Forum of Mathematics, Sigma (2019), vol. , e1, 46 pages
doi:

ENDOSCOPY AND COHOMOLOGY IN A TOWER OF CONGRUENCE MANIFOLDS FOR $U(n, 1)$

SIMON MARSHALL[1] and SUG WOO SHIN[2]

[1]Department of Mathematics
University of Wisconsin – Madison
480 Lincoln Drive
Madison
WI 53706, USA; marshall@math.wisc.edu
[2]Department of Mathematics
University of California, Berkeley
901 Evans Hall, Berkeley, CA 94720, USA / Korea Institute for Advanced Study,
Dongdaemun-gu, Seoul 130-722, Republic of Korea; sug.woo.shin@berkeley.edu

Abstract

By assuming the endoscopic classification of automorphic representations on inner forms of unitary groups, which is currently work in progress by Kaletha, Minguez, Shin, and White, we bound the growth of cohomology in congruence towers of locally symmetric spaces associated to $U(n, 1)$. In the case of lattices arising from Hermitian forms, we expect that the growth exponents we obtain are sharp in all degrees.

2010 Mathematics Subject Classification: 32N10 (primary); 32M15 (secondary)

1. Introduction

This paper studies the limit multiplicity problem for cohomological automorphic forms on arithmetic quotients of $U(N-1, 1)$. Let F be a totally real number field with ring of integers $\mathcal{O}_F$. Write $\mathbb{A}$ for the ring of adeles over F. Let $U(N) = U_{E/F}(N)$ denote the quasi-split unitary group with respect to a totally imaginary quadratic extension E of F. Let G be a unitary group over F that is an inner form of $U_{E/F}(N)$. We assume that G has signature $(N-1, 1)$ at one real place and compact factors at all other real places. Let S be a finite set of places to be defined later, and which includes all infinite places, and let $\mathfrak{n} \subset O_F$ be a nonzero ideal that is divisible only by primes away from S that split in E/F. We let $K(\mathfrak{n}) \subset G(\mathbb{A}_f)$ be the compact congruence subgroup of level $\mathfrak{n}$, and let $\Gamma(\mathfrak{n}) = G(F) \cap K(\mathfrak{n})$ be the congruence arithmetic lattice in $U(N-1,1)$ of level $\mathfrak{n}$ associated to G. Let $Y(\mathfrak{n})$ be the manifold $\Gamma(\mathfrak{n})\backslash U(N-1,1)/U(N-1)\times U(1)$, which is a connected finite volume complex hyperbolic manifold of complex dimension $N-1$. (See (12) below for the precise definition.) Write $h^d_{(2)}(Y(\mathfrak{n}))$ for the dimension of the L^2-cohomology of $Y(\mathfrak{n})$ in degree $d \geq 0$.

Theorem 1.1. *Assume the endoscopic classification for inner forms of $U(N)$ stated in Theorem 1.7.1 of [16].*[1] *If $d < N-1$, we have*

$$h^d_{(2)}(Y(\mathfrak{n})) \ll_\epsilon \mathrm{vol}(Y(\mathfrak{n}))^{Nd/(N^2-1)+\epsilon}.$$

The case $d > N-1$ follows by Poincaré duality. It is well known that $h^{N-1}_{(2)}(Y(\mathfrak{n})) \sim \mathrm{vol}(Y(\mathfrak{n}))$. Previous results of this type in the case of $U(2,1)$ and $U(2,2)$ can be found in work of the first author [20, 19]; see also [12] for the case of $U(N-1,1)$.

Theorem 1.1 fits into the general framework of estimating the asymptotic multiplicities of automorphic forms. We now recall the general formulation of this problem, and some of the previous results on it. Let G be a semisimple real algebraic group with no compact factors. We still write G for $G(\mathbb{R})$, the real group of $\mathbb{R}$-points, if there is no danger of confusion. If $\Gamma \subset G$ is a lattice and π an irreducible unitary representation of G, we let $m(\pi,\Gamma)$ be the multiplicity with which π appears in $L^2(\Gamma\backslash G)$. If we now assume that Γ is congruence arithmetic and that $\Gamma_n \subset \Gamma$ is a family of principal congruence subgroups, the limit multiplicity problem is to provide estimates for $m(\pi,\Gamma_n)$.

A general principle that has emerged from work on this problem is that, the further π is from being discrete series, the better bounds one should be able to prove for $m(\pi,\Gamma_n)$. If we define $V(n) = \mathrm{vol}(\Gamma_n\backslash G)$, the trivial bound (at least when Γ is cocompact) is $m(\pi,\Gamma_n) \ll V(n)$, and it is known from work of de George and Wallach [13] (if Γ is cocompact) and Savin [29] (if it is not) that this is realized if and only if π is in the discrete series. In the cocompact case, it also follows from [13] that if π is nontempered, then one has a bound of the

[1] The introduction of [16] clarifies the conditionality of Theorem 1.7.1. In a nutshell, we are still waiting for the remaining case in Chaudouard–Laumon's proof of the weighted fundamental lemma and the papers [A25,A26] as cited in [1]. In addition we need the sequel papers [KMSb] and [KMSa] (still in preparation) as cited in [16] to complete the proof of Theorem 1.7.1 for pure inner forms and all inner forms of $U(N)$, respectively. In the meantime, the stabilization of the twisted trace formula has been completed by Moeglin–Waldspurger [23, 24], so this is no longer an obstacle.

form $m(\pi,\Gamma_n) \ll V(n)^{1-\delta(\pi)}$ for some $\delta(\pi) > 0$; see the introduction of [28] for an explanation of this principle, and [33] for an explicit determination of such a $\delta(\pi)$ in some cases.

For the most highly nontempered representation, namely the trivial one, one has $m(\pi,\Gamma_n) = 1$. Sarnak and Xue [28] made a conjecture that interpolates between this and $m(\pi,\Gamma_n) \ll V(n)$ in the discrete series case. Define $p(\pi)$ to be the infimum over p for which the K-finite matrix coefficients of π lie in $L^p(G)$. We then have:

Conjecture 1 (Sarnak-Xue)**.** *For fixed π, we have $m(\pi,\Gamma_n) \ll_\epsilon V(n)^{2/p(\pi)+\epsilon}$.*

Note that Conjecture 1 is weaker than the trivial bound in both cases of π discrete or trivial. The point is that it is much stronger for general nontempered π than what one can prove using the methods of deGeorge–Wallach mentioned above. Sarnak and Xue established Conjecture 1 for $SL(2,\mathbb{R})$ or $SL(2,\mathbb{C})$, and proved an approximation for $SU(2,1)$ that, in our setting, implies that $h^1(Y(\mathfrak{n})) \ll \mathrm{vol}(Y(\mathfrak{n}))^{7/12+\epsilon}$ when $N = 3$ and Γ is cocompact and arises from a Hermitian form.

We show in Proposition 3.1 that the representations π of $U(N-1,1)$ contributing to $h^d_{(2)}(Y(\mathfrak{n}))$ all have $p(\pi) \geq 2(N-1)/d$. In the setting of Theorem 1.1, Conjecture 1 therefore predicts that $h^d_{(2)}(Y(\mathfrak{n})) \ll_\epsilon \mathrm{vol}(Y(\mathfrak{n}))^{d/(N-1)+\epsilon}$, so that Theorem 1.1 in fact represents a strengthening of this conjecture.

We note that there has also been significant progress recently on the problem of showing that the normalized discrete spectral measure of $L^2(\Gamma_n\backslash G)$ tends weakly to the Plancherel measure of G. This work is in some sense orthogonal to ours, and as formulated these results do not provide information on $m(\pi,\Gamma_n)$ beyond showing that $m(\pi,\Gamma_n)/V(n)$ approaches the expected value.

1.1. Outline of the proof. We go back to the unitary group G over F introduced at the very beginning. Let K_∞ denote a maximal compact subgroup of $G(F \otimes_{\mathbb{Q}} \mathbb{R})$ so that K_∞ is isomorphic to $U(N-1) \times U(1) \times U(N)^{[F:\mathbb{Q}]-1}$. It will be more convenient for us to work on the possibly disconnected arithmetic quotients $X(\mathfrak{n}) = G(F)\backslash G(\mathbb{A})/K(\mathfrak{n})K_\infty$. If q_v denotes the order of the residue field of F_v, we prove the following more precise bound.

Theorem 1.2. *Assume the endoscopic classification for inner forms of $U(N)$ stated in Theorem 1.7.1 of [16]. If $d < N-1$, we have*

$$h^d_{(2)}(X(\mathfrak{n})) \ll \prod_{v|\mathfrak{n}} (1 - 1/q_v) N\mathfrak{n}^{Nd+1},$$

except when $N = 4$ and $d = 2$ when we have $h^d_{(2)}(X(\mathfrak{n})) \ll_\epsilon N\mathfrak{n}^{Nd+1+\epsilon}$.

Theorem 1.1 follows from this, as $X(\mathfrak{n})$ contains $\gg_\epsilon N\mathfrak{n}^{1-\epsilon}$ copies of $Y(\mathfrak{n})$ and we have $\mathrm{vol}(Y(\mathfrak{n})) = N\mathfrak{n}^{N^2-1+o(1)}$. We now give an outline of the proof of Theorem 1.2. For simplicity, we shall either omit or simplify much of the notation for things like Arthur parameters and packets. Because of this, all notation introduced here is temporary. (Refer to Section 2 below for unexplained notation.)

Let $\Phi_{\mathrm{sim}}(n)$ denote the set of conjugate self-dual cusp forms on $GL(n, \mathbb{A}_E)$, and let $\nu(l)$ denote the unique irreducible (complex algebraic) representation of $SL(2, \mathbb{C})$ of dimension l. Let $U(n)$ be the quasi-split unitary group of degree n with respect to E/F. Let $\Psi_2(n)$ denote the set of square-integrable Arthur parameters for $U(n)$, which are formal sums $\psi = \phi_1 \boxtimes \nu(n_1) \boxplus \cdots \boxplus \phi_k \boxtimes \nu(n_k)$ with $\phi_i \in \Phi_{\mathrm{sim}}(m_i)$, subject to certain conditions including that $n = \sum_{i\geq 1} n_i m_i$ and that the pairs $\phi_i \boxtimes \nu(n_i)$ have to be distinct. Any $\phi \in \Phi_{\mathrm{sim}}(n)$ (resp. $\psi \in \Psi_2(n)$) has localizations ϕ_v (resp. ψ_v), which are local Langlands parameters (resp. Arthur parameters) for $U(n)$. To each $\psi \in \Psi_2(N)$ and each place v of F, there is associated a local packet $\Pi_{\psi_v}(G)$ of representations of $G(F_v)$, and a global packet $\Pi_\psi(G) = \prod \Pi_{\psi_v}(G)$. If $\psi \in \Psi_2(N)$ and $K \subset G(\mathbb{A}_f)$ we define $\dim_G(K, \psi) = \sum_{\pi \in \Pi_\psi(G)} \dim(\pi_f^K)$. Similarly, one may associate to $\psi \in \Psi_2(n)$ a packet $\Pi_\psi(U(n)) = \prod \Pi_{\psi_v}(U(n))$ for $U(n, \mathbb{A})$, and we define $\dim_{U(n)}(K, \psi)$ for $K \subset U(n, \mathbb{A}_f)$ analogously to $\dim_G(K, \psi)$.

The main result of the endoscopic classification implies that the automorphic spectrum of G is contained in the union of $\Pi_\psi(G)$ for $\psi \in \Psi_2(N)$. If we combine this classification with Matsushima's formula, we have

$$h^d_{(2)}(X(\mathfrak{n})) \leq \sum_{\psi \in \Psi_2(N)} \sum_{\pi \in \Pi_\psi(G)} \dim H^d(\mathfrak{g}, K_\infty; \pi_\infty) \dim \pi_f^{K(\mathfrak{n})}, \tag{1}$$

where $H^d(\mathfrak{g}, K_\infty; \pi_\infty)$ denotes the relative Lie algebra cohomology of π_∞ as in [6]. The main part of the proof involves using the structure of the packets $\Pi_\psi(G)$ to bound the right hand side of (1) in terms of global multiplicities on smaller quasi-split unitary groups, which we then bound using a theorem of Savin. The key fact that allows us to control the power of $N\mathfrak{n}$ we obtain is due to Bergeron, Millson, and Moeglin [3, Prop 13.2], and essentially states that if there exists $\pi \in \Pi_\psi(G)$ with $H^d(\mathfrak{g}, K_\infty; \pi_\infty) \neq 0$, then ψ must contain a representation $\nu(n)$ with $n \geq N - d$.

We define a shape to be a list of pairs $(n_1, m_1), \ldots, (n_k, m_k)$ with $\sum_{i\geq 1} m_i n_i = N$, and may naturally talk about the shape of an Arthur parameter. If $\mathcal{S} = (n_1, m_1), \ldots, (n_k, m_k)$ is a shape, we let $\Psi_2(N)_{\mathcal{S}} \subset \Psi_2(N)$ be the set of parameters having that shape. If $\psi \in \Psi_2(N)_{\mathcal{S}}$, we let $\phi_i \in \Phi_{\mathrm{sim}}(m_i)$ be the terms in the decomposition $\psi = \phi_1 \boxtimes \nu(n_1) \boxplus \cdots \boxplus \phi_k \boxtimes \nu(n_k)$. We also define $P_{\mathcal{S}}$ to be the standard parabolic in GL_N of type

$$(\underbrace{m_1, \ldots, m_1}_{n_1 \text{ times}}, \ldots, \underbrace{m_k, \ldots, m_k}_{n_k \text{ times}}).$$

We now fix $\mathcal{S}$, and bound the contribution to (1) from $\Psi_2(N)_{\mathcal{S}}$, which we denote $h^d_{(2)}(X(\mathfrak{n}))_{\mathcal{S}}$. As mentioned above, we may assume that $n_1 \geq N - d$. We may restrict our attention to those $\psi \in \Psi_2(N)_{\mathcal{S}}$ for which there is $\pi \in \Pi_\psi(G)$ with $H^d(\mathfrak{g}, K_\infty; \pi_\infty) \neq 0$. This condition restricts ψ_∞, and hence $\phi_{i,\infty}$, to finite sets which we denote Ψ_∞ and $\Phi_{i,\infty}$, so that

$$h^d_{(2)}(X(\mathfrak{n}))_{\mathcal{S}} \ll \sum_{\substack{\psi \in \Psi_2(N)_{\mathcal{S}} \\ \psi_\infty \in \Psi_\infty}} \dim_G(K(\mathfrak{n}), \psi).$$

In Section 5 we prove Proposition 5.1, which states that if the principal congruence subgroups $K_i(\mathfrak{n}) \subset U(m_i, \mathbb{A}_f)$ are chosen correctly, then one can bound $\dim_G(K(\mathfrak{n}), \psi)$ in terms of $\dim_{U(m_i)}(K_i(\mathfrak{n}), \phi_i)$. For most choices of $\mathcal{S}$, this bound has the form

$$\dim_G(K(\mathfrak{n}), \psi) \ll N\mathfrak{n}^{\dim GL_N/P_{\mathcal{S}} + \epsilon} \prod_{i=1}^{k} \dim_{U(m_i)}(K_i(\mathfrak{n}), \phi_i)^{n_i}. \tag{2}$$

We prove this bound by factorizing both sides over places of F. At nonsplit places we apply the trace identities that appear in the definition of the local packets $\Pi_{\psi_v}(G)$. At split places, $\Pi_{\psi_v}(G)$ is a singleton $\{\pi_v\}$, and we use the description of π_v as the Langlands quotient of a representation induced from $P_{\mathcal{S}}$.

We next sum the bound (2) over $\psi \in \Psi_2(N)_{\mathcal{S}}$, or equivalently we sum ϕ_i over $\Phi_{\text{sim}}(m_i)$, which gives

$$\begin{aligned} h^d_{(2)}(X(\mathfrak{n}))_{\mathcal{S}} &\ll N\mathfrak{n}^{\dim GL_N/P_{\mathcal{S}} + \epsilon} \prod_{i=1}^{k} \sum_{\substack{\phi_i \in \Phi_{\text{sim}}(m_i) \\ \phi_{i,\infty} \in \Phi_{i,\infty}}} \dim_{U(m_i)}(K_i(\mathfrak{n}), \phi_i)^{n_i} \\ &\leq N\mathfrak{n}^{\dim GL_N/P_{\mathcal{S}} + \epsilon} \prod_{i=1}^{k} \left(\sum_{\substack{\phi_i \in \Phi_{\text{sim}}(m_i) \\ \phi_{i,\infty} \in \Phi_{i,\infty}}} \dim_{U(m_i)}(K_i(\mathfrak{n}), \phi_i) \right)^{n_i}. \end{aligned} \tag{3}$$

If we define $\Theta_{i,\infty}$ to be the union of $\Pi_{\phi_{i,\infty}}(U(m_i))$ over $\phi_{i,\infty} \in \Phi_{i,\infty}$, then $\Theta_{i,\infty}$ is finite. Moreover, because the parameters ϕ_i are simple generic, the packet

$\Pi_{\phi_i}(U(m_i))$ is stable, so all representations in it occur discretely on $U(m_i)$. This implies that

$$\sum_{\substack{\phi_i \in \Phi_{\mathrm{sim}}(m_i) \\ \phi_{i,\infty} \in \Phi_{i,\infty}}} \dim_{U(m_i)}(K_i(\mathfrak{n}), \phi_i) \leq \sum_{\pi_\infty \in \Theta_{i,\infty}} m(\pi_\infty, \mathfrak{n}), \tag{4}$$

where $m(\pi_\infty, \mathfrak{n})$ denotes the multiplicity of π_∞ in $L^2_{\mathrm{disc}}(U(m_i, F) \backslash U(m_i, \mathbb{A}) / K_i(\mathfrak{n}))$. In fact it follows from the known cases of the Ramanujan conjecture that π_∞ is tempered, so π_∞ appears only in the cuspidal spectrum. Then a theorem of Savin [29] gives $m(\pi_\infty, \mathfrak{n}) \ll N\mathfrak{n}^{m_i^2}$ for all π_∞. Combining this with (3) and (4) gives a bound

$$h^d_{(2)}(X(\mathfrak{n}))_{\mathcal{S}} \ll N\mathfrak{n}^{\dim GL_N / U_{\mathcal{S}} + \epsilon}$$

where $U_{\mathcal{S}}$ is the unipotent radical of $P_{\mathcal{S}}$. Showing that $\dim GL_N / U_{\mathcal{S}} \leq Nd + 1$ completes the proof.

The role played by the cohomological degree in this argument is that $\dim U_{\mathcal{S}}$ must be large if d is small, because of the bound $n_1 \geq N - d$. However, it should be noted that the bound $\dim GL_N / U_{\mathcal{S}} \leq Nd + 1$ does not need to hold if $m_1 \leq 3$, and in these cases there are some additional steps one must take to optimize the argument to obtain the exponent $Nd + 1$. We will describe them in the course of the proof in the main body except for the following key input, which may be of independent interest. Namely we give in Lemma A.1 a uniform bound (which is significantly better than a trivial bound; see the remark below Lemma A.1) on the dimension of invariant vectors in supercuspidal representations of $GL(r)$ under principal congruence subgroups. By a uniform bound we mean a bound which is independent of the representation (and only depends on the residue field cardinality and the level of congruence subgroup). The asymptotic growth of the invariant dimension is fairly well understood if a representation is fixed but not otherwise. Analogous uniform bounds, on which our paper sheds some light, should be useful for bounding the growth of cohomology of other locally symmetric spaces.

In an earlier version of this paper, we obtained bounds for invariant vectors in supercuspidal representations of GL_2 and GL_3 (which are the only cases we need here) by a different method, which involved explicitly constructing the representations. The argument for GL_3 may be of independent interest, and we have included it in Appendix B. The argument for GL_2 is more routine, and may be found in an earlier version of this paper on the arXiv. It applied the construction described in [14, Section 7.A], which for a p-adic field L with $p \neq 2$, produces supercuspidals π for $GL_2(L)$ from a quadratic extension L'/L and a character χ of $L'^{\times}$. Moreover, all supercuspidals are obtained in this way. The construction realizes π on the Schwarz space $\mathcal{S}(L')$, and the formulas it provides for the action of $GL_2(L)$ on $\mathcal{S}(L')$ easily let one bound the fixed vectors in π. The bound we obtain is $\dim \pi^{K(n)} \leq q^n(1+1/q)$, where q is the order of the residue field of L and $K(n) < GL_2(L)$ is the principal congruence subgroup of depth n.

