Limit multiplicity for unitary groups
and the stable trace formula

Mathilde Gerbelli-Gauthier

Limit multiplicity for unitary groups and the stable trace formula

Mathilde Gerbelli-Gauthier

We give upper bounds on limit multiplicities of certain nontempered representations of unitary groups $U(a, b)$, conditionally on the endoscopic classification of representations. Our result applies to some cohomological representations, and we give applications to the growth of cohomology of cocompact arithmetic subgroups of unitary groups. The representations considered are transfers of products of characters and discrete series on endoscopic groups, and the bounds are obtained using Arthur's stabilization of the trace formula and the classification established by Mok, and Kaletha, Minguez, Shin and White.

1. Introduction

Let G be a semisimple Lie group and let $\Gamma \subset G$ be an arithmetic lattice. Such a group is an analogue of the "$\mathbb{Z}$-points of G": is realized as the intersection $\mathcal{G}(\mathbb{Q}) \cap K \subset \mathcal{G}(\mathbb{A}_f)$ for a choice of algebraic group $\mathcal{G}/\mathbb{Q}$ such that $\mathcal{G}(\mathbb{R}) = G$ and K a compact-open subgroups of the finite adelic points of G. A natural invariant to study is the group cohomology $H^i(\Gamma, \mathbb{C})$. Yet beyond some low-rank examples, the dimension of this cohomology has only been computed explicitly for specific instances of Γ; for example [11; 21]. A variant of the problem is to study this question *in towers*: one studies the asymptotic properties of $\dim H^i(\Gamma_n, \mathbb{C})$ for sequences Γ_n of nested subgroups as $n \to \infty$; see for example [10; 12; 35].

A central family of such sequences Γ_n are congruence towers $\Gamma(p^n)$. These are obtained by fixing a suitable prime p and considering sequences of subgroups $K(p^n) = K^p K_p(p^n)$. The group K^p is a fixed compact-open subgroup of $G(\mathbb{A}_f^p)$, the finite adelic points away from p, and

$$K_p(p^n) = \mathcal{G}(\mathbb{Q}_p) \cap \{g \in \mathrm{GL}_n(\mathbb{Z}_p) \mid g \equiv I \bmod p_n\}$$

for a choice of embedding $\mathcal{G}(\mathbb{Q}_p) \hookrightarrow \mathrm{GL}_n(\mathbb{Q}_p)$. The resulting nested sequence of subgroups $\Gamma(p^n) = \mathcal{G}(\mathbb{Q}) \cap K(p^n)$ are referred to as principal p-power congruence towers.

This article is motivated by the study of rates of growth of $\dim H^i(\Gamma(p^n), \mathbb{C})$ as n grows, for $\Gamma(p^n)$ cocompact.

These dimensions can be expressed representation-theoretically by using Matsushima's formula [36]:

$$H^i(\Gamma(p^n), \mathbb{C}) = \bigoplus_{\pi} m(\pi, p^n) H^i(\mathfrak{g}, K; \pi).$$

Here the sum is taken over isomorphism classes of unitary representations of G, the number $m(\pi, p^n)$ is the multiplicity of π in the regular representation of G on $L^2(\Gamma(p^n) \backslash G)$, and $H^i(\mathfrak{g}, K; \pi)$ is the $(\mathfrak{g}, K)$-cohomology of π. Following the work of Vogan and Zuckerman [49], the finitely many representations contributing nontrivially to the above sum are well-understood. Thus the question is reduced to the growth of multiplicities $m(\pi, p^n)$ of cohomological representations.

Multiplicity growth rates are best understood for discrete series representations, which contribute to cohomology only in the middle degree. In that case, DeGeorge and Wallach [19] and later Savin [43] have shown that $m(\pi, p^n)$ grows proportionally to the index $[\Gamma(1) : \Gamma(p^n)]$. This leaves open the question of multiplicity growth for cohomological representations in lower degrees. In general, these are nontempered, and DeGeorge and Wallach show that their multiplicities $m(\pi, p^n)$ satisfy

$$m(\pi, p^n)/[\Gamma(1) : \Gamma(p^n)] \xrightarrow[n \to \infty]{} 0.$$

Sarnak and Xue [42] have predicted upper bounds on growth, interpolating between the rate for discrete series and the constant multiplicity of the trivial representation. Here "$f(n) \ll g(n)$" means that for n large enough, $f(n)$ is bounded by a constant multiple of $g(n)$, and $\ll_\epsilon$ indicates that the implied constants depends on ϵ.

Conjecture 1 (Sarnak and Xue). *Let π be a unitary representation of G and let*

$$p(\pi) = \inf\{p \geq 2 \mid \text{ the } K\text{-finite matrix coefficients of } \pi \text{ are in } L^p(G)\}.$$

Then

$$m(\pi, p^n) \ll_\epsilon [\Gamma(1) : \Gamma(p^n)]^{(2/p(\pi)) + \epsilon}.$$

By definition, the representation π is tempered if $p(\pi) = 2$. Thus Sarnak and Xue expect the failure of temperedness to dictate the rate of growth of $m(\pi, p^n)$.

1.1. *Main Theorem.* In this article, we give upper bounds on the multiplicity growth of certain cohomological representations of unitary groups. The results are conditional on the endoscopic classification of representations, as discussed in Section 1.2. Let E/F be a CM extension of number fields with $F \neq \mathbb{Q}$, and $\mathfrak{p}$ a prime of F such that the cardinality $\mathrm{Nm}(\mathfrak{p})$ of the residue field is large enough, see Section 4.2. Let $a \leq N/2$ and let G be a unitary group defined from a Hermitian form of signature $((a, N-a), (N, 0), \ldots, (N, 0))$ relative to E/F. Finally, let $\Gamma(\mathfrak{p}^n)$ be a sequence of principal level

cocompact lattices in G, defined in Section 4. Denote by $\nu(n)$ the unique n-dimensional irreducible representation of $\mathrm{SL}_2(\mathbb{C})$. Our main theorem concerns cohomological representations π of G satisfying the following two conditions:

(i) π belongs to a cohomological Arthur packet associated to a parameter ψ with $\psi\mid_{\mathrm{SL}_2(\mathbb{C})} = \nu(2k) \oplus \nu(1)^{N-2k}$.

(ii) π does not appear in any other Arthur packet.

Such π are endoscopic transfers of products of discrete series with characters, from endoscopic groups of G of the form $U(2k) \times U(N - 2k)$. They exist if $a + k \geq N/2$. For example, they include a family π_k described in Section 5 and which contributes to cohomology in degrees

$$i = i(N, a, k) = \begin{cases} ((N-2)/2)^2 + 2a - k^2 & N \text{ even}, \\ ((N-1)/2)^2 + a - k^2 & N \text{ odd}. \end{cases}$$

We recall that $f(n) \ll g(n)$ means that for n large enough, f is bounded by a constant multiple of g.

Theorem 2. *Assume the endoscopic classification of representations for unitary groups stated in [26]. Let $\Gamma(\mathfrak{p}^n)$ be a tower of principal level cocompact lattices in $G = U(a, N - a)$, such that the size $\mathrm{Nm}(\mathfrak{p})$ of the residue field is large enough. Let $N/2 > k \geq N/2 - a$, and let π be a cohomological representation of G satisfying properties* (i) *and* (ii) *above. Then*

$$m(\pi, \mathfrak{p}^n) \ll \mathrm{Nm}(\mathfrak{p}^n)^{N(N-2k)}.$$

In particular, Conjecture 1 *holds for π.*

Our method of proof leads us to believe that these bounds are sharp, in the sense that one should be able to achieve them for a suitable choice of lattices. Indeed, our strategy is to decompose the multiplicity count and show that the leading term comes from a smaller group for which exact asymptotics are known. We expect that the other terms can be made to oscillate and not contribute in the limit. For $G = U(2, 1)$, this type of method was carried out successfully by Simon Marshall [34].

Our representations do not account for all of the cohomology, but in some low degrees, we expect them to do so asymptotically. For example, the smallest nonzero degree i for which $H^i(\Gamma(\mathfrak{p}^n), \mathbb{C})$ is nontrivial is $i = a$. When N is odd, the representations associated to $k = (N - 1)/2$ contribute asymptotically all the cohomology in degree a, yielding the following bounds.

Corollary 3. *Under the assumptions of* Theorem 2, *assume additionally that N is odd. Then*

$$\dim H^a(\Gamma(\mathfrak{p}^n), \mathbb{C}) \ll \mathrm{Nm}(\mathfrak{p}^n)^N.$$

In order to describe the more general range of degrees in which we predict that the representations we can control contribute all the cohomology, we state our main technical theorem. It concerns bounds on limit multiplicity for representations belonging to a prescribed archimedean Arthur packet. This result does not require that the representations be cohomological, and our most general limit multiplicity result is the following.

Theorem 4. *Under the assumptions of [Theorem 2](), let ψ_∞ be an Arthur parameter with regular infinitesimal character and such that $\psi_\infty \mid_{\mathrm{SL}_2(\mathbb{C})} = \nu(2k) \oplus \nu(1)^{N-2k}$. Let $\pi_\infty \in \Pi_{\psi_\infty}$, and let $\Psi_{\psi_\infty,1}$ be the set of Arthur parameters for G whose specialization at infinity is ψ_∞, and associated to representations with trivial central character. Then*

$$\sum_{\psi \in \Psi_{\psi_\infty,1}} \sum_{\pi = \pi_\infty \otimes \pi_f \in \Pi_\psi} m(\pi) \dim \pi_f^{K(\mathfrak{p}^n)} \ll \mathrm{Nm}(\mathfrak{p}^n)^{N(N-2k)}, \tag{1}$$

where $m(\pi)$ denotes the multiplicity of π in $L^2_{\mathrm{disc}}(G(F)\backslash G(\mathbb{A}_F), 1)$.

These types of Arthur parameters seem to control the growth of certain degrees of cohomology. The combinatorics of intersections between various Arthur packets rapidly get complicated, but here is a sample of behavior we expect.

Conjecture 5. *Let $G = U(N-a, a)$ be as above. Then*:

(i) *The representations belonging to Arthur packets attached to parameters ψ with $\psi \mid_{\mathrm{SL}_2} = \nu(N-\ell) \oplus \nu(1)^\ell$ contribute asymptotically all the cohomology in degrees $a \cdot \ell$ for $0 \leq \ell \leq N - 2(a-1)$.*

(ii) *For these degrees,*

$$\dim H^{a \cdot \ell}(\Gamma(\mathfrak{p}^n), \mathbb{C}) \ll \mathrm{Nm}(\mathfrak{p}^n)^{N \cdot \ell}.$$

The range of degrees to which the conjecture applies is larger for smaller values of a, i.e., when G is farther from being quasisplit. For $a = 1$, Marshall and Shin [35] proved (ii) under some assumptions on $\mathfrak{p}$, and conjectured (i).

1.1.1. *Outline of the proof.* The results are proved in the framework of endoscopy, Arthur parameters, and the stable trace formula. The theorem is a consequence of the endoscopic classification of representations for unitary groups. The classification is a result of Mok [38] if the group G is quasisplit, and of Kaletha, Minguez, Shin and White [26] for inner forms, building on the seminal work of Arthur [5]. It gives a decomposition of the regular representation of $G(\mathbb{A}_F)$ on the discrete spectrum

$$L^2_{\mathrm{disc}}(G(F)\backslash G(\mathbb{A}_F)) \simeq \bigoplus_\psi \bigoplus_{\pi \in \Pi_\psi} m(\pi)\pi$$

where the irreducible summands $\pi = \otimes'_v \pi_v$ are automorphic representations; they appear in the discrete spectrum with multiplicity $m(\pi)$. This decomposition is given in terms of Arthur packets Π_ψ indexed by Arthur parameters ψ. These parameters are formal objects

$$\psi = \boxplus_i (\mu_i \boxtimes \nu(m_i))$$

where each μ_i is a cuspidal automorphic representation of GL_{n_i} and $\nu(m_i)$ is the unique irreducible m_i-dimensional representation of $\mathrm{SL}_2(\mathbb{C})$. Such a ψ is associated to a packet of representations of a unitary group of rank N if $\sum_i n_i m_i = N$ and if ψ is self-dual in a suitable sense. The parameters stand in for homomorphisms

$$\psi : L_F \times \mathrm{SL}_2(\mathbb{C}) \to {}^L G$$

where $^L G$ is the L-group of G and L_F is the Langlands group of F, an object whose existence is at the present moment only hypothetical. Despite Arthur parameters being purely formal objects, one can consider the restriction $\psi \mid_{\mathrm{SL}_2(\mathbb{C})} := \oplus_i \nu(m_i)^{n_i}$ which is an actual finite-dimensional representation. The classification of parameters in terms of this restriction plays a central role in our argument, and we refer to the group $\mathrm{SL}_2(\mathbb{C})$ used to build Arthur parameters as the "Arthur SL_2".

Endoscopy is a specific instance of the principle of functoriality in the Langlands program. It concerns certain groups H, the so-called endoscopic groups of G, and states that if ψ factors through an embedding $^L H \hookrightarrow {}^L G$, then there must be trace identities between the characters of the representations $\pi \in \Pi_\psi$ and those of representations π^H of H in a corresponding packet Π_ψ^H. The character identities are witnessed through the trace formula $I_{\mathrm{disc},\psi}(f)$. In the case of our parameters with regular infinitesimal character, this distribution computes the trace of convolution by a smooth, compactly supported function f on the subspace of L_{disc}^2 spanned by the representations $\pi \in \Pi_\psi$. More specifically, the character identities appear in a decomposition of $I_{\mathrm{disc},\psi}(f)$ referred to as the stabilization of the trace formula (written here in a simplified version for exposition purposes):

$$I_{\mathrm{disc},\psi}(f) = \sum_H S_{\mathrm{disc},\psi}^H(f^H). \tag{2}$$

Here the sum runs over all endoscopic groups H such that ψ factors through $^L H$. The distributions $S_{\mathrm{disc},\psi}^H(f^H)$ are stable, meaning that they satisfy a strengthening of the conjugacy-invariance property of characters of representations.

The summands $S_{\mathrm{disc},\psi}^H(f)$, initially defined inductively, can be expanded explicitly as linear combinations of traces $\mathrm{tr}\,\pi(f)$ of the representations $\pi \in \Pi_\psi$; this is the so-called stable multiplicity formula. We write here a simplified version of the stable multiplicity formula in which we have omitted constants which can be ignored in the asymptotic questions we are concerned with

$$S_{\mathrm{disc},\psi}^H(f^H) = \sum_{\pi \in \Pi_\psi} \xi(\pi, H)\,\mathrm{tr}\,\pi(f). \tag{3}$$

The coefficients $\xi(\pi, H)$ arise from characters of a 2-group $\mathcal{S}_\psi$, the group of connected components of the centralizer of the image of ψ. More precisely, there are two mappings

$$\{\text{representations } \pi \in \Pi_\psi\} \to \{\text{characters of } \mathcal{S}_\psi\},$$

$$\{H \text{ such that } \psi \text{ factors through } {}^L H\} \to \{\text{elements of } \mathcal{S}_\psi\},$$

the second of which is a bijection. In this way, the coefficient $\xi(\pi, H)$ in the decomposition of the stable term $S_{\mathrm{disc},\psi}^H(f^H)$ is the value of the character associated to π on the group element corresponding to H.

In this context, the steps of the proof of Theorem 2 can be outlined as:

(i) (Section 5.2) Determine the parameters ψ associated to the packets containing cohomological representations. This relies on work of Arthur [4] and Adams and Johnson [1].

(ii) (Section 4.2) Write the dimension of cohomology as $\sum_\psi I_{\mathrm{disc},\psi}(f(\mathfrak{p}^n))$ for a specific function $f(\mathfrak{p}^n)$, summing over the parameters ψ computed in the first step.

(iii) (Section 2.6.2) Fix a cohomological parameter ψ. Use the stabilization of the trace formula to decompose

$$I_{\mathrm{disc},\psi}(f(\mathfrak{p}^n)) = \sum_H S^H_{\mathrm{disc},\psi}(f(\mathfrak{p}^n)^H).$$

(iv) (Section 3.2) By interpreting the coefficients $\xi(\pi, H)$ appearing in the stable multiplicity formula (3) as values of characters of $\mathcal{S}_\psi$, conclude that there is a specific endoscopic group H_ψ whose contribution bounds that of all the others in (2), i.e., such that

$$I_{\mathrm{disc},\psi}(f(\mathfrak{p}^n)) \leq K(\psi) S^{H_\psi}_{\mathrm{disc},\psi}(f(\mathfrak{p}^n)^{H_\psi})$$

for a constant $K(\psi)$ computed in terms $\psi|_{\mathrm{SL}_2(\mathbb{C})}$ and of the number of irreducible summands of ψ, and which can be uniformly bounded in terms of the rank N of the unitary group. The group H_ψ depends only on $\psi|_{\mathrm{SL}_2(\mathbb{C})}$. As such it is determined by the parameter ψ_∞ and ultimately by the choice of cohomological representations.

(v) (Sections 3.3 and 4.4) Bound the stable trace $S^{H_\psi}_{\mathrm{disc},\psi}(f(\mathfrak{p}^n)^{H_\psi})$ in terms of the multiplicity $m(\pi^{H_\psi}, \mathfrak{p}^n)$ for a family π^{H_ψ} of representations of H_ψ. This relies on the fundamental lemma, proved by Laumon and Ngô for unitary groups [33], but also on a variant for congruence subgroups due to Ferrari [17]. In order to control the discrepancy between $S^{H_\psi}_{\mathrm{disc},\psi}$ and $I^{H_\psi}_{\mathrm{disc},\psi}$, we make use of the notion of hyperendoscopy, also introduced by Ferrari.

(vi) (Sections 4.3 and 5.3) The representations π^{H_ψ} obtained via steps (i)-(v) from parameters such that $\psi|_{\mathrm{SL}_2(\mathbb{C})} = \nu(2k) \oplus \nu(1)^{N-2k}$ are the product of a discrete series representation and a character. Their limit multiplicity is thus known by results of Savin [43], which gives the desired bounds.

Remark 6. Some comments on possible extensions of the result: the proof exploits the fact that for a global A-parameter ψ, the restriction $\psi|_{\mathrm{SL}_2(\mathbb{C})}$ is determined locally at any place. Here, archimedean restrictions associated to cohomological representations propagate to global and everywhere-local restrictions and induce slow rates of growth. But there is nothing special about infinity: similar methods could provide information about automorphic representations which belong to Arthur packets with large Arthur SL_2.

The endoscopic classification was of course proved by Arthur [5] for quasisplit orthogonal and symplectic groups, and Taïbi [46] extended the key result used here, namely the stable multiplicity formula, to some classes of inner forms. It is likely the case that similar methods could be a good starting point to provide analogous bounds for these groups.

The restrictions on the types of representations we deal with are rooted in restrictions on the Arthur parameters we consider. These have two simple pieces which witness opposite extreme behaviors when restricted to the Arthur SL_2. This allows us to obtain bounds by "applying endoscopy once". To extend the results to e.g., representations associated to global parameters with an arbitrary number of simple

pieces, one could iterate the inductive process of steps (iv) and (v) and bound representations coming from hyperendoscopic groups, i.e., endoscopic groups of endoscopic groups.

Our proof method is in the lineage of a body of recent work applying the framework of endoscopy to growth of cohomology. Most notably, bounds on multiplicity growth of all nontempered cohomological representations were obtained by Marshall [34] for $G = U(2, 1)$, and Marshall and Shin [35] for $G = U(N, 1)$ and a level $\mathfrak{n}$ divisible by primes splitting in the CM extension used to define the unitary group.

1.2. Conditionality. Our results are conditional on the endoscopic classification of representations for inner forms of unitary groups, a result which remains to be fully proved in several ways. As explained in the introduction of [26] and in [38, Section 2.6] the classification depends on upcoming work of Chaudouard and Laumon on the weighted fundamental lemma. It also depends, through its dependency on [5], on several papers of Arthur not yet made public. Moreover, the proof of the classification in [26] is not itself complete: in particular, the results appearing here as Theorems 19 and 23 are only proved for generic parameters. A full proof is expected in [27].

2. *L*-groups, parameters, and the trace formula

2.1. Notation. Let E/F be a CM extension of number fields with Galois group $\Gamma_{E/F}$, algebraic closure $\overline{F}$ and absolute Galois groups Γ_F and Γ_E. We denote places of F and E by v and w respectively, and let $E_v = E \otimes_F F_v$ for v a place of F. Let $F_\infty = F \otimes_\mathbb{Q} \mathbb{R}$, the product of all archimedean completions of F. Let $\mathcal{O}_F$ and $\mathcal{O}_E$ be rings of integers, and $\mathbb{A}_F$ and $\mathbb{A}_E$ be adèle rings, with $\mathrm{Nm} : \mathbb{A}_E \to \mathbb{A}_F$ the norm map. Let $\mathbb{A}_F^f$ be the finite adèles, so that we have $\mathbb{A}_F = F_\infty \times \mathbb{A}_F^f$.

Fix χ_κ for $\kappa \in \{\pm 1\}$, a pair of Hecke characters of E. We fix χ_{+1} to be trivial and the character χ_{-1} is chosen so that its restriction to $\mathbb{A}_F/F^\times$ is the quadratic character associated to E by class field theory.

If F is a field and G/F is a reductive group, we will denote the center of G by Z_G or by $Z(G)$. If F is global, we denote $G(F_v)$ by G_v and $G(F_\infty)$ by G_∞. For $H \subset G(\mathbb{A}_F)$, we use the notation $H_f = H \cap G(\mathbb{A}_F^f)$. The complexified Lie algebra of G_∞ will be denoted $\mathfrak{g}_\infty$.

2.2. Unitary groups and their L-groups.

2.2.1. Quasisplit unitary groups. We now introduce unitary groups and their L-groups, following the exposition of Kaletha, Minguez, Shin and White [26, Section 0]. Let E/F be a quadratic algebra: either the CM extension introduced above or one of its localizations E_v/F_v, in which case we have $E_v \simeq F_v \times F_v$ when v is split. If this is the case, fix an identification $E_v = F_v \times F_v$. Let $\sigma \in \mathrm{Aut}_F(E)$ be the nontrivial element of $\Gamma_{E/F}$ if E is a field, and the involution $\sigma(x, y) = (y, x)$ if $E = F \times F$. If E is a split quadratic algebra, set $\Gamma_E := \Gamma_F$. Let Φ_N be the antidiagonal $N \times N$ matrix

$$\Phi_N = \begin{pmatrix} & & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ (-1)^{N-1} & & \end{pmatrix}. \tag{4}$$

Let $U_{E/F}(N)$ (sometimes denoted $U(N)$) be the reductive group over F with $U_{E/F}(\bar{F}) \simeq \mathrm{GL}_N(\bar{F})$, with Galois action

$$\tau_N(g) = \begin{cases} \tau(g), & \tau \in \Gamma_E, \\ \mathrm{Ad}(\Phi_N)\tau(g)^{-t}, & \tau \in \Gamma_F \setminus \Gamma_E. \end{cases}$$

We have $U_{E/F}(N, E) = \mathrm{GL}_N(E)$, and we can identify

$$U_{E/F}(N, F) = \{g \in \mathrm{GL}_N(E) \mid \mathrm{Ad}(\Phi_N)\sigma(g)^{-t} = g\}, \tag{5}$$

a quasisplit unitary group with maximal (nonsplit) torus given by the group of diagonal matrices, and a Borel subgroup consisting of upper-triangular matrices. If $E = F \times F$, we have $U(N) \simeq \mathrm{GL}_N$ and we fix an isomorphism to identify them.

If F is global, we consider the various localizations of $U(N, F)$. If v splits in E, we have $U(N, F_v) \simeq \mathrm{GL}_N(F_v)$. Otherwise $U(N, F_v)$ is a quasisplit unitary group over F_v, which determines it uniquely up to isomorphism as we shall see below.

