STATISTICS OF COHOMOLOGICAL AUTOMORPHIC REPRESENTATIONS ON UNITARY GROUPS VIA THE ENDOSCOPIC CLASSIFICATION

RAHUL DALAL[1] AND MATHILDE GERBELLI-GAUTHIER[2]

ABSTRACT. Consider the family of automorphic representations on some unitary group with fixed (possibly non-tempered) cohomological representation π_0 at infinity and level dividing some finite upper bound. We compute statistics of this family as the level restriction goes to infinity. For unramified unitary groups and a large class of π_0, we are able to compute the exact leading term for both counts of representations and averages of Satake parameters. We get bounds on our error term similar to previous work by Shin-Templier that studied the case of discrete series at infinity.

This provides many corollaries: for example, we get new exact asymptotics on the growth of certain degrees of cohomology in certain towers of locally symmetric spaces, prove an averaged Sato-Tate equidistribution law for spectral families with specific non-tempered cohomological components at infinity, and extend bounds of Marshall and Shin to prove Sarnak-Xue density for cohomological representations at infinity on all unitary groups that don't have a $U(2,2)$ factor over infinity.

The main technical tool is an extension of an inductive argument that was originally developed by Taïbi to count unramified representations on Sp and SO and used the endoscopic classification of representations (which our case requires for non-quasisplit unitary groups).

CONTENTS

[1]DEPARTMENT OF MATHEMATICS, JOHNS HOPKINS UNIVERSITY, BALTIMORE, MD
[2]DEPARTMENT OF MATHEMATICS AND STATISTICS, MCGILL UNIVERSITY, MONTREAL, QC
E-mail addresses: `dalal@jhu.edu, mathilde.gerbelli-gauthier@mcgill.ca`.

Date: February 26, 2023.
2000 *Mathematics Subject Classification.* 11F55, 11F70, 11F72, 11F75, 22E50.
The first author was supported by an National Science Foundation postdoctoral research fellowship. The second author was supported by the Charles Simonyi endowment at the Institute for Advanced Study.

1. Introduction

1.1. **Context.**

1.1.1. *Statistics in general.* Let G be a reductive group over a number field F. Automorphic representations for G are very roughly irreducible subrepresentations of $L^2(G(F)\backslash G(\mathbb{A}))$ under right multiplication by $G(\mathbb{A})$. While this definition may seem unmotivated, automorphic representations encode information important to many applications—for example Galois representations through the Langlands program, so-called expanders used for computer algorithms through constructions akin to that of Lubotzky-Phillips-Sarnak, and the geometry locally symmetric spaces through automorphic decompositions of their cohomology.

Every automorphic representation π has a tensor product decomposition:

$$\pi = \bigotimes_v{}' \pi_v$$

into unitary irreducible representations π_v of G_v over places v of F. Applications usually depend on a key question: which combinations of π_v actually tensor together into an automorphic representation; i.e. which products are represented as functions in $L^2(G(F)\backslash G(\mathbb{A}))$?

This paper broadly focuses on an easier version of this key question—that of computing *statistics*. We will consider families $\mathcal{F}_i$ of automorphic representations satisfying local conditions on each π_v. Then we try to estimate as well as possible the asymptotics of the count of such representations weighted by some parameter of the representations when the local conditions "go to infinity" in some sense. For example, when $G = \mathrm{GL}_2$, we could look at the "mth moment" of the Hecke eigenvalue at some place p averaged over weight-k, level-N holomorphic modular forms as $N \to \infty$. Problems of statistics are usually related to studying automorphic representations in spectral families in the sense of [67].

1.1.2. *Our specific case.* More specifically, we are interested in the specific case where the condition on the component π_∞ at infinity is that it is restricted to a specific representation—this can be thought of as fixing the "qualitative type" of an automorphic representation; for example, "holomorphic Siegel modular form of weight $\vec{v}$". We organize such questions by an informal ranking of the complexity of an automorphic representation based on the complexity of the component π_∞ [1].

The simplest automorphic representations correspond to the simplest components at infinity: the discrete series that can be realized as explicit subrepresentations of $L^2(G_\infty)$. Statistics of these discrete-at-infinity automorphic representations are

[1]There are of course many powerful automorphic statistics results that don't fit into this story—as a good "most general" representative [32] considers π_∞ contained in any subset of the unitary dual with finite, non-zero volume. We are not giving a full literature review here.

well-understood through Arthur's trace formula. Specifically, [7] developed an reasonably tractable trace formula for studying them. The techniques of first and most importantly [73] and second [31] as used in first author's thesis [26] build on this tractable trace formula to provide good asymptotic estimates and error terms in the general discrete-at-infinity case (as far as the authors are aware, these are the strongest bounds known for the general case).

This paper goes beyond discrete series and studies automorphic representations with cohomological component at infinity—i.e. component at infinity that has some non-trivial relative Lie algebra cohomology. There is a critical new complication here: when a group has discrete series, non-discrete cohomological representations are always non-tempered. In particular, general cohomological automorphic representations can represent violations of the Ramanujan conjecture. They can therefore be very sparse in the automorphic spectrum and difficult to isolate.

Recent results suggest that known cases of the endoscopic classification—e.g. [11], [58], and [44]—are a good way to study general cohomological automorphic representations. Work of Marshall, Shin and the second author have used it to provide good upper bounds for counts on unitary groups, see [50], [52] and [34]. In specific simpler cases, explicit counts have even been computed: [20] and [76] consider the case of level-1 representations on classical groups and [64] develops techniques that apply to the case of low-level automorphic representations on Sp_4.

This work attempts to organize and synthesize the bounds of Marshal, Shin, and the second author together with an inductive analysis used in the work of Taïbi [76] into a proposal for a general method to understand statistics of cohomological automorphic representations. While our proposed method in general depends on some wide-open and difficult problems in local representation theory, we are able to explicitly implement it in some specific cases on unitary groups. This gives an understanding of cohomological asymptotic statistics on unitary groups which is more general than any previous work.

We emphasize in particular that, in many cases, we are able to compute exact leading terms together with estimates on sub-leading terms. This gives us applications towards results like the Sarnak-Xue density conjecture, the growth of cohomology of locally symmetric spaces, and Sato-Tate equidistribution averaged over families of automorphic representations. We motivate some of these in more detail:

1.1.3. *Application: Growth of Cohomology.* A motivating problem for these statistical computations is that of the growth of cohomology in towers of arithmetic groups. Let Γ be an arithmetic lattice in $G_\infty = G(F \otimes_{\mathbb{Q}} \mathbb{R})$, and assume that Γ is both neat and cocompact for simplicity. Then the group cohomology of Γ (or equivalently, de Rham cohomlogy of the locally symmetric space $\Gamma \backslash G_\infty / K_\infty$ for K_∞ a maximal compact subgroup) is computed in terms of automorphic forms via Matsushima's formula:

$$H^*(\Gamma, \mathbb{C}) = \bigoplus_{\pi_\infty \text{ cohomological}} m(\pi_\infty, \Gamma) H^*(\mathfrak{g}, K_\infty; \pi_\infty),$$

where $m(\pi_\infty, \Gamma) = \dim \mathrm{Hom}_{G_\infty}(\pi_\infty, L^2(\Gamma \backslash G_\infty))$. Beyond the fact that locally symmetric space provides a rich class of examples of manifolds, the cohomology $H^*(\Gamma, \mathbb{C})$ is also of interest because it carries an action of Hecke operators, making it a generalization of modular forms of weight > 1. The Hecke eigensystems that arise should correspond to Galois representations via the Langlands program.

Outside of some low-rank examples, dimensions of $H^*(\Gamma, \mathbb{C})$ are only known for specific lattices Γ, see for example [12, 37]. A fruitful approach to studying the cohomology of Γ is doing so *in towers*: one fixes a sequence Γ_n of typically nested normal subgroups, and studies the asymptotics of $\dim H^*(\Gamma_n, \mathbb{C})$ as $n \to \infty$. Without giving a systematic survey of this problem, we note there have been multiple approaches: topological constructions, for example [74], non-abelian Iwasawa theory methods as in [19], and, beginning with the work of DeGeorge-Wallach [27], what can be referred to as spectral approaches.

As shown in [78], cohomological discrete series representations only contribute to cohomology in degree $(1/2)\dim(G_\infty/K_\infty)$ (where we assume G_∞ has anisotropic center for simplicity). Thus the exact asymptotics of [27], and later [21] and [70], show that the middle degree of cohomology grows like the volume of the corresponding symmetric space when G_∞ has discrete series, and that the growth of the lower degrees is slower. This leaves open the question of more precise upper bounds for other degrees of cohomology, and it seems (see [68, §1]) that this motivating question was at the heart of the discussion which led to the formulation of the Sarnak-Xue conjecture discussed in 1.1.5.

Progress on bounding growth of Betti numbers has been made in various directions, from the lower bounds of [24, 25, 71, 74] to the vanishing results of [18, 22] to extension to of the DeGeorge-Wallach results to more general sequences of lattices in [1]. Starting with [68], there has also been progress on upper bounds: [19] have obtained a power saving for lattices in any group G_∞ admitting discrete series. Most influential for us is a series [25, 50, 51] of work of Marshall and his collaborators, culminating in [52], in which Marshall-Shin give upper bounds for all degrees of cohomology for lattices in $U(N-1, 1)$. Among other results in this article, we show that the bounds of Marshall-Shin are sharp in every other degree.

1.1.4. *Application: Averaged Sato-Tate.* Let π be an automorphic representation on a reductive group G. A Sato-Tate result for π is a statement that the Satake parameters of the unramified components π_v are equidistributed over v according to a Sato-Tate distribution μ_π determined by some properties of π. This should be thought of as a generalized, automorphic-side analogue of the classical Sato-Tate conjecture (proved in [14]) for the equidistribution over p of the coefficients a_p associated to point counts over $\mathbb{Z}/p$ of elliptic curves. See [73, §1.1] for a full introduction to the problem (in particular, [72] states a very general Galois-side version of the conjecture and [13], [14] prove the conjecture for restrictions of scalars of GL_1 and GL_2).

Unfortunately, even stating what this Sato-Tate distribution should be for a single π depends on more-or-less the full conjectures of Langlands functoriality. Extremely roughly, π should correspond to another reductive group H_π and L-map $\varphi : {}^L H_\pi \hookrightarrow {}^L G$—the smallest such that the conjectural global L-parameter for π factors through φ. The law μ_π should then be thought of as pushforward of a measure μ_{H_π} from the space of Satake parameters of H_π to that of G. For general π on high-rank groups, Sato-Tate results therefore appear unapproachable. In fact, empirically measuring the Sato-Tate distribution for π is arguably one of the key currently accessible pieces of evidence for the existence of this conjectural H_π in the first place.

Following [73], we therefore instead study Sato-Tate laws averaged over some increasing sequence of families $\mathcal{F}_i$. Heuristically, representations $\pi \in \mathcal{F}$ can have many different H_π. However, for reasonable families, we roughly expect the log of the count of forms associated to $\pi \in \mathcal{F}_i$ with $H_\pi = H_0$ to be proportional to the dimension of H_0. Therefore, we should expect most $\pi \in \mathcal{F}_i$ to have H_π be some maximum value $H_\mathcal{F}^{\max}$. In particular, if we look at Satake parameters π_v over all places v *and* all $\pi \in \mathcal{F}_i$, we should expect the distribution for $H_\pi = H_\mathcal{F}^{\max}$ to dominate.

Such "averaged Sato-Tate laws" end up being far easier to establish. For example, [73] studied families of automorphic representations on G with discrete series at infinity and showed they satisfied averaged Sato-Tate laws coming from G itself. This corresponds to the heuristic that most π with discrete series at infinity should be "primitive": i.e. have $H_\pi = G$ (see e.g. Theorem 1.3 in [45] for a precise version of such a claim).

In Section 12.3, we instead study families with certain cohomological (possibly nontempered) components π_0 at infinity of a particular kind, which we refer to as "odd GSK-maxed" as defined in 11.2.11. In this case, the resulting averaged Sato-Tate laws are not those from G itself, but rather from certain H_{π_0} that we explicitly compute from π_0. The pairs (φ, H_{π_0}) are not in general endoscopic embeddings, but instead compositions of endoscopic embeddings ${}^LH \hookrightarrow {}^LG$ and tensor-product maps analogous to $\mathrm{GL}_n \times \mathrm{GL}_m \to \mathrm{GL}_{nm}$. This can be taken as evidence for a speculative interpretation that is nevertheless clearly suggested by the form of the endoscopic classification: that the decomposition of an A-parameter in terms of cuspidal parameters is literally realizing the corresponding packet as a functorial transfer from a smaller, possibly *non-endoscopic* group.

1.1.5. *Application: Sarnak-Xue Density.* The Sarnak-Xue density conjecture aims to quantify the failure of the naïve generalized Ramanujan conjecture—that all the local components of cuspidal automorphic representations are tempered. Naïve Ramanujan was the generalization and translation to representation theoretic, adelic language of the classical Ramanujan conjecture bounding Hecke eigenvalues of the Δ function—see [66] for a full introduction. It was found to be false even for split G through counterexamples constructed in [39].

Luckily, many desired applications don't require the absence of non-tempered representations: only that there to be not too many. Sarnak and Xue in [68] conjectured a precise meaning for "too many": the exact, necessary upper bound on the asymptotic growth rate of representations π with component π_v non-tempered. They also proved it in some small-rank cases.

Their original conjecture was stated in terms of classical, real locally symmetric spaces. Stating it in more modern language and focusing on the case of cohomological representations, let G/F be reductive and let U_i be a sequence of open compact subgroups of $G^\infty := G(\mathbb{A}^\infty)$ decreasing to the identity. Let $\Gamma_i = G(F) \cap U_i$ and choose a cohomological representation π_0 of G_∞. Then:

Conjecture 1.1.1 (Cohomological Sarnak-Xue)**.** *Let $m(\pi_0, \Gamma_i \backslash G_\infty)$ be the multiplicity of π_0 in $L^2(\Gamma_i \backslash G_\infty)$. Then for all $\epsilon > 0$,*

$$m(\pi_0, \Gamma_i \backslash G_\infty) \ll_\epsilon \mathrm{vol}(\Gamma_i \backslash G_\infty)^{\frac{2}{p(\pi_0)} + \epsilon}$$

where $p(\pi_0)$ is the infimum over p such that the (spherically finite) matrix coefficients of π_0 are in $L^p(G_\infty)$.

As alluded to in 1.1.3, it is known that if π_0 is discrete-series (i.e. $p(\pi_0) = 2$) then $m(\pi_0, \Gamma_i \backslash G_\infty) \sim \mathrm{vol}(\Gamma_i \backslash G_\infty)$. On the other hand, if π_0 is a character, (i.e. $p(\pi_0) = \infty$), then $m(\pi_0, \Gamma_i \backslash G_\infty) \sim 1$. This is therefore a claim that an asymptotically negligible fraction of automorphic representations have some non-tempered component π_0 at infinity and further that the quantitative strength of "asymptotically negligible" depends on the failure of temperedness $p(\pi_0)$ and is an interpolation between the cases $p = 2$ and $p = \infty$.

Many analytic applications are discussed in [35] including those from [30] relating to certain constructions (so-called golden gates and Ramunjuan complexes) used in computer science. There have been many recent breakthroughs proving the conjecture in specific cases: for example, [16] proved a version of the conjecture at infinity for Maass forms on GL_n using the Kuznetsov formula and [33] proved many versions for Maass forms on products of $\mathrm{SL}_2\mathbb{R}$ and $\mathrm{SL}_2\mathbb{C}$ using Arthur's trace formula.

Most important in our context is the work of Marshall and collaborators applying Arthur's classification to the problem at infinity for cohomological π_0 on unitary groups. The most general results are in [52] and prove the cohomological Sarnak-Xue conjecture for groups that are $U(N, 1)$ at infinity. As one implication of this project, we are able to show that, when inputted into our general framework, Marshall-Shin's bounds extend to prove the Sarnak-Xue conjecture at infinity for cohomological π_0 on *all* unitary groups that don't have a $U(2, 2)$ factor at infinity.