The proof of Theorem 1.2 in fact shows that $h^d_{(2)}(X(\mathfrak{n}))_{\mathcal{S}} \ll_\epsilon N\mathfrak{n}^{Nd+\epsilon}$, except when $\mathcal{S} = (N-d,1),(1,d)$, or in the exceptional case when $N=4$, $d=2$, and $\mathcal{S} = (2,2)$. Moreover, when $\mathcal{S} = (N-d,1),(1,d)$ we expect that the bound $h^d_{(2)}(X(\mathfrak{n}))_{\mathcal{S}} \ll_\epsilon N\mathfrak{n}^{Nd+1+\epsilon}$ is sharp when G arises from a Hermitian form, so that the majority of $H^d_{(2)}(X(\mathfrak{n}))$ comes from parameters of this shape. By [3, Theorem 10.1], these forms are theta lifted from a Hermitian space of dimension d, and it may therefore be possible to prove that Theorem 1.2 is sharp using the theta lift. Note that this is done in [12], but in a slightly different setting to Theorem 1.2. In particular, it is proved there that $h^d_{(2)}(X(\mathfrak{n})) \gg N\mathfrak{n}^{Nd+1}$ if G arises from a Hermitian form, $d < N/2$, and $\mathfrak{n}$ has the form $\mathfrak{c}\mathfrak{p}^k$, where $\mathfrak{c}$ is a fixed ideal of F that is sufficiently divisible, and $\mathfrak{p}$ is a fixed prime of F that is inert in E.

The reader may be curious as to how we can expect our bound to be sharp, when at a key point in the proof (Lemma 5.2) we seem to bound the dimension of the space of K-invariants in a quotient of an induced representation by the invariants in the whole induced representation. We remark that Lemma 5.2 is actually more efficient than this for certain $\mathcal{S}$. In particular, when $\mathcal{S} = (N-d,1),(1,d)$ (which should give the main contribution), the bound of Lemma 5.2 is sharp.

Finally we remark that it should be possible to adapt much of our arguments to unitary groups of other signatures, notwithstanding combinatorial complexity. However the bound is not going to be optimal as the case of $U(2,2)$ already shows [19]. We would need a sharper uniform bound than $\dim \pi^{K(n)} \ll q^{d(d-1)n/2}$ when π runs over non-generic representations of GL_d.

1.2. Acknowledgements. We would like to thank Nicolas Bergeron for helpful discussions. The first author was supported by the NSF under grant no. 1501230, and under grant no. 1440140 while in residence at MSRI in Berkeley, California, during the spring of 2017. He was also supported as the Neil Chriss and Natasha Herron Chriss Founders' Circle Member while at the IAS in 2017–18. The second author was supported by NSF grant no. 1501882 and a Sloan Fellowship.

2. Notation

Our notation and discussion in this section are based on [25] and [16]. (Similar summaries are given in [19] and [20] with more details in the quasi-split case.)

Let N be a positive integer. Write $GL(N)$ for the general linear group. Let F be a field of characteristic zero. Given a quadratic algebra E over F, we define $U(N) = U_{E/F}(N)$ to be the quasi-split unitary group in N variables, defined by an antidiagonal matrix J_N with $(-1)^{i-1}$ in the $(i, N+1-i)$ entry, as in [16, 0.2.2]. The compact special unitary group in two variables is denoted by $SU(2)$. Let $\nu(n)$ denote its n-dimensional irreducible representation (unique up to isomorphism).

Assume that F is a local or global field of characteristic zero. Write W_F for the Weil group of F. For any connected reductive group G over F, its Langlands dual group is denoted by $\widehat{G}$. Let ${}^LG = \widehat{G} \rtimes W_F$ denote the (Weil form of) L-group of G. Note that ${}^LGL(N) = GL(N, \mathbb{C}) \times W_F$ and that ${}^LU_{E/F}(N)$ may be explicitly described, cf. [16, 0.2.2].

Now assume that F is local. Define the local Langlands group $L_F := W_F$ if F is archimedean and $L_F := W_F \times SU(2)$ otherwise. An A-parameter is a continuous homomorphism $\psi : L_F \times SL(2, \mathbb{C}) \to {}^LG$ commuting with the projection maps onto W_F such that $\psi(L_F)$ has relatively compact image in $\widehat{G}$ and that ψ restricted to $SL(2, \mathbb{C})$ is a map of $\mathbb{C}$-algebraic groups into $\widehat{G}$. Two parameters are considered isomorphic if they are conjugate under $\widehat{G}$. Write $\Psi(G)$ or $\Psi(G, F)$ for the set of isomorphism classes of A-parameters. Define $\Psi^+(G)$ analogously without the condition on relatively compact image. Define $s_\psi := \psi(1, -1)$ for any $\psi \in \Psi^+(G)$.

An L-parameter is $\psi^+ \in \Psi^+(G)$ which is trivial on the $SL(2, \mathbb{C})$-factor (external to L_F). The subset of L-parameters (up to isomorphism) is denoted by $\Phi(G)$. Any $\psi \in \Psi^+(G)$ gives rise to an L-parameter ϕ_ψ by pulling back via the map $L_F \to L_F \times SL(2, \mathbb{C})$, $w \mapsto \left(w, \begin{pmatrix} |w|^{1/2} & 0 \\ 0 & |w|^{-1/2} \end{pmatrix}\right)$.

When $G = GL(N)$, we associate representations π_ψ and ρ_ψ of $\mathrm{GL}(N, F)$ to $\psi \in \Psi^+(G)$ by the following recipe from [16, 1.2.2]. We may decompose $\psi = \oplus_{i=1}^k \psi_i$ with $\psi_i = \phi_i \boxtimes \nu(n_i)$ such that $\phi_i : L_F \to {}^LGL(m_i)$ is an irreducible m_i-dimensional representation of L_F and $\sum_{i=1}^k m_i n_i = N$. The local Langlands correspondence associates an irreducible essentially square-integrable representation π_{ϕ_i} of $GL(m_i, F)$ to ϕ_i. Let $|\det(m)|$ denote the composition of the absolute value on $F^\times$ with the determinant map on $GL(m, F)$. Then consider the multi-set of representations

$$\{\pi_{\phi_i}|\det(m_i)|^{\frac{n_i-1}{2}}, \pi_{\phi_i}|\det(m_i)|^{\frac{n_i-3}{2}}, ..., \pi_{\phi_i}|\det(m_i)|^{\frac{1-n_i}{2}}\}_{i=1}^k. \tag{5}$$

This defines a representation of $\prod_{i=1}^k GL(m_i)^{n_i}$ viewed as a block diagonal Levi subgroup of $GL(N)$. Let ρ_ψ denote the parabolically induced representation. (The choice of parabolic subgroup does not affect our argument; we will choose the upper triangular one.) The Langlands quotient construction singles out an irreducible subquotient π_{ψ_i} of the representation of $GL(m_i n_i, F)$ induced from

(5) on the Levi subgroup $GL(m_i, F)^{n_i}$. We write π_ψ for the representation of $GL(N)$ parabolically induced from $\boxtimes_{i=1}^{k}\pi_{\psi_i}$. (In the case of interest, one actually knows that π_{ψ_i} is unitary by Lemma 6.1. Thus π_ψ is the irreducible representation corresponding to ψ by the A-packet parametrization.)

As in the introduction, from here throughout the paper, we fix a totally real field F and a totally complex quadratic extension E over F with complex conjugation c in $\mathrm{Gal}(E/F)$. The ring of adeles over F (resp. E) is denoted by $\mathbb{A}$ (resp. $\mathbb{A}_E$). We often write G^* for $U(N)$ and $G(N)$ for $\mathrm{Res}_{E/F}GL(N)$. The group $G(N)$ is equipped with involution $\theta : g \mapsto J_N{}^t c(g)^{-1} J_N^{-1}$, giving rise to the twisted group $\widetilde{G}^+(N) = G(N) \rtimes \{1, \theta\}$. Write $\widetilde{G}(N)$ for the coset $G(N) \rtimes \theta$. Let v be a place of F. Given any algebraic group H over F, we often write H_v for $H(F_v)$ or $H \otimes_F F_v$ (the context will make it clear which one we mean).

For $n \in \mathbb{Z}_{\geq 1}$, define $\widetilde{\Phi}_{\mathrm{sim}}(n)$ (a shorthand for $\Phi_{\mathrm{sim}}(\widetilde{G}(n))$) to be the set of conjugate self-dual cuspidal automorphic representations of $GL(n, \mathbb{A}_E)$. Here a representation π is considered conjugate self-dual if $\pi \circ c$ is isomorphic to the contragredient of π, or equivalently if $\pi \circ \theta$ is isomorphic to π. Fix two Hecke characters $\chi_\kappa : \mathbb{A}_E^\times / E^\times \to \mathbb{C}^\times$ with $\kappa \in \{\pm 1\}$ as follows: χ_+ is the trivial character while χ_- is an extension of the quadratic character of $\mathbb{A}^\times / F^\times$ associated to E/F by class field theory. We use χ_κ to define two base-change L-morphisms

$$\eta_{\chi_\kappa} : {}^L U(N) \to {}^L G(N), \quad \kappa \in \{\pm 1\},$$

as follows. Choose $w_c \in W_F \backslash W_E$ so that $W_F = W_E \coprod W_E w_c$. Under the identification $\widehat{U}(N) = GL(N, \mathbb{C})$ and $\widehat{G}(N) = GL(N, \mathbb{C}) \times GL(N, \mathbb{C})$, we have (where scalars stand for scalar $N \times N$-matrices whenever appropriate)

$$\begin{aligned} \eta_{\chi_\kappa}(g \rtimes 1) &= (g, J_N{}^t g^{-1} J_N^{-1}) \rtimes 1, \\ \eta_{\chi_\kappa}(1 \rtimes w) &= (\chi_\kappa(w), \chi_\kappa^{-1}(w)) \rtimes w, \quad w \in W_E, \\ \eta_{\chi_\kappa}(1 \rtimes w_c) &= (1, \kappa) \rtimes w_c. \end{aligned}$$

Let us define $\widetilde{\Psi}_{\mathrm{ell}}(N)$, the set of (formal) elliptic parameters for $\widetilde{G}(N)$. Such a parameter is represented by a formal sum $\psi = \boxplus_{i=1}^{k} \psi_i$ with $\psi_i = \mu_i \boxtimes \nu(n_i)$ such that the pairs (μ_i, n_i) are mutually distinct, where $\mu_i \in \widetilde{\Phi}_{\mathrm{sim}}(m_i)$, $\sum_{i=1}^{k} m_i n_i = N$. (Two formal sums are identified under permutation of indices.)

Mok defines the sets $\Psi_2(U(N), \eta_{\chi_\kappa})$ for $\kappa \in \{\pm 1\}$. (In [25], he writes ξ_{χ_κ} for η_{χ_κ}.) They are identified (via the map $(\psi^N, \widetilde{\psi}) \mapsto \psi^N$ of [25, Section 2.4]) with disjoint subsets of $\widetilde{\Psi}_{\mathrm{ell}}(N)$, corresponding to the two ways $U(N)$ can be viewed as a twisted endoscopic group of $\widetilde{G}(N)$ via η_{χ_κ}, characterized by a sign condition. We don't need to recall the sign condition here. It suffices to know that each $\psi \in \Psi_2(U(N), \eta_{\chi_\kappa})$ admits localizations to $\Psi^+(U(N)_v)$; see below. We write ψ^N for ψ when ψ is viewed as a member of $\widetilde{\Psi}_{\mathrm{ell}}(N)$.

A parameter ψ in $\Psi_2(U(N), \eta_{\chi_\kappa})$ is said to be generic if $n_i = 1$ for all $1 \leq i \leq k$ and simple if $k = 1$. Write $\Phi_{\mathrm{sim}}(U(N), \eta_{\chi_\kappa})$ for the subset of simple generic parameters. Theorem 2.4.2 of [25] shows that $\widetilde{\Phi}_{\mathrm{sim}}(N)$ is partitioned into $\Phi_{\mathrm{sim}}(U(N), \eta_{\chi_\kappa})$, $\kappa \in \{\pm 1\}$.

To a parameter $\psi \in \Psi_2(U(N), \eta_{\chi_\kappa})$ is associated localizations $\psi_v \in \Psi^+(G^*_v)$ such that ψ_v is carried to $\oplus_{i=1}^k \phi_{\mu_{i,v}} \boxtimes \nu(n_i)$ via the L-morphism $\eta_{\chi_\kappa} : {}^L U(N) \to {}^L G(N)$, where $\phi_{\mu_{i,v}}$ is the L-parameter for $\mu_{i,v}$ (via local Langlands for $GL(m_i)$). For each place v of F split in E, fix a place w of E above v. Then we have an isomorphism $G^*_v \simeq GL(N, E_w)$.

At every finite place v of F where G^*_v is unramified, fix hyperspecial subgroups $\widetilde{K}_v = GL(O_{F_v} \otimes_{O_F} O_E)$ of $G(N, F_v)$ and K^*_v of $G^*(F_v)$ (such that they come from global integral models away from finitely many v). When v is split as w and $c(w)$ in E we have a decomposition $\widetilde{K}_v = \widetilde{K}_w \times \widetilde{K}_{c(w)}$, and we may identify K^*_v with $\widetilde{K}_w$ via $G^*_v \simeq GL(N, E_w)$.

Finally let G be an inner form of G^* over F. It can always be promoted to an extended pure inner twist $(\xi, z) : G^* \to G$, [16, 0.3.3]. Let S be a set of places of F such that both G_v and G^*_v are unramified for every $v \notin S$. Then fix an isomorphism $G^*_v \simeq G_v$, which is $G(\overline{F})$-conjugate to (ξ, z). We have a hyperspecial subgroup $K_v \subset G_v$ by transferring K^*_v. So if $v \notin S$ is split in E then K_v and K^*_v are identified with $GL(N, O_{E,w})$ under the isomorphisms $G_v \simeq G^*_v \simeq GL(N, E_w)$.

Let $\psi_v \in \Psi(U(N)_v)$ for a place v of F. This gives rise to a distribution $f \mapsto f(\psi_v)$ on the space of smooth compactly supported functions on $U(N)_v$ [25, Theorem 3.2.1].

Given a connected reductive group H over F_v, a smooth compactly supported function f on $H(F_v)$, and an admissible representation π of $H(F_v)$, we write $\mathrm{tr}(\pi(f))$ or $f(\pi)$ for the trace value. Occasionally we also consider a twisted variant when $\tilde{\pi}$ is an admissible representation of $G^+(N, F_v)$ and $\tilde{f}$ is a smooth compactly supported function on $G(N, F_v) \rtimes \theta$. Then $\mathrm{tr}(\tilde{\pi}(f))$ will denote the (twisted) trace.

3. Cohomological representations of $U(N-1,1)$

In this section, we recall some facts about the cohomological representations of the real Lie group $U(N-1,1)$, which will imply that any global Arthur parameter that contributes to $h^d_{(2)}(X(\mathfrak{n}))$ must have a factor $\mu \boxtimes \nu(n)$ with $n \geq N-d$ by applying results of Bergeron, Millson, and Moeglin. Let $\mathfrak{g}_0$ be the real Lie algebra of $U(N-1,1)$, and K a maximal compact subgroup. Write $\mathfrak{g}$ for the complexification of $\mathfrak{g}_0$. Similarly the complexification of real Lie algebras $\mathfrak{k}_0$, $\mathfrak{p}_0$, etc will be denoted by $\mathfrak{k}$, $\mathfrak{p}$, etc below. The facts we shall need on the cohomological representations of $U(N-1,1)$ are summarized in the following proposition; recall that $p(\pi)$ is the infimum over p for which the K-finite matrix coefficients of π lie in $L^p(G)$.

Proposition 3.1. *Let a, b be a pair of integers with $a, b \geq 0$ and $a+b \leq N-1$, and let $d = a + b$. There is an irreducible unitary representation $\pi_{a,b}$ of $U(N-1,1)$ with the following properties.*

(i) We have

$$H^{p,q}(\mathfrak{g}, K; \pi_{a,b}) = \begin{cases} \mathbb{C} & \text{if } (p,q) = (a,b) + (k,k), \quad 0 \leq k \leq N-1-a-b, \\ 0 & \text{otherwise.} \end{cases}$$

(ii) Suppose that $d \leq N-2$. If $\varphi : \mathbb{C}^\times \to GL(N,\mathbb{C})$ is the restriction of the Langlands parameter of $\pi_{a,b}$ to $\mathbb{C}^\times$, then we have

$$\varphi(z) = (z/\overline{z})^{(b-a)/2}|z|^{N-d-1} \oplus (z/\overline{z})^{(b-a)/2}|z|^{-N+d+1} \oplus \bigoplus_{\substack{-N+1\leq j\leq N-1 \\ j\equiv N-1\ (2) \\ j\neq N-1-2a,-N+1+2b}} (z/\overline{z})^{j/2}.$$

(iii) We have $p(\pi_{a,b}) = 2(N-1)/d$.

Moreover, the $\pi_{a,b}$ are the only irreducible unitary representations of $U(N-1,1)$ with $H^*(\mathfrak{g}, K; \pi) \neq 0$.

3.1. The classification of Vogan and Zuckerman. We let $G = U(N-1,1)$, and realize G as the subgroup of $GL(N,\mathbb{C})$ preserving the Hermitian form $|z_1|^2 + \ldots + |z_{N-1}|^2 - |z_N|^2$. The Lie algebra $\mathfrak{g}_0$ of G is

$$\mathfrak{g}_0 = \{A \in M_N(\mathbb{C}) : {}^t\overline{A} = -I_{N-1,1}AI_{N-1,1}\}$$

where

$$I_{N-1,1} = \begin{pmatrix} I_{N-1} & \\ & -1 \end{pmatrix}.$$

The algebras $\mathfrak{k}_0$ and $\mathfrak{p}_0$ in the Cartan decomposition $\mathfrak{g}_0 = \mathfrak{k}_0 \oplus \mathfrak{p}_0$ are

$$\mathfrak{k}_0 = \left\{ \begin{pmatrix} A & 0 \\ 0 & i\theta \end{pmatrix} : {}^t\overline{A} = -A, \theta \in \mathbb{R} \right\}, \quad \mathfrak{p}_0 = \left\{ \begin{pmatrix} 0 & z \\ {}^t\overline{z} & 0 \end{pmatrix} : z \in M_{N-1,1}(\mathbb{C}) \right\}.$$

Let $\mathfrak{t}_0$ denote the Cartan subalgebra of $\mathfrak{k}_0$ consisting of diagonal matrices. The adjoint action of K on $\mathfrak{p}_0$ preserves the natural complex structure, and so we have a decomposition $\mathfrak{p} = \mathfrak{p}_+ \oplus \mathfrak{p}_-$ of K-modules. We may naturally identify $\mathfrak{g}$ with $M_n(\mathbb{C})$, and under this identification we have

$$\mathfrak{p}_+ = \left\{ \begin{pmatrix} 0 & z \\ 0 & 0 \end{pmatrix} : z \in M_{N-1,1}(\mathbb{C}) \right\}, \quad \mathfrak{p}_- = \left\{ \begin{pmatrix} 0 & 0 \\ z & 0 \end{pmatrix} : z \in M_{1,N-1}(\mathbb{C}) \right\}.$$

If τ_d is the representation of K on $\bigwedge^d \mathfrak{p}$, it is well known [6, VI 4.8-9] that there is a decomposition

$$\tau_d = \oplus_{a+b=d} \tau_{a,b}, \tag{6}$$

where $\tau_{a,b}$ is the representation of K on $\bigwedge^a \mathfrak{p}_- \otimes \bigwedge^b \mathfrak{p}_+$. Moreover, we have

$$\tau_{a,b} = \oplus_{k=0}^{\min(a,b)} \tau'_{a-k,b-k} \tag{7}$$

for $a + b \leq N - 1$, where the representations $\tau'_{a,b}$ are irreducible with highest weight

$$\sum_{i=1}^{b} \varepsilon_i - \sum_{i=N-a}^{N-1} \varepsilon_i + (a-b)\varepsilon_N. \tag{8}$$

Here, $\{\varepsilon_i\}$ is the standard basis for $\mathfrak{t}^*$ consisting of elements that are real on $i\mathfrak{t}_0$. These decompositions correspond to the Hodge-Lefschetz decomposition for the cohomology of $X(\mathfrak{n})$.

