Comment. Math. Helv. 101 (2026), 393–428
DOI 10.4171/CMH/612

A note on Sarnak's density hypothesis for Sp_4

Edgar Assing

Abstract. In this note, we prove Sarnak's (spherical) density hypothesis for the full discrete spectrum of the quotients $\Gamma_{\mathrm{pa}}(q)\backslash \mathrm{Sp}_4(\mathbb{R})$, where $\Gamma_{\mathrm{pa}}(q)$ are paramodular groups with square-free level q. To derive this estimate we upgrade a density estimate established via the Kuznetsov formula, which only accounts for the generic part, using Arthur's parametrization of the discrete spectrum.

1. Introduction

The generalized Ramanujan conjecture is a central and very difficult problem in the theory of automorphic forms. For GL_n, it predicts that all cuspidal representations are tempered. It has been known for a long time that this does not directly generalize to other groups; see [38]. The known counter examples to the naive Ramanujan conjecture are all cuspidal associated to parabolics (in short CAP). The correct generalized Ramanujan conjecture now states that all cuspidal non-CAP forms are tempered. But even for GL_2, where no CAP forms appear, the Ramanujan conjecture is still out of reach in general. Therefore, in practice, one often tries to replace the full conjecture by suitable approximations. One such approximation is Sarnak's density hypothesis, which predicts in a precise quantitative way that exceptions to the generalized Ramanujan conjecture are rare. This density hypothesis has the advantage that it can be formulated for the full discrete spectrum and it is expected to hold uniformly in great generality. In particular, it includes both CAP forms and residual forms. As we will see below, this feature is of particular importance in the context of this note.

1.1. The main result

Before continuing our discussion, let us formulate the spherical density hypothesis for Sp_4. Let $M > 2$ be a (large) parameter and $\Gamma \subseteq \mathrm{Sp}_4(\mathbb{Q})$ be a congruence lattice. To each eigenform of all invariant differential operators in $L^2_{\mathrm{disc}}(\Gamma\backslash\mathbb{H}_2)$, we associate

Mathematics Subject Classification 2020: 11F46 (primary); 11F70, 11F72 (secondary).
Keywords: exceptional eigenvalues, density hypothesis, paramodular forms, Arthur packets.

a spectral parameter $\mu_\varpi = (\mu_\varpi(1), \mu_\varpi(2)) \in \mathbb{C}^2$. Let $\mathcal{F}_\Gamma(M)$ denote a maximal orthogonal family of eigenforms ϖ with $\|\mu_\varpi\| \leq M$. The restriction of the spectral parameter to a ball of radius M is important to make the set $\mathcal{F}_\Gamma(M)$ finite. Note that the constant function **1** has spectral parameter

$$\mu_{\mathbf{1}} = \left(\frac{3}{2}, \frac{1}{2}\right),$$

and is always in $\mathcal{F}_\Gamma(M)$. We have normalized the spectral parameters so that an eigenform ϖ is tempered if and only if $\mu_\varpi \in (i\mathbb{R})^2$. Thus, we introduce

$$\sigma_\varpi = \max_{j=1,2} \left(|\Re(\mu_\varpi(j))|\right)$$

to measure how far an eigenfunction is from being tempered. Possible exceptions of badness σ are counted by

$$N_\Gamma(\sigma; M) = \sharp\{\varpi \in \mathcal{F}_\Gamma(M) : \sigma_\varpi \geq \sigma\}.$$

The two extreme cases are

$$N_\Gamma(0; M) = \sharp\mathcal{F}_\Gamma(M) \quad \text{and} \quad N_\Gamma(\sigma_{\mathbf{1}}; M) = 1.$$

Interpolating these two bounds leads to the density hypothesis

$$N(\sigma; M) \ll \sharp\mathcal{F}_\Gamma(M)^{1-\sigma/\sigma_1}.$$

We will be mainly interested in varying the lattice Γ through some suitable level family, and will therefore ignore the dependency on M, confining ourselves to the weaker hypothesis

$$N_\Gamma(\sigma; M) \ll_{M,\varepsilon} \mathrm{Vol}(\Gamma\backslash\mathbb{H}_2)^{1-\sigma/\sigma_{\mathbb{1}}+\varepsilon}. \tag{1}$$

Our main result is Theorem 5.9 below. It establishes the density hypothesis and more for the paramodular group $\Gamma_{\mathrm{pa}}(q)$ as defined in (3). We can summarize the most important part of the statement as follows.

Theorem 1.1. *For q square-free, we have*

$$N_{\Gamma_{\mathrm{pa}}(q)}(\sigma; M) \ll_{M,\varepsilon} \mathrm{Vol}\big(\Gamma_{\mathrm{pa}}(q)\backslash\mathbb{H}_2\big)^{1-\frac{\sigma}{\sigma_{\mathbb{1}}}(1+1/2)+\varepsilon} + 1. \tag{2}$$

1.2. The methods

The proof of Theorem 1.1 has two main steps. First, a density result for generic forms is proved. The main global tool here is a Kuznetsov-type formula for Sp_4, which has the restriction to generic forms built in by default. A similar argument has appeared in [31] and we follow it closely.[1] Second, we have to account for the non-generic forms manually. This is done using Arthur's classification of the discrete spectrum predicted in [1]. We will see that non-generic non-CAP cusp forms (i.e., those cusp forms that are expected to be tempered) are covered by the generic density result. The remaining task is to handle the contribution of CAP forms. These are treated by carefully studying the associate Arthur packets, which were computed explicitly in [41].

It should be mentioned that our approach to use Arthur's classification to control the non-generic spectrum is not completely new. Indeed, similar ideas were used in [33], and also later in [16].

Note that one might have hoped to transfer the density estimates directly from GL_4 to Sp_4. However, since the image of the transfer from Sp_4 is self-dual and the Plancherel density drops when restricted to self-dual forms, this turns out to be difficult. The latter phenomenon was worked out in [27].

1.3. Further discussion

The density hypothesis in rank one was a hot topic in the 1990's and is relatively well understood. We refer to [22, 23, 25, 37] and the references within for a more in depth discussion of the matter. In higher rank, recently breakthroughs have been obtained for GL_n, where (1) has been established for the standard Hecke congruence subgroup, the principal congruence subgroup and a Borel-type congruence subgroup; see [5–7]. Note that even though we have formulated the density hypothesis only for Sp_4, it should be clear how to adapt this formulation to other groups. For groups of higher rank different from GL_n strong density results are still sparse. Following partial progress in the case of Sp_4 made in [31], our result is, to the best of our knowledge, the first spherical density theorem covering the full discrete spectrum beyond GL_n. We hope that the strategy of combining a generic density hypothesis, obtained using a version of the Kuznetsov formula, with the endoscopic classification will generalize to a wider class of groups.

We conclude the introduction by briefly pointing towards some interesting points:

(1) We have made no attempt of explicating how the implicit constant in (2) depends on M. A close analysis of the argument along the lines of [6] will show that this dependency is polynomial, but obtaining a reasonable exponent would require more work.

[1] Unfortunately, the argument in [31] seems to have a gap. We explain this in more detail in Remark 5.1 below.

(2) The assumption that q is square-free enters only in the first step of our proof. Indeed, when setting up the Kuznetsov formula we make use of [10] to handle the spectral side. This is where square-freeness is required. The general case would follow from a conjecture of Lapid–Mao [29] (see also Conjecture 4.1 below) together with the corresponding local estimates (12). Note that the conjecture of Lapid and Mao is known for tempered representations due to recent work by Furusawa and Morimoto; see [18, Theorem 6.3]. Unfortunately, one would need the conjecture precisely for (the conjecturally empty set of) generic non-tempered forms.

(3) When comparing the estimate given in Theorem 1.1 with Sarnak's density hypothesis as stated in (1) we observe that there is an improvement of $1/2$ in the exponent. This breach of the density barrier has its roots in the estimate for the generic contribution, which is obtained using the Kuznetsov formula. The relevant part of the argument is carried out in Section 5.1 below. One cannot expect to obtain further improvements in the density theorem as stated in (2), because the estimate is sharp at $\sigma = 1/2$, where CAP-forms of Saito–Kurokawa type contribute. However, one can expect further improvements in the exponent on the generic part of the spectrum. This requires non-trivial estimates for Kloosterman sums and we will explore this feature for prime q. See Theorems 5.6 and 5.9 below for details.

(4) Other approximations to the generalized Ramanujan conjecture are absolute upper bounds on σ_{ϖ} for non-CAP forms ϖ. In the case of Sp_4, these can be obtained by exploiting the transfer to GL_4 established in [1, 2] (or [4] in the generic case), and applying bounds towards the Ramanujan conjecture on GL_4. The latter are well recorded in [8], for example. The CAP forms of Sp_4 all satisfy $\sigma_{\varpi} = 1/2$, so that an in-principle non-negligible part of the spectrum is heavily non-tempered. However, these forms are still covered by Theorem 1.1. This is not only convenient for possible applications, but also shows how robust Sarnak's density hypothesis is.

(5) There are different measures for the non-temperedness of ϖ, which lead to alternative formulations of the density hypothesis. In general not all of these are equivalent. For applications where the pre-trace formula is used directly, our formulation works quite well. Examples that fall in this category are lattice point counting problems and optimal lifting. See [26] or [6, Section 1.3 and Section 8] for related discussions in the context of GL_n. Another popular formulation is in terms of $p(\varpi)$, which is the infimum over real numbers $p \geq 2$ such that the matrix coefficient of ϖ is in $L^p(\mathrm{Sp}_4(\mathbb{R}))$. It is this invariant, which was used by Sarnak and Xue to formulate their multiplicity conjecture; see [39]. This aspect is considered in [16] for spaces closely related to ours. One can deduce the value of $p(\varpi)$ from the information contained in $\Re(\mu_{\varpi})$, but in general not from σ_{ϖ} alone; see [21, Lemma 3.2], for example.

(6) Our arguments apply in principle to other lattice families, but some difficulties arise in general. Indeed, the paramodular group is convenient for several reasons. First, we have the results from [10] available. This makes the generic estimate unconditional for square-free levels. Second, as shown in [36,40,41], for example, the level structure given by the paramodular group has many nice properties. In particular, it harmonizes with Arthur's classification. That said, we will take the liberty to comment on features that arise for other lattice families once in a while.

2. Notation and preliminaries

As the title suggests, we are interested in establishing (a version of) Sarnak's density hypothesis for Sp_4. As the mostly classical discussion in the introduction reveals, we are particularly interested in the spectrum of (congruence) quotients of the form $\Gamma\backslash\mathbb{H}_2$ for $\Gamma \subseteq \mathrm{Sp}_4(\mathbb{Q})$. However, in order to establish the desired result it will be useful to work adelically. It turns out that in the latter setting it is often more convenient to work with GSp_4 or PGSp_4 instead of Sp_4. Therefore, it will be important to introduce all these groups as well as some related objects.

Conventions differ when dealing with symplectic groups. We will closely follow the set-up in [31].

2.1. Matrices and subgroups

We set

$$J = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}.$$

The general symplectic group over a ring R (with unit) is given by

$$G(R) = \mathrm{GSp}_4(R) = \left\{ g \in \mathrm{Mat}_4(R) : g^\top J g = \mu \cdot J \text{ for } \mu = \mu(g) \in R^\times \right\}.$$

The map $g \mapsto \mu(g)$ is called the multiplier homomorphism. The special symplectic group is denoted by $G_0(R) = \mathrm{Sp}_4(R)$ and consists of those $g \in G(R)$ with $\mu(g) = 1$. Furthermore, we set $G(\mathbb{R})^+ = \{g \in G(\mathbb{R}) : \mu(g) > 0\}$.

Our choice of maximal torus is

$$T(R) = \left\{ \mathrm{diag}(x, y, \mu \cdot x^{-1}, \mu \cdot y^{-1}) : x, y, \mu \in R^\times \right\}$$

and we put $T_0(R) = T(R) \cap \mathrm{Sp}_4(R)$. For later reference, we write

$$t(x, y) = \mathrm{diag}(x, y, x^{-1}, y^{-1}).$$

A useful variant is given by

$$c = (c_1, c_2) \mapsto c^* = t\Big(\frac{1}{c_1}, \frac{c_1}{c_2}\Big).$$

Over the real numbers, the maximal compact subgroup is given by $K_\infty = \mathrm{O}_4(\mathbb{R}) \cap G(\mathbb{R})$. The classical Siegel upper half space (of degree 2) can be identified with

$$\mathbb{H}_2 = G_0(\mathbb{R})/\big(K_\infty \cap G_0(\mathbb{R})\big).$$

We will further need the following special coordinates in the real case. We first define the embedding $\iota\colon \mathbb{R}_+^2 \to T_0(\mathbb{R})$ by

$$\iota(y) = t\big(y_1 y_2^{1/2}, y_2^{1/2}\big).$$

The inverse of ι will be denoted by $\mathrm{y}\colon T_0(\mathbb{R}_+) \to \mathbb{R}_+^2$. We extend y to G_0 via the Iwasawa decomposition $G_0(\mathbb{R}) = U(\mathbb{R})T_0(\mathbb{R}_+)(K_\infty \cap G_0)$.

Next we define some unipotent matrices:

$$n(x) = \begin{pmatrix} 1 & x & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -x & 1 \end{pmatrix} \quad \text{and} \quad s\left(\begin{pmatrix} a & b \\ b & c \end{pmatrix}\right) = \begin{pmatrix} 1 & 0 & a & b \\ 0 & 1 & b & c \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The Borel subgroup is given by

$$B = TU = UT \quad \text{with} \quad U(R) = \big\{n(x) \cdot s(M) : x \in R \text{ and } M \in \mathrm{Sym}_{2\times 2}(R)\big\}.$$

This is the standard minimal parabolic subgroup.