1.2. **Results.** To make this all precise, let E/F be a CM extension of number fields (i.e. a totally imaginary quadratic extension of a totally real field). Let G/F be a unitary group that splits over E, so that G has discrete series at infinity. We prove two main results, both conditional on the endoscopic classification of representations as stated in [44].

First, assume E/F is unramified at all finite places. Let π_0 be a in certain class of good cohomological representations π_0 of G_∞ that satisfy a condition of being odd GSK-maxed (as in Definition 11.2.10) and a technical parity condition from Lemma 11.3.1. Theorem 11.4.1 then finds explicit constants $R(\pi_0)$ and $M(\pi_0) \neq 0$ such that for split-level principal congruence subgroups U of G^∞,

$$\mathrm{vol}(U)^{R(\pi_0)/\dim G} L(U) \sum_{\pi \in \mathcal{AR}_{\mathrm{disc}}(G)} \mathbf{1}_{\pi_\infty = \pi_0} \dim((\pi^\infty)^U) = M(\pi_0) + O(\mathrm{vol}(U)^C)$$

for some correction factor $L(U)$ made precise in the theorem statement and some constant $C \geq 1/\dim G$. The sum should be thought of as the multiplicity of π_0 inside L^2 functions on the automorphic quotient $G(F)\backslash G(\mathbb{A})/U$ (which has volume proportional to $\mathrm{vol}(U)^{-1}$).

The theorem also allows weighting the count by a Weyl-symmetric polynomial P in the Satake parameters s_{π_v} at a place v where U has hyperspecial component:

$$\begin{aligned}(1.2.1)\quad \mathrm{vol}(U)^{R(\pi_0)/\dim G} L(U) &\sum_{\pi \in \mathcal{AR}_{\mathrm{disc}}(G)} \mathbf{1}_{\pi_\infty = \pi_0} \dim((\pi^\infty)^U) P(s_{\pi_v}) \\ &= M(\pi_0) M(P) + O(\mathrm{vol}(U)^C q_v^{A + B \deg P})\end{aligned}$$

for some explicit constant $M(P)$ and inexplicit constants A, B.

The second result, Theorem 11.4.2, applies for G arbitrary and π_0 arbitrary cohomological. It only provides upper bounds

$$\sum_{\pi \in \mathcal{AR}_{\mathrm{disc}}(G)} \mathbf{1}_{\pi_\infty = \pi_0} \dim((\pi^\infty)^U) P(s_{\pi_v}) = O(\mathrm{vol}(U)^{R(\pi_0)} q_v^{A + B \deg P}). \tag{1.2.2}$$

When π_0 is discrete series, we recover the "trivial bound" of $R(\pi_0) = \dim G$. Otherwise, we get an improvement $R(\pi_0) < \dim G$.

Beyond some specific lower bounds on classical groups in [24] extended to exact asymptotics in [25, cor. 1.3] on specific cohomological representations of symplectic groups, Theorem 11.4.1 seems to be the first exact asymptotic for counts of automorphic representations with non-tempered cohomological factor at infinity. As far as the authors are aware, it is the first with an estimate on the sub-leading term. It therefore gives many new corollaries:

1.2.1. *Corollary: Cohomology.* Our results give bounds for the growth of Betti numbers in towers of arithmetic manifolds. In this context, it is traditional and simplest, though by no means necessary, to fix a G_∞ which is isomorphic to $U(p,q)$, with $p+q = N$, at one infinite place and compact at all the others. Letting $U = K(\mathfrak{n})$ be the principal congruence subgroup associated to ideal $\mathfrak{n}$ (ignoring ramified places in this discussion for simplicity), the resulting $\Gamma(\mathfrak{n}) = G(F) \cap K(\mathfrak{n})$ are then cocompact lattices in $U(p,q)$. Our results in this context imply two types of corollaries.

First, using the bounds of 11.4 and the explicit description of packets of cohomological representations in 11.2 and 11.1, we give an algorithm to compute upper bounds on the growth of the Betti numbers $h^k(\Gamma(\mathfrak{n}))$ and of the dimensions $h^{p,q}(\Gamma(\mathfrak{n}))$ of any piece of the Hodge decomposition. Though we don't expect the resulting upper bounds to be sharp in general, they should be in many cases and they always give a non-trivial power saving when compared to the volume of $X(\mathfrak{n}) = \Gamma(\mathfrak{n}) \backslash G_\infty / K_\infty$ in degrees strictly below $\frac{1}{2} \dim X(\mathfrak{n})$. The algorithm is described in 12.2.2 and can be outlined as:

(1) Given a degree of cohomology or a Hodge weight, list all representations π_∞ for which the corresponding $(\mathfrak{g}, K)$-cohomology is nonvanishing using the parameterization of 11.1.
(2) For each representation π_∞, compute the shape $\Delta^{\max}(\pi_\infty)$ as in 11.2.2.
(3) Use Theorem 11.4.2 to give upper bounds
$$\dim \mathrm{Hom}(\pi_\infty, L^2(G(F) \backslash G(\mathbb{A}) / K(\mathfrak{n}) K_\infty)) \ll |\mathfrak{n}|^{R(\Delta^{\max}(\pi_\infty))}.$$
(4) From Matsushima's formula and the growth $\gg |\mathfrak{n}_i|^{1-\epsilon}$ of components of $G(F) \backslash G(\mathbb{A}) / K(\mathfrak{n}_i)$, deduce upper bounds for $h^k(\Gamma(\mathfrak{n}))$.

In practice, the combinatorics of computing the representations which contribute to cohomology in a given degree, as well as the corresponding $\Delta^{\max}(\pi_\infty)$ rapidly get complicated, but in some cases, the bounds can be expressed succinctly. For example, when $r = \min(p,q)$ is the smallest degree carrying the cohomology of a non-trivial representation, the contributions appear in weights $(r,0)$ and $(0,r)$ so we have

$$h^{r,0}(\Gamma(\mathfrak{n})) + h^{0,r}(\Gamma(\mathfrak{n})) \ll_\epsilon |\mathfrak{n}|^N + \epsilon.$$

As second corollary, our exact asymptotics give lower bounds on a range of degrees of cohomology when E/F is unramified. For example, using again $r = \min(p,q)$, Corollary 12.2.2 exhibits lattices such that for $1 \le j \le |p-q|-1$ and $j \not\equiv N \mod 2$,

$$h^{kj,(r-k)j}(\Gamma(\mathfrak{n})) \gg |\mathfrak{n}|^{Nj}, \quad 0 \le k \le r.$$

These results apply to a wider range of the degree when $U(p,q)$ is farther away from being quasisplit. In the extremal case where the noncompact part of G is $U(N-1,1)$, we deduce that in degrees j whose parity is opposite to that of N, the upper bounds of [52] are sharp.

1.2.2. *Corollary: Averaged Sato-Tate.* The error bound on our sub-leading term in Theorem 11.4.1 is as strong as in [73] so we can mimic their argument and prove an averaged Sato-Tate law. This is Theorem 12.3.3.

More specifically, given an odd GSK-maxed representation π_0 on an unramified unitary group G, we compute an unordered sequence of pairs $((T_i, d_i))_{1 \le i \le r}$ that is common to all elements of a set $\Delta^{\max}(\pi_0)$ understood by the algorithm at the end of §11.2.2. To this list of pairs we associate a group

$$H_{\pi_0} = (U_{E/F}(T_1) \times U_{E/F}(1)) \times \cdots \times (U_{E/F}(T_r) \times U_{E/F}(1))$$

and an L-embedding $\varphi : {}^L H_{\pi_0} \to {}^L G$ constructed in three stages: first we embed the second coordinate of each pair into ${}^L U_{E/F}(d_i)$ through the cocharacter corresponding to the parameter of the trivial representation. Then we take the tensor product embedding

$${}^L U_{E/F}(T_1) \times {}^L U_{E/F}(d_i) \to {}^L U_{E/F}(T_i d_i)$$

followed by the diagonal embedding

$$\prod_i {}^L U_{E/F}(T_i d_i) \hookrightarrow {}^L U_{E/F}(T_1 r_1 + \cdots + T_r d_r) = {}^L G.$$

The group ${}^L H_{\pi_0}$ has a canonical Sato-Tate measure on the space of Satake parameters for each splitting type θ of prime in E/F. We let $\mu_\theta^{\mathrm{ST}(\pi_0)}$ be the pushforward of this measure to G.

For each finite place v of the right splitting type and open compact $U \subseteq G^\infty$, we then define empirical measure:

$$\mu_{U,v}^{\pi_0} = \sum_{\substack{\pi \in \mathcal{AR}_{\mathrm{disc}}(G) \\ \pi_\infty = \pi_0}} \dim\left((\pi^\infty)^U\right) \delta_{s_{\pi_v}}$$

as a sum of delta-measure on the space of Satake parameters s_{π_v}. Then Theorem 12.3.3 show that for certain sequences of U_i and v_i such that $\mathrm{vol}(U)^{-1}$ grows much faster than q_{v_i}, we have weak convergence

$$C(\pi_0, U_i)^{-1} \mu_{U_i, v_i}^{\pi_0} \to \mu_\theta^{\mathrm{ST}(\pi_0)} \tag{1.2.3}$$

for an appropriate scaling factor $C(\pi_0, U)$ as long as a parity condition from Lemma 11.3.1 holds. This can be heuristically interpreted as evidence that most $\pi \in \mathcal{AR}_{\mathrm{disc}}(G)$ with $\pi_\infty = \pi_0$ are functorial transfers from H_{π_0} through φ.

1.2.3. *Corollary: Sarnak-Xue.* The bounds in Theorem 11.4.2 are good enough to achieve the Sarnak-Xue bounds for cohomological component at infinity on all unitary groups without a $U(2,2)$ factor at infinity in Theorem 12.4.7. It is very important to mention that the local bounds proved in [52] were already good enough to achieve the Sarnak-Xue threshold. The new work here is inputting them into a more general framework that applies beyond $U(N,1)$.

Nevertheless, we do not expect these bounds to be optimal. Through some heuristics relating to GK-dimension, we conjecture an optimal exponent $R_0(\pi)$ in Section 9.6. By the algorithm at the end of §11.2.2, π_0 gets associated an unordered sequence of pairs $((T_i, d_i))_i$ corresponding to the "most tempered" decomposition of Arthur parameters into cuspidals that allows component π_0 at infinity. As in the previous section about the Sato-Tate conjecture, this sequence should heuristically be thought of as determining an L-embedding ${}^L H_{\pi_0} \hookrightarrow {}^L G$ with respect to which almost all forms associated to π with $\pi_\infty = \pi_0$ are primitive. Then:

Conjecture 1.2.1. *The optimal exponent on* $\operatorname{vol}(U)$ *in* (1.2.2) *is*

$$R_0(\pi) = \frac{1}{2}\left(N^2 - \sum_i T_i^2 d_i\right) + \sum_i \left(T_i^2 + \frac{1}{2} T_i (T_i - 1)(d_i^2 - 1)\right).$$

We compare our bound R, the optimal bound R_0, the Sarnak-Xue threshold, and the trival bound $\dim G$ in many cases in Table 12.4.1.

1.2.4. *Conditionality.* As a very important warning, our argument depends heavily on Mok's and Kaletha-Minguez-Shin-White's endoscopic classifications for unitary groups [58] and [44]. The first depends on some references in Arthur's book [11] (and their analogues for unitary groups) that are not yet publicly available. The second in addition pushes many technical details to a specific reference "KMSb" that is also not yet publicly available. All these missing references are expected to be completed soon.

1.3. **Summary of Argument.** We prove the result using the Arthur-Selberg trace formula (see [10] for a review). From the perspective of this project, it is an attempt to give an explicit "geometric side" formula for

$$\sum_{\pi \in \mathcal{AR}_{\mathrm{disc}}(G)} \operatorname{tr}_\pi f$$

for some compactly supported, smooth test function f on $G(\mathbb{A})$. There are two main obstacles in applying it directly to our statistical problem:

- Given our chosen π_0, we would need to find a smooth compactly supported test function f_∞ such that $\operatorname{tr}_\pi f_\infty = \mathbf{1}_{\pi = \pi_0}$ to pick out only automorphic representation with component π_0 at infinity in this sum. This would be the simplest way to understand
$$m(\pi_0, f^\infty) := \sum_{\substack{\pi \in \mathcal{AR}_{\mathrm{disc}}(G) \\ \pi_\infty = \pi_0}} \operatorname{tr}_{\pi^\infty}(f^\infty) = \sum_{\pi \in \mathcal{AR}_{\mathrm{disc}}(G)} \operatorname{tr}_\pi(f_\infty f^\infty).$$
- The Arthur-Selberg trace formula is not very explicit in general.

Both these obstacles can be removed for the case of π_0 a discrete series. Here, f_∞ can be chosen to be a pseudocoefficient of [23]. The paper [7] (with some addenda in [31]) then showed that the geometric side of Arthur's invariant trace formula $I^G_{\mathrm{disc}}(f_\infty f^\infty)$ simplifies to something tractable with f_∞ a pseudocoefficient. In [73] (with some addenda in [26]), this formula was understood well enough for the purposes of computing asymptotics with error terms as we desire.

As soon as we try to generalize to all cohomological π_0, finding f_∞ becomes a much larger issue—in fact, for non-tempered cohomological π_0, there is no such test function by the results of [23]. The endoscopic classification of [44] gives a way out: π_0 is contained some special finite sets of unitary irreducible representations called A-packets Π_{ψ_∞} attached to A-parameters ψ_∞. It turns out that we can find a pseudocoefficient associated to π_0 such that, while it doesn't isolate π_0 amongst all unitary irreducibles, it isolates it amongst those with which it shares an A-packet (Lemma 3.4.3).

Next, the endoscopic classification also gives a decomposition

$$I^G_{\mathrm{disc}}(f_\infty f^\infty) = \sum_{\psi \in \Psi(G)} I^G_\psi(f_\infty f^\infty)$$

into pieces corresponding to global A-parameters ψ. The I^G_ψ only involves traces against automorphic representations π such that $\pi_\infty \in \Pi_{\psi_\infty}$ where ψ_∞ is the associated local parameter at infinity. Understanding $m(\pi_0, f^\infty)$ is then reduced to finding an explicit, geometric expression for the part of the trace formula corresponding only to those A-parameters ψ with component ψ_∞ at infinity such that $\pi_0 \in \Pi_{\psi_\infty}$.

We do something slightly different, defining instead a global invariant Δ called the *refined shape* for those A-parameters ψ. The Δ's have two key properties:

- Δ determines the local parameter ψ_∞ at infinity if it is cohomological,
- There is a an inductive method developed in [76] to write

  $$I^G_\Delta(f_\infty f^\infty) := \sum_{\psi \text{ with inv. } \Delta} I^G_\psi(f_\infty f^\infty)$$

  as a linear combination of terms of the form $I^H_{\mathrm{disc}}(f'_\infty (f')^\infty)$ with the f'_∞ pseudocoefficients on smaller groups H—i.e. these are terms that are already understood explicitly by [73] and [26].

Summing the inductive expressions for those Δ that correspond to ψ_∞ with $\pi_0 \in \Pi_{\psi_\infty}$, we would therefore get an explicit geometric formula for our desired

$$m(\pi_0, f^\infty) := \sum_{\substack{\pi \in \mathcal{AR}_{\mathrm{disc}}(G) \\ \pi_\infty = \pi_0}} \mathrm{tr}_{\pi^\infty}(f^\infty).$$

For technical reasons, we instead work with the analogous summand S_Δ of Arthur's stable S_{disc} (see [11, §3.1-3.3]). The two end up being more or less interchangeable for asymptotics by the "hyperendoscopy" techniques of [31] as used in [26].

