THE DENSITY CONJECTURE FOR PRINCIPAL
CONGRUENCE SUBGROUPS

EDGAR ASSING AND VALENTIN BLOMER

Abstract. We prove the spherical density conjecture of Sarnak for the principal subgroup of $\text{SL}_n(\mathbb{Z})$ of squarefree level and discuss various arithmetic applications. The ingredients include new bounds for local Whittaker functions and Kloosterman sums.

1. Introduction

1.1. The density hypothesis. Perhaps the most fundamental open conjecture in the theory of automorphic forms, in some aspects comparable to the Riemann hypothesis in the theory of $L$-functions, is the generalized Ramanujan conjecture: unitary cuspidal automorphic representations for the group $\text{GL}(n)$ are tempered. Even for $n = 2$ it appears completely out of reach. See [Sa3, BB] for surveys. Many applications, however, do not need the full strength of the Ramanujan conjecture, it suffices to quantify the possible exceptions and show that “strong” violations to the Ramanujan conjecture appear “rarely” (in a sense to be made precise below). Such a quantitative approximation towards the Ramanujan conjecture is called a density theorem. This is a familiar concept from the theory of $L$-functions, the most famous application being the Bombieri–Vinogradov theorem. It can be proved by a density theorem for Dirichlet $L$-functions (see [IK, Section 10]), and its arithmetic content is (a precise version of) the statement that primes are equidistributed in residue classes to large moduli; see [IK, Section 17]. The Bombieri–Vinogradov theorem is one of the cornerstones in analytic number theory and a good substitute for the Riemann hypothesis in many cases.

We turn to the automorphic set-up. Let $\mathcal{F}$ be a finite family of automorphic forms contributing to the spectral decomposition of $L^2(\Gamma \backslash \text{SL}_n(\mathbb{R})/\text{SO}_n(\mathbb{R}))$ for some congruence lattice $\Gamma \subset \text{SL}_n(\mathbb{R})$. For $\varpi \in \mathcal{F}$ and an unramified place $v$ of $\mathbb{Q}$ we have a set of Langlands parameters $\{\mu_{\varpi, v}(j) \mid 1 \leq j \leq n\}$. We normalize these such that $\varpi$ is tempered at $v$ when all $\mu_{\varpi, v}(j)$ are purely imaginary and write $\sigma_{\varpi, v} = \max_j |\Re \mu_{\varpi, v}(j)|$. For $\sigma \geq 0$ we write

\begin{equation}
N_v(\sigma, \mathcal{F}) = \# \{\varpi \in \mathcal{F} \mid \sigma_{\varpi, v} \geq \sigma\}
\end{equation}
for the number of automorphic forms in $\mathcal{F}$ violating the Ramanujan conjecture at $v$ by $\sigma$. The trivial bound is $N_v(\sigma, \mathcal{F}) \leq N_v(0, \mathcal{F}) = \# \mathcal{F}$, and if $\mathcal{F}$ contains the trivial representation, then $N_v(\frac{1}{2}(n - 1), \mathcal{F}) = 1$.

In his 1990 ICM address [Sa2], see also [SX], Sarnak considered this scenario in the context of groups $G$ of real rank 1, the principal congruence subgroup and $v = \infty$, and coined the phrase density hypothesis for a linear interpolation between these two trivial bounds for the tempered spectrum and the trivial representation. See [Sa5, PSa] for interesting recent applications of this hypothesis in the context of Golden Gates and quantum computing.

For the rest of this paper, we consider the (finite) family $\mathcal{F} = \mathcal{F}_{\Gamma(q)}(M)$ of cusp forms $\pi$ for the principal congruence subgroup

$$\Gamma(q) = \ker(\text{SL}_n(\mathbb{Z}) \to \text{SL}_n(\mathbb{Z}/q\mathbb{Z})) = \begin{pmatrix} 1 + q\mathbb{Z} & q\mathbb{Z} & \cdots & q\mathbb{Z} \\ q\mathbb{Z} & 1 + q\mathbb{Z} & \cdots & q\mathbb{Z} \\ \vdots & \ddots & \ddots & \vdots \\ q\mathbb{Z} & \cdots & 1 + q\mathbb{Z} & q\mathbb{Z} \\ q\mathbb{Z} & \cdots & q\mathbb{Z} & 1 + q\mathbb{Z} \end{pmatrix} \cap \text{SL}_n(\mathbb{Q})$$

with bounded spectral parameter $\|\mu_{\pi, \infty}\| \leq M$. In this case, the spherical density conjecture reads

$$N_v(\sigma, \mathcal{F}) \ll_{v, \varepsilon, M, n} [\text{SL}_n(\mathbb{Z}) : \Gamma(q)]^{1 - \frac{2\sigma}{n-1} + \varepsilon}$$

where we think of $O_M([\text{SL}_n(\mathbb{Z}) : \Gamma(q)])$ as the cardinality of $\mathcal{F}$ (by an appropriate version of Weyl’s law). For applications it is desirable to have polynomial dependence in $M$. For $n = 2$, this conjecture and archimedean variations thereof are well understood and proved in many typical situations and in different ways; see e.g. [Hum, Hu, Iw, Sa1, FHMM].

For $n > 2$ the problem is completely open, even in special cases. Only very recently, the second author [Bl] established (a stronger version of) the density conjecture for the Hecke congruence subgroup $\Gamma_0(p) \subseteq \text{SL}_n(\mathbb{Z})$ for a large prime $p$, using the Kuznetsov formula and a careful analysis of Kloosterman sums. An archimedean analogue was subsequently shown in [Ja]. Other than that we have no results for $n > 3$, neither for archimedean families, nor in the level aspect for any type of congruence subgroup. The authors of [GK] argue “that it seems that the problem is harder for principal congruence subgroups, and easier for group[s] that are far from being normal.” While it is too early to formalize this and also depends to some extent on the viewpoint, the principal congruence subgroup is certainly a very natural scenario and also interesting for applications. In this paper we solve the problem for the principal congruence subgroup of squarefree level in arbitrary dimension.

**Theorem 1.1.** Let $M, \varepsilon > 0$. Let $q \in \mathbb{N}$ be squarefree and let $\mathcal{F} = \mathcal{F}_{\Gamma(q)}(M)$ be the set of cuspidal automorphic forms for $\Gamma(q) \subseteq \text{SL}_n(\mathbb{Z})$ with archimedean
spectral parameter $\|\mu\| \leq M$. Fix a place $v$ of $\mathbb{Q}$, and if $v = p$ is finite, assume that $p \nmid q$. There exists a constant $K$ depending only on $n$, such that
\[
\mathcal{N}_v(\sigma, F) \ll_{v, \varepsilon, n} M^K [\text{SL}_n(\mathbb{Z}) : \Gamma(q)]^{1 - \frac{2\varepsilon}{n^2} + \varepsilon}
\]
for $\sigma \geq 0$.

We will extend Theorem 1.1 to the entire spectrum (not only the cuspidal one) in Theorem 7.1 below. However, our method begins naturally with the cuspidal spectrum and in this way provides, as in [Bl], the option to improve the exponent beyond the interpolation between the tempered and spectrum and the trivial representation (which is not cuspidal). See [BM, Proposition 4] for a special case in this direction. This is the beauty of the Kuznetsov formula, since the spectral side involves no residual spectrum. By the Luo-Rudnick-Sarnak bounds [LRS], Theorem 1.1 by itself is nevertheless non-trivial for $\sigma < 1/2 - 1/(n^2 + 1)$.

We rephrase Theorem 1.1 in a representation theoretical way. A spherical (i.e. unramified) representation $\pi_v$ of $\text{GL}_n(\mathbb{Q}_v)$ is determined by a set of Langlands parameters $\{\mu_{\pi_v}(j) | 1 \leq j \leq n\}$. Again we normalize so that $\pi_v$ is tempered when all $\mu_{\pi_v}(j)$ are purely imaginary. We say that $\pi_v$ appears in $F$ when there is $\varpi \in F$ with $\mu_{\varpi \pi_v} = \mu_{\pi_v}$ and write $m_F(\pi_v)$ for the multiplicity of $\pi_v$ in $F$, i.e. $m_F(\pi_v) = \#\{\varpi \in F : \mu_{\varpi \pi_v} = \mu_{\pi_v}\}$ (see also (2.30) below). If $F = F_{\Gamma(q)}(M)$, it is suggestive to write $m_F = m_{\Gamma(q)}$.

An immediate corollary of Theorem 1.1 is the following upper bound on these multiplicities. This proves Conjecture 1 of Sarnak-Xue [SX] when generalized to arbitrary rank and an arbitrary place $v$.

**Corollary 1.2.** Under the same assumptions as in Theorem 1.1 and with the same notation we have for an irreducible unramified representation $\pi_v$ of $\text{GL}_n(\mathbb{Q}_v)$ the bound
\[
m_{\Gamma(q)}(\pi_v) \ll_{v, n, \varepsilon} \|\mu_{\pi_v}\|^{K} [\text{SL}_n(\mathbb{Z}) : \Gamma(q)]^{1 - \frac{2\varepsilon}{n^2} + \varepsilon}
\]
where $\|\cdot\|$ denotes the max-norm and $\sigma_{\pi_v} = \|\Re \mu_{\pi_v}\| = \max_j |\Re \mu_{\pi_v}(j)|$.

The restriction that $q$ be squarefree is used at two independent technical (but important) points in the argument: in the computation of Kloosterman sums in Lemma 4.3 and in the computation of local Whittaker functions for supercuspidal representations in Section 6. It might be possible to remove this assumption with more work and somewhat different arguments, cf. the remarks at the end of Subsection 1.2 and after Lemma 4.3.

We prove several different versions of Theorem 1.1 that we now discuss.

The condition $\sigma_{\varpi \pi_v}(v) \geq \sigma$ in the definition of $\mathcal{N}_v(\sigma, F)$ will be detected by a weight $T^{2\varpi \pi_v} \varpi$ for a parameter $T > 1$ in the trace formula. For several applications (including the ones in this paper) it is easier to work directly with this weight rather than the quantity $\mathcal{N}_v(\sigma, F)$; see Corollary 6.11 below.

It is also possible to obtain a version of Theorem 1.1 on the level of cuspidal automorphic representations, rather than cuspidal automorphic forms. This is discussed in Remark 6.1 below.
Also for applications it is useful to include non-cuspidal automorphic representations into the bound to be able to handle a full spectral decomposition. By Langlands’ classification of Eisenstein series, this is essentially an inductive process based on Theorem 1.1. This more general bound will be stated and proved in Section 7 below, and we refer to Theorem 7.1.

The prototype of our density result comes from an application of the Kuznetsov formula and therefore involves unipotent periods (Fourier coefficients) as weights; see Proposition 5.2. This might be useful for some applications where these weights come up naturally. Removing them to obtain the classical unweighted density conjecture is a non-trivial task in the present situation. This would be simple for the group $\Gamma_0(q)$ by newform theory, but in absence of multiplicity one theorems and newform theory for general subgroups like $\Gamma(q)$ one has to obtain good lower bounds/asymptotics for local Whittaker functions on average over an orthonormal basis in each given representation. This is to some extent a problem of local representation theory and should have applications elsewhere. We discuss this in more detail below; see Theorem 1.3 in the next subsection.

Finally we mention that there is a slightly different way to measure the failure of being tempered which avoids local Langlands parameters; cf. [SX]. For a local representation $\pi$ let $p(\pi)$ be the infimum over $p \geq 2$ such that $K$-finite matrix coefficients of $\pi$ are in $L^p(Z(Q_v) \backslash GL_n(Q_v))$. Then $\pi$ is tempered if $p(\pi) = 2$, and instead of $\sigma_{\pi,v}$ one can use a function containing $p(\pi)$, e.g. $1 - 2/p(\pi)$ in the definition (1.1) of $N_v$. One can express $p(\pi)$ as a function of $\Re \mu_\pi$; see [GGN, Lemma 3.2]. In principle our method could also handle this alternative formulation of the density conjecture by using more general weight parameters in (5.3).

1.2. The methods. All approaches to automorphic density conjectures use some sort of trace formula which translates the spectral problem into an arithmetic problem. If we were to apply the pretrace formula to prove Theorem 1.1, we would essentially arrive at the counting problem described in Theorem 1.4 below. This is the route that Sarnak-Xue [SX] used for $n = 2$. However, in higher rank we do not know how to solve this problem independently. Instead, we use the Kuznetsov formula to derive Theorem 1.1; which also leads to a counting problem, but a rather different one. The general framework of the Kuznetsov formula (see Section 5) was already developed in [Bl], but a trace formula is only as powerful as its input on the geometric and/or the spectral side. Here we need to develop new tools on both sides, and our application requires sharp bounds on both sides.

The arithmetic work consists in analyzing Kloosterman sums for the group $\Gamma(q)$ in higher rank. The investigation of such Kloosterman sums is, to the best of the authors’ knowledge, new. It turns out that the critical Weyl element is the element $w_\ast$ as defined in (2.2) below, and here we obtain sharp upper bounds for the corresponding Kloosterman sums. The key result is Theorem 4.4 below; see also Subsection 1.4.
Being a relative trace formula with respect to two copies of unipotent upper triangular matrices, the Kuznetsov formula contains Fourier coefficients on the spectral side. It is most convenient to work with the conjugated group
\[ \Gamma(q)^2 = \text{diag}(q^{n-1}, \ldots, q, 1) \Gamma(q) \text{diag}(q^{n-1}, \ldots, q, 1), \]
the advantage being that the standard character on the unipotent group \( U \) is trivial on \( \Gamma(q)^2 \) and \( \Gamma(q)^2 \) contains \( U(\mathbb{Z}) \). Therefore the Fourier expansion with respect to \( \Gamma(q)^2 \) is similar to the one for \( \Gamma_0(q) \) and in contrast to \( \Gamma(q) \) has no “denominators”. For a member \( \varpi \) of an orthonormal basis of \( L^2(\Gamma(q)^2 \backslash \text{SL}_n(\mathbb{R})/\text{SO}_n(\mathbb{R})) \) with respect to the inner product as defined in (2.19) below and \( y = \text{diag}(y_{n-1} \cdots y_1, y_2 y_1, y_1, 1) \), we denote by
\[ \mathcal{W}_\varpi(y) = \int_{U(\mathbb{Z}) \setminus U(\mathbb{R})} \varpi(xy) \theta^{-1}(x) dx \]
its “first” Fourier coefficient (with the standard notation to be recalled and introduced in the next section). For cuspidal \( \varpi \) with spectral parameter \( \mu = \mu_{\varpi, \infty} \) the Fourier coefficient \( \mathcal{W}_\varpi(y) \) is a multiple of the standard Whittaker function \( \mathcal{W}_\mu(y_1, \ldots, y_{n-1}) \) as defined in (2.16) below, and we write
\[ \mathcal{W}_\varpi(y) = A_{\varpi}(1) \mathcal{W}_\mu(y_1, \ldots, y_{n-1}). \]
It may well happen that \( A_{\varpi}(1) = 0 \). An important ingredient in the proof, which may have applications elsewhere, is the following essentially sharp lower bound of independent interest, which shows that in a certain average sense the first Fourier coefficient is at least as big as expected.

**Theorem 1.3.** Let \( q \) be squarefree. For an irreducible cuspidal automorphic representation \( \pi \) on \( X_q^A \) (as defined in (2.23) below) let \( V_\pi \subseteq L^2(\Gamma(q)^2 \backslash \text{SL}_n(\mathbb{R})/\text{SO}_n(\mathbb{R})) \) be the (finite dimensional) space of cusp forms whose adelic lifts generate \( \pi \).

For any \( \varepsilon > 0 \) we have
\[ \frac{1}{\dim V_\pi} \sum_{\varpi \in \text{ONB}(V_\pi)} |A_{\varpi}(1)|^2 \geq (\|\mu_{\varpi, \infty}\|_q)^{-\varepsilon} q^{n(n-1)(n-2)/6} q^{n^2-1}, \]
where ONB refers to an orthonormal of \( V_\pi \).

Here the denominator is essentially the index \( [\text{SL}_n(\mathbb{Z}) : \Gamma(q)] \) and the numerator is the index \( [\Gamma(q)^2 \cap U(\mathbb{Q}) : U(\mathbb{Z})] \). The proof of Theorem 1.3 depends ultimately on classification results for \( p \)-adic representations, and for computations it is useful to note that the principal congruence subgroup is not only normal, but also admits an Iwahori factorization with respect to all standard parabolic subgroups (so it is “good” in the sense of [NV, Section 23]). Locally, we need to understand the average
\[ \sum_{v \in \text{ONB}(\pi_p \Gamma(q)^2)} \int_{U(\mathbb{Q}_p)} \theta^{-1}(x) \langle \pi(x)v, v \rangle dx, \]
where \( \text{ONB}(\pi_p \Gamma(q)^2) \) is any orthonormal basis of \( \Gamma(q)^2 \)-fixed vectors in \( \pi_p \) and \( \int^{\text{st}} \) denotes the stable integral as defined in [LM, Definition 2.1]. Suppose for
instance that $\pi_p$ is a depth-zero supercuspidal representation. Any such representation can be obtained from a cuspidal representation $\pi_0$ of $GL_n(\mathbb{Z}/p\mathbb{Z})$. Then the integral (1.2) depends only on the character of $\pi_0$ at elements in $U(\mathbb{Z}/p\mathbb{Z})$, so one can compute (1.2) quite explicitly. For $q = p^k$ with $k > 1$, the bottleneck is the treatment of supercuspidal representations of higher depth. It should be possible to treat these using more general constructions via induction from open compact subgroups. The methods in [NV] provide robust tools in the archimedean (microlocal) case, and it should be possible to translate them to the $p$-adic setting; see also [Ne, Section 2.5]. We hope to return to this in the future.

1.3. Two arithmetic counting problems. As observed by Sarnak, closely related to the density conjecture for the group $\Gamma(q) \subseteq SL_n(\mathbb{Z})$ are two very concrete arithmetic applications that are in some sense the analogue of the distribution of primes in arithmetic progressions for density theorems of $L$-functions. Both of them only need the case $v = \infty$ in Theorem 1.1.

The first one is a rather innocent looking counting problem: Count the number of matrices $\gamma \in \Gamma(q)$ of norm $\leq T$. As long as $q$ is fixed and $T$ tends to infinity, this has been established in much broader context in [DRS]. The problem gets much harder if we are aiming for a bound that is uniform in $q$. By a volume argument, we expect something of the size $T^{n(n-1)}/q^{n^2-1}$. In analogy with primes $p \leq x$ in arithmetic progressions modulo $q$, where the Riemann hypothesis predicts a main term of size (roughly) $x/q$ and an error term of size (roughly) $x^{1/2}$, one may optimistically hope to obtain a correction term of size $T^{n(n-1)/2}$ for the matrix counting problem. This was conjectured by Sarnak-Xue [SX] in rank 1 and proved by an essentially elementary technique, which then implied the density conjecture in rank 1. In higher rank, this counting problem becomes surprisingly hard and has remained completely open until now (even for $n = 3$, say). Here we turn tables and try to derive it as a consequence of the density conjecture.\footnote{This leads to an interesting logical sequence which in rather simplified form looks as follows: in order to solve the matrix counting problem, we apply the pretrace formula to translate it into a spectral problem, add artificially unipotent periods, apply the Kuznetsov formula backwards, and obtain a very different arithmetic problem of counting double cosets in the definition of Kloosterman sums that we are able to solve in a best possible way.}

The second application is a beautiful lifting property. As is well-known, the map $SL_n(\mathbb{Z}) \to SL_n(\mathbb{Z}/q\mathbb{Z})$ is surjective. Given some $g \in SL_n(\mathbb{Z}/q\mathbb{Z})$, how small is the smallest preimage? This is the automorphic analogue of finding the smallest prime in an arithmetic progression. Since the term $T^{n(n-1)}/q^{n^2-1}$ in Theorem 1.4 comes from the volume and cannot be improved, by the pigeonhole principle we cannot ask for anything better than a preimage of norm $q^{1+1/n}$. But even this is not true in general. Already for $n = 2$ and $q$ a large power of 2, Sarnak [Sa5] constructed a matrix $g$ whose smallest preimage is as big as $q^{8/3}$. However, generically, i.e. for almost all $g \in SL_n(\mathbb{Z}/q\mathbb{Z})$, the pigeonhole principle should give the correct...
answer. This is Sarnak’s optimal lifting property. See [GK, KL, JK] from more further discussion.

A crucial input for both applications is a density theorem, and interestingly a spectral gap (even a numerically strong one) alone does not suffice. For an illuminating discussion of the various connections and logical implications of the density conjecture and the associated counting problems for co-compact subgroups see [GK].

In the present case of a non-cocompact congruence subgroup, one has to treat Eisenstein series in the pre-trace formula (see Section 8), and to this end we need some information on the $L^2$-growth of Eisenstein series. Our results are conditional on the assumption that the norms of all relevant truncated (adelic) Eisenstein series for $GL_n(\mathbb{A})$ are essentially bounded in the $q$-aspect and of at most polynomial growth in the spectral aspect:

$$\|\Lambda^T E(\cdot, \phi, \lambda)\|^2 \ll_{T, \varepsilon} q\varepsilon |\mu_{\sigma, \lambda}|^{O(1)}.$$  \hfill (1.3)

For the notation and the precise conditions we refer to Section 8 where it is formulated as Hypothesis 1. We discuss this hypothesis as well as recent developments to remove it after stating the following two theorems.

**Theorem 1.4.** Assume Hypothesis 1. Let $q$ be squarefree, $T \geq 1$, $\varepsilon > 0$. The number of matrices $\gamma \in \Gamma(q) \subseteq SL_n(\mathbb{Z})$ of norm not exceeding $T$ is

$$\ll_{\varepsilon, n} (Tq)^\varepsilon \left(\frac{T^{n(n-1)}}{q^{2n-1}} + T^{n(n-1)/2}\right).$$

As explained above, the key point of this bound is that the second order term features square-root cancellation in the $T$-aspect (and hence is “of Ramanujan quality”). With slightly more work, this could be turned into an asymptotic formula with a main term of size $T^{n(n-1)}/q^{2n-1}$ and an error term of form $T^{n(n-1)-\delta}$ for some $\delta > 0$. This has been obtained in [GN, Corollary 5.2]. It is easy to see (cf. [GK, Section 2.6]) that a lower bound is given by $T^{n(n-1)}/q^{2n-1} + (T/q)^{n(n-1)/2} + 1$.

**Theorem 1.5.** Assume Hypothesis 1. For any $\varepsilon > 0$ there exists $\delta > 0$ with the following property. Let $q$ be squarefree. Then the number of matrices $\gamma \in SL_n(\mathbb{Z}/q\mathbb{Z})$ without a lift of norm $\leq q^{1+1/n+\varepsilon}$ is at most $O(q^{n^2-1-\delta})$.

In other words, since $|SL_n(\mathbb{Z}/q\mathbb{Z})| \asymp q^{n^2-1}$, “almost all” matrices $\gamma \in SL_n(\mathbb{Z}/q\mathbb{Z})$ have the optimal lifting property.

By Langlands’ inner product formula (also referred to as the Maaß-Selberg relations), the left hand side of (1.3) can be expressed in terms of quotients of Rankin-Selberg $L$-functions $L(s, \pi_1 \times \pi_2)$ for $\Re s \geq 1$ for cusp forms $\pi_j$ on $GL_{n_j}(\mathbb{A})$ with $n_1 + n_2 \leq n$. To show that for residual Eisenstein series no values of $s$ inside the critical strip contribute requires non-trivial cancellation of terms in the Maaß-Selberg relations; see [Mi] for $n = 3$ and [JK] for the details in the general case (cf. also [GL]). In particular, in this way we see that Hypothesis 1 follows from the following weak zero-free region (and hence
EDGAR ASSING AND VALENTIN BLOMER

a fortiori from GRH): whenever $n_1 + n_2 \leq n$ there exists $A > 0$ such that we have

\[(1.4) \quad L(1 + it, \pi_1 \times \tilde{\pi}_2) \gg c_{\text{fin}}(\pi_1 \times \tilde{\pi}_2)^{-\varepsilon}((1 + |t|)c_{\infty}(\pi_1 \times \tilde{\pi}_2))^{-A}\]

for all cuspidal automorphic representations $\pi_i$ of $\text{GL}_{n_i}(\mathbb{A})$ with $c_{\text{fin}}(\pi_i) | q^{n_i}$.

In particular, by functoriality of $\text{GL}(2) \times \text{GL}(2)$ Rankin-Selberg convolution [Ra] together with the standard zero-free region [Iw, Theorem 5.10], Siegel’s bound [Si] and the non-existence of Siegel zeros for standard $L$-functions on $\text{GL}(2)$ [HL], standard $L$-functions on $\text{GL}(3)$ [Ba] and Rankin-Selberg $L$-functions on $\text{GL}(2) \times \text{GL}(2)$ [RW], this is known for $n \leq 4$ (for which Theorems 1.4 and 1.5 become unconditional). See [Br] for general zero-free regions (which however is not of the strength required in (1.4)).

An inspection of the proof of Lemma 8.1 shows that we only need a certain averaged version of Hypothesis 1. After this paper was submitted, Jana and Kamber managed to prove this, cf. [JK, Section 1.3]. In particular, Theorems 1.4 and 1.5 are now unconditional.

