd_1}{2f_1} - \alpha\Big(-\frac{1}{4}n(n+1) + \frac{3}{2} + \frac{1}{2}\Big(\frac{d_1^2}{f_1} - 2f_1 - n + d_1\Big)\Big).$$

We compute

$$\frac{\partial}{\partial d}\phi(\alpha,n,d,f)=\frac{n-\alpha(2d+f)}{2f}$$

with a unique zero at $d_0=d_0(f)=(n-\alpha f)/(2\alpha)$ which is a local maximum. If $n/f\le\alpha$, then $d_0\le 0$, so on the interval $[1,n]$ the function $d\mapsto\phi(\alpha,n,d,f)$ has its maximum at $d=1$. If $n/f>\alpha$ and $\alpha\ge 1/2$, then $d_0<n$, and so the maximum lies at $d=d_0$.

Next we compute

$$\frac{\partial}{\partial f}\phi(\alpha,n,d,f)=\frac{\alpha d^2+2\alpha f^2-dn}{2f^2}.$$

If $n/d\le\alpha$, this is always non-negative, so $f\mapsto\phi(\alpha,n,d,f)$ is increasing in f. If $n/d>\alpha$, this has a unique positive zero at $((dn-\alpha d^2)/(2\alpha))^{1/2}$ which is a local minimum. So in either case, on the interval $[1,d]$, the function $f\mapsto\phi(\alpha,n,d,f)$ is maximized at $f=1$ or $f=d$.

We conclude that for $1\le f\le d\le n$, the function $\phi(\alpha,n,d,f)$ becomes globally maximal at most at the three points

$$(f,d)\in\{(1,1),(1,d_0(1)),(n/(3\alpha),n/(3\alpha))\}$$

where $f_0=n/(3\alpha)$ is the solution to $d_0(f_0)=f_0$. Substituting, we obtain

$$\phi(\alpha,n,d,f)\le\max\Big(\frac{n^2}{4}(\alpha+2)+\frac{3\alpha n}{4}-\frac{3\alpha}{2},\frac{n^2}{4}\Big(\alpha+\frac{1}{2\alpha}+2\Big)+\frac{3n}{4}(\alpha-1)-\frac{3}{8}\alpha\Big).$$

Thus our final error term is $T^{\psi(\alpha,n)+\varepsilon}$ for

$$\psi(\alpha,n)=\max\Big(n(n-1)-\alpha,\frac{n^2}{4}(\alpha+2)+\frac{3\alpha n}{4}-\frac{3\alpha}{2},\frac{n^2}{4}\Big(\alpha+\frac{1}{2\alpha}+2\Big)+\frac{3n}{4}(\alpha-1)-\frac{3}{8}\alpha\Big)$$

where we can freely choose $0<\alpha<2$. A final exercise in calculus shows that for $n\ge 5$ the best choice is

$$\alpha_0=\begin{cases}5/\sqrt{77}, & n=5,\\ \frac{n}{2n-1-\sqrt{2n^2-10n-4}}=1+\frac{1}{\sqrt{2}}+O\big(\frac{1}{n}\big), & n>5,\end{cases}$$

(satisfying $0<\alpha_0<2$) giving

$$\psi(\alpha_0,n)=n(n-1)-\begin{cases}5(9-\sqrt{77})/4, & n=5,\\ \alpha_0, & n>5.\end{cases}\tag{22}$$

This completes the proof of Theorem 1 for $n\ge 5$.

Needless to say that these estimates are (deliberately) a bit lossy and can be slightly improved, certainly on a scale $O(1/n)$. In the following two subsections we tighten all screws to obtain ("best-possible") Selberg type exponents.

4.2. **The case $n=3$.** According to the parametrization in Section 2.5 we distinguish four cases:

1) $r=1$, $d_1=f_1=3$ (cusp forms) with the subcases of a tempered cusp form (case 1a) and a non-tempered cusp form (case 1b);
2) $r=2$, $d_1=f_1=2$, $d_2=f_2=1$ (maximal Eisenstein series with a GL_2 cusp form) with the subcases of a tempered cusp form (case 2a) and a non-tempered cusp form (case 2b);
3) $r=2$, $d_1=2$, $d_2=f_2=f_1=1$ (Epstein zeta function);

4) $r=3$, $d_1=d_2=d_3=f_1=f_2=f_3=1$ (minimal Eisenstein series).