Let $\Gamma_0 = \mathrm{Sp}_4(\mathbb{Z})$ and define the paramodular group of level q by

$$\Gamma_{\mathrm{pa}}(q) = \begin{bmatrix} \mathbb{Z} & \mathbb{Z} & q^{-1}\mathbb{Z} & \mathbb{Z} \\ q\mathbb{Z} & \mathbb{Z} & \mathbb{Z} & \mathbb{Z} \\ q\mathbb{Z} & q\mathbb{Z} & \mathbb{Z} & q\mathbb{Z} \\ q\mathbb{Z} & \mathbb{Z} & \mathbb{Z} & \mathbb{Z} \end{bmatrix} \cap \mathrm{Sp}_4(\mathbb{Q}). \tag{3}$$

Note that this is not a subgroup of Γ_0, but $U(\mathbb{Z}) \subseteq \Gamma_{\mathrm{pa}}(q) \cap U(\mathbb{Q})$. The latter is not an equality and the difference is measured by the index

$$\mathcal{N}\big(\Gamma_{\mathrm{pa}}(q)\big) = \big[\Gamma_{\mathrm{pa}}(q) \cap U(\mathbb{Q}) : U(\mathbb{Z})\big] = q. \tag{4}$$

While our main results concern only the paramodular group $\Gamma_{\mathrm{pa}}(q)$, we will often discuss general aspects of the argument for arbitrary lattices $\Gamma \subseteq \mathrm{Sp}_4(\mathbb{Q})$. This allows us to comment on certain interesting features that arise for the principal congruence subgroup or other parahoric lattices. We write X_Γ for the classical quotient

$$X_\Gamma = \Gamma \backslash \mathbb{H}_2.$$

When working adelically we will encounter a certain open compact subgroup as counterparts to the lattice Γ. Recall that strong approximation for G states that

$$G(\mathbb{A}) = G(\mathbb{Q}) \cdot \big(G(\mathbb{R})^+ \times K\big),$$

where $K \subseteq G(\mathbb{A}_{\mathrm{fin}})$ is an open compact subgroup such that

$$\mu\colon K \to \widehat{\mathbb{Z}}^\times$$

is surjective. We thus define the groups K_Γ by requiring

$$\Gamma = G(\mathbb{Q}) \cap G(\mathbb{R})^+ \times K_\Gamma.$$

Then $K_\Gamma = \prod_p K_{\Gamma,p}$, and assume that $K_{\Gamma,p} \subseteq G(\mathbb{Q}_p)$ is an open compact subgroup. For the paramodular group, it is easy to write down these groups explicitly:

$$K_{\Gamma_{\mathrm{pa}}(q),p} = \begin{bmatrix} \mathbb{Z}_p & \mathbb{Z}_p & p^{-v_p(q)}\mathbb{Z}_p & \mathbb{Z}_p \\ p^{v_p(q)}\mathbb{Z}_p & \mathbb{Z}_p & \mathbb{Z}_p & \mathbb{Z}_p \\ p^{v_p(q)}\mathbb{Z}_p & p^{v_p(q)}\mathbb{Z}_p & \mathbb{Z}_p & p^{v_p(q)}\mathbb{Z}_p \\ p^{v_p(q)}\mathbb{Z}_p & \mathbb{Z}_p & \mathbb{Z}_p & \mathbb{Z}_p \end{bmatrix} \cap G(\mathbb{Q}_p).$$

In general, one sees that there is a square-free integer D_Γ, so that $K_{\Gamma,p} = G(\mathbb{Z}_p)$ for all $p \mid D_\Gamma$. For lack of a better name, we refer to D_Γ as the discriminant of Γ. Note that if $\Gamma = \Gamma_{\mathrm{pa}}(q)$, then $D_\Gamma = \mathrm{rad}(q)$. Of course, one has $D_{\Gamma_0} = 1$.

Note that, if not explicitly stated otherwise, rational matrices (or scalars) will always be embedded diagonally in the corresponding adelic spaces. On the other hand, we will use the notation $\iota_\infty\colon G(\mathbb{R}) \to G(\mathbb{A})$ for the embedding $g \mapsto (g, 1, \ldots)$. This map will play a distinguished role in the adelization process of automorphic forms and therefore deserves a special name.

The two simple roots of G_0 are given by

$$\alpha\big(t(x,y)\big) = xy^{-1} \quad \text{and} \quad \beta\big(t(x,y)\big) = y^2.$$

The positive roots are then $\Sigma_+ = \{\alpha, \beta, \alpha+\beta, 2\alpha+\beta\}$.

We denote the Weyl group by W. It is given by

$$W = \{1, s_1, s_2, s_1s_2, s_2s_1, s_1s_2s_1, s_2s_1s_2, s_1s_2s_1s_2\},$$

for

$$s_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix} \quad \text{and} \quad s_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}.$$

Note that s_1 (resp. s_2) is the simple root reflection corresponding to α (resp. β). We set

$$U_w = w^{-1}U^{\top}w \cap U \quad \text{and} \quad \bar{U}_w = w^{-1}Uw \cap U.$$

Further, we will write ${}^w y = \mathrm{y}[w\iota(y)^{-1}w^{-1}]$ for $y \in \mathbb{R}_2^+$.

2.2. Measures and characters

We first equip $\mathbb{R}$ with the usual Lebesgue measure. The fields $\mathbb{Q}_2, \mathbb{Q}_3, \ldots$ are equipped with the additive Haar measure normalized by $\mathrm{Vol}(\mathbb{Z}_p) = 1$. These give rise to the product measure on $\mathbb{A}$, which by strong approximation satisfies $\mathrm{Vol}(\mathbb{Q}\backslash\mathbb{A}) = 1$. If R is one of these rings, we put the Tamagawa measure on $U(R)$. It is normalized by

$$\mathrm{Vol}\big(U(\mathbb{Q})\backslash U(\mathbb{A})\big) = 1.$$

Furthermore, the measures on U naturally factor through $U = U_w \cdot \bar{U}_w$.

For $\alpha \in \mathbb{C}^2$ and $y \in \mathbb{R}_+^2$, we employ the usual notation $y^\alpha = y_1^{\alpha_1} \cdot y_2^{\alpha_2}$. The exponent $\eta = (2, 3/2)$, which is used to define the measure

$$d^*y = y^{-2\eta} \frac{dy_1}{y_1} \cdot \frac{dy_2}{y_2}$$

on $\mathbb{R}_+^2$, will be important. We use ι to push this measure to $T_0(\mathbb{R}_+)$. Finally, we equip K_∞ with the Haar measure such that $\mathrm{Vol}(K_\infty \cap G_0(\mathbb{R})) = 1$. The Iwasawa decomposition allows us to define a Haar measure on $G_0(\mathbb{R})$ by

$$\int_{G_0(\mathbb{R})} f(g)\, dg = \int_{K_\infty \cap G_0(\mathbb{R})} \int_{T_0(\mathbb{R}_+)} \int_{U(\mathbb{R})} f\big(u\iota(y)k\big)\, du\, d^*y\, dk.$$

This descends to the usual measure $du\, d^*y$ on $\mathbb{H}_2$. To make the comparison of different measures easier, we recall that

$$\mathrm{Vol}\big(G_0(\mathbb{Z})\backslash G_0(\mathbb{R}), dg\big) = \frac{\zeta(2)\zeta(4)}{2\pi^3}.$$

See, for example, [28, Proposition A.3]. We can use this measure to define a measure on $\mathrm{PGSp}_4(\mathbb{R})$ and to normalize $G(\mathbb{R})$ in the obvious way.

At the finite places, we normalize the Haar measure on $Z(\mathbb{Q}_p)\backslash G(\mathbb{Q}_p)$ and $G(\mathbb{Q}_p)$, so that

$$\mathrm{Vol}\big((Z(\mathbb{Q}_p)\cap G(\mathbb{Z}_p))\backslash G(\mathbb{Z}_p), d\overline{g}\big) = 1 = \mathrm{Vol}\big(G(\mathbb{Z}_p), dg\big).$$

(Note that these normalizations are compatible if we equip $\mathbb{Q}_p^\times$ with the Haar measure satisfying $\mathrm{Vol}(\mathbb{Z}_p^\times, d^\times x) = 1$.) By [36, Lemma 3.3.3], we have

$$\mathrm{Vol}\big(K_{\Gamma_{\mathrm{pa}}(q),p}, dg\big) = q^{-2}(1+q^{-2})^{-1}.$$

Globally, we equip $G(\mathbb{A})$ with the product measure $dg = dg_\infty \cdot \prod_p dg_p$. Using strong approximation, it is easy to see that with our normalizations, we have

$$\mathrm{Vol}\big(Z(\mathbb{A})G(\mathbb{Q})\backslash G(\mathbb{A}), dg\big) = \frac{\zeta(2)\zeta(4)}{2\pi^3}.$$

For comparison, let us note that the Tamagawa measure on $G(\mathbb{A})$, which we will denote by $d^{\mathrm{ta}}g$, is normalized such that

$$\mathrm{Vol}\big(Z(\mathbb{A})G(\mathbb{Q})\backslash G(\mathbb{A}), d^{\mathrm{ta}}g\big) = 2.$$

Thus, we must have

$$d^{\mathrm{ta}}g = \frac{4\pi^3}{\zeta(2)\zeta(4)}\, dg.$$

Given a lattice Γ, we write

$$\mathcal{V}(\Gamma) = \mathrm{Vol}(X_\Gamma).$$

for the co-volume. It turns out that $\mathcal{V}(\Gamma_{\mathrm{pa}}(q)) \asymp q^2$.

We now fix the standard additive, $\mathbb{Q}$-invariant character $\psi\colon \mathbb{Q}\backslash\mathbb{A} \to S^1$. We have a factorization $\psi = \psi_\infty \cdot \prod_p \psi_p$ with $\psi_\infty(x) = e(x)$. Also note that all local characters ψ_p are trivial on $\mathbb{Z}_p$, but non-trivial on $p^{-1}\mathbb{Z}_p$ (i.e., they are unramified). This gives rise to a character

$$\boldsymbol{\psi}\colon U(\mathbb{Q})\backslash U(\mathbb{A}) \to S^1$$

given by

$$\boldsymbol{\psi}^{(X)}\left(n(x)s\left(\begin{pmatrix} * & * \\ * & c\end{pmatrix}\right)\right) = \psi(X_1 x + X_2 c),$$

for $X = (X_1, X_2) \in (\mathbb{A}^\times)^2$. Of course, we have the factorization

$$\boldsymbol{\psi}^{(X)} = \boldsymbol{\psi}_\infty^{(X_\infty)} \cdot \prod_p \boldsymbol{\psi}_p^{(X_p)},$$

where the local characters are defined in the obvious way. If X is trivial, we drop it from the notation. The archimedean character $\boldsymbol{\psi}_\infty$ on $U(\mathbb{R})$ will play a distinguished role below in our discussion of Kloosterman sums.

3. Kloosterman sums

We will now introduce Kloosterman sums following [31, 32], where the relevant theory is worked out explicitly for Sp_4. A more general discussion can be found in [12, 13]. Throughout this discussion it will be important that the lattice Γ satisfies $U(\mathbb{Z}) \subseteq \Gamma \cap U(\mathbb{Q})$. Recall that this is true for $\Gamma_{\mathrm{pa}}(q)$.

3.1. Definitions and basic properties

We start by defining the global Kloosterman set

$$\mathfrak{X}_\Gamma(c^*w) = U(\mathbb{Z})\backslash\big[U(\mathbb{Q})wc^*U_w(\mathbb{Q}) \cap \Gamma\big]/U_w(\mathbb{Z}).$$

For $M, N \in \mathbb{N}^2$, we call a tuple $(w, c) \in W \times \mathbb{N}^2$ admissible (or, more precisely, (M, N)-admissible) if

$$\psi_\infty^{(M)}\big(wc^*x(c^*)^{-1}w^{-1}\big) = \psi_\infty^{(N)}(x) \quad \text{for all } x \in \bar{U}_w.$$

This is precisely the condition [31, (4.1)], which is termed *well-definedness condition* there. The constraints imposed on $c = (c_1, c_2)$ by the admissibility of (w, c) can be computed by hand. Details can be found in [31, (4.4)] or [32, Section 5.1] (and also in a slightly different form in [11, Section 5.4]). We summarize the constraints on N, M and c arising from admissibility in Table 1.

w	1	s_1	s_2	s_1s_2
$c = (c_1, c_2)$	$M = N$	$N_2 = M_2 = 0$	$N_1 = M_1 = 0$	$N_1 = M_2 = 0$
w	s_2s_1	$s_1s_2s_1$	$s_2s_1s_2$	$s_1s_2s_1s_2$
$c = (c_1, c_2)$	$N_2 = M_1 = 0$	$N_2 = M_2 \cdot \frac{c_1^2}{c_2^2}$	$N_1 = M_1 \cdot \frac{c_2}{c_1^2}$	-

Table 1. Admissibility constraints.

Further, we call a Weyl element w relevant, if the condition on c can be satisfied without N_1, M_1, N_2, or M_2 being zero. Thus, the relevant Weyl elements are 1, $s_1s_2s_1$, $s_2s_1s_2$, and $s_1s_2s_1s_2$.

For admissible (w, c), we define the Kloosterman sum by

$$\mathrm{KL}_{\Gamma,w}(c; M, N) = \sum_{xwc^*x' \in \mathfrak{X}_\Gamma(c^*w)} \psi_\infty^{(M)}(x)\psi_\infty^{(N)}(x').$$

It is notationally convenient to set $\mathrm{KL}_{\Gamma,w}(c; M, N) = 0$ if (w, c) is not admissible.

The Kloosterman sum for the trivial Weyl element is easy to compute. Indeed, one gets

$$\mathrm{KL}_{\Gamma,1}(c; M, N) = \delta_{\substack{M=N,\\ c=(1,1)}} \cdot \mathcal{N}(\Gamma). \tag{5}$$

The general case is much harder. Note that the Kloosterman sums are actually local in nature. In particular, for $\Gamma_{\mathrm{pa}}(q)$, they factor as

$$\mathrm{KL}_{\Gamma_{\mathrm{pa}}(q),w}(c; \mathbf{1}, \mathbf{1}) = \mathrm{KL}_{\Gamma_{\mathrm{pa}}(q),w}(d; \mathbf{1}, N_{c'})\, \mathrm{KL}_{\Gamma_0,w}(c'; \mathbf{1}, N_d), \tag{6}$$

where $c_i = d_i c_i'$ with $d_i \mid q^\infty$ and $(c_i', q) = 1$.[2] Note that the coordinates of $N_{c'}$ are co-prime to q and $(N_{d,1}N_{d,2}, c_1'c_2') = 1$. This is essentially [31, (4.5)]. For more details we refer to [43] and [17, Proposition 2.4], where Kloosterman sums for the lattice $\mathrm{SL}_n(\mathbb{Z})$ are treated, but the arguments directly generalize to our setting.