1.4. Some remarks on the numerology. We end the introduction with a quick synopsis of the exponents appearing in the various statements presented so far. The spectral side of the Kuznetsov formula (Lemma 5.1) for the group $\Gamma(q)^2$ features an expression of the type

\[(1.5) \quad \int_{\mu = \ll 1} |A_{\sim}(1)|^2 T^{2\sigma_{w,v}} d\varepsilon\]

(see Section 2.5 for the notation). As long as

\[(1.6) \quad T \ll q^{n+1},\]

only two Weyl group elements contribute to the Kloosterman side: the trivial Weyl element and $w_\ast$, as defined in (2.2) below. The Kloosterman sums corresponding to the other Weyl elements vanish, since their moduli are divisible by such high $q$-powers that they are outside the support of test function on the Kloosterman side (Lemma 4.2). The contribution of the trivial Weyl element is of size $N_q = q^{n(n-1)(n-2)/6}$ as defined in (2.14). For $w_\ast$, the admissible moduli $c \in \mathbb{Z}^{n-1}$ satisfy $q^n | c_j \ll T \ll q^{n+1}$, so that they are all of the form $c_j = q^n \gamma_j$ with $\gamma_j \ll q$. For simplicity let us assume $(\gamma_j, q) = 1$. We factor the Kloosterman sums to moduli $c$ into one with moduli $(q^n, \ldots, q^n)$ which is bounded in Theorem 4.4 by $N_q q^{(n-1)^2}$. This is sharp. The remaining unramified Kloosterman sum with modulus $\gamma$ is estimated trivially. In this way we conclude that the contribution of $w_\ast$ is $\ll N_q T^{n-1} q^{1-n^2}$. By (1.6) this is not more than the trivial Weyl element (which is the reason why (1.6) is the critical bottleneck). We conclude that (1.5) is bounded by $N_q$, provided that (1.6) holds. Now Theorem 1.3 implies

\[2\text{For the group } \Gamma_0(q) \text{ one can derive the original density conjecture (not the improved version of [Bl]) without any analysis of Kloosterman sums, but it is interesting to note that this is not possible here – the analysis of } w_\ast \text{ is absolutely crucial.}\]
that the weight $|A_\omega(1)|^2$ can essentially be dropped at the cost of a factor $\mathcal{N}_q/V_q$, as defined in (2.14). Dividing both sides by this factor, we conclude

$$\int_{\mu_\omega < 1} T^{2\sigma_{\omega,v}} d\omega \leq V_q^{1+o(1)},$$

provided that (1.6) holds. Of course, on the left hand side we can replace $\Gamma(q)^2$ with $\Gamma(q)$, and the right hand side is essentially $[\text{SL}_n(\mathbb{Z}) : \Gamma(q)]$, so this bound is best possible, the only restriction being (1.6) which limits the size of the penalty $T^{2\sigma_{\omega,v}}$ for large values of $\sigma_{\omega,v}$. Theorem 1.1 is now an easy consequence (see Section 6). Any numerical improvement in Theorem 1.1 needs a relaxed version of (1.6), but as remarked above, even in the most optimistic scenario this is rather small (and has no influence on the applications discussed in Subsection 1.3).

A roadmap: Section 2 contains some basic notation. For the reader’s convenience we also establish a careful dictionary between the classical and the adelic framework. Section 3 is technical in nature and refines [Bl, Section 5]; this is needed to have polynomial control on $M$ in Theorem 1.1 and all subsequent results. Section 4 is the key input on the arithmetic side of the Kuznetsov formula: a careful analysis of Kloosterman sums for the principal congruence subgroup. Section 5 sets up the Kuznetsov formula, while Section 6 provides the key input on the spectral side of the Kuznetsov formula: lower bounds for local Whittaker functions. In Section 7 we add the non-cuspidal spectrum to the picture, while the final Section 8 deals with applications.

2. Basic notation

2.1. Basic conventions. We regard $n$ as (arbitrary, but) fixed. For convenience we assume $n \geq 3$, since all statements in the introduction are known for $n = 2$. The phrase “absolute constant” refers to a constant that may depend on $n$. In this terminology, the letter $K$ denotes a sufficiently large absolute constant, not necessarily the same on each occurrence.

We write $a | b^\infty$ to mean that $a$ has only prime factors occurring in $b$. Similarly $(a,b^\infty) = \sup_{n \in \mathbb{N}} (a,b^p)$ is the greatest divisor of $a$ whose prime divisors divide $b$. We denote by $\| \|$ the maximum norm on vectors, the maximum norm on matrices and the sup-norm on functions. We denote by $I_d$ the identity matrix of dimension $d$.

Let $U \subseteq \text{GL}_n$ be the subgroup of unipotent upper triangular matrices. Let $V \subseteq \text{GL}_n$ be the group of diagonal matrices with entries $\pm 1$. Let $T \subseteq \text{GL}_n$ be the diagonal torus, $\tilde{T}(\mathbb{R}) \subseteq \text{GL}_n(\mathbb{R})$ the set of diagonal matrices with positive entries and let $Z$ denote the center of $\text{GL}_n$. We write $Z^+ \cong \mathbb{R}_{>0}$ for the subgroup of diagonal scalar matrices with positive entries and $\mathcal{H} = U(\mathbb{R})\tilde{T}(\mathbb{R})/Z^+$. Write $T_0 = T \cap \text{SL}_n$ for the diagonal torus with determinant one.

We will write $P = MN$ for a standard parabolic subgroup and its Levi decomposition. In particular there is a partition $n = n_1 + \ldots + n_k$ so that
\( M \cong \text{GL}_{n_1} \times \ldots \times \text{GL}_{n_k} \) and \( N \subseteq U \). Further let \( P^{\text{op}} = M^{\text{op}} N^{\text{op}} \) be the standard parabolic subgroup conjugate to \( P^\top \).

Let \( W \) be the Weyl group. For our purpose it is important to choose representatives with determinant 1 (since the principal congruence group \( \Gamma(q) \) has no elements with determinant \(-1\) for \( q > 2 \)), so we identify \( W \) with permutation matrices where potentially the entry in first row (say) is \(-1\). We identify a permutation matrix \( w = (w_{ij}) \in W \) with the permutation \( i \mapsto j \) for \( w_{ij} = \pm 1 \). For \( w \in W \) we define
\[
U_w = w^{-1} U^\top w \cap U.
\]

Important Weyl elements for us are the special element \( w_s \) and the long element \( w_l \) defined by
\[
w_s = \begin{pmatrix} 1 & I_{n-2}^{-1} \\ 1 & \cdots & 1 \end{pmatrix} \quad \text{and} \quad w_l = \begin{pmatrix} \pm 1 \\ \vdots \\ 1 \end{pmatrix}.
\]

2.2. Embeddings, coordinates and measures. The Haar measure on \( U(\mathbb{R}) \) is given by \( dx = \prod_{1 \leq i < j \leq n} dx_{ij} \). We write \( dx \) also for the induced measure on the subgroup \( U_w(\mathbb{R}) \).

We embed \( y = (y_1, \ldots, y_{n-1}) \in \mathbb{C}^{n-1} \) into \( T \) as
\[
i(y) = \text{diag}(y_{n-1} \cdots y_1, y_2 y_1, y_1, 1) = \text{diag}((y_1 \cdots y_{n-j} \mid 1 \leq j \leq n))
\]
with the empty product defined as 1. We multiply two elements in \( y, y' \in \mathbb{C}^{n-1} \) componentwise, written \( y \cdot y' \), so that \( i \) is a homomorphism. The Iwasawa decomposition gives a unique decomposition \( g = xyk\alpha \in \text{GL}_n(\mathbb{R}) \) with \( x \in U(\mathbb{R}) \), \( y \in \hat{T}(\mathbb{R}) \), \( k \in O_n(\mathbb{R}) \), \( \alpha \in \mathbb{Z}^+ \). We write \( y(g) = i^{-1} y \in \mathbb{R}_{>0}^{n-1} \) for \((n-1)\)-tuple of Iwasawa \( y \)-coordinates. In particular, for \( g = \text{diag}(y_1, \ldots, y_n) \in \hat{T}(\mathbb{R}) \) we have
\[
y(g) = \left( \frac{y_{n-1}}{y_n}, \ldots, \frac{y_1}{y_2} \right) \in \mathbb{R}_{>0}^{n-1}.
\]

For \( w \in W \), \( y \in \mathbb{R}_{>0}^{n-1} \) we write
\[
y(wu(y))^{-1} w^{-1} =: wy := (w y_1, \ldots, w y_{n-1})
\]
for the Iwasawa \( y \)-coordinates of \( wu(y)^{-1} w^{-1} \).

For \( \alpha \in \mathbb{C}^{n-1} \), \( y \in \mathbb{R}_{>0}^{n-1} \) we write \( y^\alpha = y_1^{\alpha_1} \cdots y_{n-1}^{\alpha_{n-1}} \in \mathbb{C} \). Let
\[
\eta = (\eta_1, \ldots, \eta_{n-1}) = \left( \frac{1}{2} j(n-j) \right)_{1 \leq j \leq n-1} \in \mathbb{Q}^{n-1}.
\]

We define a measure on \( \mathbb{R}_{>0}^{n-1} \) by
\[
d^* y = y^{-2\eta} \frac{dy_1}{y_1} \cdots \frac{dy_{n-1}}{y_{n-1}}
\]
and correspondingly an inner product by
\[
\langle f, g \rangle = \int_{\mathbb{R}_{>0}^{n-1}} f(y) \overline{g(y)} d^* y.
\]

We denote the push forward of \( d^* y \) to \( \hat{T}(\mathbb{R}) \) by \( \iota \) also by \( d^* y \).
We define a different embedding of $\mathbb{R}_{\geq 0}^{n-1}$ into $T(\mathbb{R})$ by
\begin{equation}
(2.9) \quad c = (c_1, \ldots, c_{n-1}) \mapsto e^c = \text{diag}\left(\frac{1}{c_{n-1}}, \frac{c_{n-1}}{c_{n-2}}, \ldots, \frac{c_2}{c_1}, c_1\right).
\end{equation}

2.3. Subgroups and characters. For $N \in \mathbb{Z}^{n-1}$ we define a character
\begin{equation}
(2.10) \quad \theta_N : U(\mathbb{R})/U(\mathbb{Z}) \to S^1 \text{ by }
\theta_N(x) = e(N_{n-1}x_{12} + \ldots + N_1x_{n-1,n}).
\end{equation}
For $v \in V$ we write $\theta^v_{\mathcal{N}}(x) = \theta_N(v^{-1}xv)$.

Let $q \in \mathbb{N}$ and $\Gamma(q) \subseteq \text{SL}_n(\mathbb{Z})$ be the principal congruence subgroup. Let
\begin{equation}
(2.11) \quad D_q = \text{diag}(q^{n-1}, \ldots, q, 1)
\end{equation}
and write
\begin{equation}
(2.12) \quad \Gamma(q)^2 = D_q^{-1}\Gamma(q)D_q = \left(\begin{array}{cccc}
1 + q\mathbb{Z} & \frac{1}{q}\mathbb{Z} & \ldots & \frac{1}{q^n-1}\mathbb{Z} \\
q^2\mathbb{Z} & 1 + q\mathbb{Z} & \mathbb{Z} & \ldots & \frac{1}{q^n-1}\mathbb{Z} \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
q^{n-1}\mathbb{Z} & \ldots & 1 + q\mathbb{Z} & \mathbb{Z} & \\
q^n\mathbb{Z} & q^{n-1}\mathbb{Z} & \ldots & q^2\mathbb{Z} & 1 + q\mathbb{Z}
\end{array}\right) \cap \text{SL}_n(\mathbb{Q}).
\end{equation}
In contrast to $\Gamma(q)$, the conjugated group $\Gamma(q)^2$ has the advantage that
\begin{equation}
(2.13) \quad U(\mathbb{Z}) \subseteq \Gamma(q)^2 \quad \text{and} \quad \theta_N \text{ is trivial on } \Gamma(q)^2 \cap U(\mathbb{R}).
\end{equation}
In particular, $U(\mathbb{Z})$ acts on $\Gamma(q)^2$ both from the left and the right. Let
\begin{equation}
(2.14) \quad \mathcal{V}_q = [\text{SL}_n(\mathbb{Z}) : \Gamma(q)] = q^{(n^2-1)+o(1)}, \\
\mathcal{N}_q = [\Gamma(q)^2 \cap U(\mathbb{Q}) : U(\mathbb{Z})] = q^{n(n-1)(n-2)/6}.
\end{equation}
These are the quantities appearing in Theorem 1.3.

2.4. Whittaker functions. For $\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{C}^n$ satisfying
\begin{equation}
(2.15) \quad \mu_1 + \ldots + \mu_n = 0, \quad \{\mu_1, \ldots, \mu_n\} = \{-\bar{\mu}_1, \ldots, -\bar{\mu}_n\}
\end{equation}
we define $\nu \in \mathbb{C}^{n-1}$ by
\begin{equation}
\nu_j = \frac{1}{n}(\mu_j - \mu_{j+1}), \quad 1 \leq j \leq n-1.
\end{equation}
We denote by $W_\mu : \mathbb{R}_{\geq 0}^{n-1} \to \mathbb{C}$ the normalized Whittaker function
\begin{equation}
(2.16) \quad W_\mu(y) = \int_{U(\mathbb{R})} I_\nu(w x \iota(y)) \theta_{(1, \ldots, 1)}(x)^{-1} dx
\end{equation}
with the power function
\begin{equation}
I_\nu(g) = \prod_{i,j=1}^{n-1} |y(i,j)|^{b_{ij}(\frac{1}{n} + \nu_j)}, \quad b_{ij} =\begin{cases}
ij, & i + j \leq n, \\
(n-i)(n-j), & i + j \geq n.
\end{cases}
\end{equation}
This differs from the completed Whittaker function by a factor
\( G(\mu) := \prod_{1 \leq j \leq k \leq n-1} \Gamma_{\mathbb{R}}(1 + n(\nu_j + \ldots + \nu_k)) \), \( \Gamma_{\mathbb{R}}(s) = \Gamma(s/2)\pi^{-s/2} \).

2.5. Automorphic forms. In the interest of a compact notation, for a congruence subgroup \( \Gamma \subset \text{SL}_n(\mathbb{R}) \) we denote by
\( (2.18) \int_{\Gamma} V_\varpi d\varpi \)
a spectral decomposition of \( L^2(\Gamma \backslash \text{SL}_n(\mathbb{R})/\text{SO}_n(\mathbb{R})) \) respecting the unramified Hecke algebra. Note that the spaces \( V_\varpi \) are one-dimensional, and we use \( \varpi \) to denote a suitably normalized generator of \( V_\varpi \). The standard inner product is
\[ (2.19) \langle \varpi_1, \varpi_2 \rangle = \int_{\Gamma \backslash \mathcal{H}} \overline{\varpi_1(xy)} \varpi_2(xy) dx dy. \]

The relevant spectral decomposition is a special case of Langlands’ general theory; see e.g. [Ar1] for a convenient summary in adelic language. Each element \( \varpi \) in the decomposition above has spectral parameter \( \mu_\varpi = (\mu_1, \ldots, \mu_n) \) related to the eigenvalues of invariant different operators. This spectral parameter agrees with the local archimedean Langlands parameter \( \mu_{\pi_\infty} \) of the associated automorphic representation \( \pi = \otimes_v \pi_v \) and satisfies \( (2.15) \).

We now specialize to the group \( \Gamma = \Gamma(q)^\sharp \), defined in \( (2.12) \). Set
\[ X_q = \Gamma(q)^\sharp \backslash \text{SL}_n(\mathbb{R})/\text{SO}_n(\mathbb{R}). \]

For the adelisation process described below we will need the following decomposition of the space \( L^2(X_q) \). Let \( \Gamma_1(q) \subset \text{SL}_n(\mathbb{Z}) \) be the set of matrices that are diagonal modulo \( q \). Put \( \Gamma_1(q)^2 = D_q^{-1} \Gamma_1(q) D_q \). Given \( \chi = (\chi_1, \ldots, \chi_{n-1}) \), where \( \chi_1, \ldots, \chi_{n-1} \) are Dirichlet characters modulo \( q \), we set
\[ L^2(X_q, \chi) = \{ f \in L^2(X_q) : f(\gamma g) = \chi_1(\gamma_{11}) \cdots \chi_{n-1}(\gamma_{n-1,n-1}) \cdot f(g) \}
\text{for all } g \in \text{SL}_n(\mathbb{R}), \gamma = (\gamma_{ij}) \in \Gamma_1(q)^\sharp \}, \]

We have the spectral refinement
\[ L^2(X_q) = \bigoplus_{\chi \in \hat{T}_0(\mathbb{Z}/q\mathbb{Z})} L^2(X_q, \chi). \]

For \( \varpi \in L^2(X_q) \) and \( N \in \mathbb{Z}^{n-1} \) we denote by
\[ \mathcal{W}_\varpi(y; N) = \int_{U(\mathbb{Z}) \backslash U(\mathbb{R})} \varpi(xy) \theta_N^{-1}(x) dx \]
its \( N \)-th Fourier coefficient. This is a multiple of the Whittaker function \( W_\mu \) as in \( (2.16) \) associated with the archimedean Langlands parameters \( \mu = \mu_\varpi \).
and using (2.6) we decompose it as
\[ W_\infty(y; N) = \frac{A_\infty(N)}{N^{2n}} W_\mu(N \cdot y(y)) \]
for some constant \( A_\infty(N) \in \mathbb{C} \). If \( \lambda_\infty(m) \) denotes the Hecke eigenvalue with respect to the Hecke operator \( T_m \) for \( m \in \mathbb{N} \), \( (m, q) = 1 \), then
\[ A_\infty((m, 1, \ldots, 1)) = \lambda_\infty(m) A_\infty((1, \ldots, 1)). \]

For a prime \( p \mid q \) and \( \nu > n \) we have
\[ \max_{0 \leq j \leq n-1} |\lambda_\infty(p^{\nu-j})| \geq (2p^{\sigma_{\omega,p}})^{1-n} p^{\nu \sigma_{\omega,p}} \]
by [Bl, Lemma 4].

2.6. Adelising Automorphic Forms.

2.6.1. The adelic space. Let \( \mathbb{A} \) denote the adele ring of \( \mathbb{Q} \). Let \( \psi_\mathbb{A} = \prod_v \psi_v \) be the usual adelic character on \( \mathbb{Q} \backslash \mathbb{A} \), where \( \psi_\infty = e(\cdot) \). We lift \( \psi_\mathbb{A} \) to a character
\[ \psi_\mathbb{A}: U(\mathbb{Q}) \backslash U(\mathbb{A}) \to \mathbb{C}^\times, \quad x \mapsto \psi(x_{12} + \ldots + x_{n-1,n}), \]
so that the archimedian part is \( \psi_\mathbb{Q}_\infty = \theta(1, \ldots, 1) \).

Given a primitive Dirichlet character \( \xi \) we write \( \omega_\xi: \mathbb{Q}^\times \backslash \mathbb{A}^\times \to S^1 \) for the unique Hecke character so that the completed Dirichlet \( L \)-function associated to \( \xi \) agrees with the \( L \)-function associated to \( \omega_\xi \) by Tate. If \( \xi \) is a non-primitive Dirichlet character, then we write \( \omega_\xi \) for the Hecke character associated to the underlying primitive character (sometimes called the idelic lift of \( \xi \)). In either case we have the factorization \( \omega_\xi = \prod_v \omega_{\xi,v} \).

We set \( K_\infty = \text{SO}_n(\mathbb{R}) \), \( K_p = \text{GL}_n(\mathbb{Z}_p) \), \( K_\text{fin} = \prod_p K_p \) and \( K = K_\infty \times K_\text{fin} \).

Further, we define \( \Gamma(q) \subset K_\text{fin} \) by requiring
\[ \Gamma(q) = \text{GL}_n(\mathbb{Q}) \cap (\mathbb{Z}^+ \text{SL}_n(\mathbb{R}) \times K(q)). \]

The local components \( K_p(q) \) of \( K(q) \) are simply the local principal congruence subgroups
\[ (2.22) \quad K_p(q) = \ker(K_p \to \text{GL}_n(\mathbb{Z}_p/q\mathbb{Z}_p)). \]
Also define \( K(q) = D_q^{-1} K(q) D_q \), where we embed \( D_q \) diagonally in \( \text{GL}_n(\mathbb{A}_\text{fin}) \).

This is the adelic version of \( \Gamma(q) \). In order to apply strong approximation we enlarge \( K(q) \) and \( K(q) \) slightly. Put \( K_1(q) = T(\mathbb{Z}) K(q) \) and \( K_1(q) = D_q^{-1} K_1(q) D_q \). For this open compact subgroup the map \( \det: K_1(q) \to \mathbb{Z}^\times \) is surjective. By strong approximation we get \( \text{GL}_n(\mathbb{A}) = \text{GL}_n(\mathbb{Q}) \cdot (\mathbb{Z}^+ \text{SL}_n(\mathbb{R}) \times K_1(q)) \).

The inner product is given by
\[ (2.24) \quad \langle F, G \rangle = \int_{\text{GL}_n(\mathbb{Q}) \backslash \text{GL}_n(\mathbb{A})^1} F(g) \overline{G(g)} dg, \]
where we use the Tamagawa measure, which satisfies \( \text{Vol}(GL_n(\mathbb{Q}) \backslash GL_n(\mathbb{A})^1, dg) = 1 \). Note that we use \( GL_n(\mathbb{A}) \), which is given by those \( g \in GL_n(\mathbb{A}) \) with \( |\text{det}(g)|_\mathbb{A} = 1 \), in order to treat possibly different central characters (all of which will be trivial on \( Z^+ \)) simultaneously.

We can decompose the space \( L^2(X_q^A) \) similarly as the space \( L^2(X_q) \) using the bigger group \( K_1(q)^2 \). To do so, given an \( n \)-tuple \( \zeta = (\zeta_1, \ldots, \zeta_n) \) of Dirichlet characters modulo \( q \), we define \( \zeta : K_1(q)^2 \rightarrow S^1 \) by

\[
\zeta(D_q^{-1}kD_q) = \prod_{p|q} \omega_{\zeta_1, p}((k_p)_{11}) \cdots \omega_{\zeta_n, p}((k_p)_{nn}).
\]

Define \( L^2(X_q^A, \zeta) = \{ F \in L^2(X_q^A) : F(\gamma x) = \zeta(k)F(g) \text{ for all } g \in GL_n(\mathbb{A})^1, k \in K_1(q)^2 \} \). Note that we have the decomposition \( L^2(X_q^A) = \bigoplus_\zeta L^2(X_q^A, \zeta) \). Furthermore, given \( F \in L^2(X_q^A, \zeta) \) the central character of \( F \) is determined by \( \zeta \). Indeed we have

\[
F(zg) = \left( \prod_{i=1}^n \omega_{\zeta_i}(a) \right) F(g), \quad g \in GL_n(\mathbb{A})^1, \quad z = \text{diag}(a, \ldots, a), \quad |a|_\mathbb{A} = 1.
\]

### 2.6.2. Adelization and de-adelization.

The de-adelisation map \( \mathcal{D} : L^2(X_q^A) \rightarrow L^2(X_q) \) is given by \( [DF](x) = F(x) \), where we view \( x \in SL_n(\mathbb{R}) \subset GL_n(\mathbb{A})^1 \). Sometimes we will write \( DF = \varpi_F \). Suppose \( F \in X_q^A(\zeta) \), then one checks that \( \varpi_F(\gamma x) = [\zeta_1 \zeta_n^{-1}]([\zeta_1] \cdots [\zeta_{n-1} \zeta_n^{-1}]) [\gamma_{n-1,n-1} \cdots \gamma_{11}] \varpi_F(x) \) for all \( \gamma \in K_1(q)^2 \). Therefore the de-adelisation map restricts to

\[
\mathcal{D} : L^2(X_q^A, \zeta) \rightarrow L^2(X_q, \chi_\zeta),
\]

where \( \chi_\zeta = (\zeta_1 \zeta_n^{-1}, \ldots, \zeta_{n-1} \zeta_n^{-1}) \).

We come to the adelisation procedure. Given \( \varpi \in L^2(X_q, \chi) \) and a Dirichlet character \( \xi \) modulo \( q \) we define

\[
[A(\xi)]\varpi(g) = \zeta_{\chi, \xi}(k)\varpi(g_{\infty}) \text{ for } g = \gamma z g_{\infty} k \in GL_n(\mathbb{Q})Z^+SL_n(\mathbb{R})K_1(q)^2,
\]

for \( \zeta_{\chi, \xi} = (\chi_1, \ldots, \chi_{n-1}, \xi, \xi). \) This is a slight modification of the standard adelisation procedure for \( GL_n \) as described in [GH, Section 13.4]. We call \( F_{(\xi)}^{(\varpi)} = A(\xi)\varpi \) the \( \xi \)-adelic lift of \( \varpi \). Note that the choice of the character \( \xi \) is the only redundancy in the lifting procedure. One easily verifies that \( F_{(\xi)}^{(\varpi)} \) is well-defined, satisfies \( F_{(\xi)}^{(\varpi)}(\gamma k) = \zeta_{\chi, \xi}(k)F_{(\varpi)}(g) \) for all \( g \in GL_n(\mathbb{A})^1 \) and all \( k \in K_1(q)^2 \) and that the central character of \( F_{(\varpi)} \) is \( \omega_{\chi_1} \cdots \omega_{\chi_{n-1}} \cdot \omega_\xi^{(1)} \).

Clearly we have \( \mathcal{D} \circ A(\xi)\varpi = \varpi \). One also sees that the redundancy in the lifting procedure is manifested by twisting with \( \omega_\xi \) on the adelic side. This is reflected in the following diagram:

\[
\begin{array}{c}
L^2(X_q, \chi) \overbrace{\text{A}(\xi)}^\sim \rightarrow L^2(X_q, \chi) \\
\downarrow \quad \quad \downarrow \\
L^2(X_q^A, \zeta_{\chi, 1}) \overbrace{\omega_\xi^{(\otimes)}} \rightarrow L^2(X_q^A, \zeta_{\chi, \xi})
\end{array}
\]

Recall that the twisting operation \( \omega_\xi^{(\otimes)} \) is defined by \( F \mapsto \omega_\xi(\det(\cdot))F \).
If \( \varpi \in L^2(X_q, \chi) \) is cuspidal, then so is the adelisation \( F_{\varpi}^{(\xi)} \). A special case of this computation shows that
\[
W_{F_{\varpi}^{(\xi)}}(y) := \int_{U(\mathbb{Q}) \backslash U(\mathbb{A})} F_{\varpi}^{(\xi)}(uy) \psi_{\varpi}(u)^{-1} du = W_{\varpi}(y) = A_{\varpi}(1) W_{\varpi}(y)
\]
for \( y \in \tilde{T}(\mathbb{R}) \subseteq \text{GL}_n(\mathbb{A}) \).