We can combine the cases 1a, 2a, 4 which are all tempered. Combining Corollary 4 (again dropping $\|\Re\mu\|$ in the denominator) and Lemma 1, for each of them we obtain

$$\begin{aligned}&\int_{E(d,f)} |\tilde{\chi}_{T(1\pm\delta),\delta}(\mu_\varpi)||\varpi(z)|^2 d\varpi\\ &\ll T^{3+\varepsilon}\int_{\substack{\mu_1+\mu_2+\mu_3=0\\ \mu_j\in i\mathbb{R}}} \frac{G(\Im\mu)(1+\delta\|\mu\|)^{-A}}{(1+\|\mu\|)^2}(1+|\mu_1-\mu_2|)(1+|\mu_1-\mu_3|)(1+|\mu_2-\mu_3|)|d\mu|\\ &\ll T^{3+\varepsilon}\delta^{-2}.\end{aligned}$$

Here we used that both terms of $G(\Im\mu)$ contain the square roots of precisely two factors of $(1+|\mu_1-\mu_2|)$, $(1+|\mu_1-\mu_3|)$, $(1+|\mu_2-\mu_3|)$, so that

$$G(\Im\mu)(1+|\mu_1-\mu_2|)(1+|\mu_1-\mu_3|)(1+|\mu_2-\mu_3|)\ll(1+\|\mu\|)^2$$

and we are left with two integration variables of effective length $1/\delta$.

In the cases 1b, 2b, 3 we have $\|\Re\mu\|\le 1/2$ and by unitarity (cf. (5)) the spectral spectral parameters are of the form $(\beta+it,-\beta+it,-2it)$ with $0<\beta\le 1/2$, $t\in\mathbb{R}$, so that in each of these cases we can estimate

$$\begin{aligned}&\int_{E(d,f)} |\tilde{\chi}_{T(1\pm\delta),\delta}(\mu_\varpi)||\varpi(z)|^2 d\varpi\\ &\ll T^{9/2+\varepsilon}\int_{t\in\mathbb{R}} \frac{G((t,t,-2t))(1+\delta|t|)^{-A}}{(1+|t|)^2}(1+|t|)^2dt\ll T^{9/2+\varepsilon}\delta^{-1/2}.\end{aligned}$$

Thus we obtain a total error of

$$T^6\delta+T^{3+\varepsilon}\delta^{-2}+T^{9/2+\varepsilon}\delta^{-1/2}\ll T^{5+\varepsilon}$$

upon choosing $\delta=1/T$.

4.3. **The case $n=4$.** According to the parameterization in Section 2.5 we distinguish 10 cases:

1) $r=1$, $d_1=f_1=4$ (cusp forms) with the subcases of a tempered cusp form (case 1a), a cusp form with exactly one pair of non-tempered components (case 1b) and a cusp form with 2 pairs of non-tempered components (case 1c);
2) $r=1$, $d_1=4$, $f_1=2$ (Speh representation);
3) $r=2$, $d_1=f_1=3$, $d_2=f_2=1$ (maximal Eisenstein series) with the subcases of a tempered GL(3) cusp form (case 3a), and a non-tempered cusp form (case 3b);
4) $r=2$, $d_1=3$, $f_1=d_2=f_2=1$ (Epstein zeta function)
5) $r=2$, $d_1=d_2=f_1=f_2=2$ with the subcases of two tempered GL(2) cusp forms (case 5a), exactly one tempered cusp form (case 5b) and two non-tempered cusp forms (case 5c);
6) $r=2$, $d_1=d_2=f_1=2$, $f_2=1$ with the subcases of a tempered GL(2) cusp form (case 6a) and a non-tempered cusp form (case 6b);
7) $r=2$, $d_1=d_2=2$, $f_1=f_2=1$;
8) $r=3$, $d_1=f_1=2$, $d_2=d_3=f_2=f_3=1$ with the subcases of a tempered GL(2) cusp form (case 8a) and a non-tempered cusp form (case 8b);
9) $r=3$, $d_1=2$, $d_2=d_3=f_1=f_2=f_3=1$;

10) $r=4$, $d_j=f_j=1$ (minimal Eisenstein series).