2.2.2. *Inner forms.* An inner form of $U(N)$ is a pair (G, ξ) consisting of an algebraic group G/F together with an isomorphism $\xi : G(\bar{F}) \to U(N, \bar{F})$ such that for all $\sigma \in \Gamma_F$, the automorphism $\xi^{-1} \circ \sigma \circ \xi \circ \sigma^{-1}$ is inner. Though the choice of ξ is always present, it will sometimes be implicit in our notation. We will denote $U(N)$ by G^* when we want to highlight that it is the quasisplit form of G.

Remark 7. In this article, we always require that the inner forms be groups defined with respect to a Hermitian space over E.

We now discuss which possible groups G can arise as inner forms of $U_{E/F}(N)$ in the cases where F is local or global.

2.2.3. *Local inner forms and the Kottwitz sign.* If v is archimedean, the classification of inner forms is well-known: a unitary group over $F_v = \mathbb{R}$ is determined by its signature $p + q = N$, with $U(p, q) \simeq U(q, p)$. Since the notation $U(N)$ is reserved for quasisplit groups, we denote the compact inner form of $U(N, \mathbb{R})$ by $U_N(\mathbb{R})$.

For v nonarchimedean, the classification of unitary groups coming from Hermitian forms over quadratic algebras over F_v is due to Landherr [31]: If N is odd, there is one class of Hermitian forms up to isomorphism, so the group $U(N, F_v)$ is the unique unitary group of rank N. If N is even, there are two isomorphism classes of unitary groups, only one of which (the one containing $U(N, F_v)$) is quasisplit.

One associates to an inner form G_v of $U_{E_v/F_v}(N)$ a Kottwitz sign $e(G_v)$. We record the formulas for $e(G_v)$ as computed in [29]:

- For $F_v = \mathbb{R}$, let $q(G_v)$ be half the dimension of the symmetric space associated to the group G_v. Then $e(G_v) = (-1)^{q(G_v)-q(G_v^*)}$.

- For F_v nonarchimedean, let $r(G_v)$ be the rank of G_v. Then $e(G_v) = (-1)^{r(G_v)-r(G_v^*)}$.

Kottwitz proves [29] that for G defined over a global field, the local signs cancel out and $\prod_v e(G_v) = 1$.

2.2.4. *Global inner forms.* We describe the classification of global forms of unitary groups, following the discussion in Section 0.3.3 of [26]. For N odd, any collection of local inner twists, quasisplit at all but finitely many places, can be realized as the localization of a global inner twist.

Fo N even, the behavior of the place v in E determines cohomological invariants attached to G_v. For each v, we have $H^1(\Gamma_{F_v}, G_v^{*,\mathrm{ad}}) \simeq \mathbb{Z}/2\mathbb{Z}$. If v is split in E, the invariant of G_v depends on the division algebra D_v such that $G_v = \mathrm{Res}_{F_v}^{D_v} \mathrm{GL}_{M_v}$. Since we only consider unitary groups coming from Hermitian forms, this invariant is always 0 for us. At finite nonsplit places, the quasisplit group $U(N)_v$ and its unique inner form correspond respectively to 0 and 1 in $\mathbb{Z}/2\mathbb{Z}$. At the infinite places, the signature (p, q) determines the invariant $N/2 + q \in \mathbb{Z}/2\mathbb{Z}$. For a collection of local G_v to come from a global unitary group, the all but finitely many nonzero invariants associated to G_v must sum up to zero. Consequently, we have:

Lemma 8. *Let E/F be a CM extension of number fields. There exists an inner form G of $U_{E/F}(N)$ with any choice of signature at the infinite places. Moreover G can be chosen to be quasisplit outside of a set of places of size at most 1.*

Remark 9. The authors of [26] work with a refinement of the notion of inner form. Recall that isomorphism classes of inner forms of G are in bijection with $H^1(\Gamma_F, G^{\mathrm{ad}})$. In addition to this, they introduce the notion of pure inner form, a triple consisting of G, the map ξ, and a cocycle $z \in Z^1(\Gamma_F, G)$ compatible with the inner twist ξ. The map sending a pure inner form to z induces a bijection between isomorphism classes of pure inner forms and $H^1(\Gamma_F, G)$. Inner forms of $U(N)$ which can be realized as pure inner forms are those coming from a Hermitian space, i.e., precisely the groups we work with. We will point out dependency on z whenever it appears: in the normalization of transfer factors, and the pairings associated to local Arthur packets. Due to our rather rudimentary use of the stable trace formula, the choice of pure inner form does not affect our results.

2.2.5. *L-groups.* Throughout, we will work with the Weil group version of the L-group, primarily because it is well-suited to our description of local parameters. In terms of the actual definition of the L-group, this choice is purely cosmetic as the Galois actions involved will always factor through a quotient of order at most 2.

For G/F with F either local or global, fix a root datum. The L-group of G is a semidirect product

$$^L G = \hat{G} \rtimes W_F.$$

Here the group $\hat{G}$ is the complex dual group of G, and the action of W_F on $\hat{G}$ is induced by the Galois action on the root datum of G. As a consequence, if G is split then $^L G = \hat{G} \times W_F$, and in particular, $^L \mathrm{GL}_N(F) = \mathrm{GL}_N(\mathbb{C}) \times W_F$. If G'/F is an inner form of G then by definition $G'(\bar{F}) \simeq G(\bar{F})$ and the corresponding Galois actions differ by an inner automorphism. These induce isomorphisms of root data and Galois actions, and $^L G \simeq {}^L G'$.

For F global, we will sometimes abuse notation and denote by ${}^L G_v$ the L-group of the base change of G to a completion F_v. In this situation, the embedding $W_{F_v} \to W_F$ induces a map ${}^L G_v \to {}^L G$ which restricts to the identity on $\hat{G}$.

The L-group of $U(N)$ (and of any inner form) is defined as

$$
{}^L U(N) = \mathrm{GL}_N(\mathbb{C}) \rtimes W_F
$$

where W_F acts through the order two quotient $\Gamma_{E/F}$. The nontrivial element $\sigma \in \Gamma_{E/F}$ acts by $\sigma(g) = \Phi_N^{-1} g^{-t} \Phi_N$ of GL_N, where Φ_N is as in (4).

2.2.6. *Morphisms of L-groups.* A morphism of L-groups, or L-morphism, is a continuous morphism

$$
\eta : {}^L H \to {}^L G
$$

which commutes with the projections onto W_F. In practice, all morphisms of L-groups considered here will be admissible, i.e., induced by an algebraic map $\hat{H} \to \hat{G}$ and such that the image of elements of W_F are semisimple.

Denote the Weil restriction $\mathrm{Res}_F^E \mathrm{GL}_N$ by $G(N)$. In particular, $G(N)(F) = \mathrm{GL}_N(E)$ and $G(N)(E) \simeq \mathrm{GL}_N(E) \times \mathrm{GL}_N(E)$. As such, the connected component $\widehat{G(N)}$ of ${}^L G(N)$ is the product of two copies of $\mathrm{GL}_N(\mathbb{C})$, and W_F acts through $\Gamma_{E/F}$ via the automorphism that interchanges the two factors. Many objects associated to a unitary group $U(N)$ depend on a choice of embedding of L-groups from ${}^L U(N)$ to ${}^L G(N)$.

To define the L-embedding $\eta_\kappa : {}^L U(N) \to {}^L G(N)$, recall the characters χ_κ from Section 2.1. If F is global, we will use these characters, and if $F = F_v$ is local, we will momentarily also denote by χ_κ the restriction of χ_κ to $E_v^\times$. Let I_N be the identity matrix. For each $\kappa \in \{\pm 1\}$ we define η_κ as

$$
\begin{aligned}
\eta_\kappa(g \rtimes 1) &= (g, {}^t g^{-1}) \rtimes 1, & g &\in \hat{G}, \\
\eta_\kappa(I_N \rtimes \sigma) &= (\chi_\kappa(\sigma) I_N, \chi_\kappa^{-1}(\sigma) I_N) \rtimes \sigma, & \sigma &\in W_E, \\
\eta_\kappa(I_N \rtimes w_c) &= (\kappa \Phi_N, \Phi_N^{-1}) \rtimes w_c.
\end{aligned}
\tag{6}
$$

We consider a second class of L-embeddings $\xi_{\underline{\kappa}} : {}^L(U(N_1) \times \cdots \times U(N_r)) \to {}^L U(N)$, for $\sum N_i = N$ into ${}^L U(N)$. Put $\kappa_i = (-1)^{N - N_i}$ for each i, and let $\underline{\kappa} = (\kappa_1, \ldots, \kappa_r)$. Given $\underline{\chi}$ with signature $\underline{\kappa}$, and for a choice of w_c as above, the embedding $\xi_{\underline{\kappa}}$ is defined as

$$
\begin{aligned}
\xi_{\underline{\kappa}}(g_1, \ldots, g_r \rtimes 1) &= \mathrm{diag}(g_1, \ldots, g_r) \rtimes 1, & g_i &\in \mathrm{GL}_{N_i}(\mathbb{C}), \\
\xi_{\underline{\kappa}}(I_{N_1}, \ldots, I_{N_r} \rtimes \sigma) &= \mathrm{diag}(\chi_{\kappa_1}(\sigma) I_{N_1}, \ldots, \chi_{\kappa_r}(\sigma) I_{N_r}) \rtimes \sigma, & \sigma &\in W_E, \\
\xi_{\underline{\kappa}}(I_{N_1}, \ldots, I_{N_r} \rtimes w_c) &= \mathrm{diag}(\kappa_1 \Phi_{N_1}, \ldots, \kappa_r \Phi_{N_r}) \cdot \Phi_N^{-1} \rtimes w_c.
\end{aligned}
\tag{7}
$$

Note that the composite embedding $\eta_\kappa \circ \xi_{\underline{\kappa}}$ gives an embedding

$$
\eta_{\kappa \cdot \underline{\kappa}} : {}^L(U(N_1) \times \cdots \times U(N_r)) \to {}^L G(N)
\tag{8}
$$

with signature $\kappa \cdot \underline{\kappa} = (\kappa \kappa_1, \ldots, \kappa \kappa_r)$.

Remark 10. The need to consider several embeddings depending on $\underline{\kappa}$ stems from the possibility that parameters for the pair $(U(N), \eta_+)$ may factor through different embeddings of the products of groups $U(N_i)$ associated to different signs.

2.3. *Parameters.*

We introduce the discrete automorphic spectrum of a unitary group G, and the local and global parameters which will classify the (constituents of) automorphic representations, following [5], [26], and [38].

2.3.1. *Automorphic representations.*

Let G/F be a reductive group. Fix a closed subgroup $\mathfrak{X} \subset Z_G(\mathbb{A}_F)$ and a maximal compact subgroup of K of $G(\mathbb{A}_F)$, which in turn determines maximal compact subgroups K_v of $G_v = G(F_v)$ for any v. We consider the right-regular representation of $G(\mathbb{A}_F)$ on

$$L^2_{\text{disc}}(G(F)\backslash G(\mathbb{A}_F), \omega),$$

the discrete part of the space of square-integrable functions which transform by ω under the action of $\mathfrak{X}$. We will omit the ω when we allow for any central character. In our initial cases of interest, G/F will be an anisotropic inner form of $U_{E/F}(N)$, the group $\mathfrak{X}$ will be the full center, the central character ω will be trivial, and the entire automorphic spectrum will be discrete. However for induction purposes we will consider arbitrary central character data $(\mathfrak{X}, \omega)$ and allow for $L^2(G(F)\backslash G(\mathbb{A}_F), \omega)$ to have a continuous part. The discrete spectrum decomposes as

$$L^2_{\text{disc}}(G(F)\backslash G(\mathbb{A}_F)) = \bigoplus m(\pi)\pi$$

where $m(\pi)$ denotes the multiplicity of π, and the irreducible constituents are automorphic representations. Each π is a restricted tensor product $\pi = \otimes'_v \pi_v$ with each π_v an irreducible admissible unitary representation of G_v. All but finitely many π_v are spherical with respect to K_v. The representation π_v is said to be tempered if its K_v-finite matrix coefficients belong to $L^{2+\epsilon}(G_v)$ for all $\epsilon > 0$.

After fixing a maximal compact subgroup K_∞ of G_∞, we replace π_∞ by the dense subspace of K_∞-finite smooth vectors, which we view as an admissible $(\mathfrak{g}_\infty, K_\infty)$-module. This is no loss of information since unitary admissible representations are determined by their underlying $(\mathfrak{g}_\infty, K_\infty)$-modules [28, 9.2].

2.3.2. *Local Langlands parameters.*

Let F be a local field with Weil group W_F. The Langlands group L_F of F is defined as

$$L_F := \begin{cases} W_F & F \text{ is archimedean,} \\ W_F \times \mathrm{SU}(2, \mathbb{R}) & F \text{ is nonarchimedean.} \end{cases}$$

A local Langlands parameter for the reductive group G/F is a continuous homomorphism $\varphi : L_F \to {}^L G$ satisfying certain conditions (see [9] for a discussion):

(i) The map φ commutes with the projections $L_F \to W_F$ and ${}^L G \to W_F$.

(ii) In the nonarchimedean case, the restriction $\varphi\,|_{\mathrm{SU}(2,\mathbb{C})}$ is algebraic.

(iii) The image of W_F under φ consists of semisimple elements of ${}^L G$.

(iv) If the image of φ in $\hat{G}$ factors through a parabolic subgroup of $\hat{G}$, then this parabolic subgroup must be the dual $\hat{P}$ of a parabolic subgroup P of G.

Continuous maps that satisfy condition (i) are known as L-homomorphisms. If they additionally satisfy (ii)–(iii) they are called *admissible*. If they satisfy (iv), they are called *relevant*, or *G-relevant*. Finally, we say that φ is *bounded* if W_F has bounded image in $\hat{G}$. We will denote the collection of $\hat{G}$-conjugacy classes of Langlands parameters for G by $\Phi(G)$.

2.3.3. *Local Arthur parameters.* In order to describe the nontempered spectrum of G, we consider enhancements of Langlands parameters known as Arthur parameters. These are admissible L-homomorphisms

$$\psi : L_F \times \mathrm{SL}_2(\mathbb{C}) \to {}^L G$$

such that $\psi\,|_{L_F}$ is bounded. We denote the set of $\hat{G}$-conjugacy classes of Arthur parameters by $\Psi(G)$. We refer to the $\mathrm{SL}_2(\mathbb{C})$ factor in the above product as the "Arthur SL_2", and say that ψ is bounded if it restricts trivially to the Arthur SL_2.

Each Arthur parameter ψ determines a Langlands parameter φ_ψ as follows. Recall (e.g., [47]) that the Weil group W_F is naturally equipped with a norm homomorphism $|\cdot|$ to $\mathbb{C}^\times$. Then φ_ψ is defined as the composition

$$\varphi_\psi : W_F \to {}^L G, \quad \varphi_\psi(\sigma) = \psi\left(\sigma, \begin{pmatrix} |\sigma|^{1/2} & 0 \\ 0 & |\sigma|^{-1/2} \end{pmatrix}\right).$$

We now give a more detailed description of local Arthur parameters in the case where $G = U(N)$, following Section 2.2 of Mok [38]. Specifically, we use the map η_κ introduced in Section 2.2.6 to realize $\Psi(U(N))$ as a set of N-dimensional representations satisfying an appropriate self-duality condition.

We first describe a natural bijection between $\Psi(G(N))$, and $\Psi(\mathrm{GL}_N(E))$. To produce an element of $\Psi(G(N))$, one starts with $\psi \in \Psi(\mathrm{GL}_N(E))$, i.e., an admissible N-dimensional representation of $L_E \times \mathrm{SL}_2(\mathbb{C})$, and promotes it to a L-morphism $\psi' : L_F \times \mathrm{SL}_2(\mathbb{C}) \to {}^L G(N)$ by choosing $w_c \in W_F \setminus W_E$ and defining

$$\psi'(\sigma, g) = (\psi(\sigma, g), \psi^c(\sigma, g)) \rtimes \sigma, \quad (\sigma, g) \in L_E \times \mathrm{SL}_2(\mathbb{C})$$
$$\psi'(w_c) = (\psi(w_c^2), I_N) \rtimes w_c,$$

where $\psi^c(\sigma, g) = \psi(w_c^{-1}\sigma w_c, g)$. The resulting bijection $\Psi(G(N)) \simeq \Psi(\mathrm{GL}_N(E))$ is independent of the choice of w_c. Moreover, if $\psi^c \simeq \psi^\vee$ where $\psi^\vee$ is the contragredient of ψ, then ψ is called *conjugate self-dual*. More precisely, the parameter ψ is conjugate self-dual of parity ± 1, depending on the parity of the resulting bilinear form.

The map η_κ introduced in (6) then induces a mapping

$$\eta_{\kappa*} : \Psi(U(N)) \to \Psi(G(N)) \simeq \Psi(\mathrm{GL}_N(E)) \tag{9}$$

which is shown by Mok, following work of Gan, Gross and Prasad [18], to be an injection whose image consists precisely of the subset of $\Psi(\mathrm{GL}_N(E))$ of conjugate self-dual representations of parity $(-1)^{N+1}\kappa$.

2.3.4. *Global Arthur parameters.* In lieu of global parameters, Arthur [5, Section 1.4] introduces formal objects realized by combining cuspidal automorphic representations of GL_N and representations of the Arthur SL_2. Echoing the local discussion, global Arthur parameters are first defined in terms of $G(N)$, and Arthur parameters for $U(N)$ are the ones factoring through a fixed embedding of L-groups.

A global Arthur parameter for GL_N is an unordered sum

$$\psi^N = \boxplus_i \psi_i^{N_i}, \quad \psi_i^{N_i} = \mu_i \boxtimes \nu(m_i).$$

Here μ_i is a cuspidal automorphic representation of $\mathrm{GL}_{n_i}(\mathbb{A}_E)$ and $\nu(m_i)$ is the irreducible m_i-dimensional representation of $\mathrm{SL}_2(\mathbb{C})$, with $m_i n_i = N_i$ and $\sum_i N_i = N$. Departing from our references, we immediately restrict our attention to the set of Arthur parameters such that the $\psi_i^{N_i}$ are pairwise distinct: we denote this set $\Psi(N)$ instead of $\Psi_{\mathrm{ell}}(N)$. The collection $\Psi(N)$ contains a distinguished subset $\Psi_{\mathrm{sim}}(N)$ of simple parameters with a unique summand ψ^N. Following the theorem of Mœglin and Waldspurger [37], this subset $\Psi_{\mathrm{sim}}(N)$ parametrizes the discrete spectrum of GL_N.

We now give the construction of global Arthur parameters for a quasisplit unitary group $G = U(N)$, following Section 1.3.4 of [26]. We start by restricting our attention to the set $\tilde{\Psi}(N) \subset \Psi(N)$ of parameters for which each of the μ_i is *conjugate self-dual*, i.e., satisfies $\mu_i = \overline{\mu_i}^\vee$ where $\overline{\mu} = \mu \circ \sigma$ and $\sigma \in \Gamma_{E/F}$ is nontrivial.

To record the parameter in relation to the embedding η_κ, we introduce the group $\mathcal{L}_\psi$. If ψ^N decomposes as a sum of $\mu_i \boxtimes \nu(m_i)$, we associate to each index a pair $(U_{E/F}(n_i), \eta_{\kappa_i})$ as in Section 2.2.6. Here the choice of sign κ_i is determined by μ_i. Then $\mathcal{L}_\psi$ is the fiber product $\mathcal{L}_\psi = \prod_i ({}^L U_{E/F}(n_i) \to W_F)$. There is a natural map $\tilde{\psi}^N : \mathcal{L}_\psi \times \mathrm{SL}_2(\mathbb{C}) \to {}^L G(N)$ given by the direct sum

$$\tilde{\psi}^N = \oplus(\eta_{\kappa_i} \otimes \nu(m_i)).$$

A global Arthur parameter for $(U_{E/F}(N), \eta_\kappa)$ is then defined as a pair $\psi = (\psi^N, \tilde{\psi})$ where $\psi^N \in \tilde{\Psi}(N)$, and

$$\tilde{\psi} : \mathcal{L}_\psi \times \mathrm{SL}_2(\mathbb{C}) \to {}^L U_{E/F}(N)$$

is an L-homomorphism such that $\eta_\kappa \circ \tilde{\psi} = \tilde{\psi}^N$. It is useful to remember that ψ^N encodes the arithmetic information of the automorphic representations of GL_{n_i}, and that $\tilde{\psi}$ is an actual homomorphism. As such, we can (and will) discuss the centralizer of the image of $\tilde{\psi}$. Two Arthur parameters are equivalent if the $\tilde{\psi}$ are $\hat{U}(N)$-conjugate, and we denote the set of equivalence classes of ψ as above by $\Psi(U(N), \eta_\kappa)$. Note that we have again broken off from our references in the choice of notation: our set $\Psi(U(N), \eta_\kappa)$ is the one that the authors of [26] denote $\Psi_2(U_{E/F}(N), \eta_\kappa)$. Finally, note that the map $\eta_{\kappa,*}$ sending ψ to ψ^N is an injection: this allows us to view $\Psi(U(N), \eta_\kappa)$ as a subset of $\Psi(N)$. If $(G, \xi_{\underline{\kappa}})$ is a product as in (8), we can similarly define $\Psi(G, \xi_{\underline{\kappa}})$. Via the block-diagonal embedding $\prod_i \mathrm{GL}_{N_i} \hookrightarrow \mathrm{GL}_N$, we can identify $\Psi(G, \xi_{\underline{\kappa}}) \simeq \prod_i \Psi(U(N_i), \eta_{\kappa_i})$.

Remark 11. We have made two constraints on the set of parameters under consideration here which bear highlighting. We require:

(i) That the irreducible summands ψ_i be pairwise distinct. In Mok's description of the parameters in [38, Section 2.4] this amounts to requiring that all the $l_i = 1$.

(ii) That each irreducible summand be conjugate self-dual. This is stricter than requiring ψ to be conjugate self-dual since we could have had $\mu_i^\vee \simeq \mu_j$.

Parameters satisfying these conditions are called *elliptic*. These restrictions will give us control on the group $\mathcal{S}_\psi$ to be introduced below, whose characters determine which products of local representations occur in the discrete spectrum. It is also the case that only the parameters in the set which we denote by $\Psi(U(N), \eta_\kappa)$ correspond to packets whose members actually appear in the decomposition of L^2_{disc}, although this fact is far from obvious and is one of the main theorems in [38] and [26]. Following this result, global elliptic parameters are also called *square-integrable*.

2.3.5. *Localization.* We now describe how a global Arthur parameter $\psi \in \Psi(U(N), \eta_\kappa)$ gives rise to local Arthur parameters ψ_v at each place v. Each cuspidal representation μ of GL_N factors as a restricted tensor product $\mu = \otimes' \mu_v$ over places v of F. The μ_v are admissible representations of $\mathrm{GL}_N(F_v)$. The local Langlands correspondence for GL_N [23; 24; 44] associates to μ_v a parameter $\varphi_{\mu_v} \in \Phi(\mathrm{GL}_N)$. Following [5], we define the localization of ψ at v as the direct sum

$$\psi_v = \oplus_i \psi_{v,i}, \quad \psi_{v,i} = \varphi_{\mu_{v,i}} \otimes \nu(m_i).$$

These localizations a priori only belong to $\Psi(G(N))$. The fact that they are in the image of the map (9) is one of the central theorems of [38].