3.2. Bounds for Kloosterman sums

Below we will need several estimates for Kloosterman sums. The upshot of the factorization stated in (6) is, that it separates the unramified and the ramified contribution nicely. Note that we always have the trivial bound

$$|\mathrm{KL}_{\Gamma,w}(c; M, N)| \leq \sharp\mathfrak{X}_\Gamma(c^* w).$$

For $\Gamma = \Gamma_0$, the Kloosterman sets are sufficiently well understood by [13], and we have the estimate

$$\mathrm{KL}_{\Gamma_0,w}(c; \mathbf{1}, \mathbf{1}) \leq \sharp\mathfrak{X}_{\Gamma_0}(c^* w) \ll (c_1c_2)^{1+\varepsilon}.$$

For $\Gamma_{\mathrm{pa}}(q)$, estimating $\sharp\mathfrak{X}_{\Gamma_{\mathrm{pa}}(q)}(c^* w)$ is a more delicate matter, which we will look at in Lemma 3.2 below. Before doing so we will recall some non-trivial estimates for unramified Kloosterman sums.

Theorem 3.1 ([32]). *Let $N, c \in \mathbb{N}^2$ and suppose that $(N_1N_2, c_1c_2) = 1$. Then we have the bounds*

$$\begin{aligned}
\mathrm{Kl}_{\Gamma_0,s_1s_2s_1}(c; \mathbf{1}, N) &\ll c_1^{5/3+\varepsilon} && \text{for } c_1 = c_2,\\
\mathrm{Kl}_{\Gamma_0,s_2s_1s_2}(c; \mathbf{1}, N) &\ll c_1^{5/2+\varepsilon} && \text{for } c_1^2 = c_2,\\
\mathrm{Kl}_{\Gamma_0,s_1s_2s_1s_2}(c; \mathbf{1}, N) &\ll c_1^{1/2+\varepsilon} c_2^{3/4+\varepsilon} (c_1, c_2)^{1/2}.
\end{aligned}$$

[2] Given two numbers $a, b \in \mathbb{N}$ we write $a \mid b^\infty$ if every prime divisor of a also divides b.

Proof. This is extracted from [32, Theorem 1.2]. Note that the estimates given in [32, Theorem 1.2] are more general, so that it might be useful to make some comments. First, note that $n_w(c_1, c_2)$ used in [32] reads

$$n_w(c_1, c_2) = (c_1, c_2)^* w$$

in our notation. Second, we recall that $\mathrm{Kl}_{\Gamma_0, w}(c; \mathbf{1}, N) = 0$ unless (w, c) is $(\mathbf{1}, N)$-admissible. Since in this case the bounds are trivially true we only need to treat (w, c) that are $(\mathbf{1}, N)$-admissible. Finally, we remark that the assumption $(N_1 N_2, c_1 c_2) = 1$ allows us discard all the gcd's involving entries of N and c simultaneously from the estimates given in [32, Theorem 1.2].

The estimates are collected as follows:

- The bound for $\mathrm{Kl}_{\Gamma_0, s_1 s_2 s_1}(c; \mathbf{1}, N)$ follows from the sixth estimate in [32, Theorem 1.2] with $c_1 = c_2$, in which case $c_2 \mid c_1^2$ is obviously satisfied, and we have $(c_1, c_2) = c_1$.
- Similarly, we obtain the bound for $\mathrm{Kl}_{\Gamma_0, s_2 s_1 s_2}(c; \mathbf{1}, N)$ from the seventh estimate in [32, Theorem 1.2] with $c_1^2 = c_2$. Again, this is a special case of $c_1^2 \mid c_2$, and we also have $(c_1^2, c_2) = c_1^2$.
- The bound for the long Weyl element is given in the last estimate of [32, Theorem 1.2].

This completes the argument. ∎

Ramified Kloosterman sums were computed in [31, Section 4] for congruence subgroups of Siegel type. We adapt these results to the lattices $\Gamma_{\mathrm{pa}}(q)$.

Lemma 3.2. *Let $q > 1$. Unless $q \mid a$, we have*

$$\mathrm{KL}_{\Gamma_{\mathrm{pa}}(q), s_1 s_2 s_1}\big((a, a), \mathbf{1}, N\big) = \mathrm{KL}_{\Gamma_{\mathrm{pa}}(q), s_2 s_1 s_2}\big((a, a^2), \mathbf{1}, N\big) = 0.$$

Similarly,

$$\mathrm{KL}_{\Gamma_{\mathrm{pa}}(q), s_1 s_2 s_1 s_2}\big((a, b), \mathbf{1}, N\big) = 0$$

unless $q \mid a$ and $a \mid b$. If q is prime, we have

$$\mathrm{KL}_{\Gamma_{\mathrm{pa}}(q), s_1 s_2 s_1}\big((q, q), \mathbf{1}, N\big) = q^2,$$
$$\mathrm{KL}_{\Gamma_{\mathrm{pa}}(q), s_2 s_1 s_2}\big((q, q^2), \mathbf{1}, N\big) = 0.$$

Furthermore, if q is prime, $(N_2, q) = 1$ and $k \in \{1, 2, 3\}$, then

$$\mathrm{KL}_{\Gamma_{\mathrm{pa}}(q), s_1 s_2 s_1 s_2}\big((q, q^k), \mathbf{1}, N\big) \ll \tau(q^k) q^{2+(k-1)/2}.$$

Proof. We start by considering the Weyl element $w = s_1 s_2 s_1$ with a corresponding modulus $c = (a, a)$. We compute

$$xwc^*x' = \begin{pmatrix} 1 & x_1 & x_2 & x_3 \\ 0 & 1 & x_3' & x_4 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -x_1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & -a^{-1} & 0 \\ 0 & -1 & 0 & 0 \\ a & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & \frac{y_1}{a} & \frac{y_2}{a} & \frac{y_3}{a} \\ 0 & 1 & \frac{y_3}{a} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{y_1}{a} & 1 \end{pmatrix}$$

$$= \begin{pmatrix} ax_2 & x_2 y_1 - x_1 & -\frac{1}{a} - \frac{x_1 y_3}{a} + x_3' y_2 + \frac{x_4 y_1}{a} & x_2 y_3 - x_3 \\ ax_3' & x_3' y_1 - 1 & x_3' y_2 - \frac{y_3}{a} + \frac{x_4 y_1}{a} & y_3 x_3' - x_4 \\ a & y_1 & y_2 & y_3 \\ -ax_1 & -x_1 y_1 & -x_1 y_2 + \frac{y_1}{a} & -1 - x_1 y_3 \end{pmatrix}.$$

Now we will look at what we gain from the condition

$$\gamma = xwc^*x' \in \begin{bmatrix} \mathbb{Z} & \mathbb{Z} & q^{-1}\mathbb{Z} & \mathbb{Z} \\ q\mathbb{Z} & \mathbb{Z} & \mathbb{Z} & \mathbb{Z} \\ q\mathbb{Z} & q\mathbb{Z} & \mathbb{Z} & q\mathbb{Z} \\ q\mathbb{Z} & \mathbb{Z} & \mathbb{Z} & \mathbb{Z} \end{bmatrix} = \Gamma_{\mathrm{pa}}(q).$$

In complete generality, we only make the observation that $q \mid a$ is necessary. Now we specialize to $a = q$. First, we look at the third row and deduce that $y_1 = y_3 = 0$ and $y_2 \in \mathbb{Z}/q\mathbb{Z}$. Studying the first column gives $x_1 = x_3' = 0$ and $x_2 = q\tilde{x}_2$ with $\tilde{x}_2 \in \mathbb{Z}/q\mathbb{Z}$. Finally, from the second row we extract $x_4 = 0$. This is all the information we can get. In summary, we have

$$\mathrm{KL}_{\Gamma_{\mathrm{pa}}(q), s_1 s_2 s_1}\big((q, q), \mathbf{1}, N\big) = q^2,$$

since both x_2 and y_2 are not seen by the characters.

We turn towards $w = s_2 s_1 s_2$ and $c = (a, a^2)$. One computes

$$xwc^*x' = \begin{pmatrix} 1 & x_1 & x_2 & x_3 \\ 0 & 1 & x_3' & x_4 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -x_1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & a^{-1} \\ 0 & 0 & -a^{-1} & 0 \\ 0 & -a & 0 & 0 \\ a & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & \frac{y_1}{a^2} & \frac{y_2}{a^2} \\ 0 & 1 & \frac{y_2}{a^2} & \frac{y_3}{a^2} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} ax_3 & -ax_2 & \frac{x_3 y_1}{a} - \frac{x_2 y_2}{a} - \frac{x_1}{a} & \frac{1}{a} - \frac{x_2 y_3}{a} + \frac{x_3 y_2}{a} \\ ax_4 & -ax_3' & -\frac{1}{a} - \frac{x_3' y_2}{a} + \frac{x_4 y_1}{a} & \frac{x_4 y_2}{a} - \frac{x_3' y_3}{a} \\ 0 & -a & -\frac{y_2}{a} & -\frac{y_3}{a} \\ a & ax_1 & \frac{y_1}{a} + \frac{x_1 y_2}{a} & \frac{y_2}{a} + \frac{x_1 y_3}{a} \end{pmatrix}.$$

In general, we must have $q \mid a$. For the special case $a = q$, we further look at the third row to obtain $y_3 = 0$ and $y_2 = \widetilde{y}_2 q$ with $\widetilde{y}_2 \in \mathbb{Z}/q\mathbb{Z}$. The first column gives $x_3 = x_3' = \frac{1}{q}\widetilde{x}_3$ with $\widetilde{x}_3 \in \mathbb{Z}/q\mathbb{Z}$ and $x_4 = 0$. Considering the second column reveals $x_1 = \frac{1}{q}\widetilde{x}_1$ and $x_2 = \frac{1}{q}\widetilde{x}_2$ for $\widetilde{x}_1, \widetilde{x}_2 \in \mathbb{Z}/q\mathbb{Z}$. The remaining entries give the congruence conditions

$$\widetilde{x}_3 \cdot \widetilde{y}_2 + 1 \equiv 0 \bmod q,$$
$$\widetilde{x}_1 \cdot \widetilde{y}_2 + y_1 \equiv 0 \bmod q,$$
$$\widetilde{x}_3 y_1 - \widetilde{x}_1 - q\widetilde{x}_2\widetilde{y}_2 \equiv 0 \bmod q.$$

It turns out that the $\widetilde{x}_1$ sum is free, and by character orthogonality we conclude that

$$\mathrm{KL}_{\Gamma_{\mathrm{pa}}(q),s_1s_2s_1}\big((q,q^2),\mathbf{1},N\big) = 0.$$

Finally, we turn towards $w = s_1s_2s_1s_2$ and $c = (a,b)$. We compute

$$xwc^*x' = \begin{pmatrix} 1 & x_1 & x_2 & x_3 \\ 0 & 1 & x_3' & x_4 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -x_1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & -\frac{1}{a} & 0 \\ 0 & 0 & 0 & -\frac{a}{b} \\ a & 0 & 0 & 0 \\ 0 & \frac{b}{a} & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & \frac{y_1}{a} & \frac{y_2}{a} & \frac{y_3}{a} \\ 0 & 1 & \frac{ay_3'}{b} & \frac{ay_4}{b} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{y_1}{a} & 1 \end{pmatrix}$$
$$= \begin{pmatrix} ax_2 & x_2y_1 + \frac{bx_3}{a} & -\frac{1}{a} + \frac{x_1y_1}{b} + x_2y_2 + x_3y_3' & -\frac{ax_1}{b} + x_2y_3 + x_3y_4 \\ ax_3' & x_3'y_1 + \frac{bx_4}{a} & \frac{y_1}{b} + x_3'y_2 + x_4y_3' & -\frac{a}{b} + x_3'y_3 + x_4y_4 \\ a & y_1 & y_2 & y_3 \\ -ax_1 & \frac{b}{a} - x_1y_1 & y_3' - x_1y_2 & y_4 - x_1y_3 \end{pmatrix}.$$

The third entry in the first column shows that $q \mid a$. By looking at the third entry of the second column, we obtain $y_1 \in q\mathbb{Z}$. Since we are working modulo $\mathbb{Z}$, we can put $y_1 = 0$. The last entry in the second column gives rise to the condition $b/a \in \mathbb{Z}$. This forces $a \mid b$, as claimed.

Now we specialize to $a = q$ and $b = q^k$ for $k = 1,2,3$. Looking at the last two rows and the first two columns, we observe that

$$x_1 = x_3 = x_3' = y_1 = y_3 = y_3' = 0.$$

Furthermore, $y_2 \in \mathbb{Z}/q\mathbb{Z}$, $y_4 \in \mathbb{Z}/q^{k-1}\mathbb{Z}$, and $x_2 = \frac{1}{q}\widetilde{x}_2$ with $\widetilde{x}_2 \in \mathbb{Z}/q\mathbb{Z}$ and $x_4 = \frac{a}{b}\widetilde{x}_4$ with $\widetilde{x}_4 \in \mathbb{Z}/q^{k-1}\mathbb{Z}$. Finally, the last entry in the second row reveals

$$\widetilde{x}_4 y_4 \equiv 1 \bmod q^{k-1}.$$

Thus, we end up with

$$\mathrm{KL}_{\Gamma_{\mathrm{pa}}(q),s_1s_2s_1s_2}\big((q,q^k),\mathbf{1},N\big) = q^2 \sum_{x \in (\mathbb{Z}/q^{k-1}\mathbb{Z})^\times} e\left(\frac{x + N_2\overline{x}}{q^{k-1}}\right).$$

We identify the remaining x-sum as a classical Kloosterman sum. The desired estimate follows from standard bounds for the latter. ∎

Remark 3.3. The proof of Lemma 3.2 is essentially elementary. The same brute force strategy can be applied to any other lattice Γ (e.g., Klingen congruence subgroup, Borel-type congruence subgroup or principal congruence subgroup). Note that we only perform detailed computations for a restricted range of admissible moduli $c = (c_1, c_2)$. It is an interesting problem, which goes beyond the scope of this article, to determine how ramified Kloosterman sums behave for general moduli c.