We can combine the tempered cases 1a, 3a, 5a, 6a, 8a, 10 and estimate each of them by

$$\ll T^{6+\varepsilon}\int_{\substack{\mu_1+\mu_2+\mu_3+\mu_4=0\\ \mu_j\in i\mathbb{R}}}\frac{G(\Im\mu)(1+\delta\|\mu\|)^{-A}}{(1+\|\mu\|)^{3.5}}\prod_{1\le i<j\le 4}(1+|\mu_i-\mu_j|)|d\mu|\ll T^{6+\varepsilon}\delta^{-4}$$

since $G(\Im\mu)\prod(1+|\mu_i-\mu_j|)\ll(1+\|\mu\|)^{4.5}$.

Next we consider the cases 1b, 3b, 5b, 8b, 9. In each of these we have $\|\Re\mu\|\le 1/2$ and spectral parameters of the form $(\beta+it_1-\beta+it_1,-i(t_1-t_2),-i(t_1+t_2))$ with $0\le\beta\le 1/2$, $t_1,t_2\in\mathbb{R}$. This is the only case where we need the extra $\|\Re\mu\|$ in the exponent of the denominator in Corollary 4. Here we estimate

$$\ll\sup_{0<\beta\le 1/2}T^{6+4\beta+\varepsilon}\int_{t_1,t_2\in\mathbb{R}}\frac{G(t_1,t_1,-t_1+t_2,-t_1-t_2)(1+\delta\|t\|)^{-A}}{(1+\|t\|)^{\frac{7}{2}+\frac{\beta}{4}}}$$
$$\times(1+|2t_1-t_2|)^2(1+|2t_1+t_2|)^2(1+|t_2|)dt\ll\sup_{0<\beta\le 1/2}T^{6+4\beta+\varepsilon}\delta^{-\frac{5}{2}+\frac{\beta}{4}}.$$

Here we used that both terms of $G(...)$ contain the square roots of at two factors of five linear forms $(1+|2t_1-t_2|)^2(1+|2t_1+t_2|)^2(1+|t_2|)$, and hence

$$G(t_1,t_1,-t_1+t_2,-t_1-t_2)(1+|2t_1-t_2|)^2(1+|2t_1+t_2|)^2(1+|t_2|)\ll(1+\|t\|)^4.$$

Since $\delta\gg T^{-2}$, the worst case is clearly $\beta=1/2$.

Slightly simpler are the cases 1c, 2, 5c, 6b, 7 where again $\|\Re\mu\|\le 1/2$ and the spectral parameters are even more degenerate of the form $(\beta_1+it,-\beta_1+it,\beta_2-it,-\beta_2-it)$ with $0<\beta_1,\beta_2\le 1/2$, $t\in\mathbb{R}$. Here we have the bound

$$\ll T^{8+\varepsilon}\int_{t\in\mathbb{R}}\frac{G(t,t,-t,-t)(1+\delta|t|)^{-A}}{(1+|t|)^{3.5}}(1+|t|)^4\,dt\ll T^{8+\varepsilon}\delta^{-1/2}.$$

It remains to treat case 4 where $\|\Re\mu\|=1$, and the spectral parameter is of the form $(1+it,it,-1+it,-3it)$, so that we obtain the bound

$$\ll T^{10+\varepsilon}\int_{t\in\mathbb{R}}\frac{G(t,t,t,-3t)(1+\delta|t|)^{-A}}{(1+|t|)^{3.5}}(1+|t|)^3\,dt\ll T^{10+\varepsilon}.$$

Combining the previous bounds, we obtain a total error of

$$\ll T^{12}\delta+T^{6+\varepsilon}\delta^{-4}+T^{8+\varepsilon}\delta^{-9/4}+T^{10+\varepsilon}\ll T^{12-6/5+\varepsilon}$$

upon choosing $\delta=T^{-6/5}$. Note that this estimate is very tight and the estimate in the cases 1b, 3b, 5b, 8b, 9 just suffices.