2.3.6. *Parameters of inner forms.* Let (G, ξ) be an inner form of $G^* = U(N)$. A local Arthur parameter for G is simply a G-relevant parameter for $U(N)$, see Section 2.3.2. Globally, a parameter $\psi \in \Psi(G^*, \eta_\kappa)$ is G-relevant if it is so everywhere locally [26, Section 1.3.7]. We denote by $\Psi(G, \xi)$ the collection of G-relevant parameters in $\Psi(G^*, \eta_\kappa)$. In summary, we have the following chain of inclusions:

$$\Psi(G, \xi) \subset \Psi(G^*, \eta_\kappa) \subset \tilde{\Psi}(N) \subset \Psi(N),$$

where the parameters in $\tilde{\Psi}(N)$ are conjugate self-dual, those in $\Psi(G^*, \eta_\kappa)$ factor through the embedding η_κ, and those in $\Psi(G, \xi)$ are additionally G-relevant.

2.3.7. *Parameters and conjugacy classes.* We attach families of conjugacy classes to objects introduced above, following [5, Section 1.3]. For F global, G reductive, and any finite set S of places of F containing the archimedean ones, let $\mathcal{C}^S(G)$ denote the set of collections $c = \{c_v\}_{v \notin S}$, where each c_v is a semisimple conjugacy class in $\hat{G}$. For two sets S and S', let $c \sim c'$ if $c_v = c'_v$ for almost all v. Denote the set of such equivalence classes by $\mathcal{C}(G)$. As we did for parameters, let $\mathcal{C}(N) := \mathcal{C}(GL_N)$. We associate elements of $\mathcal{C}(G)$ to automorphic representations π of G. Factoring $\pi = \otimes'_v \pi_v$, let $c(\pi) = \{c(\pi_v)\} \in \mathcal{C}(G)$ be the Satake parameters of all the unramified π_v. Note also that an L-embedding $\eta : {}^L G \to {}^L G(N)$ such as those introduced in Section 2.2.6 gives rise to a map $\eta_* : \mathcal{C}(G) \to \mathcal{C}(N)$.

When $G = \mathrm{GL}_N$ one can associate an element of $\mathcal{C}(N)$ to each $\psi \in \Psi(N)$. Starting with simple parameters $\psi \in \Psi_{\mathrm{sim}}(N)$, use the recipe for the representation π_ψ prescribed by Moeglin and Waldspurger's

theorem [37] and let $c(\psi) := c(\pi_\psi)$. If ψ is not simple, apply the process to its simple constituents and associate to ψ the conjugacy class coming from the diagonally embedded product of the GL_{N_i} inside of GL_N. This produces a mapping

$$\Psi(N) \to \mathcal{C}(N), \quad \psi \mapsto c(\psi)$$

which is injective, following Jacquet and Shalika [25]. Denote its image by $\mathcal{C}_{\mathrm{aut}}(N)$.

2.3.8. *Stabilizers and quotients.* For ψ either local or global, we have

$$S_\psi := \mathrm{Cent}(\mathrm{Im}(\psi), \hat{G}), \quad \bar{S}_\psi := S_\psi/Z(\hat{G})^{W_F}, \quad \mathcal{S}_\psi := \pi_0(\bar{S}_\psi).$$

As mentioned previously, when ψ is global then $\mathrm{Im}(\psi)$ really means $\mathrm{Im}(\tilde{\psi})$. Localization of parameters $\psi \mapsto \psi_v$ induces a mapping $\mathcal{S}_\psi \to \mathcal{S}_{\psi_v}$. When G is unitary, the groups $\mathcal{S}_\psi$ can be readily computed, as the four authors do in [26, page 63]. In particular, for F global and $\psi \in \Psi(G^*, \eta_\kappa)$ decomposing as $\psi = \boxplus_{i=1}^r \psi_i$, we have

$$\mathcal{S}_\psi = (\mathbb{Z}/2\mathbb{Z})^{r-1}. \tag{10}$$

The reader who looks at the computations in [26] will notice that this is the point where we use the assumptions from Remark 11. Finally, we introduce the element

$$s_\psi := \psi\left(1, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}\right) \in S_\psi. \tag{11}$$

We will sometimes conflate s_ψ and its image in the quotient $\mathcal{S}_\psi$.

Remark 12. The authors of [26] work with the centralizer quotient $S_\psi^\natural$, which agrees with $\mathcal{S}_\psi$ for G local and unitary. If the local group G_v is isomorphic to $\mathrm{GL}_{N,v}$ (the only possibility for us at split places, since our unitary groups arise from Hermitian forms) then $S_\psi^\natural \simeq \mathbb{C}^\times$. However, if $G_v = \mathrm{GL}_{N,v}$, then only the trivial character of $S_\psi^\natural$ arises in the character identities, as will be discussed in Section 2.5.6. Thus there is no loss in working instead with the group $\mathcal{S}_\psi = \{1\}$. In the global situation, the characters of $S_\psi^\natural$ that arise all factor through $\mathcal{S}_\psi$ [26, page 89]. Note that we follow Arthur's convention and use the notation $\mathcal{S}_\psi$ instead of $\bar{\mathcal{S}}_\psi$ as in [26].

2.3.9. *Epsilon factors.* The last invariant attached to a global parameter ψ is the character ϵ_ψ of $\mathcal{S}_\psi$, defined by Arthur in [5, Section 1.5]. The definition involves the symplectic root number $\epsilon(1/2, \mu_\alpha)$ of an automorphic L-function $L(s, \mu_\alpha)$ for a product of general linear groups, obtained by composing ψ with the adjoint representation. As such ϵ_ψ encodes arithmetic data in the decomposition of $L^2_{\mathrm{disc}}(G(F)\backslash G(\mathbb{A}_F))$. Note that ϵ_ψ only depends on ψ and in particular is independent of the inner form of G^* under consideration, as discussed in [26, page 89].

2.4. *Endoscopic data.* An endoscopic datum for G/F is a triple (ξ, H, s) where:

- s is a semisimple element of $\hat{G}$.
- H/F is a connected, quasisplit group.
- $\xi : {}^L H \to {}^L G$ is an L-embedding.

The triple must satisfy certain conditions, see [26, Section 1.1.1], including that $\xi(\hat{H})$ is the connected component of the centralizer of s in $\hat{G}$. We will work only with *elliptic* endoscopic data, characterized by the requirement that $\xi(Z(\hat{H})^{W_F})^0 \subset (Z(\hat{G}))^{W_F}$. As such, we denote the set of conjugacy classes of elliptic endoscopic data for G by $\mathcal{E}(G)$, dropping the "ell" subscript appearing in our references. An endoscopic datum of G for which ${}^L H \not\simeq {}^L G$ will be called *proper*. We will frequently abuse notation and refer to H as a stand-in for the full datum, and denote the other two elements of the triple by ξ_H and s_H. Lastly, we will also use the formalism of endoscopic data for our unitary groups and denote by $\tilde{\mathcal{E}}(N)$ the set of pairs consisting of a product G of quasisplit unitary groups together with the L-embedding $\xi = \eta_{\kappa,\underline{\kappa}}$ from Section 2.2.6, and by $\tilde{\mathcal{E}}_{\mathrm{sim}}(N)$ the subset of $\tilde{\mathcal{E}}(N)$ for which $G = U(N)$.

For any inner form G of $U_{E/F}(N)$, the set $\mathcal{E}(G)$ consists of pairs

$$(H, \xi) = (U(N_1) \times U(N_2), \xi_{\underline{\kappa}}), \quad N_1, N_2 \geq 0, N_1 + N_2 = N,$$

where $\xi_{\underline{\kappa}}$ was defined in (7). The signature $\underline{\kappa} = ((-1)^{N-N_1}, (-1)^{N-N_2})$ depends on the respective ranks of the groups. The equivalence class of endoscopic data is then uniquely determined by N_1; see [38, Section 2.4].

2.4.1. *Endoscopic data and parameters.* Let F be global and G/F be unitary and $\psi = (\psi^N, \tilde{\psi}) \in \Psi(G^*, \eta_\kappa)$ be an Arthur parameter. Let $(H, \xi_H, s_H) \in \mathcal{E}(G)$, and let $\psi^H = (\psi^{N,H}, \tilde{\psi}^H) \in \Psi(H, \eta_\kappa \circ \xi_H)$ be an Arthur parameter for H satisfying $\psi^N = \psi^{N,H}$ and $\tilde{\psi} = \xi_H \circ \tilde{\psi}^H$. In this situation, we will abuse notation and write that $\psi = \xi_H \circ \psi^H$. Since s_H commutes with H, it also commutes with the image of $\tilde{\psi}$. We thus get a mapping

$$(H, \psi^H) \mapsto (\xi_H \circ \psi^H, s_H) \tag{12}$$

from the set of pairs (H, ψ^H) onto the set of pairs consisting of a parameter ψ for G and an element s of the centralizer S_ψ. The importance of the quotient $\mathcal{S}_\psi$ comes from the fact that for each ψ, the map (12) descends to a bijection between $\mathcal{S}_\psi$ and the set of endoscopic data such that ψ factors through ξ_H. We state this result below, under simplifying assumptions: G global unitary and ψ square-integrable.

Lemma 13. *Let F be global and $G^* = U_{E/F}(N)$. Let $\psi = (\tilde{\psi}, \psi^N) \in \Psi(G^*, \eta_\kappa)$. The map (12) induces a bijection*

$$(H, \psi^H) \leftrightarrow (\psi, s)$$

where the left-hand side runs over pairs where H stands in for an endoscopic datum (H, ξ, s) and $\psi^H = (\tilde{\psi}^H, \psi^{N,H}) \in \Psi(H, \eta_\kappa \circ \xi_H)$ with $\psi^N = \psi^{N,H}$ and $\tilde{\psi} = \xi \circ \tilde{\psi}^H$, and the right-hand side runs over elements of $\mathcal{S}_\psi$.

Proof. The proof occupies Section 1.4 of [26], and the above statement is a reformulation of Lemma 1.4.3 therein. The square-integrability assumption on ψ implies that S_ψ and a fortiori $\bar{S}_\psi$ are finite. Thus $\bar{S}_\psi = \mathcal{S}_\psi$ and we use the latter. $\square$

2.5. *Packets.* Here, we introduce A-packets of representations associated to Arthur parameters, and the character identities relating their traces to those of corresponding representations for endoscopic groups.

2.5.1. *Local Arthur packets.* Let (G, ξ) be a unitary group over a local field. The main local results of Mok [38, Theorem 2.5.1] and Kaletha, Minguez, Shin and White [26, Theorem 1.6.1] associate to each Arthur parameter $\psi \in \Psi(G, \xi)$ a finite set Π_ψ of irreducible unitary representations of $G(F)$ called a local Arthur packet. This packet Π_ψ is empty if ψ is not relevant, and contains only tempered representations when ψ is bounded. Each nonempty Π_ψ is equipped with a pairing

$$\langle \cdot, \cdot \rangle : \mathcal{S}_\psi \times \Pi_\psi \to \{\pm 1\}. \tag{13}$$

In this way, every $\pi \in \Pi_\psi$ gives rise to a character of $\mathcal{S}_\psi$. Unramified representations correspond to the trivial character. The pairing depends on the triple (G, ξ, z) realizing G as a pure inner twist, as discussed in Remark 9.

For F archimedean, all $\pi \in \Pi_\psi$ have the same infinitesimal character. We recall how to compute it from φ_ψ following [39]. The group $W_\mathbb{R}$ is an extension of $\mathbb{C}^\times$ by the group $\langle \sigma \rangle$ of order 2. For each ψ, there is a torus $\hat{T} \in \hat{G}$ such that

$$\varphi_\psi \mid_{\mathbb{C}^\times} (z) = z^\mu \bar{z}^\nu, \quad \mu, \nu \in X_*(\hat{T}).$$

The infinitesimal character of the representations $\pi \in \Pi_\psi$ is then identified with $\mu \in X_*(\hat{T}) \simeq X^*(T)$ via the Harish-Chandra isomorphism.

Lemma 14. *Let $\psi \in \Psi(G)$ be an archimedean Arthur parameter with regular infinitesimal character, and let $H \in \mathcal{E}(G)$ be such that $\psi = \xi_\underline{\kappa} \circ \psi^H$ for $\psi^H \in \Psi(H)$. Then the infinitesimal character of ψ^H is also regular.*

Proof. By assumption, $\psi((z, 1), I) = z^\mu \bar{z}^\nu$ and the weights appearing in μ are distinct. The parameter φ factors through $\xi_\underline{\kappa} : {}^L H \to {}^L G$. Referring to (6), the restriction of $\xi_\underline{\kappa}$ to $\mathbb{C}^\times \subset W_\mathbb{R} \subset {}^L H$ is trivial, since it factors through χ_κ, which takes values in ± 1. Thus the weights of the z-part of $\varphi^H \mid_{\mathbb{C}^\times}$ are also distinct, and the infinitesimal character of the corresponding packet is regular. $\qquad \square$

For any local F, we record a result initially proved by Mok about the central character of the representations in the packet Π_ψ for the quasisplit group G^*.

Proposition 15 [26, Proposition 1.5.2, 2]. *The Langlands parameter of the central character $\omega_\pi :$ $Z(G^*)(F) \to \mathbb{C}^\times$ of any $\pi \in \Pi_\psi$ is given by the composition*

$$L_F \xrightarrow{\varphi_\psi} {}^L G^* \xrightarrow{(\det \times \mathrm{id}) \circ \eta_\kappa} \mathbb{C}^\times \rtimes W_F.$$

2.5.2. *Global Arthur packets.* Let $\psi \in \Psi(G, \xi)$ be global with localizations ψ_v. The global Arthur packet Π_ψ is then defined as

$$\Pi_\psi = \{\pi = \otimes_v \pi_v \mid \pi_v \in \Pi_{\psi_v}, \langle \cdot, \pi_v \rangle_{\psi_v} = 1 \text{ for almost all } v\}.$$

It is equipped with a pairing

$$\langle \cdot, \cdot \rangle_\psi : \mathcal{S}_\psi \times \Pi_\psi \to \{\pm 1\}, \quad \langle \cdot, \pi \rangle_\psi = \prod_v \langle \cdot, \pi_v \rangle_{\psi_v} \tag{14}$$

determined by the maps $\mathcal{S}_\psi \to \mathcal{S}_{\psi_v}$ induced by localization. We note once again that this pairing depends on the full inner twist (G, ξ). However, the local dependence on the pure inner twist, i.e., the dependency on the cocycle z appearing in the local definition of the pairing, cancels out globally. This is detailed in [26, Section 1.7].

2.5.3. *Test functions.* Continuing with F global, we fix a maximal compact subgroup K of $G(\mathbb{A}_F)$. The group K determines a maximal compact subgroup $K_v \subset G_v$ at each place v: we choose K so that K_v is hyperspecial at all the unramified v. We also fix for each v a Haar measure μ_v on G_v satisfying $\mu_v(K_v) = 1$, and a corresponding measure $\mu = \prod_v \mu_v$ on $G(\mathbb{A}_F)$. The local Hecke algebra $\mathcal{H}(G_v)$ consists of smooth, compactly supported, left and right K_v-finite functions on G_v. We will call its elements local test functions. The global Hecke algebra is the restricted product $\mathcal{H}(G) = \otimes'_v \mathcal{H}(G_v)$: it consists of smooth, compactly supported, K-finite functions. Each such test function is a finite sum of factorizable test functions of the form $f = \prod_v f_v$, where each $f_v \in \mathcal{H}(G_v)$ and all but finitely many f_v are the characteristic function of K_v.

For π_v a smooth, admissible representation of G_v, each $f_v \in \mathcal{H}(G_v)$ gives rise to an operator $\pi_v(f_v)$ on the underlying vector space of π_v, defined as follows:

$$\pi_v(f_v)(x) = \int_{G_v} f_v(g)\pi_v(g)(x)\, d\mu_v.$$

This operator is of trace class, and we denote its trace by $\operatorname{tr} \pi_v(f_v)$. Likewise globally, the algebra $\mathcal{H}(G)$ acts on $L^2_{\mathrm{disc}}(G(F)\backslash G(\mathbb{A}_F))$ and on its irreducible constituents π. We denote the trace of convolution by f by $\operatorname{tr} R(f)$ (when considering the right-regular representation on $L^2_{\mathrm{disc}}(G(F), G(\mathbb{A}_F))$) or by $\operatorname{tr} \pi(f)$ (when f acts on π irreducible).

2.5.4. *Stable distributions and transfer.* We introduce stable distributions on the local and global Hecke algebras, following Sections 3.1 and 4.2 of [38] respectively. Let $\gamma \in G(F_v)$ and let $G(F_v)_\gamma$ be its centralizer. For $f \in \mathcal{H}(G(F_v))$, let $f_{G(F_v)}(\gamma) := \int_{G(F_v)/G(F_v)_\gamma} f(g\gamma g^{-1})d\mu_v$ be the *orbital integral* associated to γ and f. It only depends on the $G(F_v)$ conjugacy class of γ.

We now introduce transfer, which makes use of *stable conjugacy classes*: the union of the finitely many conjugacy classes of $G(F_v)$ that are $G(\overline{F_v})$-conjugate. Let $G(F_v)$ first be a quasisplit unitary group. Each stable conjugacy class δ gives rise to a linear functional

$$f_v^G(\delta) = \sum_\gamma \Delta_v(\delta, \gamma) f_{G(F_v)}(\gamma), \tag{15}$$

where the sum is taken over representatives γ of all the conjugacy classes of $G(F_v)$. The factor $\Delta(\delta, \gamma)$ is equal to 1 if $\gamma \in \delta$ and to 0 otherwise. This construction gives a map from $\mathcal{H}(G_v)$ to functions on stable conjugacy classes. Denote the image of this map by $\mathcal{S}(G_v)$. A linear functional on $\mathcal{H}(G_v)$ is said to be *stable* if it factors through $\mathcal{S}(G_v)$.

Now let G_v be an arbitrary unitary group. For each endoscopic group H_v of G_v, the construction of transfer factors by Langlands and Shelstad [32] and Kottwitz and Shelstad [30] gives rise to maps

$\mathcal{H}(G_v) \to \mathcal{S}(H_v)$. The transfer factors are a significantly more delicate generalization of the $\Delta(\delta, \gamma)$ above; in particular, their normalization in [26, Section 1.1.2] (and thus the notion of transfer) depends on the choice of pure inner form as in Remark 9. This provides a system of maps from the Hecke algebras to their stable counterparts, and two functions $f_v \in \mathcal{H}(G_v)$ and $f_v^{H_v} \in \mathcal{H}(H_v)$ will be said to form a *transfer pair* if their images under their respective maps to $\mathcal{S}(H_v)$ agree. Although $f_v^{H_v}$ is not uniquely determined by f_v, we may abuse terminology and refer to a choice of $f_v^{H_v}$ as the transfer of f_v.

To extend the notion of transfer to global test functions, it is first necessary to know that the transfer of characteristic functions of maximal compact subgroups of G_v are the corresponding functions on H_v. This is the fundamental lemma, now a theorem due to Laumon and Ngô [33] in the case of unitary groups, and to Ngô [40] in general, after reductions by Waldspurger [50; 51].

Theorem 16 (fundamental lemma). *Let G_v and H_v be unramified reductive groups over a nonarchimedean local field F_v. Let $K(G_v)$ and $K(H_v)$ be respective choices of hyperspecial maximal compact subgroups. Then their characteristic functions $f_v = 1_{K(G_v)}$ and $f_v^{H_v} = 1_{K(H_v)}$ form a transfer pair.*

With this in mind, the transfer of a factorizable global test function $f = \prod_v f_v \in \mathcal{H}(G_v)$ is the product $f^H = \prod_v f_v^{H_v}$ of its transfers, a definition extended linearly to all of $\mathcal{H}(G)$. We will likewise define the global stable Hecke algebra $\mathcal{S}(G^*) := \otimes'_v \mathcal{S}(G_v^*)$. A linear functional on $\mathcal{H}(G^*)$ is *stable* if it factors through $\mathcal{S}(G^*)$.

2.5.5. *Local character identities.* The transfer of representations between G and its endoscopic groups H is encoded via identities between linear combinations of characters; the coefficients are determined by the pairings (13). We start with distributions $f^G(\psi)$ on $\mathcal{H}(G)$. Let F be local and G^*/F be a quasisplit unitary group or a product thereof, and ψ be an Arthur parameter of G^*. Then Mok attaches a stable linear form to ψ.

Theorem 17 [38, Theorem 3.2.1(a)]. *Let $\psi \in \Psi(G^*)$. Then there exists a unique stable linear form $f \mapsto f^{G^*}(\psi)$ on $\mathcal{H}(G^*)$, determined by transfer properties to GL_N. If $G^* = G_1^* \times G_2^*$ and $\psi = \psi_1 \times \psi_2$, then $f^{G^*}(\psi) = f^{G_1^*}(\psi_1) \times f^{G_2^*}(\psi_2)$.*

We will not discuss in detail the character identities relating $f^{G^*}(\psi)$ to traces on GL_N, save for reminding the reader that this distribution is related to the trace $\mathrm{tr}\, \pi_{\psi,N}(f)$ where $\pi_{\psi,N}$ corresponds to ψ under the Local Langlands Correspondence. We will focus on the relation between the $f^H(\psi^H)$ for the groups $H \in \mathcal{E}(G)$ and the characters of representations in Π_ψ. If $G = G^*$, these identities were established by Mok, and for inner forms by Kaletha, Minguez, Shin and White. Recall that s_ψ is the distinguished element of $\mathcal{S}_\psi$ defined in (11).

Theorem 18 [38, Theorem 3.2.1(b)]. *Let G^* be a quasisplit unitary group, let $\psi \in \Psi(G^*)$, and let Π_ψ be the associated Arthur packet equipped with the pairing of equation (13). Let $s_H \in \mathcal{S}_\psi$ be such that (H, ψ^H) correspond to (ψ, s_H) under the correspondence of Lemma 13. Then for a transfer pair (f, f^H) we have*

$$f^H(\psi^H) = \sum_{\pi \in \Pi_\psi} \langle s_\psi s_H, \pi \rangle \, \mathrm{tr}\, \pi(f).$$

Theorem 19 [26, Theorem 1.6.1]. *Let (G, ξ) be an inner form of $U(N)$ and let ψ, Π_ψ, H, s_H, and (f, f^H) be as above. Let $e(G)$ be the Kottwitz sign. Then*

$$f^H(\psi^H) = e(G) \sum_{\pi \in \Pi_\psi} \langle s_\psi s_H, \pi \rangle \operatorname{tr} \pi(f).$$

Remark 20. Let us recall a discussion from the introduction: the proofs in [26] are not given in full generality. For example, Theorem 19 is only proved for bounded parameters. The authors of [26] anticipate that the proof will appear in a pair of papers, the first of which [27] should contain the results we use here.

2.5.6. *Local packets for general linear groups.* As discussed in Section 2.2.1, if F is local and corresponds to a place splitting in our global CM extension, then $G \simeq \mathrm{GL}_N$. In this situation the local Arthur packet and the pairing are especially simple.

Theorem 21 [38, Section 2]. *If $G = \mathrm{GL}_N$ and ψ is an Arthur parameter for G, then the packet Π_ψ contains one element*: *the irreducible representation associated to φ_ψ by the local Langlands correspondence. The character $\langle \cdot, \pi_\psi \rangle$ is trivial.*

We now consider character of identities between representations of G and those of its endoscopic groups. They are alluded to in [38] and [26], but we give a more explicit description based on [45, Section 3.3]. For $G = \mathrm{GL}_N$, stable and regular conjugacy classes coincide, so $\mathcal{S}(G) = \mathcal{H}(G)$. Since the global extension giving rise to our unitary group is CM, we may assume that F is nonarchimedean. If $H = \mathrm{GL}_{N_1} \times \mathrm{GL}_{N_2}$ with $N_1 + N_2 = N$, then the embedding ξ_κ realizes H as a Levi subgroup of G. Let $P = HN$ be a parabolic subgroup of G containing H. Given $f \in \mathcal{H}(G)$, define the constant term along P as

$$f^P(h) := \delta_P^{1/2}(h) \int_N \int_K f(khnk^{-1}) \, dk \, dn, \quad h \in H(F).$$

Here the integrals are taken with respect to suitably normalized Haar measures and δ_P is the modulus character. The function f_v^P is smooth and compactly supported, and by results of van Dijk [15], it satisfies the requisite orbital integrals identities to be a transfer of f, so we let $f^H := f^P$. If f is unramified, then f^H is the image of f under the map $\mathcal{H}(G)^{ur} \to \mathcal{H}(H)^{ur}$ induced by the Satake isomorphism. Thus this notion of transfer satisfies the fundamental lemma.