We now recall the classification of cohomological representations of G due to Vogan and Zuckerman [31]. We choose an element $H \in i\mathfrak{t}_0$, so that $\mathrm{ad}(H)$ has real eigenvalues. We let $\mathfrak{q} \subset \mathfrak{g}$ be the parabolic subalgebra $\mathfrak{l} + \mathfrak{u}$, where $\mathfrak{l} = Z_{\mathfrak{g}}(H)$ and $\mathfrak{u}$ is the sum of all the eigenspaces for $\mathrm{ad}(H)$ with positive eigenvalues. Because $\mathfrak{k}$ and $\mathfrak{p}_\pm$ are stable under $\mathrm{ad}(H)$, we have $\mathfrak{u} = \mathfrak{u} \cap \mathfrak{k} + \mathfrak{u} \cap \mathfrak{p}_- + \mathfrak{u} \cap \mathfrak{p}_+$. We define $R_\pm = \dim(\mathfrak{u} \cap \mathfrak{p}_\pm)$ and $R = R_+ + R_-$, and let $\mu = 2\rho(\mathfrak{u} \cap \mathfrak{p})$, which is the sum of the roots of $\mathfrak{t}$ in $\mathfrak{u} \cap \mathfrak{p}$. We fix a set of positive roots for $\mathfrak{t}$ in $\mathfrak{l} \cap \mathfrak{k}$ so that a positive root system for $\mathfrak{t}$ in $\mathfrak{k}$ is determined (together with $\mathfrak{u} \cap \mathfrak{k}$). Then μ is a highest weight for the positive root system.

The main theorem of Vogan and Zuckerman is that there is a unique irreducible unitary representation $A_{\mathfrak{q}}$ of G [2] with the following properties:

- $A_{\mathfrak{q}}$ has the same infinitesimal character as the trivial representation.
- $A_{\mathfrak{q}}$ contains the K-type with highest weight μ.

They also show that any irreducible unitary representation of G with nonzero $(\mathfrak{g}, K)$-cohomology (with trivial coefficients) must be of the form $A_{\mathfrak{q}}$ for some $\mathfrak{q}$. It is clear that $A_{\mathfrak{q}}$ only depends on $\mathfrak{u} \cap \mathfrak{p}$. Moreover, we have [31, Prop 6.19]

$$H^{R_+ + p, R_- + p}(\mathfrak{g}, K; A_{\mathfrak{q}}) \simeq \mathrm{Hom}_{\mathfrak{l} \cap \mathfrak{k}}(\wedge^{2p}(\mathfrak{l} \cap \mathfrak{p}), \mathbb{C}), \quad p \geq 0, \tag{9}$$

and

$$H^{p,q}(\mathfrak{g}, K; A_{\mathfrak{q}}) = 0 \tag{10}$$

[2] Note that the general unitarity of the representations $A_{\mathfrak{q}}$ is proved in [30].

for other (p,q), i.e. if $p-q \neq R_+ - R_-$.

Write $H_1, \dots, H_N$ for the entries of the real diagonal matrix H. Because $A_{\mathfrak{q}}$ only depends on the orbit of H under the Weyl group of K, we may assume that $H_1 \geq \cdots \geq H_{N-1}$. The subspace $\mathfrak{u} \cap \mathfrak{p}$, and hence $A_{\mathfrak{q}}$, only depends on the number of $H_i - H_N$ that are positive, negative, and zero. Therefore, if a and b are the number of $H_i - H_N$ that are positive and negative respectively, then we have $H_{a+1} = \cdots = H_{N-1-b} = H_N$, while we may assume that all the remaining H_i are distinct. It may be seen that $R_+ = a$ and $R_- = b$, and

$$\mu = \sum_{i=1}^{a} \varepsilon_i - \sum_{i=N-b}^{N-1} \varepsilon_i - (a-b)\varepsilon_N.$$

The representation $A_{\mathfrak{q}}$ depends only on a and b, and we denote it by $\pi_{a,b}$.

To prove (iii), we will need the description of $\pi_{a,b}$ as a Langlands quotient when $a+b < N-1$, which is given by Vogan and Zuckerman in [31, Theorem 6.16]. Define

$$V = \begin{pmatrix} & & 1 \\ & 0_{N-2} & \\ 1 & & \end{pmatrix},$$

and let $\mathfrak{a}_0 = \mathbb{R}V$ so that $\mathfrak{a}_0$ is a maximal abelian subalgebra of $\mathfrak{p}_0$. Let $A = \exp(\mathfrak{a}_0)$ be the corresponding subgroup. Define $\alpha \in \mathfrak{a}^*$ by $\alpha(V) = 1$. The roots of $\mathfrak{a}$ in $\mathfrak{g}$ are $\pm\alpha$ and $\pm 2\alpha$ with multiplicities $2(N-2)$ and 1 respectively, so that $\rho = (N-1)\alpha$. Let U be the unipotent subgroup corresponding to the positive roots. Let $M = Z_K(V)$, so that

$$M = \left\{ \begin{pmatrix} e^{i\theta} & & \\ & X & \\ & & e^{i\theta} \end{pmatrix} : X \in U(N-2), \theta \in \mathbb{R} \right\}.$$

Let $\mathfrak{t}_M \subset \mathfrak{t}$ be the diagonal Cartan subalgebra in $\mathfrak{m}$. Let σ be the irreducible representation of M with highest weight given by the restriction to $\mathfrak{t}_M$ of

$$\sum_{i=2}^{a+1} \varepsilon_i - \sum_{i=N-b}^{N-1} \varepsilon_i + (b-a)\varepsilon_1.$$

Let $\nu = (N-1-d)\alpha$. We define $I_{\nu,\sigma}$ to be the unitarily normalized induction from $P = MAU$ to G of the representation $\sigma \otimes e^{\nu} \otimes 1$. Then $\pi_{a,b}$ is the Langlands quotient of $I_{\nu,\sigma}$.

3.2. Proof of Proposition 3.1 . The assertion that $\pi_{a,b}$ are the only representations with nonzero cohomology is clear, because any such representation is isomorphic to $A_{\mathfrak{q}}$ for some $\mathfrak{q}$. The calculation of $H^{p,q}(\mathfrak{g}, K; \pi_{a,b})$ in condition (i) follows from (9) and (10) after we compute $\mathrm{Hom}_{\mathfrak{l}\cap\mathfrak{k}}(\wedge^{2p}(\mathfrak{l}\cap\mathfrak{p}), \mathbb{C})$. Our assumption on H implies that $\mathfrak{l}_0 \simeq \mathfrak{u}(N-d-1, 1)\times\mathfrak{u}(1)^d$, and $\mathfrak{l} = \mathfrak{l}\cap\mathfrak{k}\oplus\mathfrak{l}\cap\mathfrak{p}$ is the standard Cartan decomposition of $\mathfrak{l}$. We wish to show that the trivial representation of $\mathfrak{l}\cap\mathfrak{k}$ occurs exactly once in $\wedge^{2p}(\mathfrak{l}\cap\mathfrak{p})$ for all $0 \leq p \leq N-d-1$, but this follows from the decompositions (6) and (7) for $\mathfrak{u}(N-d-1, 1)$, combined with the fact that $\tau'_{p,q}$ is trivial if and only if $p = q = 0$ as one sees from the highest weight formula (8).

The description of the Langlands parameter of $\pi_{a,b}$ in (ii) follows from [2, Section 5.3].

To prove assertion (iii), we may assume that $a + b < N - 1$ as otherwise $\pi_{a,b}$ lies in the discrete series. When $a + b < N - 1$, the assertion follows from our description of $\pi_{a,b}$ as a Langlands quotient, and well-known asymptotics for matrix coefficients, which we recall from Knapp [17]. Let $\overline{P} = MA\overline{U}$ be the opposite parabolic to P, and let $\overline{I}_{\sigma,\nu}$ be the normalized induction of $\sigma\otimes e^{\nu}\otimes 1$ from $\overline{P}$ to G. Let $A(\sigma,\nu) : I_{\sigma,\nu} \to \overline{I}_{\sigma,\nu}$ be the intertwiner

$$A(\sigma,\nu)f(g) = \int_{\overline{U}} f(ug)du,$$

which converges by [17, VII, Prop 7.8]. Then the image of $A(\sigma,\nu)$ is isomorphic to the Langlands quotient $\pi_{a,b}$ of $I_{\sigma,\nu}$. We introduce the pairing on $I_{\sigma,\nu}$ given by

$$\langle f, g\rangle = \int_K \langle f(k), g(k)\rangle_\sigma dk$$

where $\langle\cdot,\cdot\rangle_\sigma$ denotes a choice of inner product on σ. If we choose $g \in I_{\sigma,\nu}$ to pair trivially with the kernel of $A(\sigma,\nu)$, then $\langle I_{\sigma,\nu}(\cdot)f, g\rangle$ is a matrix coefficient of $\pi_{a,b}$, and all coefficients are realized in this way. The asymptotic behaviour of the coefficients is given by [17, VII, Lemma 7.23], which states that

$$\lim_{a\to\infty} e^{(\rho-\nu)\log a}\langle I_{\sigma,\nu}(a)f, g\rangle = \langle A(\sigma,\nu)f(1), g(1)\rangle_\sigma. \tag{11}$$

As $\nu = (N-d-1)\alpha$, [17, VIII, Theorem 8.48] implies that $p(\pi_{a,b}) \leq 2(N-1)/d$. It also follows from that theorem that to prove $p(\pi_{a,b}) = 2(N-1)/d$, we need only show that the right hand side of (11) is nonzero for some choice of f and g, subject to the condition that g pairs trivially with $\ker A(\sigma,\nu)$. To do this, choose $f \in I_{\sigma,\nu}$ such that $A(\sigma,\nu)f \neq 0$, and some nonzero g of the required type. Because $A(\sigma,\nu)$ is an intertwiner, after translating f by K we may assume that $A(\sigma,\nu)f(1) \neq 0$. Because $\ker A(\sigma,\nu)$ is an invariant subspace, we may likewise assume that $g(1) \neq 0$. Because σ was irreducible, translating by M we may also assume that $\langle A(\sigma,\nu)f(1), g(1)\rangle_\sigma \neq 0$ as required.

4. Application of the global classification

As in the notation section, $(\xi, z) : G^* \to G$ is an extended pure inner twist of the quasi-split unitary group $G^* = U(N)$ over F. We always assume that G_{v_0} is isomorphic to $U(N-1,1)$ at a real place v_0 of F and that G_v is compact at all other real places v. Although much of our argument works for general inner forms, the assumption significantly simplifies some combinatorial and representation-theoretic arguments (especially of Section 3) and ensures that we obtain expectedly optimal upper bounds in all degrees in the main theorem.

Let $\psi \in \Psi_2(G^*, \eta_{\chi_\kappa})$. The main local theorem of [16] defines local packets $\Pi_{\psi_v}(G,\xi)$ consisting of finitely many (possibly reducible and non-unitary[3]) representations of G_v such that $\Pi_{\psi_v}(G,\xi)$ contains an unramified representation (relative to K_v) at all but finitely many v. The global packet $\Pi_\psi(G,\xi)$ consists of restricted tensor products $\pi = \otimes'_v \pi_v$ with $\pi_v \in \Pi_{\psi_v}(G,\xi)$. The parameter ψ determines a sign character ϵ_ψ on a certain centralizer group (in $\widehat{G}$) attached to ψ, and [16] defines a subset $\Pi_\psi(G,\xi,\epsilon_\psi)$ of $\Pi_\psi(G,\xi)$ by imposing a sign condition. We need not recall the condition as it will be soon ignored along the way to an upper bound. Theorem 1.7.1 of [16] asserts the following.

Theorem 4.1. *There is a $G(\mathbb{A})$-module isomorphism*

$$L^2_{\mathrm{disc}}(G(F)\backslash G(\mathbb{A})) \simeq \bigoplus_{\psi\in\Psi_2(G^*,\eta_{\chi_\kappa})} \bigoplus_{\pi\in\Pi_\psi(G,\xi,\epsilon_\psi)} \pi.$$

Let S be a finite set of finite places of F containing all places at which E or G ramify. Let $\mathfrak{n} \subset \mathcal{O}_F$ be a nonzero ideal whose prime factors are split in E and don't lie in S. In Section 2 we have introduced hyperspecial subgroups K_v of $G(F_v)$ when $v \notin S$. For $v \in S$ let K_v be an arbitrary open compact subgroup of $G(F_v)$. Now we define the congruence subgroup $K(\mathfrak{n}) = \prod_v K(\mathfrak{n})_v$, where $K(\mathfrak{n})_v$ is given as follows for each finite place v. Define $K(\mathfrak{n})_v$ to be K_v if v does not divide $\mathfrak{n}$. If $v|\mathfrak{n}$ then we have fixed an isomorphism $K_v \simeq GL(N, \mathcal{O}_{E_w})$, and $K(\mathfrak{n})_v$ is the subgroup of K_v consisting of elements congruent to the identity modulo $\mathfrak{n}$. Let K_∞ denote a maximal compact subgroup of $G(F \otimes_{\mathbb{Q}} \mathbb{R})$. Often we write $[G]$ for the quotient $G(F)\backslash G(\mathbb{A})$, and likewise when G is replaced with quasi-split unitary groups.

We would like to investigate the cohomology of the arithmetic manifold

$$X(\mathfrak{n}) = G(F)\backslash G(\mathbb{A})/K(\mathfrak{n})K_\infty. \tag{12}$$

[3] The issue is that $\psi_v \in \Psi^+(G_v^*)$ is not known to be in $\Psi(G_v^*)$ in general although it is expected. However this is actually known for parameters contributing to cohomology from the known cases of the Ramanujan conjecture, see Section 6. It follows that all representations in the local packets we will consider are irreducible and unitary.

Since $G(F \otimes_{\mathbb{Q}} \mathbb{R})/K_\infty$ is isomorphic to the symmetric space $U(N-1,1)/(U(N-1) \times U(1))$, which has complex dimension $N-1$, we see that the complex dimension of $X(\mathfrak{n})$ is also $N-1$.

We take the first step in proving Theorem 1.2 on bounding the L^2-Lefschetz numbers $h^d_{(2)}(X(\mathfrak{n}))$ in degrees $0 \le d < N-1$, as $\mathfrak{n}$ varies. Write $h^d(\mathfrak{g}, K_\infty; \pi_\infty)$ for $\dim H^d(\mathfrak{g}, K_\infty; \pi_\infty)$. Matsushima's formula (see [6] in the noncompact case) gives

$$h^d_{(2)}(X(\mathfrak{n})) = \sum_{\pi \subset L^2_{\mathrm{disc}}([G])} m(\pi) h^d(\mathfrak{g}, K_\infty; \pi_\infty) \dim \pi_f^{K(\mathfrak{n})},$$

where the sum runs over irreducible $G(\mathbb{A})$-subrepresentations π of $L^2_{\mathrm{disc}}([G])$ up to isomorphism, and $m(\pi) := \dim_{G(\mathbb{A})}(\pi, L^2_{\mathrm{disc}}([G]))$ denotes the multiplicity of π.

Combining this with Theorem 4.1 gives

$$h^d_{(2)}(X(\mathfrak{n})) \le \sum_{\psi \in \Psi_2(G^*, \eta_{\chi_\kappa})} \sum_{\pi \in \Pi_\psi(G,\xi)} h^d(\mathfrak{g}, K_\infty; \pi_\infty) \dim \pi_f^{K(\mathfrak{n})}. \tag{13}$$

5. Bounding the contribution of a single parameter

In this section, we bound the contribution of a single parameter ψ to the right hand side of (13). The form of our bound will depend on the shape of ψ, and so throughout this section we shall fix a shape $\mathcal{S} = (n_1, m_1), \dots, (n_k, m_k)$ and define $\Psi_2(G^*, \eta_{\chi_\kappa})_{\mathcal{S}}$ to be the set of parameters with that shape. If $\psi \in \Psi_2(G^*, \eta_{\chi_\kappa})_{\mathcal{S}}$, we define $\mu_i \in \widetilde{\Phi}_{\mathrm{sim}}(m_i)$ to be such that $\psi^N = \boxplus_{i \ge 1} \mu_i \boxtimes \nu(n_i)$. Each μ_i represents a simple generic parameter $\phi_i \in \Phi_{\mathrm{sim}}(U(m_i), \eta_{\chi_{\kappa_i}})$ for a unique sign $\kappa_i \in \{\pm 1\}$ determined as in [25, (2.4.8)]. We define

$$\begin{aligned} \tau(\mathcal{S}) &= \binom{N}{2} - \sum_{i \ge 1} n_i \binom{m_i}{2}, \qquad (14) \\ \tau_1(\mathcal{S}) &= \binom{N}{2} - \binom{n_1}{2} - \sum_{i \ge 2} n_i \binom{m_i}{2}, \\ \tau_2(\mathcal{S}) &= \tau(\mathcal{S}) + (n_1 - 1) + \epsilon, \\ \tau_3(\mathcal{S}) &= \tau(\mathcal{S}) + 3(n_1 - 1) + \epsilon. \end{aligned}$$

Here, $\epsilon > 0$ is an arbitrarily small constant that may vary from line to line. Any implied constants in bounds for quantities containing $\tau_2(\mathcal{S})$ or $\tau_3(\mathcal{S})$ will be assumed to depend on ϵ. We also define

$$\sigma(\mathcal{S}) = \sigma_2(\mathcal{S}) = \sigma_3(\mathcal{S}) = \sum_{i=1}^k n_i - 1, \quad \sigma_1(\mathcal{S}) = \sum_{i=2}^k n_i.$$

For each $1 \le i \le k$ and finite place v, define a compact open subgroup $K_{i,v}$ of $U(m_i)_v$ as follows. If $v \notin S$, then $K_{i,v}$ is the standard hyperspecial subgroup, and if $v \in S$ then $K_{i,v}$ will be chosen during the proof of Proposition 5.1. Let $K_i = \prod_v K_{i,v}$, and let $K_i(\mathfrak{n})$ be the principal congruence subgroup of K_i of level $\mathfrak{n}$. Let $P \subset GL(N)$ be the standard parabolic subgroup with Levi $\prod_{i=1}^k GL(m_i)^{n_i}$.

Proposition 5.1. *There is a choice of $K_{i,v}$ for $v \in S$ with the following property. Let $\psi \in \Psi_2(G^*, \eta_{\chi_\kappa})_{\mathcal{S}}$, and assume that ϕ_i (arising from ψ as above) is bounded everywhere for each $1 \le i \le k$. Then*

$$\sum_{\pi \in \Pi_\psi(G,\xi)} \dim \pi_f^{K(\mathfrak{n})} \ll \prod_{v|\mathfrak{n}} (1 + 1/q_v)^{\sigma(\mathcal{S})} N\mathfrak{n}^{\tau(\mathcal{S})} \prod_{i \ge 1} \left(\sum_{\pi_i \in \Pi_{\phi_i}(U(m_i))} \dim \pi_{i,f}^{K_i(\mathfrak{n})} \right)^{n_i}. \quad (15)$$

Moreover, if $m_1 = l$ with $l = 1,2,3$, we have

$$\sum_{\pi \in \Pi_\psi(G,\xi)} \dim \pi_f^{K(\mathfrak{n})} \ll \prod_{v|\mathfrak{n}} (1 + 1/q_v)^{\sigma_l(\mathcal{S})} N\mathfrak{n}^{\tau_l(\mathcal{S})}$$
$$\sum_{\pi_1 \in \Pi_{\phi_1}(U(m_1))} \dim \pi_{1,f}^{K_1(\mathfrak{n})} \prod_{i \ge 2} \left(\sum_{\pi_i \in \Pi_{\phi_i}(U(m_i))} \dim \pi_{i,f}^{K_i(\mathfrak{n})} \right)^{n_i}.$$

The first step in proving Proposition 5.1 is to write both sides as a product over the finite places. We describe this in the case of the first inequality, as the second is similar. We have

$$\sum_{\pi \in \Pi_\psi(G,\xi)} \dim \pi_f^{K(\mathfrak{n})} = \prod_{v \nmid \infty} \sum_{\pi_v \in \Pi_{\psi_v}(G_v,\xi_v)} \dim \pi_v^{K_v(\mathfrak{n})} \quad (16)$$

and

$$\prod_{i \ge 1} \left(\sum_{\pi_i \in \Pi_{\phi_i}(U(m_i))} \dim \pi_{i,f}^{K_i(\mathfrak{n})} \right)^{n_i} = \prod_{v \nmid \infty} \prod_{i \ge 1} \left(\sum_{\pi_{i,v} \in \Pi_{\phi_{i,v}}(U(m_i))} \dim \pi_{i,v}^{K_{i,v}(\mathfrak{n})} \right)^{n_i}.$$

It therefore suffices to prove that

$$\sum_{\pi_v \in \Pi_{\psi_v}(G_v,\xi_v)} \dim \pi_v^{K_v(\mathfrak{n})} \leq C_v(1+O(q_v^{-2})) \prod_{i \geq 1} \left(\sum_{\pi_{i,v} \in \Pi_{\phi_{i,v}}(U(m_i))} \dim \pi_{i,v}^{K_{i,v}(\mathfrak{n})} \right)^{n_i} \tag{17}$$

for all finite v. Here, the constant C_v may be arbitrary for $v \in S$, while for $v \notin S$ it is either 1 if v is inert in E/F, or the v-component of the constant in (15) if v is split. It is important to keep C_v independent of ψ_v and $\mathfrak{n}$ (but it could depend on K_v) for the application to the proof of the main theorem. We divide the proof of (17) into four cases, depending on whether v is split in E, and whether $v \in S$. Thus the proof of Proposition 5.1 will be complete by Lemmas 5.2, 5.3, 5.4, and 5.5 below.