Sections 2 and 3 give the necessary background material; §2 focused on the endoscopic classification and §3 on the real representation theory surrounding cohomological representations, their A-packets, and the pseudocoefficients. The definition of refined shape is discussed in §4 while the inductive procedure to understand S_Δ is explained in §5. Going from S_Δ to $m(\pi_0, f^\infty)$ is the work of §10.

Unfortunately, serious technical obstacles intrude. The inductive procedure of §5 in general requires the construction and computation of certain transfers of f^∞: the conjectural stable and "Speh" transfers described in §6.4. We do not even attempt to understand these in general. Instead, throughout §6, we compute them in various already-known cases while otherwise proving only inequalities. This is the main barrier towards producing exact asymptotics for either E/F ramified or for representations π_0 beyond GSK-maxed. It is also the main barrier to generalizing our techniques to quasisplit symplectic and orthogonal groups.

Computing transfers isn't the only technical obstruction: the sign $\epsilon_\psi(s_\psi)$ in the stable multiplicity formula 2.6.3 also confounds the inductive expansion of S_Δ. We are therefore only able to compute exact asymptotics for Δ such that this sign is always positive, only proving inequalities otherwise. This is the main barrier towards Theorem 11.4.1 applying to all GSK-maxed π_0 instead of just the odd GSK-maxed.

Sections 7 and 8 make up the technical work of squeezing as much information about S_Δ as possible from this partial information about the inductive expansion. The key conclusions are Propositions 7.3.3 on general Δ and 8.2.2 on a odd GSK Δ. A reformulation of the main technical trick of [34] plays an important role as Proposition 7.1.2. Finally, Section 9 completes the full inductive analysis of S_Δ, concluding with the main technical results of the paper: Theorems 9.3.2 and 9.5.1. The latter requires one final detail: a strengthening of 9.3.2 to Corollary 9.4.4 using local bounds from [52] reformulated as Lemmas 9.4.2 and 9.4.3. In §9.6, we highlight a heuristic for and possible strategy to prove conjectural optimal versions (Conjectures 9.6.3 and 9.6.4) of Corollary 9.4.4.

All that remains is the previously mentioned work in §10 writing $m(\Delta, f^\infty)$ in terms of S_Δ and the computations in §11 using a parameterization of cohomological representations on unitary groups to get explicit numbers for $R(\pi_0)$ and $M(\pi_0)$. This produces our main results: exact asymptotic Theorem 11.4.1 and upper bound Theorem 11.4.2. The last section, 12, first computes more details for specific example π_0 and second, applies the two main theorems to growth of cohomology, Sato-Tate equidistribution in families, and Sarnak-Xue density.

1.3.1. *Possible Extensions.* We highlight the barriers towards generalizing our result further. The first is generalizing the computation of the stable transfer of indicators of congruence subgroups in Lemma 6.1.2 to non-split places. This would of course allow us to extend theorems 11.4.1 and 11.4.2 to $\mathfrak{n}$ divisible by non-split primes. More importantly, it would allow for extending the techniques here to the case of quasisplit symplectic and orthogonal G, where there are no "split places" v such that $G_v \cong \mathrm{GL}_n$.

Next, improving the bounds 6.1.3, 9.4.1, 9.4.2, and 9.4.3 on Speh transfers of indicators of congruence subgroups would allow tightening the bound in Theorem 11.4.2, possibly even to the conjectural optimal value 9.6.3. As explained in §9.6, it would help tremendously to have a good enough understanding of the local character expansions of Speh representations through the rich interplay of ideas involving generalized Whittaker models and A-parameters as studied in [55] and [40]. Beyond even this, in the dream case where exact formulas can be computed, Theorem 11.4.1 may even be extended to general π_0 instead of just those that are GSK-maxed.

Relatedly, 11.4.1 is restricted to π_0 that are odd GSK-maxed instead of generally GSK-maxed because of our inability to control signs $\epsilon_\psi(s_\psi)$ that appear in the stable multiplicity formula 2.6.3. Proving that these signs cause any cancellation at all in S_Δ for Δ where they are non-trivial would allow us to remove the "odd" restriction. See the discussion around Conjecture 9.6.4.

Finally, just proving the existence of Stable and Speh transfers for functions at ramified places as in §6.4 would allow extending Theorem 11.4.1 to unitary groups for ramified quadratic extensions E/F.

1.4. **How to Read.** Sections 2 and 3 are background material that can mostly be skipped by experts. Sections 4 and 5 are the conceptual heart of the paper and should be understood very well in full detail before moving on. Sections 6, 7, and 8 are technical details involved in implementing the strategy of §5. We recommend skipping them on a first read through and referring back depending on need/interest while reading §9, 10. The most important results from the technical sections to understand the statements of are Propositions 7.3.3, 8.2.2, and Corollary 8.1.2.

Finally, Section 11.2 contains many pages of very involved but extremely elementary combinatorial arguments with the parameterization of cohomological representations on unitary groups. Unless very interested, we recommend reading enough to understand the definitions and statements while ignoring proofs. Subsection 12.4 on Sarnak-Xue density contains some sub-subsections that should be treated similarly.

Due to the length of the write-up and density of cross-references in later sections, we highly recommend reading this work electronically on a PDF reader that can handle intra-document hyperlinks and that has a back button.

1.5. **Acknowledgements.** The idea for this project started in two places: when the first author was taught about [76] by Olivier Taïbi at the CIRM workshop "Periods, functoriality and L-functions" and in conversations between the two authors at the 2022 Arizona Winter School. Many helpful conversations also happened at the 2022 Midwest Representation Theory Conference, the ESI workshop "Minimal Representations and Theta Correspondence", the IHES summer school on the Langlands Program, and the "Community Building in the Langlands Program" conference in Bonn.

Masao Oi provided us the full argument of Lemma 6.3.6 and Jeffrey Adams explained to us the argument of §3.2.3. In addition to many useful exchanges, Simon Marshall suggested the strategy to compute of the "Sarnak-Xue invariants" in section 12.4. We would also like to thank Patrick Allen, Alexander Bertoloni-Meli, Antonio Cauchi, Gaëtan Chenevier, Andrea Dotto, Peter Dillery, Matt Emerton, Melissa Emory, Shai Evra, Jessica Fintzen, Solomon Friedberg, Wee Teck Gan, Radhika Ganapathy, Henrik Gustafsson, Alexander Hazeltine, Pol van Hoften, Ashwin Iyengar, Tasho Kaletha, Gil Moss, Samuel Mundy, Alberto Minguez, Yiannis Sakellaridis, Peter Sarnak, Gordan Savin, David Schwein, Sug Woo Shin, Joel Specter, Loren Spice, and Tian An Wong for pointing out many useful arguments and also for pointing out when to consider changing strategies away from previous attempts at the proof that might not have been the most feasible.

1.6. **Notation.**

1.6.1. *Global Variables:* As some notation used throughout:
Basics:

- $F/\mathbb{Q}$ a totally real number field
- ∞ the set of infinite places of F
- $\mathcal{O}_F$ the ring of integers of F
- places of F will be denoted by v, with completion F_v
- q_S is the product of residue field degrees over a finite set of finite places v
- E/F an imaginary quadratic extension
- Γ_F the absolute Galois group of F
- $\Gamma(E/F) = \langle\sigma\rangle$, the Galois group of E over F
- $\omega_{E/F}$ is the order-2 character associated to the quadratic extension E/F
- $\star_v$ is the local component at v of structure $\star$.
- $\star^S$ and $\star_S$ are components at S and away from S of structure $\star$ for S some set of places
- "Irreps" are irreducible representations

- "Unirreps" are unitary irreducible representations
- $v_1 \boxtimes v_2$ is the corresponding representation of $H_1 \times H_2$ if v_i is a representation of H_i
- $[d]$ is the d-dimesional irrep of SL_2
- $\mathbf{1}_X$ is the indicator function of the set X
- $\bar{\mathbf{1}}_X$ is the indicator distribution of the set X (indicator function normalized by volume to have integral 1)
- $\mathbf{1}_{x=y}$ is an indicator function if $x = y$

Adelic Groups:

- $G_v = G(F_v)$ for v a place of F and G/F reductive
- $G_S, G^S = G(\mathbb{A}_S), G(\mathbb{A}^S)$ respectively for S a set of places of F and G/F reductive
- $\Omega_G, \Omega_{G,F}$ is the (geometric, F-rational) Weyl group of G
- ρ_G is the half-sum of positive roots of G
- K_v^G is a chosen hyperspecial of G/F reductive at unramified place v.
- $\mathscr{H}^{\mathrm{ur}}(G_v)$ is the unramfied Hecke algebra for G/F with respect to K_v^G
- $K_v^G(q_v^k)$ is the kth Moy-Prasad filtration group of K_v^G for G/F reductive and unramified at place v.
- $K^G(\mathfrak{n})$ for $\mathfrak{n}$ an ideal of $\mathcal{O}_F$ at which reductive G/F is unramified is the product of $K_v^G(q_v^k)$ that is the congruence subgroup corresponding to $\mathfrak{n}$
- $\Pi_{\mathrm{disc}}^G(\lambda)$ is the discrete series L-packet with infinitesimal character λ for the real group G
- φ_{π_d} is the pseudocoefficient of the discrete series representation π_d
- EP_λ is the (endoscopically normalized) Euler-Poincaré function for infinitesimal character λ

The Endoscopic Classification:

- $G(N)$ is the GL_N-like group defined in §2.1
- $\widetilde{G}(N)$ is the GL_N-like twisted group defined in §2.1
- $U(N) := U_{E/F}(N)$ is the quasisplit unitary group as in §2.1
- $U(p,q)$ is the indefinite unitary group with signature (p,q) over $\mathbb{R}$
- $\mathcal{E}_{\mathrm{ell}}(N), \mathcal{E}_{\mathrm{ell}}(G)$ are the elliptic endoscopic groups of $\widetilde{G}(N), G$ respectively as in §2.1.3
- $\mathcal{E}_{\mathrm{sim}}(N)$ are the simple endoscopic groups of $\widetilde{G}(N)$ as in §2.1.3
- f^H for f a test function on reductive G/F is a choice of endoscopic transfer to some $H \in \mathcal{E}_{\mathrm{ell}}(G)$
- f^N for f a test function on $G \in \mathcal{E}_{\mathrm{ell}}(N)$ is a test function on $\widetilde{G}(N)$ that transfers to f
- $\psi = \oplus \tau_i[d_i]$ is an Arthur parameter with cuspidal components τ_i as in §2.2.1
- $\Psi(N), \Psi(G)$ are the sets of parameters associated to $\widetilde{G}(N)$ and G respectively as in §2.2.1 and §2.2.4
- The "Arthur-SL_2" of a parameter is an unordered partition Q representing its restriction to the Arthur-SL_2
- π_ψ is the automorphic representation of $G(N)$ corresponding to ψ as in §2.2.2
- $\tilde{\pi}_\psi$ is the extension of π_ψ to $\widetilde{G}(N)$ as in §2.2.3
- φ_ψ is the L-parameter associated to A-parameter ψ as in equation (2.3.1)

- $\mathcal{S}_\psi, \mathcal{S}_{\psi_v}$ are Arthur's component groups associated to global or local parameters as in §2.4.1, 2.4.2
- $S^\natural_\psi, S^\natural_{\psi_v}$ are Kaletha's larger component groups associated to global or local parameters as in §2.4.1, 2.4.2
- s_ψ, s_{ψ_v} are the special elements identified in these component groups as in §2.4.1, 2.4.2
- ϵ_ψ is the identified character on global $\mathcal{S}_\psi$ as in §2.4.3
- $\Pi_\psi(G), \Pi_{\psi_v}(G_v)$ are the A-packets associated to global or local A-parameters on group G, G_v.
- $\eta^{\psi_v}_{\pi_v}$ is the local character of $S^\natural_{\psi_v}$ associated to $\pi_v \in \Pi_{\psi_v}$
- η^ψ_π is the character of $\mathcal{S}_\psi$ associated to $\pi \in \Pi_\psi$
- $\mathrm{tr}_{\psi_v} := \mathrm{tr}^{G_v}_{\psi_v}$ is the stable packet trace for parameter ψ_v on group G_v

Shapes:

- $\Delta = (T_i, d_i, \lambda_i, \eta_i)$ is a refined shape as in §4.2
- $\psi \in \Delta$ means that A-parameter ψ has refined shape Δ
- $\Sigma_{\lambda,\eta}$ is the simple refined shape as in §4.2
- $\mathcal{S}_\Delta, s_\Delta$ are the component groups and special elements associated to refined shape Δ
- ψ^Δ_∞ is the local component at infinity associated to refined shape Δ
- $H(\Delta)$ is the $H \in \mathcal{E}_{\mathrm{ell}}(N)$ such that $\psi \in \Delta$ implies that $\psi \in \Psi(H)$
- $\Delta(\pi_0)$ is the set of refined shapes Δ such that $\pi_0 \in \Pi_{\psi^\Delta_\infty}$
- "GSK" and "odd GSK" are conditions on shapes defined in 8.2.1

Trace Formulas:

- $\mathcal{AR}_{\mathrm{disc}}(G)$ is the set of discrete automorphic representations of reductive G/F
- I^G, S^G are Arthur's invariant and stable trace formulas for group G
- $I^G_{\mathrm{disc}}, S^G_{\mathrm{disc}}$ are their discrete parts
- R^G is the trace against $\mathcal{AR}_{\mathrm{disc}}(G)$
- I^G_ψ, S^G_ψ are the summands of $I^G_{\mathrm{disc}}, S^G_{\mathrm{disc}}$ associated to parameter ψ
- I^G_Δ, S^G_Δ are the summands of $I^G_{\mathrm{disc}}, S^G_{\mathrm{disc}}$ associated to refined shape Δ
- $\star^N$ is the version of any of the variants above associated to $\widetilde{G}(N)$

Asymptotics:

- $|\mathfrak{n}|$ is the norm of the ideal $\mathfrak{n}$ of $\mathcal{O}_F$.
- $\Gamma_{n_1,\ldots,n_k}(\mathfrak{n}_i)$ is an Euler factor associated to $\mathfrak{n}$ and the list $n_1, \ldots, n_k$ in §9.1
- $\bar{R}(\Delta)$ is an upper bound on growth rate associated to refined shape Δ in Theorem 9.3.2
- $R(\Delta)$ is a tighter upper bound on growth rate associated to refined shape Δ in corollary 9.4.4
- $R_0(\Delta)$ is a conjectural optimal growth rate associated to refined shape Δ in conjecture 9.6.3
- $L_\Delta(\mathfrak{n})$ is an Euler factor associated to $\mathfrak{n}$ and GSK shape Δ in Theorem §9.5.1.
- $\tau'(G)$ is a modified Tamagawa number of reductive G/F as in [73, (9.5)].

Cohomological irreps of unitary groups

- $\mathcal{P}(N), \mathcal{P}(p,q), \mathcal{P}_1(p,q)$ are combinatorial parameter sets defined in 11.1.1
- β, δ are reduction maps between these parameterizing sets defined in (11.1.2)

- $\Delta^{\max}(\pi_0)$ is a subset of $\Delta(\pi_0)$ determined by 11.2.1
- $R(\pi_0)$ is the common value of $R(\Delta)$ for $\Delta \in \Delta^{\max}(\pi_0)$
- $Q^{\max}(\pi_0)$ is the set of Arthur-SL_2's of elements of $\Delta^{\max}(\pi_0)$ as in 11.2.4
- $Q_{\mathrm{can}}(\pi_0)$ is a certain unordered partition assigned to cohomological representation π_0 in 11.2.8
- "GSK-maxed", "odd GSK-maxed" are conditions on cohomological representations π_0 of unitary groups defined in 11.2.10, 11.2.11

1.6.2. *Shorthand for Non-Factorizable Functions.* There is a technical annoyance that certain transfer maps from functions on a group G to a functions on a product of groups $H_1 \times H_2$ may not always have image in factorizable functions.