For a parameter ψ of G, we let $f^G(\psi) = \operatorname{tr} \pi_\psi(f)$ [26, Section 1.5] for the unique $\pi_\psi \in \Pi_\psi$ and extend this definition multiplicatively to products of general linear groups. Let π_ψ^H be the unique representation in the packet associated to ψ^H. It follows from the local Langlands correspondence (see for example [23, page 6] and note that the twist therein is accounted for here in the definition of the embedding ξ_κ) that $\pi_\psi = \mathcal{I}_P(\pi_\psi^H)$, where $\mathcal{I}_P$ denotes normalized parabolic induction with respect to P. In view of this and of Theorem 21, the local character identities for GL_N amount to an equality of traces between $\operatorname{tr} \pi(f^H)$ and the trace of f on the corresponding induced representation. Again this is a result of van Dijk, which we record below.

Theorem 22 [15, Section 5]. *Let G, H, P, and f^H be as above. Let π be a unitary irreducible representation of H and let $\mathcal{I}_P(\pi)$ be its normalized parabolic induction with respect to P. Then $\operatorname{tr}\pi(f^H) = \operatorname{tr}\mathcal{I}_P(\pi)(f)$.*

2.6. *The trace formula and its stabilization.* We now introduce Arthur's trace formula following [26, Section 3] (see [5, Section 3] for a more detailed exposition), focusing on the statements needed for our applications. The rough picture is as follows: for F a number field, and a connected reductive group G/F, the trace formula I_{disc} (sometimes denoted I_{disc}^G) is a distribution on the Hecke algebra $\mathcal{H}(G)$, defined in terms of the traces of intertwining operators on variants of $L^2_{\mathrm{disc}}(G(F)\backslash G(\mathbb{A}_F))$ indexed by a system of Levi subgroups of G. The contribution of the group G itself is the trace $\operatorname{tr} R_{\mathrm{disc}}(f) := \operatorname{tr} R(f)$ introduced in Section 2.5.3. The trace formula admits two decompositions: a spectral one into a sum over the contributions of $\psi \in \Psi(G, \xi)$, and an endoscopic one (or stabilization) into a sum of stable distributions on endoscopic groups. Our proof will follow from the interplay of these two decompositions.

2.6.1. *Contribution of a parameter.* We start by directly introducing the distributions given by the contribution of each Arthur parameter as in [26, Section 3.3]. For following paragraphs, let $(G, \eta_{\kappa \cdot \underline{\kappa}})$ be a pair consisting an inner form of a (possible product of) unitary groups, and an embedding $\eta_{\kappa \cdot \underline{\kappa}} : {}^L G \to {}^L G(N)$ as in Section 2.2.6. When G is an inner form of $U(N)$, we have $\eta_{\kappa \cdot \underline{\kappa}} = \eta_\kappa$.

Recall $\mathcal{C}(G)$, the set of families of conjugacy classes introduced in Section 2.3.7. To any automorphic representation π of G, we can associate an element $c_\pi \in \mathcal{C}(G)$ by letting $c_{\pi,v}$ be the Satake parameter of π_v at all the unramified places v. Likewise, we associate to π an infinitesimal character μ_π. Then for $c \in \mathcal{C}(G)$ and a positive real number t, the distribution $I_{\mathrm{disc},t,c}$ is described in [26, Section 3.1]. It is the restriction of the traces defining I_{disc} to representations π such that $c = c_\pi$, and such that μ_π satisfies $|\operatorname{Im}\mu_\pi| = t$ under a suitable metric. To go from conjugacy classes to parameters, recall that in Section 2.3.7 we identified $\Psi(N)$ with $\mathcal{C}_{\mathrm{aut}}(N) \subset \mathcal{C}(N)$. To each $\psi^N \in \Psi(N)$ is thus associated an element $c(\psi^N) \in \mathcal{C}_{\mathrm{aut}}(N)$ as well as a positive real number $t(\psi^N)$ coming from the infinitesimal character of ψ^N. For each parameter $\psi^N \in \Psi(N)$, we follow [26, Section 3.3] and define

$$I_{\mathrm{disc},\psi^N,\eta_{\kappa \cdot \underline{\kappa}}} = \sum_{\substack{c \mapsto c(\psi^N) \\ t \mapsto t(\psi^N)}} I_{\mathrm{disc},t,c}.$$

The sum runs over the $c \in \mathcal{C}(G)$ that map to $c(\psi^N)$ under the map $\mathcal{C}(G) \to \mathcal{C}(N)$ induced by $\eta_{\kappa \cdot \underline{\kappa}}$. When $G^* = U(N)$, we follow [26, Section 3.3] and shorten $I_{\mathrm{disc},\psi^N,\eta_\kappa}$ to $I_{\mathrm{disc},\psi}$ when $\psi = (\psi^N, \tilde{\psi}) \in \Psi(G^*, \eta_\kappa)$, using the injection $\eta_{\kappa,*}$ of Section 2.3.4. We similarly obtain distributions $\operatorname{tr} R_{\mathrm{disc},c,t}$, $R_{\mathrm{disc},\psi^N,\eta_{\kappa \cdot \underline{\kappa}}}$ and $\operatorname{tr} R_{\mathrm{disc},\psi^N,\eta_\kappa} := \operatorname{tr} R_{\mathrm{disc},\psi}$. If we have $G^* \in \tilde{\mathcal{E}}_{\mathrm{sim}}(N)$, as well as $(H, \xi_{\underline{\kappa}}) \in \mathcal{E}(G)$ and $\psi \in \Psi(G^*, \eta_\kappa)$, we will also shorten notation and denote $I_{\mathrm{disc},\psi}^H = I_{\mathrm{disc},\psi^N,\xi_{\underline{\kappa}} \circ \eta_\kappa}^H$.

An essential step in the proof of the endoscopic classification of representations is showing that $\operatorname{tr} R_{\mathrm{disc},\psi}$ computes the traces of the representations in Π_ψ, provided that ψ satisfies the two conditions of Remark 11.

Theorem 23 [38, (5.7.27); 26, proof of Theorem 5.0.5]. *Let $\psi \in \Psi(G, \xi)$ be a square-integrable parameter associated to Π_ψ, and let $f \in \mathcal{H}(G)$. Then*

$$\operatorname{tr} R_{\mathrm{disc},\psi}(f) = \sum_{\pi \in \Pi_\psi} m(\pi) \operatorname{tr} \pi(f).$$

In the notation of Section 2.5, the multiplicity $m(\pi)$ is equal to 1 if $\langle \pi, \cdot \rangle_\psi = \epsilon_\psi$ as characters of $\mathcal{S}_\psi$, and 0 otherwise; see [26, Section 1.7]. Note again that [26, Theorem 5.0.5] is stated, but not fully proved, in the case of nongeneric parameters, as mentioned in the introduction and in Remark 20.

Following a result of Bergeron and Clozel, the distributions $\operatorname{tr} R_{\mathrm{disc},\psi}$ and $I_{\mathrm{disc},\psi}$ agree if the infinitesimal character is regular.

Theorem 24 [7, Theorem 6.2]. *Let G be a connected reductive group. Let $\psi \in \Psi(G)$ be a global Arthur parameter such that ψ_∞ has regular infinitesimal character. Then the contributions of the Levi subgroups $M \neq G$ to the distribution $I_{\mathrm{disc},\psi}$ vanish. In particular for all $f \in \mathcal{H}(G)$ we have $I_{\mathrm{disc},\psi}(f) = \operatorname{tr} R_{\mathrm{disc},\psi}(f)$.*

If $H = U(N_1) \times U(N_2)$, and $\psi^H = \psi_1 \times \psi_2$ we can write

$$R_{\mathrm{disc},\psi^H}(f) = R_{\mathrm{disc},\psi_1}(f_1) \cdot R_{\mathrm{disc},\psi_2}(f_2), \quad f = f_1 \times f_2 \in \mathcal{H}(H).$$

Following the above result, we can also write this as $I_{\mathrm{disc},\psi^H}(f)$ provided that ψ^H has regular infinitesimal character.

2.6.2. *Stabilization.* We now recall the identity that drives our theorems: the stabilization of $I_{\mathrm{disc},\psi}$, i.e., its decomposition into sum of stable traces of the transfers f^H of f for the endoscopic groups $H \in \mathcal{E}(G)$. Our references are to Arthur [5], but the versions for unitary groups are formally identical; see for example [26, (3.3.2)] and [38, (4.2.1)]. Recall that $\tilde{\Psi}(N)$ is the set of conjugate self-dual parameters, and $\Psi(G, \xi) \subset \tilde{\Psi}(N)$.

Theorem 25 [5, Corollary 3.3.2(b)]. *Suppose that $\psi \in \tilde{\Psi}(N)$ and let $f \in \mathcal{H}(G)$. Then for each endoscopic datum $(H, \xi_H) \in \mathcal{E}(G)$ there is a constant $\iota(G, H)$ and stable distributions $S^H_{\mathrm{disc},\psi}$ on $\mathcal{H}(H)$, defined inductively, such that*

$$I_{\mathrm{disc},\psi}(f) = \sum_{H \in \mathcal{E}(G)} \iota(G, H) S^H_{\mathrm{disc},\psi}(f^H). \tag{16}$$

Remark 26. For unitary groups, the global factor $\iota(G, H)$ is introduced in [38, Section 4.2] and [26, Section 3.1]. It is independent of the inner form G. If $G = U(N)$ and $H = U(N_1) \times U(N_2)$, then following [38, 4.2] we have

$$\iota(G, H) = \begin{cases} 1, & N_1 N_2 = 0, \\ \frac{1}{2}, & N_1, N_2 \neq 0, \ N_1 \neq N_2, \\ \frac{1}{4}, & N_1 = N_2 \neq 0. \end{cases} \tag{17}$$

3. Upper bounds from the stabilization

In this section we unpack the summands of the stabilization of $I_{\mathrm{disc},\psi}$ and extract upper bounds on the trace of test functions from the character identities.

3.1. *The stable multiplicity formula.* Let $\psi = (\psi^N, \tilde{\psi}) \in \Psi(G^*, \eta_\kappa)$. Recall the decomposition from (16):

$$I_{\mathrm{disc},\psi}(f) = \sum_{H \in \mathcal{E}(G)} \iota(G, H) S^H_{\mathrm{disc},\psi}(f^H). \tag{18}$$

The stable multiplicity formula expresses each $S^H_{\mathrm{disc},\psi}$ as a sum of traces. If f_v is a local test function and ψ_v a local parameter, the formula for $f^{H_v}(\psi_v)$ was given in Section 2.5.5. If $f = \prod_v f_v$ and ψ are global, we write $f^H(\psi) := \prod_v f^{H_v}(\psi_v)$. The group $\mathcal{S}_\psi$ and the element s_ψ were defined in Section 2.3.8, and ϵ_ψ in Section 2.3.9. The stable multiplicity formula, only defined for quasisplit groups, is the following expression:

Theorem 27 [38, Theorem 5.1.2]. *For $\psi \in \Psi(G, \eta_\kappa)$, we have*

$$S^G_{\mathrm{disc},\psi}(f) = |\mathcal{S}_\psi|^{-1} \epsilon^G_\psi(s_\psi) \sigma(\bar{S}^0_\psi) f^G(\psi).$$

For any connected reductive group S, the quantity $\sigma(S)$ was defined by Arthur in [5, Section 4.1]. The centralizers S_ψ of our ψ are always finite, so $\bar{S}^0_\psi$ is trivial and $\sigma(\bar{S}^0_\psi) = 1$; see [38, Remark 5.1.4]. The stable multiplicity formula is stated for G a unitary group (in which case the map $\psi \mapsto \psi^N$ is injective), but can be extended to products $H \in \mathcal{E}(G)$. Let $\Psi(H, \psi^N)$ be the set consisting of parameters $\psi^H = (\psi^{N,H}, \tilde{\psi}^H)$ with $\psi^{N,H} = \psi^N$. The stable multiplicity formula for H, given in [38, (5.6.3)], is

$$S^H_{\mathrm{disc},\psi}(f^H) = \sum_{\psi^H \in \Psi(H, \psi^N)} \frac{1}{|\mathcal{S}_{\psi^H}|} \epsilon^H_\psi(s^H_{\psi^H}) \sigma(\bar{S}^0_{\psi^H}) f^H(\psi^H). \tag{19}$$

We combine (18) and (19) and rewrite the resulting expression as a sum over pairs (H, ψ^H) to get

$$I_{\mathrm{disc},\psi}(f) = \sum_{(H, \psi^H)} \iota(G, H) \frac{1}{|\mathcal{S}_{\psi^H}|} \epsilon^H_\psi(s^H_{\psi^H}) \sigma(\bar{S}^0_{\psi^H}) f^H(\psi^H). \tag{20}$$

We now collect the terms that can be bounded uniformly, and let

$$C(\psi, H) := \iota(G, H) \sigma(\bar{S}^0_{\psi^H}) |\mathcal{S}_{\psi^H}|^{-1}. \tag{21}$$

Lemma 28. *Let $\psi \in \Psi(G^*, \eta_\kappa)$ and let $(H, \xi_H, s_H) \in \mathcal{E}(G)$ be an endoscopic datum such that ψ factors through ξ_H. Then*:

(i) *The contribution of (H, ψ^H) to the sum (20) is equal to $C(\psi, H) \epsilon^H_\psi(s^H_\psi) f^H(\psi^H)$.*

(ii) *The constant $C(\psi, H)$ is bounded uniformly in ψ and H: it always satisfies $2^{-(N+1)} \leq C(\psi, H) \leq 1$, where N is the rank of G.*

Proof. Part (i) follows immediately from (20) and it suffices to exhibit the bound on $C(\psi, H)$. As stated above, we have $\sigma(\bar{S}^0_\psi) = 1$ since ψ is elliptic. We also gave uniform bounds on $\iota(G, H)$ in (17) and on $|\mathcal{S}_\psi|$ in (10). $\qquad\square$

In Lemma 13, we gave a bijection between $\mathcal{S}_\psi$ and the set of pairs (H, ψ^H). We use it to reindex the sum (20) and obtain the expression

$$I_{\text{disc},\psi}(f) = \sum_{s_H \in \mathcal{S}_\psi} C(\psi, s_H) \epsilon_\psi^H(s_{\psi^H}^H) f^H(\psi^H). \tag{22}$$

This sum depends on parameters and representations of H, which we want to rewrite in terms of G. For ϵ_ψ, we use Mok's so-called *endoscopic sign lemma*.

Lemma 29 [38, Lemma 5.6.1]. *Let $(H, \xi, s_H) \in \mathcal{E}(G)$ and $\psi \in \Psi(G^*, \eta_\kappa)$ be such that (H, ψ^H) corresponds to (ψ, s_H). Let $\epsilon_\psi^{G^*}$ and ϵ_ψ^H be the respective characters of ψ and ψ^H. Let $s_{\psi^H}^H$ be the image of $\psi^H(-I)$ in the quotient $\mathcal{S}_\psi^H$ associated to H. Then we have*

$$\epsilon_\psi^H(s_{\psi^H}^H) = \epsilon^{G^*}(s_\psi s_H).$$

We can now rewrite $I_{\text{disc},\psi}(f)$ in a form conducive to extracting bounds.

Proposition 30. *Let $\psi \in \Psi(G, \xi)$, and let $f \in \mathcal{H}(G)$ be factorizable. Then*

$$I_{\text{disc},\psi}(f) = \sum_{s_H \in \mathcal{S}_\psi} C(\psi, s_H) \epsilon_\psi^{G^*}(s_\psi s_H) \prod_v \left(\sum_{\pi_v \in \Pi_{\psi_v}} \langle s_{\psi_v} s_{H_v}, \pi_v \rangle \operatorname{tr} \pi_v(f_v) \right)$$

$$= \sum_{s_H \in \mathcal{S}_\psi} C(\psi, s_H) \sum_{\pi \in \Pi_\psi} \epsilon_\psi^{G^*}(s_\psi s_H) \langle s_\psi s_H, \pi \rangle \operatorname{tr} \pi(f). \tag{23}$$

Proof. We start from the equality (22):

$$I_{\text{disc},\psi}(f) = \sum_{s_H \in \mathcal{S}_\psi} C(\psi, s_H) \epsilon_\psi^H(s_\psi^H) f^H(\psi^H).$$

The distribution $f^H(\psi^H)$ was defined as $f^H(\psi^H) = \prod_v f_v^{H_v}(\psi_v^H)$. Each local factor can be written in terms of the trace of representations in Π_{ψ_v} by Theorems 18, 19, and 22. In all cases, the identity is

$$f_v^{H_v}(\psi_v^{H_v}) = e(G_v) \sum_{\pi_v \in \Pi_{\psi_v}} \langle s_{\psi_v} s_{H_v}, \pi_v \rangle \operatorname{tr} \pi_v(f_v).$$

The local Kottwitz signs cancel out globally, and using Lemma 29, we rewrite

$$I_{\text{disc},\psi}(f) = \sum_{s_H \in \mathcal{S}_\psi} C(\psi, s_H) \epsilon_\psi^{G^*}(s_\psi s_H) \prod_v \left(\sum_{\pi_v \in \Pi_{\psi_v}} \langle s_{\psi_v} s_{H_v}, \pi_v \rangle \operatorname{tr} \pi_v(f_v) \right).$$

At all but finitely many v, we have $f_v = 1_{K_v}$ for a hyperspecial maximal compact subgroup K_v. At these places, $\operatorname{tr} \pi_v(f_v)$ is only nonzero on K_v-unramified representations π_v. Unramified local packets contain

exactly one unramified representation following [26, Proposition 1.5.2(5)] so we interchange the sum and product to get

$$\prod_v \left(\sum_{\pi_v \in \Pi_{\psi_v}} \langle s_{\psi_v} s_{H_v}, \pi_v \rangle \operatorname{tr} \pi_v(f_v) \right) = \sum_{\pi \in \Pi_\psi} \left(\prod_v \langle s_{\psi_v} s_{H_v}, \pi_v \rangle \right) \operatorname{tr} \pi(f)$$

for $\pi = \otimes_v \pi_v$. Using the definition $\langle \cdot, \pi \rangle := \prod_v \langle \cdot, \pi_v \rangle$, we rewrite

$$I_{\mathrm{disc},\psi}(f) = \sum_{s_H \in \mathcal{S}_\psi} C(\psi, s_H) \sum_{\pi \in \Pi_\psi} \epsilon_\psi^{G^*}(s_\psi s_H) \langle s_\psi s_H, \pi \rangle \operatorname{tr} \pi(f). \qquad \square$$

3.2. Upper bounds and the dominant group.

Recall once more the bijection $(H, \psi^H) \leftrightarrow (\psi, s_H)$ from Section 2.4.1. We will single out one object on either side, and show that for certain f, its contribution to the distribution $I_{\mathrm{disc},\psi}(f)$ bounds the others. Recall that $s_\psi \in \mathcal{S}_\psi$ was the image of the matrix $-I \in \mathrm{SL}_2$ under ψ.

Definition 31. Let (H_ψ, ψ^{H_ψ}) be the pair corresponding to the pair (ψ, s_ψ) containing the distinguished element s_ψ under the bijection $(H, \psi^H) \leftrightarrow (\psi, s_H)$.

Note that it is possible that $H_\psi = G$, for example when ψ is bounded.

Definition 32. Let $\psi \in \Psi(G, \xi)$ and let (H, ξ_H, s_H) be such that ψ factors through ξ_H. Let f be a global test function. Then define

$$S(\psi, s_H, f) = C(\psi, s_H) \sum_{\pi \in \Pi_\psi} \epsilon_\psi^{G^*}(s_\psi s_H) \langle s_\psi s_H, \pi \rangle \operatorname{tr} \pi(f).$$

Proposition 30 can then be reformulated as stating that

$$I_{\mathrm{disc},\psi}(f) = \sum_{s_H \in \mathcal{S}_\psi} S(\psi, s_H, f). \tag{24}$$

Lemma 33. If $H = H_\psi$, then

$$S(\psi, s_\psi, f) = C(\psi, s_\psi) \sum_{\pi \in \Pi_\psi} \operatorname{tr}(\pi)(f). \tag{25}$$

Proof. This follows since $s_{H_\psi} = s_\psi$ by definition. Since $\mathcal{S}_\psi \simeq (\mathbb{Z}/2\mathbb{Z})^{r-1}$, this implies that $\epsilon_\psi^{G^*}(s_\psi^2) = 1$ and $\langle s_\psi^2, \pi \rangle = 1$ for all π. $\qquad \square$

This allows us to state our main application of the stable trace formula.

Theorem 34. Let G be a unitary group, let $\psi \in \Psi(G, \xi)$, and let $f \in \mathcal{H}(G)$ be a factorizable test function with $\operatorname{tr} \pi(f)$ real and nonnegative for all $\pi \in \Pi_\psi$. Then there exist a constant $C(\psi)$ such that

$$I_{\mathrm{disc},\psi}(f) \leq C(\psi) S(\psi, s_\psi, f).$$

The constant $C(\psi)$ satisfies $2^{-(N+1)} \leq C(\psi) \leq 2^{2N}$; it is thus bounded above and below independently of ψ.

Proof. We compare the various terms appearing in (24):

$$I_{\text{disc},\psi}(f) = \sum_{s_H \in \mathcal{S}_\psi} S(\psi, s_H, f).$$

Ignoring for a moment the constants $C(\psi, s_H)$, the summands only differ from one another via the signs $\epsilon_\psi^{G^*}(s_\psi s_H)\langle s_\psi s_H, \pi \rangle \in \{\pm 1\}$ appearing as coefficients of the traces $\operatorname{tr} \pi(f)$. In the term coming from s_ψ, we get

$$S(\psi, s_\psi, f) = C(\psi, s_\psi) \sum_{\pi \in \Pi_\psi} \operatorname{tr}(\pi)(f)$$

from Lemma 33. For any other $s_H \in \mathcal{S}_\psi$, the coefficients $\epsilon_\psi^{G^*}(s_\psi s_H)\langle s_\psi s_H, \pi \rangle$ have the potential to be equal to -1. Thus if $\operatorname{tr} \pi(f) \geq 0$ for all $\pi \in \Pi_\psi$, we have

$$S(\psi, s_H, f) = C(\psi, s_H) \sum_{\pi \in \Pi_\psi} \epsilon_\psi^{G^*}(s_\psi s_H)\langle s_\psi s_H, \pi \rangle \operatorname{tr} \pi(f)$$

$$\leq C(\psi, s_H) \sum_{\pi \in \Pi_\psi} \operatorname{tr}(\pi)(f) = \frac{C(\psi, s_H)}{C(\psi, s_\psi)} \cdot S(\psi, s_\psi, f).$$

Summing over the s_H we get

$$I_{\text{disc},\psi}(f) \leq \left(\frac{\sum_{s_H \in \mathcal{S}_\psi} C(\psi, s_H)}{C(\psi, s_\psi)} \right) S(\psi, s_\psi, f) := C(\psi) S(\psi, s_\psi, f).$$

For the bounds, we showed in Lemma 28 that $2^{-(N+1)} \leq C(\psi, s_H) \leq 1$. As for the cardinality of $\mathcal{S}_\psi$, it is bounded between 1 and 2^{N-1} as we saw in Section 2.3.8. □

In practice, the group H_ψ is easily computed from $\psi \mid_{\text{SL}_2}$.