5.1. A remark on the representations ρ_ψ. In Section 5, we will need to modify the definition of the local representation ρ_ψ given in Section 2 to make it compatible with our global arguments. We first assume that v is split in E/F. Let ψ be as in Proposition 5.1, and consider the localization ψ_v, which is a parameter for $GL(N, E_w)$. In Section 2, we associated a representation ρ_{ψ_v} to ψ_v, by decomposing ψ_v into irreducibles and applying the recipe (5). We may modify this definition, by instead using the decomposition $\psi_v = \oplus_{i \geq 1} \phi_{v,i} \otimes \nu(n_i)$ with $\phi_{v,i} \in \Phi_v(m_i)$ associated to the shape $\mathcal{S}$ (so that the $\phi_{v,i}$ are not necessarily irreducible), and in Section 5 we will let ρ_{ψ_v} denote the representation of $GL(N, E_w)$ obtained in this way. In particular, ρ_{ψ_v} is induced from a parabolic of $GL(N)$ with Levi $\prod_{i=1}^k GL(m_i)^{n_i}$. If v is not split in E/F, then we have a local parameter ψ_v^N for $GL(N, E_w)$, and we modify the definition of the induced representation $\rho_{\psi_v^N}$ in the same way.

If we denote the result of the old construction by ρ'_{ψ_v}, then ρ_{ψ_v} and ρ'_{ψ_v} have the same composition series. This implies that π_{ψ_v} is a subquotient of ρ_{ψ_v}, which is all that matters for most of our arguments. The only exception to this is in Lemma 5.2, which will be discussed in the course of the proof.

5.2. v split in E/F, $v \notin S$. These v are the only ones which require us to consider the special cases $m_1 = 1, 2, 3$ of Proposition 5.1 separately. In this case, the local packets under consideration each contain a single representation of $GL(N, F_v)$ or $GL(m_i, F_v)$. The bound we prove, Lemma 5.2, is an application of the fact that the representation in $\Pi_{\psi_v}(G_v, \xi_v)$ is a subquotient of an explicit induced representation.

Lemma 5.2. *Let* $\Pi_{\psi_v}(G, \xi) = \{\pi_v\}$ *and* $\Pi_{\phi_{i,v}}(U(m_i)) = \{\pi_{i,v}\}$. *We have*

$$\dim \pi_v^{K(\mathfrak{n})_v} \leq (1+1/q_v)^{\sigma(\mathcal{S})}(1+O(q_v^{-2}))N\mathfrak{n}_v^{\tau(\mathcal{S})} \prod_{i \geq 1} \left(\dim \pi_{i,v}^{K_i(\mathfrak{n})_v} \right)^{n_i}, \tag{18}$$

and if $m_1 = l$ *with* $l = 1, 2, 3$ *we have*

$$\dim \pi_v^{K(\mathfrak{n})_v} \leq C(\epsilon, q_v)(1+1/q_v)^{\sigma_l(S)}(1+O(q_v^{-2}))N\mathfrak{n}_v^{\tau_l(S)} \dim \pi_{1,v}^{K_1(\mathfrak{n})_v} \prod_{i\geq 2} \left(\dim \pi_{i,v}^{K_i(\mathfrak{n})_v}\right)^{n_i}. \tag{19}$$

The terms involving $1 + 1/q_v$ *only need to be included if* $v|\mathfrak{n}$. *The term* $C(\epsilon, q_v) = 1$ *if* $l = 1$ *or* q_v *is greater than a constant depending on* ϵ.

Proof. We recall the identification $G_v = GL(N, E_w)$, which carries K_v to $\widetilde{K}_w$. View ψ_v as a member of $\Psi(GL(N, E_w))$. As in Section 5.1, we have an irreducible subquotient $\pi_v = \pi_{\psi_v}$ of an induced representation $\rho_w = \rho_{\psi_v}$ of $GL(N, E_w)$. Let P_w denote the block upper triangular parabolic subgroup from which ρ_w is induced. (So the Levi factor of P_w is $\prod_{i=1}^k GL(m_i)^{n_i}$.)

We shall prove the first bound using $\dim \pi_v^{K(\mathfrak{n})_v} \leq \dim \rho_w^{\widetilde{K}(\mathfrak{n})_w}$. We have

$$\dim \rho_w^{\widetilde{K}(\mathfrak{n})_w} = [\widetilde{K}_w : \widetilde{K}_w \cap \widetilde{K}(\mathfrak{n})_w P_w] \prod_{i\geq 1} \left(\dim \mu_{i,w}^{\widetilde{K}_i(\mathfrak{n})_w}\right)^{n_i}.$$

The result then follows from the fact that $[\widetilde{K}_w : \widetilde{K}_w \cap \widetilde{K}(\mathfrak{n})_w P_w] = 1$ if $v \nmid \mathfrak{n}$, while if $v|\mathfrak{n}$ we have

$$[\widetilde{K}_w : \widetilde{K}_w \cap \widetilde{K}(\mathfrak{n})_w P_w] = (1 + 1/q_v)^{\sigma(S)}(1 + O(q_v^{-2}))N\mathfrak{n}_v^{\tau(S)},$$

and the fact that $\pi_{i,v}$ are isomorphic to $\mu_{i,w}$ so that $\dim \mu_{i,w}^{\widetilde{K}_i(\mathfrak{n})_w} = \dim \pi_{i,v}^{K_i(\mathfrak{n})_v}$.

The case $m_1 = 1$ is the only place that we need to know the exact definition of π_v, not just that it is a subquotient of ρ_w. Let ρ'_w be the induced representation of $GL(N, E_w)$ associated to ψ_v in Section 2. Because $\phi_{i,v}$ are bounded, π_v is the unique irreducible quotient of ρ'_w. Becuase $m_1 = 1$, ρ'_w is induced from a standard parabolic P'_w with Levi factor $L'_w = GL(1)^{n_1} \times \prod_j GL(t_j)$ for some t_j. Moreover, the representation one induces is given on the $GL(1)^{n_1}$ factor of L'_w by

$$|\det|^{\frac{n_1-1}{2}}\mu_{1,w} \otimes \ldots \otimes |\det|^{\frac{-n_1+1}{2}}\mu_{1,w}.$$

Let $(P_w^1)'$ be the parabolic obtained by modifying P'_w in the upper-left $n_1 \times n_1$ block so that the $GL(1)^{n_1}$ factor in the Levi is replaced by $GL(n_1)$. Let $(\rho_w^1)'$ be the representation induced from $(P_w^1)'$ using the same data as ρ'_w, except that one takes the representation $\mu_{1,w} \circ \det(n_1)$ on the new Levi factor $GL(n_1, E_w)$. As $(\rho_w^1)'$ is a quotient of ρ'_w, π_v is also a quotient of $(\rho_w^1)'$.

We may perform a similar modification to ρ_w, to define a representation ρ_w^1 induced from the standard parabolic P_w^1 with Levi $GL(n_1)\times\prod_{i\geq 2} GL(m_i)^{n_i}$. As ρ_w^1 and $(\rho_w^1)'$ have the same composition series, π_v is a subquotient of ρ_w^1. It follows that $\dim \pi_v^{K(\mathfrak{n})_v} \leq \dim(\rho_w^1)^{\widetilde{K}(\mathfrak{n})_w}$, and

$$\dim(\rho_w^1)^{\widetilde{K}(\mathfrak{n})_w} = [\widetilde{K}_w : \widetilde{K}_w \cap \widetilde{K}(\mathfrak{n})_w P_w^1] \dim \mu_{1,w}^{\widetilde{K}_1(\mathfrak{n})_w} \prod_{i \geq 2} \left(\dim \mu_{i,w}^{\widetilde{K}_i(\mathfrak{n})_w} \right)^{n_i}.$$

The result follows as before, after calculating $[\widetilde{K}_w : \widetilde{K}_w \cap \widetilde{K}(\mathfrak{n})_w P_w^1]$.

In the cases $m_1 = 2, 3$, we bound all but one of the factors of $\dim \pi_{1,v}^{K_1(\mathfrak{n})_v}$ in (18) using the representation theory of $GL(m_1, F_v)$. In particular, applying Corollary A.4 in (18) gives

$$\dim \pi_v^{K(\mathfrak{n})_v} \leq C(\epsilon, q_v)(1 + 1/q_v)^{\sigma(S)}(1 + O(q_v^{-2}))N\mathfrak{n}_v^{\tau(S)+\frac{m_1(m_1-1)}{2}(n_1-1)+\epsilon}$$
$$\dim \pi_{1,v}^{K_1(\mathfrak{n})_v} \prod_{i \geq 2} \left(\dim \pi_{i,v}^{K_i(\mathfrak{n})_v} \right)^{n_i},$$

which gives (19).

□

5.3. v split in E/F, $v \in S$. In this case, $G_v^* \simeq GL(N, F_v) \simeq GL(N, E_w)$ and G_v is an inner form of $GL(N, F_v)$. It is known (see [16, Theorem 1.6.4] for instance) that the packet $\Pi_{\psi_v}(G^*)$ contains exactly one element whereas $\Pi_{\psi_v}(G, \xi)$ has one or zero elements.

For $v \in S$, the constant C_v can be arbitrary. This means that to prove (17), we need to know that the left hand side is bounded independently of ψ_v, and that if it is nonzero, then the right hand side is also nonzero. Both facts are provided by the following local lemma, where we consider $\psi_v \in \Psi(G_v^*)$ and bounded $\phi_{i,v} \in \Phi(GL(m_i, F_v))$ with $\psi_v = \oplus_{i=1}^k \phi_{i,v} \boxtimes \nu(n_i)$. The unique representations in $\Pi_{\psi_v}(G, \xi)$ and $\Pi_{\phi_{i,v}}(U(m_i)_v) = \Pi_{\phi_{i,v}}(GL(m_i, F_v))$ are denoted by π_v and $\pi_{i,v}$, respectively.

Lemma 5.3. *There is $C(K_v) > 0$ such that $\dim \pi_v^{K_v} \leq C(K_v)$. For each i there exists an open compact subgroup $K_{i,v} \subset U(m_i, F_v)$ depending only on K_v such that the following is true for every ψ_v and $\phi_{i,v}$ as above: if $\pi_v^{K_v} \neq 0$, then $\pi_{i,v}^{K_{i,v}} \neq 0$ for all i.*

Proof. The first claim is Bernstein's uniform admissibility theorem [5]. (We need it just for unitary representations, but the proof there shows the theorem for irreducible admissible representations of general p-adic reductive groups.)

To prove the second claim, recall that ψ_v gives rise to representations ρ_{ψ_v} and π_{ψ_v} of $G_v^* \simeq GL(N, F_v)$ as in Section 5.1. So π_{ψ_v} is an irreducible subquotient of ρ_{ψ_v}.

The hypothesis $\pi_v^{K_v} \neq 0$ means that $1_{K_v}(\pi_v) \neq 0$. If we transfer 1_{K_v} to a function $1^*_{K_v}$ on G^*_v, we have the character identity $1^*_{K_v}(\pi_{\psi_v}) = e(G_v)a_{\psi_v}1_{K_v}(\pi_v) \neq 0$ by Theorem 1.6.4 (1) of [16] with certain signs $e(G_v), a_{\psi_v} \in \{\pm 1\}$. If we let $K'_v \subset GL(N, F_v)$ be an open compact subgroup such that $1^*_{K_v}$ is bi-invariant under K'_v, this implies that $\pi_{\psi_v}^{K'_v} \neq 0$ and thus $\rho_{\psi_v}^{K'_v} \neq 0$. This gives $\pi_{i,v}^{K_{i,v}} \neq 0$ for suitable $K_{i,v} \subset GL(m_i, F_v)$, which implies the claim. (To see this, one uses a description of invariant vectors in an induced representation under an open compact subgroup as in the first display of [4, p.26], noting that the double coset $P\backslash G/K$ there is finite.)

□

5.4. v nonsplit in E/F, $v \notin S$. In this case, for each $\psi_v \in \Psi(G^*_v)$ we have $\psi_v^N = \eta_{\chi_\kappa} \circ \psi_v \in \Psi(\tilde{G}(N)) = \Psi(GL(N, E_w))$. This gives rise to representations $\pi_{\psi_v^N}$ and $\rho_{\psi_v^N}$ of $GL(N, E_w)$ as in Section 5.1. Similarly $\phi_{i,v} \in \Phi(U(m_i)_v)$ gives a representation $\pi_{\phi_{i,v}^{m_i}}$ of $GL(m_i, E_w)$ for the parameter $\eta_{\chi_{\kappa_i}} \circ \phi_{i,v}$. If ψ_v and $\phi_{i,v}$ arise from global data as at the start of Section 5 then $\pi_{\phi_{i,v}^{m_i}}$ is nothing but $\mu_{i,w}$.

Inequality (17) in this case follows from the lemma below.

Lemma 5.4. *Consider $\psi_v \in \Psi(G^*_v)$ and $\phi_{i,v} \in \Phi(U(m_i)_v)$ as above such that $\psi_v^N = \oplus_{i\geq 1}\phi_{i,v}^{m_i} \boxtimes \nu(n_i)$. Then we have*

$$\sum_{\pi_v \in \Pi_{\psi_v}(G,\xi)} \dim \pi_v^{K_v} \leq 1. \tag{20}$$

If equality holds, then

$$\sum_{\pi_{i,v} \in \Pi_{\phi_{i,v}}(U(m_i))} \dim \pi_{i,v}^{K_{i,v}} = 1 \tag{21}$$

for all i.

Proof. Suppose first that $s_{\psi_v} \in \{\pm 1\}$. We have a hyperspecial subgroup $\widetilde{K}_v$ of $G(N)_v \simeq GL(N, E_w)$. The twisted fundamental lemma implies that the functions 1_{K_v} and $1_{\widetilde{K}_v \rtimes \theta}$ are related by transfer.

Applying the character identity for $U(N)$ (Theorem 3.2.1 (b) of [25]) with $s = 1$ gives

$$1_{K_v}^{U(N)}(\psi_v) = \sum_{\pi_v \in \Pi_{\psi_v}(G,\xi)} \dim \pi_v^{K_v},$$

and combining this with the twisted character identity [25, Theorem 3.2.1 (a)] and the twisted fundamental lemma gives

$$\sum_{\pi_v \in \Pi_{\psi_v}(G,\xi)} \dim \pi_v^{K_v} = \mathrm{tr}(\widetilde{\pi}_{\psi_v^N}(1_{\widetilde{K}_v \rtimes \theta})).$$

The twisted trace $\mathrm{tr}(\widetilde{\pi}_{\psi_v^N}(1_{\widetilde{K}_v \rtimes \theta}))$ is equal to the trace of $\widetilde{\pi}_{\psi_v^N}(\theta)$ on $\pi_{\psi_v^N}^{\widetilde{K}_v}$, so we have

$$\mathrm{tr}(\widetilde{\pi}_{\psi_v^N}(1_{\widetilde{K}_v \rtimes \theta})) \leq \dim \pi_{\psi_v^N}^{\widetilde{K}_v}.$$

Since $\pi_{\psi_v^N}$ is a subquotient of $\rho_{\psi_v^N}$, we have

$$\dim \pi_{\psi_v^N}^{\widetilde{K}_v} \leq \dim \rho_{\psi_v^N}^{\widetilde{K}_v} \leq 1$$

which gives (20).

If equality holds, then ψ_v^N is unramified. So all $\phi_{i,v}$ are unramified as well. Applying [25, Theorem 3.2.1 (b)] to the parameter $\phi_{i,v}$ and the function $1_{K_{i,v}}$ for $U(m_i)$ gives

$$\sum_{\pi_{i,v} \in \Pi_{\phi_{i,v}}(U(m_i))} \dim \pi_{i,v}^{K_{i,v}} = 1_{K_{i,v}}^{U(m_i)}(\phi_{i,v}).$$

If $\widetilde{\pi}_{\phi_{i,v}^{m_i}}$ is the canonical extension of $\pi_{\phi_{i,v}^{m_i}}$ to $\widetilde{G}(m_i)_v$ (via Whittaker normalization),

$$\sum_{\pi_{i,v} \in \Pi_{\phi_{i,v}}(U(m_i))} \dim \pi_{i,v}^{K_{i,v}} = \mathrm{tr}(\widetilde{\pi}_{\phi_{i,v}^{m_i}}(1_{\widetilde{K}_{i,v} \rtimes \theta})).$$

$\mathrm{tr}(\widetilde{\pi}_{\phi_{i,v}^{m_i}}(1_{\widetilde{K}_{i,v} \rtimes \theta}))$ is the trace of θ on the one-dimensional space $\pi_{\phi_{i,v}^{m_i}}^{\widetilde{K}_{i,v}}$, so we have $\mathrm{tr}(\widetilde{\pi}_{\phi_{i,v}^{m_i}}(1_{\widetilde{K}_{i,v} \rtimes \theta})) = \pm 1$, and (21) follows from positivity.

Now suppose that $s_{\psi_v} \notin \{\pm 1\}$, and let $(G^{\mathfrak{e}}, s^{\mathfrak{e}}, \eta^{\mathfrak{e}})$ be the elliptic endoscopic triple for G with $s^{\mathfrak{e}} = s_{\psi_v}$. We have $G^{\mathfrak{e}} = U(a) \times U(b)$ for some $a, b > 0$. There is an Arthur parameter $\psi^{\mathfrak{e}}$ for $G^{\mathfrak{e}}$ such that $\eta^{\mathfrak{e}} \circ \psi^{\mathfrak{e}} = \psi$, which we may factorise as $\psi^{\mathfrak{e}} = \psi_1 \times \psi_2$. We let $K_v^{\mathfrak{e}} \subset G^{\mathfrak{e}}(F_v)$ be a hyperspecial subgroup, and let $1_{K_v}^{\mathfrak{e}}$ be the characteristic function of $K_v^{\mathfrak{e}}$. The Fundamental Lemma implies that $1_{K_v} \in \mathcal{H}(G_v)$ and $1_{K_v}^{\mathfrak{e}} \in \mathcal{H}(G_v^{\mathfrak{e}})$ have $\Delta[\mathfrak{e}, \xi, z]$-matching orbital integrals. Applying [25, Theorem 3.2.1 (b)] with $s = s_{\psi_v}$ gives

$$\sum_{\pi_v \in \Pi_{\psi_v}(G,\xi)} 1_{K_v}(\pi_v) = 1_{K_v}^{\mathfrak{e}}(\psi_v^{\mathfrak{e}}) = 1_{K_{1,v}}(\psi_{1,v}) 1_{K_{2,v}}(\psi_{2,v}).$$

The result now follows by applying the result in the case $s_{\psi_v} \in \{\pm 1\}$ to the groups $U(a)$ and $U(b)$.

□

5.5. v nonsplit in E/F, $v \in S$. Here we prove a result analogous to Lemma 5.3.