However, they will always land in linear combinations of factorizable functions. Therefore, we will use the "mystical gate" notation

$$\prod_i^{\oplus} f_i.$$

to represent the sum of the factored term over this linear combination.

At some points, we also for simplicity want to elide the fact that a transfer to a group like H^d may not be the same on each H-factor. Therefore we will use the even more abusive notation

$$(f_1)^{d\oplus}$$

to represent a sum over the factorizable pieces of the product over factors for each H in the H^d.

1.6.3. *Sequences.* Some objects we will consider will be indexed by finite sequences $n_1, \ldots, n_k$. As shorthand, we will define the sequence

$$n_1^{(r_1)}, \ldots, n_k^{(r_k)} := \overbrace{n_1, \ldots, n_1}^{r_1}, \ldots, \overbrace{n_k, \ldots, n_k}^{r_k}.$$

Also, if L_1 and L_2 are sequences, "L_1, L_2" will represent their concatenation. We also represent concatenation of multiple lists by a disjoint union symbol:

$$\bigsqcup_i L_i.$$

Finally, if $P = (p_1, \ldots, p_k)$ is an ordered partition of n and $a = (a_i)_i$ is a list of length n, the P-parts of a are defined by partitioning a in order according to P:

$$\overbrace{\xi_1 \ldots, \xi_{n_1}}^{a_1^P}, \overbrace{\xi_{n_1+1} \ldots, \xi_{n_1+n_2}}^{a_2^P}, \cdots, \overbrace{\xi_{N-n_k+1} \ldots, \xi_N}^{a_k^P}.$$

2. A-parameters and the Classification

We attempt to summarize as concisely as possible the parts of endoscopic classification that are relevant to this project.

An Arthur/endoscopic classification for a group G is conceptually a "transfer" of two known facts about automorphic representations on GL_n—

- The classification of the discrete spectrum in [57],
- The local Langlands correspondence for local components

—to a parameterization of automorphic representations of G and their local components.

We will focus on the example of Mok's and Kaletha-Minguez-Shin-White's versions from [58] and [44] for quasisplit and general unitary groups respectively. Our summary will be in two pieces:

- A formalism of local and global parameters which encapsulates the known information on the GL_n side.
- A description of how the automorphic spectrum on G decomposes into pieces that correspond to each parameter together with a description of the structure of each of these pieces.

We will not go over background for endoscopy or the stable trace formula since sections 2.1 and 3.1-3 of [11] already give a good, relatively concise introduction with an eye towards the endoscopic classification.

2.1. **Groups Considered.** We begin needing to define certain groups and L-embeddings.

Fix a totally real number field F and totally complex quadratic extension E/F. For each $N > 0$, consider the group

$$G(N) = \mathrm{Res}^E_F \mathrm{GL}_{N,E}$$

Let θ_N be the automorphism of $G(N)$ in the outer class of conjugate inverse transpose that fixes the standard pinning. In particular, it is an involution. It can be written as $\theta_N(g) = \mathrm{Ad}(J_N)(\bar{g}^{-t})$ for a choice of J_N.

2.1.1. *Unitary Groups.* Let $U_{E/F}(N)/F$ be the reductive group

$$U_{E/F}(N, F) = \{g \in GL_N(E) : \theta(g) = g\} \tag{2.1.1}$$

It is a quasisplit unitary group and therefore a form of GL_N/F. We can choose a Borel and maximal torus (B, T) to be the upper triangular and diagonal θ-fixed matrices respectively

For any place v of F, we consider $U_{E/F}(N, F_v)$. When v is split in E, we have $U_{E/F}(N, F_v) \simeq GL_N(F_v)$. Otherwise, $U_{E/F}(N, F_v)$ is the unique quasisplit unitary group of rank N over F_v.

Finally, because of our choice of E/F, all inner forms of $U_{E/F}(N, F_\infty)$ will have discrete series.

2.1.2. *L-groups and embeddings.* All our groups split over E so in all our L-groups, the action of Γ_F factors through $\Gamma(E/F)$. We have

$${}^L G(N) = (GL_N(\mathbb{C}) \times GL_N(\mathbb{C})) \rtimes \Gamma_F$$

with σ swapping the two copies of $GL_N(\mathbb{C})$. We also have

$${}^L U_{E/F}(N) = GL_N(\mathbb{C}) \rtimes \Gamma_F,$$

with $\sigma(g) = \mathrm{Ad}(J_N)(g^{-t})$.

For $\kappa \in \pm 1$, we have L-embeddings

$$\xi_\kappa : {}^L U_{E/F}(N) \to {}^L G(N). \tag{2.1.2}$$

The explicit coordinates are not important to us and can be found in §2.1 of [58].

2.1.3. *Endoscopic data.* We are interested in the twisted endoscopic groups of $\widetilde{G}(N) = G(N) \rtimes \theta_N$. As in §2.4.1 in [58], these are parameterized as

$$\mathcal{E}_{\mathrm{ell}}(N) = \{U_{\kappa_1}(N_1) \times U_{\kappa_2}(N_2) : \kappa_i = \pm 1, N = N_1 + N_2, \kappa_1\kappa_2 = (-1)^{N-1}\}$$

with each $U_\pm(N_i)$ isomorphic to the quasisplit unitary group $U_{E/F}(N_i)$ and the κ_i determining the specific L-embedding ξ_κ. Among these we highlight the simple endoscopic groups:

$$\mathcal{E}_{\mathrm{sim}}(N) := \{U_+(N), U_-(N)\}.$$

Note that $U_\pm$ are isomorphic as groups.

We are also interested in the endoscopic groups of $G = U_{E/F}(N)$. As enumerated in [44, §1.1.1], these are parameterized as

$$\mathcal{E}_{\mathrm{ell}}(G) = \{U_{E/F}(N_1) \times U_{E/F}(N_2) : N = N_1 + N_2, N_1 \geq N_2\}.$$

We do not need the full information of the endoscopic triples involved. Beware that our $\mathcal{E}_{\mathrm{ell}}(G)$ is the $\overline{\mathcal{E}}_{\mathrm{ell}}(G)$ of [44].

2.1.4. *Inner Forms.* We will also consider extended pure inner forms of $U_{E/F}(N)$ as in [44, §0.3.3]. Since the general precise definition is complicated and not relevant to our computation, we simply recall their enumeration of possibilities for unitary groups.

Let $G \in \mathcal{E}_{\mathrm{ell}}(N)$. Then in the local case:

- If v is non-Archimedean and split in E, the extended pure inner forms of $G_v \cong \mathrm{GL}_{N,v}$ are of the form $\mathrm{Res}^{D_v}_{F_v} \mathrm{GL}_m$ for D_v a division algebra over F. They are associated invariant $a_v = N \cdot \mathrm{inv}(D_v)$.
- If v is non-Archimedean non-split[2] in E, the extended pure inner forms of $G_v \cong U_{E/F}(N)_v$ are:
  - $U_{E/F}(N)_v$ itself, with associated invariant $a_v = 0$,
  - Another form associated to $a_v = 1$. If N is odd, this is isomorphic as a group to $U_{E/F}(N)_v$. If N is even, it is the unique non-quasisplit inner form of $U_{E/F}(N)_v$.
- If v is Archimedean real in F, then the extended pure inner forms of $G_v \cong U_{\mathbb{C}/\mathbb{R}}(N)$ are the $U(p,q)$ for $p + q = N$ and associated invariant $a_v = N(N-1)/2 + q$. Note that $U(p,q) \neq U(q,p)$ as extended pure inner forms even though they are isomorphic as groups.

A choice of local extended pure inner form and each v comes from a global extended pure inner form if and only if:

- Almost all a_v are trivial,
- The sum of the a_v is even.

Note that if we only care about inner forms as groups, the second condition is irrelevant for N odd: we can always switch an infinite-place $G_v = U(p,q)$ to $U(q,p)$, which is isomorphic as a group but has opposite a_v.

We are particularly interested in the isomorphism-as-groups classes of extended pure inner forms that are unramified at all finite places. This is only possible when E/F is unramified at all finite places—for example

$$E/F = \mathbb{Q}[\sqrt{3}, i]/\mathbb{Q}[\sqrt{3}].$$

[2]While the current as-of-this-comment draft of [44] only says inert, this seems to be a typo since the arguments they give work for ramified places as well.

Casework with respect to the parity of N then gives:

Lemma 2.1.1. *Assume E/F is unramified at all finite places. Then every isomorphism-as-groups class of extended pure inner forms G of $G^* \in \mathcal{E}_{\mathrm{ell}}(N)$ that is unramified at all finite places has a representative that satisfies $G^\infty = (G^*)^\infty$ and is therefrore determined by*

$$G_\infty = \prod_{v \in \infty} U(p_v, q_v).$$

Such a choice is valid if and only if

$$\frac{N(N-1)}{2}|\infty| + \sum_{v \in \infty} q_v$$

is even.

We will henceforth only consider these representatives to normalize local transfer factors later on so that they are consistent with the fundamental lemma at all finite places.

2.2. **Global Parameters.**

2.2.1. *Definitions.*

Definition 2.2.1. A global A-parameter of rank N is a conjugate self-dual (through θ_N) formal expression

$$\psi = \tau_1[d_1] \oplus \cdots \oplus \tau_k[d_k]$$

up to reordering the summands and where each τ_i is a cuspidal automorphic representation of $G(T_i)$ (equivalently, one of GL_{T_i}/E), $d_i \in \mathbb{Z}^+$, and $\sum_i T_i d_i = N$.

Definition 2.2.2. We say $\psi = \tau_1[d_1] \oplus \cdots \oplus \tau_k[d_k]$ is

- *cuspidal* if $k = 1$ and $d_1 = 1$,
- *simple* or *stable* if $k = 1$,
- *generic* if each $d_i = 1$,
- *elliptic* if each τ_i is itself conjugate self-dual and the $\tau_i[d_i]$ are distinct.

Definition 2.2.3. Let $\Psi(N)$ be the set of elliptic parameters in $G(N)$.

All global parameters henceforth considered will be elliptic.

2.2.2. *Representations.* As explained in [11, §1.3], the main result of [57] associates a unique discrete automorphic representation π_ψ of GL_n/E to each Arthur parameter ψ. First, for simple $\tau[d]$, consider the parabolic induction

$$\mathrm{Ind}_{P(\mathbb{A})}^{\mathrm{GL}_N(\mathbb{A})}(\tau|\det|^{(d-1)/2} \boxtimes \tau|\det|^{(d-3)/2} \boxtimes \cdots \boxtimes \tau|\det|^{-(d-1)/2})$$

where P is the parabolic associated to ordered partition $(\dim \tau, \dots, \dim \tau)$. We let $\pi_{\tau[d]}$ be the unique Langlands quotient of this induction: it exists and is unitary.

For general $\psi = \bigoplus_i \tau_i[d_i]$, we let

$$\pi_\psi := \mathrm{Ind}_{P(\mathbb{A})}^{\mathrm{GL}_N(\mathbb{A})}(\boxtimes_i \pi_{\tau_i[d_i]})$$

where P is the appropriate parabolic. This is always unitary and irreducible.

2.2.3. *Canonical Extensions to $\widetilde{G}(N)$.* Fix a Whittaker datum ω for $G(N)$ giving local Whittaker data ω_v on each $G_v(N)$. Then each π_ψ for $\psi \in \Psi(N)$ has a canonical extension $\tilde{\pi}_\psi := \tilde{\pi}_{\psi,\omega}$ to $\widetilde{G}(N)$ as explained in §2.2 of [11] or §3.2 of [58]. We warn that this "choice of sign" is extremely important conceptually and not just a technicality to be ignored[3]—it enters crucially into the computation of various sign characters in the works of Arthur and Mok and is the main difficulty in understanding Conjecture 6.4.2 on stable transfer.

The $\tilde{\pi}_\psi$ is a product of extensions $\tilde{\pi}_{\psi,v}$ of each $\pi_{\psi,v}$. By the Langlands classification, each $\pi_{\psi,v}$ is the Langlands quotient of an induction

$$\mathrm{Ind}_{P_v}^{G(N)_v}(\sigma_1|\det|^{r_i} \boxtimes \cdots \boxtimes \sigma_k|\det|^{r_k})$$

where each σ_i is tempered and therefore generic. We choose the θ-action on σ_i to be the one that acts as $+1$ instead of -1 on its one-dimensional space of Whittaker functionals with respect to ω_v.

Finally, since ψ is conjugate self-dual, we necessarily have that $r_j = -r_{k-j}$ and that σ_j and σ_{r-j} are conjugate-duals of each other. Therefore we can choose P to be fixed by θ and can define the action of θ on $\pi_{\psi,v}$ as coming from the induction of the actions on each σ_i.

2.2.4. *Assignment to Groups in $\mathcal{E}_{\mathrm{ell}}(N)$.* As in remark 2.4.6 of [58], every $\psi \in \Psi(N)$ can be assigned to a unique element of $\mathcal{E}_{\mathrm{ell}}(N)$ through which it should be thought of as "factoring": If $\tau \in \Psi(N)$ is cuspidal, then Mok assigns it a parity δ. Then τ is assigned to $U_\delta(N)$. More generally, $\tau[d]$ is assigned to $U_\kappa(Nd)$ where

$$\kappa = \delta(-1)^{(N-1)(d-1)}.$$

Finally, if

$$\psi = \bigoplus_i \tau_i[d_i]$$

with each $\tau_i \in \Psi(T_i)$, let N_O be the sum of $T_i d_i$ such that $\tau_i[d_i]$ are *orthogonal*: i.e. $\delta_i(-1)^{T_i+d_i} = 1$. Similarly, let N_S be defined similarly for the $\tau_i[d_i]$ that are the opposite: *symplectic*. The discussion after 2.4.6 in [58] assigns to ψ the group

$$U_{(-1)^{N_O-1}}(N_O) \times U_{(-1)^{N_S}}(N_S) \in \mathcal{E}_{\mathrm{ell}}(N_O + N_S).$$

Definition 2.2.4. For $G \in \mathcal{E}_{\mathrm{ell}}(N)$, let $\Psi(G)$ be the subset of $\Psi(N)$ assigned to G.

2.2.5. *As Morphisms.* We can interpret global parameters as morphisms into ${}^L G(N)$. This is a technical replacement for not having access to the conjectural global Langlands group and will be useful for discussing component groups later.

Given a parameter

$$\psi = \bigoplus_i \tau_i[d_i] \in \Psi(N),$$

let $\tau_i \in \Psi(H_i) \subseteq \Psi(T_i)$ for H_i a group and T_i a number. Let μ_i be the embedding

$$\mu_i : {}^L H_i \hookrightarrow {}^L G(T_i)$$

from (2.1.2). Define the fiber product

$$\mathcal{L}_\psi := \prod_i ({}^L H_i \to W_F).$$

[3]As the authors learned at their own expense

Then we define map

$$\psi' : \mathcal{L}_\psi \times \mathrm{SL}_2 \hookrightarrow {}^L G(N) : \psi' = \bigoplus_i \mu_i \boxtimes [d]$$

where $[d]$ is the d-dimensional irreducible representation of SL_2.

Note by construction that for the $G \in \mathcal{E}_{\mathrm{ell}}(N)$ such that $\psi \in \Psi(G)$, ψ' factors through ${}^L G$. Finally we will refer to the restriction of ψ' to SL_2 as the Arthur-SL_2 of ψ. This can be represented as an unordered partition encoding its decomposition into irreps.

2.3. **Local Parameters.**

2.3.1. *Definitions.* We first define a formalism of local parameters.