Lemma 35. *Let* $\psi = \boxplus_i (\mu_i \boxtimes \nu(m_i)) \in \Psi(N)$ *be a global square-integrable Arthur parameter, and let* $N_1 = \sum_{m_i \equiv 1 \bmod 2} m_i$. *Then the group* H_ψ *is*

$$H_\psi = U(N_1) \times U(N - N_1).$$

Proof. By definition $s_\psi = \psi(1, -I) \in \text{GL}_N$. The image of $-I$ under the m-dimensional representation of SL_2 is $(-1)^{m+1} I_m$. Thus $s_\psi = \operatorname{diag}(-I_{N_1}, I_{N_2})$, where $N_1 = \sum_{m_i \equiv 1 \bmod 2} m_i$ and $N_2 = N - N_1$, with centralizer $\text{GL}_{N_1} \times \text{GL}_{N_2}$. □

The image $\psi(\text{SL}_2)$ and the group H_ψ are determined by any localization $\psi_v(\text{SL}_2)$. In Section 5, we will use this, together with the known (archimedean) parameters of cohomological representations, to bound growth of cohomology.

3.3. *Hyperendoscopy.* We recall the notion of hyperendoscopic datum first introduced by Ferrari [17]. We will use it to bound the expression $S(\psi, s_\psi, f)$. As pointed out by Dalal [13], the results of [17] do not quite hold in full generality, but they do hold for unitary groups, which have simply connected derived subgroups.

Definition 36. A chain of hyperendoscopic data for the (local or global) group G is a collection

$$\mathcal{H} = (G, H_1, \ldots, H_q),$$

where H_1 is a proper endoscopic datum for G and H_{i+1} is a proper endoscopic datum for H_i.

The integer $p(\mathcal{H}) = q$ is the *depth* of the datum. Denote

$$\iota(\mathcal{H}) = (-1)^{p(\mathcal{H})} \iota(G, H_1) \cdot \iota(H_1, H_2) \cdots \cdots \iota(H_{q-1}, H_q),$$

and $I_{\text{disc}}^{\mathcal{H}} := I_{\text{disc}}^{H_q}$. As with endoscopic data, two chains of hyperendoscopic data will be considered equivalent if they are conjugate under $\hat{G}$. Ferrari denotes by $\mathcal{HE}(G)$ the collection of equivalence classes of chains of hyperendoscopic data for G. If $\psi \in \Psi(G)$ is an Arthur parameter, we will denote by $\mathcal{HE}(G, \psi)$ the collection of equivalence classes of chains of hyperendoscopic data $\mathcal{H} \in \mathcal{HE}(G)$ such that ψ factors through the embedding $\xi_{p(\mathcal{H})}$ associated to $H_{p(\mathcal{H})}$. Note that the depth of chains in $\mathcal{HE}(G, \psi)$ is bounded above by the number of simple constituents of ψ.

If $\mathcal{H} \in \mathcal{HE}(G)$ is a chain of hyperendoscopic data, and f^G is a test function, we inductively define $f^{H_{i+1}} = (f^{H_i})^{H_{i+1}}$. The function $f^{H_{p(\mathcal{H})}}$ depends on a choice of transfer f^{H_i} at each step. We allow this, but require that our choice of f^{H_i} be consistent over chains that are truncations of one another. The following is the specialization to a parameter ψ of a trick initially discovered by Ferrari [17, 3.4.2].

Proposition 37. *Let $(G, \eta) \in \tilde{\mathcal{E}}(N)$ be quasisplit and let $\psi^N \in \Psi(N)$. Then*

$$S_{\text{disc}, \psi^N, \eta}^G(f) = \sum_{\mathcal{H} \in \mathcal{HE}(G, \psi)} \iota(\mathcal{H}) I_{\text{disc}, \psi^N, \eta \circ \xi_{p(\mathcal{H})}}^{H_q}(f^{H_q}).$$

Proof. If $\psi^N \in \Psi_{\text{sim}}(N)$, then $I_{\text{disc}, \psi^N, \eta}^G(f) = S_{\text{disc}, \psi^N, \eta}^G(f)$ and the result holds trivially. Otherwise, we have that

$$S_{\text{disc}, \psi^N, \eta}^G(f) = I_{\text{disc}, \psi^N, \eta}^G(f) - \sum_{H \in \mathcal{E}(G, \psi^N)} \iota(G, H) S_{\text{disc}, \psi^N, \eta \circ \xi}^H(f^H). \tag{26}$$

By induction, for each H in $\mathcal{E}(G, \psi)$, we have

$$S_{\text{disc}, \psi^N, \eta \circ \xi}^H(f^H) = \sum_{\mathcal{H} \in \mathcal{HE}(G, \psi)} \iota(\mathcal{H}) I_{\text{disc}, \psi^N, \eta \circ \xi_{p(\mathcal{H})}}^{H_q}(f^{H_q}). \tag{27}$$

By construction, each $\mathcal{H} \in \mathcal{HE}(G, \psi^N)$ is obtained from a hyperendoscopic datum $\mathcal{H}' \in \mathcal{HE}(H, \psi^N)$ for some $H \in \mathcal{E}(G, \psi^N)$, and $p(\mathcal{H}) = p(\mathcal{H}') + 1$. Substituting (27) into (26) yields the result. □

Recall that when $G \in \tilde{\mathcal{E}}_{\text{sim}}(N)$, the map $\psi \mapsto \psi^N$ is injective. On the other hand, if H is an product of unitary groups, there could be several parameters ψ^H for H such that $\psi^H \mapsto \psi^N$ under $\eta_{\kappa \cdot \underline{\kappa}}$. From [38, Section 5.6] we see that if $H = H_1 \times H_2$ with $H_i = U(N_i)$, and $f^H = f^{H_1} \times f^{H_2}$, then

$$S_{\text{disc}, \psi^N, \eta_{\kappa \cdot \underline{\kappa}}}^H(f^H) = \sum_{\psi^H = \psi_1 \times \psi_2, \psi^H \mapsto \psi^N} S_{\text{disc}, \psi_1}^{H_1}(f^{H_1}) \times S_{\text{disc}, \psi_2}^{H_2}(f^{H_2}).$$

The expression $S(\psi, s_H, f)$ of Definition 32 picks out one of these summands.

Definition 38. Let $H = H_1 \times H_2$ as above, and let $\mathcal{H} \in \mathcal{HE}(H)$ with $p(\mathcal{H}) = q$. Then $H_q = H_{q_1} \times H_{q_2}$ with $H_{q_i} \in \mathcal{HE}(H_i)$. Let $\psi^H = \psi_1 \times \psi_2 \in \Psi(H, \eta_{\kappa \cdot \underline{\kappa}})$. For a test function $f^{H_q} = f^{H_{q_1}} \times f^{H_{q_2}}$, define

$$I_{\mathrm{disc}, \psi^H}^{H_q}(f^H) = I_{\mathrm{disc}, \psi_1}^{H_{q_1}}(f^{H_{q_1}}) \times I_{\mathrm{disc}, \psi_2}^{H_{q_2}}(f^{H_{q_2}}).$$

Here we use the notation $I_{\mathrm{disc}, \psi_i}^{H_{q_i}}$ as in Section 2.6 since $U(N_i) \in \tilde{\mathcal{E}}_{\mathrm{sim}}(N_i)$.

Corollary 39. *Let $H = H_1 \times H_2$ as above and let $\psi^H = \psi_1 \times \psi_2 \in \Psi(H, \xi_{\underline{\kappa}})$ so that (H, ψ^H) corresponds to (ψ^N, s_H) under the correspondence of Lemma 13. Assume that $f^H = f^{H_1} \times f^{H_2}$. Then*

$$S(\psi, s_H, f) = \iota(G, H) \sum_{\mathcal{H} \in \mathcal{HE}(H, \psi)} \iota(\mathcal{H}) I_{\mathrm{disc}, \psi^H}^{H_q}(f^{H_q}).$$

Proof. We see in [38, Section 5.6] that the term $S_{\mathrm{disc}, \psi_1}^{H_1}(f^{H_1}) \times S_{\mathrm{disc}, \psi_2}^{H_2}(f^{H_2})$ is equal to

$$\frac{1}{|\mathcal{S}_{\psi_1}||\mathcal{S}_{\psi_1}|} \epsilon^{H_1}(\psi_1) \epsilon^{H_2}(\psi_2) f^{H_1}(\psi_1) f^{H_2}(\psi_2) = \frac{1}{|\mathcal{S}_\psi|} \epsilon^H(\psi) f^H(\psi).$$

By the argument of Proposition 30, the last expression is equal to

$$\frac{1}{|\mathcal{S}_\psi|} \sum_{\pi \in \Pi_\psi} \epsilon_\psi^{G^*}(s_\psi s_H) \langle s_\psi s_H, \pi \rangle \operatorname{tr} \pi(f) = \frac{S(\psi, s_H, f) \cdot}{|\mathcal{S}_\psi| C(\psi, s_H)} = \frac{S(\psi, s_H, f)}{\iota(G, H)}.$$

Applying Proposition 37 to each factor of $S_{\mathrm{disc}, \psi_1}^{H_1}(f^{H_1}) \times S_{\mathrm{disc}, \psi_2}^{H_2}(f^{H_2})$, we get

$$\frac{S(\psi, s_H, f)}{\iota(G, H)} = \left(\sum_{\mathcal{H}_1 \in \mathcal{HE}(H_1, \psi_1)} \iota(\mathcal{H}_1) I_{\mathrm{disc}, \psi_1}^{H_{q_1}}(f^{H_{q_1}}) \right) \cdot \left(\sum_{\mathcal{H}_2 \in \mathcal{HE}(H_2, \psi_2)} \iota(\mathcal{H}_2) I_{\mathrm{disc}, \psi_2}^{H_{q_2}}(f^{H_{q_2}}) \right)$$

$$= \sum_{\mathcal{H} \in \mathcal{HE}(H, \psi)} \iota(\mathcal{H}) I_{\mathrm{disc}, \psi^H}^{H_q}(f^{H_q}). \qquad \square$$

4. Limit multiplicity

Here we apply the results of the previous section to the limit multiplicity problem.

4.1. *Level structures.* Let $\mathcal{O}_E$ and $\mathcal{O}_F$ be the rings of integers of the global fields E and F. We introduce sets of places of F:

- S_f is a finite set of finite places of F, containing the places which ramify in E as well as the places below those where the character χ_- introduced in Section 2.1 is ramified.

- S_∞ is the set of all infinite places of F.

- $S_0 \subsetneq S_\infty$ is a nonempty subset of the infinite places.

- $S = S_f \cup S_\infty$.

Note that the third requirement implies that $F \neq \mathbb{Q}$. Let $\mathfrak{p}$ be an ideal of F with residue characteristic strictly greater than $N^2[F : \mathbb{Q}] + 1$, corresponding to a place $v_\mathfrak{p} \notin S$. For each finite place v of F, denote by $\mathcal{O}_{F_v}$ the ring of integers of F_v, and let $\hat{\mathcal{O}}_F = \prod_v \mathcal{O}_{F_v}$, and similarly for $\hat{\mathcal{O}}_E$. We define the subgroups $U(N, \mathfrak{p}^n) \subset U(N, \mathbb{A}_F^f)$ to be

$$U(N, \mathfrak{p}^n) := \{g \in U(N, \hat{\mathcal{O}}_F) \subset \mathrm{GL}_N(\hat{\mathcal{O}}_E) \mid g \equiv I_N(\mathfrak{p}^n \mathcal{O}_E)\}.$$

For any finite place v of F, let $U(N, \mathfrak{p}^n)_v = U(N, \mathfrak{p}^n) \cap U(N)_v$. At the expense of possibly enlarging the set S_f, note that for all $v \notin S \cup \{v_\mathfrak{p}\}$, the subgroup $U(N, \mathfrak{p}^n)_v$ is a hyperspecial maximal compact subgroup of $U(N)_v$. This gives level structures on the quasisplit group $U(N)$. If $H = U(N_1) \times \cdots \times U(N_r)$ is a product of quasisplit unitary groups, we define level subgroups $H(\mathfrak{p}^n) = U(N_1, \mathfrak{p}^n) \times \cdots \times U(N_r, \mathfrak{p}^n)$.

Let (G, ξ) be an inner form of $U(N, F)$ defined with respect to a Hermitian inner product and with prescribed signatures $U(a_v, b_v)$ at the archimedean places. We require that G_v be compact at the archimedean places contained in S_0: this ensures that the group G is anisotropic. Following Lemma 8, if N is odd, the group G can be chosen so that G_v is quasisplit at all finite places. If N is even, then G is determined by choosing at most one place $v \in S_f$, up to again enlarging S_f. Once that choice is made, the group G can be chosen to be quasisplit away from $\{v\} \cup S_\infty$. In both cases, this group G is realized as an inner form (G, ξ) as in Section 2.2.2.

For each finite $v \notin S_f$, the inner twist induces isomorphisms $\xi_v : G_v \simeq U(N)_v$. For each natural number n, we fix a compact subgroup $K(\mathfrak{p}^n) = \prod_v K_v(\mathfrak{p}^n)$ of $G(\mathbb{A}_F)$ as follows: at all finite $v \notin S$, we let $K_v(\mathfrak{p}^n) = \xi_v^{-1}(U(N, \mathfrak{p}^n)_v)$; at $v \in S_f$, the subgroup $K_v(\mathfrak{p}^n)$ is an arbitrary open compact subgroup fixed once and for all independently of n; at the archimedean places we let $K_v(\mathfrak{p}^n) \simeq U_{a_v}(\mathbb{R}) \times U_{b_v}(\mathbb{R})$ be a maximal compact subgroup. Let $K_f(\mathfrak{p}^n) = \prod_{v < \infty} K_v(\mathfrak{p}^n)$ and $K_\infty(\mathfrak{p}^n) = \prod_{v \mid \infty} K_v(\mathfrak{p}^n)$. We may use the notation K_v instead of $K_v(\mathfrak{p}^n)$ for $v \neq v_\mathfrak{p}$. We extend these definitions to products of unitary groups.

We now define the (cocompact since G is anisotropic) lattices

$$\Gamma(\mathfrak{p}^n) := G(F) \cap K_f(\mathfrak{p}^n).$$

Recall that $G_\infty = \prod_{v \mid \infty} G_v$ and let $X_G = G_\infty / K_\infty Z_{G_\infty}$. Assume that G_∞ has at least one noncompact factor. The diagonal embedding $\Gamma(\mathfrak{p}^n) \hookrightarrow \prod_{v \mid \infty} G_v$ induces an action $\Gamma(\mathfrak{p}^n) \curvearrowright X_G$, and we let $X(\mathfrak{p}^n) := \Gamma(\mathfrak{p}^n) \backslash X_G$. We start by comparing them to their disconnected counterparts realized as adelic double quotients. Let

$$Y(\mathfrak{p}^n) = G(F) \backslash G(\mathbb{A}_F) / K(\mathfrak{p}^n) Z_G(\mathbb{A}_F).$$

The quotient $Y(\mathfrak{p}^n)$ is a disjoint union of finitely many connected locally symmetric spaces, each associated to a conjugate ot $K(\mathfrak{p}^n)$. In particular, the summand corresponding to $K(\mathfrak{p}^n)$ is $X(\mathfrak{p}^n)$.

Proposition 40. *Let G be an inner form of $U(N)$ and $Y(\mathfrak{p}^n)$ be defined as above. The cardinality of the set of components $\pi_0(Y(\mathfrak{p}^n))$ is bounded independently of n.*

Proof. We adapt an argument from [16, Section 2]. Considering G as a subgroup of GL_N / E, let $\det : G \to U(1, E/F)$ be the determinant map and let $G^1 = \ker(\det)$. This map induces a fibering of

$Y(\mathfrak{p}^n)$ over

$$U(1, F)\backslash U(1, \mathbb{A}_F)/\det(Z(\mathbb{A}_F)K(\mathfrak{p}^n)).$$

The fibers are adelic double quotients for the group G^1, which is simply connected and has at least one noncompact factor at infinity. So by [41, 7.12], the group G^1 satisfies strong approximation with respect to the set S_∞ and $G^1(F)$ is dense in $G^1(\mathbb{A}_F^f)$, making the fibers connected. Thus we find that

$$\pi_0(Y(\mathfrak{p}^n)) \simeq U(1, F)\backslash U(1, \mathbb{A}_F)/\det(Z(\mathbb{A}_F)K(\mathfrak{p}^n)) = E^1\backslash \mathbb{A}_E^1/\det(Z(\mathbb{A}_F)K(\mathfrak{p}^n)).$$

Now the image $\det(Z(\mathbb{A}_F))$ is the subgroup $(\mathbb{A}_E^1)^N$ of $\mathbb{A}_E^1$. For each finite place w, the factor corresponding to E_w in the quotient $\mathbb{A}_E^1/(\mathbb{A}_E^1)^N$ is a finite set. It follows that by increasing the level in powers of a single prime $\mathfrak{p}$, one can only produce a bounded number of components. $\qquad\square$

We now fix a unitary irreducible admissible representation $\pi_\infty = \otimes_{v|\infty}\pi_v$ of G_∞ with trivial central character and such that π_v is the trivial representation if G_v is compact. Denote

$$m(\pi_\infty, \mathfrak{p}^n) := \dim \operatorname{Hom}_{G_\infty}(\pi_\infty, L^2(\Gamma(\mathfrak{p}^n)\backslash G_\infty)). \tag{28}$$

Since $X(\mathfrak{p}^n)$ is one of the connected components of $Y(\mathfrak{p}^n)$, we have

$$m(\pi_\infty, \mathfrak{p}^n) \leq \dim \operatorname{Hom}_{G_\infty}(\pi_\infty, L^2(Y(\mathfrak{p}^n))) = \sum_{\pi=\pi_\infty\otimes\pi_f} m(\pi)\dim\pi_f^{K_f(\mathfrak{p}^n)}. \tag{29}$$

We will be interested in the asymptotics of the multiplicities $m(\pi_\infty, \mathfrak{p}^n)$ as $n \to \infty$.

4.2. *Choice of test functions.* We define test functions whose traces will compute the multiplicity of archimedean representations at level $\mathfrak{p}^n$. Recall that μ_v denotes the Haar measure on G_v.

Definition 41. At each finite place v, let $f_v(\mathfrak{p}^n) := 1_{K_v(\mathfrak{p}^n)}/\mu_v(K_v(\mathfrak{p}^n))$.

Definition 42. Let $v \in S_0$ be an archimedean place such that G_v is compact. Let $f_v(\mathfrak{p}^n)$ be equal to the constant function $f_v = \mu_v(G_v)^{-1}$.

The traces of these test functions count the dimension of spaces of $K(\mathfrak{p}^n)$-fixed vectors. At $v \in S_0$, they only detect the trivial representation and have vanishing trace on all other representations of G_v. We want functions that play the same role at the noncompact archimedean places: they should detect representations π_v contained in a specific subset $\Pi_{\psi_v}^0 \subset \Pi_{\psi_v}$ and vanish on $\Pi_{\psi_v} \setminus \Pi_{\psi_v}^0$. The key is that we will only be working with Arthur packets attached to parameters ψ all having one specific ψ_∞. As such, the test function at an infinite place v only needs to isolate $\pi_v \in \Pi_v^0$ from the other finitely many representations in the same packet.

Lemma 43. *Let ψ_v be a local Arthur parameter with associated Arthur packet Π_{ψ_v}. Fix a subset $\Pi_{\psi_v}^0 \subset \Pi_{\psi_v}$. Then there exists a function $f_v^0 \in \mathcal{H}(G_v)$ such that*

$$\operatorname{tr}\pi_v(f_v^0) = \begin{cases} 1, & \pi_v = \Pi_{\psi_v}^0, \\ 0, & otherwise, \end{cases} \quad \pi_v \in \Pi_v.$$

Proof. This follows directly from linear independence of characters for admissible representations. If v is archimedean this was proved by Harish-Chandra in [22]. $\quad\square$

Definition 44. Let v be a noncompact archimedean place, let ψ_v be an Arthur parameter and fix a subset $\Pi^0_{\psi_v} \subset \Pi_{\psi_v}$. Let $f_v(\mathfrak{p}^n) = f_v(\mathfrak{p}^n, \Pi^0_{\psi_v})$ be the function f^0_v described above.

Definition 45. Let the function $f(\mathfrak{p}^n)$ be defined as $f(\mathfrak{p}^n) = \prod_v f_v(\mathfrak{p}^n_v)$. We will also denote $f_f(\mathfrak{p}^n) = \prod_{v \nmid \infty} f_v(\mathfrak{p}^n_v)$.

Given a choice of ψ_∞ and $\Pi^0_{\psi_\infty}$, the function $f(\mathfrak{p}^n)$ satisfies the assumption of Theorem 34: it is factorizable and has nonnegative trace on $\pi \in \Pi_\psi$.

Proposition 46. *Let $\psi \in \Psi(G, \xi)$. For each $v \in S_\infty \setminus S_0$, fix a subset $\Pi^0_{\psi_v}$ and a corresponding function $f(\mathfrak{p}^n) = f(\mathfrak{p}^n, \Pi^0_{\psi_v})$. Then we have*

$$\operatorname{tr} R_{\mathrm{disc},\psi}(f(\mathfrak{p}^n)) = \sum_\pi m(\pi) \dim \pi_f^{K_f(\mathfrak{p}^n)}$$

where the sum is taken over representations $\pi = (\otimes_{v|\infty} \pi_v) \otimes \pi_f \in \Pi_\psi$ such that for archimedean v, the representation π_v is trivial if $v \in S_0$ and $\pi_v \in \Pi^0_v$ otherwise.

Proof. As stated in Theorem 23, the distribution $\operatorname{tr} R_{\mathrm{disc},\psi}(f)$ computes the sum of $\operatorname{tr} \pi(f) = \prod_v \operatorname{tr} \pi_v(f_v)$ over all representations in the packet Π_ψ. At the finite places, the trace of convolution by the characteristic function of a compact open subgroup K_v is equal to the product $\mu_v(K_v) \cdot \dim \pi_v^{K_v}$. For archimedean places $v \in S_0$, the representations π_v are finite-dimensional so the only representation with a K_v-fixed vector is the trivial representation. At $v \in S_\infty \setminus S_0$, the function $f_v(\mathfrak{p}^n)$ was chosen precisely to detect $\pi_v \in \Pi^0_{\psi_v}$. $\quad\square$

The key input allowing us to compare multiplicity growth on G and $H \in \mathcal{E}(H)$ is a fundamental lemma for congruence subgroups, proved by Ferrari [17].

Theorem 47 [17, Theorem 3.2.3]. *Let $\mathfrak{p}$ be a prime of F with localization $F_{v_\mathfrak{p}}$ and residue field $k_\mathfrak{p}$. Let $\mathrm{Nm}(\mathfrak{p})$ be the cardinality of $k_\mathfrak{p}$ and let p be its characteristic. Assume that $p > N^2[F : \mathbb{Q}] + 1$. Let $H \in \mathcal{E}(G)$, and $d(G, H) = (\dim G - \dim H)/2$. Then the functions*

$$f_{v_\mathfrak{p}}(\mathfrak{p}^n) = \frac{1_{K_{v_\mathfrak{p}}(\mathfrak{p}^n)}}{\mu_{v_\mathfrak{p}}(K_{v_\mathfrak{p}}(\mathfrak{p}^n))} \quad and \quad f^H_{v_\mathfrak{p}}(\mathfrak{p}^n) = \mathrm{Nm}(\mathfrak{p})^{-n \cdot d(G,H)} \frac{1_{K_{v_\mathfrak{p}}(\mathfrak{p}^n)^H}}{\mu_{v_\mathfrak{p}}(K_{v_\mathfrak{p}}(\mathfrak{p}^n)^H)}$$

form a transfer pair.