Lemma 5.5. *There exist open compact subgroups $K_{i,v} \subset U(m_i)_v$ depending only on K_v such that the following holds: given $\psi_v \in \Psi(G_v)$ and $\phi_{i,v} \in \Phi(U(m_i)_v)$ such that $\psi_v^N = \oplus_{i=1}^k \phi_{i,v}^N \boxtimes \nu(n_i)$ (thus $\phi_{i,v}$ are bounded), if*

$$\sum_{\pi_v \in \Pi_{\psi_v}(G,\xi)} \dim \pi_v^{K_v} \neq 0 \quad \textit{then} \quad \sum_{\pi_{i,v} \in \Pi_{\phi_{i,v}}(U(m_i))} \dim \pi_{i,v}^{K_{i,v}} \neq 0.$$

Moreover there is a constant $C(K_v) > 0$ which is independent of ψ_v such that

$$\sum_{\pi_v \in \Pi_{\psi_v}(G,\xi)} \dim \pi_v^{K_v} \leq C(K_v).$$

Proof. We begin with the first claim. Suppose $s_{\psi_v} \in \{\pm 1\}$. Let $1_{K_v}^*$ be the transfer of 1_{K_v} to G_v^*. The character identity of [16, Thm 1.6.1 (4)] gives

$$0 \neq e(G_v) \sum_{\pi_v \in \Pi_{\psi_v}(G,\xi)} \dim \pi_v^{K_v} = 1_{K_v}^*(\psi_v),$$

where $e(G_v) \in \{\pm 1\}$; note that the coefficients $\langle \pi, 1 \rangle$ appearing in the cited theorem are all 1 (where we take $s^{\mathfrak{e}} = s_{\psi_v}$). Using the surjectivity result of Mok [25, Prop. 3.1.1 (b)], there is a function $\widetilde{1_{K_v}}$ on $\widetilde{G}(N)_v$ whose twisted transfer to G_v^* is $1_{K_v}^*$, and so we have $1_{K_v}^*(\psi_v) = \mathrm{tr}(\widetilde{\pi}_{\psi_v}(\widetilde{1_{K_v}}))$. Let $\widetilde{K}_v \subset G(N)_v$ be a compact open subgroup such that $\widetilde{1_{K_v}}$ is bi-invariant under $\widetilde{K}_v$. It follows that we must have $\pi_{\psi_v}^{\widetilde{K}_v} \neq 0$, and hence there are compact open $\widetilde{K}_{i,v} \subset G(m_i)_v$ depending only on K_v such that $\pi_{\phi_{i,v}}^{\widetilde{K}_{i,v}} \neq 0$. The result now follows from Lemma 5.6 below.

Now suppose that $s_{\psi_v} \notin \{\pm 1\}$, and let $(G^{\mathfrak{e}}, s^{\mathfrak{e}}, \eta^{\mathfrak{e}})$ be the elliptic endoscopic triple for G with $s^{\mathfrak{e}} = s_{\psi_v}$ and so $G^{\mathfrak{e}} = U(a) \times U(b)$ for some $a, b > 0$. There is an Arthur parameter $\psi^{\mathfrak{e}}$ for $G^{\mathfrak{e}}$ such that $\eta^{\mathfrak{e}} \circ \psi^{\mathfrak{e}} = \psi$, which we may factorise as $\psi^{\mathfrak{e}} = \psi_1 \times \psi_2$. Let $1_{K_v}^{\mathfrak{e}}$ be the function obtained by transferring 1_{K_v} to $G_v^{\mathfrak{e}}$. Applying the trace identity

$$e(G_v) \sum_{\pi_v \in \Pi_{\psi_v}(G,\xi)} \dim \pi_v^{K_v} = 1_{K_v}^{\mathfrak{e}}(\psi_v^{\mathfrak{e}})$$

gives $1_{K_v}^{\mathfrak{e}}(\psi_v^{\mathfrak{e}}) \neq 0$. Because $1_{K_v}^{\mathfrak{e}}(\psi_v^{\mathfrak{e}})$ is equal to a sum of traces there is a compact open $K_{1,v} \times K_{2,v} \subset G_v^{\mathfrak{e}}$ such that

$$1_{K_{1,v} \times K_{2,v}}(\psi_v^{\mathfrak{e}}) = 1_{K_{1,v}}(\psi_{1,v}) 1_{K_{2,v}}(\psi_{2,v}) \neq 0$$

and the result now follows from the case $s_{\psi_v} \in \{\pm 1\}$ for the groups $U(a)$ and $U(b)$.

We now prove the second claim. Suppose $s_{\psi_v} \in \{\pm 1\}$. We again use the identity

$$e(G_v) \sum_{\pi_v \in \Pi_{\psi_v}(G,\xi)} \dim \pi_v^{K_v} = \mathrm{tr}(\widetilde{\pi}_{\psi_v}(\widetilde{1}_{K_v})),$$

and let $\widetilde{K}_v \subset G(N)_v$ be a compact open subgroup such that $\widetilde{1}_{K_v}$ is bi-invariant under $\widetilde{K}_v$. The trace $\mathrm{tr}(\widetilde{\pi}_{\psi_v}(\widetilde{1}_{K_v}))$ is equal to the trace of $\widetilde{\pi}_{\psi_v}(\widetilde{1}_{K_v})$ on the space $\pi_{\psi_v}^{\widetilde{K}_v}$, and the operator norm of $\widetilde{\pi}_{\psi_v}(\widetilde{1}_{K_v})$ is at most $\|\widetilde{1}_{K_v}\|_1 = C(K_v)$. We therefore have $\left|\mathrm{tr}(\widetilde{\pi}_{\psi_v}(\widetilde{1}_{K_v}))\right| \leq C(K_v) \dim \pi_{\psi_v}^{\widetilde{K}_v}$, and the result follows as in Lemma 5.3. If $s_{\psi_v} \notin \{\pm 1\}$, we reduce to the case of $U(a) \times U(b)$ as before.

□

Recall that $\eta_{\chi_{\kappa_i}} \circ \phi_{i,v} \in \Phi(G(m_i)_v)$ corresponds to $\mu_{i,w}$ via local Langlands under the isomorphism $G(m_i)_v \simeq GL(m_i, E_w)$, where w is the unique place of E above v.

Lemma 5.6. *If $\widetilde{K}_{i,w} \subset GL(m_i, E_w)$ is a compact open subgroup, then there is a compact open subgroup $K_{i,v} \subset U(m_i)_v$ with the following property: For any bounded parameter $\phi_{i,v} \in \Phi(U(m_i)_v)$ and the representation $\mu_{i,w}$ of $GL(m_i, E_w)$ corresponding as above, if $\mu_{i,w}^{\widetilde{K}_{i,w}} \neq 0$ then*

$$\sum_{\pi_{i,v} \in \Pi_{\phi_{i,v}}(U(m_i))} \dim \pi_{i,v}^{K_{i,v}} \neq 0. \tag{22}$$

Proof. The only nontrivial part of the lemma is the assertion that $K_{i,v}$ may be chosen independently of $\mu_{i,w}$. To this end, we will show that $\phi_{i,v}$ (or $\mu_{i,w}$) varies over a compact domain and that $K_{i,v}$ as in the lemma can be chosen in open neighborhoods. Then the proof will be complete by taking intersection of the finitely many $K_{i,v}$ for a finite open covering.

By a theorem of Jacquet, our assumption that $\mu_{i,w}$ was tempered implies that $\mu_{i,w}$ belongs to a family of full induced representations from some

$$\mu_1' |\cdot|^{is_1} \otimes \cdots \otimes \mu_k' |\cdot|^{is_k} \tag{23}$$

with μ_j' square integrable and $s_j \in \mathbb{R}/(2\pi/\log q_w)\mathbb{Z}$. Our assumption that $\mu_{i,w}$ had bounded depth implies that the set of tuples $\mu_1', \ldots, \mu_k'$ we must consider is finite, and so we only need to consider one. We then need to show that the set of s_j such that $\mu_{i,w}$ is conjugate self-dual (i.e. $\mu_{i,w} \simeq \mu_{i,w} \circ \theta$) is compact. Because $\mu_j' |\cdot|^{is_j} \circ \theta = (\mu_j' \circ \theta) |\cdot|^{-is_j}$, this condition is equivalent to saying that the multisets

$\{\mu'_j|\cdot|^{is_j}\}$ and $\{(\mu'_j\circ\theta)|\cdot|^{-is_j}\}$ are equivalent. This in turn is equivalent to the existence of a permutation $\sigma\in S_k$ such that $\mu'_j|\cdot|^{is_j}\simeq(\mu'_{\sigma(j)}\circ\theta)|\cdot|^{-is_{\sigma(j)}}$ for all j. For each σ the set of s_j satisfying this is closed, and hence the set of s_j such that $\mu_{i,w}$ is conjugate self-dual is closed and compact.

For fixed $s_1,\ldots,s_k$, there is some $K_{i,v}$ such that $1_{K_{i,v}}(\phi_{i,v})\neq 0$, where $1_{K_{i,v}}(\phi_{i,v})$ is equal to the left hand side of (22) by definition. Also, if we transfer $1_{K_{i,v}}$ to $\widetilde{1}_{K_{i,v}}$ on $\widetilde{G}(m_i)$ using the surjectivity theorem of Mok [25, Prop. 3.1.1 (b)] then the character identity tells us that $1_{K_{i,v}}(\phi_{i,v})=\mathrm{tr}(\widetilde{\pi}^{m_i}_{\phi_{i,v}}(\widetilde{1}_{K_{i,v}}))$ (where the twisted trace is Whittaker normalized). The point is that $\mathrm{tr}(\widetilde{\pi}^{m_i}_{\phi_{i,v}}(\widetilde{1}_{K_{i,v}}))$ varies continuously in the s_j (see [27]) so we still have $1_{K_{i,v}}(\phi_{i,v})\neq 0$ around an open neighborhood of $s_1,...,s_k$ (where $\phi_{i,v}$ varies as $s_1,...,s_k$ vary). The result now follows by compactness.

□

6. Archimedean control on parameters

In this section, we prove some useful conditions on the parameters ψ that contribute to the cohomology of $X(\mathfrak{n})$.

Given $\phi_\infty=\otimes_{v|\infty}\phi_v\in\Phi(U(n)_\infty)$ for $n\geq 1$, note that the restriction of ϕ_v to $W_{\mathbb{C}}=\mathbb{C}^\times$ (for a fixed isomorphism $\overline{F}_v\simeq\mathbb{C}$), viewed as an n-dimensional representation via $\widehat{U(n)}=GL(n,\mathbb{C})$, is a direct sum of n characters $z\mapsto z^{a_{i,v}}\overline{z}^{b_{i,v}}$ with $a_{i,v},b_{i,v}\in\mathbb{C}$ and $a_{i,v}-b_{i,v}\in\mathbb{Z}$ for $i=1,...,n$. We say that ϕ_∞ is C-algebraic if n is odd and all $a_{i,v}\in\mathbb{Z}$ or if n is even and all $a_{i,v}\in\frac{1}{2}+\mathbb{Z}$. We say ϕ_v is regular if $a_{i,v}$ are distinct. If π_∞ is a member of the L-packet for ϕ_∞ then π_∞ is said to be regular or C-algebraic if ϕ_∞ is. (This is Clozel's definition and coincides with the general definition [8, Definition 2.3.3] for general reductive groups.)

Let $\mathcal{S}=(n_1,m_1),\ldots,(n_k,m_k)$ be a shape as in Section 5. If $\psi\in\Psi_2(G^*,\eta_{\chi_\kappa})_{\mathcal{S}}$ then ψ gives rise to $\mu_i\in\widetilde{\Phi}_{\mathrm{sim}}(m_i)$ and $\phi_i\in\Phi_{\mathrm{sim}}(U(m_i),\eta_{\chi_{\kappa_i}})$ as before.

Lemma 6.1. *Let $\psi\in\Psi_2(G^*,\eta_{\chi_\kappa})_{\mathcal{S}}$. If there is $\pi_\infty\in\Pi_{\psi_\infty}(G,\xi)$ with $h^d(\mathfrak{g},K_\infty;\pi_\infty)\neq 0$ then $\phi_{i,\infty}$ or $\phi_{i,\infty}\otimes\chi_{-,\infty}$ is C-algebraic. Moreover every $\phi_{i,v}$ is bounded at every place v (equivalently $\mu_{i,v}$ is tempered at every place v).*

Proof. Since π_∞ contributes to cohomology, its infinitesimal character is equal to that of the trivial representation. In particular it is regular C-algebraic, cf. [8, Lemmas 7.2.2, 7.2.3]. Hence ϕ_{ψ_∞} is regular C-algebraic. (Here we use the simple recipe to determine the infinitesimal character of π_∞ from ϕ_{ψ_∞} by differentiation, as described in [22, Section 2.1].) For each infinite place v, the representation $\phi_{i,v}|_{W_{\mathbb{C}}}$ is the direct sum of m_i characters, say $\eta_{i,1}$, ..., η_{i,m_i}. Then $\oplus_{j=1}^{m_i}\oplus_{l=0}^{n_i-1}\eta_{i,j}|\cdot|^{\frac{n_i-1}{2}-l}$ appears as a subrepresentation of ϕ_{ψ_∞}. As the latter is regular

C-algebraic, we see that for each $v|\infty$, $\phi_{i,v}|_{W_{\mathbb{C}}}$ is regular and that either $\phi_{i,v}|_{W_{\mathbb{C}}}$ or $\phi_{i,v}|_{W_{\mathbb{C}}} \otimes |\cdot|^{1/2}$ is C-algebraic, depending on the parity of $N - n_i$. It follows that $\phi_{i,\infty}|_{W_{\mathbb{C}}}$ or $\phi_{i,\infty}|_{W_{\mathbb{C}}} \otimes |\cdot|^{1/2}$ is regular C-algebraic. By the definition of χ_- in Section 2, $\phi_{i,\infty}|_{W_{\mathbb{C}}} \otimes |\cdot|^{1/2}$ is C-algebraic if and only if $\phi_{i,\infty}|_{W_{\mathbb{C}}} \otimes \chi_{-,\infty}$ is.

The key point is that μ_i or $\mu_i \otimes \chi_-$ is an automorphic representation with regular C-algebraic component at ∞ (recalling that $\mu_{i,\infty}$ lies in the packet for $\phi_{i,\infty}$). Both μ_i and $\mu_i \otimes \chi_-$ are cuspidal and conjugate self-dual, so either μ_i or $\mu_i \otimes \chi_-$ (whichever is C-algebraic at ∞) is essentially tempered at all finite places by [9, Theorem 1.2] (the cohomological condition in *loc. cit.* follows from regular C-algebraicity, cf. [11, Lemme 3.14]) and at all infinite places by [11, Lemme 4.9]. In either case, twisting by χ_- if necessary, we deduce that μ_i is essentially tempered everywhere. Since the central character of μ_i is unitary, we see that μ_i is tempered everywhere. By the local Langlands correspondence [25, Theorem 2.5.1 (b)], this is equivalent to $\phi_{i,v}$ being bounded at every v. □

Lemma 6.2. *For each i, there is a finite set of parameters $\mathcal{P}_{i,\infty} \subset \Phi(U(m_i)_\infty)$ with the following property: If $\psi \in \Psi_2(G^*, \eta_{\chi_\kappa})_{\mathcal{S}}$, and there exists $\pi_\infty \in \Pi_{\psi_\infty}(G,\xi)$ with $h^d(\mathfrak{g}, K_\infty; \pi_\infty) \neq 0$, then $\phi_{i,\infty} \in \mathcal{P}_{i,\infty}$.*

Proof. The infinitesimal character of π_∞ is determined by the condition that $h^d(\mathfrak{g}, K_\infty; \pi_\infty) \neq 0$ (to be the half sum of all positive roots of G), thus there are finitely many such π_∞. So they are contained in finitely many Arthur packets, whose parameters form a finite subset $\mathcal{P} \subset \Psi(U(N)_\infty)$. If ψ gives rise to ϕ_i then $\eta_{\chi_{\kappa_i}} \circ (\phi_{i,v} \boxtimes \nu(n_i))$ should appear as a factor of $\eta_{\chi_\kappa} \circ \psi_v$ for every infinite place v. Since we have the constraint $\otimes_{v|\infty}\psi_v \in \mathcal{P}$, it is clear that there are finitely many possibilities for $\phi_{i,\infty}$.

□

7. Summing over parameters

In this section we continue the proof of Theorem 1.2 from the end of Section 4 and finish the proof. In the preliminary bound (13), we will fix a shape $\mathcal{S}$ and bound the contribution to $h^d_{(2)}(X(\mathfrak{n}))$ from parameters in $\Psi_2(G^*, \eta_{\chi_\kappa})_{\mathcal{S}}$, which we denote by $h^d_{(2)}(X(\mathfrak{n}))_{\mathcal{S}}$. Clearly it suffices to establish a bound for $h^d_{(2)}(X(\mathfrak{n}))_{\mathcal{S}}$ as in Theorem 1.2.

Suppose $\psi \in \Psi_2(G^*, \eta_{\chi_\kappa})_{\mathcal{S}}$ has the property that there is $\pi \in \Pi_\psi(G,\xi)$ with $h^d(\mathfrak{g}, K_\infty; \pi_\infty) \neq 0$. Proposition 3.1 implies that π_{v_0} must be a Langlands quotient of a standard representation with an exponent of the form $(z/\overline{z})^{p/2}(z\overline{z})^{(a-1)/2}$ for some $a \geq N-d$. Proposition 13.2 of [3] implies that there is i such that $n_i \geq N-d$, and we assume that this is n_1. Note that [3, Prop 13.2] implicitly assumes that the other archimedean components of π have regular infinitesimal character, but this is satisfied in our case.

Apply Lemma 6.2 to obtain finite sets $\mathcal{P}_{i,\infty} \subset \Phi(U(m_i)_\infty)$ for all i such that if $\psi \in \Psi_2(G^*, \eta_{\chi_\kappa})_S$, and there is $\pi \in \Pi_\psi(G, \xi)$ with $h^d(\mathfrak{g}, K_\infty; \pi_\infty) \neq 0$, then $\phi_{i,\infty} \in \mathcal{P}_{i,\infty}$. Let Ψ_{rel} be the set of A-parameters $\psi \in \Psi_2(G^*, \eta_{\chi_\kappa})_S$ with ϕ_i bounded everywhere and $\phi_{i,\infty} \in \mathcal{P}_{i,\infty}$ for all i. By Lemmas 6.1 and 6.2 we have

$$h^d_{(2)}(X(\mathfrak{n}))_S \leq \sum_{\psi \in \Psi_{\text{rel}}} \sum_{\pi \in \Pi_\psi(G,\xi)} h^d(\mathfrak{g}, K_\infty; \pi_\infty) \dim \pi_f^{K(\mathfrak{n})},$$

and because $h^d(\mathfrak{g}, K_\infty; \pi_\infty)$ is bounded we may simplify this to

$$h^d_{(2)}(X(\mathfrak{n}))_S \ll \sum_{\psi \in \Psi_{\text{rel}}} \sum_{\pi \in \Pi_\psi(G,\xi)} \dim \pi_f^{K(\mathfrak{n})}.$$

Because ϕ_i is bounded everywhere for every i, we may apply Proposition 5.1 to obtain

$$h^d_{(2)}(X(\mathfrak{n}))_S \ll \prod_{v|\mathfrak{n}} (1 + 1/q_v)^{\sigma(S)} N\mathfrak{n}^{\tau(S)} \sum_{\psi \in \Psi_{\text{rel}}} \prod_{i \geq 1} \left(\sum_{\pi_i \in \Pi_{\phi_i}(U(m_i))} \dim \pi_{i,f}^{K_i(\mathfrak{n})} \right)^{n_i}.$$

Let $\Phi^{\text{bdd}}_{\text{sim}}(U(m_i), \eta_{\chi_{\kappa_i}})$ denote the set of simple parameters that are bounded everywhere. Taking a sum over $\psi \in \Psi_{\text{rel}}$ corresponds to taking a sum over the possibilities for $\phi_i \in \Phi^{\text{bdd}}_{\text{sim}}(U(m_i), \eta_{\chi_{\kappa_i}})$ with $\phi_{i,\infty} \in \mathcal{P}_{i,\infty}$. We may therefore factorize the sum over ψ to ones over ϕ_i, which gives

$$\begin{aligned} h^d_{(2)}(X(\mathfrak{n}))_S &\ll \prod_{v|\mathfrak{n}} (1 + 1/q_v)^{\sigma(S)} N\mathfrak{n}^{\tau(S)} \prod_{i \geq 1} \sum_{\substack{\phi_i \in \Phi^{\text{bdd}}_{\text{sim}}(U(m_i), \eta_{\chi_{\kappa_i}}) \\ \phi_{i,\infty} \in \mathcal{P}_{i,\infty}}} \left(\sum_{\pi_i \in \Pi_{\phi_i}(U(m_i))} \dim \pi_{i,f}^{K_i(\mathfrak{n})} \right)^{n_i} \\ &\leq \prod_{v|\mathfrak{n}} (1 + 1/q_v)^{\sigma(S)} N\mathfrak{n}^{\tau(S)} \prod_{i \geq 1} \left(\sum_{\substack{\phi_i \in \Phi^{\text{bdd}}_{\text{sim}}(U(m_i), \eta_{\chi_{\kappa_i}}) \\ \phi_{i,\infty} \in \mathcal{P}_{i,\infty}}} \sum_{\pi_i \in \Pi_{\phi_i}(U(m_i))} \dim \pi_{i,f}^{K_i(\mathfrak{n})} \right)^{n_i}. \end{aligned} \tag{24}$$