4.4. **Proof of Theorem 2.** The strategy is the same as in Subsection 4.1 except that we replace (14) with (15), so that in place of (19) we have to estimate

$$T^{\frac{1}{2}n(n-1)+n\max_j(\frac{1}{2}(\frac{d_j}{f_j}-1)+\delta_{f_j\ge 2}\frac{1}{2})+\varepsilon}\Big(1+\delta^{\frac{1}{4}n(n-1)+\frac{1}{2}-\frac{1}{2}\sum_j f_j(f_j-1)+1-\sum_j f_j}\Big).$$

The optimization procedure is again somewhat tedious. We assume that the maximum is attained at $j = 1$. Then clearly for all indices $j \geq 2$ the worst case is $f_j = d_j$. For fixed d_1, the remaining sum $\sum_{j\geq 2} d_j = n - d_1$ is fixed, so that

$$\frac{1}{2}\sum_{j=2}^{r} d_j(d_j+1)$$

becomes maximal if $r = 2$ (the degenerate case $d_1 = n$ would formally correspond to $r = 1$). Thus we bound the previous expression by

$$T^{\frac{1}{2}n(n-1)+\frac{n}{2}(\frac{d_1}{f_1}-\delta_{f_1=1})+\varepsilon}\Big(1+\delta^{\frac{1}{4}n(n-1)+\frac{3}{2}-\frac{1}{2}f_1(f_1+1)-\frac{1}{2}(n-d_1)(n-d_1+1)}\Big).$$

We weaken $f_1(f_1+1)$ to $f_1^2+d_1$, and consider the "exponent" function

$$\begin{aligned}\tilde{\phi}(\alpha,n,d,f) =& \frac{1}{2}n(n-1)+\frac{n}{2}\Big(\frac{d_1}{f_1}-\delta_{f_1=1}\Big)\\ &+\max\Big[0,-\alpha\Big(\frac{1}{4}n(n-1)+\frac{3}{2}-\frac{1}{2}(f_1^2+d_1)-\frac{1}{2}(n-d_1)(n-d_1+1)\Big)\Big]\end{aligned}$$

where as before $\delta = T^{-\alpha}$ with $0 < \alpha < 2$. It is easy to see that the function $f \mapsto nd/(2f) - f^2/2$ has its maximum at the boundary, so the worst case options for f_1 are $f_1 \in \{1, 2, d_1\}$. We have

$$\begin{aligned}\tilde{\phi}(\alpha,n,d,1) =& \frac{n(n+d-2)}{2}+\max\Big[0,-\alpha\Big(\frac{n(n-1)}{4}+1-\frac{d}{2}-\frac{(n-d)(n-d+1)}{2}\Big)\Big],\\ \tilde{\phi}(\alpha,n,d,2) =& \frac{n(n+\frac{1}{2}d-1)}{2}+\max\Big[0,-\alpha\Big(\frac{n(n-1)}{4}-\frac{1}{2}-\frac{d}{2}-\frac{(n-d_1)(n-d+1)}{2}\Big)\Big],\\ \tilde{\phi}(\alpha,n,d,d) =& \frac{n^2}{2}+\max\Big[0,-\alpha\Big(\frac{n(n-1)}{4}+\frac{3}{2}-\frac{d^2+d}{2}-\frac{(n-d)(n-d+1)}{2}\Big)\Big]\end{aligned}$$

(the last case if $d > 1$, while $d = 1$ is implicit in the first case). All three functions are non-concave as functions of d, so the maximum can be attained only at the boundary and it suffices to consider

$$\begin{aligned}\tilde{\phi}(\alpha,n,1,1) &= \max\Big[\frac{n(n-1)}{2},\frac{2+\alpha}{4}(n^2-n)-\frac{\alpha}{2}\Big],\\ \tilde{\phi}(\alpha,n,n-1,1) &= \max\Big[n^2-\frac{3}{2}n,\frac{4-\alpha}{4}n^2-\frac{6-\alpha}{4}n+\frac{\alpha}{2}\Big],\\ \tilde{\phi}(\alpha,n,2,2) &= \max\Big[\frac{2n^2-1}{4},\frac{2+\alpha}{4}n^2-\frac{5\alpha}{4}(n-2)\Big],\\ \tilde{\phi}(\alpha,n,n,2) &= \max\Big[\frac{3n^2-2n}{4},\frac{3-\alpha}{4}n^2+\frac{\alpha}{4}(n-2)+\frac{3\alpha}{2}\Big],\\ \tilde{\phi}(\alpha,n,n,n) &= \max\Big[\frac{n^2}{2},\frac{2+\alpha}{4}n^2+\frac{3\alpha}{4}(n-2)\Big].\end{aligned}$$