Definition 2.3.1. Let G/F be a reductive group. A local A-parameter for G at v, $\psi_v \in \Psi_v(G)$, is an L-morphism

$$\psi_v : L_{F_v} \times \mathrm{SL}_2 \to {}^L G$$

where

- L_{F_v} is the Weil group W_{F_v} if v is Archimedean and the Weil-Deligne group WD_{F_v} is v is non-Archimedean,
- $\psi_v|_{L_v}$ is a bounded L-parameter,

and up to conjugacy in $\widehat{G}$.

A generalized local A-parameter, $\psi_v \in \Psi_v^+(G)$, is the same object without the boundedness condition. As shorthand, we also write $\Psi_v^\star(N) := \Psi_v^\star(\widetilde{G}(N))$.

As we will see below, we need to consider the set $\Psi_v^+(N)$ since the Ramanujan conjecture is at present unknown. We call L_{F_v} the local Langlands group and the SL_2 factor the Arthur-SL_2. As in the global case, we can represent the Arthur-SL_2 by an unordered permutation Q.

Next, every local parameter ψ_v has an associated L-parameter φ_{ψ_v} :

$$\varphi_{\psi_v} : L_{F_v} \to {}^L G : w \mapsto \psi_v \left(w, \begin{pmatrix} |w| & 0 \\ 0 & |w|^{-1} \end{pmatrix} \right). \tag{2.3.1}$$

We may also write $\psi_v \in \Psi_v(N)$ as:

$$\psi_v = \bigoplus_i \tau_i[d_i] := \bigoplus_i \tau_i \boxtimes [d],$$

where $[d]$ represents the d-dimensional representation of SL_2 and each τ_i is a representation of L_{F_v}.

As in Section 2.2.4, we have decompositions

$$\Psi_v(N) = \bigsqcup_{G \in \mathcal{E}_{\mathrm{ell}}(N)} \Psi_v(G), \qquad \Psi_v^+(N) = \bigsqcup_{G \in \mathcal{E}_{\mathrm{ell}}(N)} \Psi_v^+(G)$$

determined by parities $\eta_{i,v}$ assigned to irreducible τ_i (see §2.2 in [58]). We thereby extend the definitions of simple, stable, generic, and elliptic from Definition 2.2.2 to local parameters.

2.3.2. *Localization.* There is a localization map $\Psi(N) \to \Psi_v^+(N)$. Consider

$$\psi = \bigoplus_i \tau_i[d_i] \in \Psi(N)$$

with each $\tau_i \in \Psi(T_i)$ cuspidal. By local Langlands, the component $\pi_{\tau_i,v}$ at v of π_{τ_i} corresponds to an L-parameter

$$\varphi_{i,v} : L_{F_v} \to {}^L G(N)_v.$$

Then we define local A-parameter

$$\psi_v := \bigoplus_i \varphi_{i,v} \boxtimes [d_i].$$

This ψ_v is currently only known to be an element of $\Psi_v^+(N)$. However, the Ramanujan conjecture would imply that each $\varphi_{i,v}$ is bounded since they come from local components of cuspidal automorphic representations. This would make $\psi_v \in \Psi(N)$.

Localization is consistent with the global picture: first, comparing with the construction of π_ψ shows that $(\pi_\psi)_v$ has L-parameter φ_{ψ_v}. Moreover, Corollary 2.4.11 in [58] and the subsequent discussion show that if $\psi \in \Psi(G)$ for $G \in \mathcal{E}_{\mathrm{ell}}(N)$, then ψ_v factors through ${}^L G$—in other words, the localization map restricts to

$$\Psi(G) \to \Psi_v^+(G)$$

for $G \in \mathcal{E}_{\mathrm{ell}}(N)$.

In addition, the discussion after Corollary 2.4.11 explains how to produce localization maps

$$L_{F_v} \to \mathcal{L}_\psi \tag{2.3.2}$$

for any $\psi \in \Psi(N)$, such that ψ_v is the pullback of ψ through the localization.

2.4. **Centralizer Subgroups and ϵ-Characters.** Now we start defining notions needed to understand the structure of the part of the automorphic spectrum of $G \in \mathcal{E}_{\mathrm{ell}}(N)$ corresponding to some ψ.

2.4.1. *Global Centralizers.* To each global parameter $\psi \in \Psi(G)$, Mok attaches a component group $\mathcal{S}_\psi$ defined as follows:

$$S_\psi(G) := Z_{\widehat{G}}(\operatorname{im} \psi'),$$
$$\mathcal{S}_\psi := \pi_0(S_\psi / Z(\widehat{G})^{\Gamma_F}).$$

In addition, [44] attaches a larger component group $S_\psi^\natural$. By the discussion around (1.3.6) there, for our special case of unitary groups we may use the formula:

$$S_\psi^\natural = \pi_0(S_\psi).$$

Also define a distinguished element

$$s_\psi := \psi'(1 \times -1) \in S_\psi^\natural.$$

These groups are explicitly computed in [44] around (1.3.6): if

$$\psi = \bigoplus_{i \in I} \tau_i[d_i]$$

with τ_i cuspidal, then the I^+ mentioned is all of I since ellipticity of ψ means that the $\tau_i[d_i]$ all have multiplicity 1. Then there are canonical isomorphisms

$$S_\psi^\natural = (\mathbb{Z}/2)^I, \qquad \mathcal{S}_\psi = (\mathbb{Z}/2)^I / (\mathbb{Z}/2)^{\mathrm{diag}} \tag{2.4.1}$$

in which
$$s_\psi = \bigoplus_{\substack{i \in I \\ d_i \text{ even}}} 1.$$
Note that s_ψ is trivial in $\mathcal{S}_\psi$ if the d_i all have the same parity.

2.4.2. *Local Centralizers.* Mok also defines local component groups:
$$S_{\psi_v}(G) := Z_{\widehat{G}_v}(\operatorname{im}\psi_v),$$
$$\mathcal{S}_{\psi_v} := \pi_0(S_{\psi_v}/Z(\widehat{G_v})^{\Gamma}_{F_v}).$$
We also similarly have an $S^\natural_{\psi_v}$. As explained at the end of §1.2.4 in [44], we may again use the formula
$$S^\natural_{\psi_v} := \pi_0(S_{\psi_v})$$
in our special case of unitary groups. We also define
$$s_{\psi_v} := \psi_v(1 \times -1) \in \mathcal{S}_{\psi_v}.$$
Just as in the global case, $\mathcal{S}_{\psi_v}$ and $S^\natural_{\psi_v}$ can be computed explicitly, though the lack of a corresponding "elliptic" condition makes this slightly more complicated—see the end of §1.2.4 in [44] again for details. We will only need to worry about local component groups explicitly for very specific parameters at ∞, so these details aren't relevant here.

Finally, the localization maps $L_{F_v} \to \mathcal{L}_\psi$ give localization maps $\mathcal{S}_\psi \to \mathcal{S}_{\psi_v}$ and $S^\natural_\psi \to S^\natural_{\psi_v}$. Under these maps, we can identify s_{ψ_v} with s_ψ.

2.4.3. *ϵ characters.* Fix $H \in \mathcal{E}_{\mathrm{ell}}(N)$. The third structure we need to understand for global parameter $\psi \in \Psi(H)$ is a character ϵ_ψ on $\mathcal{S}_\psi$.

Definition 2.4.1. Let $\psi \in \Psi(H)$ and (ρ, V) a finite-dimensional representation of ${}^LH$. Then there is an action $\rho_\psi : S_\psi \times L_\psi \times \mathrm{SL}_2$ on V. Let ρ_ψ factor into irreducibles:
$$\sum_{j \in J} \sigma_j \otimes \gamma_j \otimes \delta_j.$$
Let J' be the set of indices such that

- γ_j is symplectic
- $\epsilon(1/2, \gamma_j) = -1$ (see Arthur for how to define this from local ϵ-factors).

Then we define
$$\epsilon^\rho_\psi : S_\psi \to \mathbb{C} : s \mapsto \prod_{j \in J'} \det(\sigma_j(s)).$$

Definition 2.4.2. Let $\epsilon_\psi := \epsilon^H_\psi$ be ϵ^ρ_ψ for ρ the adjoint representation of ${}^LH$ on $\operatorname{Lie}\widehat{H}$. Note that it factors through $\mathcal{S}_\psi$ itself.

The adjoint representation ρ on $\operatorname{Lie}(\hat{H})$ preserves the Killing form and is as such orthogonal, thus only the summands such that δ_i is symplectic, i.e. even dimensional, contribute non-trivially to ϵ_ψ. If $\widehat{H} = GL_N$ and $\psi = \oplus_i \tau_i[d_i]$, then $\rho_\psi|_{SL_2\mathbb{C}} = \bigoplus_i \bigoplus_j \nu(d_i) \otimes \nu(d_j)$. A computation of the dimensions of the irreducible constituents of $\nu(d) \otimes \nu(d')$ then shows:

Lemma 2.4.3. *If $\psi = \oplus_i \tau_i[d_i]$ and all the d_i have the same parity, then $\epsilon_\psi \equiv 1$.*

2.5. **Main Theorems of the Classification.** Now we can state the two main theorems of the endoscopic classification:

2.5.1. *Local Packets.* First, we state the existence of local A-packets. Let $G^* \in \mathcal{E}_{\mathrm{ell}}(G)$. We recall from (0.3.1) in [44] that every extended pure inner form G_v of G^*_v gets associated a character

$$\chi_{G_v} : Z(\widehat{G}_v)^\Gamma \to \mathbb{C}^\times.$$

This association is a bijection if v is non-Archimedean, and $\chi_{G^*_v}$ is trivial.

We define $\mathrm{Irr}(S^\natural_G, \chi)$ to be the set of trace characters of irreps of $S^\natural_G$ that pullback to χ through $Z(\widehat{G}_v)^\Gamma \to S^\natural_G$ (beware that this set can be empty—there is a condition of being "relevant" defined in [44, §0.4, 1.2] discussing when this happens).

Theorem 2.5.1 ([44, 1.6.1])**.** *Let $G^* \in \mathcal{E}_{\mathrm{ell}}(N)$ and $\psi_v \in \Psi(G^*_v)$. Fix a Whittaker datum on G^*_v. Then for each extended pure inner form G_v of G^*_v, there is an associated set $\Pi_{\psi_v}(G_v)$ of unitary representations of G_v together with a map*

$$\eta := \eta_{G_v} : \Pi_{\psi_v} \to \mathrm{Irr}(S^\natural_{\psi_v}, \chi_{G_v}), \quad \pi_v \mapsto \eta^{\psi_v}_{\pi_v}.$$

These satisfy:

- *Assume ψ is generic. If v is non-Archimedean, then η is a bijection. If v is Archimedean, then the maps $\eta_{G'_v}$ for G'_v such that $\chi_{G'_v} = \chi_{G_v}$ jointly give a bijection from the disjoint union of the $\Pi_{\psi_v}(G'_v)$.*
- *The $\Pi_{\psi_v}(G_v)$ for generic ψ_v partition the set of tempered unirreps of G_v.*

We also make a definition:

Definition 2.5.2. Let ψ_v be a local A-parameter for $G^* \in \mathcal{E}_{\mathrm{ell}}(N)$, G_v an extended pure inner form of G^*_v, and f_v a test function on G_v. We define the stable packet trace:

$$\mathrm{tr}_{\psi_v}(f_v) := \mathrm{tr}^{G_v}_{\psi_v}(f_v) := \sum_{\pi_v \in \Pi_{\psi_v}(G_v)} \eta^{\psi_v}_{\pi_v}(s_{\psi_v})\,\mathrm{tr}_\pi(f_v).$$

Theorem 2.5.3. *Let ψ_v, G_v as in the above definition. Then $\mathrm{tr}^{G_v}_{\psi_v}$ is a stable distribution on G_v.*

2.5.2. *Global Packets.* As before, let $G^* \in \mathcal{E}_{\mathrm{ell}}(G)$ and G be an extended pure inner form of G^*. If $\psi \in \Psi(G^*)$ we define

$$\Pi^G_\psi := \Pi_\psi = \prod_v{}' \, \Pi^G_{\psi_v},$$

where the product is restricted so that $\eta^{\psi_v}_{\pi_v} = 1$ at almost all places.

We recall from (0.3.3) in [44] that

$$\prod_v \chi_{G_v} = 1,$$

so that for any $\pi \in \Pi_\psi$,

$$\eta^\psi_\pi := \prod_v \eta^{\psi_v}_{\pi_v}$$

is a character on $\mathcal{S}_\psi$.

The characters η^ψ_π determine the decomposition of the discrete automorphic spectrum $\mathcal{AR}_{\mathrm{disc}}(G)$:

Theorem 2.5.4 (Arthur's Multiplicity Formula, [44] thm. 1.7.1)**.** *We have that*

$$\mathcal{AR}_{\mathrm{disc}}(G) = \bigoplus_{\psi \in \Psi(G^*)} \bigoplus_{\pi \in \Pi_\psi^G} m_\pi^\psi \pi$$

where

$$m_\pi^\psi = \langle \epsilon_\psi, \eta_\pi^\psi \rangle_{\mathcal{S}_\psi}$$

is the trace-character pairing.

2.6. **Trace Formula Decompositions.**

2.6.1. *Defintions.* Let I^G and S^G be Arthur's invariant and stable trace formulas for reductive G/F respectively and I^G_{disc} and S^G_{disc} their discrete parts. Also define distribution

$$R^G_{\mathrm{disc}} := \sum_{\pi \in \mathcal{AR}_{\mathrm{disc}}(G)} \mathrm{tr}_\pi \, .$$

Arthur's classification gives various decompositions of these into smaller parts and identities between the parts.

Now let $G^* \in \mathcal{E}_{\mathrm{ell}}(N)$ and G an extended pure inner form. For each $\psi \in \Psi(G^*)$, [44, §3.1,3.3] defines

$$I^G_\psi, S^G_\psi$$

as summands of I^G_{disc} and S^G_{disc} (In fact, these are defined for more than just elliptic parameters but we won't need the more general definitions).

It turns out (e.g. from the stable multiplicity formula 2.6.3 or an argument like [7, (3.9)]) that for elliptic $\psi \in \Psi(G)$:

$$I^G_\psi := \sum_{\pi \in \Pi^G_\psi} m^\psi_\pi \, \mathrm{tr}_\pi$$

so that

$$R^G_{\mathrm{disc}} = \sum_{\psi \in \Psi(G^*)} I^G_\psi . \tag{2.6.1}$$

Finally, recall the stabilization

$$I^G_{\mathrm{disc}}(f) = \sum_{H \in \mathcal{E}_{\mathrm{ell}}(G)} \iota(G,H) S^H_{\mathrm{disc}}(f^H)$$

for constants $\iota(G,H)$ and endoscopic transfers f^H. If

$$f = \prod_v f_v$$

then we can take

$$f = \prod_v f_v^{H_v}$$

for the $f_v^{H_v}$ defined up to non-canonical scalars that multiply to one and depend on choices of local transfer factors as in [41].

2.6.2. *Results.* The key point we will use is that the stabilization of I^G_{disc} in terms of S^H_{disc} for $H \in \mathcal{E}_{\mathrm{ell}}(G)$ descends to the level of I^G_ψ and S^G_ψ. First:

Proposition 2.6.1 ([44, §1.4])**.** *Let $\psi \in \Psi(G^*)$. Then there is a bijection from $\mathcal{S}_\psi$ to the set of pairs $(H(s), \psi^H(s))$ with $H(s) \in \mathcal{E}_{\mathrm{ell}}(G)$, $\psi^H(s) \in \Psi(H)$ that pushes forward to ψ, up to conjugation by the the subset of* $\mathrm{Out}(\widehat{H})$ *produced by conjugation in $\widehat{G}$. In this bijection $1 \mapsto (G^*, \psi)$.*

The analogous statement also holds for $\psi_v \in \Psi(G_v)$.