4.3. *Adaptation of previous limit multiplicity results.*
Here, we collect all the results so far and import known upper bounds from the literature to prove our main limit multiplicity results. We start with a discussion of central characters.

4.3.1. *Central character data.* Our initial discussion of the discrete spectrum in Section 2.3.1 included the prescription of a subgroup $\mathfrak{X} \subset Z_G(\mathbb{A}_F)$, and we recalled in Proposition 15 that in the case where $G = U_{E/F}(N)$, the central character of representations in a packet Π_ψ is determined explicitly in terms of ψ. We now need to extend this discussion to central characters of $H \in \mathcal{E}(G)$. For this, we will denote $\mathfrak{X}_G = Z_G(\mathbb{A}_F)$, and $\mathfrak{X}_H = Z_H(\mathbb{A}_F)$. As described in [5, Section 3.2], the group $\mathfrak{X}_G$ can be viewed canonically as a subgroup of $\mathfrak{X}_H$, and we can speak of $(\mathfrak{X}_G, \omega)$ as a central character datum of H, though it is not properly speaking a character $Z_H(\mathbb{A}_F)$. This can be extended inductively to $H \in \mathcal{EH}(G)$. From Proposition 15 and the definition of the embeddings in (7), we get the following.

Lemma 48. *Let G be a unitary group and $(H, \xi_H) \in \mathcal{E}(G)$. Let $\psi \in \Psi(G, \xi)$ be associated to the central character datum $(\mathfrak{X}_G, \omega)$, and be such that $\psi = \xi_H \circ \psi^H$. Then the central character datum $(\mathfrak{X}_G, \omega')$ associated to ψ^H is determined by ω and ξ_H.*

4.3.2. *Sets of parameters.*

Definition 49. Let G be a reductive group, ψ_∞ a parameter of G_∞, and $(\mathfrak{X}, \omega)$ a central character. We denote by $\Psi(G, \psi_\infty, \omega)$ the set of $\psi \in \Psi(G)$ such that $(\psi)_\infty = \psi_\infty$ and such that the associated representations in the packet of ψ have central character ω.

Definition 50. Letting G be as above, and $(H, \xi) \in \mathcal{E}(G)$ or $H = G$, we define for $f \in \mathcal{H}(H)$,

$$I_{\mathrm{disc}, \psi_\infty, \omega}^H(f) = \sum_{\psi \in \Psi(G, \psi_\infty, \omega)} I_{\mathrm{disc}, \psi}^H(f). \tag{30}$$

We will need to rewrite the right-hand side of (30) as a sum over parameters of H. By definition, for each $\psi \in \Psi(G, \eta)$, we have

$$I_{\mathrm{disc}, \psi}^H(f) = \sum_{c^H \overset{\eta \circ \xi}{\to} c(\psi)} I_{\mathrm{disc}, c^H}^H(f),$$

for $c^H \in \mathcal{C}(H)$. But following the main theorem of the spectral expansion of the trace formula [38, Proposition 4.3.4] applied to each of the simple factors of H, shows that $I_{\mathrm{disc}, c^H}^H(f) = 0$ unless $c^H = c(\psi^{H,N})$ is attached to a parameter $\psi^H = (\tilde{\psi}^H, \psi^{N,H})$. By the assumption $c^H \overset{\eta \circ \xi}{\longrightarrow} c(\psi)$, we mush have $\psi^{H,N} = \psi^N$, and $\tilde{\psi}$ must factor through ^LH. Thus we can rewrite

$$\sum_{\psi \in \Psi(G, \psi_\infty, \omega)} I_{\mathrm{disc}, \psi}^H(f) = \sum_{\psi \in \Psi(G, \psi_\infty, \omega)} \sum_{\psi^H \mapsto \psi} I_{\mathrm{disc}, \psi^H, \omega}^H(f).$$

In the arguments of this section, we will work with three families of groups, and three sets of parameters, which we describe now.

- The group $G \in \tilde{\mathcal{E}}_{\mathrm{sim}}(N)$ is the inner form of $U_{E/F}(N)$ for which we ultimately want to produce bounds. These will be obtained in Theorem 56 by taking a sum over $\Psi(G, \psi_\infty, 1)$, for the trivial central character of Z_G.

- The group $H = H_\psi = U(N_1) \times U(N_2)$ with $N_1 + N_2 = N$ belongs to the endoscopic datum $(H, \xi, s) \in \mathcal{E}(G)$ whose stable trace gives the upper bounds in Theorem 34. In the lead-up to Theorem 34, for each $\psi \in \Psi(G, \xi)$, we singled out a parameter ψ^H such that $\xi \circ \psi^H = \psi$. Through this choice, the parameter ψ_∞ of G_∞ determines a unique parameter ψ_∞^H of H_∞. In Proposition 54, the sum will be taken over the set $\Psi(H, \psi_\infty^H, \omega)$ for a suitable central character ω.

- The difference between the distribution giving the upper bounds in Theorem 34 and $I_{\mathrm{disc}, \psi^H}^H$ is expressed in Corollary 39 in terms of hyperendoscopic data (H_q, ξ_q) for H. We will consider parameters such that $\xi_q(\psi^{H_q}) \in \Pi(H, \psi_\infty^H, \omega)$. We have $\psi_\infty^{H_q} = \psi_\infty^{H_q^1} \times \psi_\infty^{H_q^2}$ and we will give upper bounds on the multiplicities of representations associated to each of these factors in Propositions 51 and 52.

4.3.3. *Upper bounds for hyperendoscopic groups.* We start by adapting limit multiplicity results of Savin [43], which will form the basis for our inductive proof. Since Savin's result applies to semisimple groups, we pay attention to the central characters and components of locally symmetric spaces. We first give bounds for bounded parameters. The result is stated in terms of any G and H, but will be specialized to $G = U(N_2)$ and $H = H_q^2$.

Proposition 51. *Let $G \in \tilde{\mathcal{E}}_{\mathrm{sim}}(N)$ and $\mathcal{H} \in \mathcal{HE}(G)$ be a hyperendoscopic group. Let $\psi_\infty \in \Psi(G_\infty)$ be a bounded parameter with regular infinitesimal character. Let $(\mathfrak{X}_G, \omega)$ be a central character datum for G such that $\omega|_{(\mathfrak{X}_G \cap G_\infty)}$ is the central character associated to ψ_∞ by Proposition 15. Let $v_\mathfrak{p}$ be an unramified finite place of F, associated to the prime $\mathfrak{p}$, and let $f(\mathfrak{p}^n) = \prod_v f_v(\mathfrak{p}^n) \in \mathcal{H}(H)$ be such that:*

- *$f_v(\mathfrak{p}^n)$ is independent of n if $v \neq v_\mathfrak{p}$.*
- *$f_{v_\mathfrak{p}}(\mathfrak{p}^n) = 1_{K(\mathfrak{p}^n)}/\mu(K(\mathfrak{p}^n))$, for $K(\mathfrak{p}^n)$ as in Section 4.1.*
- *$f(\mathfrak{p}^n)$ satisfy the assumptions of Theorem 34.*

Then

$$|I_{\mathrm{disc}, \psi_\infty, \omega}^H(f(\mathfrak{p}^n))| \ll Nm(\mathfrak{p}^n)^{\dim H - 1}.$$

Proof. Since the infinitesimal character of ψ_∞ is regular, we can equate

$$I_{\mathrm{disc}, \psi_\infty, \omega}^H(f(\mathfrak{p}^n)) = \mathrm{tr}\, R_{\mathrm{disc}, \psi_\infty, \omega}^H(f(\mathfrak{p}^n)) = \sum_{\psi_\infty^H \mapsto \psi_\infty} \sum_{\psi^H \in \Psi(H, \psi_\infty^H, \omega')} \sum_{\pi \in \Pi_{\psi^H}} m(\pi)\, \mathrm{tr}\, \pi(f(\mathfrak{p}^n)),$$

where $(\mathfrak{X}, \omega')$ is the central character datum associated to ψ as in Lemma 48. Since the first sum is finite, we will bound

$$\left| \sum_{\psi^H \in \Psi(H, \psi_\infty^H, \omega')} \sum_{\pi \in \Pi_{\psi^H}} m(\pi)\, \mathrm{tr}\, \pi(f(\mathfrak{p}^n)) \right|.$$

For each π, we have $\mathrm{tr}\, \pi(f(\mathfrak{p}^n)) = \prod_v \mathrm{tr}\, \pi_v(f_v(\mathfrak{p}^n))$. At $v \mid \infty$, the packet is always the same, so $|\mathrm{tr}\, \pi_\infty(f_\infty)|$ is uniformly bounded. For each finite v, there is an open compact subgroup $K_v' \subset G_v$, depending on n only if $v = v_\mathfrak{p}$, such that $|\mathrm{tr}\, \pi_v f_v(\mathfrak{p}^n)| \neq 0 \implies \dim \pi_v^{K_v'} \neq 0$. Indeed, since $f_v(\mathfrak{p}^n)$ is

K_v-finite, where K_v is a maximal compact subgroup, there is a subgroup $K'_v \leq K_v$ of finite index such that $f_v(\mathfrak{p}^n)$ is K'_v-invariant, so that convolution by f_v is a projection onto $\pi_v^{K'_v}$. At all but finitely many places the group K_v is hyperspecial, we have $f_v(\mathfrak{p}^n) = 1_{K_v}$ and we can take $K'_v = K_v$. Thus we have $|\operatorname{tr} \pi_v f_v| < C(f_v) \dim \pi^{K'_v}$ and by Bernstein's uniform admissibility [8], the right-hand side is bounded uniformly, with $C(f_v(\mathfrak{p}^n)) = 1$ at $v \notin S$. At $v = v_p$, we have $K'_v = K_v(\mathfrak{p}^n)$. Let $K'(\mathfrak{p}^n) = \prod_{v < \infty} K'_v$. From our restriction on the central character, we thus have

$$\left| \sum_{\Psi(H, \psi_\infty^H, \omega')} \sum_{\pi \in \Pi_{\psi H}} m(\pi) \operatorname{tr} \pi(f(\mathfrak{p}^n)) \right| \leq C(\psi_\infty, S) \sum_{\substack{\pi : \pi_\infty \in \Pi_{\psi_\infty^H} \\ \omega(\pi) = \omega'}} m(\pi) \dim \pi_f^{K'(\mathfrak{p}^n)},$$

where $C(\psi_\infty, S)$ is a constant depending only on ψ_∞ and S.

Since ψ_∞^H is bounded, any representation $\pi_\infty \in \Pi_{\psi_\infty^H}$ is tempered, which implies that $\pi \in \Pi_\psi^H$ occur in the cuspidal part of the discrete spectrum [52, Theorem 4.3]. Thus for each $\pi_\infty \in \Pi_{\psi_\infty^H}$ we have

$$\sum_{\substack{\pi = \pi_\infty \cdot \pi_f \\ \omega(\pi) = \omega'}} m(\pi) \dim \pi_f^{K'(\mathfrak{p}^n)} \leq \dim \operatorname{Hom}_{H_\infty}(\pi_\infty, L^2_{\mathrm{cusp}}(H(F)\backslash H(\mathbb{A}_F), \omega')^{K'(\mathfrak{p}^n)}).$$

The right-hand side of the inequality is equal to

$$m(\pi_\infty, \mathfrak{p}^n, \omega') := \dim \operatorname{Hom}_{H_\infty}(\pi_\infty, L^2_{\mathrm{cusp}}(H(F)\backslash H(\mathbb{A}_F)/K'(\mathfrak{p}^n), \omega')). \tag{31}$$

The space $Y_H^*(\mathfrak{p}^n) := H(F)\backslash H(\mathbb{A}_F)/K'(\mathfrak{p}^n)$ carries commuting actions of H_∞ and $Z_H(\mathbb{A}_F)$, inducing representations on $L^2_{\mathrm{cusp}}(Y_H^*(\mathfrak{p}^n))$. For n large enough, the character ω' is trivial on $\mathfrak{X}_G \cap K'(\mathfrak{p}^n)$, and thus appears in the representation of $\mathfrak{X}_G$ on $L^2_{\mathrm{cusp}}(Y_H^*(\mathfrak{p}^n))$. It is this ω'-isotypic subspace that we denote by $L^2_{\mathrm{cusp}}(Y_H^*(\mathfrak{p}^n), \omega')$.

To bound $m(\pi_\infty, \mathfrak{p}^n, \omega')$, consider first the case where the central character datum for $\mathfrak{X}_H$ is trivial: this setup is similar to that of Proposition 40. We have $H = U(N_1) \times \cdots \times U(N_r)$; let $H^1 = \mathrm{SU}(N_1) \times \cdots \times \mathrm{SU}(N_r)$. The representation π_∞ of H_∞ restricts to an irreducible representation ρ_∞ of H_∞^1; see [2, Section 2]. Let

$$X_H(\mathfrak{p}^n) = H^1(F)\backslash H^1(\mathbb{A})/K^1(\mathfrak{p}^n), \quad K^1(\mathfrak{p}^n) = K'(\mathfrak{p}^n) \cap G^1(\mathbb{A}).$$

The group H^1 is simply connected and has no compact factors at infinity, so $X_H(\mathfrak{p}^n)$ is connected [41]. Following a result of Savin [43], we have

$$m(\rho_\infty, \mathfrak{p}^n) := \dim \operatorname{Hom}_{H_\infty^1}(\rho_\infty, L^2_{\mathrm{cusp}}(X_H(\mathfrak{p}^n))) \asymp \operatorname{Vol}(X_H(\mathfrak{p}^n)) \asymp Nm(\mathfrak{p})^{n \cdot \dim H^1}.$$

We now consider general central characters. The space $Y_H^*(\mathfrak{p}^n)$ is a disjoint union of finitely many locally symmetric spaces, associated to conjugates of $K^1(\mathfrak{p}^n)$, and the theorem of Savin applies to each of them. Let $T = H/H^1$, and let ν denote the quotient map, through which all central characters factor, see Proposition 15. Following [14, 2.7.1], the set $\pi_0(Y_H^*(\mathfrak{p}^n))$ is a torsor for the finite group

$$T_{\mathfrak{p}^n} := T(\mathbb{A}_F)/T(F)\nu(K'(\mathfrak{p}^n)).$$

Denote by $\mathfrak{X}_{G,\mathfrak{p}^n}$ the image of $\mathfrak{X}_G$ in this quotient. The action of $\mathfrak{X}_{G,\mathfrak{p}^n}$ on $\pi_0(Y^H(\mathfrak{p}^n))$ is induced by multiplication in $T_{\mathfrak{p}^n}$, thus $\pi_0(Y_H^*(\mathfrak{p}^n))$ is a finite union of $[T_{\mathfrak{p}^n} : \mathfrak{X}_{G,\mathfrak{p}^n}]$ many principal homogeneous spaces for $\mathfrak{X}_{G,\mathfrak{p}^n}$. Thus as a $\mathfrak{X}_{G,\mathfrak{p}^n}$-representation, the space $\mathrm{Hom}(\pi_\infty, L^2_{\mathrm{cusp}}(Y_H^*(\mathfrak{p}^n)))$ caries finitely many copies of the regular representation of $\mathfrak{X}_{G,\mathfrak{p}^n}$, and all characters of $\mathfrak{X}_G$ factoring through $\mathfrak{X}_{G,\mathfrak{p}^n}$ occur with equal multiplicity. The group $\mathfrak{X}_G$ is the adelic points of torus of diagonal matrices isomorphic to $U(1)$, and $T \simeq U(1)^r$. So each character ω' of $\mathfrak{X}_G$ factoring through $\mathfrak{X}_{G,\mathfrak{p}^n}$ does so with multiplicity

$$m(\pi_\infty, \mathfrak{p}^n, \omega') = m(\rho_\infty, \mathfrak{p}^n)[T_{\mathfrak{p}^n} : T_{\mathfrak{p}^n}^H] = m(\rho_\infty, \mathfrak{p}^n)\frac{[T(1):T(\mathfrak{p}^n)]}{[\mathfrak{X}_{G,1} : \mathfrak{X}_{G,\mathfrak{p}^n}]} \asymp \mathrm{Nm}(\mathfrak{p}^n)^{\dim H - 1}.$$

Summing over all π_∞ in Π_{ψ_∞}, we conclude. $\qquad\square$

We now give bounds for parameters where $\psi(\mathrm{SL}_2)$ is maximally large. In the final proof, the group G will be specialized to $U(N_1)$.

Proposition 52. *Let $G \in \tilde{\mathcal{E}}_{\mathrm{sim}}(N)$, and $\psi_\infty \in \Psi(G_\infty)$. Let $f(\mathfrak{p}^n) = \prod_v f_v(\mathfrak{p}^n)$ satisfy the same assumptions as in Proposition 51. Let $\psi_\infty \in \Psi(G_\infty)$ be a parameter with regular infinitesimal character and such that $\psi_\infty|_{\mathrm{SL}_2} = \nu(N)$. Let $(\mathfrak{X}_G, \omega)$ be a central character datum. Then there is a constant M depending only on G, ψ_∞, and the set S of bad places (and in particular neither on n nor on ω) such that*

$$|I^G_{\mathrm{disc},\psi_\infty,\omega}(f(\mathfrak{p}^n))| < M.$$

Proof. The proof is a simplified version of that of Proposition 51. The restriction of the infinitesimal character and on the possible representations at infinity gives

$$|I^G_{\mathrm{disc},\psi_\infty,\omega}(f(\mathfrak{p}^n))| \leq C(\psi_\infty, S) \sum_{\substack{\pi:\pi_\infty \in \Pi_{\psi_\infty} \\ \omega(\pi)=\omega}} m(\pi) \dim \pi_f^{K'(\mathfrak{p}^n)}.$$

The assumption on the Arthur SL_2 implies that the representations $\pi_\infty \in \Pi_{\psi_\infty}$ are one-dimensional; see, e.g., [4, Section 5]. Thus they factor through the determinant map ν, and, as above, through the action of the quotient $T(\mathbb{A}_F)/T(F)$. It follows that in this case the multiplicity $m(\pi_\infty, \mathfrak{p}^n, \chi)$ is bounded above by $|T(\mathbb{A}_F)/T(F)\nu(K(\mathfrak{p}^n))|$. Recall here that $\mathfrak{X}_G = Z(\mathbb{A}_F)$. If ω were trivial, then the representations contributing to $m(\pi_\infty, \mathfrak{p}^n, \omega)$ would be bounded above by the size of the quotient $|T(\mathbb{A}_F)/T(\mathbb{Q})\nu(K(\mathfrak{p}^n) \cdot Z(\mathbb{A}_F))|$, which we showed in Proposition 15 to be bounded independently of n. But by the proof of Proposition 51, the representation of $\mathfrak{X}_G$ on $\mathrm{Hom}(\pi_\infty, L^2_{\mathrm{cusp}}(Y^*H(\mathfrak{p}^n)))$ factors through a sum of copies of the regular representation of a finite quotient $\mathfrak{X}_{G,\mathfrak{p}^n}$. As such, all characters of $\mathfrak{X}_G$ appearing in the quotient do so with equal multiplicity. Thus the bound M also holds for ω. $\qquad\square$

4.4. *Limit multiplicity for G.* Before we assemble the results for various endoscopic groups, we bound the number of central character data $(\mathfrak{X}_H, \omega)$ of a given level and restriction to $\mathfrak{X}_G$.

Lemma 53. *Let $H \in \mathcal{E}(G)$. Let $(\mathfrak{X}_G, \omega)$ be a central character datum for G. For each n, fix a level structure $K_f^H(\mathfrak{p}^n)$ as in Section 4.1. Define*

$$\Xi(\omega, \mathfrak{p}^n) = \{(\mathfrak{X}_H, \omega_H) : \omega_H|_{\mathfrak{X}_G} = \omega, \omega_H(\mathfrak{X}_H \cap K_f^H(\mathfrak{p}^n)) = \omega_H(\mathfrak{X}_H \cap Z_H(F)) = 1\}.$$

Then we have $|\Xi(\omega, \mathfrak{p}^n)| \ll \mathrm{Nm}(\mathfrak{p}^n)$.

Proof. Central characters of $H = U_{E/F}(N_1) \times U_{E/F}(N_2)$ are products $\omega_H = \omega_1 \times \omega_2$ of characters of the respective centers. The condition upon restriction to $Z_H(F)$ implies that these are of the form $\omega_i = \theta_i \circ \det$ for θ_i a Hecke character of $\mathbb{A}_E^1$. Given a choice of θ_1, the condition that $\omega_H \mid_{\mathfrak{x}_G} = \omega$ restricts θ_2 to at most N_2 different characters. The restriction on conductor thus implies that $|\Xi(\chi, \mathfrak{p}^n)| \ll |\Xi(\mathfrak{a}\mathfrak{p}^n)|$, where $\Xi(\mathfrak{a}\mathfrak{p}^n)$ consists of Hecke characters θ_1 of $\mathbb{A}_E^1/E^1$ whose conductor divides $\mathfrak{a}\mathfrak{p}^n$; the presence of a conductor away from $\mathfrak{p}$ comes from the possibility that at places $v \in S_f$, the (fixed) subgroup $K_v^H(\mathfrak{p}^n)$ is not maximal. The number of such characters grows like $\mathrm{Nm}(\mathfrak{p}^n)$. $\square$

We now assemble the results of Section 4.3 to give upper bounds for the contribution of parameters where all but one summand have trivial Arthur SL_2. We start by bounding the contribution of each hyperendoscopic group of H_ψ.

Proposition 54. *Let $G \in \tilde{\mathcal{E}}_{\mathrm{sim}}(N)$, and $H = (U_{E/F}(N_1) \times U_{E/F}(N_2), \xi) \in \mathcal{E}(G)$. Let ψ_∞ and ψ_∞^H be such that*:

(1) $\psi_\infty = \xi \circ \psi_\infty^H$.

(2) $\psi_\infty = \psi_{\infty,1} \oplus \psi_{\infty,2}$ *with* $\psi_{\infty,1} \mid_{\mathrm{SL}_2} = \nu(N_1)$ *and* $\psi_{\infty,2} \mid_{\mathrm{SL}_2} = \nu(1)^{N_2}$.

(3) *Each* $\psi_{i,\infty}$ *factors through* $^L U(N_i)$.