4. Automorphic preliminaries

We turn to the automorphic side of the medal. After introducing the basic objects, namely, Hecke–Siegel–Maaß forms, we consider their adelic lifts and the corresponding representations. Finally, we recall Arthur's parametrization of the discrete spectrum and draw some first conclusions.

4.1. Automorphic forms and representations

We equip the space $L^2(X_\Gamma)$ with the inner product

$$\langle f, g\rangle_\Gamma = \int_{X_\Gamma} f(x)\overline{g(x)}\, dx.$$

Due to Langlands, it is well known that we have the decomposition

$$L^2(X_\Gamma) = L^2_{\mathrm{disc}}(X_\Gamma) \oplus L^2_{\mathrm{Eis}}(X_\Gamma).$$

In this note we focus exclusively on the discrete part, which further decomposes in a space of cusp forms and residues of Eisenstein series:

$$L^2_{\mathrm{disc}}(X_\Gamma) = L^2_{\mathrm{cusp}}(X_\Gamma) \oplus L^2_{\mathrm{res}}(X_\Gamma).$$

Each element $\varpi \in L^2_{\mathrm{disc}}(X_\Gamma)$, which is an eigenfunction of all invariant differential operators, comes with a spectral parameter $\mu_\varpi \in \mathbb{C}^2$. Below, in Section 4.2, we will give a finer decomposition of the discrete part based on Arthur's classification of automorphic representations for G.

The (classical) Jacquet period of $\varpi \in L^2(X_\Gamma)$ is given by

$$\mathcal{W}_\varpi(g) = \int_{U(\mathbb{Z})\backslash U(\mathbb{R})} \varpi(ug)\overline{\psi_\infty(u)}\, du.$$

If this does not vanish identically as a function on g, then we call ϖ generic. Now let ϖ be an eigenform with spectral parameter μ_ϖ. Then we can write

$$\mathcal{W}_\varpi\big(u\iota(y)k\big) = A_\varpi(\mathbf{1})W_{\mu_\varpi}(y)\boldsymbol{\psi}_\infty(u), \tag{7}$$

where $A_\varpi(\mathbf{1})$ is the first Fourier coefficient of ϖ and W_{μ_ϖ} is the standard Whittaker function as defined in [24].

Adelically, one considers functions in $L^2(Z(\mathbb{A})G(\mathbb{Q})\backslash G(\mathbb{A}))$. Note that we can identify

$$L^2\big(Z(\mathbb{A})G(\mathbb{Q})\backslash G(\mathbb{A})\big) = L^2\big(G'(\mathbb{Q})\backslash G'(\mathbb{A})\big),$$

where $G' = \mathrm{PGSp}_4$. The adelic Petersson norm is defined by

$$\langle \phi, \phi \rangle = \int_{Z(\mathbb{A})G(\mathbb{Q})\backslash G(\mathbb{A})} |\phi(g)|^2 \, d^{\mathrm{ta}}g \quad \text{for } \phi \in L^2\big(Z(\mathbb{A})G(\mathbb{Q})\backslash G(\mathbb{A})\big).$$

Further, we have the global Whittaker period given by

$$\mathcal{W}_\phi^{(\mathbb{A})}(g) = \int_{U(\mathbb{Q})\backslash U(\mathbb{A})} \phi(ug)\overline{\boldsymbol{\psi}(u)}\, du.$$

We call $\phi \in L^2(Z(\mathbb{A})G(\mathbb{Q})\backslash G(\mathbb{A}))$ an adelic lift of $\varpi \in L^2(X_\Gamma)$, and write $\phi = \phi_\varpi$ if

$$\varpi(g) = \phi_f\big(\iota_\infty(g)\big) \quad \text{for } g \in G_0(\mathbb{R}).$$

Note that ϕ_ϖ is necessarily right-K_Γ-invariant. For the paramodular group $\Gamma_{\mathrm{pa}}(q)$ of level q, the lifting procedure is well defined. For more context, we refer to [3], where the case of classical Siegel modular forms is explained in detail.

By strong approximation, we can relate the classical Jacquet period and the global Whittaker period as follows:

$$\mathcal{W}_\phi^{(\mathbb{A})}\big(\iota_\infty(g)\big) = \mathcal{W}_\varpi(g) \tag{8}$$

for $\phi = \phi_\varpi$ and $g \in G_0(\mathbb{R})$. Similarly, one gets

$$\langle \phi_\varpi, \phi_\varpi \rangle = 2\frac{\langle \varpi, \varpi \rangle_\Gamma}{\mathcal{V}(\Gamma)}. \tag{9}$$

It will be convenient to package an orthogonal basis of $L^2_{\mathrm{dis}}(X_\Gamma)$ into pieces associated to (irreducible) automorphic representations π. We write $\pi \mid X_\Gamma$ if π has non-trivial K_Γ invariant elements. The finite-dimensional space generated by these elements will be denoted by π^{K_Γ}. If π contributes to the discrete spectrum of $L^2(X_\Gamma)$, we put

$$V_\Gamma(\pi) = \big\{g \mapsto \pi\big(\iota_\infty(g)\big)v : v \in \pi^{K_\Gamma}\big\} \subseteq L^2_{\mathrm{disc}}(X_\Gamma). \tag{10}$$

We write $\mathcal{O}_\Gamma(\pi)$ for an orthogonal basis of $V_\Gamma(\pi)$.

By Flath's theorem, we can factor

$$\pi = \pi_\infty \otimes \bigotimes_p \pi_p$$

into local representations. In complete analogy to the global case, we write $\pi_p^{K_{\Gamma,p}}$ for the space generated by right-$K_{\Gamma,p}$-invariant elements of π_p, and we let $\mathcal{O}_{K_{\Gamma,p}}(\pi_p)$ be an orthogonal basis of $\pi_p^{K_{\Gamma,p}}$. Note that if $p \nmid D_\Gamma$, then $\pi_p^{K_{\Gamma,p}}$ is generated by the (up to scaling) unique spherical element v_p°. Thus, globally

$$\left\{ v_\infty^\circ \otimes \bigotimes_{p \nmid D_\Gamma} v_p^\circ \otimes \bigotimes_{p \mid D_\Gamma} v_p : v_p \in \mathcal{O}_{K_{\Gamma,p}}(\pi_p) \text{ for } p \mid D_\Gamma \right\}$$

can be taken as an orthogonal basis of π^{K_Γ}. We will always choose $\mathcal{O}_\Gamma(\pi)$ by descending the above basis of π^{K_Γ} to $V_\Gamma(\pi)$.

Note that each $\varpi \in V_\Gamma(\pi)$ is an eigenfunction of all invariant differential operators (and even of all but finitely many Hecke-operators). The spectral parameter of ϖ only depends on π_∞. Thus, we can write

$$\mu_\varpi = \mu(\pi_\infty).$$

Indeed, since π_∞ is spherical, it can be realized as Langlands' quotient of an induced representation

$$\pi_\infty = L(\chi_1 \times \chi_2 \rtimes \sigma).$$

Note that by assumption on the central character χ, we have $\chi_1\chi_2\sigma^2 = \chi_\infty \in \{\mathbf{1}, \mathrm{sgn}\}$. We write $\chi_i = \mathrm{sgn}^{\rho_i} |\cdot|^{\alpha_i}$ and $\sigma = \mathrm{sgn}^{\rho'} |\cdot|^{\beta}$. Note that we have $\alpha_1 + \alpha_2 + 2\beta = 0$. We associate the spectral parameter

$$\mu(\pi_\infty) = \left(\frac{\alpha_1 + \alpha_2}{2}, \frac{\alpha_1 - \alpha_2}{2} \right) \in \mathbb{C}^2.$$

Note that the transfer of π_∞ to GL_4 is self-dual and has spectral parameter

$$\left(\frac{\alpha_1 + \alpha_2}{2}, \frac{\alpha_1 - \alpha_2}{2}, \frac{-\alpha_1 + \alpha_2}{2}, \frac{-\alpha_1 - \alpha_2}{2} \right) \in \mathbb{C}^4.$$

With this notation at hand, we can express

$$N_\Gamma(\sigma; M) = \sum_{\substack{\pi \mid X_\Gamma, \\ \|\mu(\pi_\infty)\| \leq M, \\ \sigma(\pi_\infty) \geq \sigma}} \dim_{\mathbb{C}}\big(V_\Gamma(\pi)\big).$$

Recall that π is called (globally) generic if the global Whittaker period $\mathcal{W}_\phi^{(\mathbb{A})}$ is non-zero (as function on $G(\mathbb{A})$) for some element ϕ of π. On the other hand, we call π abstractly generic (or locally everywhere generic) if π_∞ and π_p are generic for all p in the sense that they admit a non-trivial Whittaker functional. According to [2, Proposition 8.3.2], π is generic if and only if π is abstractly generic.[3]

To each generic representation π_p we fix a Whittaker functional leading to an isomorphism $v \mapsto W_v$ between π_p and its Whittaker model $\mathcal{W}(\pi_p, \psi_p)$. If π_p is unramified we fix (once and for all) the normalization such that

$$W_{v_p^\circ}(1) = 1.$$

We want to make similar accommodations at the archimedean place. To do so we note that if π_∞ is generic and spherical, then π_∞ is a principal series. This is [11, Proposition 3.4.1]. In particular, we have

$$\pi_\infty|_{G_0(\mathbb{R})} = \mathrm{Ind}_{B_0(\mathbb{R})}^{G_0(\mathbb{R})}\big(|\cdot|^{\mu_1} \boxtimes |\cdot|^{\mu_2}\big),$$

where $\mu(\pi_\infty) = (\mu_1, \mu_2)$. This is the setting also described in [10]. We normalize the Whittaker functional so that

$$W_{v_\infty^\circ}(1) = W_{\mu(\pi_\infty)}(1).$$

If $\varpi \in \mathcal{O}_\Gamma(\pi)$ corresponds to $v_\infty^\circ \otimes \bigotimes_{p\nmid D_\Gamma} v_p^\circ \otimes \bigotimes_{p\mid D_\Gamma} v_p$, then by uniqueness of the Whittaker functional, we must have

$$\begin{aligned}\frac{|A_\varpi(1)W_{\mu_\varpi}(1)|^2}{\langle \varpi, \varpi\rangle} = \frac{|\mathcal{W}_\varpi(1)|^2}{\langle \varpi, \varpi\rangle} &= \frac{|\mathcal{W}_{\phi_w}^{(\mathbb{A})}(1)|^2}{\mathcal{V}(\Gamma)\cdot\langle \phi_\varpi, \phi_\varpi\rangle} \\ &= \frac{C(\pi)}{\mathcal{V}(\Gamma)} \cdot \frac{|W_{\mu(\pi_\infty)}(1)|^2 \cdot \prod_{p\mid D_\Gamma} |W_{v_p}(1)|^2}{\langle \phi_\varpi, \phi_\varpi\rangle}.\end{aligned}$$

The constant (as well as the normalization of the Whittaker functionals at the places where $p \mid D_\Gamma$) is usually determined via the Rankin–Selberg method. Unfortunately, there are some hurdles when carrying this out in generally. These can be resolved under assumption of a conjecture from Lapid and Mao given in [29]. Here we give a formulation of this conjecture along the lines of [35, Section 4] adapted to our notation.

Conjecture 4.1 (Lapid–Mao). *Let $\pi = \pi_\infty \otimes \bigotimes_p \pi_p$ be an irreducible, unitary, generic cuspidal automorphic representation with trivial central character. Suppose*

[3] The implication "globally generic" implies everywhere locally generic is of course essentially trivial. The other direction, however, is highly non-trivial. As an alternative to Arthur's seminal work, one can argue as in [42, (6.8)] to establish it.

that $\pi \mid X_\Gamma$. *Then, for* $\varpi \in \mathcal{O}_\Gamma(\pi)$, *we have*

$$\frac{|A_\varpi(1)|^2}{\langle\varpi,\varpi\rangle} = \frac{2^{5-c}}{\mathcal{V}(\Gamma)} \cdot \frac{\Lambda^{D_\Gamma}(2)\Lambda^{D_\Gamma}(4)}{\Lambda^{D_\Gamma}(1,\pi,\mathrm{Ad})} \cdot \prod_{p\mid D_\Gamma} J_{\pi_p}(v_p),$$

where $c = 1$ *if* π *is stable (i.e., of general type) and* $c = 2$ *if* π *is endoscopic. Furthermore,* $\Lambda^{D_\Gamma}(s,\pi,\mathrm{Ad})$ *is the completed adjoint* L*-function of* π *with Euler factors at primes dividing* D_Γ *omitted. (Similarly,* Λ^{D_Γ} *is the completed Riemann zeta function with Euler factors at* $p \mid D_\Gamma$ *omitted.) Finally, the local period* $J_{\pi_p}(v_p)$ *is defined by*

$$J_{\pi_p}(v_p) = \int_{U(\mathbb{Q}_p)}^{\mathrm{st}} \frac{\langle\pi_p(u)v_p, v_p\rangle}{\langle v_p, v_p\rangle}\overline{\psi_p(u)}\,du.$$

A special case of Conjecture 4.1 was solved in [10]. Since their result is important for the unconditional part of Theorem 1.1, we recall it here.