Then we have a local formula:

Theorem 2.6.2 (Local Character Relation, [44] theorem 1.6.1 (4))**.** *We have that for any test function f_v on G_v and $s \in \mathcal{S}_{\psi_v}$:*

$$\sum_{\pi \in \Pi_{\psi_v}} \eta^{\psi_v}_{\pi_v}(s' s_{\psi_v}) \operatorname{tr}_\pi(f_v) = \operatorname{tr}_{\psi_v^H(s)}(f_v^{H(s)}).$$

Here, s' is a lift of s to $S^\natural_\psi$ that together with the chosen Whittaker datum on G_v determines the transfer factors for endoscopic transfer $f_v^{H(s)}$ (this is related to the difference we are ignoring between $\mathcal{E}_{\mathrm{ell}}(G)$ and $\overline{\mathcal{E}}_{\mathrm{ell}}(G)$ in [44]*).*

We also have a global formula:

Theorem 2.6.3 (Stable Multiplicity Formula [58, thm. 5.1.2])**.** *Let $\psi \in \Psi(G)$. Then for all test functions f on G*

$$S^{G^*}_\psi(f) = |\mathcal{S}_\psi|^{-1} \epsilon_\psi(s_\psi) \operatorname{tr}^{G^*}_\psi(f).$$

Proof. We can ignore the σ term since we are restricting to elliptic ψ. □

As a useful addendum:

Proposition 2.6.4 (Endoscopic Sign Lemma [58, lem. 5.6.1])**.** *Let $\psi \in \Psi(G^*)$ and let $s \in \mathcal{S}_\psi$ correspond to (H, ψ^H) as in Proposition 2.6.1. Then*

$$\epsilon^H_{\psi^H}(s_{\psi^H}) = \epsilon^G_\psi(s s_\psi).$$

We can use all the above to compute:

Theorem 2.6.5. *For any test function f on G*

$$I^G_\psi(f) = \sum_{(H, \psi_H)} \iota(G, H) S^H_{\psi^H}(f^H)$$

where (H, ψ_H) ranges over $H \in \mathcal{E}_{\mathrm{ell}}(G)$ and $\psi^H \in \Psi(H)$ that pushes forward to ψ (up to equivalence in H). Here, f^H is the endoscopic transfer and $\iota(G, H)$ is the constant that appears in the stabilization of the trace formula.

Furthermore, for any $s \in \mathcal{S}_\psi$

$$S^{H(s)}_{\psi^H(s)}(f^{H(s)}) = 2|\mathcal{S}_\psi|^{-1} \epsilon_\psi(s s_\psi) \sum_{\pi \in \Pi^G_\psi} \eta^\psi_\pi(s s_\psi) \operatorname{tr}_\pi(f).$$

Proof. The first statement follows from the exact definition of S_ψ and I_ψ in [44, §3.3]. The second can be computed from the stable multiplicity formula, the local character relation, the endoscopic sign lemma that $\epsilon^H_{\psi^H}(s_{\psi^H}) = \epsilon^G_\psi(s s_\psi)$, and noting that $|\mathcal{S}_{\psi^H}|^{-1} = 2|\mathcal{S}_\psi|^{-1}$. □

3. AJ-packets and Pseudocoefficients

We recall some more background at the real place. First, we define cohomological representations on G_∞ and the A-packets that contain them. Second, we introduce some specific test functions to eventually plug into the trace formula.

3.1. **Cohomological Representations.**

3.1.1. *Infinitesimal characters.* Let G_∞ be a reductive group over $\mathbb{R}$ and $\mathfrak{t}$ be a Cartan subalgebra of $G_{\infty,\mathbb{C}}$. To irreducible representations of G_∞, one associates an infinitesimal character $\lambda \in \Omega_{G_{\infty,\mathbb{C}}} \backslash \operatorname{Hom}(\mathfrak{t}, \mathbb{C})$. This data is the same as a map from Weyl orbits of $X^*(\widehat{\mathfrak{t}}) = X_*(\mathfrak{t})$ to $\mathbb{C}$, which is further the same as a semisimple conjugacy class in $\widehat{\mathfrak{g}}$.

Definition 3.1.1. We say that the infinitesimal character λ is regular integral if it is that of a finite dimensional representation of $G_{\infty,\mathbb{C}}$.

If φ_∞ is a Langlands parameter for G_∞, then $\varphi|_{W_\mathbb{C}}$ is of the form $z \mapsto (z\bar{z})^\mu (z/\bar{z})^\nu$ for cocharacters $\mu, \nu \in X_*(\widehat{\mathfrak{t}})$ on some Cartan $\widehat{\mathfrak{t}}$ of $\widehat{G}$. All $\pi \in \Pi^G_{\varphi_\infty}$ then have infinitesimal character $\mu + \nu$.

3.1.2. *Cohomological Representations.* Choose a maximal compact K of G_∞. Given a finite dimensional representation V_λ with highest weight λ and a unirrep π of G_∞, we can define the $(\mathfrak{g}, K)$-cohomology groups $H^i(\mathfrak{g}, K; \pi \otimes V_\lambda)$ as in [17].

Definition 3.1.2. A unirrep π of G_∞ is called cohomological of weight λ if there is an i such that $H^i(\mathfrak{g}, K; \pi \otimes V_\lambda) \neq 0$.

Every unirrep that is cohomological of weight λ necessarily has infinitesimal character $\lambda + \rho_G$, which is regular integral by definition. For any real group G_∞ and finite-dimensional representation V_λ, there are only finitely many cohomological representations; an algorithm to construct them and compute their cohomology is given in [78] where they are realized as "cohomologically induced" representations $A_\mathfrak{q}(\lambda)$ attached to θ-stable parabolic subalgebras $\mathfrak{q} \subset \mathfrak{g}$.

Finally, [65] proves that all unirreps with regular, integral infinitesimal character are cohomological. We therefore suggest using "regular, integral infinitesimal character" as a simpler working definition of cohomological.

3.2. **AJ-packets.**

Both our multiplicity computations and subsequent cohomological applications will be phrased in terms of Adams-Johnson parameters, whose definition we now recall following [48]; see also [3, 8].

3.2.1. *AJ Parameters.* Denote the Weil group of $\mathbb{R}$ by $W_\mathbb{R} = W_\mathbb{C} \sqcup jW_\mathbb{C}$.

Definition 3.2.1. Let $\psi_\infty : W_\mathbb{R} \times SL_2(\mathbb{R}) \to {}^L G_\infty$ be an A-parameter, and denote by $\widehat{L}$ the centralizer of $\psi_\infty(W_\mathbb{C})$ in $\widehat{G}$. Then ψ is an *Adams-Johnson* parameter if:

(i) $\psi_\infty(SL_2(\mathbb{C}))$ contains a principal unipotent element of $\widehat{L}$,
(ii) the identity component of $Z(\widehat{L})^{W_\mathbb{R}}$ is contained in $Z(\widehat{G})$,
(iii) the infinitesimal character of the parameter φ_{ψ_∞} is regular.

In condition (ii), the action of $W_\mathbb{R}$ on $Z(\widehat{L})$ is defined as conjugation by $\psi_\infty(W_\mathbb{R}) \subset {}^L G$. Additionally, condition (i) implies that $\psi_\infty(W_\mathbb{C}) \subset Z(\widehat{L})$. It follows from [59, Theorem 5], that any parameter with regular integral infinitesimal character is an Adams-Johnson parameter.

We recall a more explicit description of Adams-Johnson parameters, summarizing the discussion in [4, §8]. Let $(\widehat{T}, \widehat{B}, X_\alpha)$ be a $W_{\mathbb{R}}$-stable pinning of $\widehat{G}$, so that ${}^L G_\infty$ is defined via this pinning. Then $\widehat{L}$ can be conjugated to be the Levi of a parabolic standard with respect to $\widehat{B}$. We additionally make a choice of a Borel pair (T, B) of $G_{\infty,\mathbb{C}}$ such that T is a compact Cartan defined over $\mathbb{R}$. Via these splittings, $\widehat{L}$ is identified with the dual group of a Levi subgroup L of G containing T. One can then construct an embedding $\xi : {}^L L \to {}^L G_\infty$ such that $\psi_\infty = \xi \circ \psi_L$ for ψ_L the A-parameter associated to a unitary one-dimensional representation of L whose differential we will denote by ω; this is done in [8, §5].

Finally, if ψ_∞ is an AJ-parameter, let I_∞ be the set of blocks of the Levi ${}^L L$ constructed in this way. We then have a decomposition $\psi = \bigoplus_{i \in I_\infty} \psi_i$. All of the ψ_i are necessarily conjugate self-dual because they are pairwise distinct by regularity of the infinitesimal character. In particular, the multiplicity of each ψ_i is 1. As such, the computations of [44, §1.2.4] give a canonical isomorphism

$$S^\natural_{\psi_\infty} \simeq (\mathbb{Z}/2)^{I^+_\infty} = (\mathbb{Z}/2)^{I_\infty}. \tag{3.2.1}$$

Furthermore, the localization map $S^\natural_\psi \to S^\natural_{\psi_\infty}$ is the diagonal embedding induced by the surjection $I_\infty \to I$ and the canonical isomorphism (2.4.1).

3.2.2. *AJ Packets.* Adams-Johnson [3] construct packets attached to the above parameters. As described above, an AJ-parameter ψ_∞, gives rise to a pair (L, ω) of a Levi subgroup $L \subset G_\infty$ and ω the differential of a one-dimensional unitary representation of L such that $\omega + \rho_L$ is the infinitesimal character of ψ_∞.

Let $\Omega(G, T)$, $\Omega(L, T)$, and $\Omega_{\mathbb{R}}(G, T)$ be the Weyl groups of $G_{\infty,\mathbb{C}}$, $L_{\mathbb{C}}$, and G_∞ respectively. Then the elements of the packet $\Pi_\psi = \Pi(L, \omega)$ constructed by Adams-Johnson are in bijection with

$$\Sigma_L = \Omega(L, T) \backslash \Omega(G, T) / \Omega_{\mathbb{R}}(G, T).$$

For each $w \in \Sigma_L$, consider the inner form $L_w = w^{-1} L w$; it is the centralizer of an element $x \in i \operatorname{Lie}(T)$, itself giving rise to a θ-stable parabolic subalgebra $\mathfrak{q}_w \subset \mathfrak{g}$. Similarly, let $\omega_w = w^{-1}\omega$. Then Adams-Johnson define

$$\Pi(L, \omega) = \{A_{\mathfrak{q}_w}(\omega_w); w \in \Sigma_L\},$$

where $A_{\mathfrak{q}_w}(\omega_w)$ is the cohomologically induced representation from [78].

We will give more details on Adams-Johnson's packets of cohomological representations for unitary groups in §11.1, including a combinatorial parameterization of the representations in a packet, their relation with cohomology of locally symmetric spaces, and explicit formulas for the characters of S_ψ associated to each parameter.

3.2.3. *Compatibility of Descriptions.* We need to check that Adams' and Johnson's construction of packets matches that in [44]. We thank Jeffrey Adams for explaining this point to us.

By Theorem 4.18 in [5], Adams-Johnson's packets are a special case of the ABV-packets defined in [2]. These ABV-packets have further recently been shown by Arancibia-Mezo [6] to agree with the packets built by Mok [58] in the quasisplit case for any given parameter. Finally, the packets of [44] on non-quasisplit unitary groups are determined by trace identities comparing them to the quasisplit inner form. These trace identities are automatically satisfied by ABV-packets, so we also get that ABV-packets match the packets of [44].

In total, we may use the combinatorial description of Adams and Johnson to understand the structure of AJ-packets on all groups we are considering.

3.3. **Pseudocoefficients and Euler-Poincaré Functions.** We also recall the definitions of certain special test functions. Recall that a *standard module* of G_∞ is the full (possibly reducible) parabolic induction of a discrete series or limit of discrete series representation on a standard Levi M.

If π_d is a discrete series representation of G_∞, the paper [23] constructs *pseudo-coefficients* φ_{π_d} satisfying

$$\mathrm{tr}_\sigma(\varphi_{\pi_d}) = \mathbf{1}_{\sigma=\pi_d}$$

for all standard modules σ. We also define the Euler-Poincaré function

$$\mathrm{EP}_\lambda = \frac{1}{|\Pi_{\mathrm{disc}}(\lambda)|} \sum_{\pi_d \in \Pi_{\mathrm{disc}}(\lambda)} \varphi_{\pi_d},$$

where $\Pi_{\mathrm{disc}}(\lambda)$ is the discrete series L-packet of infinitesimal character λ. Beware that our "endoscopic normalization" of Euler-Poincaré functions is different from the usual one in the literature to work better with endoscopic transfer.

3.4. **Trace Identities.** We also collect some useful computations about traces against AJ-packets:

3.4.1. *Character Formulas.* First, let $\psi_\infty : W_\mathbb{R} \times \mathrm{SL}_2 \to {}^L G_\infty$ be a parameter for an AJ-packet with infinitesimal character λ. We recall some combinatorial results about the character formulas of elements in Π_ψ.

Recall that the *character formula* for a representation π of G_∞ is its expansion in the Grothendieck group as a linear combination of standard modules.

Lemma 3.4.1. *Let π_d be a discrete series representation of G_∞ with infinitesimal character λ. Then there is a unique $\pi \in \Pi_\psi$ such that π_d appears in the character formula for π.*

Proof. This is a consequence of Lemma 8.8 in [3]. □

Next, let $\psi_{\mathrm{disc}} = \psi_{\mathrm{disc}}(\lambda)$ be the discrete parameter with infinitesimal character λ (this is the AJ-parameter with trivial Arthur-SL_2; its associated packet $\Pi_{\mathrm{disc}}(\lambda)$ is a discrete series L-packet). We have an inclusion $S_\psi^\natural \subseteq S_{\psi_{\mathrm{disc}}}^\natural$.

Lemma 3.4.2. *Let $\pi \in \Pi_\psi$ and π_d any discrete series appearing in the character formula for π. Then $\eta_\pi^\psi = \eta_{\pi_d}^{\psi_{\mathrm{disc}}}|_{S_\psi^\natural}$.*

Proof. The values of η_π^ψ on the lifts s' appearing in the endoscopic character identity 2.6.2 are determined by the realization of π as $A(w\lambda)$ and the values $\kappa(w)$ in Theorem 2.21 of [3]. It is therefore determined on all of $S_\psi^\natural$ since we also know that it also restricts to the character χ_{G_∞} of Theorem §2.5.1. However, the character formula in Theorem 8.2 of [3] can be seen to show that π_d corresponds to the same w relative to a given choice of Whittaker datum. It therefore gets assigned the same values $\kappa(w)$.

See also the discussion on page 57 of [76] summarizing parts of [8, §5] for a more explicit computation of these characters in the quasisplit case. By a parenthetical note there, the argument should generalize to non-quasisplit groups through the methods of [42, §5.6]. □

3.4.2. *Identities.* Now we can prove our trace identities:

Lemma 3.4.3. *Let π_d be a discrete series of G_∞ with infinitesimal character λ.*

(1) Let π_0 be a a representation of G_∞ such that π_d appears in its character formula with coefficient ε. Then for all parameters ψ_∞ such that $\pi_0 \in \Pi_{\psi_\infty}$: for all $\pi \in \Pi_{\psi_\infty}$,

$$\mathrm{tr}_{\pi_{\psi_\infty}}(\varphi_{\pi_d}) = \begin{cases} \varepsilon & \pi = \pi_0 \\ 0 & else \end{cases}.$$

(2) Let ψ_∞ be a parameter with infinitesimal character λ. Then for all $s \in S^\natural_{\psi_\infty}$,

$$\sum_{\pi_\infty \in \Pi_{\psi_\infty}} \eta^{\psi_\infty}_{\pi_\infty}(s)\,\mathrm{tr}_{\pi_\infty}(\varphi_{\pi_d}) = \eta^{\psi_{\mathrm{disc}}(\lambda)}_{\pi_d}(s_{\psi_\infty} s)$$

(where we implicitly use that $S^\natural_{\psi_\infty} \subseteq S^\natural_{\psi_{\mathrm{disc}}(\lambda)}$).