Let $(\mathfrak{X}_G, \omega)$ be a central character datum for H consistent with to ψ_∞^H. Assume that $\mathfrak{p}$ is large enough to apply the results of Theorem 47. Let $H_q \in \mathcal{HE}(H)$, and let $f(\mathfrak{p}^n)$ be the sequence of test functions defined in Definition 45. Then

$$|I_{\mathrm{disc}, \psi_\infty^H, \omega}^{H_q}(f^{H_q}(\mathfrak{p}^n))| \ll Nm(\mathfrak{p}^n)^{N(N-N_1)}.$$

Proof. Since H_q is a hyperendoscopic group of H we have $H_q = H_q^1 \times H_q^2$, where ψ_i factors through H_q^i. The growth rate in the theorem is defined up to constants, so we can assume that $f^{H_q}(\mathfrak{p}^n) = f^{H_q^1}(\mathfrak{p}^n) \times f^{H_q^2}(\mathfrak{p}^n)$; this is true locally almost everywhere by the Fundamental Lemma. Indeed, at all $v \notin S \cup \{v_\mathfrak{p}\}$, the function $f_v^{H_q}(\mathfrak{p}^n)$ can be taken to be the characteristic function of a hyperspecial maximal compact subgroup of $H_{q,v}$. At $v = v_\mathfrak{p}$, we iterate the conclusion of Theorem 47 to get

$$f_{v_\mathfrak{p}}^{H_q}(\mathfrak{p}^n) = \mathrm{Nm}(\mathfrak{p})^{-n \cdot d(G, H_q)} \frac{1_{K_{v_\mathfrak{p}}(\mathfrak{p}^n)^{H_q}}}{\mu(K_{v_\mathfrak{p}}(\mathfrak{p}^n))} := \frac{\mathrm{Nm}(\mathfrak{p})^{-n \cdot d(G, H_q)}}{\mu(K_{v_p})(\mathfrak{p}^n)/\mu(K_{v_\mathfrak{p}}(\mathfrak{p}^n)^{H_q})} \varphi_{v_\mathfrak{p}}(\mathfrak{p}^n). \tag{32}$$

Write $\varphi_{v_\mathfrak{p}}(\mathfrak{p}^n) = \varphi_{v_\mathfrak{p}}^1(\mathfrak{p}^n) \times \varphi_{v_\mathfrak{p}}^2(\mathfrak{p}^n)$, and for $i = 1, 2$, let

$$\varphi^{H_q^i}(\mathfrak{p}^n) = \varphi_{v_\mathfrak{p}}^i(\mathfrak{p}^n) \cdot \prod_{v \neq v_\mathfrak{p}} f_v^{H_q^i}(\mathfrak{p}^n), \quad \varphi^{H_q}(\mathfrak{p}^n) = \varphi^{H_q^1}(\mathfrak{p}^n) \times \varphi^{H_q^2}(\mathfrak{p}^n).$$

Each of the two functions $\varphi^{H_q^i}(\mathfrak{p}^n)$ satisfies the identical assumptions of Propositions 51 and 52. We also recall that $H = U(N_1) \times U(N_2)$, and we shorten $U(N_i) = H^i$. We also have $\psi^H = \psi_1 \oplus \psi_2$ with ψ_i landing in $^L H^i$. Thus if we fix data $(\mathfrak{X}_{H^1}, \omega_1)$ and $(\mathfrak{X}_{H^2}, \omega_2)$ coming from H^1 and H^2 respectively,

we find that

$$\left| \sum_{\Psi(H, \psi_\infty^H, \omega_1 \times \omega_2)} I_{\mathrm{disc}, \psi^H}^{H_q}(\varphi^{H_q}(\mathfrak{p}^n)) \right| \leq \sum_{\Psi(H, \psi_\infty^H, \omega_1 \times \omega_2)} |I_{\mathrm{disc}, \psi_1}^{H_q^1}(\varphi^1(\mathfrak{p}^n))| \cdot |I_{\mathrm{disc}, \psi_2}^{H_q^2}(\varphi^2(\mathfrak{p}^n))|$$

$$= |I_{\mathrm{disc}, \psi_{\infty,1}, \omega_1}^{H_q^1}(\varphi^1(\mathfrak{p}^n))| \times |I_{\mathrm{disc}, \psi_{\infty,2}, \omega_2}^{H_q^2}(\varphi^2(\mathfrak{p}^n))|$$

$$\ll M \cdot Nm(\mathfrak{p}^n)^{\dim H_q^2 - 1}.$$

The quantity in the left-hand side above isn't quite what we want to measure. First, we want to replace the choice of a pair of central characters $\omega_1 \times \omega_2$ by a sum over all parameters with central character datum $(\mathfrak{X}_G, \omega)$. In Lemma 53, we saw that the number of products $\omega_1 \times \omega_2$ of level $\mathfrak{p}^n$ which restrict to ω on $\mathfrak{X}_G$ is $\ll Nm(\mathfrak{p}^n)$. Second, we slightly modify the test functions. From (32), we have

$$f^{H_q}(\mathfrak{p}^n) = C(G, H_q, n)\varphi^{H_q}(\mathfrak{p}^n), \quad C(G, H_q, n) \asymp Nm(\mathfrak{p})^{n \cdot d(G, H_q)}.$$

Thus combining our upper bounds with these modifications we obtain

$$|I_{\mathrm{disc}, \psi_\infty^H, \omega}^{H_q}(f^{H_q}(\mathfrak{p}^n))| \ll Nm(\mathfrak{p}^n)^{(1 + d(G, H_q))} \left| \sum_{\psi^H \in \Psi(H, \psi_\infty^H, \omega_1 \times \omega_2)} I_{\mathrm{disc}, \psi^H}^{H_q}(\varphi^{H_q}(\mathfrak{p}^n)) \right|$$

$$\ll Nm(\mathfrak{p}^n)^{d(G, H_q) + \dim H_q^2}$$

$$= Nm(\mathfrak{p}^n)^{\dim(G)/2 - \dim(H_q^1)/2 + \dim(H_q^2)/2}.$$

Recall that $\dim G = N^2$, and that since the dual group of H_q^1 receives an N_1-dimensional irreducible representation of SL_2, we have $\dim(H_q^1) = N_1^2$. Finally, it follows that $\dim(H_q^2) \leq (N - N_1)^2$, which gives us the desired upper bounds. $\square$

Remark 55. Note that the only situation in which this upper bound has a chance of being sharp is when $\dim H_q^2 = (N - N_1)^2$, i.e., when $H_q = H$.

We have now collected all the facts leading up to our limit multiplicity theorem.

Theorem 56. *Let ψ_∞ be an Arthur parameter with regular infinitesimal character, and such that $\psi_\infty |_{\mathrm{SL}_2(\mathbb{C})} = \nu(2k) \oplus \nu(1)^{N-2k}$. Let $(\mathfrak{X}_G, 1)$ be the trivial central character. Fix $\Pi_{\psi_\infty}^0 \subset \Pi_{\psi_\infty}$. For each $\psi \in \Psi(G, \psi_\infty, 1)$, let*

$$\Pi_\psi^0 = \{\pi = \otimes_v' \pi_v \in \Pi_\psi \mid \pi_\infty \in \Pi_{\psi_\infty}^0.\}$$

Then

$$\sum_{\psi \in \Psi(G, \psi_\infty, 1)} \sum_{\pi \in \Pi_\psi^0} m(\pi) \dim \pi_f^{K(\mathfrak{p}^n)} \ll Nm(\mathfrak{p}^n)^{N(N-2k)}. \tag{33}$$

Proof. From Proposition 46, we take $f(\mathfrak{p}^n)$ as in Section 4.2 and write

$$\sum_{\psi \in \Psi(G, \psi_\infty, 1)} \sum_{\pi \in \Pi_\psi^0} m(\pi) \dim \pi_f^{K(\mathfrak{p}^n)} = \sum_{\psi \in \Psi(G, \psi_\infty, 1)} \operatorname{tr} R_{\text{disc}, \psi}(f(\mathfrak{p}^n))$$

$$= \sum_{\psi \in \Psi(G, \psi_\infty, 1)} I_{\text{disc}, \psi}(f(\mathfrak{p}^n)) \qquad \text{(Theorem 24)}$$

$$\leq \sum_{\psi \in \Psi(G, \psi_\infty, 1)} C(\psi) S(\psi, s_{H_\psi}, f(\mathfrak{p}^n)),$$

where the last inequality follows from the results of Section 3.2, where the notation $C(\psi)$ was defined, since $f(\mathfrak{p}^n)$ takes only positive values. The group H_ψ is determined by the localization ψ_v of ψ at any place v, and in particular by ψ_∞. Thus H_ψ is the same for any $\psi \in \Psi(G, \psi_\infty, 1)$ since by definition they all localize to the same ψ_∞. By the assumption on ψ_∞, we have $H_\psi = U_{E/F}(2k) \times U_{E/F}(N - 2k)$, and the parameters satisfy the assumptions of Proposition 54. To lighten the notation, we denote H_ψ by H and s_ψ by s_H for the end of the proof.

The parameter ψ^H corresponding to s_H under the bijection of Lemma 13, we have shown in Corollary 39 that

$$S(\psi, s_H, f(\mathfrak{p}^n)) = \iota(G, H) \sum_{\mathcal{H} \in \mathcal{HE}(H, \psi)} \iota(\mathcal{H}) I_{\text{disc}, \psi^H}^{H_q}(f^{H_q}(\mathfrak{p}^n)).$$

For each summand on the right-hand side, we sum over $\Psi(H, \psi^H, \omega)$, where $(\mathfrak{X}_G, \omega)$ is determined by ψ^H as in Lemma 48. We then apply Proposition 54 with $N_1 = 2k$. Note that we have ensured in Lemma 14 that the infinitesimal character of the representations of $H_{q,\infty}$ associated to all $\psi_\infty^{H_q}$ are regular. This gives us the following bounds:

$$\left| \sum_{\psi \in \Psi(H, \psi_\infty, 1)} I_{\text{disc}, \psi^H}^{H_q}(f^{H_q}(\mathfrak{p}^n)) \right| \ll Nm(\mathfrak{p}^n)^{N(N-2k)}.$$

We conclude by summing over the finitely many $H_q \in \mathcal{HE}(H, \psi)$. $\qquad\square$

5. Applications to growth of cohomology

We now apply the results of Section 4 to cohomology of arithmetic groups. This section is concerned with local questions at infinity, and the notation is different from the rest of the paper. From now until Section 5.3, G will be a Lie group.

5.1. *Cohomological representations.* Given a Lie group G, let $\tilde{G}$ denote the unitary dual of G.

Theorem 57 (Matsushima's formula [36]). *Let G be a connected semisimple Lie group with maximal compact subgroup K and complexified Lie algebra $\mathfrak{g}$. Let $\Gamma \subset G$ be a cocompact lattice and let $X_\Gamma = \Gamma \backslash G / K$. For $\pi \in \tilde{G}$, denote by $m(\pi, \Gamma)$ the multiplicity of π in the right-regular representation of*

G on $L^2(\Gamma \backslash G)$. Then

$$\dim(H^i(X_\Gamma, \mathbb{C})) = \sum_{\pi \in \tilde{G}} m(\pi, \Gamma) \dim(H^i(\mathfrak{g}, K; \pi)).$$

The $H^i(\mathfrak{g}, K; \pi)$ which appear in the right-hand side are the so-called $(\mathfrak{g}, K)$ cohomology groups of π. We say that π is *cohomological* if $H^*(\mathfrak{g}, K; \pi) \neq 0$; such representations were characterized by Vogan and Zuckerman [49].

Theorem 58 [49]. *Let G, $\mathfrak{g}$ be as above. Let K a maximal compact subgroup of G and $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{s}$ be the corresponding Cartan decomposition with $\mathfrak{k}$ the Lie algebra of K. The group G has finitely many cohomological representations π, and $H^i(\mathfrak{g}, K; \pi) \neq 0$ if and only if:*

(i) *π has the infinitesimal character of the trivial representation of G.*

(ii) *$\mathrm{Hom}_K(\pi, \wedge^i \mathfrak{s}) \neq 0$.*

Where the action of K on $\wedge^i \mathfrak{s}$ is induced by the adjoint representation.

The results apply only to semisimple groups: they are extended to $U(a, b)$ in [48], and condition (ii) above implies that cohomological representations have trivial central character. Below, we give a concrete parametrization of cohomological representations of $U(a, b)$ in terms of refinements of partitions of $a + b$ which are compatible with the signature (a, b). More details can be found in [6] and [20].

5.1.1. *Cohomological representations and ordered bipartitions.* In [49], cohomological representations $A_{\mathfrak{q}}$ are built from so-called θ-stable parabolic subalgebras $\mathfrak{q} = \mathfrak{l} \oplus \mathfrak{u}$. In [6, Section 5], Bergeron and Clozel show that for $U(a, b)$, the data of the algebra $\mathfrak{q}$ can be encoded in a choice of centralizing Levi subgroup $L(\mathfrak{q}) = \prod U(a_i, b_i) \subset U(a, b)$ whose Lie algebra is $\mathfrak{l}$. Thus $\mathfrak{q}$'s are parametrized by ordered tuples

$$B = ((a_1, b_1), \ldots, (a_r, b_r))$$

of pairs of nonnegative integers with $\sum a_i = a$ and $\sum b_i = b$. We call these tuples B *ordered bipartitions* of (a, b) and denote the associated Levi subgroup L_B, and the corresponding representation by π_B.

The ordered bipartitions of (a, b) *almost* parametrize the cohomological representation of $U(a, b)$, but there is redundancy. Specifically, $\pi_B \simeq \pi_{B'}$ if B' has adjacent pairs of the form $(a_1, 0)$, $(a_2, 0)$ (resp. $(0, b_1)(0, b_2)$) which are collapsed into $(a_1 + a_2, 0)$ (resp. $(0, b_1 + b_2)$) in B. We will say that an ordered bipartition is *reduced* if all pairs in which one entry is zero are maximally broken up.

Example 59. The following ordered bipartition is not reduced:

$$((3, 1), (2, 0), (1, 0), (0, 3)).$$

It is associated to the same cohomological representation as the following reduced ordered bipartition:

$$((3, 1)(1, 0)(1, 0)(1, 0)(0, 1)(0, 1)(0, 1)).$$

The cohomology of π_B can be expressed in terms of B.

Proposition 60 [49], Proposition 3.2]. *Let* $\mathfrak{q} = \mathfrak{l} \oplus \mathfrak{u}$ *be a* θ-*stable parabolic subalgebra and let* $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{s}$ *be the Cartan decomposition. Let* $R = \dim \mathfrak{u} \cap \mathfrak{s}$. *Then*

$$H^i(\mathfrak{g}, K, A_{\mathfrak{q}}) \simeq \operatorname{Hom}_{\mathfrak{l}\cap\mathfrak{k}}(\wedge^{i-R}\mathfrak{s}, \mathbb{C}).$$

In particular, the smallest nonvanishing degree of cohomology of $A_{\mathfrak{q}}$ is R, which, writing $A_{\mathfrak{q}} = \pi_B$ and referring once more to [6] and [20], is equal to

$$R = \frac{\dim(\mathfrak{s}) - \dim(\mathfrak{s} \cap \mathfrak{l})}{2} = ab - \sum_{i=1}^{r} a_i b_i. \tag{34}$$

In particular, if $a_i b_i = 0$ for all pairs, i.e., if L is compact, then π_B is a discrete series representation and only has cohomology in degree ab.

5.2. *Arthur parameters of cohomological representations.* We turn our attention to archimedean parameters ψ whose associated Arthur packets contain cohomological representations. These are obtained via embedding of L-groups from parameters associated to the trivial representation of Levi subgroups of $G = U(a, b)$. The packets associated to these parameters were constructed by Adams and Johnson [1] in conversation with work of Arthur [4], in a language predating the current formulation of the endoscopic classification of representations. Arancibia, Moeglin and Renard [3] have shown that Adams and Johnson's construction yields the same packets as those appearing in the endoscopic classification in [38] and [26].

To begin, note that there is a natural way to associate to an ordered bipartition B of (a, b) an ordered partition P_B of N, namely by letting

$$P_B = (N_1, \ldots, N_r), \quad N_i = a_i + b_i.$$

Let B be an ordered bipartition, and L_B be the associated Levi subgroup. Then $\hat{L} \simeq \prod_i \operatorname{GL}_{N_i}(\mathbb{C}) \hookrightarrow \hat{G}$, is determined by P_B. The description of $^L L$, i.e., of the Galois action on $\hat{L}$, is given in Section 2.2.5. Cohomological Arthur parameters depend on an embedding $\xi_{\hat{L},\hat{G}} : {}^L L \hookrightarrow {}^L G$ extending the map $\hat{L} \hookrightarrow \hat{G}$. To define $\xi_{\hat{L},\hat{G}}$, it suffices to give the image of $W_{\mathbb{R}}$ inside of $^L G$. Recall that $W_{\mathbb{R}}$ is an extension of $\mathbb{C}^\times$ by a group of order 2, which we write as $\mathbb{C}^\times \sqcup \sigma \mathbb{C}^\times$ with $\sigma^2 = -1$. We give Arthur's construction from Section 5 of [4]. The construction of $A_{\mathfrak{q}}$ in [49] depends on an element α of the Lie algebra $\mathfrak{t}$ of a compact torus. Let T be the torus with Lie algebra $\mathfrak{t}$ and let $\psi_{\hat{L},\hat{G}} : W_{\mathbb{R}} \to {}^L G$ be the map sending $\mathbb{C}^\times$ into $\hat{T}$ so that for any $\lambda^\vee \in X_*(T)$, we have

$$\lambda^\vee(\psi_{\hat{L},\hat{G}}(z)) = z^{\langle \rho_Q, \lambda^\vee \rangle} \bar{z}^{-\langle \rho_Q, \lambda^\vee \rangle}$$

where $\rho_Q = \rho_{\hat{G}} - \rho_{\hat{L}}$, the difference of half-sums of positive roots. Let the element $(1 \rtimes \sigma)$ map to $n_Q \rtimes \sigma$, where for any group G, n_G is an element in the derived group of $\hat{G}$ such that $\operatorname{ad} n_G$ interchanges the positive and negative roots of $(\hat{G}, \hat{T})$, and with $n_Q = n_L^{-1} n_G$. Putting this together and denoting the embedding of $\hat{L}$ into $\hat{G}$ by ι, define $\xi_{\hat{L},\hat{G}}(g, w) = \iota(g)\psi_{\hat{L},\hat{G}}(w)$.

Now let $\psi_{0,\hat{L}} : \mathrm{SL}_2(\mathbb{C}) \times W_{\mathbb{R}} \to {}^L L$ be the Arthur parameter of the packet containing the trivial representation of L. It is trivial on $W_{\mathbb{R}}$ and sends SL_2 to the principal SL_2 of $\hat{L}$. Then the Arthur parameter of G corresponding to the Levi subgroup $\hat{L}$ is the composition

$$\psi_{\hat{L}} := \xi_{\hat{L},\hat{G}} \circ \psi_{0,\hat{L}} : \mathrm{SL}_2 \times W_{\mathbb{R}} \to {}^L G.$$

Adams and Johnson [1] and more recently Nair and Prasad [39] have given a description of the packets attached to the parameters $\psi_{\hat{L}}$.

Proposition 61 [1, Section 3.3]. *Let $\hat{L}$ be a Levi subgroup of $\hat{G}$, dual to a Levi $L(\mathfrak{q})$ attached to a θ-stable parabolic subalgebra $\mathfrak{q}$. The parameter $\psi_{\hat{L}} = \xi_{\hat{G},\hat{L}} \circ \psi_{0,\hat{L}}$ corresponds to a packet Π_ψ consisting of the representations $A_\mathfrak{q}$ such that $\hat{L}(\mathfrak{q}) = \hat{L}$.*

We now translate the descriptions of the packets $\Pi_{\psi_{\hat{L}}}$ given in [1] and [4] into our parametrization by ordered bipartitions.

Proposition 62. *Let $P = (N_1, \ldots, N_r)$ be an ordered partition of N and $\psi_P := \psi_{\hat{L}_P}$ be the corresponding parameter. Then the packet $\Pi_P := \Pi_{\psi_P}$ consists precisely of the cohomological representations π_B associated to bipartitions B such that $P_B = P$.*

Proof. We explained above how L_B gives rise to $\psi_{\hat{L}_B} = \psi_{\hat{L}_{P_B}}$; the parameters $\psi_{\hat{L}_B}$ and $\psi_{\hat{L}_{B'}}$ are equivalent if they are $\hat{G}$-conjugate. The isomorphism classes of representations π_B correspond to Levi subgroups L_B containing the fixed torus T, so we need only consider conjugation by $N_{\hat{G}}(\hat{T})$. This action induces an action of the Weyl group $W(\hat{T}, \hat{G})$ on $\hat{T}$ and on the root datum $(X_*(T), \Delta(T), X_*(\hat{T}), \Delta(\hat{T}))$. Note that the action of conjugation by $\hat{T}$ on cohomological Arthur parameters will only modify $\psi_{\hat{L}}$ by scaling the entries of n_Q. This has no impact on the parameter since n_Q was only specified up to scalars in the construction of $\psi_{\hat{L}}$.

Thus to determine which Levi subgroups $L_{B'}$ give rise to the conjugacy class of $\hat{L}_P$, we consider the action of $W(\hat{G}, \hat{T})$ (denoted $W(\mathfrak{g}, \mathfrak{t})$ in [1]) on the set of ordered bipartitions. Following the description of L_B given in [6], ordered bipartitions are determined ultimately by an element $\alpha \in \mathfrak{t}$. The entries of conjugate elements $w \cdot \alpha$ will have the same values, but these values will be distributed differently among the two pieces of $\mathfrak{t}$ belonging to $U(a)$ and $U(b)$. We denote the values appearing in the entries of α by z_i. The data being preserved by conjugation is the number of entries $a_i + b_i$ which are associated to the same value z_i, as well as the ordering of the z_i. Transitivity of the Weyl group action then ensures that all the possible B such that $P_B = P$ give rise to ψ_P. $\qquad \square$

5.3. *Limit multiplicity for cohomological representations.* We now give results on growth of cohomology. We return our usual notation, in which F is global, $\mathfrak{p}$ is a prime of F, and the subscript "∞" denotes the collection of all the archimedean places. Fix the set S_0 as in Section 4.1 so that it contains all but one archimedean place v_0. Let G be the inner form of $U_{E/F}(N)$ such that $G_{v_0} \simeq U(a, b)$ and all the other factors at infinity are compact. Define the groups $K(\mathfrak{p}^n)$ and $\Gamma(\mathfrak{p}^n)$ as in Section 4.1. By Matsushima's

formula and the inequality (29), we have

$$h^i(\mathfrak{p}^n) := \dim(H^i(\Gamma(\mathfrak{p}^n), \mathbb{C})) \le \sum_{\pi = \mathbf{1}^{|S_0|} \otimes \pi_{v_0} \otimes \pi_f} m(\pi) h^i(\mathfrak{g}_{v_0}, K_{v_0}; \pi_{v_0}) \dim \pi_f^{K_f(\mathfrak{p}^n)}.$$

We can now give our theorem for growth of cohomology.

Theorem 63. *Let ψ_∞ be the cohomological parameter of G_∞ associated to a reordering of $(2k, 1, \ldots, 1)$. Let*

$$h^i_{\psi_\infty}(\mathfrak{p}^n) = \sum_{\psi \in \Psi(\psi_\infty)} \sum_{\pi \in \Pi_\psi} m(\pi) h^i(\mathfrak{g}_{v_0}, K_{v_0}; \pi_{v_0}) \dim \pi_f^{K_f(\mathfrak{p}^n)}.$$

Then

$$h^i_{\psi_\infty}(\mathfrak{p}^n) \ll \mathrm{Nm}(\mathfrak{p}^n)^{N(N-2k)}.$$

Proof. The possible contribution to cohomology of a given representation π_{v_0} is bounded, so we need only bound the contribution to $m(\pi_\infty, \mathfrak{p}^n)$ coming from packets attached to parameters specializing to ψ_∞, for each $\pi_\infty \simeq \pi_{v_0} \otimes \mathbf{1}^{[F:\mathbb{Q}]-1} \in \Pi^0_{\psi_\infty}$. The result then follows from Theorem 56, provided that cohomological parameters satisfy its assumptions. From Theorem 58 and the following comment, cohomological representations have the (regular) infinitesimal character of the trivial representation, and trivial central character. From the discussion in Section 5.2, reorderings of $(2k, 1, \ldots, 1)$ correspond to parameters for which $\psi(\mathrm{SL}_2) = \nu(2k) \oplus \nu(1)^{N-2k}$. Thus the assumptions are satisfied and the result holds. $\square$

Note that the theorem does not in fact bound $m(\pi_\infty, \mathfrak{p}^n)$ for a general $\pi_\infty \in \Pi_{\psi_\infty}$. Indeed, since Arthur packets are not disjoint, the representation π_∞ could also appear in a different Arthur packet whose growth we do not bound. More specifically, if $\pi_\infty = \pi_B \otimes \mathbf{1}^{[F:\mathbb{Q}]-1} \in \Pi^0_{\psi_\infty}$ where B is an ordered bipartition described in Section 5.1.1, it could be the case that B is the reduction of an ordered bipartition B', for example if we had

$$B = ((1, 1), (1, 0), (1, 0), (0, 1)), \quad B' = ((1, 1), (2, 0), (0, 1)).$$

On the other hand, if B is not the reduction of another ordered bipartition, then $\pi_\infty = \pi_B$ does not appear in any other archimedean Arthur packet and the theorem produces upper bounds for $m(\pi_\infty, \mathfrak{p}^n)$. We record this discussion below.