Theorem 4.2 ([10, Theorem 2.1]). *Let* $q \geq 1$ *be square-free, then we have*

$$\frac{|A_\varpi(\mathbf{1})|^2}{\langle\varpi,\varpi\rangle_{\Gamma_{\mathrm{pa}}(q)}} = \frac{2^{5-c}}{\mathcal{V}(\Gamma_{\mathrm{pa}}(q))} \cdot \frac{\Lambda^q(2)\Lambda^q(4)}{\Lambda^q(1,\pi,\mathrm{Ad})} \cdot \prod_{p\mid q} \frac{p}{(1-p^{-2})^2}$$

for $\varpi \in \mathcal{O}_{\Gamma_{\mathrm{pa}}(q)}(\pi)$ *and* $\pi \mid X_{\Gamma_{\mathrm{pa}}(q)}$ *generic.*

Proof. To see how this follows from [10, Theorem 2.1], we first combine (7) with (8) to see that

$$\mathcal{W}_{\phi_\varpi}^{(\mathbb{A})}(1) = \mathcal{W}_\varpi(1) = A_\varpi(\mathbf{1})W_{\mu_\varpi}(1).$$

Note that our $\mathcal{W}_{\phi_\varpi}^{(\mathbb{A})}$ is simply denoted by W in [10, Section 2]. We also recall (9) and note that $\langle\phi_\varpi,\phi_\varpi\rangle$ is normalized, as in [10, Section 2].

Next we observe that since $\varpi \in \mathcal{O}_\Gamma(\pi)$ our form ϕ_ϖ has locally the transformation behavior described in [10, (2.4)]. In particular, we can use [10, Theorem 2.1] to compute the Petersson norm of ϖ. Indeed, using notation from [10], we obtain

$$\begin{aligned}\frac{|A_\varpi(\mathbf{1})W_{\mu_\varpi}(1)|^2}{\langle\varpi,\varpi\rangle_{\Gamma_{\mathrm{pa}}(q))}} &= \frac{2}{\mathcal{V}(\Gamma_{\mathrm{pa}}(q))} \cdot \frac{|\mathcal{W}_{\phi_\varpi}^{(\mathbb{A})}(1)|^2}{\langle\phi_\varpi,\phi_\varpi\rangle} \\ &= \frac{2}{\mathcal{V}(\Gamma_{\mathrm{pa}}(q))} \cdot |W_{\mu_\varpi}(1)|^2 \cdot \left(2^c \frac{L(1,\pi,\mathrm{Ad})}{\Delta_{\mathrm{PGSP}_4}} \prod_v C(\pi_v)\right)^{-1}. \qquad (11)\end{aligned}$$

Taking the quotient makes the expression scaling invariant, so that there is no problem in using [10, (2.5) and (2.6)] to compute it. We are done after dividing both sides by $|W_{\mu_\varpi}(1)|^2$ and explicating the right hand side of (11) using the exact values for $C(\pi_v)$ given in the statement of [10, Theorem 2.1]. ∎

For more general Γ, estimating the first Fourier coefficient from below is a hard problem. By assuming Conjecture 4.1, it reduces to a purely local computation, which is still hard. We are, however, tempted to conjecture that

$$\prod_{p \mid D_\Gamma} \sum_{v_p \in \mathcal{O}_{K_{\Gamma,p}}(\pi_p)} J_{\pi_p}(v_p) \asymp \dim V_\pi(\Gamma) \cdot \mathcal{N}(\Gamma) \tag{12}$$

for generic $\pi = \otimes_v \pi_v$. Of course, this is speculation, but we have the following *evidence*:

- For $\Gamma = \Gamma_{\mathrm{pa}}(q)$ with square-free q, this follows from the main result of [10]. Furthermore, if π_p is a simple supercuspidal representation (this forces $p^5 \mid q$), the local integral $J_{\pi_p}(v_p)$ has recently been computed for the local new-vector v_p in [35].
- For congruence subgroups associated to parabolic subgroups (i.e., Siegel congruence subgroups, Klingen congruence subgroups or Borel-type congruence subgroups) of square-free level local computations in the spirit of [14], lead to the desired result.
- For the principal congruence subgroup of square-free level, the arguments from [6] can be adapted to prove (12) in the setting at hand. Note that carrying this out requires a little bit of care, since not all depth zero supercuspidal representations are generic.

4.2. The parametrization of the discrete spectrum

The parametrization of the discrete spectrum of GSp_4 in terms of discrete automorphic forms of GL_n has been announced in [1]. Building on the seminal monograph [2] this has now been established in [20]. This parametrization will be important for us in order to upgrade the density estimate for spherical, generic representations to a density theorem for the full (spherical) discrete spectrum. This will require a good understanding of the local constituents of the Arthur packets. For trivial central character the representations factor through $G' = \mathrm{PGSp}_4$ and an explicit parametrization was worked out in [40, 41]. We will follow the notation within these references.

Using the exceptional isomorphism $\mathrm{SO}_5 \cong G'$, we can express the central result from [1] as

$$L^2_{\mathrm{disc}}\big(G'(\mathbb{Q})\backslash G'(\mathbb{A})\big) \equiv \bigoplus_{\psi \in \boldsymbol{\psi}_2(G')} \bigoplus_{\pi \in \Pi_\psi : \langle \cdot, \pi \rangle = \varepsilon_\psi} \pi. \tag{13}$$

See also [40, (1.2)]. We will introduce the missing notation on the way of gathering all the necessary information contained in (13). In the following, we will classify the parameters ψ by type. Relevant types are **(G)**, **(Y)**, **(Q)**, **(P)**, **(B)**, and **(F)**. Accordingly, we decompose

$$L^2_{\mathrm{disc}}\big(G'(\mathbb{Q})\backslash G'(\mathbb{A})\big) = \bigoplus_{*\in\{\mathbf{G},\mathbf{Y},\mathbf{Q},\mathbf{P},\mathbf{B},\mathbf{F}\}} L^2_{(*)}\big(G'(\mathbb{Q})\backslash G'(\mathbb{A})\big).$$

We will now spend the rest of this section describing the Arthur parameters of each type in detail.

First, the set $\boldsymbol{\psi}_2(G')$ consists of formal expressions

$$\psi = (\mu_i \boxtimes \nu_1) \boxplus \cdots \boxplus (\mu_r \boxtimes \nu_r),$$

where μ_i is a self-dual, cuspidal automorphic representation of $\mathrm{GL}_{m_i}(\mathbb{A})$ and ν_i is the irreducible representation of $\mathrm{SL}_2(\mathbb{C})$ of dimension n_i. Furthermore, the following conditions need to be satisfied:

(1) $m_1 n_1 + \cdots + m_r n_r = 4$;

(2) $\mu_i \boxtimes \nu_i \neq \mu_j \boxtimes \nu_j$ for $i \neq j$;

(3) if n_i is odd (resp. even) then μ_i is symplectic (resp. orthogonal).

Analyzing these conditions as in [40] leads to different types of parameters. They are as follows:

(1) General type (**G**): $\psi = \mu \boxtimes 1$ for a self-dual symplectic (i.e., $L(s,\mu,\wedge^2)$ has a pole at $s = 1$) unitary cuspidal automorphic representation of $\mathrm{GL}_4(\mathbb{A})$.

(2) Yoshida type (**Y**): $\psi = (\mu_1 \boxtimes 1) \boxplus (\mu_2 \boxtimes 1)$ for distinct unitary cuspidal automorphic representations of $\mathrm{GL}_2(\mathbb{A})$ with trivial central character.

(3) Soundry type (**Q**): $\psi = \mu \boxtimes \nu(2)$ for a self-dual unitary cuspidal automorphic representation μ of $\mathrm{GL}_2(\mathbb{A})$ with non-trivial central character ω_μ. ($\nu(2)$ is the two-dimensional irreducible representation of $\mathrm{SL}_2(\mathbb{C})$.)

(4) Saito–Kurokawa type (**P**): $\psi = (\mu \boxtimes 1) \boxplus (\sigma \boxtimes \nu(2))$ for a unitary cuspidal automorphic representation μ of $\mathrm{GL}_2(\mathbb{A})$ with trivial central character and a quadratic Hecke character σ.

(5) Howe–Piatetski–Shapiro type (**B**): $\psi = (\chi_1 \boxtimes \nu(2)) \boxplus (\chi_2 \boxtimes \nu(2))$ for two distinct quadratic Hecke characters.

(6) Finite type (**F**): $\psi = \xi \boxtimes \nu(4)$ for a quadratic Hecke character ξ. ($\nu(4)$ is the four-dimensional irreducible representation of $\mathrm{SL}_2(\mathbb{C})$.)

To each packet we attach the group $\mathcal{S}_\psi$. For all practical purposes, it is sufficient to know that $\mathcal{S}_\psi = 1$ if ψ is of type (**G**), (**Q**), or (**F**). For the remaining types, namely (**Y**), (**P**), and (**B**), we have $\mathcal{S}_\psi = \{\pm 1\}$.

The global parameter ψ has a localization at each place $v \in \{\infty, 2, 3, \ldots\}$. Formally, these are maps

$$\psi_v\colon L_{\mathbb{Q}_v} \times \mathrm{SU}_2 \to \mathrm{Sp}_4(\mathbb{C}),$$

satisfying certain compatibility conditions. In our setting they can be given quite explicitly in terms of the local Langlands parameters of the GL_n objects that appear in the corresponding global parameter. We omit their explicit description and refer to [40, Section 1.2] for details. As in the global case, each local parameter comes with a local centralizer group $\mathcal{S}_{\psi_v}$, and we have canonical maps

$$\mathcal{S}_\psi \to \mathcal{S}_{\psi_v}.$$

Now we attach to each local parameter a finite packet of admissible representations Π_{ψ_v} as in [2, Theorem 1.5.1]. Each packet comes with a canonical map

$$\Pi_{\psi_v} \ni \pi_v \mapsto \langle \cdot, \pi_v \rangle \in \widehat{\mathcal{S}}_{\psi_v}.$$

Note that if π_v is unramified, we have $\langle \cdot, \pi_v \rangle = 1$. The global packet is then given by

$$\Pi_\psi = \{\pi = \oplus_v \pi_v : \pi_v \in \Pi_{\psi_v}\}.$$

This is all we want to say about the parameters ψ in general. We now turn towards the individual analysis of packets for the different types.

4.2.1. Parameters of type (G) and (Y). For a parameter ψ of type **(G)** or **(Y)**, the local parameter ψ_v is trivial on SU_2. We can therefore interpret it as a classical L-parameter and it turns out that the local packet Π_{ψ_v} coincides with the packets defined by the local Langlands correspondence. See [19] for the latter in the case of GSp_4. In particular, the elements of Π_{ψ_v} are irreducible and unitary.

Lemma 4.3. *Let ψ be a parameter of type* **(G)** *or* **(Y)**. *Then Π_ψ contains precisely one generic member π_ψ^{gen}. For $q > 1$ arbitrary, we have*

$$\sum_{\pi \in \Pi_\psi : \langle \cdot, \pi \rangle = \varepsilon_\psi} \dim V_{\Gamma_{\mathrm{pa}}(q)}(\pi) = \dim V_{\Gamma_{\mathrm{pa}}(q)}(\pi_\psi^{\mathrm{gen}}). \tag{14}$$

Proof. That Π_ψ contains precisely one generic member follows from [2, Proposition 8.3.2]. (See also [40, Theorem 1.1].) Note that each L-packet for GSp_4 has at most two elements. Of course, it suffices to check those L-packets with exactly two elements. These are

$$\{\mathrm{VIa}, \mathrm{VIb}\}, \quad \{\mathrm{VIIIa}, \mathrm{VIIIb}\}, \quad \{\mathrm{Va}, \mathrm{Va}^*\}, \quad \{\mathrm{XIa}, \mathrm{XIa}^*\}. \tag{15}$$

The desired equality follows if we can show that

$$\dim_{\mathbb{C}} (\pi')^{K_{\Gamma_{\mathrm{pa}}(q),p}} = 0, \tag{16}$$

where $\Pi_{\psi_p} = \{\pi_{\psi_p}^{\mathrm{gen}}, \pi'\}$ is one of the L-packets given in (15). For the paramodular group $\Gamma_{\mathrm{pa}}(q)$ and $p \mid q$, this follows from [40, Theorem 1.1]. A related argument can be found in the proof of [40, Lemma 2.5]. ■

Remark 4.4. The exact equality (14) is special for the paramodular group. Indeed, for more general Γ, the vanishing result (16) will fail. As a consequence the global Arthur packets of type **(G)** and **(Y)** will have non-generic members that contribute to the cuspidal spectrum of X_Γ. However, we expect the estimate

$$\sum_{\pi \in \Pi_\psi : \langle \cdot, \pi \rangle = \varepsilon_\psi} \dim V_\Gamma(\pi) \ll 2^{\omega(D_\Gamma)} \cdot \dim V_\Gamma(\pi_\psi^{\mathrm{gen}}),$$

to hold in great generality. For example, by using [36, Table A.15], one can easily verify this for Siegel congruence subgroups, Klingen congruence subgroups and Borel-type congruence subgroups. Furthermore, a similar estimate can be given for the principal congruence subgroup. The corresponding local dimensions have been computed in [9].

4.2.2. Parameters of type (B) and (Q). It turns out that packets of these types do not contribute to the spectrum of $X_{\Gamma_{\mathrm{pa}}(q)}$. This is due to [41, Proposition 5.1]. We record this fact formally in the form of the following lemma.

Lemma 4.5. *Let ψ be an Arthur parameter of type* **(B)** *or* **(Q)** *and let $q > 1$ be arbitrary. We have $V_{\Gamma_{\mathrm{pa}}(q)}(\pi) = \{0\}$ for all $\pi \in \Pi_\psi$.*

Proof. Suppose there is $\pi \in \Pi_\psi$ with $V_{\Gamma_{\mathrm{pa}}(q)}(\pi) \neq \{0\}$. Then π would be paramodular at every finite place, which is impossible by [41, Proposition 5.1]. ■

Remark 4.6. One should not expect a vanishing result as in Lemma 4.5 to hold for arbitrary families of congruence lattices. For example, if Γ is a principal congruence subgroups, then packets of both types will contribute to the discrete spectrum of X_Γ. It should be noted that in this case one encounters representations with non-trivial central character. In particular, one has to use [20] and the explicit description of the local packets cannot be directly obtained from [41]. For the types **(B)** and **(P)**, this can be resolved using [34], but for packages of type **(Q)** a little more work seems to be necessary.[4]

4.2.3. Packets of type (P). Fix a parameter $\psi = (\mu \boxtimes 1) \boxplus (\sigma \boxtimes \nu(2))$ of type **(P)** given by a unitary cuspidal automorphic representation μ with trivial central character and a quadratic Hecke character σ. The global base point π_ψ^+, as defined below [41, (12)], is the isobaric constituent of the globally induced representation

$$|\cdot|^{1/2}\sigma\mu \ltimes |\cdot|^{-1/2}\sigma.$$

[4] I would like to thank R. Schmidt for pointing me to the result in [34] and for other very useful tips concerning the (local) Arthur packets in the presence of a non-trivial central character.