Proof. For (1), ψ_∞ then has infinitesimal character λ and is therefore necessarily an AJ-parameter. Therefore π_0 is then the unique $\pi \in \Pi_{\psi_\infty}$ such that π_d appears in the character formula for π by 3.4.1.

For (2), choose π_0 to be the unique $\pi \in \Pi_{\psi_\infty}$ with π_d in its character formula. Let it appear with coefficient ε. By (1) the sum of traces is $\varepsilon\eta^{\psi_\infty}_{\pi_0}(s)$.

To compute ε, we use that by [3, (8.10)] all discrete series appear in the character formula of the stable sum

$$\sum_{\pi \in \Pi_{\psi_\infty}} \eta^{\psi_\infty}_{\pi}(s_{\psi_\infty})\pi \tag{3.4.1}$$

with multiplicity 1. Therefore, $\varepsilon = \eta^{\psi}_{\pi_0}(s_{\psi_\infty})$. Lemma 3.4.2 finishes the proof. □

As an important special case:

Corollary 3.4.4. *Let π_d be a discrete series of G_∞ with infinitesimal character λ and let ψ_∞ be a parameter with infinitesimal character λ. Then*

$$\mathrm{tr}_{\psi_\infty}(\mathrm{EP}_\lambda) = \mathrm{tr}_{\psi_\infty}(\varphi_{\pi_d}) = 1.$$

4. Refined Shapes

We now come to the key definition of this paper: the invariant of a *refined shape* Δ attached to an an Arthur parameter ψ. . We construct the invariant to satisfy:

- Δ determines the $H \in \mathcal{E}_{\mathrm{ell}}(N)$ attached to ψ.
- Δ determines the local factor ψ_∞ (and in particular the Arthur-SL_2), as well as the pair $(\mathcal{S}_\psi, s_\psi)$.
- *Cuspidal* parameters ψ have refined shapes Δ such that there is a well-understood geometric-side expression for the trace against the Δ-part of the automorphic spectrum of H.
- The stable multiplicity formula for the ψ-part S_ψ of S_{disc} can be understood in terms of the S_{ψ_i} for ψ_i the cuspidal parameters that build up ψ in enough cases. This understanding should be through algorithms that are uniform over all ψ with refined shape Δ. In particular, the character ϵ_ψ and the local stable traces can be understood inductively this way.

These properties are in turn exactly the ones needed to run the inductive argument of §5.

4.1. **Infinitesimal and Central Characters.** We need some preliminary details on infinitesimal and central characters:

4.1.1. *Infinitesimal characters in the classification.* Let $G \in \mathcal{E}_{\mathrm{sim}}(N)$. Then $G_\infty = U_N^*(F_\infty)$ where the U^* represents the quasisplit inner form. The Lie algebra $\widehat{\mathfrak{g}}_\infty = \mathfrak{gl}_n(F_\infty \otimes_{\mathbb{R}} \mathbb{C})$, so consider an infinitesimal character of G as a semisimple matrix up to conjugacy, or in other words, an unordered sequence

$$\xi = (\xi_1, \xi_2, \dots, \xi_n)$$

with $\xi_j \in F_\infty \otimes_{\mathbb{R}} \mathbb{C}$. We can further expand out each ξ_j as a list of complex numbers $\xi_{j,v}$ for each place $v \in \infty$ of F (which is necessarily real).

It is also sometimes useful to think of each local component ξ_v as the generating function $\sum_j X^{\xi_{j,v}}$. In this way, if each $\tau_{i,v}$ has infinitesimal character $\xi_v^{(i)}$, we have infinitesimal character assignment

$$\left(\bigoplus_i \tau_i[d_i]\right)_v \mapsto \sum_i \xi_v^{(i)} \sum_{l=1}^{d_i} X^{\frac{d+1}{2}-l} \tag{4.1.1}$$

It can be seen from this that the character of $\tau[d]$ determines that of τ since the group ring $\mathbb{Z}[\mathbb{C}]$ is an integral domain.

Finally, recall from §3.1.1 that we call ξ regular integral if it is the infinitesimal character of a finite-dimensional representation. Equivalently, for each $v \in \infty$ the $\xi_{i,v}$ are distinct values that are all integers if N is odd or all half-integers if N is even.

4.2. **Definitions.** Our notion of refined shape is built off of the details of how Mok assigns a parameter to an element of $\mathcal{E}_{\mathrm{ell}}(N)$. First, recall the assigned parity δ_i of each simple parameter τ_i in §2.2.4:

Lemma 4.2.1. *Let $\tau \in \Psi(N)$ be a cuspidal parameter with regular integral infinitesimal character ξ at infinity and of parity $\delta = \pm 1$. Then there exists a unique $H = H(\delta) \in \mathcal{E}_{\mathrm{sim}}(N)$ depending only on δ such that τ is a parameter for $H(\delta)$.*

Furthermore, ξ is the infinitesimal character of a finite dimensional representation of H.

Proof. The first claim is a rephrasing of §2.2.4. The second is by definition. □

Motivated by the above, we define:

Definition 4.2.2. A *refined shape* is a sequence

$$\Delta = (T_i, d_i, \xi_i, \eta_i)_i$$

up to permutation and where the T_i and d_i are positive integers, ξ_i is an infinitesimal character of rank T_i, and $\eta_i = \pm 1$.

We say that $\psi \in \Delta$, or that ψ has refined global shape Δ, if ψ is elliptic and $\psi = \bigoplus_i \tau_i[d_i]$ with each τ_i of rank T_i and such that each τ_i has infinitesimal character ξ_i at infinity and is of parity $\delta_i = \eta_i$.

We let $\Psi(\Delta) \subseteq \Psi(N)$ be the set of all elliptic, self-dual parameters on $G(N)$ of refined shape Δ.

Note. Unitary groups G may be realized as twisted endoscopic groups of $\widetilde{G}(N)$ in two different ways. The η_i data is a technical necessity to select which of these realizations the shape corresponds to—it is not very important conceptually. To define analogous refined shapes for orthogonal and symplectic groups, the η_i would record the central characters of the τ_i instead of their parity.

Definition 4.2.3. Let ξ be an infinitesimal character of rank N and $\eta = \pm 1$. Then $\Sigma_{\xi,\eta}$ is the refined shape $(N, 1, \xi, \eta)$.

Definition 4.2.4. The refined shape Δ is *integral* if its *total* infinitesimal character as determined by formula (4.1.1) is regular integral.

In particular, if Δ is integral, then each of the ξ_i must be regular integral, though this isn't sufficient.

4.3. **Properties.** We list the key properties of refined shapes:

Proposition 4.3.1. *Let the refined shape Δ have rank N such that all the ξ_i are integral. Then there is a group $H(\Delta) \in \mathcal{E}_{\mathrm{ell}}(N)$ such that all $\psi \in \Delta$ are parameters for $H(\Delta)$.*

Proof. The assignment described in §2.2.4 only depends on T_i, d_i, and η_i. □

This gives us two quick corollaries:

Corollary 4.3.2. *Let $G \in \mathcal{E}_{\mathrm{ell}}(N)$. Then*

$$\Psi(G) = \bigsqcup_{\Delta : H(\Delta) = G} \Psi(\Delta).$$

Proof. This follows from the above since every parameter has a refined shape. □

Corollary 4.3.3. *Let ξ be an integral infinitesimal character of rank N. Then*

$$H(\Sigma_{\xi,\eta}) = U_{\eta(-1)^{N-1}}(N).$$

In particular, for every $G \in \mathcal{E}_{\mathrm{sim}}(N)$ and infinitesimal character ξ of a finite dimensional representation on G, there is η such that $H(\Sigma_{\xi,\eta}) = G$.

As three more facts, if Δ is a refined shape:

- All $\psi \in \Delta$ correspond to the same pairs $(S^\natural_\psi, s_\psi)$ by formula (2.4.1). Call the common values $S^\natural_\Delta$ and s_Δ. We can similarly define a common $\mathcal{S}_\Delta$.
- All $\psi \in \Delta$ have the same Arthur-SL_2, so the Arthur-SL_2 of Δ is well-defined.
- The infinitesimal character at infinity of $\psi \in \Delta$ is determined by (4.1.1).

Finally:

Lemma 4.3.4. *Let Δ be an integral refined shape. There exists an AJ-parameter ψ^Δ_∞ such that for all $\psi \in \Delta$, $\psi_\infty = \psi^\Delta_\infty$. Furthermore, the induced localization map $S^\natural_\Delta \to S^\natural_{\psi^\Delta_\infty}$ is also determined by Δ.*

Proof. The shape Δ determines the localization ψ_∞ of any parameter $\psi \in \Delta$: for $\psi_\infty|_{SL_2} = \oplus_i \nu(d_i)^{T_i}$, this is immediate. For the restriction $\psi_\infty|_{W_\mathbb{R}}$, let $\psi = \tau_i[d_i]$ such that each term has infinitesimal character ξ_i. Then by a known case of the Ramanujan conjecture as explained in [52, Lem. 6.1], the parameter associated to $\tau_{i,\infty}$ is bounded with infinitesimal character ξ_i matching that of a finite dimensional representation, so it is uniquely determined. The parameter ψ_∞ is determined from this data following the constructions of §2.3.2.

By construction, the resulting ψ_∞ has a regular integral infinitesimal character. Additionally, it maps the Arthur-SL_2 to a principal SL_2 of the Levi $\widehat{L} = \prod_i GL_{d_i}^{T_i} \subset$

$\hat{G}$, with $\widehat{L} = Z_{\widehat{G}}(\psi_\infty(W_\mathbb{C}))$ following the assumption that the total infinitesimal character is regular. Then $\psi_\infty(W_\mathbb{C}) \subset Z(\widehat{L})$, and since ψ_∞ was built from parameters of discrete series, $\psi_\infty|_{W_\mathbb{C}}$ factors through $z \mapsto \frac{z}{\bar{z}}$ and $W_\mathbb{R}$ acts on $Z(\widehat{L})$ by inversion. Thus the identity component of $Z(\widehat{L})^{W_\mathbb{R}}$ is trivial, and ψ_∞ is an AJ-parameter following Definition 3.2.1.

For localization statement, the map $I_\infty^+ \to I^+$ described after formula (3.2.1) can be seen to depend only on Δ. This determines the localization map. □

5. The Trace Formula with Fixed Shape

Let $G \in \mathcal{E}_{\mathrm{ell}}(N)$ (the case of $G \in \mathcal{E}_{\mathrm{sim}}(N)$ suffices for us). If Δ is a refined shape such that $H(\Delta) = G$, define

$$S^G_\Delta := \sum_{\psi \in \Delta} S^G_\psi.$$

This S^G_Δ is the main technical building block in our applications, and understanding it is the key step in our argument.

For an infinitesimal character λ, let EP_λ be the Euler-Poincaré function introduced in §3.3. The overarching goal for this section is:

Goal. *Let λ be the infinitesimal character of a finite dimensional representation of G. Understand $S^G_\Delta(\mathrm{EP}_\lambda f^\infty)$ as a linear combination of terms $I^H_{\mathrm{disc}}(\mathrm{EP}_{\lambda'}(f^\infty)')$ on other groups H.*

The terms $I^H_{\mathrm{disc}}(\mathrm{EP}_{\lambda'}(f^\infty)')$ are well-understood—they were given an explicit formula in [7], which was in turn studied in great detail and bounded in [73]. Achieving the goal would therefore give fine control over S_Δ.

Taïbi in [76] gave such a description in the level-1 case. This work is basically extending his method as much as possible to deeper levels.

5.1. Overall Strategy.

We build up $S^G_\Delta(\mathrm{EP}_\lambda f^\infty)$ from $I^H_{\mathrm{disc}}(\mathrm{EP}_{\lambda'}(f^\infty)')$ terms in stages:

(1) Switch from I to S to allow for various transfers in the endoscopic classification using that $S^H_{\mathrm{disc}}(\mathrm{EP}_\lambda f^\infty)$ can be expanded as a linear combination of $I^{H'}_{\mathrm{disc}}(\mathrm{EP}_{\lambda'}(f^\infty)')$ terms through "hyperendoscopy" as in [31].

(2) Define a sum of traces on each $\widetilde{G}(M)$ called $S^M(\lambda, \eta, f^\infty)$ satisfying that if $H = H(\Sigma_{\lambda,\eta})$, $S^M(\lambda, \eta, f^\infty) = S^H_{\mathrm{disc}}(\mathrm{EP}_\lambda(f^\infty)^H)$. The use of traces on $\widetilde{G}(M)$ allows inducting by decomposing Δ down to its simple constituents.

(3) Define sub-sums $S^N_\Delta(f^\infty)$ of $S^N(\lambda, \eta, f^\infty)$ corresponding to individual refined shapes. All the $S^N_\Delta(f^\infty)$ can be computed from terms $S^M(\lambda, \eta, (f^\infty)')$ by an induction in two steps:

   - $$S^N_{\Sigma_{\lambda,\eta}}(f^\infty) = S^N(\lambda, \eta, f^\infty) - \sum_{\substack{H(\Delta) = H(\Sigma_{\lambda,\eta}) \\ \text{Inf. Char}(\Delta) = \lambda \\ \Delta \neq \Sigma_{\lambda,\eta}}} S^N_\Delta(f^\infty).$$

   When $N = 1$, the sum of "correction terms" vanishes—this is the base case of our induction.

- If $\Delta = (T_i, d_i, \lambda_i, \eta_i)_i$ with infinitesimal character λ, express $S^N_\Delta(f^\infty)$ in terms of the $S^{T_i}_{\Sigma_{\lambda_i,\eta_i}}$. This is the hardest part of the argument and the partial results needed for our applications will be postponed to Sections 7 and 8.

(4) Transfer back to the classical group—$S^G_\Delta(\mathrm{EP}_\lambda f^\infty)$ can be written as $S^{N(G)}_\Delta(f^\infty_1)$ as long as we can find $(f^\infty_1)^G = f^\infty$.

In our actual argument, steps 2 and 3 will not be so clearly separated. We will use step 2's Proposition 5.3.1 to switch freely between the perspective of traces on $G \in \mathcal{E}_{\mathrm{ell}}(N)$ and traces on $\widetilde{G}(N)$ as convenient to perform the induction of step 3.

Since there is a lot of notation introduces this section, we summarize all the parts of the trace formula defined:

- I^G_Δ, I^G_ψ are the pieces of the spectral decomposition of Arthur's I^G_{disc} on some $G \in \mathcal{E}_{\mathrm{ell}}(N)$ corresponding to a refined shape Δ or individual parameter ψ.
- S^G_Δ, S^G_ψ are the same for Arthur's S_{disc}.

5.2. **Step 1: Understanding S^H.** The S^H terms can be understood through the hyperendoscopy formula of [31]. We use notation from [26], although the full generality there isn't necessary since computation of the endoscopy of classical groups in [79] show that we will never have to take a z-extension.

Theorem 5.2.1 (Ferrari's Hyperendoscopy Formula). *Let $\mathcal{HE}_{\mathrm{ell}}(H)$ be the set of non-trivial elliptic hyperendoscopic paths of H as in* [26, §4]. *Then,*

$$S^H(\mathrm{EP}_\lambda f^\infty) = I^H(\mathrm{EP}_\lambda f^\infty) + \sum_{\mathcal{H} \in \mathcal{HE}_{\mathrm{ell}}(H)} \iota(G, \mathcal{H}) I^{\mathcal{H}}(\mathrm{EP}^{\mathcal{H}}_\lambda (f^\infty)^{\mathcal{H}})$$

for constants $\iota(G,\mathcal{H})$ and transfers $\star^{\mathcal{H}}$ defined there.