We may bound the sums using the global limit multiplicity formula of Savin [29]. Indeed, because ϕ_i is a simple generic parameter, the packet $\Pi_{\phi_i}(U(m_i))$ is stable, so that every representation $\pi_i \in \Pi_{\phi_i}(U(m_i))$ occurs in the discrete spectrum of $U(m_i)$ with multiplicity one. In fact, π_i must actually lie in the cuspidal spectrum by [32, Theorem 4.3], because $\pi_{i,\infty}$ is tempered. Because the archimedean components of ϕ_i are restricted to finite sets, there is a finite set $\Pi_{i,\infty}$

of representations of $U(m_i)_\infty$ such that if $\pi_i \in \Pi_{\phi_i}(U(m_i))$ then $\pi_{i,\infty} \in \Pi_{i,\infty}$. If $m_{\text{cusp}}(\pi_\infty, Y_i(\mathfrak{n}))$ denotes the multiplicity with which an irreducible representation π_∞ of $U(m_i)_\infty$ occurs in the L^2-space of cuspforms $L^2_{\text{cusp}}(Y_i(\mathfrak{n}))$, where $Y_i(\mathfrak{n}) = U(m_i, F)\backslash U(m_i, \mathbb{A})/K_i(\mathfrak{n})$, we have

$$\sum_{\substack{\phi_i \in \Phi^{\text{bdd}}_{\text{sim}}(U(m_i), \eta_{\chi_{\kappa_i}}) \\ \phi_{i,\infty} \in \mathcal{P}_{i,\infty}}} \sum_{\pi_i \in \Pi_{\phi_i}(U(m_i))} \dim \pi_{i,f}^{K_i(\mathfrak{n})} \leq \sum_{\substack{\pi_i \subset L^2_{\text{cusp}}([U(m_i)]) \\ \pi_{i,\infty} \in \Pi_{i,\infty}}} \dim \pi_{i,f}^{K_i(\mathfrak{n})} = \sum_{\pi_\infty \in \Pi_{i,\infty}} m_{\text{cusp}}(\pi_\infty, Y_i(\mathfrak{n})).$$

For each π_∞, Savin [29] gives

$$m_{\text{cusp}}(\pi_\infty, Y_i(\mathfrak{n})) \ll [K_i : K_i(\mathfrak{n})] \ll \prod_{v|\mathfrak{n}} (1 - 1/q_v) N\mathfrak{n}^{m_i^2},$$

and combining this with (24) gives

$$h^d_{(2)}(X(\mathfrak{n}))_{\mathcal{S}} \ll \prod_{v|\mathfrak{n}} (1 - 1/q_v) N\mathfrak{n}^{\tau'(\mathcal{S})},$$

where $\tau'(\mathcal{S}) = \tau(\mathcal{S}) + \sum_{i \geq 1} n_i m_i^2$. If $1 \leq m_1 \leq 3$ then applying Proposition 5.1 and working as above gives

$$h^d_{(2)}(X(\mathfrak{n}))_{\mathcal{S}} \ll \prod_{v|\mathfrak{n}} (1 + 1/q_v)^{\sigma'_l(\mathcal{S})} N\mathfrak{n}^{\tau'_l(\mathcal{S})},$$

where $l = m_1$, $\tau'_l(\mathcal{S}) = \tau_l(\mathcal{S}) + m_1^2 + \sum_{i \geq 2} n_i m_i^2$, and $\sigma'_l(\mathcal{S}) = \sigma_l(\mathcal{S}) - 1 - \sum_{i \geq 2} n_i$.

The bounds for the functions τ' and τ'_j given by Lemma 7.1 below then imply that

$$h^d_{(2)}(X(\mathfrak{n}))_{\mathcal{S}} \ll_\epsilon N\mathfrak{n}^{Nd+\epsilon},$$

unless we are in one of the two cases listed there. In the exceptional case (ii) we have $h^d_{(2)}(X(\mathfrak{n}))_{\mathcal{S}} \ll N\mathfrak{n}^{Nd+1+\epsilon}$, and in case (i), which should give the general main term, we have

$$h^d_{(2)}(X(\mathfrak{n}))_{\mathcal{S}} \ll \prod_{v|\mathfrak{n}} (1 - 1/q_v) N\mathfrak{n}^{Nd+1}.$$

This completes the proof of Theorem 1.2. □

It remains to prove the lemma used in the above proof.

Lemma 7.1. *If $m_1 \geq 4$, we have $\tau'(\mathcal{S}) \leq Nd$. If $m_1 = l$, $l = 1, 2, 3$, we have $\tau'_l(\mathcal{S}) \leq Nd + \epsilon$, except in the following cases.*

(i) $\mathcal{S} = (N - d, 1), (1, d)$, in which case $\tau'_1(\mathcal{S}) = Nd + 1$.

(ii) $\mathcal{S} = (2,2)$ *and* $d = 2$*, in which case* $\tau_2'(\mathcal{S}) = Nd + 1 + \epsilon = 9 + \epsilon$.

Proof. We begin with the case $m_1 \geq 4$. The inequality $d \geq N - n_1$ implies that it suffices to prove $\tau'(\mathcal{S}) \leq N(N - n_1)$. Substituting the definition of τ' and simplifying, we must show that

$$\binom{N}{2} + \sum_{i \geq 1} n_i \binom{m_i + 1}{2} \leq N(N - n_1). \tag{25}$$

We next eliminate the variables other than N, m_1 and n_1. The identity $\binom{n+1}{2} = 1 + \ldots + n$ implies that if $A = \sum a_i$, then $\binom{A+1}{2} \geq \sum \binom{a_i+1}{2}$, and applying this to the m_i with multiplicity n_i for $i \geq 2$ gives

$$\sum_{i \geq 2} n_i \binom{m_i + 1}{2} \leq \binom{N - n_1 m_1 + 1}{2}. \tag{26}$$

Note that equality occurs above if and only if $\sum_{i \geq 2} n_i$ is either 0 or 1. After applying this in (25), we are reduced to showing that

$$\binom{N}{2} + n_1 \binom{m_1 + 1}{2} + \binom{N - n_1 m_1 + 1}{2} \leq N(N - n_1).$$

Simplifying gives

$$\begin{aligned} N(N-1) + n_1(m_1+1)m_1 + (N - n_1 m_1 + 1)(N - n_1 m_1) &\leq 2N(N - n_1) \\ -2m_1 n_1 N + 2Nn_1 + m_1^2 n_1^2 &\leq -m_1^2 n_1 \\ 0 &\leq 2m_1 N - 2N - m_1^2 n_1 - m_1^2 \end{aligned}$$

As $N \geq m_1 n_1$, we have $m_1 N \geq m_1^2 n_1$ so that

$$2m_1 N - 2N - m_1^2 n_1 - m_1^2 \geq (m_1 - 2)N - m_1^2.$$

Because $n_1 \geq 2$ we have $N \geq n_1 m_1 \geq 2m_1$, so that

$$(m_1 - 2)N - m_1^2 \geq m_1^2 - 4m_1 \geq 0,$$

where we have used $m_1 \geq 4$ at the last step.

In the case $m_1 = 1$, we have

$$\tau_1'(\mathcal{S}) = \binom{N}{2} - \binom{n_1}{2} + \sum_{i \geq 2} n_i \binom{m_i + 1}{2} + 1,$$

and applying (26) gives

$$\tau_1'(\mathcal{S}) \leq \binom{N}{2} - \binom{n_1}{2} + \binom{N - n_1 + 1}{2} + 1.$$

It may be seen that the right hand side of this simplifies to $N(N - n_1) + 1$ as required. Equality occurs when $d = N - n_1$ and we have equality in (26), which is equivalent to the conditions given in (i).

In the case $m_1 = 2$, simplifying the definition of $\tau_2'(\mathcal{S})$ gives

$$\tau_2'(\mathcal{S}) = \binom{N}{2} + \sum_{i \geq 2} n_i \binom{m_i + 1}{2} + 3 + \epsilon,$$

and after applying (26) we have

$$\tau_2'(\mathcal{S}) \leq \binom{N}{2} + \binom{N - 2n_1 + 1}{2} + 3 + \epsilon.$$

We must therefore show that

$$\binom{N}{2} + \binom{N - 2n_1 + 1}{2} + 3 \leq N(N - n_1) + 1.$$

Simplifying this gives

$$\begin{aligned} N(N-1) + (N - 2n_1 + 1)(N - 2n_1) + 4 &\leq 2N(N - n_1) \\ 4 &\leq 2Nn_1 - 4n_1^2 + 2n_1 \\ 2 &\leq n_1(N - 2n_1 + 1). \end{aligned}$$

The result now follows from $n_1 \geq 2$ and $N \geq n_1 m_1 = 2n_1$, and equality occurs exactly in case (ii).

When $m_1 = 3$, after simplifying the definition of $\tau_3'(\mathcal{S})$ and dropping the ϵ term, we must show that

$$\binom{N}{2} + 6 + \sum_{i \geq 2} n_i \binom{m_i + 1}{2} \leq \binom{N}{2} + 6 + \binom{N - 3n_1 + 1}{2} \leq N(N - n_1),$$

where the first inequality is (26). Simplifying this gives

$$\begin{aligned} N(N-1) + 12 + (N - 3n_1 + 1)(N - 3n_1) &\leq 2N(N - n_1) \\ 12 &\leq n_1(4N - 9n_1 + 3). \end{aligned}$$

We have $n_1 \geq 2$ and $N \geq 3n_1$, so that $N \geq 6$ and $4N - 9n_1 + 3 \geq N + 3 \geq 9$ as required.

□

A. Bounds for fixed vectors in representations of GL_d

In this section we prove bounds for the dimension of the vectors in an irreducible representation of GL_d that are invariant under a congruence subgroup, with no assumption on d or the residue characteristic, by applying results of Lapid [18] on the support of Whittaker functions for supercuspidal representations. These results were used in the proof of Lemma 5.2 (thus in the proof of Proposition 5.1) when $m_1 = 2, 3$.

Let F be a p-adic field. Let R be the ring of integers of F, ϖ a uniformizer, k the residue field, and q its cardinality. Write $v : F^\times \to \mathbb{Z}$ for the additive valuation such that $v(\varpi) = 1$. Let $d \geq 2$, and let $G = GL_d(F)$. Let N, A, and $K = GL_d(R)$ be the usual upper triangular unipotent, diagonal, and maximal compact subgroups of G, respectively. For $n \geq 0$ let $K(n)$ be the level ϖ^n principal congruence subgroup of K consisting of matrices which reduce to the identity matrix modulo ϖ^n.

We first prove the following result, which bounds $\dim \pi^{K(n)}$ for π supercuspidal, before using it to deduce a bound for a general π in Corollary A.3. Note that the constant c in Theorem A.2 can be made explicit, which implies that the constants C in Lemma A.1 and Corollary A.3 can be also.

Lemma A.1. *There is a constant C depending only on d such that for any $n \geq 1$ and any irreducible supercuspidal representation π of G, we have $\dim \pi^{K(n)} \leq Cn^{d-1}q^{nd(d-1)/2}$.*

Remark. We may obtain a uniform bound on $\dim \pi^{K(n)}$ of order $p^{(d^2-1)n}$ using the Plancherel theorem. Indeed, if we let Z_K be the center of K, and ω be the central character of π, we may define f to be the function supported on $Z_K K(n)$ and given by $f(zk) = \omega^{-1}(z)$. Applying the Plancherel theorem to f gives $\dim \pi^{K(n)} \leq d(\pi)^{-1}\mathrm{vol}(Z_K K(n))^{-1}$ for any Haar measure on G and for any supercuspidal π, where $d(\pi)$ is the formal degree of π. By normalizing Haar measure, we can arrange that $d(\pi)$ is a positive integer, which gives $\dim \pi^{K(n)} \ll q^{(d^2-1)n}$. This is considered a trivial bound. On the other hand, for a fixed π (either supercuspidal or any generic representation of G), the asymptotic growth of $\dim \pi^{K(n)}$ is well known to be of order $q^{nd(d-1)/2}$. (Such an asymptotic growth is known for general reductive p-adic groups either by character expansion or by a building argument [21, Theorem 8.5].) So the bound of Lemma A.1 is close to optimal (and more than enough for our global application). On a general reductive group, it is an interesting question whether a uniform bound can be established to the same order as the bound for an individual representation.

Proof of Lemma A.1. For all $n \geq 1$, define $A(n)$ to be the subset of A given by

$$A(n) = \{\mathrm{diag}(t_1, \dots, t_d) \in A : -n \leq v(t_i/t_{i+1}) \leq n, \ i = 1, \dots, d-1\}.$$

Let ψ_0 be an additive character of F with conductor R, and let ψ be the character of N defined by $\psi(n) = \psi_0(n_{1,2} + \ldots + n_{d-1,d})$. For a generic irreducible representation π of G let $\mathcal{W}(\pi)$ be the Whittaker model of π with respect to ψ. In [18, Thm, 2.1], Lapid proves the following.

Theorem A.2 (Lapid)**.** *There exists a constant $c = c(d) > 0$ with the following property. Let π be an irreducible supercuspidal representation of G with Whittaker model $\mathcal{W}(\pi)$, and $n \geq 1$. Then the support of any $W \in \mathcal{W}(\pi)^{K(n)}$ is contained in $NA(cn)K$.*

Let $G_0 = GL_{d-1}(F)$, embedded in G as the upper left block. Let N_0, K_0, and $K_0(n)$ be the corresponding subgroups of G_0. Define T_0 to be the diagonal subgroup of G_0, and for $n \geq 1$, define $T_0(n)$ to be $T_0 \cap A(n)$. We are going to deduce Lemma A.1 from Theorem A.2 by combining it with the result of Bernstein-Zelevinsky that the restriction map $\mathcal{W}(\pi) \to C(G_0)$ is injective.

Let $W \in \mathcal{W}(\pi)^{K(n)}$, and let W_0 be its restriction to G_0. Theorem A.2 implies that W_0 is supported on $N_0 T_0(cn) K_0$, and it is left equivariant under N_0 and right invariant under $K_0(n)$. Let $\mathcal{V}_0$ be the space of functions on $N_0 T_0(cn) K_0$ satisfying these conditions. Any $V \in \mathcal{V}_0$ is determined by its values on $T_0(cn)K_0/K_0(n)$, and $\dim \mathcal{V}_0$ is the number of left N_0-orbits on $T_0(cn)K_0/K_0(n)$. If $T_{0,c}$ is the maximal compact subgroup of T_0, there is a surjective map $T_0(cn)K_0/K_0(n) \to T_0(cn)/T_{0,c}$ that is constant on the orbits of N_0. The fiber of this map above $t \in T_0(cn)$ is naturally identified with $K_0/K_0(n)$, and under this identification the orbits of N_0 on the fiber are the same as those of $N_0 \cap K_0$ on $K_0/K_0(n)$. It follows that the number of N_0-orbits on $T_0(cn)K_0/K_0(n)$ is equal to the product of $\#T_0(cn)/T_{0,c}$ and $\#(N_0 \cap K_0)\backslash K_0/K_0(n)$. We have

$$\#T_0(cn)/T_{0,c} = (2cn+1)^{d-1} \leq C(d) n^{d-1}$$

for some $C(d) > 0$, and $\#(N_0 \cap K_0)\backslash K_0/K_0(n) \leq q^{nd(d-1)/2}$, so $\dim \mathcal{V}_0 \leq C(d) n^{d-1} q^{nd(d-1)/2}$. As the map $\mathcal{W}(\pi)^{K(n)} \to \mathcal{V}_0$ given by restriction to G_0 is injective, this implies the Lemma.

□

Corollary A.3. *There is a constant C depending only on d such that for any $n \geq 1$ and any irreducible representation π of G, we have* $\dim \pi^{K(n)} \leq C n^{d-1} q^{nd(d-1)/2}$.

Proof. There is a standard parabolic P with Levi $\prod_{i=1}^k GL_{d_i}$, and irreducible supercuspidal representations σ_i of GL_{d_i}, such that π embeds in $\mathrm{Ind}_P^G(\sigma_1 \times \ldots \times \sigma_k)$. We have

$$\begin{aligned}\dim \pi^{K(n)} &\leq \dim \mathrm{Ind}_P^G(\sigma_1 \times \ldots \times \sigma_k)^{K(n)} \\ &= \#(P \cap K \backslash K/K(n)) \prod_{i=1}^k \dim \sigma_i^{K_i(n)},\end{aligned}$$

where $K_i(n)$ are the level ϖ^n principal congruence subgroups of GL_{d_i}. Applying Lemma A.1 gives

$$\dim \pi^{K(n)} \leq C(d)\#(P \cap K \backslash K/K(n)) n^{d-1} q^{n \sum \binom{d_i}{2}},$$

and the bound $\#(P \cap K\backslash K/K(n)) \leq q^{n \dim(G/P)}$ finishes the proof.

□

We may rewrite Corollary A.3 in a form which is less sharp, but better suited to the proof of our main theorem.

Corollary A.4. *If π is an irreducible admissible representation of G, then for every $\epsilon > 0$ there is a constant $C(\epsilon, q) > 0$ such that*

$$\dim \pi^{K(n)} \leq C(\epsilon, q) q^{(d(d-1)/2+\epsilon)n}.$$

Moreover, for any $\epsilon > 0$ there is $q(\epsilon) > 0$ such that we may take $C(\epsilon, q) = 1$ for all $q > q(\epsilon)$.

B. Bounds for fixed vectors in representations of GL_3

Let F be a p-adic field. Throughout this section we assume $p \neq 2, 3$. Let R be the ring of integers of F, ϖ a uniformizer, k the residue field, and q its cardinality. Write $v : F^\times \to \mathbb{Z}$ for the additive valuation such that $v(\varpi) = 1$. Let $G = GL_3(F)$, $K = GL_3(R)$, and $A = M_3(R)$. Let K_j be the subgroup of K containing all elements congruent to 1 modulo ϖ^j. Put $U_j = 1 + \pi^j R$, a subgroup of $F^\times$. The following result was used in an earlier version of this paper as a substitute for the results of Appendix A. We have decided to leave it in, as it may be of independent interest.

Theorem B.1. *Assume $p \neq 2, 3$. If π is an irreducible supercuspidal representation of G, then*

$$\dim \pi^{K_n} \leq 9n^2 q^{4n}(1 + 1/q)^3.$$

The proof uses the construction of supercuspidal representations of $GL_n(F)$ by Howe in [15], in the case $n = 3$. It was shown in [26] that these exhaust all supercuspidal representations of G when p is not 3.

Remark. We rely on [15], where it is essential to assume $p \neq 3$ (or more generally $p \nmid r$ if we study GL_r). It may be superfluous to require $p \neq 2$, but as p is odd in [15], we keep this assumption. When $p \in \{2, 3\}$, one could try to adapt our argument using the construction of supercuspidals in [10] or [7], but we have not investigated this as Appendix A gives the desired result for our purpose for all p via a simpler approach.

B.1. An overview of Howe's construction. We now describe the construction of Howe in more detail, including the features we shall use to prove Theorem B.1. Howe's representations $V(\psi')$ are parametrized by a degree 3 extension F'/F and a character ψ' of $F'^{\times}$, satisfying a condition called admissibility. Fix such an F' and ψ', and let R', ϖ', and k' be the ring of integers, uniformizer, and residue field of F'. Write $N = N(F'/F)$ for the norm map from F' to F. Choose a basis for R' as a free R-module, which defines an embedding $F' \subset M_3(F)$. We shall identify F' with a subalgebra of $M_3(F)$ from now on. We define the order $A' = \cap_{x \in F'^{\times}} xAx^{-1}$, which is characterized as the set of matrices M such that $M\varpi'^i R^3 \subset \varpi'^i R^3$ for all i. We define $K' = \cap_{x \in F'^{\times}} xKx^{-1} = A' \cap GL_3(R)$, which is the subgroup of matrices preserving the lattices $\varpi'^i R^3$. For $i \geq 1$ we define $K'_i = 1 + \varpi'^i A'$ and $U'_i = 1 + \varpi'^i R'$. Let j be the conductor of ψ', that is the minimal j such that ψ' is trivial on U'_j. The admissibility condition placed on ψ' implies that $j \geq 1$.