Note that here $\varepsilon_\psi = \varepsilon(1/2, \sigma^{-1} \otimes \mu)$, so that we have the global compatibility condition

$$\prod_v \varepsilon(\pi_v) = \langle -1, \pi \rangle = \varepsilon\Big(\frac{1}{2}, \sigma^{-1} \otimes \mu\Big).$$

The local packets depend on the factorization $\mu = \otimes_v \mu_v$. We briefly summarize [41, Table 2] in our Table 2. (Note that we have written the local base point as first entry in each packet.)

μ_v	$\chi \times \chi^{-1}$	$\chi \cdot \mathrm{St}_{\mathrm{GL}_2}$	$\sigma_v \cdot \mathrm{St}_{\mathrm{GL}_2}$	supercuspidal
Π_{ψ_v}	{IIb}	{Vb, Va*}	{VIc, VIb}	{XIb, XIa*}

Table 2. Summary of [41, Table 2].

A finer investigation of the properties of packets of type (**P**) will naturally lead us to related properties of the underlying GL_2 representation μ. In this context, we write $\Gamma_H(q)$ for the standard Hecke congruence subgroup of level q. In analogy with (10) we can define the space $V_{\Gamma_H(q)}(\mu)$ in the GL_2-setting. (Here the classical upper half plane $\mathbb{H}$ takes over the role of $\mathbb{H}_2$.) The dimension of $V_{\Gamma_H(q)}(\mu)$ is well understood by the newform theory of Atkin and Lehner. It depends only on the local conductor exponent of μ_v, which we denote by $a(\mu_v)$. The (global) conductor of μ is given by $c(\mu) = \prod_p p^{a(\mu_p)}$. We are now ready to prove the following result.

Lemma 4.7. *Let* $\psi = (\mu \boxtimes 1) \boxplus (\sigma \boxtimes \nu(2))$ *be an Arthur parameter of type* (**P**). *For* $q > 1$ *arbitrary, we have*

$$\sum_{\substack{\pi \in \Pi_\psi \\ \langle -1,\pi\rangle = \varepsilon(1/2,\mu)}} \dim V_{\Gamma_{\mathrm{pa}}(q)}(\pi) = \delta_{\varepsilon(1/2,\mu)=1} \dim V_{\Gamma_{\mathrm{pa}}(q)}(\pi_\psi^+).$$

Furthermore, we have $\dim V_{\Gamma_{\mathrm{pa}}(q)}(\pi_\psi^+) \leq \dim V_{\Gamma_H(q)}(\mu) \ll q^\varepsilon$ *and* $V_{\Gamma_{\mathrm{pa}}(q)}(\pi_\psi^+) = \{0\}$, *unless the conductor of* μ *divides* q *and* σ *is trivial. Finally, if* $\varpi \in V_{\Gamma_{\mathrm{pa}}(q)}(\pi_\psi^+)$ *is non-zero, then* $\sigma_\varpi = 1/2$.

Proof. We can follow the proof of [41, Proposition 5.2] very closely. The main difference for us is that, since $\pi \mid X_{\Gamma_{\mathrm{pa}}(q)}$, the representation π_∞ must be spherical. In particular, we automatically have $\pi_\infty = (\pi_\psi^+)_\infty$. More precisely, if $\pi \in \Pi_\psi$ is spherical, then $\mu_\infty = \chi \times \chi^{-1}$ (i.e., μ is of Maaß type). Looking at [41, Table 2], we find that

$$\pi_\infty = \chi\sigma 1_{\mathrm{GL}_2} \ltimes \chi^{-1}.$$

The associated real part of the spectral-parameter is $\Re\mu(\pi_\infty) = (\beta, 1/2)$, where $|\chi(x)| = |x|^\beta$. Note that $\beta = \sigma(\mu_\infty)$ and $\beta < 1/2$. In particular, we have $\sigma(\pi_\infty) = 1/2$.

We continue as in [41] by noting that, if there is a finite place v where σ is ramified, then no representation in Π_{ψ_v} has vectors fixed by the paramodular subgroup. Thus, we can assume that σ is trivial. In this case, we see that only the local base point at finite places has vectors fixed by the paramodular group. In other words, $\dim V_{\Gamma_{\mathrm{pa}}(q)}(\pi) \neq 0$ implies that $\pi = \pi_\psi^+$. But $\langle -1, \pi_\psi^+ \rangle = 1$, so that the parity condition forces $\varepsilon(1/2, \mu) = 1$. Looking closer at [41, Table 2], we find that:

- If $\mu_v = \chi \times \chi^{-1}$, then $(\pi_\psi^+)_v$ is of type IIb. More precisely, in the notation of [36] we have $(\pi_\psi^+)_v = \chi\sigma_v 1_{\mathrm{GL}_2} \ltimes \chi^{-1}$. Recall that σ_v is unramified, so that $a(\sigma_v\chi) = a(\chi)$. Finally, note that $a(\mu_v) = 2a(\chi)$. In this case the σ in [36, Table A.12] corresponds to our χ^{-1}.
- If $\mu_v = \chi\,\mathrm{St}_{\mathrm{GL}_2}$ for $\chi \neq \sigma_v$ quadratic, then $(\pi_\psi^+)_v = L(\nu^{1/2}\chi\sigma_v\,\mathrm{St}_{\mathrm{GL}_2}, \nu^{-1/2}\sigma_v)$ is of type Vb. Here we have $a(\mu_v) = 2$. In this case the ξ in [36, Table A.12] corresponds to our $\chi\sigma_v$.
- If $\mu_v = \sigma_v\,\mathrm{St}_{\mathrm{GL}_2}$, then $(\pi_\psi^+)_v = L(\nu^{1/2}\,\mathrm{St}_{\mathrm{GL}_2}, \nu^{-1/2}\sigma_v)$ is of type VIc. Since σ_v is unramified, one has $a(\mu_v) = 1$. In this case, the σ in [36, Table A.12] corresponds to our σ_v.
- If μ_v is supercuspidal, then $(\pi_\psi^+)_v = L(\nu^{1/2}\sigma_v\nu_v, \nu^{-1/2}\sigma_v)$ is of type XIb. By unramifiedness of σ_v, we have $a(\sigma_v\mu_v) = a(\mu_v)$. In this case, π (resp. σ) in [36, Table A.12] corresponds to our $\sigma_v\mu_v$ (resp. σ_v).

Using these observations, one can study [36, Table A.12] to obtain

$$\dim(\pi_{\psi_v}^+)^{K_{\Gamma_{\mathrm{pa}}(q),v}} = \left\lfloor \frac{v(q) - a(\mu_v)}{2} \right\rfloor.$$

This completes the proof. ∎

Remark 4.8. A Similar result can be established for other lattice families as well. The most important ingredients to generalize the argument are the corresponding formulae for the dimensions of local fixed vectors. For example, if Γ is a congruence subgroup associated to a parabolic (e.g., Siegel congruence subgroup, Klingen congruence subgroup or Borel-type congruence subgroup) with square-free level, then these can be found in [36, Table A.15]. For principal congruence subgroups, the relevant local dimension formulae are given in [9].

5. The density theorem

After gathering all the ingredients, we are now ready to prove our density theorem. We do so in two steps. First, we use the Kuznetsov formula to establish a bound for the generic part of the discrete spectrum. Second, we account for the missing pieces by hand.

5.1. The generic contribution

As mentioned previously in the introduction, the generic spectrum can be treated using a Kuznetsov formula following the strategy pioneered in [7]. The argument was adapted to Siegel congruence subgroups of $\mathrm{Sp}_4(\mathbb{Z})$ in [31]. We will start by discussing the overall strategy of this argument for a quite general class of lattices Γ. Later we will restrict our attention to the paramodular groups $\Gamma_{\mathrm{pa}}(q)$, where we can carry out all the details.

Remark 5.1. Unfortunately, it seems that the proof of the density theorem given in [31] has the following gap. As we will see below the non-vanishing of the first Fourier coefficient $A_{\varpi}(\mathbf{1})$ for generic Siegel–Maaß cusp forms is crucial to the approach using the Kuznetsov formula. For the Siegel congruence subgroup, this is stated in [31, (6.3)] and is supposed to follow from [10, Theorem 1.1]. However, [10] works with the paramodular group in place of the Siegel congruence subgroup. It is not clear to us if and with how much work the methods form [10] adapt to generic Siegel–Maaß cusp forms for the Siegel congruence subgroup. An alternative would be to establish the desired estimate conditional on the conjecture of Lapid–Mao and some extensive local computations. See Conjecture 4.1, (12), and the discussion below.

It is easy to see that the density result follows from the weighted estimate

$$\sum_{\substack{\pi|X_\Gamma,\\ \text{generic}}} \dim_{\mathbb{C}} V_\Gamma(\pi)\cdot Z^{2\sigma_\pi} \ll \mathcal{V}(\Gamma)^{1+\varepsilon} \tag{17}$$

for sufficiently large Z. We call the value of Z necessary to establish the density hypothesis $Z_0(\Gamma)$. It is given by

$$Z_0(\Gamma) = \mathcal{V}(\Gamma)^{1/2\sigma_1} = \mathcal{V}(\Gamma)^{1/3}.$$

If the estimate holds for larger values of Z, then one obtains a subconvex density theorem. For example, we have $Z_0(\Gamma_{\mathrm{pa}}(q)) = q^{2/3}$, since $\mathcal{V}(\Gamma_{\mathrm{pa}}(q)) = q^2$.

We will now recall the Kuznetsov formula from [31, Section 7] and modify it slightly to apply to all lattices Γ in question. Note that one can alternatively use the relative trace formula developed in [11].

For the following argument, we assume that $U(\mathbb{Z}) \subseteq \Gamma$. Given a function

$$E\colon \mathbb{R}_+^2 \to \mathbb{C}$$

with compact support and a parameter $X \in \mathbb{R}_+^2$, we set

$$E^{(X)}(y_1, y_2) = E(X_1 y_1, X_2 y_2).$$

This function is then lifted to a function $F^{(X)}$ on $\mathrm{Sp}_4(\mathbb{R})$ by

$$F^{(X)}(xyk) = \psi(x) E^{(X)}\big(\mathrm{y}(y)\big) \quad \text{for } x \in U(\mathbb{R}),\, y \in T(\mathbb{R}_+),\, k \in K_\infty.$$

We are now ready to define the Poincaré series

$$P_\Gamma^{(X)}(g) = \sum_{\gamma \in U(\mathbb{Z})\backslash\Gamma} F^{(X)}(\gamma g) \quad \text{for } g \in \mathrm{Sp}_4(\mathbb{R}).$$

The usual unfolding trick shows that

$$\langle \varpi, P_\Gamma^{(X)} \rangle_\Gamma = A_\varpi(\mathbf{1}) \langle W_\mu, E^{(X)} \rangle.$$

On the other hand, we can compute the Fourier coefficient of the Poincaré series directly using the Bruhat decomposition. One gets

$$\begin{aligned} &\int_{U(\mathbb{Z})\backslash U(\mathbb{R})} P_\Gamma^{(X)}(xg) \overline{\psi_\infty(x)}\, dx \\ &\qquad = \sum_{w \in W} \sum_{c \in \mathbb{N}^2} \mathrm{KL}_{\Gamma,w}(c; 1, 1) \cdot \int_{U_w(\mathbb{R})} F^{(X)}(c^* w x g) \overline{\psi_\infty(x)}\, dx. \end{aligned}$$

Taking the inner product of two Poincaré series and applying Parseval gives the following lemma.

Lemma 5.2. *Let $Z \in \mathbb{R}_+$ and let $E\colon \mathbb{R}_+^2 \to \mathbb{C}$ be compactly supported. Then we have*

$$\begin{aligned} &\sideset{}{'}\sum_{\substack{\pi | X_\Gamma, \\ \text{generic}}} \sum_{\varpi \in \mathcal{O}_\Gamma(\pi)} \frac{|A_\varpi(\mathbf{1})|^2}{\langle \varpi, \varpi \rangle_\Gamma} \cdot Z^{2\sigma_\pi} \ll_M \sum_{w \in W} \sum_{c \in \mathbb{N}^2} \frac{\mathrm{KL}_{\Gamma,w}(c; 1, 1)}{c_1 c_2} \\ &\quad \cdot \mathrm{y}(A_c)^{-\eta} \int_{T(\mathbb{R}_+)} \int_{U_w(\mathbb{R})} F^{(X)}\big(\iota^{-1}(X)^{-1} A_c w x y\big) \psi_\infty^{(X^{-1})}(x^{-1}) \overline{E\big(\mathrm{y}(y)\big)}\, dx\, d^*y \end{aligned}$$

for $A_c = \iota(X) c^* w \iota(X)^{-1} w^{-1} \in T(\mathbb{R}_+)$ and $X = (1, Z)$.

Proof. This follows directly from a version of [31, Lemma 7.1] for Γ after inserting [31, Lemma 5.2] and dropping the continuous spectrum. Note that in [31, Lemma 7.1] the factor $X^{2\eta} = Z^{2\eta_2}$ appears on the geometric side. However, the same factor enters the spectral side via the estimate [31, Lemma 5.2] and the contributions cancel out. ∎

The geometric side can further be brought in the following form.