2.6.3. Spectral decomposition. Suppose that \( \varpi \) is a cuspidal joint eigenfunction of the ring of invariant differential operators and the unramified Hecke algebra, then \( F_{\varpi}^{(\text{triv})} \) generates a cuspidal automorphic representation \( \pi_{\varpi} \) and \( F_{\varpi}^{(\xi)} \) generates \( \omega_{\xi} \otimes \pi_{\varpi} \). Given a cuspidal automorphic representation \( \pi \) we abuse notation and write
\[
(2.25) \quad \pi | L^2(X_q)
\]
if there is \( \varpi \in L^2(X_q, \chi) \subset L^2(X_q) \) with \( \omega_{\xi} \otimes \pi_{\varpi} \cong \pi \) for some Dirichlet character \( \xi \) modulo \( q \). In order to exploit the language of (adelic) automorphic representations, in particular the inherited factorisation properties, we want to construct a spectral decomposition for \( L^2(X_q) \) parametrized by automorphic representations \( \pi | L^2(X_q) \). To do so it is important to keep the twist redundancy in mind. Let us make this precise. Each irreducible unitary automorphic representation \( \pi \) of \( \text{GL}_n(\mathbb{A}) \) has a finite dimensional subspace \( \pi^{K(q)^2} \) of right \( K(q)^2 \)-invariant elements that are spherical at infinity. If \( \pi^{K(q)^2} \neq \{0\} \), then \( \pi | L^2(X_q) \). In general we write \( \text{ONB}(\pi^{K(q)^2}) \) for an orthonormal basis of this space and we put
\[
V_{\pi} = D(\pi^{K(q)^2}) \subset L^2(X_q).
\]
This space \( V_{\pi} \) is precisely the linear subspace of \( L^2(X_q) \) associated to a generic irreducible automorphic representation that is studied in Theorem 1.3. (Note that it is finite-dimensional even though \( \pi \) itself is infinite-dimensional.) We write the cuspidal part of spectral decomposition of \( L^2(X_q) \) as
\[
(2.26) \quad L^2_{\text{cusp}}(X_q) = \bigoplus'_{\pi | L^2(X_q)} V_{\pi},
\]
where \( \bigoplus' \) stands for the direct sum of the corresponding automorphic representations modulo character twists. It now becomes clear why we introduced different versions of the adelisation process. Indeed, in order to go from \( V_{\pi} \) to \( \pi^{K(q)^2} \) we need to combine all of them as follows. Write \( V_{\pi}(\chi) = V_{\pi} \cap L^2(X_q, \chi) \) and \( \pi^{(K_1(q)^2, \zeta)} = \pi \cap L^2(X_q, \zeta) \). If \( V_{\pi}(\chi) \neq \{0\} \), then there is a unique adelisation map \( A: V_{\pi} \to \pi^{K(q)^2} \) given fibrewise by \( A(\xi): V_{\pi}(\chi) \to \pi^{(K_1(q)^2, \zeta_{\chi, \xi})} \) for a suitable Dirichlet characters \( \xi \) modulo \( q \). depending on \( \chi \).
2.6.4. \textit{Inner products}. Finally we need to investigate how the inner products behave under adélation. This is standard, but since the exact $q$-dependence is absolutely crucial, we give some details. First note that if $\chi_1 \neq \chi_2$, then

$$\langle F^{(\xi)}_{\omega_1}, F^{(\xi)}_{\omega_2} \rangle = 0 = \langle \omega_1, \omega_2 \rangle$$

for $\omega_1 \in L^2(X_q, \chi_1)$ and $\omega_2 \in L^2(X_q, \chi_2)$.

Thus we assume $\chi_1 = \chi_2$. Using strong approximation for $K_1(q)^{\dagger}$ one obtains

$$\langle F^{(\xi)}_{\omega_1}, F^{(\xi)}_{\omega_2} \rangle = \text{Vol}(K_\infty K_1(q)^{\dagger}) \cdot \int_{\text{GL}_n(Q) \backslash \text{GL}_n(A)^1 / K_\infty K_1(q)^{\dagger}} \overline{F^{(\xi)}_{\omega_1}(x) F^{(\xi)}_{\omega_2}(x)} dx$$

$$= C_{\text{meas}} \text{Vol}(K_\infty K_1(q)^{\dagger}) \cdot \int_{\Gamma_1(q)^{\dagger} \backslash \mathcal{H}} \overline{\omega_1(xy)} \overline{\omega_2(xy)} dx dy$$

$$= C_{\text{meas}} \text{Vol}(K_\infty K_1(q)^{\dagger}) \left[ \Gamma_1(q)^{\dagger} \right]^{-1} \cdot \langle \omega_1, \omega_2 \rangle$$

for some proportionality constant $C_{\text{meas}}$. We can evaluate both sides with $\omega_1 = \omega_2 \equiv 1$ leading to

$$C_{\text{meas}} \text{Vol}(K_\infty K_1(q)^{\dagger}) \left[ \Gamma_1(q)^{\dagger} \right]^{-1} = \text{Vol}(\Gamma(q)^{\dagger} \backslash \mathcal{H})^{-1}$$

$$= |\text{SL}_n(\mathbb{Z}) : \Gamma(q)|^{-1} \text{Vol}(\text{SL}_n(\mathbb{Z}) \backslash \mathcal{H})^{-1}.$$ 

In summary we have

$$\langle F^{(\xi)}_{\omega_1}, F^{(\xi)}_{\omega_2} \rangle = n \cdot \left( \prod_{l=2}^{n} \frac{\pi^{l/2}}{\Gamma(l/2) \zeta(l)} \right) \cdot \frac{\langle \omega_1, \omega_2 \rangle}{V_q}, \quad \omega_j \in L^2(X_q, \chi_j).$$

Let $\pi \mid L^2(X_q)$ be a $\psi_A$-generic (irreducible) cuspidal representation realised in $L^2_{\text{cusp}}(\text{GL}_n(Q) \backslash \text{GL}_n(A)^1)$. We realise the contragredient $\pi$ as $\{ F : F \in \pi \}$. The inner product defined in (2.24) restricts to an invariant inner product on $\pi$ which we denote by $\langle \cdot, \cdot \rangle_{\pi}$. In particular $\| F \|_{L^2}^2 = (F, F)_{\pi}$. By Flath’s factorization theorem we can write $\pi = \otimes_v \pi_v$ and also the Whittaker model $\text{Wh}(\pi) = \bigotimes_v \text{Wh}(\pi_v)$ factors. Note that for $\omega \in V_\pi$ we have $\sigma_{\omega, v} = \sigma_{\pi_v}$ for all unramified places $v$. This always makes sense since $\sigma_{\pi_v}$ is independent under unitary unramified twists.

We fix any invariant inner product $\langle \cdot, \cdot \rangle_{\text{Wh}(\pi_v)}$ on the local Whittaker model $\text{Wh}(\pi_v)$. For $W_v \in \text{Wh}(\pi_v)$ (whose $L^2$-norm is nonzero) we define

$$I_v(W_v) = \int_{U(Q_v)} \frac{\langle \pi_v(u) W_v, W_v \rangle_{\text{Wh}(\pi_v)} \text{Wh}(\pi_v)}{\langle W_v, W_v \rangle_{\text{Wh}(\pi_v)}} \psi_q(u) \cdot u^{-1} du.$$ 

This is independent of the choice of the invariant inner product and automatically normalises $W_v$. Suppose now that $F \in \pi$ corresponds to a pure tensor in $\otimes_v \pi_v$. In this case the Whittaker coefficient factors as $W_F = \bigotimes_v W_{F,v}$. Fix a finite set $S$ of places containing $\infty$ and all places where $F$ is unramified. According to [LM, Lemma 4.4] and the Rankin-Selberg method we have

$$\frac{|W_F(1)|^2}{\langle F, F \rangle} = \lim_{s \to 1} \frac{\Delta_S^\pi(q^n)(s)}{L^S(s, \pi \times \pi)} \prod_{v \in S} I_v(W_{F,v}),$$
where $\Delta_{\text{GL}_n}(s) = \prod_{j=1}^{\infty} \zeta^s(s + j - 1)$. We apply this formula to $F = A(\varpi)$ with $\varpi \in \mathcal{V}_n$. In this case we can take $S = \{\infty\} \cup \{p \mid q\}$ and get

$$\langle W_{\mu_{\varpi}}, W_{\mu_{\varpi}} \rangle^{-1} = C(\pi) \cdot \prod_{p \mid q} I_{p}(W_{F,p}),$$

for

$$C(\pi) = n \cdot \left( \frac{\pi^{1/2}}{\Gamma(l/2) \zeta(l)} \right) \cdot \lim_{s \to 1} \frac{\Delta_{\text{GL}_n}(s)}{L^S(s, \pi \times \tilde{\pi})} \cdot \langle W_{\mu_{\varpi}}, W_{\mu_{\varpi}} \rangle^{-1} \cdot \langle W_{\mu_{\varpi}}, W_{\mu_{\varpi}} \rangle_{\text{Wh}(\pi_{\infty})},$$

where we recall (2.27). Equation (2.28) will be crucial in the proof of Theorem 1.3. We will also need a lower bound for $C(\pi)$, which can be obtained as follows. First note that

$$C(\pi) \gg q^{-\varepsilon} \cdot |L(1, \pi, \text{Ad})|^{-1} \cdot \langle W_{\mu_{\varpi}}, W_{\mu_{\varpi}} \rangle_{\text{Wh}(\pi_{\infty})}^{-1}. $$

The remaining archimedean inner product in the Whittaker model is $\asymp 1$ by Stade’s formula (see the remark after (3.4) below). Finally the adjoint $L$-function can be bounded from above using [Li] and we obtain

$$C(\pi) \gg (\|\mu_{\infty}\|q)^{-\varepsilon}. $$

(Note that we only need upper bounds for $C(\pi)$, lower bounds are in general not available unconditionally for $n \geq 3$.)

Our discussion was focused on adelising the cuspidal part of $L^2(X_q)$. Essentially the same discussion applies to the $L^2$-space of

$$X(q) = \Gamma(q) \backslash \text{SL}_n(\mathbb{R})/\text{SO}_n(\mathbb{R})$$

as well. We can now give a new description of $m_{\Gamma(q)}(\pi_v)$ in terms of automorphic representations. Indeed, for an irreducible unramified representation $\eta$ of $\text{GL}_n(\mathbb{Q}_v)$, we have

$$m_{\Gamma(q)}(\eta_v) = \sum' \dim(V_{\pi}).$$

The non-cuspidal part of $L^2(X(q))$ can be adelised and de-adelised similarly.

2.7. Spherical functions and spherical transform on $\text{SL}_n(\mathbb{R})$. The content of this subsection will only be used in Section 8. In this subsection, we write $G = \text{SL}_n(\mathbb{R})$, $K = \text{SO}_n(\mathbb{R})$ and $A \subseteq \text{SL}_n(\mathbb{R})$ for the group of diagonal matrices with positive entries and determinant 1, and we decompose $G = KAU(\mathbb{R}) = U(\mathbb{R})AK$. Accordingly, we have the Iwasawa projections $^3 A, H : G \to A$ such that $g \in K \exp(H(g))U(\mathbb{R}) = U(\mathbb{R}) \exp(A(g))K$. We denote by $du, da, dk, dg$ the usual Haar measures. As usual, we write

---

^3The double use of $A$ will not lead to confusion.
\[ \rho = \left( \frac{n-1}{2}, \frac{n-3}{2}, \ldots, \frac{3-n}{2}, \frac{1-n}{2} \right) \in \mathfrak{a}^* \text{ and for } \mu \in \mathfrak{a}^*_C \text{ we define the spherical function (cf. [He, p. 418 & p. 435] with } \mu \text{ in place of } i\lambda) \]

\[ (2.31) \quad \phi_\mu(g) = \int_K e^{-(\rho+\mu)H(gk)}dk = \int_K e^{(\rho+\mu)A(kg)}dk. \]

For a smooth, 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

\[ A_f(a) = e^{\rho(\log a)} \int_{U(\mathbb{R})} f(au)du, \quad a \in A, \]

and the Fourier transform

\[ \tilde{f}(\mu) = \int_A A_f(a)e^{-\mu \log a}da. \]

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 \( z, w \in \Gamma(q)\backslash SL_n(\mathbb{R})/SO_n(\mathbb{R}) = \Gamma(q)\backslash \mathcal{H} \) 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

\[ (2.32) \quad \sum_{\gamma \in \Gamma(q)} f(w^{-1}\gamma z) = \int_{\Gamma(q)} \tilde{f}(\mu)\varpi(\gamma \varpi)\varpi(\gamma)\varpi(d\varpi). \]

We will apply this with the following test function. Let \( T \geq 2 \) and \( f_0: \mathbb{R}_{>0} \to [0,1] \) be a fixed smooth, non-negative function with \( f_0(x) = 1 \) for \( x \in [0,n] \) and \( f_0(x) = 0 \) for \( x > 2n \). Let \( \|g\|_F = \text{tr}(g^\top g)^{1/2} \geq \sqrt{n} \) denote the bi-K-invariant Frobenius norm, and define

\[ (2.33) \quad f(g) = f_0\left( \frac{\|g\|_F}{T} \right), \quad g \in G. \]

Note that \( \|g\| \leq \|g\|_F \leq n\|g\| \) for \( g \in G \). For \( a = \text{diag}(a_1, \ldots, a_{n-1}, (a_1 \cdots a_{n-1})^{-1}) \in A \subseteq G \) it is easy to see that

\[ \partial_{a_1}^{i_1} \cdots \partial_{a_{n-1}}^{i_{n-1}} A_f(a) \ll T^{n(n-1)/2}T^{-i_1-\cdots-i_{n-1}}\delta_{\|a\| \leq 2n^2T} \]

for \( i_1, \ldots, i_{n-1} \in \mathbb{N}_0 \). For \( \mu \) satisfying (2.15) and \( a \in A \) with \( \|a\| \ll T \) we have \( e^{\mu \log a} \ll T^{n\|\rho\|_F} \), and so by partial integration we obtain

\[ (2.34) \quad \tilde{f}(\mu) \ll \varepsilon, B (1 + \|\mu\|)^{-B} T^n \left( \frac{1}{\|a\|} + \|\rho\| + \varepsilon \right) \]

for any fixed constants \( B, \varepsilon > 0 \). On the other hand, since \( \phi_{\pm \rho}(g) = 1 \), we have

\[ (2.35) \quad \tilde{f}(\rho) = \int_G f(g)dg \gg T^{n(n-1)}. \]
3. Bounds for Whittaker Functions

In this section we quote the analysis of Whittaker functions $W_\mu$ from [Bl, Section 5] and make sure that all bounds are polynomial in $\mu$. The main result is Lemma 3.2 below that states in a precise form that a large value of $\sigma_\pi(\infty)$ is reflected in the growth of the Whittaker function near the origin. This is the archimedean analogue of (2.21). We always assume that $\|\Re \mu\| < 1/2$. As in [Bl], we follow Stade [St1, St2] and renormalize the Whittaker function (2.16) by

$$W_\mu^*(y) = \pi^{(n-1)n(n+1)/12} y^{-n/2} W_{2\mu} \left(\sqrt{y_1/\pi}, \ldots, \sqrt{y_{n-1}/\pi}\right).$$

The inclusion of the normalization factor (2.17) (which is independent of $y$) is now very useful, because it ensures that the Whittaker function is on a polynomial scale in $\mu$: by [BHM, Theorem 2] we have

$$W_\mu^*(y) \ll \left(1 + \|\mu\|\right)^K \left(\|y\| + \frac{1}{\min y_j}\right) \exp\left(-\frac{\|y\|}{K(1 + \|\mu\|)}\right).$$

Since $W_\mu^*(y)$ is holomorphic in $\mu$ in $\|\Re \mu\| < 1/2$, by Cauchy’s formula the same bound holds for the derivatives with respect to $\mu$:

$$\nabla_\mu W_\mu^*(y) \ll \left(1 + \|\mu\|\right)^K \left(\|y\| + \frac{1}{\min y_j}\right) \exp\left(-\frac{\|y\|}{K(1 + \|\mu\|)}\right).$$

Stade’s important formula [St2] reads

$$\int_{\mathbb{R}^{n-1}_+} |W_\mu^*(y)|^2 \det(\ell(y)) s^* y = \frac{C \pi^{\ell(s,\mu)} \prod_{j,k=0}^n \Gamma(s + \mu_j - k \mu)}{\Gamma(n s) G(2\mu) G(-2\mu)}$$

with $G$ as in (2.17) and $d^* y$ as in (2.7), where $C \in \mathbb{R}_{>0}$ and $\ell$ is a linear form with real coefficients, and the left hand side converges absolutely for $\Re s > \max_{j,k} \Re(\mu_j - \mu_k)$. The key point of our normalization here is that the exponential behaviour in $\mu$ in the numerator and denominator cancels, and in fixed vertical strips the right hand side is on a polynomial scale in $\mu$. In particular, we can now justify (2.29) since (recall (3.1))

$$\langle W_\mu, W_\mu \rangle = \int_{\mathbb{R}^{n-1}_+} |W_\mu(y)|^2 \det(\ell(y)) s^* y \times \frac{\prod_{j,k=1}^n \Gamma(\frac{\ell(s,\mu)}{2}(1 + \mu_j - k \mu))}{G(\mu) G(-\mu)} \asymp 1.$$

For $\mu$ satisfying (2.15) and $\|\Re \mu\| < 1/2$ we define $\mathcal{I}_n(\mu)$ to be the set of $y \in \mathbb{R}^{n-1}_{>0}$ such that

$$(2 + \|\mu\|)^{-K} \leq \min_j y_j \leq \max_j y_j \leq (2 + \|\mu\|)^K,$$

$$(3.5) |W_\mu^*(y)| \geq (2 + \|\mu\|)^{-K} \text{ for } |\tilde{\mu} - \mu| \leq (2 + \|\mu\|)^{-K}.$$

Here we recall our basic convention on $K$ from Section 2.1. Choosing $s$ (real and sufficiently large) it is easy to conclude from (3.2), (3.4) and (3.3) that

$$\int_{\mathcal{I}_n(\mu)} |W_\mu^*(y)|^2 d^* y \gg (2 + \|\mu\|)^{-K}.$$
We now consider the Mellin transform
\[ \widehat{W}_\mu^*(s) = \int_{\mathbb{R}_+^n} W_\mu^*(y) y^s \prod_{j=1}^n \frac{dy_j}{y_j} \]
which by [FG] is meromorphic in \( s \in \mathbb{C}^{n-1} \). For \( \Re s_j \geq K \) it follows from (3.2) and from [St1, Theorem 3.1] that
\[ \widehat{W}_\mu^*(s) \ll \min \left( (1 + \|\mu\|)^K, \exp \left( K \|\mu\| - \|s\| \right) \right). \]
By [FG, (2.1)], this bound (possibly with a different value of \( K \)) remains true for all \( s \) with distance \( \geq \varepsilon \) from poles. It follows from [St1, Section 3] that
\[ \widehat{W}_\mu^*(s) = \widehat{W}_{\mu^*}^*(s) \prod_{j=1}^n (s_1 + \mu_j) \]
is holomorphic for \( \Re s_1 > \sigma_\pi(\infty) - 1 \) and \( \Re s_2, \ldots, \Re s_{n-1} \geq K \), and
\[ \widehat{W}_\mu^*(s) = \prod_{j=1}^n (s_1 + \mu_j). \]
where
\[ s^{(j)} = (s_2, \ldots, s_{n-1}) + \left( \frac{n-2}{n-1}, \ldots, \frac{1}{n-1} \right) \mu_j, \]
\[ \mu^{(j)} = (\mu_1, \ldots, \mu_{j-1}, \mu_{j+1}, \ldots, \mu_{n-1}) + \frac{\mu_j}{n-1} \cdot 1. \]
For \( \beta \in \mathbb{C} \) let \( \mathcal{D}_\beta = -y^j \partial_y + \beta \). This is a commutative family of differential operators that under Mellin transformation correspond to multiplication with \( s + \beta \). Assume (without loss of generality by (2.15) and Weyl group symmetry) that \( \Re(-\mu_1) = \sigma_\pi(\infty) \), and let
\[ \widehat{W}_\mu(s) := \frac{\widehat{W}_\mu^*(s)}{s_1 + \mu_1} = \widehat{W}_{\mu^*}^*(s) \prod_{j=2}^n (s_1 + \mu_j). \]
Taking inverse Mellin transforms, we obtain
\[ W_\mu(y) = \mathcal{D}_{\mu_2} \cdots \mathcal{D}_{\mu_n} W_\mu^*(y) \]
where the differential operators are applied to the first variable. By Mellin inversion, (3.8) and (3.7), a contour shift past the simple pole at \( s_1 = -\mu_1 \) while keeping \( \Re s_2 = \ldots = \Re s_{n-1} = K \) fixed shows
\[ y_j^j \partial_{y_j} W_\mu(y) = \mu_1 y_1^{\mu_1} W_\mu^{**}(y_2, \ldots, y_{n-1}) + O \left( (1 + \|\mu\|)^K y_1^{\Re \mu_1 + 1/2} (y_2 \cdots y_{n-1})^{-K} \right) \]
for \( j \in \{0, 1\} \) where
\[ W_\mu^{**}(y_2, \ldots, y_{n-1}) = W_{\mu^{(1)}}^{*}(y_2, \ldots, y_{n-1}) \prod_{j=2}^{n-1} y_j^{\mu_1} \left( \frac{G(\mu^{(j)})}{G(\mu)} \prod_{k=2}^n \Gamma(1 + \mu_k - \mu_1) \right). \]
Note that the last big parenthesis is \( \gg (1 + \|\mu\|)^{-K} \). Let us assume that \( (y_2, \ldots, y_{n-1}) \in I_{n-1}(\mu^{(1)}) \) as defined in (3.5) and that \( y_1 \asymp 1/Z^2 \) for some
\( Z > (2 + \|\mu\|)^{K^2} \) where \( K \) is so large that (3.10) is an asymptotic formula for \( j = 0 \). To transform this asymptotic formula of \( \mathcal{W}_\mu \) into a statement for \( W_\mu^* \) by means of (3.9), we quote [Bl, Lemma 5].

**Lemma 3.1.** Let \( \alpha \geq 0 \), \( c_0, c_1, c_2 > 0 \), \( \beta \in \mathbb{C} \). Let \( I = [a, b] \subseteq (0, 1) \) be an interval with \((1 + c_0)a \leq b \leq 2a\) and \( w : I \to \mathbb{C} \) a smooth function satisfying

\[
|D_\beta w(y)| \geq c_1 y^{-\alpha}, \quad |\partial_y D_\beta w(y)| \leq c_2 \|D_\beta w\| y^{-1}
\]

for \( y \in I \). Then there exist constants \( c'_0, c'_1, c'_2 > 0 \) depending only on \( c_0, c_1, c_2, \beta \) (but not on \( a, b \)) and an interval \( I' = [a', b'] \subseteq I \) with \((b' - a') \geq c'_0(b - a)\) such that

\[
|w(y)| \geq c'_1 y^{-\alpha}, \quad |w'(y)| \leq c'_2 \|w\| y^{-1}
\]

for \( y \in I' \).

An inspection of the proof shows immediately that

\[
c'_0 \geq \frac{c_0}{K(c_2 + 1 + |\beta|)}, \quad c'_1 \geq \frac{c_1}{K(c_2 + 1 + |\beta|)}, \quad c'_2 \geq K(c_2 + 1 + |\beta|).
\]

Applying this repeatedly with \( \beta = \mu_2, \ldots, \mu_n \) and recalling (3.6), we see that

\[
|W_\mu^*(y)| \gg y_{1-\sigma_\pi(\infty)}(2 + \|\mu\|)^{-K}
\]

for \( y_1 \in [\gamma_1/Z^2, \gamma_2/Z^2] \) with \( 1/2 \leq \gamma_1 < \gamma_2 \leq 1 \), \( \gamma_2 - \gamma_1 \geq (2 + \|\mu\|)^{-K} \) and \((y_2, \ldots, y_{n-1}) \in \mathcal{I}_{n-1}(\mu^{(1)}) \). Note that by definition of \( \mathcal{I}_{n-1}(\mu^{(1)}) \), the bound in the previous display continues to hold for \( \tilde{\mu} \) in a small ball of radius \((2 + \|\mu\|)^{-K}\) about \( \mu \) (possibly after changing the constants in (3.5)). Thus we can cover the set of \( \|\mu\| \leq M \) with \( M^K \) balls and for each one pick a (measurable) function \( E_j^{**} : [M^{-K}, M^K]^{n-2} \to [0, 1] \) such that \( \sum_j |\langle E_j^{**}, W_\mu^* \rangle|^2 \gg M^{-K} \) for \( \|\mu\| \leq K \), the inner product being restricted to the last \( n - 2 \) coordinates. Next define

\[
E_j^*(y_1, \ldots, y_{n}) = \delta_{\gamma_1 \leq y_1 \leq \gamma_2} E_j^{**}(y_2, \ldots, y_{n-1}),
\]

so that

\[
\sum_j \left( \int_{\mathbb{R}_{\geq 0}^{n-1}} E_j^*(Z^2 y_1, y_2, \ldots, y_{n-1}) W_\mu^*(y) \frac{dy_1}{y_1} \cdots \frac{dy_{n-1}}{y_{n-1}} \right)^2 \gg Z^{4\sigma_\pi(\infty)} M^{-K}.
\]

Finally changing variables \( y_j \leftarrow y_j^{1/2} / \pi \) as in (3.1), the left hand side equals up to numerical constants

\[
\sum_j \left( \int_{\mathbb{R}_{>0}} y^{-\eta} E_j^*(Z^2, \pi^2 y_1^2, \pi^2 y_2^2, \ldots, \pi^2 y_{n-1}^2) W_{2\mu}(y) \frac{dy_1}{y_1} \cdots \frac{dy_{n-1}}{y_{n-1}} \right)^2
\]

\[
= Z^{-2\eta} \sum_j |\langle E_j^{(Z, 1, \ldots, 1)}, W_{2\mu} \rangle|^2
\]

upon defining \( E_j(y_1, \ldots, y_{n-1}) = y^n E_j^{**}(\pi^2 y_1^2, \pi^2 y_2^2, \ldots, \pi^2 y_{n-1}^2) \). Re-normalizing \( \mu \) and \( \sigma_\pi(\infty) \) by division by 2, we obtain the following:
Lemma 3.2. Let $M \geq 2$ and $Z > M^{K^2}$. There exists $r \in \mathbb{N}$, $r \leq M^K$ and a collection of (measurable) functions $E_1, \ldots, E_r : [M^{-K}, M^K]^{n-1} \to [0,1]$ such that
\[
\sum_{j=1}^{r} |(E_j^{[Z,1,...,1]}, W_\mu)|^2 \gg Z^{2\eta} M^{-K}
\]
for $\mu$ satisfying $\|\mu\| \leq M$, $\|\Re \mu\| < 1/2$ and $\eta$ as in (2.6).