In [15, Lemma 12], Howe constructs a representation $W(\psi')$ of $K'F'^{\times}$[4], and defines the supercuspidal representation $V(\psi')$ associated to F' and ψ' to be the compact induction of $W(\psi')$ to G. We know that $\dim \pi^{K_n}$ is at most $\dim W(\psi')$ times the number of double cosets of the form $K'F'^{\times}gK_n$ that support K_n-invariant vectors, that is such that $W(\psi')^{gK_n g^{-1} \cap K'F'^{\times}} \neq 0$. Bounding $\dim W(\psi')$ is easy, while bounding the number of these double cosets requires a feature of $W(\psi')$ from Howe's paper that we now describe.

We first assume that $j \geq 2$. The representation $W(\psi')$ is trivial on K'_j, and because K'_{j-1}/K'_j is abelian, $W(\psi')|_{K'_{j-1}}$ decomposes into characters. Howe defines a character ψ of K'_{j-1}/K'_j by taking the natural extension of ψ' from U'_{j-1}, and shows that $W(\psi')|_{K'_{j-1}}$ contains exactly the characters lying in the K'-orbit of ψ for the natural action of K' on $\widehat{K'_{j-1}/K'_j}$.

We use this fact to control those g supporting invariant vectors by observing that if $W(\psi')^{gK_n g^{-1} \cap K'F'^{\times}} \neq 0$, then $W(\psi')^{gK_n g^{-1} \cap K'_{j-1}} \neq 0$. However, if $g \in K\lambda(\varpi)K$ with $\lambda \in X_*(T)$ too large, then $gK_n g^{-1} \cap K'_{j-1}$ will contain the intersection of K'_{j-1} with a unipotent subgroup of G, and this will turn out to be incompatible with the

[4] Lemma 12 only defines $W(\psi')$ as a representation of K', but it can be extended to $K'F'^{\times}$ by the remarks at the start of [15, Thm 2].

description of $W(\psi')|_{K'_{j-1}}$. In the case $j = 1$, F' is unramified over F and $W(\psi')$ is inflated from a cuspidal representation of $K'/K'_1 \simeq GL_3(k)$, and we may use this to argue in a similar way.

In the case of GL_3, Howe's construction may be naturally divided into the cases where F'/F is ramified or unramified. We shall therefore divide our proof into these two cases, after introducing some more notation and defining the character ψ.

B.2. The character ψ. We now assume that $j \geq 2$, and define the character ψ of K'_{j-1}/K'_j that appears in the description of $W(\psi')|_{K'_{j-1}}$.

Let B' be the group of prime to p roots of unity in $F'^{\times}$, which is naturally identified with $k'^{\times}$. Let C' be the group generated by B' and ϖ'. Let $\langle\ ,\ \rangle$ be the pairing $\langle S,T\rangle = \mathrm{tr}(ST)$ on $M_3(F)$. Let χ be a character of F of conductor R, which defines an isomorphism $\theta : M_3(F) \to \widehat{M_3(F)}$ by $\theta(S)(T) = \chi(\langle S,T\rangle)$. Let e denote the degree of ramification of F'/F. Because the dual lattice to A' under $\langle,\rangle$ is $\varpi'^{1-e}A'$ by [15, Lemma 2], the map θ gives an isomorphism between the character group of $\varpi'^{i-1}A'/\varpi'^{i}A'$ and $\varpi'^{-i-e+1}A'/\varpi'^{-i-e+2}A'$. We may combine θ with the isomorphism $K'_{j-1}/K'_j \simeq \varpi'^{j-1}A'/\varpi'^{j}A'$ to obtain $\mu : \widehat{K'_{j-1}/K'_j} \to \varpi'^{-j-e+1}A'/\varpi'^{-j-e+2}A'$. If $\mu(\varphi) = y + \varpi'^{-j-e+2}A'$, we say that y represents φ. If φ has a representative $y \in F'^{\times}$, we see that φ also has a unique representative $c \in C'$, which is called the standard representative of φ.

The map θ restricts to a map $F' \to \widehat{F'}$, which is also given by $\theta(x)(y) = \chi(\mathrm{tr}_{F'/F}xy)$. We may combine θ with the isomorphism $U'_{j-1}/U'_j \simeq \varpi'^{j-1}R'/\varpi'^{j}R'$ to obtain $\mu' : \widehat{U'_{j-1}/U'_j} \to \varpi^{-j-e+1}R'/\varpi^{-j-e+2}R'$. If $\mu'(\varphi) = y + \varpi^{-j-e+2}R'$, we say that y represents φ. We see that a nontrivial φ has a unique representative $c \in C'$, which is called the standard representative of φ.

We define ψ by taking the standard representative c for ψ' on U'_{j-1}, and letting ψ be the character represented by c. If we let Ad^* denote the natural action of K' on $\widehat{K'_{j-1}/K'_j}$, given explicitly by $[\mathrm{Ad}^*(k)\psi](g) = \psi(k^{-1}gk)$, then [15, Lemma 12] states that $W(\psi')|_{K'_{j-1}}$ contains exactly the characters in $\mathrm{Ad}^*(K')\psi$.

B.3. Reduction to the case $c \notin F$. We may carry out the argument sketched in Section B.1 once we have reduced to the case where either $j = 1$ or $c \notin F$. We perform this reduction by observing that if $j \geq 2$ and $c \in F$, then $V(\psi')$ is a twist of $V(\psi_1)$ for some ψ_1 of smaller conductor. Indeed, by [15, Lemma 11], if $c \in F$ then we may write $\psi' = \psi_1\psi_2$, where ψ_1 is trivial on U'_{j-1} and $\psi_2 = \psi'' \circ N(F'/F)$ for some character ψ'' of $F^{\times}$.

Lemma B.2. *We have* $V(\psi') = V(\psi_1) \otimes \psi'' \circ \det$.

Proof. This follows by examining the construction of $W(\psi')$ in [15, Lemma 12]. In the case $c \in F$, the groups H_i defined by Howe are equal to K_i', and the group $GL_l(F'')$ appearing in the proof of [15, Lemma 12] is equal to $GL_3(F)$. Howe constructs $W(\psi')$ by taking the representation $W(\psi_1)$ of K' associated to ψ_1 (which he denotes $W''(\psi_1')$, and whose construction can be assumed as ψ_1 has smaller conductor) and forming the twist $W(\psi_1) \otimes \psi'' \circ \det$. He then obtains $W(\psi')$ by applying the correspondence of [15, Thm 1] to $W(\psi_1) \otimes \psi'' \circ \det$, which in this case is trivial so that $W(\psi') = W(\psi_1) \otimes \psi'' \circ \det$. As $V(\psi')$ and $V(\psi_1)$ are the inductions of $W(\psi')$ and $W(\psi_1)$, the lemma follows.

□

The next lemma shows that it suffices to consider $V(\psi_1)$.

Lemma B.3. *We have* $\dim V(\psi')^{K_n} \leq \dim V(\psi_1)^{K_n}$.

Proof. Because $N(U_i') = U_{\lceil i/e \rceil}$, if a character φ of $F^\times$ has conductor $i+1$, then $\varphi \circ N(F'/F)$ has conductor $ei+1$. Because $\psi'' \circ N$ has conductor $j \geq 2$, this implies that there is some $i \geq 1$ such that $j = ei+1$ and ψ'' has conductor $i+1$.

If $n \geq i+1$ then $\det K_n \subset U_{i+1}$. This implies that $\psi'' \circ \det$ is trivial on K_n, which gives the lemma. Suppose that $n \leq i$ and $V(\psi')^{K_n} \neq 0$. As the central character of $V(\psi')$ is $\psi'|_{F^\times}$, this implies that ψ' is trivial on U_n, and hence on U_i. As ψ_1 is trivial on $U_{j-1}' \cap F = U_i$, this implies that $\psi_2 = \psi'' \circ N(F'/F)$ is trivial on U_i. This implies that ψ'' is trivial on U_i, which contradicts it having conductor $i+1$.

□

By replacing ψ' with ψ_1, multiple times if necessary, we may assume for the rest of the proof that $j = 1$ or $c \notin F$.

B.4. The unramified case. Here, the groups K' and K_j' are equal to K and K_j respectively, and so we omit the $'$ in this section. We also take $\varpi' = \varpi$. The embedding $F' \subset M_3(F)$ has the property that $R' = F' \cap M_3(R)$, so that it induces an embedding $k' \subset M_3(k)$. It also satisfies $R'^\times = F'^\times \cap K$ and $U_i' = F'^\times \cap K_i$.

We need to bound $\dim W(\psi')$, and the number of double cosets $KF'^\times gK_n$ such that $W(\psi')^{gK_ng^{-1} \cap KF'^\times} \neq 0$, and we begin with the second problem. We note that $F'^\times \subset KZ$ in the unramified case, where Z is the center of G, so that $KF'^\times = KZ$. The dimension of $W(\psi')^{gK_ng^{-1} \cap KZ}$ depends only on the double coset $KZgK$. By the Cartan decomposition, we may therefore break the problem into finding those $\lambda \in X_*(T)^+$ such that $ZK\lambda(\varpi)K$ supports invariant vectors, where T is the diagonal torus in G, and then count the (ZK, K_n)-double cosets in a given $ZK\lambda(\varpi)K$. These steps are carried out by Lemmas B.4 and B.5. We write $\lambda \in X_*(T)$ as $(\lambda_1, \lambda_2, \lambda_3)$.

Lemma B.4. *If $\lambda \in X_*(T)^+$ is such that $W(\psi')^{\lambda(\varpi)K_n\lambda(\varpi)^{-1}\cap KZ} \neq 0$, then* $\max\{\lambda_1 - \lambda_2, \lambda_2 - \lambda_3\} \leq n - j$. *(It follows that $j \leq n$.)*

Proof. We will prove $\lambda_1 - \lambda_2 \leq n - j$ by contradiction; the argument for $\lambda_2 - \lambda_3$ is exactly analogous.

First we treat the case $j > 1$. The hypothesis implies that $W(\psi')^{\lambda(\varpi)K_n\lambda(\varpi)^{-1}\cap K_{j-1}} \neq 0$, and we use the description of $W(\psi')$ restricted to K_{j-1}. We identify $K_{j-1}/K_j \simeq M_3(k)$. If $c = b\varpi^{-j}$ with $b \in B'$, and we identify b with an element of $k'^{\times} \subset M_3(k)$, then $\psi|_{K_{j-1}}$ under this identification corresponds to the character of $M_3(k)$ given by $x \to \chi(\mathrm{tr}(bx)/\varpi)$.

If $\lambda_1 - \lambda_2 \geq n - j + 1$, a simple calculation shows that the image of $\lambda(\varpi)K_n\lambda(\varpi)^{-1} \cap K_{j-1}$ in $K_{j-1}/K_j \simeq M_3(k)$ contains the subgroup

$$Y = \begin{pmatrix} & & \\ * & & \\ * & & \end{pmatrix} \subset M_3(k).$$

There must be a character in the orbit $\mathrm{Ad}^*K(\psi)$ which is trivial on Y, which means that there is $k \in K$ such that $\mathrm{tr}(y\mathrm{Ad}(k)b) = 0$ for all $y \in Y$. The annihilator of Y under the trace pairing is

$$Y^{\perp} = \begin{pmatrix} * & & \\ * & * & * \\ * & * & * \end{pmatrix} \subset M_3(k),$$

so that $\mathrm{Ad}(k)b \in Y^{\perp}$. Any $y \in Y^{\perp}$ has eigenvalues that lie in the quadratic extension of k, while the eigenvalues of b lie in $k' - k$ because $c \notin F$, which is a contradiction.

Next we consider the case $j = 1$. In this case, the proof of [15, Lemma 12] states that $W(\psi')$ is inflated from a cuspidal representation of $GL_3(k)$. If $\lambda_1 - \lambda_2 \geq n$, then the image of $\lambda(\varpi)K_n\lambda(\varpi)^{-1} \cap K$ in $K/K_1 \simeq GL_3(k)$ contains the subgroup

$$Y = \begin{pmatrix} 1 & & \\ * & 1 & \\ * & & 1 \end{pmatrix} \subset GL_3(k).$$

However, a cuspidal representation cannot have any vectors invariant under Y, because then it would be a subrepresentation of a representation induced from a parabolic of type (2,1).

□

Lemma B.5. *Let $\lambda \in X_*(T)^+$ satisfy* $\max\{\lambda_1-\lambda_2, \lambda_2-\lambda_3\} \leq n-j$ *as in Lemma B.4. The number of (ZK, K_n)-double cosets in $ZK\lambda(\varpi)K$ is at most* $q^{4n-4j}(1 + 1/q)^3$.

Proof. Any double coset $ZKgK_n$ contained in $ZK\lambda(\varpi)K$ has a representative with $g \in \lambda(\varpi)K$. It may be seen that $\lambda(\varpi)k_1$ and $\lambda(\varpi)k_2$ represent the same double coset if and only if $k_1 \in \lambda(\varpi)^{-1}K\lambda(\varpi)k_2K_n$, and so if we define $K_\lambda = \lambda(\varpi)^{-1}K\lambda(\varpi) \cap K$ then the number of double cosets is equal to $|K_\lambda \backslash K/K_n|$.

If $j = n$ then $\lambda_1 = \lambda_2 = \lambda_3$ so $K_\lambda = K$, hence the lemma is trivial. So we may assume $j \leq n-1$. Then K_λ contains any matrix $g = (g_{a,b}) \in M_3(R)$ such that $v(g_{2,1}) \geq n-j$, $v(g_{3,2}) \geq n-j$, $v(g_{3,1}) \geq 2n-2j$, and $v(g_{i,i}) = 0$ for $1 \leq i \leq 3$ (the last condition ensures that $g \in K$ as g mod ϖ is upper triangular). This implies that the image of K_λ in $K/K_n \simeq GL_3(R/\varpi^n)$ has cardinality at least $q^{5n+4j}(1-1/q)^3$. Therefore

$$\begin{aligned}|K_\lambda \backslash K/K_n| &\leq |K/K_n|/|\mathrm{image}(K_\lambda)| \\ &\leq q^{9n}(1-q^{-3})(1-q^{-2})(1-q^{-1})/q^{5n+4j}(1-1/q)^3 \\ &= q^{4n-4j}(1+q^{-1}+q^{-2})(1+q^{-1}) \leq q^{4n-4j}(1+1/q)^3.\end{aligned}$$

□

Lemma B.4 implies that there are at most n^2 choices of $\lambda \in X_*^+(T)/X_*(Z)$ such that $KZ\lambda(\varpi)K$ supports invariant vectors, and combining this with Lemma B.5 shows that there are at most $n^2q^{4n-4j}(1+1/q)^3$ double cosets $KZgK_n$ that support invariant vectors. This gives

$$\dim V(\psi')^{K_n} \leq n^2q^{4n-4j}(1+1/q)^3 \dim W(\psi').$$

We now bound $\dim W(\psi')$. We first assume that $j \geq 2$. Our assumption that $c \notin F$ implies that the field F'' in [15, Lemma 12] is the same as F', and the groups H_i are given by $H_0 = R'^\times$ and $H_i = U_i'$ for $i \geq 1$. Following the proof of that lemma, we see that $W(\psi')$ is the representation associated to the character ψ' on $R'^\times$ by [15, Thm 1]. When j is even, that theorem implies that $W(\psi')$ is the induction of a character of $R'^\times K_{j/2}$ to K, so that $\dim W(\psi') = |K : R'^\times K_{j/2}|$. We have $|K : R'^\times K_{j/2}| = q^{3j}(1-1/q)(1-1/q^2) \leq q^{3j}$.

When j is odd, we let $j = 2i+1$. The construction of $W(\psi')$ in this case is described on [15, p. 448], and is given by inducing a representation J from $R'^\times K_i$ to K. The discussion on p. 448 implies that J has the same dimension as the two representations denoted $V(\widetilde{\varphi}'')$ and $V(\psi)$ there, and Howe states that $\dim V(\psi) = (\#\widetilde{\mathcal{H}}/\widetilde{\mathcal{Z}})^{1/2}$ for two groups $\widetilde{\mathcal{H}}$ and $\widetilde{\mathcal{Z}}$. Moreover, on p. 447 he states that $\widetilde{\mathcal{H}}/\widetilde{\mathcal{Z}} \simeq \mathcal{H}/\mathcal{Z} \simeq K_i/U_i'K_{i+1}$. As $i \geq 1$, we have $\dim J = |K_i/U_i'K_{i+1}|^{1/2} = q^3$. We then have $\dim W(\psi') = q^3|K : R'^\times K_i| = q^3q^{6i}(1-1/q)(1-1/q^2) \leq q^{6i+3} = q^{3j}$.

In the remaining case $j = 1$, $W(\psi')$ is inflated from a cuspidal representation of $GL_3(k)$. Such a cuspidal representation has dimension $(q^2-1)(q-1)$. Therefore $\dim W(\psi') = (q^2-1)(q-1) \leq q^3 = q^{3j}$.

In all cases we have verified $\dim W(\psi') \leq q^{3j}$. Hence

$$\dim V(\psi')^{K_n} \leq n^2q^{4n-4j}(1+1/q)^3 \cdot q^{3j} \leq n^2q^{4n}(1+1/q)^3.$$

B.5. The ramified case. In this case we must have $j \geq 2$. Moreover, F' is tamely ramified over F (since $p \neq 3$) and generated by a cube root of a uniformizer ϖ of F. Thus we may assume $\varpi'^3 = \varpi$. Choose our basis for R' as a free R-module to be $\{1, \varpi', \varpi'^2\}$. With this choice, the image of ϖ' in G is

$$\varpi' = \begin{pmatrix} & & \varpi \\ 1 & & \\ & 1 & \end{pmatrix}.$$

We see that $\varpi'^i A'$ is given by

$$\varpi'^{3i} A' = \varpi^i \begin{pmatrix} * & \varpi * & \varpi * \\ * & * & \varpi * \\ * & * & * \end{pmatrix}, \tag{B.1}$$

$$\varpi'^{3i+1} A' = \varpi^i \begin{pmatrix} \varpi * & \varpi * & \varpi * \\ * & \varpi * & \varpi * \\ * & * & \varpi * \end{pmatrix}, \tag{B.2}$$

$$\varpi'^{3i+2} A' = \varpi^i \begin{pmatrix} \varpi * & \varpi * & \varpi^2 * \\ \varpi * & \varpi * & \varpi * \\ * & \varpi * & \varpi * \end{pmatrix}, \tag{B.3}$$

where the *'s lie in R. As $K' = A'^{\times}$, K' is the lower triangular Iwahori subgroup.

The proof may be naturally broken into cases depending on the residue class of j modulo 3. We may assume that $j \not\equiv 1$ (3), as in this case we have $c \in F$. Note that we are using our assumption that $\varpi'^3 = \varpi$ here.

As in the unramified case, we begin by observing that it suffices to bound $\dim W(\psi')$ and the number of double cosets $K'F'^{\times} g K_n$ such that $W(\psi')^{gK_n g^{-1} \cap K'F^{\times}} \neq 0$. This condition depends only on $K'F'^{\times} gK$, and the following lemma gives a convenient set of representatives for these double cosets.

Lemma B.6. *If* $\Sigma = \{\lambda \in X_*(T) : \lambda_1 + \lambda_2 + \lambda_3 = 0\}$, *we have* $G = \bigcup_{\lambda \in \Sigma} K'F'^{\times} \lambda(\varpi) K$.

Proof. We use the Bruhat decomposition. Let T^1 and $N(T)$ be the maximal compact subgroup and normalizer of T. We define the Weyl group $W = N(T)/T$ and affine Weyl group $\widetilde{W} = N(T)/T^1$. We identify W with the group of permutation matrices in K, and hence with a subgroup of $\widetilde{W}$. We then have $\widetilde{W} \simeq X_*(T) \rtimes W$, and $\widetilde{W}$ may be identified with matrices of the form $\lambda(\varpi)w$ with $\lambda \in X_*(T)$ and $w \in W$.

We have $\varpi' \in N(T)$, and it may be seen that $\widetilde{W} = \langle \varpi' \rangle \Sigma W$. Indeed, the action of ϖ' on $\widetilde{W}/W \simeq X_*(T)$ by left multiplication is given by

$$\varpi'(\lambda_1, \lambda_2, \lambda_3) = (\lambda_3 + 1, \lambda_1, \lambda_2),$$

so every orbit contains a unique element of Σ. The Bruhat decomposition then gives

$$G = \bigcup_{w \in \widetilde{W}} K'wK' = \bigcup_{\lambda \in \Sigma} K'\langle \varpi' \rangle \lambda(\varpi) W K' = \bigcup_{\lambda \in \Sigma} K' F'^{\times} \lambda(\varpi) K$$

as required.