Lemma 5.3. *For $Z \in \mathbb{R}_+$, we have*

$$\sideset{}{'}\sum_{\substack{\pi \mid X_\Gamma,\\ \text{generic}}} \sum_{\varpi \in \mathcal{O}_\Gamma(\pi)} \frac{|A_\varpi(\mathbf{1})|^2}{\langle \varpi, \varpi \rangle} \cdot Z^{2\sigma_\pi} \ll_{M,\varepsilon} \mathcal{N}(\Gamma) + Z^\varepsilon \sum_{\substack{1 \neq w \in W,\\ \text{relevant}}} S_\Gamma(w; Z)$$

for

$$\begin{aligned} S_\Gamma(s_1 s_2 s_1; Z) &= \sum_{\substack{c=(c_1,c_1)\in\mathbb{N}^2,\\ c_1 \ll Z}} \frac{\mathrm{KL}_{\Gamma,w}(c; 1, 1)}{c_1^2}, \\ S_\Gamma(s_2 s_1 s_2; Z) &= \sum_{\substack{c=(c_1,c_1^2)\in\mathbb{N}^2,\\ c_1 \ll Z}} \frac{\mathrm{KL}_{\Gamma,w}(c; 1, 1)}{c_1^3}, \\ S_\Gamma(s_1 s_2 s_1 s_2; Z) &= \sum_{\substack{c=(c_1,c_2)\in\mathbb{N}^2,\\ c_1 \ll Z, c_2 \ll Z^2}} \frac{\mathrm{KL}_{\Gamma,w}(c; 1, 1)}{c_1 c_2}. \end{aligned} \tag{18}$$

Proof. First, we note that for (w, c) non-admissible the corresponding Kloosterman sum vanishes. Thus we drop those tuples from the geometric side. Furthermore, the contribution of $w = 1$ is given by (5).

We consider the remaining cases. Let $w \in W$ be a non-trivial relevant Weyl element. The corresponding part of the geometric side of the Kuznetsov formula is estimated as follows. We first apply [31, Lemma 3.1] as in [31, Section 8] to truncate the c-sum. The result is the contribution

$$\sum_{\substack{c\in\mathbb{N}^2,\\ c_1 \ll Z, c_2 \ll Z^{\kappa(w)}}} \frac{\mathrm{KL}_{\Gamma,w}(c; 1, 1)}{c_1 c_2} \cdot \mathrm{y}(A_c)^{-\eta} \int_{T(\mathbb{R}_+)} \int_{U_w(\mathbb{R})} \Big(F^{(X)}\big(\iota^{-1}(X)^{-1} A_c w x y\big) \cdot \psi_\infty^{(X^{-1})}(x^{-1}) \overline{E\big(\mathrm{y}(y)\big)} \Big)\, dx\, d^*y \tag{19}$$

for $\kappa(s_1 s_2 s_1) = 1$ and $\kappa(s_2 s_1 s_2) = \kappa(s_1 s_2 s_1 s_2) = 2$. Now one uses [31, Lemmas 3.1 and 3.3] to estimate the orbital integrals as in [31, Section 8].[5] This gives

$$\left| \int_{T(\mathbb{R}_+)} \int_{U_w(\mathbb{R})} F^{(X)}\big(\iota^{-1}(X)^{-1} A_c w x y\big) \psi_\infty^{(X^{-1})}(x^{-1}) \overline{E\big(\mathrm{y}(y)\big)}\, dx\, d^*y \right| \ll_{E,\varepsilon} \mathrm{y}(A_c)^{\eta(1+\varepsilon)}$$

[5] Note that [31, Lemma 3.3] is stated only for $w = s_2 s_1 s_2$. However, the statement remains true in general. For Sp_4, this can be checked by hand in the remaining two cases. Otherwise, the argument from [5] applies here as well.

as on the bottom of [31, p. 2071]. Thus, (19) is

$$\ll_{M,\varepsilon} Z^{\varepsilon} \cdot \sum_{\substack{c\in\mathbb{N}^2,\\ c_1\ll Z,\, c_2\ll Z^{\kappa(w)}}} \frac{\mathrm{KL}_{\Gamma,w}(c;1,1)}{c_1c_2}.$$

The desired result follows after inserting the admissibility constraints for c as recorded in Table 1. ∎

For the paramodular group $\Gamma = \Gamma_{\mathrm{pa}}(q)$, we can use the computations from Section 3 to deduce explicit estimates for the geometric side.

Lemma 5.4. *We have*

$$S_{\Gamma_{\mathrm{pa}}(q)}(s_1s_2s_1;Z) = S_{\Gamma_{\mathrm{pa}}(q)}(s_2s_1s_2;Z) = S_{\Gamma_{\mathrm{pa}}(q)}(s_1s_2s_1s_2;Z) = 0$$

for $Z \ll q^{1-\varepsilon}$. Furthermore, if q is prime and $Z \ll q^{2-\varepsilon}$, then we have

$$\begin{aligned} S_{\Gamma_{\mathrm{pa}}(q)}(s_1s_2s_1;Z) &\ll q,\\ S_{\Gamma_{\mathrm{pa}}(q)}(s_2s_1s_2;Z) &= 0,\\ S_{\Gamma_{\mathrm{pa}}(q)}(s_1s_2s_1s_2;Z) &\ll \frac{Z^2}{q^{1+1/4}}. \end{aligned}$$

Proof. If $Z \ll q^{1-\varepsilon}$, everything follows from the definition of $S_{\Gamma_{\mathrm{pa}}(q)}(w;Z)$ together with Lemma 3.2. Thus, we assume that q is prime and that $Z \ll q^{2-\varepsilon}$. This implies that $(c_1, q^{\infty}) = q$ for all moduli $c = (c_1, c_2)$ that appear in the contribution of relevant Weyl elements. With this simple observation the sums $S_{\Gamma_{\mathrm{pa}}(q)}(*;Z)$ can be easily estimated. The case $S_{\Gamma_{\mathrm{pa}}(q)}(s_1s_2s_1s_2;Z)$ is most interesting. From the definition (see (18)) and the factorization formula (6), we get

$$\begin{aligned} S_{\Gamma_{\mathrm{pa}}(q)}(s_1s_2s_1s_2;Z) &= \sum_{\substack{c=(c_1,c_2)\in\mathbb{N}^2,\\ c_1\ll Z,\, c_2\ll Z^2}} \frac{\mathrm{KL}_{\Gamma_{\mathrm{pa}}(q),w}(c;1,1)}{c_1c_2}\\ &= \sum_{i=1,2,3} \frac{|\,\mathrm{KL}_{\Gamma_{\mathrm{pa}}(q),w}((q,q^i);1,N')|}{q^{i+1}}\\ &\qquad \cdot \sum_{\substack{c=(c_1,c_2)\in\mathbb{N}^2,\\ c_1\ll Z/q,\, c_2\ll Z^2/q^i,\\ (c_1c_2,q)=1}} \frac{|\,\mathrm{KL}_{\Gamma_{\mathrm{pa}}(q),w}(c;1,N'')|}{c_1c_2}. \end{aligned}$$

Inserting Lemma 3.2 and Theorem 3.1 yields

$$S_{\Gamma_{\mathrm{pa}}(q)}(s_1s_2s_1s_2;Z) \ll \sum_{i=1,2,3} q^{1+\varepsilon-(i+1)/2} \sum_{\substack{c=(c_1,c_2)\in\mathbb{N}^2,\\ c_1\ll Z/q,\, c_2\ll Z^2/q^i,\\ (c_1c_2,q)=1}} \frac{(c_1,c_2)^{1/2}}{c_1^{1/2}c_2^{1/4}}.$$

Estimating the c-sum is now routine. Indeed, for $i\in\{1,2,3\}$, we have

$$\sum_{\substack{c=(c_1,c_2)\in\mathbb{N}^2,\\ c_1\ll Z/q,\, c_2\ll Z^2/q^i,\\ (c_1c_2,q)=1}} \frac{(c_1,c_2)^{1/2}}{c_1^{1/2}c_2^{1/4}} = \sum_{\substack{a\leq Z/q,\\ (a,q)=1}} a^{-1/4} \sum_{\substack{c_1\ll Z/q,\\ a\mid c_1,\,(c_1,q)=1}} c_1^{-1/2} \sum_{\substack{c_2\ll Z^2/q^i,\\ a\mid c_2,\,(c_2/a,qc_1/a)=1}} c_2^{-1/4}$$

$$\ll \frac{Z^2}{q^{1/2+3i/4}}.$$

The result follows, with the bottleneck being the contribution of $i=1$. ∎

To complement the geometric estimates above we need to handle the spectral side of the Kuznetsov formula. More precisely, we need to show that

$$\sum_{\varpi\in V_\Gamma(\pi)} |A_\varpi(\mathbf{1})|^2$$

is not too small. Establishing this estimate is not an easy task. Assuming Conjecture 4.1 this reduces to a purely local computation, which can in principle be carried out for many families of congruence subgroups. However, these local computations can be rather involved. Luckily, in the case of $\Gamma_{\mathrm{pa}}(q)$ with q square-free, the desired result is available in [10].

Lemma 5.5. *Let $q\in\mathbb{N}$ be square-free and let π be a generic cuspidal automorphic representation. Then we have*

$$\mathcal{V}(\Gamma_{\mathrm{pa}}(q))^{\varepsilon} \sum_{\varpi\in\mathcal{O}_{\Gamma_{\mathrm{pa}}(q)}(\pi)} \frac{|A_\varpi(\mathbf{1})|^2}{\langle\varpi,\varpi\rangle} \gg_\varepsilon \frac{\mathcal{N}(\Gamma_{\mathrm{pa}}(q))}{\mathcal{V}(\Gamma_{\mathrm{pa}}(q))}\cdot \dim_{\mathbb{C}} V_{\Gamma_{\mathrm{pa}}(q)}(\pi).$$

Proof. Recall from (4) that $\mathcal{N}(\Gamma_{\mathrm{pa}}(q))=q$. Further, we observe that

$$\dim V_{\Gamma_{\mathrm{pa}}(q)}(\pi) \ll \mathcal{V}(\Gamma_{\mathrm{pa}}(q))^{\varepsilon}.$$

(The latter result can easily be extracted from [36, Table A.12] and the local nature of $V_{\Gamma_{\mathrm{pa}}(q)}(\pi)$.) From Theorem 4.2, we get

$$\frac{\mathcal{N}(\Gamma_{\mathrm{pa}}(q))}{\mathcal{V}(\Gamma_{\mathrm{pa}}(q))}\cdot\frac{1}{\Lambda^q(1,\pi,\mathrm{Ad})} \ll \sum_{\varpi\in\mathcal{O}_{\Gamma_{\mathrm{pa}}(q)}(\pi)} \frac{|A_\varpi(\mathbf{1})|^2}{\langle\varpi,\varpi\rangle}.$$

We conclude the proof by applying [30, Theorem 2] to obtain an upper bound for $\Lambda^q(1,\pi,\mathrm{Ad})$. ■

Combining the results of this section gives the following theorem for the paramodular group.

Theorem 5.6. *Let q be square-free. We have*

$$N^{\mathrm{gen}}_{\Gamma_{\mathrm{pa}}(q)}(\sigma;M) \ll_{M,\varepsilon} \mathcal{V}(\Gamma_{\mathrm{pa}}(q))^{1-\frac{2\sigma}{3}(1+\delta)+\varepsilon} \tag{20}$$

for $\delta = 1/2$. If q is prime, then we have (20) with $\delta = 11/16$. Furthermore, assuming Conjecture 4.1 and (12) yields (20) for arbitrary q with $\delta = 1/2$.

Proof. We first observe that, if (17) holds with $Z \asymp \mathcal{V}(\Gamma_{\mathrm{pa}}(q))^{\alpha-\varepsilon}$, then we obtain (20) with $\delta = 3\alpha - 1$. In particular, we need $\alpha \geq 1/3$ to achieve the density hypothesis. We continue by combining Lemma 5.3 and Lemma 5.5 to obtain

$$\sum_{\substack{\pi | X_{\Gamma_{\mathrm{pa}}(q)},\\ \text{generic}}} \dim V_{\Gamma_{\mathrm{pa}}(q)}(\pi) \cdot Z^{2\sigma_\pi}$$

$$\ll_M \mathcal{V}\big(\Gamma_{\mathrm{pa}}(q)\big)^{1+\varepsilon} + \frac{\mathcal{V}(\Gamma_{\mathrm{pa}}(q))^{1+\varepsilon}}{\mathcal{N}(\Gamma_{\mathrm{pa}}(q))} \sum_{\substack{1\neq w \in W,\\ \text{relevant}}} S_{\Gamma_{\mathrm{pa}}(q)}(w;Z).$$

This holds for q square-free. In general, we need to assume Conjecture 4.1 and (12) in order to replace Lemma 5.5. We conclude the proof by using Lemma 5.4 to estimate $S_{\Gamma_{\mathrm{pa}}(q)}(w;Z)$. Let us briefly explain this. First, if $\alpha = 1/2$ (i.e., $Z = q$), then $S_{\Gamma_{\mathrm{pa}}(q)}(w;Z) = 0$ for all admissible $w \neq 1$. (This applies to all q not just square-free ones.) Second, if q is prime, then we can take $\alpha = 9/16$ (i.e., $Z = q^{1+1/8}$). ■

Remark 5.7. We expect versions of Theorem 5.6 to hold in great generality. Indeed the main global tool, which is the Kuznetsov formula, is very flexible. Thus one is left with the problem of estimating the first Fourier coefficient on the spectral side and the ramified Kloosterman sums on the geometric side. The spectral problem can, at least conditionally on Conjecture 4.1, be reduced to the purely local problem of establishing (12). The geometric problem also reduces to a local problem. Namely to a suitable estimate for the ramified Kloosterman sets. We believe that for (20) with $\delta = 0$ only very weak bounds for the ramified Kloosterman sets suffice.

5.2. The CAP and residual contribution

We will first decompose the counting function $N_\Gamma(\sigma;M)$ as

$$N_\Gamma(\sigma;M) = \sum_{* \in \{\mathbf{G},\mathbf{Y},\mathbf{Q},\mathbf{P},\mathbf{B},\mathbf{F}\}} N_\Gamma^{(*)}(\sigma;M),$$

where

$$N_{\Gamma}^{(*)}(\sigma; M) = \sum_{\psi \text{ of type } (*)} \sum_{\substack{\pi \in \Pi_{\psi}, \\ \langle \cdot, \pi \rangle = \varepsilon_{\psi}}} \dim V_{\Gamma}(\pi).$$

Note that $\dim V_{\Gamma}(\pi) = 0$ unless $\pi \mid X_{\Gamma}$.