4. Kloosterman sums

We start by compiling the definition and basic properties of Kloosterman sums, following [Fr] in the case of the group $\text{SL}_n(\mathbb{Z})$ that generalizes readily to the congruence subgroups $\Gamma(q)^2$ defined in (2.12); see also [Bl, Section 4] and recall (2.13). The Bruhat decomposition gives $\text{GL}_n(\mathbb{Q}) = \bigcup_{w \in \mathcal{W}} G_w(\mathbb{Q})$ with $G_w := U \wedge U_w$ as a disjoint union. Let $N, M \in \mathbb{Z}^{n-1}$, $c \in \mathbb{N}^{n-1}$, $w \in W$, $v \in V$. Then provided that
\[
\theta_M(c^* wxw^{-1}(c^*)^{-1}) = \theta_N^v(x)
\]
for all $x \in w^{-1}U(\mathbb{Q})w \cap U(\mathbb{Q})$, the Kloosterman sum
\[
S_{q,w}(M, N, c) = \sum_{xc^* wy \in U(\mathbb{Z}) \cap G_w(\mathbb{Q}) \cap \Gamma(q)^2 / U_w(\mathbb{Z})} \theta_M(x) \theta_N^v(y)
\]
with $U_w$ as in (2.1) is well-defined; see [Fr, Proposition 1.3]. If (4.1) is not met, we define $S_{q,w}(M, N, c) = 0$. If $v = \text{id}$, we drop it from the notation. By [Fr, p. 175], the Kloosterman sum is non-zero only if $w$ is of the form (possibly up to sign in the first row)
\[
w = \begin{pmatrix}
I_{d_1} & & \\
& \ddots & \\
& & I_{d_r}
\end{pmatrix}
\]
where $d_1 + \ldots + d_r = n$. The condition (4.1) is equivalent to
\[
M_{n-i}^{c_n+1, c_n-1} = \frac{v_{w(i)+1}}{v_{w(i)}} N_{n-w(i)}
\]
for all $1 \leq i \leq n-1$ satisfying
\[
w(i) + 1 = w(i + 1)
\]
with the convention $c_0 = c_n = 1$ and $v = \text{diag}(v_1, \ldots, v_n)$. If $w$ is of the form (4.2), then the $i$'s satisfying (4.4) are precisely the $i \notin \{d_1, d_1 + d_2, \ldots, d_1 + d_2 + \ldots + d_{r-1}\}$.

The components $c_j$ of $c$ can be recovered from $xc^* wy$ by taking suitable $j$-by-$j$ minors containing the last $j$ rows; see [Fr, Proposition 3.1]. Since each of them is integral for $xc^* wy \in \Gamma(q)^2$ we see that $c \in (\mathbb{Z} \setminus \{0\})^{n-1}$, and modulo $V$ we may assume $c \in \mathbb{N}^{n-1}$.

The sum in the definition of the Kloosterman sum $S_{q,w}(N, M, c)$ runs over the so called Kloosterman set
\[
X_{q,w}(c) = \{ xc^* wy \in U(\mathbb{Z}) \cap G_w(\mathbb{Q}) \cap \Gamma(q)^2 / U_w(\mathbb{Z}) \}$
of modulus \(c\) (for the Weyl element \(w\)). It will be convenient to also define the corresponding \(p\)-adic versions

\[
X_{q,w}^{(p)}(c) = \{ x^*wy \in U(\mathbb{Z}_p) \setminus G_w(Q_p) \cap K_p(q)^s / U_w(\mathbb{Z}_p) \}.
\]

By the Chinese remainder theorem, Kloosterman sums enjoy certain multiplicativity properties in the moduli, cf. \([Fr\), Proposition 2.4\). However, in \([Fr\) only Kloosterman sums for \(\text{SL}_n(\mathbb{Z})\) are considered. While the same factorization properties remain true for a large class of lattices, the situation for the principal congruence subgroup (and its conjugate \(\Gamma(q)^s\)) is slightly more complicated. Therefore we will only give a weak statement in one particular case of interest:

**Lemma 4.1.** Suppose that \((rs, q) = 1\) and \(u, v \mid q^\infty\). Write \(d = (u, uv, \ldots, uv^{n-2}),\)
\(c = (r, rs, \ldots, rs^{n-2})\) and denote the coordinate wise product by \(dc\). Then

\[
|S_{q,w}^v(M, N, dc)| \leq |S_{1,w}^v(M, N', c)| \cdot \prod_{p \mid q} \sharp X_{p,w}^{(p)}((u_p, u_p\cdot v_p S_p), \ldots, u_p\cdot v_p S_p)^n - 2)
\]

with \(u_p = (u, p^\infty),\) \(v_p = (v, p^\infty)\) and a representative \(S_p\) of \(s \cdot \frac{w}{v_p}\) modulo \(p\.\)
The new parameter \(N'\) can be described explicitly in terms of \(u, v\) and \(N\).

**Proof.** The proof relies essentially on finding an explicit bijection between the involved Kloosterman sets. By strong approximation for \(U\) and \(U_{w}\) the diagonal embedding

\[
\psi: X_{q,w}^v(c') \to \prod_{p\mid q} X_{q,w}^{(p)}(c')
\]

is a bijection. Recall that the adelic character \(\psi_{\mathbb{A}}\) (resp. its restriction to \(U_{w}\)) is invariant under \(U(\mathbb{Q})\) (resp. \(U_{w}(\mathbb{Q})\)) and factors into local characters

\[
\psi_{\mathbb{A}} = \theta_{1,\ldots,1} \cdot \prod_{p} \psi_{\mathbb{Q}_p}.
\]

Since \(X_{q,w}^{(p)}(c') = U(\mathbb{Z}_p)(c')^s w_u U_w(\mathbb{Z}_p),\) we see that

\[
S_{q,w}^v(M, N, c') = \prod_{p \mid q, c_0 \ldots c_{n-1}} S_{q,w}^{(p)}(\psi_{\mathbb{Q}_p}^{N,v}, \psi_{\mathbb{Q}_p}^{M,u}, c')
\]

into local Kloosterman sums. These are defined by

\[
S_{q,w}^{(p)}(\psi_{\mathbb{Q}_p}^{N,v}, \psi_{\mathbb{Q}_p}^{M,u}, c') = \sum_{xc^*w,y \in X_{q,w}^{(p)}(c)} \psi_{\mathbb{Q}_p}^{(M,1)}(x) \cdot \psi_{\mathbb{Q}_p}^{(N,v)}(y),
\]

with \(\psi_{\mathbb{Q}_p}^{(N,v)}(y) = \psi_{\mathbb{Q}_p}^{(N,1)}(v^{-1}yv)\) and \(\psi_{\mathbb{Q}_p}^{(N,1)}(y) = \psi_p(-N_{n-1}y_{1,2} - \ldots - N_1y_{n-1,n})\). We estimate

\[
|S_{q,w}^v(M, N, cd)| \leq \prod_{p \mid rs} S_{q,w}^{(p)}(\psi_{\mathbb{Q}_p}^{(M,1)}, \psi_{\mathbb{Q}_p}^{(N,v)}, cd) \cdot \prod_{p \mid q} \sharp X_{q,w}^{(p)}(cd).
\]

For \(p \nmid q\) we have \(K_p(q)^s = \text{GL}_n(\mathbb{Z}_p),\) so that in particular

\[
X_{q,w}^{(p)}(dc) \cong X_{1,w}^{(p)}(c)
\]
via the usual bijection. This allows us to rewrite
\[ \prod_{p \mid rs} S_{q,u}^{(p)}(\psi_{\mathcal{Q}_p}^{(M,1)}, \psi_{\mathcal{Q}_p}^{(N,v)}, cd) = \prod_{p \mid rs} S_{1,u}^{(p)}(\psi_{\mathcal{Q}_p}^{(M,1)}, \psi_{\mathcal{Q}_p}^{(N',v)}, c) = S_{1,u}^{w}(M, N', c), \]
for a suitable \( N' \in \mathbb{N} \) depending on \( N \) and \( d \).

Suppose now that \( p \mid q \). We claim that
\[ X_{q,w}^{(p)}(dc) \equiv X_{q,w}^{(p)}((u_p, u_p \cdot (v_p \overline{s}_p), \ldots, u_p \cdot (v_p \overline{s}_p)^{n-2})). \]
This can be seen by defining the two torus elements
\[ t_1 = \text{diag}(1, \bar{s}, \ldots, \bar{s}, 1) \in T(1 + q\mathbb{Z}_p) \quad \text{and} \quad t_2 = \text{diag}(\frac{u_p}{ut}, 1, \ldots, 1) \in T(\mathbb{Z}_p) \]
where \( \bar{s}s_{v_p} = \pi \) and using the bijection
\[ K_{p}(q)^{2} \ni k \mapsto t_{1}t_{2}k^{-1} \in K_{p}(q)^{2}. \]
This establishes the claim and with it the desired estimate. \( \square \)

Remark 4.1. The proof shows explicitly that \( p \)-adically the main obstruction to the usual factorization properties is that \( T(\mathbb{Z}_p) \cdot K_{p}(q)^{2} \neq K_{p}(q)^{2} \) for \( p \mid q \).

By [DR, Theorem 0.3(i)] we have the trivial bound
\[ |S_{1,w}^{w}(M, N, c)| \leq \#X_{1,w}(c) \ll (c_1 \cdot \ldots \cdot c_{n-1})^{1+\varepsilon}. \]
If \( w = \text{id} \), then \( U_{w}(\mathbb{Q})/U(\mathbb{Z}) = \{ I_{n} \} \) and \( xc^{*} \in \Gamma(q)^{i} \) with \( x \in U(\mathbb{Z}) \setminus U(\mathbb{Q}) \) implies \( c_1 = \ldots = c_{n-1} = 1 \), and so
\[ S_{q,\text{id}}^{w}(M, N, c) = \delta_{c=(1,\ldots,1)}\delta_{N=M}N_{q} \]
using the notation (2.14).

For the following lemma let \( w_{1} = \left( \begin{array}{c} I_{n-1} \\ 1 \end{array} \right) \) (possibly up to sign in the first row) be the “Voronoi element” and \( w_{*} \) as in (2.2).

Lemma 4.2. Let \( x \in U(\mathbb{Q}), \ y \in U_{w}(\mathbb{Q}), \ c = (c_1, \ldots, c_{n-1}) \in \mathbb{N}^{n-1}. \)
(a) Let \( w \in W \setminus \{ \text{id}, w_{*}, w_{1} \} \) be a Weyl element of the form (4.2). If \( xc^{*}wy \in \Gamma(q)^{i} \), then \( q^{n+1} \mid c_{j} \) for some \( j \).
(b) If \( xc^{*}w_{1}y \in \Gamma(q)^{5} \) and \( c = (m\gamma^{n-1}, \gamma^{n-2}, \ldots, \gamma^{2}, \gamma) \) for some \( \gamma \in \mathbb{N} \) and some \( m \in \mathbb{N} \) with \( (m, q) = 1 \), then \( q^{n+1} \mid c_{n} \).
(c) If \( xc^{*}w_{*}y \in \Gamma(q)^{2} \), then \( q^{n} \mid c_{j} \) for all \( j \).

Remarks: 1) For the congruence subgroup \( \Gamma_{0}(q) \subseteq \text{SL}_{2}(\mathbb{Z}) \), the moduli \( c \) in the Kuznetsov formula need to be divisible by \( q \). More generally, for \( \Gamma_{0}(q) \subseteq \text{SL}_{n}(\mathbb{Z}) \), the moduli \( c_1, \ldots, c_{n-d_{1}} \) in the notation of (4.2) need to be divisible by \( q \), cf. [Bl, (4.3)]. This lemma generalizes such statements for the group \( \Gamma(q)^{2} \) and shows strong divisibility statements for the moduli.
2) The proofs show the slightly stronger statements \( q^{\min(n+2,2n-2)} \mid c_{j} \) in part a) and
\[ (n-1)v_{p}(\gamma) \geq \left( n+1 + \frac{2}{n-2} \right)v_{p}(q) \]
in part b).
The following technical lemma bounds the Kloosterman set for $w = w_*$ and certain moduli.

**Lemma 4.3.** Let $p$ be a prime, $\alpha, \beta \in \mathbb{N}_0$, $s_p \in \mathbb{Z}_p^\times$ and
\[ c = (p^{n+\alpha}, s_p p^{n+\alpha+\beta}, \ldots, s_p^{n-2} p^{n+\alpha+\beta(n-2)}) \in \mathbb{Z}_p^{n-1}. \]
Then the number $C_{\alpha, \beta}$ of $(x, y) \in U(\mathbb{Z}_p) \setminus U(Q_p) \times U_{w_*}(Q_p) / U_{w_*}(\mathbb{Z}_p)$ such that $xc^s w_*$ is at most
\[ 2p^{\frac{1}{2}(n^3 + 3n^2 - 10n + 6) + 2\alpha(n-1) + (n-1)(n-2)\beta}. \]
Remarks: We expect that the true order of magnitude of the previous count is \( p^{\frac{1}{3}(n^3 + 3n^2 - 10n + 6) + \alpha(n-1) + \frac{1}{2}(n-1)(n-2)\beta} \), so the constant part of the exponent is sharp, but the \( \alpha \) and \( \beta \)-dependence loses a factor of 2. This is just as much as we can afford without penalty in the proof of Proposition 5.2 below. The significance of the cubic polynomial \( p^{\frac{1}{6}(n^3 + 3n^2 - 10n + 6)} \) is that

\[ q_p^\frac{1}{6}(n^3 + 3n^2 - 10n + 6) = N_q q^{(n-1)^2}. \]

A generalization to not necessarily squarefree \( q \) in Proposition 5.2 would require more general exponents of \( p \) in the entries of \( c \).

Proof. Recall that \( U(\mathbb{Z}_p) \) acts on \( K_p(p)^2 \) both from the left and the right. As a system of representatives for \( U(\mathbb{Z}_p) \setminus U(\mathbb{Q}_p) \) we choose the set of matrices whose entries above the diagonal are in \( \mathbb{Q}_p \setminus \mathbb{Z}_p \). Similarly we restrict the relevant entries in \( U_{w_+}(\mathbb{Q}_p) \setminus U_{w_+}(\mathbb{Z}_p) \). Note that

\[ c^* = \text{diag}(s_p^{n+2} p^{-n-\beta(n-2)}, s_p p^\beta, \ldots, s_p p^\beta, p^{n+\alpha}). \]

We claim that \( xc^*w_+y \in K_p(p)^2 \) implies the following:

\( (4.10) \quad x_{ij} \in \frac{1}{p^{j-1+\alpha+\beta}} \mathbb{Z}_p, \quad j = 2, \ldots, n-1, \)

\( (4.11) \quad y_{ij} \in \frac{1}{p^{j-1+\alpha}} \mathbb{Z}_p, \quad j = 2, \ldots, n-1, \)

\( (4.12) \quad x_{in} \in \frac{1}{p^{n+\alpha}} \mathbb{Z}_p, \quad i = 2, \ldots, n-1, \)

\( (4.13) \quad y_{in} \in \frac{1}{p^{1+\alpha+\beta(n-i)}} \mathbb{Z}_p, \quad i = 2, \ldots, n-1, \)

\( (4.14) \quad x_{1n}, y_{1n} \in \frac{1}{p^{n+\alpha}} + \frac{1}{p^{n+1+\alpha}} \mathbb{Z}_p, \)

\( (4.15) \quad x_{ij} \in \frac{1}{p^{j-i+\alpha+\beta}} \mathbb{Z}_p, \quad 2 \leq i < j \leq n-1. \)

To show \((4.10) - (4.15)\), we compute suitable entries of \( xc^*w_+y = (A_{ij}) \), say. First we see that

\[ A_{i1} = p^{n+\alpha} x_{in} \quad (1 \leq i \leq n-1), \quad A_{nj} = p^{n+\alpha} y_{1j} \quad (2 \leq j \leq n) \]

in the first column and last row. Since \( A_{i1} \in p^i \mathbb{Z}_p \) for \( i \geq 2 \) and \( A_{nj} \in p^{n-j+1} \mathbb{Z}_p \) for \( j \leq n-1 \) and \( A_{11}, A_{nn} \in 1 + p \mathbb{Z}_p \) by \((2.12)\), this implies \((4.11)\), \((4.12)\), \((4.14)\). Next we have

\[ \frac{1}{p^{j-i+1}} \mathbb{Z}_p \ni A_{ij} = s_p p^\beta x_{ij} + p^{n+\alpha} x_{in} y_{1j}, \quad 1 \leq i < j \leq n-1. \]
By induction on $x$ and $(\ref{condition})$, we conclude $(\ref{established})$, and for $i = 1$ a similar argument using $(\ref{last_column})$ shows $(\ref{we_have})$. To establish $(\ref{follows})$, we use the last column:

$$
\frac{1}{p^{n-1}} \mathbb{Z}_p \ni A_{in} = p^{n+\alpha} x_{in} y_{1n} + \overline{x}_p p^\beta y_{in} + \overline{y}_p p^\beta \sum_{j=i+1}^{n-1} x_{ij} y_{jn}, \quad 2 \leq i \leq n-1.
$$

By induction on $i = n-1, n-2, \ldots, 2$ we have

$$
p^{n+\alpha} x_{in} y_{1n} + \overline{x}_p p^\beta \sum_{j=i+1}^{n-1} x_{ij} y_{jn} \in \frac{1}{p^{n-i+\alpha}} \mathbb{Z}_p + \frac{1}{p(1+\alpha+\beta)(n-i-1)+\alpha} \mathbb{Z}_p \subseteq \frac{p^\beta}{p(1+\alpha+\beta)(n-i)} \mathbb{Z}_p
$$

and $(\ref{follows})$ follows for $i-1$. We have now established $(\ref{we_have}) - (\ref{established})$.

By choosing a suitable fundamental domain for $\mathbb{Q}_p/\mathbb{Z}_p$ in which the entries $x_{ij}, y_{ij}$ must lay, we re-write these conditions as

(\ref{re-written})

$$
x_{1j} = \frac{x'_{1j}}{p^{j-1+\alpha+\beta}}, \quad j = 2, \ldots, n-1, \quad x'_{1j} \in \{0, 1, \ldots, p^{j-1+\alpha+\beta} - 1\},
$$

$$
y_{1j} = \frac{y'_{1j}}{p^{j-1+\alpha}}, \quad j = 2, \ldots, n-1, \quad y'_{1j} \in \{0, 1, \ldots, p^{j-1+\alpha} - 1\},
$$

$$
x_{in} = \frac{x'_{in}}{p^{n-i+\alpha}}, \quad i = 2, \ldots, n-1, \quad x'_{in} \in \{0, 1, \ldots, p^{n-i+\alpha} - 1\},
$$

$$
y_{in} = \frac{y'_{in}}{p(1+\alpha+\beta)(n-i)}, \quad i = 2, \ldots, n-1, \quad y'_{in} \in \{0, 1, \ldots, p(1+\alpha+\beta)(n-i) - 1\},
$$

$$
x_{1n} = \frac{1 + px'_{1n}}{p^{n+\alpha}}, \quad y_{1n} = \frac{1 + py'_{1n}}{p^{n+\alpha}}, \quad x'_{1n}, y'_{1n} \in \{0, 1, \ldots, p^{n+\alpha} - 1\},
$$

$$
x_{ij} = \frac{x'_{ij}}{p^{j-1+\alpha+\beta}}, \quad 2 \leq i < j \leq n-1, \quad x'_{ij} \in \{0, 1, \ldots, p^{j-i+\alpha+\beta} - 1\}.
$$

The number of choices of such $x, y$ is

(\ref{number_of_choices})

$$
p^{\frac{1}{2}(n-1)(n-2)+(\alpha+\beta)(n-2)} \cdot p^{\frac{1}{2}(n-1)(n-2)+\alpha(n-2)} \cdot p^{\frac{1}{2}(n-1)(n-2)+\alpha(n-2)}
$$

$$
\cdot p^{\frac{1}{2}(n-1)(n-2)(1+\alpha+\beta)} \cdot p^{2(n-1+\alpha)} \cdot p^{\frac{1}{2}(n-2)(n-3)(n-4)+\frac{1}{2}(n-2)(n-3)(\alpha+\beta)}
$$

$$
= p^{\frac{1}{2}(n^3+3n^2+2n-12)+\alpha(n-1)+\beta(n-1)(n-2)}.
$$
In the new coordinates we compute again
\[
\frac{1}{p^{j-2}} z_p \supset A_{ij} = \frac{\bar{\tau}_p x'_{ij} + y'_{ij}(1 + px'_{1n})}{p^{j-1+\alpha}}, \quad 2 \leq j \leq n - 1,
\]
\[
\frac{1}{p^{j-i-1}} z_p \supset A_{ij} = \frac{\bar{\tau}_p x'_{ij} + x'_{in}y'_{1j}}{p^{j-i-1+\alpha}}, \quad 2 \leq i < j \leq n - 1,
\]
\[
\frac{1}{p^{n-i-1}} z_p \supset A_{in} = \frac{x'_{in}(1 + py'_{1n})}{p^{n-i+\alpha}} + \frac{\bar{\sigma}_p y'_{1n}}{p^{(1+\alpha)(n-i)+\beta(n-i-1)}} + \sum_{j=i+1}^{n-1} \frac{\bar{\sigma}_p x'_{ij} y'_{jn}}{p^{n-i-1+\alpha(n-j)+\beta(n-j)}},
\]
\[
2 \leq i \leq n - 1,
\]
\[
\frac{1}{p^{n-2}} z_p \supset A_{1n} = -\frac{\bar{\sigma}_p^{-n+2}}{p^{2+\alpha+\beta(n-2)}} + \frac{(1 + px'_{in})(1 + py'_{1n})}{p^{2+\alpha}} + \frac{1}{p} \sum_{j=2}^{n-1} \frac{\bar{\sigma}_p x'_{ij} y'_{jn}}{p^{n-j+1+\alpha+\beta(n-j)}} \in \mathbb{Z}_p.
\]

which implies the congruences
\[
\bar{\tau}_p x'_{ij} + y'_{ij}(1 + px'_{1n}) \equiv 0 \pmod{p^{1+\alpha}}, \quad 2 \leq j \leq n - 1,
\]
\[
\bar{\tau}_p x'_{ij} + x'_{in}y'_{1j} \equiv 0 \pmod{p^{\alpha}}, \quad 2 \leq i < j \leq n - 1,
\]
\[
\bar{\sigma}_p y'_{in} + x'_{in}(1 + py'_{1n})p^{(\alpha+\beta)(n-i-1)} + p \sum_{j=i+1}^{n-1} p^{(\alpha+\beta)(j-i-1)} \bar{\sigma}_p x'_{ij} y'_{jn}
\equiv 0 \pmod{p^{1+\alpha(n-i)+\beta(n-i-1)}}, \quad 2 \leq i \leq n - 1,
\]
as well as
\[
\bar{\tau}_p^{-n+2} \frac{(1 + px'_{in})(1 + py'_{1n})}{p^{2+\alpha}} + \frac{1}{p} \sum_{j=2}^{n-1} \frac{\bar{\sigma}_p x'_{ij} y'_{jn}}{p^{n-j+1+\alpha+\beta(n-j)}} \in \mathbb{Z}_p.
\]

We are now ready to count the number of \((x, y)\) satisfying (4.16), (4.18) and (4.19).

Our first count drops the condition (4.19). Fixing \(y'_{1j}, \quad 2 \leq j \leq n, \) and \(x'_{in}, \quad 1 \leq i \leq n - 1, \) the first set of congruences in (4.18) fixes \(x'_{ij} \) modulo \(p^{1+\alpha}\) for \(2 \leq j \leq n - 1, \) the second set fixes \(x'_{ij} \) modulo \(p^{\alpha} \) for \(2 \leq i < j \leq n - 1, \) and once these choices are made the last set of congruences fixes inductively \(y'_{jn} \) modulo \(p^{1+\alpha(n-i)+\beta(n-i-1)} \) for \(i = n - 1, n - 2, \ldots, 2. \) Recalling (4.17), we obtain
\[
C_{\alpha, \beta} \leq \frac{p^{\frac{1}{2}(n^3 + 3n^2 + 2n - 12) + \alpha(n-1) + \beta(n-1)(n-2)}}{p^{(1 + \alpha)(n-2)} \cdot p^{\frac{3}{2}(n-2)(n-3)} \cdot p^{n-2 + \frac{1}{2}(n-1)(n-2)+\frac{1}{2}(n-2)(n-3)\beta}}
\]
\[
= p^{\frac{1}{2}(n^3 + 3n^2 - 10n + 12) + 2(n-1) - (n-1)} \cdot p^{\frac{1}{2}(n-1)(n-2)\beta}.
\]

This is sufficient for the claim of the lemma provided that \(\beta \geq 1 \) and \(n \geq 4. \)

We now consider the case \(\beta = 0, \) in which case (4.19) simplifies to
\[
(4.20) \quad p^{3\alpha} (x'_{1n} + y'_{1n} + px'_{1n}y'_{1n}) + \sum_{j=2}^{n-1} p^{\alpha(j-2)} \bar{\sigma}_p x'_{ij} y'_{jn} \equiv 0 \pmod{p^{1+(n-1)\alpha}},
\]
and our task is to gain an additional factor $p$ from this congruence. We start with the subcase $\alpha = 0$, in which (4.20) becomes
\[ x_1' + y_1' + \sum_{j=2}^{n-1} \pi_p x_{1j}' y_{jn}' \equiv 0 \pmod{p}, \]
while (4.18) becomes
\[ \pi_p x_{1j}' + y_{1j}' \equiv \pi_p y_{jn}' + x_{jn}' \equiv 0 \pmod{p}, \quad 2 \leq i, j \leq n - 1. \]
By the same argument as above we have
\[ C_{0,0} \leq \frac{p^{\frac{1}{2}(n^2+3n^2+2n-12)}}{p^{n-2} \cdot p^{n-2} \cdot p} = \frac{p^{\frac{1}{2}(n^2+3n^2-10n+6)}}{p}, \]
which establishes the bound of the lemma for $\alpha = \beta = 0$.