□

The next lemma bounds those $\lambda \in \Sigma$ such that $K' F'^{\times} \lambda(\varpi) K$ supports invariant vectors.

Lemma B.7. *If $\lambda \in \Sigma$ satisfies $W(\psi')^{\lambda(\varpi) K_n \lambda(\varpi)^{-1} \cap K' F'^{\times}} \neq 0$, then*

$$\max\{\lambda_1 - \lambda_2, \lambda_2 - \lambda_3, \lambda_3 - \lambda_1 + 1\} \leq n - i - 1 \quad \textit{if} \quad j = 3i + 2, \tag{B.4}$$
$$\max\{\lambda_2 - \lambda_1, \lambda_3 - \lambda_2, \lambda_1 - \lambda_3 - 1\} \leq n - i - 1 \quad \textit{if} \quad j = 3i. \tag{B.5}$$

In either case, summing the three bounds gives $j \leq 3n - 2$.

Proof. We may naturally identify K'/K'_1 and K'_{j-1}/K'_j with $(k^\times)^3$ and k^3 using the coordinate entries in such a way that the adjoint action of K'/K'_1 on K'_{j-1}/K'_j is given by

$$\mathrm{Ad}(x_1, x_2, x_3)(y_1, y_2, y_3) = \begin{cases} (x_1 x_2^{-1} y_1, x_2 x_3^{-1} y_2, x_3 x_1^{-1} y_3), & j \equiv 0 \ (3), \\ (x_1^{-1} x_2 y_1, x_2^{-1} x_3 y_2, x_3^{-1} x_1 y_3), & j \equiv 2 \ (3). \end{cases}$$

Moreover, if c is equal to $\varpi'^{-j-2} b$ with $b \in B' \simeq k^\times$, then the character ψ of K'_{j-1}/K'_j is given by $\psi(y_1, y_2, y_3) = \chi(b(y_1 + y_2 + y_3)/\varpi)$. This implies that $\mathrm{Ad}^*(h)\psi$ is nontrivial on every coordinate subgroup in $K'_{j-1}/K'_j \simeq k^3$ for any $h \in K'$.

If $W(\psi')^{\lambda(\varpi) K_n \lambda(\varpi)^{-1} \cap K' F'^{\times}} \neq 0$, then $W(\psi')^{\lambda(\varpi) K_n \lambda(\varpi)^{-1} \cap K'_{j-1}} \neq 0$. Because $W(\psi')|_{K'_{j-1}}$ is a sum of characters of the form $\mathrm{Ad}^*(h)\psi$ with $h \in K'$, one such character must be trivial on $\lambda(\varpi) K_n \lambda(\varpi)^{-1} \cap K'_{j-1}$, which implies that the image of $\lambda(\varpi) K_n \lambda(\varpi)^{-1} \cap K'_{j-1}$ in K'_{j-1}/K'_j does not contain a coordinate subgroup. By combining the definition $K'_{j-1} = 1 + \varpi'^{j-1} A'$ with (B.1)–(B.3), we see that this implies the inequalities (B.4) and (B.5).

□

Lemma B.8. *Let $\lambda \in \Sigma$ satisfy (B.4) or (B.5). The number of $(K'F'^{\times}, K_n)$-double cosets in $K'F'^{\times}\lambda(\varpi)K$ is at most $q^{3n-2i}(1+1/q)^3$ when $j = 3i$, and $q^{3n-2i-2}(1+1/q)^3$ when $j = 3i+2$.*

Proof. Any double coset $K'F'^{\times}gK_n$ contained in $K'F'^{\times}\lambda(\varpi)K$ has a representative of the form $\lambda(\varpi)k$, and two elements $\lambda(\varpi)k_1$ and $\lambda(\varpi)k_2$ represent the same double coset if and only if $k_2 \in \lambda(\varpi)^{-1}K'F'^{\times}\lambda(\varpi)k_1K_n$. Therefore if we define $K'_\lambda = \lambda(\varpi)^{-1}K'\lambda(\varpi) \cap K$, the number of double cosets is bounded by $|K'_\lambda\backslash K/K_n|$. It may be seen that K'_λ contains any matrix $g = (g_{a,b}) \in M_3(R)$ satisfying the conditions

$$\begin{aligned} v(g_{a,b}) &\geq \max\{\lambda_b - \lambda_a + 1, 0\}, \quad a < b \\ v(g_{i,i}) &= 0, \quad 1 \leq i \leq 3 \\ v(g_{a,b}) &\geq \max\{\lambda_b - \lambda_a, 0\}, \quad a > b \end{aligned}$$

on the upper triangular, diagonal, and lower triangular entries respectively. The reader should note that for each pair $a \neq b$, one may order (a,b) so that the inequalities on $v(g_{a,b})$ and $v(g_{b,a})$ have the form $v(g_{a,b}) \geq c$, $v(g_{b,a}) \geq 0$ for some $c > 0$. Moreover, the set of entries for which the inequality above reads $v(g_{a,b}) \geq 0$ form the unipotent radical of a Borel subgroup B_λ containing the diagonal matrices. It follows that K'_λ lies between $B_\lambda \cap GL_3(R)$ and $(B_\lambda \cap GL_3(R))K_1$.

We will divide the proof into six cases depending on the possibilities for B_λ. We treat one case in detail, and describe the modifications to be made in the others.

Case 1: $B_\lambda = \begin{pmatrix} * & * & * \\ & * & * \\ & & * \end{pmatrix}$

In this case, the significant congruence conditions imposed on $g = (g_{a,b}) \in K'_\lambda$ are

$$v(g_{2,1}) \geq \lambda_1 - \lambda_2, \quad v(g_{3,1}) \geq \lambda_1 - \lambda_3, \quad \text{and} \quad v(g_{3,2}) \geq \lambda_2 - \lambda_3. \tag{B.6}$$

The image of K'_λ in K/K_n therefore has cardinality at least $q^{9n+2\lambda_3-2\lambda_1}(1-1/q)^3$, and so as in Lemma B.5 we have $|K'_\lambda\backslash K/K_n| \leq |K/K_n|/|\mathrm{image}(K'_\lambda)| \leq (1+1/q)^3q^{2\lambda_1-2\lambda_3}$. If $j = 3i$ then Lemma B.7 gives $2\lambda_1 - 2\lambda_3 \leq 2n - 2i$ as required. If $j = 3i+2$, Lemma B.7 does not provide a strong enough bound for $\lambda_1 - \lambda_3$, and so we instead observe that the image of K'_λ in K/K_n contains those matrices satisfying

$$v(g_{2,1}) \geq \lambda_1 - \lambda_2, \quad v(g_{3,1}) \geq n, \quad \text{and} \quad v(g_{3,2}) \geq \lambda_2 - \lambda_3, \tag{B.7}$$

with the other conditions unchanged. This group has cardinality at least $q^{8n+\lambda_3-\lambda_1}(1-1/q)^3$, and so $|K'_\lambda\backslash K/K_n| \leq (1+1/q)^3 q^{n+\lambda_1-\lambda_3}$. Lemma B.7 gives

$$\lambda_1 - \lambda_3 = (\lambda_1 - \lambda_2) + (\lambda_2 - \lambda_3) \leq 2n - 2i - 2,$$

which gives the Lemma in this case.

In the other five cases, we may apply the same method to produce a bound of the form $|K'_\lambda\backslash K/K_n| \leq (1+1/q)^3 q^\tau$, where τ depends on the residue class of j modulo 3. We describe the underlying recipe for finding τ in the case above, and then show what it gives in each remaining case. When $j \equiv 0$ (3), we added the right hand sides of (B.6), and the resulting expression $2(\lambda_1 - \lambda_3)$ could be bounded using one application of Lemma B.7, which gave τ. When $j \equiv 2$ (3), we modified the bound in (B.6) corresponding to the non-simple positive root for B_λ to obtain (B.7), added the right hand sides, and bounded the result using two applications of Lemma B.7 to give τ.

We now find τ in the remaining 5 cases, and check that

$$\tau \leq \begin{cases} 3n - 2i - 2, & j = 3i + 2, \\ 3n - 2i, & j = 3i. \end{cases}$$

Note that in some cases we may need to use the assumption that $n \geq 1$, which we are free to make.

Case 2: $B_\lambda = \begin{pmatrix} * & * & * \\ & * & \\ & * & * \end{pmatrix}$

The analog of (B.6) is

$$v(g_{2,1}) \geq \lambda_1 - \lambda_2, \quad v(g_{3,1}) \geq \lambda_1 - \lambda_3, \quad v(g_{2,3}) \geq \lambda_3 - \lambda_2 + 1,$$

which is modified to $v(g_{2,1}) \geq n$. We have

$$\begin{aligned} 2\lambda_1 - 2\lambda_2 + 1 \leq 2n - 2i - 1 = \tau \quad &\text{when} \quad j = 3i + 2, \\ n + \lambda_1 - \lambda_2 + 1 \leq 3n - 2i = \tau \quad &\text{when} \quad j = 3i. \end{aligned}$$

Case 3: $B_\lambda = \begin{pmatrix} * & & * \\ * & * & * \\ & & * \end{pmatrix}$

The analog of (B.6) is

$$v(g_{1,2}) \geq \lambda_2 - \lambda_1 + 1, \quad v(g_{3,1}) \geq \lambda_1 - \lambda_3, \quad v(g_{3,2}) \geq \lambda_2 - \lambda_3,$$

which is modified to $v(g_{3,2}) \geq n$. We have

$$\begin{aligned} 2\lambda_2 - 2\lambda_3 + 1 &\leq 2n - 2i - 1 = \tau \quad \text{when} \quad j = 3i+2, \\ n + \lambda_2 - \lambda_3 + 1 &\leq 3n - 2i = \tau \quad \text{when} \quad j = 3i. \end{aligned}$$

Case 4: $B_\lambda = \begin{pmatrix} * & * & \\ & * & \\ * & * & * \end{pmatrix}$

The analog of (B.6) is

$$v(g_{2,1}) \geq \lambda_1 - \lambda_2, \quad v(g_{1,3}) \geq \lambda_3 - \lambda_1 + 1, \quad v(g_{2,3}) \geq \lambda_3 - \lambda_2 + 1,$$

which is modified to $v(g_{2,3}) \geq n$. We have

$$\begin{aligned} 2\lambda_3 - 2\lambda_2 + 2 &\leq 2n - 2i = \tau \quad \text{when} \quad j = 3i, \\ n + \lambda_3 - \lambda_2 + 1 &\leq 3n - 2i - 2 = \tau \quad \text{when} \quad j = 3i+2. \end{aligned}$$

Case 5: $B_\lambda = \begin{pmatrix} * & & \\ * & * & * \\ * & & * \end{pmatrix}$

The analog of (B.6) is

$$v(g_{1,2}) \geq \lambda_2 - \lambda_1 + 1, \quad v(g_{1,3}) \geq \lambda_3 - \lambda_1 + 1, \quad v(g_{3,2}) \geq \lambda_2 - \lambda_3,$$

which is modified to $v(g_{1,2}) \geq n$. We have

$$\begin{aligned} 2\lambda_2 - 2\lambda_1 + 2 &\leq 2n - 2i = \tau \quad \text{when} \quad j = 3i, \\ n + \lambda_2 - \lambda_1 + 1 &\leq 3n - 2i - 2 = \tau \quad \text{when} \quad j = 3i+2. \end{aligned}$$

Case 6: $B_\lambda = \begin{pmatrix} * & & \\ * & * & \\ * & * & * \end{pmatrix}$

The analog of (B.6) is

$$v(g_{1,2}) \geq \lambda_2 - \lambda_1 + 1, \quad v(g_{1,3}) \geq \lambda_3 - \lambda_1 + 1, \quad v(g_{2,3}) \geq \lambda_3 - \lambda_2 + 1,$$

which is modified to $v(g_{1,3}) \geq n$. We have

$$\begin{aligned} 2\lambda_3 - 2\lambda_1 + 3 &\leq 2n - 2i - 1 = \tau \quad \text{when} \quad j = 3i+2, \\ n + \lambda_3 - \lambda_1 + 2 &\leq 3n - 2i = \tau \quad \text{when} \quad j = 3i. \end{aligned}$$

□

There are at most $9n^2$ choices of $\lambda \in \Sigma$ satisfying the bounds of Lemma B.7. Indeed, if $j = 3i + 2$ then the Lemma gives $n - 1 \geq \lambda_1 - \lambda_2, \lambda_2 - \lambda_3 \geq -2n + 3$, and these two values determine $\lambda \in \Sigma$ uniquely. If $j = 3i$, we have $i \geq 1$ so the Lemma likewise gives $2n - 3 \geq \lambda_1 - \lambda_2, \lambda_2 - \lambda_3 \geq -n + 2$. Moreover, the bound of Lemma B.8 may be written as $q^{3n-j+i}(1 + 1/q)^3$ in either case $j = 3i$ or $j = 3i + 2$. We therefore have at most $9n^2 q^{3n-j+i}(1 + 1/q)^3$ double cosets $K'F'^{\times}gK_n$ that support invariant vectors, and

$$\dim V(\psi')^{K_n} \leq 9n^2 q^{3n-j+i}(1 + 1/q)^3 \dim W(\psi').$$

If j is even, $W(\psi')$ is again obtained by inducing a character from $R'^{\times}K'_{j/2}$ to K', and we have $\dim W(\psi') = |K' : R'^{\times}K'_{j/2}| = (1 - 1/q)^2 q^j$.

If j is odd, set $j = 2l + 1$. As before, Howe defines $W(\psi')$ to be the induction from $R'^{\times}K'_l$ to K' of a representation of dimension $|K'_l : U'_l K'_{l+1}|^{1/2} = q$. This gives $\dim W(\psi') \leq q|K' : R'^{\times}K'_l| = (1 - 1/q)^2 q^{2l+1} = (1 - 1/q)^2 q^j$.

In either case, the bound $\dim W(\psi') \leq q^j$ gives

$$\dim V(\psi')^{K_n} \leq 9n^2 q^{3n-j+i}(1 + 1/q)^3 \cdot q^j = 9n^2 q^{3n+i}(1 + 1/q)^3.$$

If $j = 3i$ then the bound $j \leq 3n - 2$ from Lemma B.7 gives $i \leq n - 1$, while if $j = 3i + 2$ then $j \leq 3n - 2$ gives $i \leq n - 2$. In either case, this completes the proof of Theorem B.1.

References

[1] James Arthur. *The endoscopic classification of representations*, volume 61 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2013. Orthogonal and symplectic groups.

[2] Nicolas Bergeron and Laurent Clozel. Spectre automorphe des variétés hyperboliques et applications topologiques. *Astérisque*, (303):xx+218, 2005.

[3] Nicolas Bergeron, John Millson, and Colette Moeglin. The Hodge conjecture and arithmetic quotients of complex balls. *Acta Math.*, 216(1):1–125, 2016.

[4] J. N. Bernstein. Le "centre" de Bernstein. In *Representations of reductive groups over a local field*, Travaux en Cours, pages 1–32. Hermann, Paris, 1984. Edited by P. Deligne.

[5] I. N. Bernšteĭn. All reductive $\mathfrak{p}$-adic groups are of type I. *Funkcional. Anal. i Priložen.*, 8(2):3–6, 1974.

[6] A. Borel and N. Wallach. *Continuous cohomology, discrete subgroups, and representations of reductive groups*, volume 67 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, second edition, 2000.

[7] Colin J. Bushnell and Philip C. Kutzko. *The admissible dual of* GL(N) *via compact open subgroups*, volume 129 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, NJ, 1993.

[8] Kevin Buzzard and Toby Gee. The conjectural connections between automorphic representations and Galois representations. In *Automorphic forms and Galois representations. Vol. 1*, volume 414 of *London Math. Soc. Lecture Note Ser.*, pages 135–187. Cambridge Univ. Press, Cambridge, 2014.

[9] Ana Caraiani. Local-global compatibility and the action of monodromy on nearby cycles. *Duke Math. J.*, 161(12):2311–2413, 2012.

[10] H. Carayol. Représentations cuspidales du groupe linéaire. *Ann. Sci. École Norm. Sup. (4)*, 17(2):191–225, 1984.

[11] Laurent Clozel. Motifs et formes automorphes: applications du principe de fonctorialité. In *Automorphic forms, Shimura varieties, and L-functions, Vol. I (Ann Arbor, MI, 1988)*, volume 10 of *Perspect. Math.*, pages 77–159. Academic Press, Boston, MA, 1990.

[12] Mathieu Cossutta and Simon Marshall. Theta lifting and cohomology growth in p-adic towers. *Int. Math. Res. Not. IMRN*, (11):2601–2623, 2013.

[13] David L. de George and Nolan R. Wallach. Limit formulas for multiplicities in $L^2(\Gamma\backslash G)$. *Ann. of Math. (2)*, 107(1):133–150, 1978.

[14] Stephen S. Gelbart. *Automorphic forms on adèle groups*. Princeton University Press, Princeton, N.J.; University of Tokyo Press, Tokyo, 1975. Annals of Mathematics Studies, No. 83.

[15] Roger E. Howe. Tamely ramified supercuspidal representations of Gl_n. *Pacific J. Math.*, 73(2):437–460, 1977.

[16] Tasho Kaletha, Alberto Minguez, Sug Woo Shin, and Paul-James White. Endoscopic classification of representations: Inner forms of unitary groups. *arXiv:1409.3731*.

[17] Anthony W. Knapp. *Representation theory of semisimple groups*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 2001. An overview based on examples, Reprint of the 1986 original.

[18] Erez Lapid. On the support of matrix coefficients of supercuspidal representations of the general linear group over a local non-archimedean field. *arXiv:1802.07154*.

[19] Simon Marshall. Endoscopy and cohomology growth on $U(3)$. *Compos. Math.*, 150(6):903–910, 2014.

[20] Simon Marshall. Endoscopy and cohomology growth on a quasi-split $U(4)$. In *Families of Automorphic Forms and the Trace Formula*, Simons Symposia, pages 297–325. Springer International Publishing Switzerland, 2016.

[21] Ralf Meyer and Maarten Solleveld. Characters and growth of admissible representations of reductive p-adic groups. *J. Inst. Math. Jussieu*, 11(2):289–331, 2012.

[22] Colette Moeglin and David Renard. Sur les paquets d'Arthur aux places réels, translation. *arXiv:1704.05096*.

[23] Colette Moeglin and Jean-Loup Waldspurger. *Stabilisation de la formule des traces tordue. Vol. 1*, volume 316 of *Progress in Mathematics*. Birkhäuser/Springer, Cham, 2016.

[24] Colette Moeglin and Jean-Loup Waldspurger. *Stabilisation de la formule des traces tordue. Vol. 2*, volume 317 of *Progress in Mathematics*. Birkhäuser/Springer, Cham, 2016.

[25] Chung Pang Mok. Endoscopic classification of representations of quasi-split unitary groups. *Mem. Amer. Math. Soc.*, 235(1108):vi+248, 2015.

[26] Allen Moy. Local constants and the tame Langlands correspondence. *Amer. J. Math.*, 108(4):863–930, 1986.

[27] Jonathan D. Rogawski. Analytic expression for the number of points mod p. In *The zeta functions of Picard modular surfaces*, pages 65–109. Univ. Montréal, Montreal, QC, 1992.

[28] Peter Sarnak and Xiao Xi Xue. Bounds for multiplicities of automorphic representations. *Duke Math. J.*, 64(1):207–227, 1991.

[29] Gordan Savin. Limit multiplicities of cusp forms. *Invent. Math.*, 95(1):149–159, 1989.

[30] David A. Vogan, Jr. Unitarizability of certain series of representations. *Ann. of Math. (2)*, 120(1):141–187, 1984.

[31] David A. Vogan, Jr. and Gregg J. Zuckerman. Unitary representations with nonzero

cohomology. *Compositio Math.*, 53(1):51–90, 1984.

[32] N. R. Wallach. On the constant term of a square integrable automorphic form. In *Operator algebras and group representations, Vol. II (Neptun, 1980)*, volume 18 of *Monogr. Stud. Math.*, pages 227–237. Pitman, Boston, MA, 1984.

[33] Xiao Xi Xue. On the first Betti numbers of hyperbolic surfaces. *Duke Math. J.*, 64(1):85–110, 1991.