Corollary 64. *Let $B = ((a_1, b_1), \ldots, (a_r, b_r))$ be a reduced ordered bipartition such that:*

(i) *The associated partition P_B is a reordering of $(2k, 1, \ldots, 1)$.*

(ii) *We have $(a_i, b_i) \ne (a_{i+1}, b_{i+1})$ for all i.*

Then

$$m(\pi_B, \mathfrak{p}^n) \ll \mathrm{Nm}(\mathfrak{p}^n)^{N(N-2k)}.$$

Example 65. Assume that $a < b$ in the signature of $U(a, b)$. A family of partitions satisfying the conditions of Corollary 64 are the suitable reorderings of

$$B_j = \begin{cases} ((0, 1), \ldots, (1, 0), (0, 1), (a - j, b - j - 2), (0, 1), (1, 0), \ldots, (0, 1)) & N \text{ even}, \\ ((a - j, b - j - 1), (0, 1), (1, 0), \ldots, (0, 1), (1, 0), (0, 1)) & N \text{ odd}, \end{cases}$$

where $1 \le j \le a - 1$ if N is even (resp. $0 \le j \le a - 1$ is N is odd.) The computations of Section 5.1.1 show that their lowest degree of cohomology is

$$i = i(N, a, j) = \begin{cases} j(N - j - 2) + 2a & N \text{ even}, \\ j(N - j - 1) + a & N \text{ odd}. \end{cases}$$

Note that $j = \frac{1}{2}(N - 2k - 2)$ for N even (resp. $j = \frac{1}{2}(N - 2k - 1)$ for N odd) which gives the family alluded to in the introduction.

Additionally, the smallest $i > 0$ for which $h^i(\mathfrak{p}^n) \ne 0$ is $i = a$. When N is odd, one can check that representations as above with $j = 0$ are the only source of cohomology in degree a. In this situation, we get bounds on Betti numbers.

Corollary 66. *Keeping the assumptions of Theorem 63, assume additionally that N is odd and that $a < b$. Then*

$$h^a(\mathfrak{p}^n) \ll \mathrm{Nm}(\mathfrak{p}^n)^N.$$

Proof. This follows from Theorem 63 with $2k = N - 1$ provided that

$$h^a(\mathfrak{p}^n) = \sum_{\psi_\infty} h^a_{\psi_\infty}(\mathfrak{p}^n),$$

where the sum is taken over finitely many parameters ψ_∞ associated to a reordering of $(1, N - 1)$. Since there are finitely many Arthur of $U(a, b)$ with cohomological representations, this amounts to showing that representations with cohomology in degree a belong only to packets Π_{ψ_∞} associated to these partitions. Going back to Proposition 60 and the following discussion, in particular to (34), we find that the only representations with cohomology in degree a are of the form π_B for B a reordering of

$$((0, 1), (a, b - 1)).$$

These cannot be reduced, nor are they the reduction of other ordered bipartitions so by Proposition 62, they only belong to packets associated to parameters corresponding ordered partitions are reorderings of $(1, a + b - 1) = (1, N - 1)$, which was exactly our requirement. □

5.4. Comparison with the Sarnak–Xue conjecture. Finally, we compare our results with the conjecture of Sarnak and Xue [42] relating multiplicity growth to decay of matrix coefficients. For an irreducible unitary representation π_∞ of a Lie group G, Sarnak and Xue define

$$p(\pi_\infty) = \inf\{p \ge 2 \mid K\text{-finite matrix coefficients of } \pi_\infty \text{ are in } L^p(G)\}.$$

They then conjecture the following bounds for unitary π_∞:

$$m(\pi_\infty, \mathfrak{p}^n) \ll_\epsilon Vol(X(\mathfrak{p}^n))^{(2/p(\pi_\infty))+\epsilon}.$$

We will now show that the Sarnak and Xue conjecture holds for the representations for which we have proved upper bounds on multiplicity growth.

Proposition 67. *Let $\pi_\infty = \pi_B$ be as in Corollary 64. Then*

$$\frac{2}{p(\pi_B)} \geq \frac{N - 2k}{N - 1}.$$

Since $m(\pi_B, \mathfrak{p}^n) \ll \mathrm{Nm}(\mathfrak{p}^n)^{N(N-2k)}$ and the volume of $X(\mathfrak{p}^n)$ grows like $\mathrm{Nm}(\mathfrak{p}^n)^{N^2-1}$ we obtain the following.

Corollary 68. *For π_B as in Corollary 64, we have*

$$m(\pi_B, \mathfrak{p}^n) \ll \mathrm{Vol}(X(\mathfrak{p}^n))^{N(N-2k)/(N^2-1)} \ll \mathrm{Vol}(X(\mathfrak{p}^n))^{2/p(\pi_B)}$$

and the Sarnak–Xue conjecture holds.

The remainder of the section sets up and gives a proof of Proposition 67.

5.4.1. *Computation of the rate of decay.* For cohomological representations, we will give bounds on $p(\pi_B)$ from the descriptions of π_B as Langlands quotients given in [49]. For this section we follow the notation of Knapp [28, Sections 7 and 8]. We start by bounding $p(\pi)$ for π an arbitrary Langlands quotient in terms of the inducing data.

We recall the setup for the definition of Langlands quotients. First, fix an Iwasawa decomposition

$$G = K A_0 N_0, \quad g = k_g a_g n_g \tag{35}$$

of G. Here K a maximal compact subgroup, A_0 a maximal split torus, N_0 unipotent. By letting $M_0 = Z_K(A_0)$, this gives rise to a minimal parabolic subgroup $S_0 = M_0 A_0 N_0$. Let $S = MAN$ be parabolic and regular with respect to S_0, which is to say that $A \subset A_0$ is split, M is the no longer necessarily compact Levi component, and $N \subset N_0$ unipotent. Denote by $\mathfrak{a}$ (resp. $\mathfrak{a}_0$) the Lie algebra of A (resp. A_0), and by $\mathfrak{a}_M$ the Lie algebra of the maximal split torus of M, so that $\mathfrak{a}_0 = \mathfrak{a} \oplus \mathfrak{a}_M$. Let ρ_0 be the half-sum of the positive roots of $\mathfrak{a}_0$ in $\mathfrak{g}$. Let σ be a discrete series representation of M and $\nu \in \mathfrak{a}^*$ be real-valued and in the open positive Weyl chamber. Denote by ν_0 the extension of ν to $\mathfrak{a}_0$ by setting it to be zero on $\mathfrak{a}_M$, and let $\rho_0 \in \mathfrak{a}_0^*$ be the half-sum of the positive roots of $\mathfrak{a}_0$ in $\mathfrak{g}$. Denote by $U(S, \sigma, \nu)$ the corresponding parabolically induced representation, and by $J(S, \sigma, \nu)$ its Langlands quotient.

The proof of the upcoming proposition depends on a collection of results from [28]; before stating it we recall the setup and some nomenclature. When studying the decay of matrix coefficients, one introduces a class of so-called *spherical functions* φ_ν^G associated to $\nu \in \mathfrak{a}_0^*$, and defined by

$$\varphi_\nu^G(g) = \int_K e^{-(\nu + \rho_0)H(g^{-1}k)} \, dk$$

where $H(g) = \log(a_g) \in \mathfrak{a}_0$ is defined using the Iwasawa decomposition (35). Paraphrasing Knapp, these are "useful yardsticks" to measure the decay of matrix coefficients. For example, it is known that the K-finite matrix coefficients of discrete series are dominated by φ_0^G. As their name indicates, the spherical functions are K-invariant. A generalization of this notion is that of a τ-*spherical function* associated to a pair $\tau = (\tau_1, \tau_2)$ of representations of K: a τ-spherical function is valued in the space $U_1 \otimes U_2^\vee$ and is left τ_1- and right τ_2-equivariant.

In [28, Section VII.8], Knapp studies representations π by producing and studying an asymptotic expansion of the τ-spherical functions associated to the K-types of π. The functions $F_{\lambda - \rho_0}$ associated to $\lambda \in \mathfrak{a}_0^*$ appearing in this asymptotic expansion control rates of decay of the τ-spherical function in various directions along A_0. The $\lambda - \rho_0$ such that $F_{\lambda - \rho_0}$ contributes nontrivially to the decomposition of the τ-spherical functions of π are called exponents of π. A *leading exponent* of π is an exponent $\mu - \rho_0$ of π, maximal in the sense that for any comparable exponent $\lambda - \rho_0$, the difference $\mu - \lambda$ is a linear combination of simple roots with nonnegative integer coefficients.

Proposition 69. *Let $\omega_1, \ldots, \omega_{\dim \mathfrak{a}_0}$ denote the basis of $\mathfrak{a}_0$ dual to the basis of $\mathfrak{a}_0^*$ consisting of the simple roots. Then*

$$p(J(S, \sigma, \nu)) \leq \inf\{p \geq 2 \mid p\langle \nu_0 - \rho_0, \omega_j \rangle < -2\langle \rho_0, \omega_j \rangle \text{ for all } \omega_j\}. \tag{36}$$

Proof. The lemma follows from the combination of results in [28]. From [28, 8.48], the K-finite matrix coefficients of $J(S, \sigma, \nu)$ belong to $L^p(G)$ if and only if for all ω_i and all leading exponents $\mu - \rho_0$ of $J(S, \sigma, \nu)$, the following inequality is satisfied:

$$p\langle \operatorname{Re}\mu - \rho_0, \omega_j \rangle < -2\langle \rho_0, \omega_j \rangle. \tag{37}$$

In [28, 8.47], we see that all leading exponents μ satisfy $\langle \operatorname{Re}\mu, \omega_j \rangle \leq \langle \nu_0, \omega_j \rangle$ provided there exists an integer $q \geq 0$ such that the K-finite matrix coefficients of $J(S, \sigma, \nu)$ are bounded above on $\overline{A_0^+}$ by a multiple of $e^{(\nu_0 - \rho)(\log a)}(1 + \operatorname{Nm} a)^q$. This upper bound is established for $U(S, \sigma, \nu)$ and a fortiori for $J(S, \sigma, \nu)$ by propositions 7.14 and 7.15 of [28], together with the fact that as a discrete series, the K-finite matrix coefficients of σ are dominated by a multiple of the spherical function φ_0^M. $\square$

We now give an explicit bounds for $p(\pi_B)$ in terms of B, for a class of representations including those of Proposition 67.

Proposition 70. *Let $G = U(a, b)$ with $a + b = N$, and let $B = ((a_1, b_1), \ldots, (a_r, b_r))$ be a reduced ordered bipartition of (a, b). Assume that there is a single index k such that $\min\{a_k, b_k\} \neq 0$, and let $N_k = a_k + b_k$. Then*

$$\frac{2}{p(\pi_B)} \geq \frac{N - N_k}{N - 1}.$$

Proof. In light of Proposition 69, we will realize π_B as a Langlands quotient $J(S, \sigma, \nu)$, and show that for the corresponding ν_0 we have

$$\inf\{p \geq 2 \mid p\langle \nu_0 - \rho_0, \omega_j \rangle < -2\langle \rho_0, \omega_j \rangle \text{ for all } \omega_j\} = \frac{2(N - 1)}{N - N_k}.$$

To simplify the computations, note that it is equivalent to show that

$$\max_{\omega_j}\left\{\frac{\langle \nu_0, \omega_j\rangle}{\langle \rho_0, \omega_j\rangle}\right\} = \frac{N_k - 1}{N - 1}. \tag{38}$$

We recall the descriptions of cohomological representations as Langlands quotients is given in [49, Section 6]. Let $L_B = \prod_i U(a_i, b_i)$ be the Levi subgroup attached to the representation π_B and fix an Iwasawa decomposition $L = (K \cap L)AN$. By assumption, A has rank $c_k = \min\{a_k, b_k\}$; denote its Lie algebra by $\mathfrak{a}$, and let ν be the half-sum of the roots of $\mathfrak{a}$ in the Lie algebra $\mathfrak{n}$ of N. Let M be the centralizer of A in G, and fix S a choice of parabolic subgroup with Levi MA, such that ν is in the open positive Weyl chamber. Then by [49, Theorem 6.16], there is a discrete series representation σ of M such that $\pi_B \simeq J(S, \sigma, \nu)$.

To conclude, we put ourselves back in the framework of Proposition 69. Let $S_0 = A_0 M_0 N_0 \subset S$ be a minimal parabolic subgroup. Then if $\mathfrak{a}_M$ is the Lie algebra of a maximal split torus in M, we have $\mathfrak{a}_0 = \mathfrak{a} \oplus \mathfrak{a}_M$ and $\dim \mathfrak{a}_0 = c := \min\{a, b\}$. Let $\alpha_1, \ldots, \alpha_c$ be the simple roots of $\mathfrak{a}_0$ in $\mathfrak{g}$. Recall that ν_0 is obtained by extending ν by 0 to $\mathfrak{a}_M$. Thus we can write $\nu_0 = \sum_{j=1}^{c_k} j(N_k - j)\alpha_j$ and $\rho_0 = \sum_{j=1}^{c} j(N - j)\alpha_j$. Since the ω_i are by construction the dual basis to the α_i, we have

$$\frac{\langle \nu_0, \omega_j\rangle}{\langle \rho_0, \omega_j\rangle} = \begin{cases} (N_k - j)/(N - j), & j \le c_k, \\ 0, & j > c_k. \end{cases}$$

The maximum is achieved when $j = 1$. $\qquad\square$

Acknowledgements

References

[1] J. Adams and J. F. Johnson, "Endoscopic groups and packets of nontempered representations", *Compositio Math.* **64**:3 (1987), 271–309. MR

[2] J. D. Adler and D. Prasad, "On certain multiplicity one theorems", *Israel J. Math.* **153** (2006), 221–245. MR Zbl

[3] N. Arancibia, C. Mœglin, and D. Renard, "Paquets d'Arthur des groupes classiques et unitaires", *Ann. Fac. Sci. Toulouse Math.* (6) **27**:5 (2018), 1023–1105. MR Zbl

[4] J. Arthur, "Unipotent automorphic representations: conjectures", pp. 13–71 in *Orbites unipotentes et représentations, II: Groupes p-adiques et réels*, Astérisque **171–172**, Soc. Math. de France, 1989. MR Zbl

[5] J. Arthur, *The endoscopic classification of representations: orthogonal and symplectic groups*, American Mathematical Society Colloquium Publications **61**, Amer. Math. Soc., Providence, RI, 2013. MR Zbl

[6] N. Bergeron and L. Clozel, "Spectre automorphe des variétés hyperboliques et applications topologiques", 303 (2005), xx+218. MR Zbl

[7] N. Bergeron and L. Clozel, "Sur la cohomologie des variétés hyperboliques de dimension 7 trialitaires", *Israel J. Math.* **222**:1 (2017), 333–400. MR Zbl

[8] I. N. Bernshtein, "All reductive p-adic groups are tame", *Functional Analysis and Its Applications* **8**:2 (1974), 91–93. Zbl

[9] A. Borel, "Automorphic *L*-functions", pp. 27–61 in *Automorphic forms, representations and L-functions* (Corvallis, OR, 1977), vol. 2, edited by A. Borel and W. Casselman, Proc. Sympos. Pure Math. **33**, Amer. Math. Soc., Providence, RI, 1979. MR Zbl

[10] F. Calegari and M. Emerton, "Bounds for multiplicities of unitary representations of cohomological type in spaces of cusp forms", *Ann. of Math.* (2) **170**:3 (2009), 1437–1446. MR Zbl

[11] G. Chenevier and O. Taïbi, "Discrete series multiplicities for classical groups over **Z** and level 1 algebraic cusp forms", *Publ. Math. Inst. Hautes Études Sci.* **131** (2020), 261–323. MR Zbl

[12] M. Cossutta and S. Marshall, "Theta lifting and cohomology growth in *p*-adic towers", *Int. Math. Res. Not.* **2013**:11 (2013), 2601–2623. MR Zbl

[13] R. Dalal, "Sato–Tate equidistribution for families of automorphic representations through the stable trace formula", *Algebra Number Theory* **16**:1 (2022), 59–137. MR Zbl

[14] P. Deligne, "Travaux de Shimura", exposé no. 389, 123–165 in *Séminaire Bourbaki*, 1970/1971, Lecture Notes in Mathematics **244**, Springer, Berlin, 1971. MR Zbl

[15] G. van Dijk, "Computation of certain induced characters of p-adic groups", *Math. Ann.* **199** (1972), 229–240. MR Zbl

[16] M. Dimitrov and D. Ramakrishnan, "Arithmetic quotients of the complex ball and a conjecture of Lang", *Doc. Math.* **20** (2015), 1185–1205. MR Zbl

[17] A. Ferrari, "Théorème de l'indice et formule des traces", *Manuscripta Math.* **124**:3 (2007), 363–390. MR Zbl

[18] W. T. Gan, B. H. Gross, and D. Prasad, "Symplectic local root numbers, central critical *L* values, and restriction problems in the representation theory of classical groups", pp. 1–109 in *Sur les conjectures de Gross et Prasad*, vol. 1, Astérisque **346**, Soc. Math. de France, 2012. MR Zbl arXiv 0909.2999

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

[20] M. Gerbelli-Gauthier, *Growth of cohomology of arithmetic groups and the stable trace formula*, Ph.D. thesis, University of Chicago, 2020, available at https://www.proquest.com/docview/2426240362/.

[21] P. E. Gunnells, M. McConnell, and D. Yasaki, "On the cohomology of congruence subgroups of GL$_3$ over the Eisenstein integers", *Exp. Math.* **30**:4 (2021), 499–512. MR Zbl

[22] Harish-Chandra, "Representations of semisimple Lie groups, III", *Trans. Amer. Math. Soc.* **76** (1954), 234–253. MR

[23] M. Harris and R. Taylor, *The geometry and cohomology of some simple Shimura varieties*, Ann. of Math. Stud. **151**, Princeton Univ. Press, 2001. MR Zbl

[24] G. Henniart, "Une preuve simple des conjectures de Langlands pour GL(n) sur un corps *p*-adique", *Invent. Math.* **139**:2 (2000), 439–455. MR Zbl

[25] H. Jacquet and J. A. Shalika, "On Euler products and the classification of automorphic forms, II", *Amer. J. Math.* **103**:4 (1981), 777–815. MR Zbl

[26] T. Kaletha, A. Minguez, S. W. Shin, and P.-J. White, "Endoscopic classification of representations: Inner forms of unitary groups", preprint, 2014. arXiv 1409.3731

[27] T. Kaletha, A. Minguez, and S. W. Shin, "Endoscopic classification of representations: inner forms of unitary groups", in preparation.

[28] A. W. Knapp, *Representation theory of semisimple groups: an overview based on examples*, Princeton Mathematical Series **36**, Princeton University Press, 1986. MR Zbl

[29] R. E. Kottwitz, "Sign changes in harmonic analysis on reductive groups", *Trans. Amer. Math. Soc.* **278**:1 (1983), 289–297. MR Zbl

[30] R. E. Kottwitz and D. Shelstad, *Foundations of twisted endoscopy*, Astérisque **255**, Soc. Math. de France, 1999. MR Zbl

[31] W. Landherr, "Äquivalenz Hermitescher Formen über einem beliebigen algebraischen Zahlkörper", *Abh. Math. Sem. Univ. Hamburg* **11**:1 (1935), 245–248. MR

[32] R. P. Langlands and D. Shelstad, "On the definition of transfer factors", *Math. Ann.* **278**:1-4 (1987), 219–271. MR

[33] G. Laumon and B. C. Ngô, "Le lemme fondamental pour les groupes unitaires", *Ann. of Math.* (2) **168**:2 (2008), 477–573. MR Zbl

[34] S. Marshall, "Endoscopy and cohomology growth on $U(3)$", *Compos. Math.* **150**:6 (2014), 903–910. MR Zbl

[35] S. Marshall and S. W. Shin, "Endoscopy and cohomology in a tower of congruence manifolds for $U(n, 1)$", *Forum Math. Sigma* **7** (2019), Art. Id. e19. MR Zbl

[36] Y. Matsushima, "A formula for the Betti numbers of compact locally symmetric Riemannian manifolds", *J. Differential Geometry* **1** (1967), 99–109. MR Zbl

[37] C. Mœglin and J.-L. Waldspurger, "Le spectre résiduel de GL(n)", *Ann. Sci. École Norm. Sup.* (4) **22**:4 (1989), 605–674. MR

[38] C. P. Mok, "Endoscopic classification of representations of quasi-split unitary groups", 1108 (2015), 1–248. MR Zbl

[39] A. N. Nair and D. Prasad, "Cohomological representations for real reductive groups", *J. Lond. Math. Soc.* (2) **104**:4 (2021), 1515–1571. MR Zbl

[40] B. C. Ngô, "Le lemme fondamental pour les algèbres de Lie", *Publ. Math. Inst. Hautes Études Sci.* 111 (2010), 1–169. MR

[41] V. Platonov and A. Rapinchuk, *Algebraic groups and number theory*, Pure and Applied Mathematics **139**, Academic Press, 1994. MR Zbl

[42] P. Sarnak and X. X. Xue, "Bounds for multiplicities of automorphic representations", *Duke Math. J.* **64**:1 (1991), 207–227. MR Zbl

[43] G. Savin, "Limit multiplicities of cusp forms", *Invent. Math.* **95**:1 (1989), 149–159. MR Zbl

[44] P. Scholze, "The local Langlands correspondence for GL$_n$ over p-adic fields", *Invent. Math.* **192**:3 (2013), 663–715. MR Zbl

[45] S. W. Shin, "Galois representations arising from some compact Shimura varieties", *Ann. of Math.* (2) **173**:3 (2011), 1645–1741. MR Zbl

[46] O. Taïbi, "Arthur's multiplicity formula for certain inner forms of special orthogonal and symplectic groups", *J. Eur. Math. Soc. (JEMS)* **21**:3 (2019), 839–871. MR Zbl

[47] J. Tate, "Number theoretic background", pp. 3–26 in *Automorphic forms, representations and L-functions* (Corvallis, OR, 1977), vol. 2, edited by A. Borel and W. Casselman, Proc. Sympos. Pure Math. **33**, Amer. Math. Soc., Providence, RI, 1979. MR Zbl

[48] D. A. Vogan, Jr., "Cohomology and group representations", pp. 219–243 in *Representation theory and automorphic forms* (Edinburgh, 1996), edited by T. N. Bailey and A. W. Knapp, Proceedings of Symposia in Pure Mathematics **61**, Amer. Math. Soc., Providence, RI, 1997. MR Zbl

[49] D. A. Vogan, Jr. and G. J. Zuckerman, "Unitary representations with nonzero cohomology", *Compositio Math.* **53**:1 (1984), 51–90. MR Zbl

[50] J.-L. Waldspurger, "Endoscopie et changement de caractéristique", *J. Inst. Math. Jussieu* **5**:3 (2006), 423–525. MR Zbl

[51] J.-L. Waldspurger, "L'endoscopie tordue n'est pas si tordue", pp. 1–261 Mem. Amer. Math. Soc. **908**, Amer. Math. Soc., Providence, RI, 2008. MR Zbl

[52] N. R. Wallach, "On the constant term of a square integrable automorphic form", pp. 227–237 in *Operator algebras and group representations* (Neptun, Romania, 1980), vol. 2, edited by G. Arsene, Monogr. Stud. Math. **18**, Pitman, Boston, 1984. MR Zbl

mathilde.gerbelli-gauthier@mcgill.ca   *Department of Mathematics and Statistics, McGill University, Montreal, Canada*

Algebra & Number Theory

msp.org/ant

See inside back cover or msp.org/ant for submission instructions.

Algebra & Number Theory