If $\beta = 0, \alpha \geq 1$, we first count the contribution of those pairs $(x, y)$ with $x_{12} \neq 0 \pmod{p}$. As before, we use (4.18) to fix $x_{1j}'$ (mod $p^{1+\alpha}$) for $2 \leq j \leq n - 1$, then $x_{1j}'$ (mod $p^{\alpha}$) for $2 \leq i < j \leq n - 1$ and then inductively $y_{in}'$ modulo $p^{1+\alpha(n-i)}$ for $i = n-1, n-2, \ldots, 3$. But in contrast to the general case, we now use (4.20) to fix $y_{2n}'$ modulo $p^{1+(n-1)\alpha}$ (which gains a factor $p^\alpha$). On the other hand, if $x_{12}' \equiv 0 \pmod{p}$, then by the first congruence in (4.18) we must have $y_{12}' \equiv 0 \pmod{p}$, so up front we save a factor $p$, since $y_{12}'$ is determined modulo $p$. We conclude that
\[ C_{\alpha,0} \leq \frac{p^{\frac{1}{2}(n^2+3n^2+2n-12)+\alpha(n-1)}}{p^{(1+\alpha)(n-2)} \cdot p^{\frac{1}{2}(n-2)(n-3)\alpha} \cdot p^{n-2+\frac{1}{2}(n-1)(n-2)\alpha} \cdot \left( \frac{1}{p^\alpha} + \frac{1}{p} \right)} \leq 2p^{\frac{1}{2}(n^2+3n^2-10n+6)+2(n-1)\alpha} \]
for $\alpha \geq 1$.

The last remaining case is $n = 3$ and $\beta \geq 1$. In this case (4.19) becomes
\[ p^\alpha (p^\beta - \pi_p) + p^{1+\alpha+\beta}(x_{13}' + y_{13}' + px_{13}' y_{13}') + p\pi_p x_{12}' y_{23}' \equiv 0 \pmod{p^{2+2\alpha+\beta}}. \]
The case $\alpha = 0$ is obviously impossible, and for $\alpha \geq 1$ the previous display implies
\[ x_{12}' y_{23}' \equiv \begin{cases} 0 \pmod{p}, & \alpha \geq 2, \\ 1 \pmod{p}, & \alpha = 1. \end{cases} \]
As before this saves a factor $2/p$ for the number of choices for $(x_{12}', y_{23}')$, and after these choices are made we use (4.18) to determine $y_{12}', x_{23}'$ modulo $p^{1+\alpha}$. This suffices to establish the lemma also for $n = 3$ and completes the proof.

**Theorem 4.4.** Let $N = (*, 1, \ldots, 1, *)$, $M = (*, 1, \ldots, 1, *) \in \mathbb{N}^{n-1}$, $v \in V$ and $q \in \mathbb{N}$. Then $S_{v, \omega, q}(M, N, c)$ vanishes unless
\[ c = (q^n r, q^n r, q^n r, q^n r, q^n r^2, \ldots, q^n r^{n-2}) \quad \text{or} \quad (q^n r^{n-2}, q^n r^{n-1}, \ldots, q^n r, q^n r) \]
for some \(r, s \in \mathbb{N}\). We write \(r = r_1r_2\) and \(s = s_1s_2\) with \(r_1s_1 \mid q^\infty\), \((r_2s_2, q) = 1\). If \(q\) is squarefree, then with this notation we have
\[
S_{q,w_*}^u(M, N, c) \ll N_q \left( \frac{c_1 \cdots c_{n-1}}{q^{n-1}} \right)^{1+\varepsilon} \left( \frac{c_1 \cdots c_{n-1}}{q^n} \right)^{q^\infty}.
\]

**Proof.** That all entries of \(c\) must be divisible by \(q^n\) follows from Lemma 4.2(c), and the special shape of \(c\) follows from (4.3) since we must have \(c_i^2 = c_{i-1}c_{i+1}\) for \(2 \leq i \leq n - 2\). The standard involution \(g \mapsto w_1g^{-\top}w_1\) leaves \(w_*\), \(\Gamma(q)^2\), the shape of \(N\) and \(M\), \(U(\mathbb{Z})\) and \(U_{w_*}(\mathbb{Z})\) invariant and switches the diagonal matrices \((q^nr, q^ns, q^nrs, \ldots, q^nrs^{n-2})^*\) and \((q^nr, q^nrs, q^nrs^{n-1}, \ldots, q^nrs, q^n)^*\). Hence it suffices to establish the bound for one of the two forms of \(c\), say the first one (cf. [Ste, Theorem 3.2]). By (4.5), we first split off the portion containing \(r_2\) and \(s_2\) for which we use the trivial bound (4.7). For the remaining portions we may apply Lemma 4.3 with \(q = p\) (since \(q\) is squarefree), \(r = p^\alpha\), \(s = p^\beta\) to obtain
\[
S_{q,w_*}^u(M, N, c) \ll N_q q^{(n-1)^2+\varepsilon}(r_1^2r_2)^{n-1}(s_1^2s_2)^{1(n-1)(n-2)},
\]
and the claim follows.

5. The Kuznetsov formula

Let \(E\) be a fixed compactly supported function on \(\mathbb{R}^{n-1}_>\), \(X \in \mathbb{R}^{n-1}_>\) a “parameter” and define
\[
E^{(X)}(y_1, \ldots, y_{n-1}) = E(X_1y_1, \ldots, X_{n-1}y_{n-1}).
\]
Next we define the right \(O_n(\mathbb{R})Z^+\) invariant function \(F^{(X)} : \text{GL}_n(\mathbb{R}) \to \mathbb{C}\) by
\[
F^{(X)}(xy\kappa) = \theta(x)E^{(X)}(y(y))
\]
for \(x \in U(\mathbb{R}), y \in \tilde{T}(\mathbb{R}), \kappa \in O_n(\mathbb{R}), \alpha \in Z^+, \theta = \theta_{(1, \ldots, 1)}\) as in (2.10) and \(y(y)\) as in (2.4). Recalling (2.13), from the Fourier expansion of Poincaré series [Fr, Theorem A] one concludes as in [Bl, Lemma 7] the following Kuznetsov-type formula, using the notation (2.3), (2.4), (2.5), (2.6), (2.8), (2.9).

**Lemma 5.1.** Let \(M, N \in \mathbb{N}^{n-1}, X \in \mathbb{R}^{n-1}_>\), \(E\) a compactly supported function on \(\mathbb{R}^{n-1}_>\) and define \(F^{(X)}\) as in (5.1). Let
\[
A = \iota(X \cdot M)c^w wv(X \cdot N)^{-1}w^{-1} = \iota(X \cdot M \cdot w(X \cdot N))c^* \in T(\mathbb{R}).
\]
Then
\[
\int_{\Gamma(q)^2} A_\varpi(M)A_\varpi(N)|(W_\mu_\varpi, E^{(X)})|^2d\varpi = \sum_{w \in W} \sum_{u \in V} \sum_{c \in \mathbb{N}^{n-1}} \sum_{c_1 \cdots c_{n-1}} S_{q,w}^u(M, N, c)
\]
\[
X^{2\eta} \int_{\tilde{T}(\mathbb{R})} \int_{U_\varpi(\mathbb{R})} F^{(X)}(\iota(X)^{-1}Awxy)\theta^w(-x)E(y(y))dx d^*y.
\]
Note that $A_{\infty}(M) = 0$ if $\infty$ belongs to the residual spectrum.

We analyze the right hand side of (5.2) for a test function $E$ with support in some rectangle $[r, R]^{n-1} \subseteq \mathbb{R}_{>0}^{n-1}$, $N = M = (m, 1, \ldots, 1)$ for some $(m, q) = 1$ and

$$(5.3) \quad X = (Z, 1, \ldots, 1)$$

for some $Z \geq 1$ where

$$(5.4) \quad mZ \leq K^{-1}(1 + 1/r + R)^{-K} q^{n+1}$$

with the convention from Section 2.1. We introduce the notation

$$A \preccurlyeq B \quad \iff \quad A \ll (1 + 1/r + R)^K B.$$  

For $w = id$ we have $A = I_n$, $U_w(\mathbb{R}) = \{I_n\}$ in Lemma 5.1, and by (4.8) we conclude that the contribution of $w = id$ is

$$(5.5) \quad \ll N_q Z^{2n}.$$  

Assume from now on $w \neq id$ and $w$ as in (4.2).

We quote [Bl, Lemma 1]: for $w \in W$, $x \in U_w(\mathbb{R})$, $y, c, B \in \mathbb{R}_{>0}^{n-1}$ write $y(\iota(B)c^x w x \iota(y)) = Y \in \mathbb{R}_{>0}^{n-1}$ and $A = \iota(B)c^x$. Then

$$c_j \ll_{y,Y} \prod_{i=1}^{n-1} B_i^{s(i,j)}$$

for $1 \leq j \leq n - 1$ if the right hand side of (5.2) is nonzero. From [Bl, (3.3), (3.5), (3.6)] it is obvious that the implied constant is of the form $(1/y + 1/Y + y + Y)^K$.

As in [Bl, (7.1)] we apply this with $B = X \cdot M \cdot w(X \cdot N)$ to see that the right hand side of (5.2) vanishes unless $c_j \ll mZ$ for $1 \leq j \leq n - 1$. We now apply Lemma 4.2 and note that by (4.3) the moduli for $w_1$ must be of the form as in part (b) of that lemma. By (5.4) (with $K$ sufficiently large) we conclude that the contribution of all $w$ except $w_1$ vanishes. Our next aim is to estimate the $x, y$-integral in (5.2) for $w = w_1$ which we estimate trivially by

$$\int_{\hat{T}(\mathbb{R})} \int_{U_{w_1}(\mathbb{R})} |E(y(Aw_1 x y))E(y(y))| dx \, dx \, y.$$  

Here we argue verbatim as in [Bl, Section 7] and invoke [Bl, Lemma 1] and [Bl, Lemma 3] (with implied constants that are polynomial in the support of $E$) to conclude

$$\int_{\hat{T}(\mathbb{R})} \int_{U_{w_1}(\mathbb{R})} F(X) \langle \iota(X)^{-1} Aw_1 x y \rangle \theta_\nu(-x) \overline{E(y(y))} dx \, dx \, y \ll y(A)^{n(1+\epsilon)}.$$  

Thus by Theorem 4.4 we see that the contribution of $w_1$ in (5.2) is

$$\ll q^s \frac{N_q}{q^{n-1}} Z^{2n} \sum_{q^n | c_1, \ldots, c_{n-1} < q^{n+1}} \left(\frac{c_1 \ldots c_{n-1}}{q^n}, q^{\infty}\right) \leq q^s N_q Z^{2n} \left(\sum_{\gamma < q} 1\right)^{n-1} \ll N_q Z^{2n}.$$
by a standard application of Rankin’s trick in the last step. This bound majorizes (5.5) and bounds the right hand side of (5.2) under the current hypotheses on $N,M,X,E$. From (2.20) we therefore obtain

$$\int_{\Gamma(q)^{\perp}} |A_{\pi}(1)|^2 |\lambda_\pi(m)|^2 |(W_{\mu_{\pi}}, E^{(Z,1,\ldots,1)})|^2 d\varpi \lesssim q^s N_q Z^{mn},$$

provided that $(m,q) = 1$ and (5.4) holds where $r,R$ are determined by the support of $E$. Combining this with (2.21) and Lemma 3.2, we conclude the following.

**Proposition 5.2.** There exists an absolute constant $K > 0$ with the following property. Fix $M,T > 1$, $q$ squarefree, and suppose that $T \leq M^{-K} q^{n+1}$. Fix a place $v$ of $\mathbb{Q}$. If $v = p$ is finite, assume that $p \nmid q$. Then

$$\int_{\|\mu_{\pi}\| \leq M} |A_{\pi}(1)|^2 T^{2\sigma_{\pi.v}} d\varpi \ll_{v, \varepsilon} M^K q^s N_q.$$

6. Fourier coefficients

In this section we prove Theorem 1.3 and Theorem 1.1.

6.1. Towards local computations. Recall that given a cuspidal automorphic representation $\pi \mid L^2(X_q)$ as in (2.25) we defined in Section 2.6 the associated subspace $V_\pi$ in $L^2(X_q)$ by de-adelisation. The first step in proving Theorem 1.3 is to reduce it to certain local computations. We factor $\pi = \otimes_v \pi_v$. For finite places $v = p$ we write $\pi_{p}^{K_p(p^m)}$ for the space of $K_p(p^m)$-invariant vectors in $\pi_p$ with $K_p(q)$ as in (2.22). We also fix an invariant inner product $\langle \cdot, \cdot \rangle_{\pi_p}$ on $\pi_p$. Recall, from [LM, Definition 2.1], that the stable integral $I_{U(\mathbb{Q}_p)}^s$ is defined by requiring that the corresponding integrals over sufficiently large open compact subgroups of $U(\mathbb{Q}_p)$ stabilizes. Define the local quantity

$$S_{\pi,p^m}(y) = \sum_{v \in \text{ONB}(\pi_{p}^{K_p(p^m)})} \int_{U(\mathbb{Q}_p)} \langle \pi_p(u)v, v \rangle \psi_{\mathbb{Q}_p} (yuy^{-1})^{-1} du.$$ 

Note that this quantity is independent of the choice of $\langle \cdot, \cdot \rangle_{\pi_p}$. Let $\text{Wh}(\pi_p)$ be the local Whittaker model for $\pi_p$ and fix an isomorphism $\pi_p \ni F_p \mapsto W_{F_p} \in \text{Wh}(\pi_p)$. This amounts to choosing (the normalization of) the Whittaker functional. Recall the definition (2.11) of $D_q$ and set $\psi_{\mathbb{Q}_p}^5 (u) = \psi_{\mathbb{Q}_p} (D_q u D_q^{-1})$.

We start by inserting (2.28) to obtain

$$\sum_{\varpi \in \text{ONB}(V_\pi)} |A_{\pi}(1)|^2 = \frac{C(\pi)}{V_q} \sum_{F \in \text{ONB}(\pi^{K(q)^{\perp}})} \prod_{p | q} I_p(W_{F_p})$$

$$= \frac{C(\pi)}{V_q} \prod_{p | q} \sum_{F_p \in \text{ONB}(\pi_{p}^{K_p(q)})} I_p(W_{\pi_p(D_q)F_p}),$$
We compute
\[
I_p(W_{\pi_p(D_q)}F_p) = \int_{U(\mathbb{Q}_p)}^{\text{st}} \frac{\langle \pi_p(uD_q)F_p, \pi_p(D_q)F_p \rangle_{\pi_p} \psi_{\mathbb{Q}_p}(u)^{-1}}{\langle F_p, F_p \rangle_{\pi_p}} \, du.
\]
Thus we have
\[
\sum_{\omega \in \text{ONB}(V_\pi)} |A_{\omega}(1)|^2 = C(\pi) \cdot \frac{\mathcal{N}_q q^{n(n-1)/2}}{V_q} \cdot \prod_{p|q} \mathcal{S}_{\pi_p, p^n(q)}(D_q).
\]

Recall the lower bound $C(\pi) \gg (|\mu_\pi|q)^{-\epsilon}$ from (2.29). Since $\dim_{\mathbb{C}}(V_\pi) = \prod_{p|q} \dim_{\mathbb{C}}(\pi_p^{K_p(p^{n(q)})})$, we conclude the following.

**Proposition 6.1.** Let $q \in \mathbb{N}$ be an integer. Assume that
\[
\mathcal{S}_{\pi_p, p^n(D_q)} \gg 1.
\]
and
\[
\dim_{\mathbb{C}}(\pi_p^{K_p(p^{m})}) \ll p^{m(n-1)/2}
\]
for all $p^m|q$. For any cuspidal automorphic representation $\pi \mid L^2(X_q)$ we have
\[
\sum_{\omega \in \text{ONB}(V_\pi)} |A_{\omega}(1)|^2 \gg (|\mu_\pi|q)^{-\epsilon} \dim(V_\pi) q^{n(n-1)(n-2)/6} q^{n^2-1}.
\]

We will now proceed by showing (6.1) and (6.2) for all primes $p$ and $m = 1$. Thus the conclusion of the proposition above holds for all squarefree $q$; which is precisely what is claimed in Theorem 1.3. We expect that (6.1) and (6.2) hold for general $m$ and most likely the bounds are essentially sharp for generic irreducible representations $\pi_p$. In the toy case $n = 2$ computations verifying these assertions can be found in [As] and [MY]. We also refer to [La2] for some results on the support of matrix coefficients.

From now on we work locally at a place $v = p$. Throughout $\pi_p$ will be an irreducible generic smooth admissible (unitary) representation of $\text{GL}_n(\mathbb{Q}_p)$. We write $c(\pi_p)$ for the exponent-conductor. This is the smallest non-negative integer $m$ such that $\pi_p$ has non-trivial $K_{1,p}(m)$-fixed vectors where $K_{1,p}(m)$ is the standard congruence subgroup consisting of matrices in $\text{GL}_n(\mathbb{Z}_p)$ with bottom row congruent to $(0, \ldots, 0, 1)$ modulo $p^m$. We write $\rho(\pi_p)$ for the depth of $\pi_p$. Finally, given any open compact subgroup $C \subset \text{GL}_n(\mathbb{Q}_p)$ we slightly abuse notation and write $\pi_p^C$ for the space of $C$-fixed vectors in $\pi_p$. Let $\omega_{\pi_p}$ be the central character of $\pi_p$. Without loss of generality we can assume that $\omega_{\pi_p}(p) = 1$, as this property can be achieved by twisting $\pi_p$ with an unramified character if necessary.

First we recall that there is a standard parabolic subgroup $P = MN \subset \text{GL}_n$ such that
\[
\pi_p = \text{Ind}^G_P(\sigma),
\]
where $\sigma$ is an irreducible essentially square integrable representation of $M$. By the classification of standard parabolic subgroups we find a partition $(n_1, \ldots, n_k)$ of $n$ such that $M \cong \text{GL}_{n_1}(\mathbb{Q}_p) \times \ldots \times \text{GL}_{n_k}(\mathbb{Q}_p)$. Thus we can write $\sigma = \tau_1 \otimes \ldots \otimes \tau_k$, for irreducible essentially square integrable representations $\tau_i$ of $\text{GL}_{n_i}(\mathbb{Q}_p)$ for $i = 1, \ldots, k$ with the usual adjectives.

The following result classifies representations with $K_p(p^m)$-fixed vectors in terms of conductors. This is certainly well-known to experts, but has been explicitly written down in general only recently in the preprint [MY, Theorem 1.2].

**Lemma 6.2.** Let $\pi$ be as above. Then $\pi^K_{p(p^m)} \neq \{0\}$ if and only if $c(\tau) \leq m \cdot n_i$.

**Proof.** For $\pi$ essentially square integrable (i.e. $k = 1$) this was shown in [BH, Proposition 2]. The general situation ($k > 1$) is easily reduced to this case by using [NV, Lemma 23.2]. \qed

Given an irreducible admissible representation $\tau_i$ of $\text{GL}_{n_i}(\mathbb{Q}_p)$, we can associate an $n_i$-dimensional Weil-Deligne representation $\rho_i = (r_i, N_i)$ via the local Langlands correspondence. Note that $\tau_i$ is essentially square integrable if and only if $\rho_i$ is indecomposable. Furthermore, $\tau_i$ is supercuspidal if and only if $\rho_i$ is irreducible; see [We, Section 1.2.3]. According to [Ro, Section 5], any indecomposable Weil-Deligne representation is of the form $\rho_i \cong (r_i, 0) \otimes \text{sp}(k_i)$, where $k_i \mid n_i$ and $r_i$ is an irreducible $\frac{n_i}{k_i}$-dimensional representation of the Weil-group (of $\mathbb{Q}_p$). We write $c(r_i)$ for the exponent conductor of $r_i$. This conductor can be defined on the Weil-Deligne side, but for our purposes it suffices to define $c(r_i)$ for irreducible $r_i$ via the local Langlands correspondence. The conductor of essentially square integrable representations has been computed in [Ro, Section 10]. We have

$$c(\tau_i) = \begin{cases} k_i \cdot c(r_i) & \text{if } r_i \text{ is ramified,} \\ k_i - 1 & \text{if } r_i \text{ is unramified.} \end{cases}$$

If $k_i = n_i$, then $r_i$ is nothing but a character $r_i = \chi$ and $c(r_i)$ coincides with the usual exponent conductor of $\chi$. In this case $\tau_i = \chi \text{St}_{n_i}$ is a twist of the Steinberg representation. This gives us a good understanding of the representations we need to consider in view of resolving (6.1) and (6.2) case by case. The situations we need to consider are given in the following lemma, which follows directly from our discussion so far.

**Lemma 6.3.** Suppose $\pi_p$ is irreducible, generic and $\pi^K_{p(p^m)}$ is non-trivial. Then we must be in one of the following cases.

- **Case 1:** $\pi_p$ is an (essentially) square integrable representation with $c(\pi_p) \leq n$. This includes the following two sub-cases:

  4Note that the authors in loc. cit. make other assumptions on the additive character $\psi'$. This implies that their conductor $f(\pi, \psi')$, which is defined via the epsilon factor, differs from ours, namely $f(\pi, \psi') = c(\pi) + n$.\footnote{Note that the authors in loc. cit. make other assumptions on the additive character $\psi'$. This implies that their conductor $f(\pi, \psi')$, which is defined via the epsilon factor, differs from ours, namely $f(\pi, \psi') = c(\pi) + n$.}
- **Case 1.a:** $\pi_p$ is supercuspidal with $c(\pi_p) = n$ (i.e. depth-zero).
- **Case 1.b:** $\pi_p$ is a special representation with $c(\pi_p) \leq n$. Thus there is $d \mid n$ and $\tau$ supercuspidal with $c(\tau) \leq \frac{n}{d}$ so that $\pi_p$ is the special representation $St(\tau, d)$. (If $d = n$, then $\tau = \chi$ is a character with $c(\chi) \leq 1$ and we have $\pi_p = \chi \cdot St_n$, where we write $St_n$ for the Steinberg representation.)

- **Case 2:** There is a parabolic subgroup $P = MN$ and an essentially square integrable representation $\sigma = \tau_1 \otimes \ldots \otimes \tau_k$ of $M$ with $c(\tau_i) \leq n_i$ such that $\text{Ind}_{P(Q_p)}^{GL_n(Q_p)}(\sigma)$ is irreducible and $\pi_p = \text{Ind}_{P(Q_p)}^{GL_n(Q_p)}(\sigma)$.

### 6.2. The case of depth-zero supercuspidal representations (Case 1.a)

Depth-zero supercuspidal representations of $GL_n(Q_p)$ can be constructed as follows. We consider a cuspidal representation $(\mu, W)$ of $GL_n(Z/pZ)$, inflate it to a representation of $K_p$ and extend it to $Z(Q_p)K_p$ (taking the central character into account). The so obtained representation will also be denoted by $\mu$. We can now consider the compact induction $\pi_p = c\text{-Ind}_{ZK_p}^{GL_n(Q_p)}(\mu)$. This representation acts by right translation on the space of functions

$$V_c(\mu) = \{ f : GL_n(Q_p) \to W : \text{compactly supported mod } Z, f(hg) = \mu(h)f(g) \text{ for all } g \in GL_n(Q_p) \text{ and } h \in Z(Q_p)K_p \}.$$ 

The inner product is given by

$$\langle f, g \rangle_{\pi_p} = \sum_{x \in ZK_p \backslash GL_n(Q_p)} \langle f(x), g(x) \rangle_W.$$ 

Note that the sum-notation is justified, since $\text{supp}(f) \cup \text{supp}(g)$ is compact so that we are summing a finite number of terms. For $w \in W$ and $y \in GL_n(Q_p)$ we can define

$$f_{y,w}(g) = \begin{cases} 
\mu(h)w & \text{if } g = hy \in Z(Q_p)K_p \cdot y, \\
0 & \text{else}.
\end{cases}$$ 

The family $\{ f_{y,w} : y \in GL_n(Q_p), w \in W \}$ spans $V_c(\mu)$ and we have

$$\langle f_{y,w}, f_{x,v} \rangle_{\pi_p} = \langle w, v \rangle_W \cdot \delta_{x \in Z(Q_p)K_p} y.$$ 

With these preliminaries set up we can construct the following basis.

**Lemma 6.4.** Let $v_1, \ldots, v_l$ with $l = \dim CW$ be an orthonormal basis of $W$. Then $(f_{1,v_i})_{i=1}^l$ is an orthonormal basis of $V_c(\mu)^{K_p(p)}$.

**Proof.** In view of the above inner product relation it is clear that the functions $f_{1,v_i}$ are orthonormal. It is also clear that they are $K_p(p)$-invariant. Thus we only need to show that they exhaust the space $\pi_p^{K_p(p)}$. However, to see this it suffices to show that $f_{y,w}$ is not $K_p(p)$-invariant if $y \notin Z(Q_p)K_p$. To this end write $g = zky \in Z(Q_p)K_p \cdot y$ for $k \in K_p$ and take $\gamma \in K_p(p)$. Then, writing $g\gamma = zk(y\gamma y^{-1})y$ yields the condition

$$y\gamma y^{-1} \in K_p(p),$$
since otherwise we can choose \( k \) so that \( f_{y,v_i}(g) \neq f_{y,v_i}(g\gamma) \). We are done since the normalizer of \( K_p(p) \) in \( \text{GL}_n(\mathbb{Q}_p) \) is \( \mathbb{Z}(\mathbb{Q}_p)K_p \).

\[ \square \]

**Lemma 6.5.** For any \( n \geq 2 \) and \( \pi_p \) depth-zero supercuspidal, we have

\[ S_{\pi_p}(D_q) = 1. \]

Furthermore,

\[
\dim \pi_p^{K_p(p)} = \prod_{i=1}^{n-1} (p^i - 1) = p^{\frac{n(n-1)}{2}}(1 + O(p^{-1})).
\]

In particular, (6.1) and (6.2) hold for depth-zero supercuspidal representations \( \pi_p \) and \( m = 1 \).