For n sufficiently large and $1 < \alpha < 2$, the maximum of these values is

$$\max\Big[n^2-\frac{3}{2}n,\frac{2+\alpha}{4}n^2+\frac{3\alpha}{4}(n-2)\Big],$$

so that our final error term becomes

$$(T^{n(n-1)-\alpha}+T^{n^2-3n/2}+T^{\frac{2+\alpha}{4}n^2+\frac{3\alpha}{4}(n-2)})T^{\varepsilon}.$$

The optimal choice for α is

$$\alpha = \frac{2(n^2-2n)}{n^2+3n-2} = 2 + O\Big(\frac{1}{n}\Big)$$

as desired.

References

[Bl] V. Blomer, *Epstein zeta-functions, subconvexity, and the purity conjecture*, J. Inst. Math. Jussieu **19** (2020), 581-596

[BM] V. Blomer, P. Maga, *Subconvexity for sup-norms of cusp forms on* $\mathrm{PGL}(n)$, Selecta Math. **22** (2016), 1269-1287

[DRS] W. Duke, Z. Rudnick, P. Sarnak, *Density of integer points on affine homogeneous varieties*, Duke Math. J. **71** (1993), 143-179

[GH] J. Getz, H. Hahn, *An introduction to automorphic representations* (2019)

[GNY] A. Gorodnik, A. Nevo, G. Yehoshua, *Counting lattice points in norm balls on higher rank simple Lie groups*, Math. Res. Lett. **24** (2017), 1285-1306.

[GR] I. S. Gradshteyn, I. M. Ryzhik, *Table of integrals, series, and products*, 7th ed., Academic Press 2007

[He] S. Helgason, *Groups and geometric analysis. Integral geometry, invariant differential operators, and spherical functions*, Mathematical Surveys and Monographs **83**. American Mathematical Society, Providence, RI, 2000.

[Iw] H. Iwaniec, *Spectral methods of automorphic forms*, Grad. Stud. Math. **53** (2002), AMS

[IK] H. Iwaniec, E. Kowalski, *Analytic Number Theory*, AMS Colloquium Publications **53** (2004)

[JK1] S. Jana, A. Kamber, *On the local* L^2*-bound of the Eisenstein series*, `arXiv:2210.16291`

[JK2] S. Jana, A. Kamber, *Optimal diophantine exponents for* $\mathrm{SL}(n)$, `arXiv:2211.05106`

[LP] P. Lax, R. Phillips, *The asymptotic distribution of lattice points in Euclidean and non-Euclidean spaces*, J. Funct. Anal. **46** (1982), 280-350.

[MW1] C. Mœglin, J.-L. Waldspurger, *Le spectre résiduel de* $\mathrm{GL}(n)$, Ann. Sci École Norm. Sup. **22** (1989), 605-674

[MW2] C. Mœglin, J.-L. Waldspurger, *Spectral decomposition and Eisenstein series*, Cambridge Tracts in Mathematics **113** (1995)

[Mu] W. Müller, *Weyl's law for the cuspidal spectrum of* SL_n, Ann. of Math. **165** (2007), 275-333

[PR] Y. Petridis, M. Risager, *Local average in hyperbolic lattice point counting*, with an appendix by N. Laaksonen, Math. Z. **285** (2017), 1319-1344

[Se] A. Selberg, *Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series*, J. Indian Math. Soc. B **20** (1956), 47-87

[Sz] W. Szerpiński, Über ein Problem aus der analytischen Zahlentheorie, Prace mat.-fiz. **17** (1906), 77-118

Mathematisches Institut, Endenicher Allee 60, 53115 Bonn, Germany
E-mail address: `blomer@math.uni-bonn.de`

Institut fr Mathematik, Universitt Zrich, Winterthurerstrasse 190, CH-8057 Zrich, Switzerland
E-mail address: `christopher.lutsko@uzh.ch`