Theorem 5.8. *Let $q > 1$ be arbitrary. The contribution of CAP-representations of types* (**B**) *and* (**Q**) *to the spectrum of $X_{\Gamma_{\mathrm{pa}}(q)}$ is*

$$N_{\Gamma_{\mathrm{pa}}(q)}^{(\mathbf{B})}(\sigma; M) = N_{\Gamma_{\mathrm{pa}}(q)}^{(\mathbf{Q})}(\sigma; M) = 0.$$

Furthermore, the contribution of CAP-representations of type (**P**) *is*

$$N_{\Gamma_{\mathrm{pa}}(q)}^{(\mathbf{P})}(\sigma; M) \ll q^{1+\varepsilon}.$$

Proof. The first part follows directly from Lemma 4.5. We turn towards the type (**P**) contribution. In this case, an application of Lemma 4.7 yields

$$N_{\Gamma}^{(\mathbf{P})}(\sigma; M) \ll q^{\varepsilon} \sharp \{\psi = (\mu \boxtimes 1) \boxplus (\mathbf{1} \boxtimes \nu(2)) \in \boldsymbol{\psi}_2(G, \mathbf{1}) : c(\mu) \mid q\}$$

for $\sigma \leq 1/2$. Thus we are essentially counting cuspidal automorphic representations μ of GL_2 with trivial central character, conductor dividing q (and bounded spectral parameter). The result follows from the appropriate Weyl law; see, for example, [15, Theorem 1.1]. ∎

5.3. The endgame

We are finally ready to put all the pieces together and assemble the final density theorem.

Theorem 5.9. *For square-free q, we have the density result*

$$N_{\Gamma_{\mathrm{pa}}(q)}(\sigma; M) \ll_{M,\varepsilon} \mathcal{V}(\Gamma_{\mathrm{pa}}(q))^{1-\frac{2}{3}\sigma(1+1/2)+\varepsilon} + 1.$$

More precisely, for each $* \in \{\mathbf{G}, \mathbf{Y}, \mathbf{B}, \mathbf{P}, \mathbf{Q}, \mathbf{F}\}$, there is $0 \leq \sigma(*) \leq 3/2$ such that $N_{\Gamma_{\mathrm{pa}}(q)}^{(*)}(\sigma; M) = 0$ for $\sigma > \sigma(*)$, and for $\sigma \leq \sigma(*)$, we have good estimates of the form

$$N_{\Gamma_{\mathrm{pa}}(q)}(\sigma; M) \ll_{M,\varepsilon} \mathcal{V}\big(\Gamma_{\mathrm{pa}}(q)\big)^{C_{(*)}(\sigma,\varepsilon)}.$$

We summarize the values for $\sigma(*)$ and $C_{(*)}(\sigma, \varepsilon)$ in Table 3.

Proof. We start by observing that, by Lemma 4.3 (in particular (14)), we have

$$N_{\Gamma_{\mathrm{pa}}(q)}^{(\mathbf{G})}(\sigma; M) + N_{\Gamma_{\mathrm{pa}}(q)}^{(\mathbf{Y})}(\sigma; M) = N_{\Gamma_{\mathrm{pa}}(q)}^{\mathrm{gen}}(\sigma; M).$$

$*$	Condition on q	$\sigma(*)$	$C_{(*)}(\sigma,\varepsilon)+\varepsilon$
G	square-free	$9/22$	$1-\frac{2}{3}\sigma(1+1/2)+\varepsilon$
G	prime	$9/22$	$1-\frac{2}{3}\sigma(1+11/16)+\varepsilon$
Y	square-free	$7/64$	$1-\frac{2}{3}\sigma(1+1/2)\varepsilon$
Y	prime	$7/64$	$1-\frac{2}{3}\sigma(1+11/16)+\varepsilon$
B	none	0	-
P	none	$1/2$	$1+\varepsilon$
Q	none	0	-
F	none	$3/2$	0

Table 3. Summary of exponents.

The desired bound follows from Theorem 5.6. The vanishing results can be deduced from absolute bounds towards the Ramanujan Conjecture for GL_4 and are given in [4, Corollary 3.4]. The results for the CAP-contributions are provided in Theorem 5.8. ■

This result directly implies Theorem 1.1 (and more).

Acknowledgments. I would like to thank V. Blomer for his encouragement while working on the project. Further I would like to thank R. Schmidt for patiently answering my questions related to Arthur parameters for GSp_4. We also thank the anonymous referee for a very carefully reading the manuscript and for making many useful suggestions.

Publisher's note. The date of submission was corrected from 9 May 2024 to 9 May 2023 on 18 March 2026.

References

[1] J. Arthur, Automorphic representations of GSp(4). In *Contributions to automorphic forms, geometry, and number theory*, pp. 65–81, Johns Hopkins University Press, Baltimore, MD, 2004 Zbl 1080.11037 MR 2058604

[2] J. Arthur, *The endoscopic classification of representations*. Amer. Math. Soc. Colloq. Publ. 61, American Mathematical Society, Providence, RI, 2013 Zbl 1310.22014 MR 3135650

[3] M. Asgari and R. Schmidt, Siegel modular forms and representations. *Manuscripta Math.* **104** (2001), no. 2, 173–200 Zbl 0987.11037 MR 1821182

[4] M. Asgari and F. Shahidi, Generic transfer from GSp(4) to GL(4). *Compos. Math.* **142** (2006), no. 3, 541–550 Zbl 1112.11027 MR 2231191

[5] E. Assing, A density theorem for Borel-type congruence subgroups and arithmetic applications. 2023, arXiv:2303.08925v1, to appear in *Algebra Number Theory*

[6] E. Assing and V. Blomer, The density conjecture for principal congruence subgroups. *Duke Math. J.* **173** (2024), no. 7, 1359–1426 Zbl 1564.11055 MR 4757534

[7] V. Blomer, Density theorems for $\mathrm{GL}(n)$. *Invent. Math.* **232** (2023), no. 2, 683–711 Zbl 1530.11052 MR 4574662

[8] V. Blomer and F. Brumley, On the Ramanujan conjecture over number fields. *Ann. of Math. (2)* **174** (2011), no. 1, 581–605 Zbl 1322.11039 MR 2811610

[9] J. Breeding, II, Irreducible characters of $\mathrm{GSp}(4, q)$ and dimensions of spaces of fixed vectors. *Ramanujan J.* **36** (2015), no. 3, 305–354 Zbl 1369.11036 MR 3317862

[10] S.-Y. Chen and A. Ichino, On Petersson norms of generic cusp forms and special values of adjoint L-functions for GSp_4. *Amer. J. Math.* **145** (2023), no. 3, 899–993 Zbl 1536.11079 MR 4596180

[11] F. Comtat, A relative trace formula approach to the Kuznetsov formula on GSp_4. 2021, arXiv:2107.08755v2

[12] R. Dąbrowski, Kloosterman sums for Chevalley groups. *Trans. Amer. Math. Soc.* **337** (1993), no. 2, 757–769 Zbl 0790.11060 MR 1102221

[13] R. Dąbrowski and M. Reeder, Kloosterman sets in reductive groups. *J. Number Theory* **73** (1998), no. 2, 228–255 Zbl 0919.11055 MR 1658031

[14] M. Dickson, A. Pitale, A. Saha, and R. Schmidt, Explicit refinements of Böcherer's conjecture for Siegel modular forms of squarefree level. *J. Math. Soc. Japan* **72** (2020), no. 1, 251–301 Zbl 1476.11079 MR 4055095

[15] H. Donnelly, On the cuspidal spectrum for finite volume symmetric spaces. *J. Differential Geometry* **17** (1982), no. 2, 239–253 Zbl 0494.58029 MR 0664496

[16] S. Evra, M. Gerbelli-Gauthier, and H. P. A. Gustafsson, The cohomological Sarnak-Xue density hypothesis for SO_5. [v1] 2023, [v4] 2025, arXiv:2309.12413v4

[17] S. Friedberg, Poincaré series for $\mathrm{GL}(n)$: Fourier expansion, Kloosterman sums, and algebreo-geometric estimates. *Math. Z.* **196** (1987), no. 2, 165–188 Zbl 0612.10020 MR 0910824

[18] M. Furusawa and K. Morimoto, On the Gross–Prasad conjecture with its refinement for $(\mathrm{SO}(5), \mathrm{SO}(2))$ and the generalized Böcherer conjecture. *Compos. Math.* **160** (2024), no. 9, 2115–2202 Zbl 1564.11048 MR 4797111

[19] W. T. Gan and S. Takeda, The local Langlands conjecture for GSp(4). *Ann. of Math. (2)* **173** (2011), no. 3, 1841–1882 Zbl 1230.11063 MR 2800725

[20] T. Gee and O. Taïbi, Arthur's multiplicity formula for $\mathbf{G}Sp_4$ and restriction to $\mathbf{S}p_4$. *J. Éc. polytech. Math.* **6** (2019), 469–535 Zbl 1468.11115 MR 3991897

[21] A. Ghosh, A. Gorodnik, and A. Nevo, Diophantine approximation and automorphic spectrum. *Int. Math. Res. Not. IMRN* (2013), no. 21, 5002–5058 Zbl 1370.11077 MR 3123673

[22] P. Humphries, Density theorems for exceptional eigenvalues for congruence subgroups. *Algebra Number Theory* **12** (2018), no. 7, 1581–1610 Zbl 1444.11105 MR 3871503

[23] M. N. Huxley, Exceptional eigenvalues and congruence subgroups. In *The Selberg trace formula and related topics (Brunswick, Maine, 1984)*, pp. 341–349, Contemp. Math. 53, American Mathematical Society, Providence, RI, 1986 Zbl 0601.10019 MR 0853564

[24] T. Ishii, On principal series Whittaker functions on $\mathrm{Sp}(2,\mathbf{R})$. *J. Funct. Anal.* **225** (2005), no. 1, 1–32 Zbl 1078.11031 MR 2149916

[25] H. Iwaniec, Small eigenvalues of Laplacian for $\Gamma_0(N)$. *Acta Arith.* **56** (1990), no. 1, 65–82 Zbl 0702.11034 MR 1067982

[26] S. Jana and A. Kamber, On the local L^2-bound of the Eisenstein series. *Forum Math. Sigma* **12** (2024), article no. e76 Zbl 1564.11053 MR 4794440

[27] V. Kala, Density of self-dual automorphic representations of $\mathrm{GL}_n(\mathbb{A}_{\mathbb{Q}})$. PhD thesis, Purdue University, 2014

[28] A. Knightly and C. Li, On the distribution of Satake parameters for Siegel modular forms. *Doc. Math.* **24** (2019), 677–747 Zbl 1470.11109 MR 3960116

[29] E. Lapid and Z. Mao, A conjecture on Whittaker–Fourier coefficients of cusp forms. *J. Number Theory* **146** (2015), 448–505 Zbl 1396.11081 MR 3267120

[30] X. Li, Upper bounds on L-functions at the edge of the critical strip. *Int. Math. Res. Not. IMRN* (2010), no. 4, 727–755 Zbl 1219.11136 MR 2595006

[31] S. H. Man, A density theorem for $\mathrm{Sp}(4)$. *J. Lond. Math. Soc. (2)* **105** (2022), no. 4, 2047–2075 Zbl 1539.11079 MR 4440530

[32] S. H. Man, Symplectic Kloosterman sums and Poincaré series. *Ramanujan J.* **57** (2022), no. 2, 707–753 Zbl 1489.11114 MR 4372237

[33] S. Marshall and S. W. Shin, Endoscopy and cohomology in a tower of congruence manifolds for $U(n,1)$. *Forum Math. Sigma* **7** (2019), article no. e19 Zbl 1421.32030 MR 3981603

[34] I. I. Piatetski-Shapiro, On the Saito–Kurokawa lifting. *Invent. Math.* **71** (1983), no. 2, 309–338 Zbl 0515.10024 MR 0689647

[35] A. Pitale, A. Saha, and R. Schmidt, Simple supercuspidal representations of GSp_4 and test vectors. 2023, arXiv:2302.05148v1

[36] B. Roberts and R. Schmidt, *Local newforms for* $\mathrm{GSp}(4)$. Lecture Notes in Math. 1918, Springer, Berlin, 2007 Zbl 1126.11027 MR 2344630

[37] P. Sarnak, Diophantine problems and linear groups. In *Proceedings of the International Congress of Mathematicians, Vol. I, II (Kyoto, 1990)*, pp. 459–471, Mathematical Society of Japan, Tokyo, 1991 Zbl 0743.11018 MR 1159234

[38] P. Sarnak, Notes on the generalized Ramanujan conjectures. In *Harmonic analysis, the trace formula, and Shimura varieties*, pp. 659–685, Clay Math. Proc. 4, American Mathematical Society, Providence, RI, 2005 Zbl 1146.11031 MR 2192019

[39] P. Sarnak and X. X. Xue, Bounds for multiplicities of automorphic representations. *Duke Math. J.* **64** (1991), no. 1, 207–227 Zbl 0741.22010 MR 1131400

[40] R. Schmidt, Packet structure and paramodular forms. *Trans. Amer. Math. Soc.* **370** (2018), no. 5, 3085–3112 Zbl 1440.11075 MR 3766842

[41] R. Schmidt, Paramodular forms in CAP representations of $\mathrm{GSp}(4)$. *Acta Arith.* **194** (2020), no. 4, 319–340 Zbl 1452.11056 MR 4103275

[42] F. Shahidi, Arthur packets and the Ramanujan conjecture. *Kyoto J. Math.* **51** (2011), no. 1, 1–23 Zbl 1238.11060 MR 2784745

[43] G. Stevens, Poincaré series on $\mathrm{GL}(r)$ and Kloostermann sums. *Math. Ann.* **277** (1987), no. 1, 25–51 Zbl 0597.12017 MR 0884644

Received 9 May 2023.

Edgar Assing
Department of Mathematics, Universität Bonn, Endenicher Allee 60, 53115 Bonn, Germany;
assing@math.uni-bonn.de