**Proof.** First we reduce the situation to a problem in the realm of representations of \( \text{GL}_n \) over finite fields. To do so we note that since \( \pi_p \) is supercuspidal the matrix coefficients have compact support and we do not need to worry about the stabilization of the integral. We simply have

\[
S_{\pi_p,p}(D_q) = \sum_{i=1}^l \int_{U(\mathbb{Q}_p)} \langle \pi_p(n)f_{1,v_i}, f_{1,v_i}\rangle_{\pi_p} \psi_{\mathbb{Q}_p}(u)^{-1}du.
\]

Note that \( \psi_{\mathbb{Q}_p}(u) \) is trivial for \( u \in U(p\mathbb{Z}_p) \). Thus the character \( \tilde{\psi} : U(\mathbb{Z}/p\mathbb{Z}) \to \mathbb{C}^\times \) defined by

\[
\tilde{\psi}(\bar{u}) = \psi_{\mathbb{Q}_p}(u) \text{ for } \bar{u} = [u] \in U(\mathbb{Z}_p)/U(p\mathbb{Z}_p) \cong U(\mathbb{Z}/p\mathbb{Z})
\]

is well defined and non-degenerate. Note that

\[
\langle \pi_p(u)f_{1,v_i}, f_{1,v_i}\rangle_{\pi_p} = \langle \mu(u)v_i, v_i \rangle_W \delta_{u \in K_p}.
\]

Thus we get

\[
S_{\pi_p,p}(D_q) = \int_{U(\mathbb{Q}_p) \cap K_p} \sum_{i=1}^l \langle \mu(u)v_i, v_i \rangle_W \cdot \psi_{\mathbb{Q}_p}(u)^{-1}du.
\]

Since \( (v_i)_{i=1}^l \) is an orthonormal basis of \( W \), we get \( \sum_i \langle \mu(u)v_i, v_i \rangle_W = \text{tr}(\mu(u)) \).

Note that \( U(\mathbb{Q}_p) \cap K_p = U(\mathbb{Z}_p) \), the integrand is \( U(p\mathbb{Z}_p) = U(\mathbb{Q}_p) \cap K_p(p) \) invariant and \( \text{Vol}(U(p\mathbb{Z}_p), du) = p^{-\frac{n(n-1)}{2}} \), so we obtain

\[ S_{\pi_p,p}(D_q) = p^{-\frac{n(n-1)}{2}} \sum_{u \in U(\mathbb{Z}/p\mathbb{Z})} \text{tr}(\mu(u))\tilde{\psi}(u). \]

We start by observing that \( \text{tr}\mu(g) = \chi_\mu(g) \) is the character of \( \mu \) at \( g \). We get that

\[
\dim \pi_p^{K_p(p)} = \dim(W) = \chi_\mu(1) = \prod_{i=1}^{n-1} (p^i - 1).
\]

This proves the last part of our statement. (For the evaluation of the character at \( 1 \) see for example [Pr, Theorem 2]). To deduce the rest we need
some more character theory. Given any finite group $H$ we define the inner product
\[
\langle f, g \rangle_H = \frac{1}{\#H} \sum_{h \in H} f(h)g(h).
\]
With this at hand we can rephrase the statement of (6.3) as
\[
S_{\pi}(a) = \langle \chi_{\mu}|_{U(\mathbb{Z}/p\mathbb{Z})}, \tilde{\psi}_{U(\mathbb{Z}/p\mathbb{Z})} \rangle.
\]
We write $\tilde{\psi}^G$ for the character of the representation obtained by inducing $\tilde{\psi}$ from $U(\mathbb{Z}/p\mathbb{Z})$ to $GL_n(\mathbb{Z}/p\mathbb{Z})$. Let $G_{\tilde{\psi}}$ denote the space of this induced representation. By Frobenius reciprocity we have
\[
\langle \chi_{\mu}|_{U(\mathbb{Z}/p\mathbb{Z})}, \tilde{\psi}_{U(\mathbb{Z}/p\mathbb{Z})} \rangle = \langle \chi_{\mu}, \tilde{\psi}^G \rangle_{GL_n(\mathbb{Z}/p\mathbb{Z})} = \text{Hom}_{GL_n(\mathbb{Z}/p\mathbb{Z})}(W, G_{\tilde{\psi}}).
\]
Recall that $\tilde{\psi}$ is non-degenerate. In this case it is a well known theorem (multiplicity one for $G_{\tilde{\psi}}$) due to Gelfand-Graev that
\[
\dim \text{Hom}_{GL_n(\mathbb{Z}/p\mathbb{Z})}(W, G_{\tilde{\psi}}) \leq 1.
\]
However, by [SZ, Corollary 5.4] the dimension is precisely one for cuspidal representations $(\mu, W)$ (i.e. cuspidal representations are $\tilde{\psi}$-generic). This completes the proof. □

Thus for depth-zero supercuspidal representations $\pi_p$ we could compute $S_{\pi_p}(D_q)$ and $\dim_{\mathbb{C}}(\pi_p^{K_p(p^m)})$ explicitly showing that (6.1) and (6.2) hold on the nose. Here we crucially used the construction of $\pi_p$ from certain representations of $GL_n(\mathbb{Z}/p\mathbb{Z})$ and the representation theory thereof. To treat general supercuspidal representations $\pi_p$ a different argument is needed.

6.3. Parabolically induced representations. Suppose we have a partition $n = n_1 + \ldots + n_k$ and (essentially) square integrable representations $\tau_1, \ldots, \tau_k$ of $M \cong GL_{n_1}(\mathbb{Q}_p) \times \ldots \times GL_{n_k}(\mathbb{Q}_p)$. If $P = MN$ is the standard parabolic subgroup with Levi component $M$, we consider the parabolically induced representation
\[
\xi = \text{Ind}_{P(\mathbb{Q}_p)}^{GL_n(\mathbb{Q}_p)}(\sigma).
\]
Note that at the moment we are not assuming that this representation is unitary or irreducible.

**Lemma 6.6.** For $m \geq 1$ we have
\[
\dim_{\mathbb{C}}(\xi^{K_p(p^m)}) = \#P(\mathbb{Q}_p)\backslash GL_n(\mathbb{Q}_p)/K_p(p^m) \cdot \prod_{i=1}^k \dim \tau_i^{K_p(p^m)}.
\]
In particular, (6.2) holds for $m = 1$ and all smooth admissible representations $\pi_p$ of $GL_n(\mathbb{Q}_p)$.

**Proof.** The dimension formula is given in [MY, Lemma 3.1] or [NV, Lemma 23.2]. The rest is a direct consequence of Lemma 6.5 and the Bernstein-Zelevinsky classification. □
To make things more precise, we fix representatives \( \gamma_1, \ldots, \gamma_l \) of the double quotient \( P(\mathbb{Q}_p) \backslash \text{GL}_n(\mathbb{Q}_p)/K_p(p^m) \) and an orthonormal basis \((v_i)_{i=1}^r\) of \( \sigma^{K_p(p^m)} \). We can parametrize a basis for \( \xi^{K_p(p^m)} \) by tuples \((\gamma_i, v_j)\) with \( 1 \leq i \leq l \) and \( 1 \leq j \leq r \). In particular \( \dim \xi^{K_p(p^m)} = lr \). We model \( \xi \) in the induced picture as follows. Suppose the representation \( \sigma \) is modeled on a space \( W \). Then we consider the space \( V(\sigma) \) of smooth functions \( f: \text{GL}_n(\mathbb{Q}_p) \to W \) such that\footnote{The double use of \( m \in \mathbb{N} \) and \( m \in M(\mathbb{Q}_p) \) should not lead to confusion.}

\[
f(mng) = \delta_p(m)^{\frac{1}{2}} \sigma(m) f(g)
\]
for all \( g \in \text{GL}_n(\mathbb{Q}_p) \), \( n \in N(\mathbb{Q}_p) \) and \( m \in M(\mathbb{Q}_p) \) where \( \delta_p \) is the modular function of \( P \).

**Lemma 6.7.** A basis for \( \xi^{K_p(p^m)} \) in the induced model is given by

\[
B(\xi^{K_p(p^m)}) = \{ \xi(\gamma_i^{-1}) f_{1,j}: 1 \leq j \leq r \text{ and } 1 \leq i \leq l \},
\]

with

\[
f_{1,j}(g) = \text{Vol}(P(\mathbb{Z}_p)K_p(p^m), dk)^{-\frac{1}{2}} \begin{cases} 
\delta_p(m)^{\frac{1}{2}} \sigma(m)v_j & \text{if } g = nmk \text{ for } k \in K_p(p^m), n \in N(\mathbb{Q}_p), m \in M(\mathbb{Q}_p), \\
0 & \text{else.}
\end{cases}
\]

**Proof.** Showing that \( f_{1,j} \) is well defined follows since \( K_p(p^m) \) has an Iwahori factorization with respect to \( P \). Verifying the rest of the statement is similarly straightforward. \( \square \)

Note that \( \hat{\xi} \cong \text{Ind}_{P(\mathbb{Q}_p)}^{\text{GL}_n(\mathbb{Q}_p)}(\hat{\sigma}) \). If we write \((\cdot, \cdot): W \times \hat{W} \to \mathbb{C}\) for the canonical pairing between \((\sigma, W)\) and its contragredient \((\hat{\sigma}, \hat{W})\), we obtain a canonical pairing between the induced representations given by

\[
(f, g) = \int_{K_p} (f(k), g(k)) dk
\]
for \( f \in V(\sigma) \) and \( g \in V(\hat{\sigma}) \). Let \( \hat{v}_1, \ldots, \hat{v}_r \) be a basis of \( \hat{\sigma}^{K_p(p^m)} \), dual to \( v_1, \ldots, v_r \) with respect to \((\cdot, \cdot)\). Then we obtain a basis \( \hat{f}_{i,j} \) with \( 1 \leq i \leq l \) and \( 1 \leq j \leq r \) of \( V(\hat{\sigma})^{K_p(p^m)} \) with

\[
(f_{i_1,j_1}, \hat{f}_{i_2,j_2}) = \delta_{i_1=j_2, j_1=j_2}.
\]

If \( \xi \) is tempered and irreducible, then we can replace the pairing \((\cdot, \cdot): \sigma \times \hat{\sigma} \to \mathbb{C}\) by an inner product on \( W \). We then obtain an invariant inner product on \( V(\sigma) \). In this case we can naturally identify \( f_{i,j} \) with \( \hat{f}_{i,j} \), so that the basis we constructed is orthonormal with respect to this inner product.
Lemma 6.8. Let \( \xi \) be as above and consider the basis \((f_{i,j})_{1 \leq j \leq r, 1 \leq i \leq l}\) of \( \xi_{K_p(p^m)} \) in the induced model. Then we have
\[
(\xi(u)f_{i,j}, \hat{f}_{i,j}) = \frac{\# [P(\mathbb{Z}_p)/(K_p(p^m) \cap P(\mathbb{Q}_p))]^{1/2}}{\text{Vol}(K_p(p^m), dk)^{1/2}} \int_{K_p(p^m)} (f_{i,j}(k'\gamma_iu\gamma_i^{-1}), \hat{\psi}_j)dk',
\]
for \( u \in U(\mathbb{Q}_p) \).

Proof. By the definition of the pairing \( (\cdot, \cdot) \) in the induced picture we have
\[
(\xi(u)f_{i,j}, \hat{f}_{i,j}) = \int_{K_p} (f_{i,j}(ku), \hat{f}_{i,j}(k))dk = \int_{K_p} (f_{i,j}(k'\gamma_iu\gamma_i^{-1}), \hat{f}_{i,j}(k))dk.
\]
We can write \( k \in K_p \) as \( k = \delta k' \) with \( k' \in K_p(p^m) \) and a representative \( \delta \) in \( K_p/K_p(p^m) \). By the definition of \( \hat{f}_1 \) we find that \( f_{i,j}(k) = f_{i,j}(\delta k') \neq 0 \) unless \( \delta \in P(\mathbb{Q}_p)K_p(p^m) \).

Thus we get
\[
(\xi(u)f_{i,j}, \hat{f}_{i,j}) = \sum_{\delta \in P(\mathbb{Z}_p)/(K_p(p^m) \cap P)} \left| \int_{K_p(p^m)} (f_{i,j}(\delta k'\gamma_iu\gamma_i^{-1}), \hat{f}_{i,j}(\delta k'))dk' \right|
\]
\[
= \frac{\# [P(\mathbb{Z}_p)/(K(p^m) \cap P)]}{\text{Vol}(P(\mathbb{Z}_p)K_p(p^m), dk)^{1/2}} \int_{K_p(p^m)} (f_{i,j}(k'\gamma_iu\gamma_i^{-1}), \hat{\psi}_j)dk'
\]
and the proof is complete.

By [LM, Proposition 2.3] the function \( f(u) = (\xi(u)f_{i,j}, \hat{f}_{i,j})\psi_{\mathbb{Q}_p}(u) \) has a stable integral, so that we can define
\[
S_{\xi,p^m}(D_q) = \sum_{1 \leq i \leq l, 1 \leq j \leq r} \int_{U(\mathbb{Q}_p)} (\xi(u)f_{i,j}, \hat{f}_{i,j})\psi_{\mathbb{Q}_p}(u)du,
\]
which extends the definition of \( S_{\pi,p^m}(D_q) \) made for irreducible unitary representations \( \pi_p \). Note that [LM, Proposition 2.3] is stated for irreducible representations \( \xi \), however the proof does not require irreducibility.

We are now going to reduce the computation of \( S_{\xi,p^m}(D_q) \) to the corresponding averages \( S_{\tau_i,p^m}(D_q) \). For convenience we set
\[
S_{\sigma,p^m}(D_q) = S_{\tau_1,p^m}(D_q^{(m_1)}) \cdots S_{\tau_k,p^m}(D_q^{(m_k)}),
\]
where \( D_q = \text{diag}(D_q^{(m_1)}, \ldots, D_q^{(m_k)}) \) and \( S_{\tau_i,p^m}(D_q^{(n_i)}) = 1 \) if \( \tau_i \) is a character (i.e. \( n_i = 1 \)).

Lemma 6.9. In the notation above with \( p^m \parallel q \) we have
\[
S_{\xi,p^m}(D_q) = S_{\sigma,p^m}(D_q).
\]
Proof. Let $P^{\text{op}} = M^{\text{op}}N^{\text{op}}$ denote the parabolic opposite to $P = MN$, and set $T = T_P = P \cap P^{\text{op}}$. We have $P^\top = MN^\top$. Given a representation $\sigma = \tau_1 \otimes \cdots \otimes \tau_k$ we write $\sigma^{\text{op}} = \tau_k \otimes \cdots \otimes \tau_1$ for the natural representation on $P^{\text{op}}$. There is a Weyl element $w$ such that $w^{-1}_P MW_P = M^{\text{op}}$ and $w^{-1}_P NW_P = (N^{\text{op}})^\top$. The long Weyl element would do the job but we can even require that

$$\sigma(m) = \sigma^{\text{op}}(w^{-1}_P mw_P)$$

for all $m \in M(\mathbb{Q}_p)$.

Our first goal is to show that the $U$-integral vanishes for many $i$. In the end only $1 \leq i \leq l$ where $\gamma_i$ is of a very nice shape will contribute. Note that this behavior is special to $D_q$. Indeed we will use the following fact. For each simple root $\alpha$ we have

$$\Psi_{\gamma_i}^\wedge(\cdot)|_{U_\alpha(p^{-1}M^{\text{op}})} \neq 1 \quad \text{and} \quad \Psi_{\gamma_i}^\wedge(\cdot)|_{U_\alpha(p^mZ_p)} \equiv 1.$$

In particular, if there is a simple root $\alpha$ so that the matrix coefficient $(\pi(\cdot)f_{i,j}, f_{i,j})$ is right (or left) $(u')^{-1}U_\alpha(p^{-1}Z_p)u'$-invariant for some $u' \in U$, then the $U$-integral will vanish for this $i$. We will show this invariance property for some of the $i$'s. Recall that $\gamma_1, \ldots, \gamma_l$ are representatives for $P(\mathbb{Q}_p)\backslash GL_n(\mathbb{Q}_p)/K_p(p^m)$, which we can choose to lie inside $K_p$.

Note that $f_{1,j}$ is $N(Z_p)$-invariant. Turning to the more general case, we take $\gamma_i \in K_p$ and consider the map

$$K_p \to K_p/K_p(p) = \text{GL}_n(\mathbb{Z}/p\mathbb{Z}) = \bigcup_{w \in W^M \setminus W} P(\mathbb{Z}/p\mathbb{Z}) w U(\mathbb{Z}/p\mathbb{Z})$$

and write $\gamma_i$ for the image of $\gamma_i$. Suppose $\gamma_i \in P(\mathbb{Z}/p\mathbb{Z}) w U(\mathbb{Z}/p\mathbb{Z})$ with $w \neq w_M$, so that we can assume that $\gamma_i = w_i + pA$ for $u_i \in U(\mathbb{Z}_p)$ and $A \in M_n(\mathbb{Z}_p)$. There is a simple root $\alpha$ with $U_\alpha \subset U \cap w^{-1}NW$. We take $u_i u_\alpha u_i^{-1} \in U_\alpha(p^{-1}Z_p)$. Recall that $\pi(u_\alpha)f_{i,j} = \pi(u_\alpha \gamma_i^{-1})f_{i,j}$. Thus in order to see that $f_{i,j}$ is $U_\alpha(p^{-1}Z_p)$-invariant it is enough to show that $\gamma_i u_\alpha \gamma_i^{-1} \in N(Z_p) + p^mM_n(\mathbb{Z}_p)$. To see this we compute

$$\gamma_i u_\alpha \gamma_i^{-1} = w_i u_i u_\alpha u_i^{-1} - 1 + pA u_\alpha u_i^{-1} w^{-1} + p w_i u_\alpha A + p^2 A u_\alpha A'$$

$$\subseteq w_i u_i u_\alpha u_i^{-1} - 1 + p A u_\alpha^{-1} w^{-1} + p w_i A + p^2 A A' + p^m M_n(\mathbb{Z}_p)$$

Here we used $u_\alpha \in 1 + p^{-m}M_n(\mathbb{Z}_p)$ and

$$1 = (w_i + pA)(u_i^{-1} w^{-1} + p A') = 1 + p A u_i^{-1} w^{-1} + p w_i A + p^2 A A'.$$

However, our choice of $u_\alpha$ ensures that $w_i u_\alpha u_i^{-1} w^{-1} \in N(\mathbb{Z}_p)$. Thus we have seen that $u_\alpha \gamma_i^{-1} \in \gamma_i^{-1} N(\mathbb{Z}_p) K_p(p^m)$, so that $f_{i,j}$ is $u_i^{-1} U_\alpha(p^{-1}Z_p) u_i$-invariant and the corresponding $U$-integral vanishes. In other words $\gamma_i$ does not contribute to $S_{\pi,p^m}(D_q)$.

We have seen that only representatives $\gamma_i$ of $P(\mathbb{Q}_p)\backslash GL_n(\mathbb{Q}_p)/K_p(p^m)$ of the shape $\gamma_i \in w_M N(Z_p) K_p(p)$ contribute to the $i$-sum in $S_{\pi,p^m}(D_q)$. However, due to the Iwahori factorisation of $K_p(p)$ we can write them as $\gamma_i \in P(\mathbb{Z}_p) w_M N(Z_p)$. Thus we have seen that only $i$ with $\gamma_i \in P(\mathbb{Z}_p) w_M U(\mathbb{Z}_p)/K_p(p^m)$
At this point we consider the Iwahori decomposition

\[ n = \prod_{i=1}^{\infty} n_i \]

Inserting this in our condition we obtain

\[ S = u_M n_i \text{ with } n_i \in N^{op}(\mathbb{Z}_p/p^n\mathbb{Z}_p). \]

Here we used the identification

\[ P(\mathbb{Z}_p) \backslash P(\mathbb{Z}_p)w_M U(\mathbb{Z}_p) / K_p(p^m) = P(\mathbb{Z}_p) \backslash P(\mathbb{Z}_p)w_M P^{op}(\mathbb{Z}_p) / K_p(p^m) \]

\[ \cong w_M N^{op}(\mathbb{Z}_p/p^n\mathbb{Z}_p). \]

For these representatives we can compute the \( U \)-integral starting from the expression for the matrix coefficients given in Lemma 6.8. Note that a change of variables \( u \mapsto n_i^{-1} u n_i \) leaves the character invariant. Thus we can assume without loss of generality that \( n_i = n_1 = 1 \) getting

\[
\int_{U(\mathbb{Q}_p)}^{st} (\xi(u) f_{i,j}, \hat{f}_{i,j}) \psi_{Q_p}^\natural(u)^{-1} du
\]

\[
= \int_{U(\mathbb{Q}_p)}^{st} \# \left[ P(\mathbb{Z}_p) / (K_p(p^m) \cap P(\mathbb{Q}_p)) \right] \frac{1}{\text{Vol}(K_p(p^m), dk)^{\frac{1}{2}}} \int_{K_p(p^m)} (f_{1,j}(k' \gamma_i u \gamma_i^{-1}, \hat{v}_j)) dk' \psi_{Q_p}^\natural(u)^{-1} du
\]

\[
= \int_{U(\mathbb{Q}_p)}^{st} \# \left[ P(\mathbb{Z}_p) / (K_p(p^m) \cap P(\mathbb{Q}_p)) \right] \frac{1}{\text{Vol}(K_p(p^m), dk)^{\frac{1}{2}}} \int_{K_p(p^m)} (f_{1,j}(k' u_M u w^{-1}_M, \hat{v}_j)) dk' \psi_{Q_p}^\natural(u)^{-1} du
\]

\[
= \int_{U(\mathbb{Q}_p)}^{st} (\xi(u) f_{1,j}, \hat{f}_{1,j}) \psi_{Q_p}^\natural(u)^{-1} du.
\]

Thus the integral is independent of \( 1 \leq i \leq j_0 \). In summary,

\[
S_{\xi,p^m}(D_q) = \sum_{1 \leq i \leq j_0} \int_{U(\mathbb{Q}_p)}^{st} (\xi(u) f_{i,j}, \hat{f}_{i,j}) \psi_{Q_p}^\natural(u) du
\]

\[
= \sum_{1 \leq i \leq j_0} \int_{U(\mathbb{Q}_p)}^{st} (\xi(u) f_{i,j}, \hat{f}_{i,j}) \psi_{Q_p}^\natural(u) du = j_0 \sum_{1 \leq j \leq r} \int_{U(\mathbb{Q}_p)}^{st} (\xi(u) f_{1,j}, \hat{f}_{1,j}) \psi_{Q_p}^\natural(u) du.
\]

In order to compute the remaining integrals we consider the condition

\[ \gamma_1 k' \gamma_1^{-1} = pk \in P(Q_p) K_p(p^m), \]

which can be rewritten as

\[ k' u = q \tilde{k} \in (P^{op})^\top Q_p K_p(p^m), q = w^{-1}_M p w_M \in (P^{op})^\top Q_p, \tilde{k} = w^{-1}_M k w_M \in K_p(p^m). \]

At this point we consider the Iwahori decomposition \( k' = n_- t n_+ \) with \( n_- \in (N^{op})^\top Q_p \cap K_p(p^m), t \in M^{op}(Q_p) \cap K_p(p^m) \) and \( n_+ \in N^{op}(Q_p) \cap K_p(p^m). \)

Inserting this in our condition we obtain \( n_+ u \in (P^{op})^\top Q_p K_p(p^m) \). We can now suppose without loss of generality that \( u = u_{N^{op}} \cdot u_{M^{op}} \) with \( u_{N^{op}} \in N^{op} \) and \( u_{M^{op}} \in U \cap M^{op}. \) We obtain the condition \( u_{N^{op}} \in K_p(p^m). \) This leads to

\[ f_{1,j}(\gamma_1^{-1} k' u_{N^{op}} u_{M^{op}} \gamma_1) = \text{Vol}(P(\mathbb{Z}_p) K_p(p^m), dk)^{-\frac{1}{2}} \delta_{u_{N^{op}} K_p(p^m)} \sigma(w_M u_{M^{op}} w^{-1}_M) v_j. \]

We now insert this observation in our formula for the matrix coefficient and get

\[
(\xi(u_{N^{op}} u_{M^{op}}) f_{1,j}, \hat{f}_{1,j}) = \delta_{u_{N^{op}} K_p(p^m)} (\sigma(w_M u_{M^{op}} w^{-1}_M) v_j, \hat{v}_j).
\]
Factor $U = (U \cap N^{\text{op}}) \cdot (U \cap M^{\text{op}})$ and note that this factorization induces a factorization of the measures. We thus have
\[
\int_{U(Q_p)} (\xi(u)f_{1,j}, \hat{f}_{1,j}) \psi_{\xi,p}^2(u) du = \text{Vol}(N^{\text{op}} \cap K_p(p^m), dn^{\text{op}}) \int_{U(Q_p) \cap M^{\text{op}}(Q_p)} (\sigma(w_p u' w_p^{-1})v_j, \hat{v}_j) \psi_{\xi,p}^2(u') du'.
\]

We now compute
\[
S_{\xi,p^m}(D_q) = j_0 \text{Vol}(N^{\text{op}} \cap K_p(p^m), dn^{\text{op}}) \sum_{j=1}^r \int_{U \cap M^{\text{op}}} (\sigma^{\text{op}}(u')v_j, \hat{v}_j) \psi_{\xi,p}^2(u') du'.
\]

Since the definition of $S_{\sigma,p^m}(D_q)$ is independent of the bases of $\tau_i K_p(p^m)$ we recognize this as
\[
\sum_{j=1}^r \int_{U} (\pi(u)f_{1,j}, \hat{f}_{1,j}) \psi_{\xi,p}^2(u) du = \text{Vol}(N^{\text{op}} \cap K_p(p^m), dn^{\text{op}}) S_{\sigma,p^m}(D_q).
\]

Note that $S_{\sigma,p^m}(D_q) = S_{\sigma,p^m}(D_q)$. Since the same formula holds for any $i$ in place of 1, we conclude by summing over $1 \leq i \leq j_0$ and realising that $j_0 = \text{Vol}(N^{\text{op}}(Q_p) \cap K_p(p^m), dn^{\text{op}})^{-1}$.

**Remark:** The previous lemma allows us to reduce the computation of $S_{\pi,p^m}$ to the case of supercuspidal representations. Note that if $\xi$ is unitary, then the integrals
\[
I_{i,j} = \int_{U(Q_p)} (\xi(u)f_{i,j}, f_{i,j}) \psi_{\xi,p}^2(u) du
\]
are all positive. Thus in this case one can simply drop all the tuples $(i, j)$ from the definition of $S_{\xi,p^m}$, where $\gamma_i$ does not lie in $P(Q_p) w_M N(Z_p) K_p(p^m)$. The same argument as above then yields the lower bound
\[
(6.4) \quad S_{\xi,p^m}(D_q) \geq S_{\sigma,p^m}(D_q).
\]

However, for our treatment of (non-supercuspidal) square integrable representations $\pi_p$ we need to consider non-unitary $\xi$. In this case the positivity of the integrals $I_{i,j}$ is not guaranteed, so that it is impossible to simply drop unwanted $i$’s from the sum. This motivates the additional work in the proof that went into actually proving equality in (6.4), which is essential. Note that this equality will not hold for $S_{\xi,p^m}(y)$ in general. Indeed it is important in the proof that $\psi_{\xi,p}^2$ is trivial on $U(Q_p) \cap K_p(p^m)$ but non-trivial on $U(Q_p) \cap K_p(p^{m-1})$.

**Lemma 6.10.** Let $q$ be a positive integer with $p^m \parallel q$. Suppose (6.1) holds for $m \in \mathbb{N}$ and for all supercuspidal representations of $\text{GL}_4(Q_p)$ with $2 \leq l \leq n$. Then we have
\[
S_{\pi,p^m}(D_q) \gg 1,
\]
for all smooth irreducible unitary representations $\pi_p$ of $GL_n(\mathbb{Q}_p)$. In particular (6.1) holds unconditionally for $m = 1$.

Proof. We start with the case of essentially square integrable representations. In this case there is $d \mid n$ and a supercuspidal representation $\tau$ of $GL_n(\mathbb{Q}_p)$ so that $\pi_p = St_n(\tau)$. In particular $\pi_p$ is the unique irreducible submodule of $\xi = Ind_{P}^{GL_n}(|\cdot|^{\frac{d-1}{2}} \tau \otimes \cdots \otimes |\cdot|^{\frac{d-1}{2}} \tau)$, where $P$ is the parabolic subgroup associated to the partition $(\frac{n}{d}, \ldots, \frac{n}{d})$. We write $\sigma = |\cdot|^{\frac{d-1}{2}} \tau \otimes \cdots \otimes |\cdot|^{\frac{d-1}{2}} \tau$.

Note that $\pi$ (resp. $\tilde{\pi}$) is the unique generic sub-quotient of $\xi$ (resp $\tilde{\xi}$). Thus the claim follows since $S_{\xi,p^m}(D_q)$ is independent of the choice of basis and the pairing $\xi \times \tilde{\xi} \to \mathbb{C}$ given by

$$(f, f') \mapsto \int_{U(\mathbb{Q}_p)}^{st} (\pi(u) f, f') \psi_{\mathbb{Q}_p}(u) du$$

descends to a pairing between the twisted Jacquet modules $j_{\psi_{\mathbb{Q}_p}}(\xi)$ and $j_{(\psi_{\mathbb{Q}_p})^{-1}}(\tilde{\xi})$. This is given in [LM, Proposition 2.3] and made more precise in [LM, Proposition 2.8].

According to the Langlands classification it remains to consider irreducible representations (parabolically) induced from square-integrable representations. More precisely we can assume that $\pi_p = Ind_{P(\mathbb{Q}_p)}^{GL_n}(\sigma)$, where $P = MN$ is a standard parabolic subgroup and $\sigma$ is an (essentially) square integrable representation of $M(\mathbb{Q}_p)$. However, since we have established the result for square integrable representations, it holds for $\sigma$. We conclude by applying Lemma 6.6 and Lemma 6.9 respectively.

This completes the proof of Theorem 1.3, since we have established (6.1) and (6.2) in all relevant cases and can therefore apply Proposition 6.1. Note that the assumption $q$ squarefree is only used in the treatment of supercuspidal representations. Indeed we crucially use that if $q$ is squarefree only depth-zero supercuspidal representations are relevant. On the other hand the argument reducing to the case of supercuspidal representations is completely general.

6.4. The final results. Dropping everything but the cuspidal part from Proposition 5.2 and combining the resulting inequality with Theorem 1.3 and (2.14), we obtain an unweighted version of Proposition 5.2 for the cuspidal spectrum. At this point there is no difference between the groups $\Gamma(q)$ and $\tilde{\Gamma}(q)$.

Corollary 6.11. There exists an absolute constant $K > 0$ with the following property. Let $M, T > 1$, $q$ squarefree and suppose that $T \leq M^{-K} q^{n+1}$. Fix
a place \( v \) of \( \mathbb{Q} \). If \( v = p \) is finite, assume that \( p \nmid q \). Then
\[
\sum_{\omega \in \text{ONB}(L^2_{\text{cusp}}(X(q))) \atop \|\mu\| \leq M} T^{2\sigma_{\omega,v}} \ll_{v,\epsilon} M^K q^\epsilon \mathcal{V}_q,
\]
where \( \text{ONB} \) stands for any orthonormal basis of eigenforms for the ring of invariant differential operators and the Hecke algebra at \( v \) (if \( v \) is finite).

Theorem 1.1 is now an easy consequence. For \( T = M^{-K}q^{n+1} \) we have (the first inequality is a version of Rankin’s trick)
\[
\mathcal{N}(\sigma, \mathcal{F}) \leq \sum_{\omega \in \text{ONB}(L^2_{\text{cusp}}(X(q))) \atop \|\mu\| \leq M} T^{2\sigma_{\omega,v}} q^{-2\sigma} \ll M^K q^{-2\sigma(n+1)+\epsilon},
\]
and Theorem 1.1 follows from (2.14).

Remark 6.1. This directly yields the following density estimate on the level of cuspidal automorphic representations
\[
\sum_{\pi \mid L^2_{\text{cusp}}(X(q)) \atop \|\mu_\pi\| \leq M, \sigma_{\pi_\infty} > \sigma} \dim_{\mathbb{C}}(\pi^K(q)) \ll M^K \cdot q^{1+\epsilon} \cdot V_q^{1-\frac{2\sigma}{n-1}}.
\]
Note that the additional factor of \( q \) is natural, since twisting \( \pi \) by a character with conductor \( q \) changes the isomorphism class of \( \pi \) but leaves \( \mu_\pi \) and \( \sigma_{\pi_\infty} \) invariant. For comparison one can look at the main term in the Weyl law for principal congruence subgroup given in [Mu, Theorem 0.2].

The weights \( \dim_{\mathbb{C}}(\pi^K(q)) \) can be removed with a little more work. Indeed, we would expect that for cuspidal representations one has \( q^{\frac{n(n-1)}{2} - \epsilon} \ll \dim_{\mathbb{C}}(\pi^K(q)) \ll q^{\frac{n(n-1)}{2} + \epsilon} \). This follows from the discussion below Proposition 6.1 and the fact that cuspidal representations are generic. More precisely, assuming (6.2) is sharp for all generic representations, one obtains the weight-free density estimate
\[
\#\{\pi \mid L^2_{\text{cusp}}(\text{GL}_n(\mathbb{Q})/\text{GL}_n(\mathbb{A})) : \pi^K(q) \neq \{0\}, \|\mu_\pi\| \leq M, \sigma_{\pi_\infty} > \sigma\}
\ll M^K [q^{\frac{n(n+1)}{2}}]^{-\frac{2\sigma}{n}} + \epsilon.
\]

7. The non-cuspidal spectrum

For our applications, we need an upgrade of Corollary 6.11 which includes Eisenstein series and the residual spectrum. The aim of this section is a proof of the following bound.

**Theorem 7.1.** There exists an absolute constant \( K > 0 \) with the following property. Let \( M, T > 1, q \) squarefree, and suppose that \( T \leq M^{-K}q^{n+1} \). Fix a place \( v \) of \( \mathbb{Q} \). If \( v = p \) is finite, assume that \( p \nmid q \). Then
\[
\int_{\Gamma(q)} T^{2\sigma_{\omega,v}} d\omega \ll_{v,\epsilon} M^K q^\epsilon \mathcal{V}_q.
\]
The proof of this is an inductive argument using Corollary 6.11, Langlands' spectral decomposition of $L^2(\Gamma_n(\mathbb{Q}) \setminus \mathbb{GL}_n(\mathbb{A}))$ and the description of the residual spectrum for $\mathbb{GL}_n$ due to Mœglin and Waldspurger. A concise summary of the relevant theory can be found in [GeH, Chapter 10] or in [Ar1]. Full proofs are given in [MW1] and [MW2].

**Proof.** Let us denote the $n$-dependence explicitly by writing

$$X(q) = X^{(n)}(q) = \Gamma(q) \setminus \mathbb{GL}_n(\mathbb{R})/\mathbb{SO}_n(\mathbb{R}).$$

The basic strategy is to use Langlands' spectral decomposition for $L^2(\Gamma_n(\mathbb{Q}) \setminus \mathbb{GL}_n(\mathbb{A}))$ to decompose

$$L^2(X^{(n)}(q)) = \bigoplus_{P/\sim} L^2_P(X^{(n)}(q)),$$

where $P$ runs through all standard parabolic subgroups of $\mathbb{GL}_n$ up to association, including $P = \mathbb{GL}_n$ itself. Furthermore each standard parabolic subgroup of $\mathbb{GL}_n$ is determined by a partition $n = n_1 + \ldots + n_k$, and we can describe each of the spaces $L^2_P(X^{(n)}(q))$ in terms of

$$L^2_{\text{disc}}(X^{(n)}(q)) = L^2_{\mathbb{GL}_n}(X^{(n)}(q))$$

with $1 \leq i \leq k$. The nature of the spaces $L^2_P(X^{(n)}(q))$ will be described below in our proof, but we will use the notation $\int_{\Gamma(q)}^{(P)} T^{2\sigma_{\infty, v}} d\varpi \ll_{v, \varepsilon} M^{K_q} q^{\varepsilon} V_q,$

for each standard parabolic subgroup $P$.

We start by considering the contribution from $P = \mathbb{GL}_n$, i.e. the discrete spectrum. To treat this piece we recall the parametrisation of the discrete spectrum given by Mœglin and Waldspurger. Given $d | n$ and a cuspidal automorphic representation $\pi$ of $GL_\mathfrak{d} \mathbb{A}^1$, we define the Speh representation $\text{Speh}(\pi, d)$ of $GL_n(\mathbb{A})^1$ as the unique irreducible subrepresentation of

$$\text{Ind}_{P_d(\mathbb{A})}^{GL_n(\mathbb{A})}(| \cdot |^{-\frac{d-1}{2}}_{\mathbb{A}} \pi \otimes \ldots | \cdot |^{-\frac{d-1}{2}}_{\mathbb{A}} \pi),$$

where $P_d$ is the standard parabolic subgroup associated to the partition $\frac{n}{d} + \ldots + \frac{n}{d} = n$. It turns out that for each Speh representation (i.e. each tuple $(\pi, d)$) there is a unique irreducible subrepresentation of $L^2(\Gamma_n(\mathbb{Q}) \setminus \mathbb{GL}_n(\mathbb{A}))$ which is isomorphic to $\text{Speh}(\pi, d)$. For convenience we will denote this automorphic representation also by $\text{Speh}(\pi, d)$. One obtains the spectral decomposition

$$L^2_{\text{disc}}(\mathbb{GL}_n(\mathbb{Q}) \setminus \mathbb{GL}_n(\mathbb{A})) = \bigoplus_{d|n} \bigoplus_{\pi \subset L^2_{\text{cusp}}(\mathbb{GL}_\mathfrak{d}(\mathbb{Q}) \setminus \mathbb{GL}_\mathfrak{d}(\mathbb{A}))} \text{Speh}(\pi, d)$$
of the discrete spectrum of $L^2(\text{GL}_n(\mathbb{Q})\backslash \text{GL}_n(\mathbb{A}))$. By de-adelisation we obtain the decomposition

$$L^2_{\text{disc}}(X^{(n)}(q)) = \bigoplus_{d|n} \bigoplus_{\pi|L^2_{\text{cusp}}(X^{n/d}(q))} V_{\text{Speh}(\pi,d)}.$$

Note that the piece given by $d = 1$ is nothing but the cuspidal part of the spectrum. Furthermore we have $\dim V_{\text{Speh}(\pi,d)} = \dim \text{Speh}(\pi,d)^K(q)$ and the factorization $\text{Speh}(\pi,d) = \bigotimes_v \text{Speh}(\pi_v,d)$, where $\text{Speh}(\pi_v,d)$ is the Langlands quotient of $\xi_v = \text{Ind}_{\text{P}_d(Q_v)}^{\text{GL}_n(Q_v)}(| \cdot |^{\frac{d-1}{2}} \pi_v \otimes \ldots | \cdot |^{\frac{d-1}{2}} \pi_v)$. Applying Lemma 6.6 locally we obtain

$$\dim V_{\text{Speh}(\pi,d)} \leq \dim(V_\pi)^d q^{\frac{n(n-1)}{2} - \frac{n(n/d-1)}{2}}.$$

Probably (7.1) is far from optimal but for what follows it is sufficient. Finally, observe that

$$\sigma_{\text{Speh}(\pi,d)}(v) = \sigma_\pi(v) + \frac{d-1}{2}.$$

Note that for a Dirichlet character $\chi$, the representation $\text{Speh}(\omega_\chi,n)$ is precisely the one-dimensional representation given by $\omega_\chi \circ \text{det}$, and we have $\sigma_{\text{Speh}(\omega_\chi,v,n)} = (n-1)/2$. We compute (recall the notation (2.26))

$$\int_{\Gamma(q)} T^{2\sigma_{\omega_\chi,v}} d\varpi \leq T^{n-1} + \sum_{d|n} T^{d-1} q^{\frac{n(n-1)}{2} - \frac{n(n/d-1)}{2}} \sum_{\pi|L^2_{\text{cusp}}(X^{n/d}(q))} \dim(V_{\text{Speh}(\pi,d)}) T^{2\sigma_{\pi,d}}\frac{\dim(V_\pi)^d}{\|\pi\|_\infty \leq M}.$$

We recall our assumption $T \leq M^{-K} q^{n+1}$ where $K = K_n$ is a sufficiently large constant. We apply Corollary 6.11 to each term in the last parenthesis for degree $n/d$ and $T^{1/d}$ in place of $T$. As long as $K_n \geq dK_n/d$, this application of Corollary 6.11 is admissible. By making $K_n$ larger if necessary, the assumption $K_n \geq dK_n/d$ can always be ensured. We thus obtain

$$\int_{\Gamma(q)} T^{2\sigma_{\omega_\chi,v}} d\varpi \ll q^{n-1} + \sum_{d|n} M^{dK_n \frac{d-1}{2} - \frac{n(n-1)}{2}} q^{(d-1)(n+1) + \frac{n(n-1)}{2} - \frac{n(n/d-1)}{2} + \frac{n^2}{2} - d + \epsilon} \ll M^{K_n} q^{\frac{2}{2}} \mathcal{V}_q \left(1 + \sum_{d|n, \alpha \neq n} q^{\alpha_d} \right), \quad \alpha_d = \frac{(d-1)n(2d-n)}{2d}.$$
Note that for $1 \leq d \leq n/2$ we have $\alpha_d \leq 0$. This completes the contribution
the discrete spectrum (i.e. $P = \text{GL}_n$). To summarise, for $T \leq M^{-K}q^{n+1}$ we
have seen that

$$
\sum_{\tau \mid L^2_{\text{disc}}(\chi^{(n)}(q))} \dim(V_{\tau}) T^{2\sigma_{\tau}} = \int_{\Gamma(q) \backslash \text{GL}_n} T^{2\sigma_{\tau},\nu} d\varpi \ll M^{K_n} q^{n^2 - 1 + \varepsilon}.
$$

We will need this bound for $n' < n$ in the process of treating the Eisenstein
contribution.

Before we treat the remaining contributions we study a certain globally
induced representation. Given a standard parabolic subgroup $P = MN$ let
$Z^+_M$ denote the intersection $\tilde{T}(\mathbb{R})$ and the center of $M(\mathbb{A})$. We write $R_{M,\text{disc}}$
for the right regular representation on $L^2_{\text{disc}}(Z^+_M \mathbb{Q} \backslash M(\mathbb{A}))$. We define the Hilbert space

$$
\mathcal{H}_P = \{ \phi: Z^+_M N(\mathbb{A}) M(\mathbb{Q}) \backslash \text{GL}_n(\mathbb{A}) \to \mathbb{C} \text{ measurable:} \}
$$

$$
\{ |m \mapsto \phi(mx)| \in L^2_{\text{disc}}(Z^+_M \mathbb{Q} \backslash M(\mathbb{A})), \}
$$

$$
\|\phi\|^2 = \int_K \int_{Z^+_M \mathbb{Q} \backslash M(\mathbb{A})} |\phi(mk)|^2 dm dk < \infty \}.
$$

For $\lambda \in \mathfrak{a}_{P,\mathbb{C}}$ let $\mathcal{I}_P(\lambda)$ act on $\mathcal{H}_P$ by

$$
[\mathcal{I}_P(\lambda, g)\phi](x) = \phi(xg)e^{(\lambda + \rho_P)(H_P(xg))} e^{-(\lambda + \rho_P)(H_P(x))},
$$

where $\rho_P \in \mathfrak{a}^*_{P,\mathbb{C}}$ is the half sum of positive roots of $\text{GL}_n$
contained in $P$ and $H_P: \text{GL}_n(\mathbb{A}) \to \mathfrak{a}_{P,\mathbb{C}}$ satisfies $\delta_P(p) = e^{2\rho_P(H_P(p))}$ and
$H_P(pk) = H_P(p)$ for all $p \in P(\mathbb{A})$ and all $k \in K$. Write $\mathcal{H}_P^\tau$ for the subspace of $K$-finite elements. For $\lambda = 0$, this is nothing but the representation
obtained by lifting $R_{M,\text{disc}}$ to a representation of $P = MN$ and parabolically
inducing to $\text{GL}_n$ from there. We have the decomposition

$$
R_{M,\text{disc}} = \bigoplus_{\tau_1 \subset L^2_{\text{disc}}(\text{GL}_n(\mathbb{Q}) \backslash \text{GL}_n(\mathbb{A}))} \bigoplus_{\tau_2 \subset L^2_{\text{disc}}(\text{GL}_{n_k}(\mathbb{Q}) \backslash \text{GL}_{n_k}(\mathbb{A}))} \tau_1 \otimes \cdots \otimes \tau_k.
$$

To shorten notation we write $\tau = (\tau_1, \ldots, \tau_k)$ for a tuple of irreducible automorphic representations as in the display and we set $\sigma(\tau) = \tau_1 \otimes \cdots \otimes \tau_k$.

The latter defines an irreducible automorphic representation of $Z^+_M \mathbb{Q} \backslash M(\mathbb{A})$, which factors $\sigma(\tau) = \otimes_v \sigma(\tau)_v$. We define

$$
\sigma(\tau)_{\nu,\lambda}(m) = \sigma(\tau)_v(m) e^{\lambda(H_P(m))} \text{ for } \lambda \in \mathfrak{a}_{P,\mathbb{C}} \text{ and } m \in M(\mathbb{Q}_\nu).
$$

We then have

$$
\mathcal{I}_P(\lambda) = \bigoplus_{\tau} \text{Ind}^{\text{GL}_n}_{P}(\sigma(\tau)_\lambda) = \bigoplus_{\tau \subset v} \text{Ind}^{\text{GL}_n}_{P}(\sigma(\tau)_{\nu,\lambda}).
$$
By Lemma 6.6 we get
\[
\dim \text{Ind}_{P}^{GL_n}(\sigma(\tau)\lambda)^K(q) = \left( \prod_{p|q} \#P(\mathbb{Q}_p) \backslash GL_n(\mathbb{Q}_p)/K_p(q) \right) \prod_{i=1}^{k} \dim(V_{\tau_i})
\]
\[
= q^{\frac{n(n-1)}{2} - \sum_{i=1}^{k} \frac{n_i(n_i-1)}{2}} \prod_{i=1}^{k} \dim(V_{\tau_i}) = q^{\sum_{1 \leq i < j \leq n} n_i n_j} \prod_{i=1}^{k} \dim(V_{\tau_i}).
\]
Furthermore, if \( \lambda \in \mathfrak{a}_P \), then
\[
\sigma_{\text{Ind}_{P}^{GL_n}(\sigma(\tau)\lambda)}(v) = \max_{i=1,\ldots,k} \sigma_{\tau_i}(v).
\]

Langlands’ theory of Eisenstein series yields the orthogonal decomposition
\[
L^2(GL_n(\mathbb{Q}) \backslash GL_n(\mathbb{A})^1) = \bigoplus_{P/\sim} L^2_P(GL_n(\mathbb{Q}) \backslash GL_n(\mathbb{A})^1),
\]
where
\[
L^2_P(GL_n(\mathbb{Q}) \backslash GL_n(\mathbb{A})^1) \cong \int_{\mathfrak{a}_P^+} I_P(\lambda) d\lambda.
\]
The intertwiner can be given more explicitly, but this is not necessary for our purposes. The space \( X_P^{(n)}(q) \) is precisely the space arising by adelisation \( L^2_P(GL_n(\mathbb{Q}) \backslash GL_n(\mathbb{A})^1) \). With this at hand we can write
\[
\int_{\Gamma(q)} T^{2\sigma_{\mu,\omega}} d\omega = \sum_{\tau_1|L^2_{\text{disc}}(X^{(n_1)}(q))} \cdots \sum_{\tau_k|L^2_{\text{disc}}(X^{(n_k)}(q))} \dim \text{Ind}_{P}^{GL_n}(\sigma(\tau)\lambda)^K(q) \cdot T^{2\sigma_{\text{Ind}_{P}^{GL_n}(\sigma(\tau)\lambda)}} \int_{\|\mu\|_{\text{Ind}_{P}^{GL_n}(\sigma(\tau)\lambda)} \leq M} \lambda \in \mathfrak{a}_P^+ d\lambda.
\]

Since the integral can be treated trivially, we proceed by handling all the \( \tau_i \)-sums. Note that even though adelisation works only up to twist (modulo \( q \)) we can use this twist redundancy only to reduce one of the smaller representations, say \( \tau_k \), modulo twists. Note that it also suffices to treat the case \( \max_{i=1,\ldots,k} \sigma_{\tau_i}(v) = k \). For \( i = 1, \ldots, k-1 \) we treat the \( \tau_i \)-sum only using a suitable Weyl law (which is a standard application of the trace formula or alternatively a consequence of (7.2) with \( T = 1 \)):
\[
\sum_{\|\mu\|_{\text{Ind}_{P}^{GL_n}(\sigma(\tau)\lambda)} \leq M} \dim V_{\tau_i} \ll q \sum_{\|\mu\|_{\text{Ind}_{P}^{GL_n}(\sigma(\tau)\lambda)} \leq M} \dim V_{\tau_i} \ll M^{K_{n_i} q^{n_i^2 + \varepsilon}}.
\]

We can also estimate
\[
T^{2\sigma_{\text{Ind}_{P}^{GL_n}(\sigma(\tau)\lambda)}(v)} \leq \left( T^{\frac{n_k}{n}} \right)^{2\sigma_{\tau_k,v}} \cdot T^{\frac{(n-n_k)(n_k-1)}{n}}.
\]
where we use the trivial bound $\sigma_{\tau_k,v} \leq \frac{n_k - 1}{2}$. With this at hand we obtain

$$\int_{\|\mu\| \leq M} (P) T^{2\sigma_{\omega,v}} d\omega \ll M^{\sum_{i \leq k-1} K_{n_i} \sum_{q \leq k-1} n_i^2 + \frac{1}{2} \sum_{1 \leq i \neq j \leq k} n_i n_j + \varepsilon} \cdot \sum_{\tau_k \in \text{disc}(X(n_k)(q))} \|\mu_{\tau_k,\infty}\| \ll M^{\dim V_{\tau_k} \sum_{i \leq k-1} K_{n_i} \sum_{q \leq k-1} n_i^2 + 1} \cdot \sum_{1 \leq i \neq j \leq k} n_i n_j - \sum_{1 \leq i < j < k} n_i n_j + \varepsilon \frac{(n-n_k)(n_k-1)}{n} \cdot \frac{(n-n_k)(n_k-1)}{n}.$$  

Here we have estimated the $\tau_k$-sum using (7.2). Inserting the estimate $T \ll q^{n+1}$ gives us

$$(7.3) \quad \int_{\|\mu\| \leq M} (P) T^{2\sigma_{\omega,v}} d\omega \ll M^K q^{\alpha_P + \varepsilon} V_q,$$

for $\alpha_P = -\frac{1}{2} \sum_{1 \leq i \neq j \leq k} n_i n_j + (n+1) \frac{(n-n_k)(n_k-1)}{n}$. We are done as soon as we can show that $\alpha_P \leq 0$. To see this we recall $n = n_1 + \ldots + n_k$ and write

$$\alpha_P = (n-n_k) \frac{(n_k-1-n)}{n} - \sum_{1 \leq i < j < k} n_i n_j \leq -\sum_{1 \leq i < j < k} n_i n_j \leq 0.$$  

**Remark 7.1.** Note that the main contribution in our density estimate above comes from the discrete (even the cuspidal) part of the spectrum. The continuous contribution coming from some standard parabolic subgroup $P$ has been estimated quite wastefully in (7.3). Indeed our proof shows that

$$\int_{\|\mu\| \leq M} (P) T^{2\sigma_{\omega,v}} d\omega \ll M^K q^{\alpha_P + \varepsilon} V_q,$$

where $\alpha_P < 0$ for all proper parabolic subgroups $P$. One sees that $\alpha_P \leq -\frac{2}{n}$, the worst case being a maximal parabolic subgroup where this bound is attained.

We close this section by recalling that congruence subgroups possess a spectral gap. For a cuspidal representation $\pi$ we have $\|\Re \mu_{\pi}\| \leq 1/2$ (in fact even $\leq \frac{1}{2} - \frac{1}{n^2 - 1}$). If we include the continuous and residual spectrum, then for $n \geq 3$ the worst case appears for the Eisenstein series associated to the $(n-1,1)$ parabolic with the trivial representation on the $(n-1)$-block, in which case

$$(7.4) \quad \|\Re \mu_{\pi}\| \leq \mu^* = \frac{n-2}{2}.$$
8. Applications

In this final section we prove Theorems 1.4 and 1.5. We start with an auxiliary result.

8.1. Bounding eigenfunctions. For each \( \bar{g} \in SL_n(\mathbb{Z}/q\mathbb{Z}) \) choose \( g \in SL_n(\mathbb{Z}) \) such that \( g \) projects onto \( \bar{g} \) modulo \( q \). These matrices form a set of representatives for \( \Gamma(1)/\Gamma(q) \). Note in particular that \( \Gamma(q) \) is normal in \( \Gamma(1) \).

Besides \( L^2 \)-eigenfunctions it is also necessary to deal with certain Eisenstein series that appear in the spectrum of \( X_q \). We start by recalling their (adelic) construction. We use the same notation as in the proof of Theorem 7.1. Let \( P = MN \) be a standard parabolic subgroup of \( GL_n \). Given \( x \in GL_n(\mathbb{A}), \lambda \in a_P^* \) and \( \phi \in H \phi_P \) we define the Eisenstein series by

\[
E(x, \phi, \lambda) = \sum_{\gamma \in P(\mathbb{Q}) \backslash G(\mathbb{Q})} \phi(\gamma x) e^{(\lambda + \rho_P)(H_P(\gamma x))}.
\]

Note that this series converges absolutely for \( \Re(\lambda) \in \rho_P + a^+_P \) and is understood by analytic continuation otherwise. Recall the action of \( I_P(\lambda, \cdot) \) on \( H_P \) and observe that

\[
E(x, I_P(\lambda, g) \phi, \lambda) = E(xg, \phi, \lambda).
\]

In particular, the de-adelisation of \( E(\cdot, \phi, \lambda) \) contributes to the spectral decomposition of \( X_q \) precisely when \( I_P(\lambda, k) \phi = \phi \) for all \( k \in K(q) \). This property is independent of \( \lambda \) and we write \( H_P^{K(q)} \) for the space of \( \phi \in H_P \) with this property. Recall that we can decompose

\[
H_P = \bigoplus_{\sigma \in L^2_{disc}(\mathbb{Z}_M^+ M(\mathbb{Q}) \backslash M(\mathbb{A}))} H_P(\sigma)^\circ
\]

and identify \( H_P(\sigma)^\circ \) with the induced representation \( \text{Ind}_P^{\text{GL}_n}(\sigma) \).

Finally we recall the definition of the truncation operator \( \Lambda^T \) from [Ar2, Section 1], for some fixed \( T \in a^+ \) with \( \alpha(T) \gg 1 \) for all simple roots \( \alpha \) of \( \text{GL}_n \). The latter condition is referred to as “sufficiently regular”. Below we will need the following assumption which is a precise version of (1.3) and discussed at the end of Subsection 1.3.

**Hypothesis 1.** For all standard parabolic subgroups \( P = MN \) of \( \text{GL}_n(\mathbb{A}) \), all irreducible automorphic representations \( \sigma \subset L^2_{disc}(\mathbb{Z}_M^+ M(\mathbb{Q}) \backslash M(\mathbb{A})) \) and all \( \phi \in H_P(\sigma)^{K(q)} \) with \( \| \phi \| = 1 \) we have

\[
\| \Lambda^T E(\cdot, \phi, \lambda) \|^2 = \int_{\text{GL}_n(\mathbb{Q}) \backslash \text{GL}_n(\mathbb{A})^1} |\Lambda^T E(x, \phi, \lambda)|^2 dx \ll_{T, \epsilon} q^\epsilon \| \mu_{\sigma_{\infty, \lambda}} \|^{K(q)}.
\]

We proceed to prove the following lemma.

**Lemma 8.1.** As in (2.18) let \( \varpi \) be a member of an ONB of the spectral decomposition \( X_q \) with spectral parameter \( \mu \). If \( \varpi \) is not discrete, assume
that Hypothesis 1 holds. Then

\[
\sum_{\tilde{g} \in \text{SL}_n(\mathbb{Z}/q\mathbb{Z})} |\varpi(g)|^2 \ll q^f \|\mu\|^K.
\]

**Proof.** This is a standard “local” argument based on properties of spherical functions; see e.g. [Sa4, pp. 11-12]. Fix \( \delta > 0 \) and let \( B_\delta(g) \) denote the ball about \( g \) of radius \( \delta \). For \( k \in g^{-1} \text{SO}_n(\mathbb{R})g \) we have

\[
\int_{B_\delta(g)} \varpi(x) \overline{\phi_\mu(g^{-1}x)} dx = \int_{B_\delta(g)} \varpi(kx) \overline{\phi_\mu(g^{-1}x)} dx
\]

by changing variables and using the left-\text{SO}_n(\mathbb{R})-invariance of the spherical function \( \phi_\mu \) defined in (2.31). Integrating both sides over \( g^{-1} \text{SO}_n(\mathbb{R})g \) and using the mean value property of \( \phi_\mu \), we obtain

\[
\int_{B_\delta(g)} \varpi(x) \overline{\phi_\mu(g^{-1}x)} dx = \varpi(g) \int_{B_\delta(g)} |\phi_\mu(g^{-1}x)|^2 dx.
\]

By the Cauchy-Schwarz inequality we obtain

\[
|\varpi(g)|^2 \leq \left( \int_{B_\delta(g)} |\phi_\mu(g^{-1}x)|^2 dx \right)^{-1} \int_{B_\delta(g)} |\varpi(x)|^2 dx
\]

\[
= \left( \int_{B_\delta(I_n)} |\phi_\mu(x)|^2 dx \right)^{-1} \int_{B_\delta(g)} |\varpi(x)|^2 dx.
\]

For fixed \( 0 < \delta < 1 \) we have from the uniform asymptotic behavior of the spherical function [Va, (5.5)] that

\[
|\phi_\mu(x)|^2 \gg \|\mu\|^{-n(n-1)}, \quad x \in B_\delta(I_n)
\]

if \( \|\mu\| \) is larger than some absolute constant. For bounded \( \mu \), this remains true after possibly reducing \( \delta \) by a compactness argument since \( \phi_\pi(I_n) = 1 \), so that

\[
|\varpi(g)|^2 \ll \|\mu\|^K \int_{B_\delta(g)} |\varpi(x)|^2 dx.
\]

After possibly reducing \( \delta \) further, we may assume that the union of the \( B_\delta(g) \) for \( \tilde{g} \in \text{SL}_n(\mathbb{Z}/q\mathbb{Z}) \) is disjoint and a compact portion \( \Omega \) of \( \Gamma(q) \backslash \text{SL}_n(\mathbb{R}) \). We conclude

\[
(8.1) \quad \sum_{\tilde{g} \in \text{SL}_n(\mathbb{Z}/q\mathbb{Z})} |\varpi(g)|^2 \ll \|\mu\|^K \int_{\Omega} |\varpi(x)|^2 dx \leq \|\mu\|^K
\]

if \( \varpi \) is square integrable (i.e. contributes to the discrete spectrum of \( X_q \)).

Turning towards the case of non-discrete \( \varpi \) we assume (without loss of generality) that \( \varpi(x) = K_q^{-1} \cdot E(x, \phi, \lambda) \), for a standard parabolic subgroup \( P \), an irreducible automorphic representation \( \sigma \subset L^2_{\text{disc}}(Z^1_M(\mathbb{Q}) \backslash M(\mathbb{A})) \) and \( \phi \in \mathcal{H}_P(\sigma)^{K(q)} \) with \( \|\phi\| = 1 \). Here \( K_q \) is a constant with \( K_q \asymp V_q^{1/2} \) coming from the different normalization of the inner products of \( L^2(\text{GL}_n(\mathbb{Q}) \backslash \text{GL}_n(\mathbb{A})) \) and \( X_q \); see (2.27).
We return to proving the desired pointwise bound. Our starting point is the first inequality in (8.1), which remains true in this case. Indeed we have
\[
\sum_{\bar{g} \in \text{SL}_n(\mathbb{Z}/q\mathbb{Z})} |\varpi(g)|^2 \ll \|\mu\|^K \int_\Omega |\varpi(x)|^2 dx \ll \frac{\|\mu\|^K}{K_q^2} \int_\Omega |E(x_\infty, \phi, \lambda)|^2 dx_\infty,
\]
where we indicate that the remaining integral sees only the archimedean place. Without loss of generality we can assume that \( \phi \) transforms under \( K_1(q) \) with respect to some tuple of character \( \zeta \) as in Section 2.6. In particular \(|E(x_\infty, \phi, \lambda)|\) is invariant under the open compact subgroup \( K(q) \subset K'(q) \subset K_1(q) \) where \( K'(q) = \text{diag}(\mathbb{Z}^\times, 1, \ldots, 1)K(q) \). We obtain
\[
\frac{1}{K_q^2} \int_\Omega |E(x_\infty, \phi, \lambda)|^2 dx_\infty = \text{Vol}(K'(q))^{-1} K_q^{-2} \int_{\Omega \times K'(q)} |E(x, \phi, \lambda)|^2 dx_\infty.
\]
Note that
\[
\text{Vol}(K'(q))^{-1} K_q^{-2} \times \frac{[K_{\text{fin}} : K'(q)]}{V_q} \times 1.
\]
Put \( \Omega_k = B_\delta(I_n) \times K_{\text{fin}} \), recall \( \Omega = \bigsqcup_{\tau \in \text{SL}_n(\mathbb{Z}/q\mathbb{Z})} g B_\delta(I_n) \) and observe \( K_{\text{fin}} = \bigsqcup_{\tau \in \text{SL}_n(\mathbb{Z}/q\mathbb{Z})} g^{-1} K'(q) \). By some changes of variables we obtain
\[
\sum_{\bar{g} \in \text{SL}_n(\mathbb{Z}/q\mathbb{Z})} |\varpi(g)|^2 \ll \frac{\|\mu\|^K}{K_q^2} \int_\Omega |E(x_\infty, \phi, \lambda)|^2 dx_\infty \ll \frac{\|\mu\|^K}{\Omega_k} \int_{\Omega_k} |E(x, \phi, \lambda)|^2 dx.
\]
The point of this manoeuvre is that we have removed any \( q \)-dependence from the domain of integration \( \Omega_k \). Instead of extending the integral to all \( \text{GL}_n(\mathbb{Q}) \backslash \text{GL}_n(\mathbb{A}) \), which is not possible since \( E(\cdot, \phi, \lambda) \) is not square integrable, we apply the truncation operator \( \Lambda^T \) first. Indeed, without loss of generality we can assume that \( \Omega_k \) is contained in a Siegel set \( \mathcal{S} \) so that after making \( \delta \) smaller or \( T \) bigger [La1, Lemma 2.6] implies
\[
\int_{\Omega_k} |E(x, \phi, \lambda)|^2 dx = \int_{\Omega_k} |\Lambda^T E(x, \phi, \lambda)|^2 dx.
\]
At this point we can extend the integral and apply Hypothesis 1 to get
\[
\sum_{\bar{g} \in \text{SL}_n(\mathbb{Z}/q\mathbb{Z})} |\varpi(g)|^2 \ll \frac{\|\mu\|^K}{\text{GL}_n(\mathbb{Q}) \backslash \text{GL}_n(\mathbb{A})} \int_{\text{GL}_n(\mathbb{Q}) \backslash \text{GL}_n(\mathbb{A})} |\Lambda^T E(x, \phi, \lambda)|^2 dx \ll q^4 \|\mu\|^K.
\]
as desired.

8.2. Optimal lifting. We are now ready to prove Theorem 1.5, following the strategy of [Sa5]. Let \( T \geq 1 \) be a parameter and let \( f : \text{SL}_n(\mathbb{R}) \to \mathbb{C} \) be as in (2.33). For \( z, w \in \Gamma(q) \backslash \text{SL}_n(\mathbb{R}) \) and \( \bar{g} \in \text{SL}_n(\mathbb{Z}/q\mathbb{Z}) \) define
\[
F(z, w; \bar{g}) = \sum_{\gamma \in \Gamma(q)} f(w^{-1} g \gamma z)
\]
and
\[
W(z) = \sum_{\bar{g} \in \text{SL}_n(\mathbb{Z}/q\mathbb{Z})} \int_{\Gamma(q) \backslash \text{SL}_n(\mathbb{R})} F(z, w; \bar{g}) \frac{d\mu(w)}{\text{vol}(\Gamma(q) \backslash \mathbb{H})}.
\]
where \(d\mu\) is the usual Haar measure on \(\text{SL}_n(\mathbb{R})\). We will estimate \(W(I_n)\) from above and below. On the one hand, let \(N(q, T)\) denote the number of \(\tilde{g} \in \text{SL}_n(\mathbb{Z}/q\mathbb{Z})\) without a lift of norm \(\leq 4nT\). Then for \(w\) in a small ball about \(I_n\) we have \(F(I_n, w; \tilde{g}) = 0\), and hence by (2.35) and (2.14) we conclude

\[
W(I_n) \gg N(q, T) \cdot \frac{T^{2n(n-1)}}{V_q^2}.
\]

On the other hand, we apply the pretrace formula (2.32) to \(F(z, w; \tilde{g})\), subtract the contribution of the trivial representation (constant function), open the square and combine the \(\tilde{g}\)-sum with the integral to obtain

\[
W(z) = \int_{\Gamma(q)}^* |\tilde{f}(\mu_\varpi)|^2 |\varpi(z)|^2 d\varpi
\]

where the * indicates that the trivial representation is omitted. We let \(P = (qT)^\varepsilon\) and distinguish the contribution \(\|\mu_\varpi\| \leq P\) and \(\|\mu_\varpi\| > P\). Correspondingly we write \(W(I_n) = W(I_n)_{\leq P} + W(I_n)_{> P}\). Using (2.34) with sufficiently large \(B\) (in terms of \(\varepsilon\)) and the simple bound

\[
\int_{\|\mu\| \leq R} |\varpi(I_n)|^2 d\varpi \ll R^{n(n+1)/2-1}
\]

for \(R \geq 1\) (which is another application of the pretrace formula – in fact any polynomial bound would suffice), we obtain

\[
W(I_n)_{> P} \ll T^{n(n-1)+\varepsilon}.
\]

Since \(\Gamma(q)\) is normal in \(\Gamma(1)\), we have \(W(I_n) = W(g)\) for \(g \in \Gamma(1) = \text{SL}_n(\mathbb{Z})\), and so

\[
W(I_n)_{\leq P} = \frac{1}{V_q} \sum_{\tilde{g} \in \text{SL}_n(\mathbb{Z}/q\mathbb{Z})} W(g)_{\leq P} = \frac{1}{V_q} \sum_{\tilde{g} \in \text{SL}_n(\mathbb{Z}/q\mathbb{Z})} \int_{\Gamma(q)}^* |\tilde{f}(\mu_\varpi)|^2 |\varpi(g)|^2 d\varpi.
\]

By Lemma 8.1 and (2.34) we conclude

\[
W(I_n)_{\leq P} \ll (Tq)^{\varepsilon} \frac{T^{n(n-1)}}{V_q^2} \int_{\|\mu_\varpi\| \leq P} T^{2n\|\Re \mu_\varpi\|} d\varpi.
\]

In order to apply Theorem 7.1, we choose \(T_0 = (P-Kq^{n+1})^{1/n} \gg P^{-K/n}V_q^{1/n(n-1)}\) and write

\[
T^{2n\|\Re \mu_\varpi\|} \leq T_0^{2n\|\Re \mu_\varpi\|} \left(\frac{T}{T_0}\right)^{2n\mu^*}
\]

using the spectral gap (7.4). Thus by Theorem 7.1 we get

\[
W(I_n)_{\leq P} \ll (Tq)^{\varepsilon} \frac{T^{n(n-1)}}{V_q} \frac{T^{2n\mu^*}}{V_q} \left(\frac{T}{T_0}\right)^{2n\mu^*}.
\]
This dominates (8.3). Combining this with (8.2), we obtain
\[
\frac{N(q, T)}{V_q} \leq (Tq)^\varepsilon \frac{V_q}{T^{n(n-1)}} \left( \frac{T}{T_0} \right)^{2n\mu^*} \ll (Tq)^\varepsilon \left( \frac{V_q}{T^{n(n-1)}} \right)^{1/(n-1)} \ll V_q^{-\delta}
\]
if \( T \gg q^{1+1/n+\varepsilon} \) and \( \delta = \delta(\varepsilon) > 0 \) is sufficiently small.

8.3. A counting problem. The argument for the proof of Theorem 1.4 is very similar to the preceding proof. With the same choice of \( f \) as in (2.33) we have
\[
\sum_{\gamma \in \Gamma(q)} \frac{1}{\|\gamma\|} \leq \sum_{\gamma \in \Gamma(q)} f(\gamma) = \frac{1}{V_q} \sum_{\tilde{g} \in SL_n(\mathbb{Z}/q\mathbb{Z})} \sum_{\gamma \in \Gamma(q)} f(g^{-1}\gamma g)
\]
where as before we used that \( \Gamma(q) \) is normal in \( SL_n(\mathbb{Z}) \). Applying the pretrace formula (2.32) and using also Lemma 8.1 and (2.34), we obtain
\[
\sum_{\gamma \in \Gamma(q)} \frac{1}{\|\gamma\|}, \sum_{\tilde{g} \in SL_n(\mathbb{Z}/q\mathbb{Z})} \int_{\Gamma(q)} |\varpi(g)|^2 \tilde{f}(\mu_\varpi) d\varpi \ll \frac{(Tq)^\varepsilon}{V_q} \int_{\Gamma(q)} T^{n\left(\frac{n-1}{2} + \|\mu_\varpi\|\right)} d\varpi.
\]
Again we choose \( T_0 = \left(P^{-K} q^{n+1}/q^n \gg P^{-2K/n} V_q^{2/(n-1)} \right) \) and apply Theorem 7.1 in combination with (7.4). Then the above is
\[
\ll (Tq)^\varepsilon T^{n(n-1)/2} V_q \left(1 + \frac{T}{T_0}\right)^{n(n-1)/2} = (Tq)^\varepsilon \left( T^{n(n-1)/2} + \frac{T^{n(n-1)-1}}{V_q} \right)
\]
as claimed.

Acknowledgement: The authors would like to thank S. Jana and A. Kamber for useful discussions on their work [JK]. We thank the referees for a careful reading of the manuscript.

References

[Ar1] J. Arthur, *Eisenstein series and the trace formula*, in: Automorphic forms, representations and \( L \)-functions, Corvallis/Oregon 1977, Proc. Symp. Pure Math. 33 (1979), 253-274

[Ar2] J. Arthur, *A trace formula for reductive groups. II. Applications of a truncation operator*, Compos. Math. 40 (1980), 87-121

[As] E. Assing, *The sup-norm problem beyond the newform*, arXiv:2111.01893

[Ba] W. Banks, *Twisted symmetric-square \( L \)-functions and the nonexistence of Siegel zeros on \( GL(3) \)*, Duke Math. J. 87 (1997), 343-353.

[Bl] V. Blomer, *Density theorems for \( GL(n) \)*, Inventiones Math. 232 (2023), 683-711

[BB] V. Blomer, F. Brumley, *The role of the Ramanujan conjecture in analytic number theory*, Bull. Amer. Math. Soc. 50 (2013), 267-320

[BHM] V. Blomer, G. Harcos, P. Maga, *Analytic properties of spherical cusp forms on \( GL(n) \)*, J. Anal. Math. 140 (2020), 483-510

[BM] V. Blomer, S. H. Man, *Bounds for Kloosterman sums on \( GL(n) \)*, arXiv:2208.01295

[Br] F. Brumley, *Effective multiplicity one on \( GL_N \) and narrow zero-free regions for Rankin-Selberg \( L \)-functions*, Amer. J. Math. 128 (2006), 1455-1474
[BH] C. J. Bushnell, G. Henniart, *Counting the discrete series for GL(n)*, Bull. Lond. Math. Soc. **39** (2007), 133-137

[DR] R. Dabrowski, M. Reeder, *Kloosterman sets in reductive groups*, J. Number Theory **73** (1998), 228-255

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

[FHMM] M. Frączyk, G. Harcos, P. Maga, D. Milićević, *The density hypothesis for horizontal families of lattices*, arXiv:2007.13961

[FR] S. Friedberg, *Poincaré series for GL(n): Fourier expansion, Kloosterman sums, and algebro-geometric estimates*, Math. Z. **196** (1987), 165-188

[FG] S. Friedberg, D. Goldfeld, *Mellin transforms of Whittaker functions*, Bull. Soc. Math. France **121** (1993), 91-107

[GL] S. Gelbart, E. Lapid, *Lower bounds for L-functions at the edge of the critical strip*, Amer. J. Math. **128** (2006), 619-638

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

[GGN] A. Ghosh, A. Gorodnik, Nevo, *Diophantine approximation and automorphic spectrum*, Int. Math. Res. Not. IMRN 2013, 5002-5058

[GH] D. Goldfeld, J. Hundley, *Automorphic representations and L-functions for the general linear group, II*, Camb. Studies Adv. Math. **130** (2011)

[GK] K. Golubev, A. Kamber, *On Sarnak’s density conjecture and its applications*, arXiv:2004.00373

[GN] A. Gorodnik, A. Nevo, *Quantitative ergodic theorems and their number-theoretic applications*, Bull. Amer. Math. Soc. (N.S.) **52** (2015), 65-113

[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.

[HL] J. Hoffstein, P. Lockhart, *Coefficients of Maass forms and the Siegel zero*. With an appendix by D. Goldfeld, J. Hoffstein and D. Lieman, Ann. of Math. **140** (1994), 161-181

[Hum] P. Humphries, *Density Theorems for Exceptional Eigenvalues for Congruence Subgroups*, Algebra & Number Theory **12** (2018), 1581-1610

[Hu] M. Huxley, *Exceptional eigenvalues and congruence subgroups*, in: The Selberg trace formula and related topics, Contemp. Math. **53** (1986), 341-349

[Iw] H. Iwaniec, *Small eigenvalues of Laplacian for Γ_0(N)*, Acta Arith. **56** (1990), 65-82

[IK] H. Iwaniec, E. Kowalski, *Analytic number theory*, AMS Colloquium Publications **53**, Amer. Math. Soc. 2004.

[Ja] S. Jana, *Applications of analytic newvectors for GL(n)*, Math. Ann. **380** (2021), 915-952

[JK] S. Jana, A. Kamber, *L^2 growth of Eisenstein series and optimal diophantine exponent*, arXiv:2210.16291

[KL] A. Kamber, H. Lavner, *Optimal lifting for the projective action of SL_2(Z)*, Algebra and Number Theory **17** (2023), 749-774

[La1] E. M. Lapid, *On the fine spectral expansion of Jacquet’s relative trace formula*, J. Inst. Math. Jussieu **5** (2006), 263-308

[La2] E. M. Lapid, *On the support of matrix coefficients of supercuspidal representations of the general linear group over a local non-archimedean field*, in: Representations of reductive groups, 211-218, Proc. Sympos. Pure Math. **101**, Amer. Math. Soc. 2019

[LM] E. Lapid, Z. Mao, *A conjecture on Whittaker-Fourier coefficients of cusp forms*, J. Number Theory **146** (2015), 448-505

[Li] X. Li, *Upper bounds on L-functions at the edge of the critical strip*, IMRN 2010, 727-755
On the generalized Ramanujan conjecture for \( GL(n) \), in: Automorphic forms, automorphic representations, and arithmetic (Fort Worth, TX, 1996), Proc. Sympos. Pure Math. 66 (1999), 301-310, Amer. Math. Soc., Providence, RI

S. D. Miller, On the existence and temperedness of cusp forms for \( SL_3(\mathbb{Z}) \), J. Reine Angew. Math. 533 (2001), 127-169.

M. Miyauchi, T. Yamauchi, A remark on conductor, depth and principal congruence subgroups, arXiv:2107.08130

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

C. Mœglin, J.-L. Waldspurger, Le spectre résiduel de \( GL(n) \), Ann. Sci École Norm. Sup. 22 (1989), 605-674

W. Müller, Weyl’s law for the cuspidal spectrum of \( SL_n \), Ann. of Math. (2) 165 (2007) 275-333

P. D. Nelson, Spectral aspect subconvex bounds for \( U_{n+1} \times U_n \), arXiv:2012.02187

P. D. Nelson, A. Venkatesh, The orbit method and analysis of automorphic forms, Acta Math. 226 (2021), 1-209

O. Parzanchevsky, P. Sarnak, Super-Golden-Gates for \( PU(2) \), Adv. Math. 327 (2018), 869-901

D. Prasad, The space of degenerate Whittaker models for general linear groups over a finite field, Internat. Math. Res. Notices 11 (2000), 579-595

D. Ramakrishnan, Modularity of the Rankin-Selberg \( L \)-series, and multiplicity one for \( SL(2) \), Ann. of Math. 152 (2000), 45-111

D. Ramakrishnan, S. Wang, On the exceptional zeros of Rankin-Selberg \( L \)-functions, Compositio Math. 135 (2003), 211-244

D. E. Rohrlich, Elliptic curves and the Weil-Deligne group, Elliptic curves and related topics, CRM Proc. Lecture Notes 4 (1994), 125-157

P. Sarnak, Statistical properties of eigenvalues of the Hecke operators, in: Analytic number theory and Diophantine problems, Progr. Math. 70 (1987), 321-331

P. Sarnak, Diophantine Problems and Linear Groups, Proceedings of the ICM Kyoto (1990), 459-471

P. Sarnak, Notes on the generalized Ramanujan conjectures, in: Harmonic analysis, the trace formula, and Shimura varieties, Clay Math. Proc. 4 (2005), 659-685

P. Sarnak, Letter to Morawetz, https://publications.ias.edu/node/480

P. Sarnak, Letter on the Solvay-Kitaev Theorem and Golden Gates. Appendix: Letter on optimal strong approximation, https://publications.ias.edu/sarnak/paper/2637

P. Sarnak, X. Xue, Bounds for multiplicities of automorphic representations, Duke Math. J. 64 (1991), 207-227

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

C. L. Siegel, Über die Classenzahl quadratischer Zahlkörper, Acta Arith. 1 (1936), 83-86

A. J. Silberger, E.-W. Zink, The characters of the generalized Steinberg representations of finite general linear groups on the regular elliptic set, Trans. Amer. Math. Soc. 352 (2000), 3339-3356

E. Stade, Mellin transforms of \( GL(n, \mathbb{R}) \) Whittaker functions, Amer. J. Math. 123 (2001), 121-161

E. Stade, Archimedean \( L \)-factors on \( GL(n) \times GL(n) \) and generalized Barnes integrals, Israel J. Math. 127 (2002), 201-219

G. Stevens, Poincarè series on \( GL(r) \) and Kloostermann sums, Math. Ann. 277 (1987), 25-51
[Va] V. S. Varadarajan, The method of stationary phase and applications to geometry and analysis on Lie groups, in: Algebraic and analytic methods in representation theory, Perspect. Math. 17 (1997), 167-242

[We] T. Wedhorn, The local Langlands correspondence for GL(n) over p-adic fields, School on Automorphic Forms on GL(n), ICTP Lect. Notes 21 (2008), 237-320

Mathematisches Institut, Endenicher Allee 60, 53115 Bonn
Email address: blomer@math.uni-bonn.de
Email address: assing@math.uni-bonn.de