Proof. See [26, Thm. 4.2.3], and note that using EP_λ at infinity lets us elide the distinction between trace formulas and their discrete parts. □

Ferrari gives an explicit formula to write the transfers EP^H_λ as linear combinations of Euler-Poincaré functions. See [26, §5.1] for an English-language presentation though beware that there is a ρ-shift between the parameterization EP_λ used here and the parameterization η_λ used there.

5.3. **Step 2: Understanding S^M.** Fix an infinitesimal character λ of rank M. Using Proposition 4.3.1, define:

$$\Psi(\lambda, \eta) = \bigcup_{\substack{\Delta : H(\Delta) = H(\Sigma_{\lambda,\eta}) \\ \text{Inf. Char}(\Delta) = \lambda}} \Psi(\Delta)$$

and define pieces of the spectral expansion:

$$\begin{aligned} S^M(\lambda, \eta, f^\infty) &:= \sum_{\Delta : H(\Delta) = H(\Sigma_{\lambda,\eta})} S^M_\Delta(f^\infty) \\ &:= \sum_{\psi \in \Psi(\lambda,\eta)} S^M_\psi(f^\infty) \\ &:= \sum_{\psi \in \Psi(\lambda,\eta)} \epsilon^H_\psi(s^H_\psi) m_\psi |S_\psi|^{-1} \operatorname{tr}_{\widetilde{\pi}^\infty_\psi}(f^\infty). \end{aligned} \tag{5.3.1}$$

Here, π_ψ is the automorphic representation of $\mathrm{GL}_M(\mathbb{A}_E)$ corresponding to ψ as in §2.2.2 and $\widetilde{\pi}_\psi$ is its extension to $\widetilde{G}(M)$ as in §2.2.3.

Proposition 5.3.1. *Let $H = H(\Sigma_{\lambda,\eta})$ which is necessarily in $\mathcal{E}_{\mathrm{sim}}(M)$. Then for a test function f^∞ at the finite places,*

$$S^M(\lambda, \eta, f^\infty) = S^H(\mathrm{EP}_\lambda(f^\infty)^H).$$

Proof. Fix $\psi \in \Psi(H)$ with infinitesimal character λ which is necessarily elliptic. Then the stable multiplicity formula 2.6.3 shows that

$$S^H_\psi(\mathrm{EP}_\lambda(f^\infty)^H) = \epsilon_\psi(s_\psi)|S_\psi|^{-1}\left(\sum_{\pi_\infty \in \Pi_{\psi_\infty}} \eta^{\psi_\infty}_{\pi_\infty}(s_\psi)\,\mathrm{tr}_{\pi_\infty}(\mathrm{EP}_\lambda)\right) \left(\sum_{\pi \in \Pi_{\psi^\infty}} \eta^{\psi^\infty}_{\pi^\infty}(s_\psi)\,\mathrm{tr}_{\pi^\infty}((f^\infty)^H)\right).$$

After recalling that S_ψ is a 2-group, Lemma 3.4.3 shows that for each pseudocoefficient φ_{π_d} with infinitesimal character λ:

$$\sum_{\pi_\infty \in \Pi_{\psi_\infty}} \eta^{\psi_\infty}_{\pi_\infty}(s_\psi)\,\mathrm{tr}_{\pi_\infty}(\varphi_{\pi_d}) = 1.$$

Averaging over pseudocoefficients:

$$\sum_{\pi_\infty \in \Pi_{\psi_\infty}} \eta^{\psi_\infty}_{\pi_\infty}(s_\psi)\,\mathrm{tr}_{\pi_\infty}(\mathrm{EP}_\lambda) = 1.$$

Furthermore, multiplying together the local character relation 2.6.2 over all finite places shows that:

$$\sum_{\pi \in \Pi_{\psi^\infty}} \eta^{\psi^\infty}_{\pi^\infty}(s_\psi)\,\mathrm{tr}_{\pi^\infty}((f^\infty)^H) = \mathrm{tr}_{\widetilde{\pi}^\infty_\psi}(f^\infty).$$

In total $S^H_\psi(\mathrm{EP}_\lambda(f^\infty)^H) = S^M_\psi(\lambda, \eta, f^\infty)$.

Since $S_\psi(\mathrm{EP}_\lambda(f^\infty)^H) = 0$ for all ψ with infinitesimal character not equal to λ, summing over $\Psi(\lambda, \eta)$ and using Corollary 4.3.2 gives that

$$S^M(\lambda, \eta, f^\infty) = \sum_{\psi \in \Psi(H)} S^H_\psi(\mathrm{EP}_\lambda(f^\infty)^H).$$

By [7, (3.9)] and equation (2.6.1), we know that

$$I^H(\mathrm{EP}_\lambda(f^\infty)^H) = R^H(\mathrm{EP}_\lambda(f^\infty)^H) = \sum_{\psi \in \Psi(H)} I^H_\psi(\mathrm{EP}_\lambda(f^\infty)^H).$$

By a hyperendoscopy argument using the expansion in Theorem 2.6.5, the same sum expansion holds for S^H. Therefore, we can conclude that

$$S^M(\lambda, \eta, f^\infty) = S^H(\mathrm{EP}_\lambda(f^\infty)^H).$$

This finishes the argument. $\square$

5.4. **Step 3: The Induction.** We give a heuristic overview to keep in mind for understanding step 3. All precise results will be postponed to §7,8.

Recall the decomposition

$$S^N(\lambda, \eta, f^\infty) = \sum_{\Delta : H(\Delta) = H(\Sigma_{\lambda,\eta})} S^N_\Delta(f^\infty)$$

which we were using for the induction in step 3. We now want to understand S_Δ in terms of smaller groups.

For the sake of heuristic understanding, we will instead consider the simpler

$$S^{|H(\Delta)|}_\Delta(\eta_\lambda (f^\infty)^{H(\Delta)}) = S^{|N|}_\Delta(f^\infty) := \sum_{\psi \in \Delta} m_\psi |S_\psi|^{-1} \operatorname{tr}_{\widetilde{\pi}^\infty_\psi}(f^\infty)$$

without the ϵ-sign. Consider $\psi = \tau_1[d_1] \oplus \cdots \oplus \tau_k[d_k] \in \Delta$. Motivated by the way $\widetilde{\pi}_\psi$ is defined through parabolically inducing determinant twists of the τ_i, assume we could define a "generalized constant term" map

$$f^\infty \mapsto (f^\infty_i)_{\Delta,i}$$

such that for all $\psi \in \Delta$

$$\operatorname{tr}_{\widetilde{\pi}^\infty_\psi}(f^\infty) = \prod_i \operatorname{tr}_{\widetilde{\pi}^\infty_{\tau_i}}(f^\infty_{\Delta,i}).$$

Because the infinitesimal character of Δ is regular and disallows repeated τ_i factors, the possible (elliptic) $\psi \in \Delta$ are exactly the $\tau_1[d_1] \oplus \cdots \oplus \tau_k[d_k]$ for all choices of $\tau_i \in \Sigma^{T_i}_{\lambda_i,\eta_i}$. Therefore we get a heuristic factorization

$$S^{|N|}_\Delta(f^\infty) = C_\Delta \prod_i S^{|T_i|}_{\Sigma_{\lambda_i,\eta_i}}(f^\infty_{\Delta,i}), \tag{5.4.1}$$

for C_Δ a constant depending on the various S_ψ, S_{τ_i}, and m_ψ's that only depend on Δ.

Obviously, we cannot simply ignore the ϵ-sign and we do not have a actual definition of this generalized constant term. In fact, the definition of this generalized constant term would allow us to define the long-desired "stable transfer" between G and its endoscopic groups, so it is likely very difficult.

However, §§7,8 will discuss enough partial results that an application to limit multiplicities at specifically split level can be completed.

5.5. **Step 4: Understanding S^G_Δ.** This step comes from a corollary to the arguments in step 2:

Corollary 5.5.1. *Fix a refined shape Δ of rank N and let $G = H(\Delta)$ as in Proposition 4.3.1. Then for any test function f^∞ on $G(\mathbb{A}^\infty)$, there is f^∞_1 on $\widetilde{G}(N)^\infty$ such that $(f^\infty_1)^G = f^\infty$. Furthermore,*

$$S^G_\Delta(\eta_\lambda f^\infty) = S^N_\Delta(f^\infty_1).$$

Proof. The existence of f^∞_1 comes from [58, Prop. 3.1.1(b)]. Then, arguing as in 5.3.1 gives $S^G_\psi(\eta_\lambda (f^\infty_1)^G) = S^N_\psi(f^\infty_1)$ for any $\psi \in \Delta$. Summing over all $\psi \in \Delta$ produces the result. □

6. Local Transfer

The argument of §5.4 requires constructing certain transfers of functions. We will not be able to do this in full generality—this section focuses on either constructing various special cases of these or approximate versions satisfying good enough bounds. The last Subsection 6.4 will state the full desired conjectures.

We will heavily use the shorthand from Section 1.6.2 throughout.If G_v is an unramified group over F_v then $K_v := K_v^G$ will be a hyperspecial subgroup for G_v.

6.1. **Split Places.**

We first discuss some results that only hold at split places. Extending Lemmas 6.1.2 and 6.1.3 to non-split places would similarly extend Theorems 11.4.1 and 11.4.2.

Fix $G \in \mathcal{E}_{\mathrm{ell}}(N)$. Let v be a finite place of F which splits in E, so that $E \otimes_F F_v \simeq F_v \times F_v$ with σ permuting the two copies of F_v. Then $\widetilde{G}(N, F_v) \simeq GL_N(F_v) \times GL_N(F_v)$ and the action of θ_N on $\widetilde{G}(N)_v$ becomes

$$(g_1, g_2) \mapsto (\mathrm{Ad}(J_N)g_2^{-t}, \mathrm{Ad}(J_N)g_1^{-t}),$$

where $G(N)_v \simeq GL_N(F_v)$ and is embedded in $\widetilde{G}(N)_v$ as the fixed points of θ. A possible embedding is $g \mapsto (g, \mathrm{Ad}(J_N)g^{-t})$.

For the reader's convenience, we now summarize some results that show Arthur's classification for $G(N)_v$ as a element of $\mathcal{E}_{\mathrm{ell}}(N)$ matches Arthur's classification coming from the isomorphism $G(N)_v \simeq \mathrm{GL}_N(F_v)$:

Lemma 6.1.1. *Let v be such that the global extension E/F splits at v and denote $E \otimes_F F_v$ by E_v. Let π_v be the irreducible conjugate self-dual representation of $G(N)(E_v) \simeq GL_N(F_v) \times GL_N(F_v)$ coming from the Arthur parameter ψ_v. Then:*

1. *(1) π_v is of the form $\pi_v^0 \otimes (\pi_v^0)^\vee$ for an irrep π_v^0 of $\mathrm{GL}_N(F_v)$,*
2. *(2) The canonical extension of π_v to $\widetilde{G}(N)$ as in* [11, §2.2] *has θ acting on $\pi_v^0 \otimes (\pi_v^0)^\vee$ through $x \otimes y \mapsto y \otimes x$,*
3. *(3) For $\tilde{f}_v = (f_v^1, f_v^2) \in \mathcal{H}(G(N)(E_v) \rtimes \theta)$, we can choose transfer $\tilde{f}_v^G = f_v^1 \star {}^\theta f_v^2$ where we define ${}^\theta f_v(g) = f_v(\mathrm{Ad}(J_n)g^{-t})$.*
4. *(4) For $\tilde{f}_v = (f_v^1, f_v^2) \in \mathcal{H}(G(N)(E_v) \rtimes \theta)$, $\mathrm{tr}_{\tilde{\pi}_v}(\tilde{f}_v) = \mathrm{tr}_{\pi_v^0}(f_v^1 \star {}^\theta f_v^2)$*
5. *(5) $\mathrm{tr}_{\psi_v}(f_v) = \mathrm{tr}_{\pi_v^0}(f_v)$.*

Proof. (1) follows from the description of irreps of $G_1 \times G_2$ and self-duality.

For (2), first assume π_v and therefore π_v^0 is tempered. A Whittaker functional on π_v is a product of a pair of functionals on π_v^0 and $(\pi_v^0)^\vee$. This product is preserved by the claimed θ since the space of Whittaker functionals is one dimensional. On the other hand, if π isn't tempered, the statement can be checked by the construction of θ through parabolic induction. This gives the second statement in all cases.

For (3), this is a special case of corollary 1.1.6 in [44].

For (4), $\tilde{f}_v(x \otimes y) = (f_v^1 x \otimes {}^\theta f_v^2 y)$. By admissibility of π_v and smoothness of the f_v^i, we can compute the trace by choosing a basis for the finite-dimensional vector space $V = (\pi_v)^U$ for some open compact U. The result follows from computing the action on the standard induced bases for $\mathrm{Sym}^2 V \oplus \wedge^2 V = V \otimes V$.

Finally, (5) follows from (3), (4), and the endoscopic character relation after choosing f_v^1 and ${}^\theta f_v^2$ such that $f_v^1 \star {}^\theta f_v^2 = f_v$. □

We derive two consquences of Lemma 6.1.1:

Lemma 6.1.2. *Let v be a place that splits in E and $\psi_v = \psi_{1,v} \oplus \psi_{2,v}$ be an Arthur parameter for $U_{E/F}(N)(F_v) \simeq \mathrm{GL}_N(F_v)$. Then ψ factors through (the L-dual of) a Levi subgroup $M = \mathrm{GL}_{N_1}(F_v) \times \mathrm{GL}_{N_2}(F_v)$ and*

$$\mathrm{tr}_{\psi_v} f = \prod^{\oplus}_{i=1,2} \mathrm{tr}_{\psi_{i,v}} f_{M,i},$$

where $f_{M,1}$ and $f_{M,2}$ are the factors of the constant term map to M.

Proof. Then $\tilde{\pi}_{\psi_v}$ is a parabolic induction of $\tilde{\pi}_{\psi_{1,v}} \otimes \tilde{\pi}_{\psi_{2,v}}$, so in the notation of Lemma 6.1.1 (1), $\pi^0_{\psi_v}$ is a parabolic induction of $\pi^0_{\psi_{1,v}} \otimes \pi^0_{\psi_{2,v}}$. The result follows from 6.1.1 (5). $\square$

Lemma 6.1.3. *Let v be a place that splits in E and $\psi_v = \psi_{i,v}[d]$ be an Arthur parameter for $U_{E/F}(N)(F_v) \simeq \mathrm{GL}_N(F_v)$. Then $\psi|_{W_F}$ factors through (the L-dual of) a Levi subgroup $M = (\mathrm{GL}_{N_1})^d$. Furthermore, for all test functions f satisfying:*

- *f is supported on the kernel of $|\det|_v$,*
- *for all unirreps π_v of $\mathrm{GL}_N(F_v)$, $tr_{\pi_v}(f) \geq 0$,*

we have

$$\mathrm{tr}_{\psi_v} f \leq (\mathrm{tr}_{\psi_{1,v}} f_{M,1})^{d\oplus},$$

where $f_{M,1}$ is the constant term to M restricted to the first factor.

Proof. This is a similar argument to 6.1.2 using that $\tilde{\pi}_{\psi_v}$ is a summand in the Grothendieck group of the parabolic induction of

$$\tilde{\pi}_{\psi_{1,v}} |\det|^{\frac{d-1}{2}} \otimes \cdots \otimes \tilde{\pi}_{\psi_{2,v}} |\det|^{\frac{1-d}{2}}.$$

We may ignore the determinant factors by the support condition. The inequality comes from the positivity condition applied to traces against the other summands in the Grothendieck group expansion. $\square$

6.2. **Transfers at Unramified Places.** We next discuss the unramified places. This is slightly trickier than split places, but unramified transfer through any L-map is very explicit through the use of Satake parameters. This gives us very good control.

6.2.1. *Basic Transfer Result.* First we recall an "Arthur packet fundamental lemma" that was the key result making the strategy of §5 work at the level-1 in [76]:

Lemma 6.2.1 (Fundamental Lemma for A-packets)**.** *Let v be a place unramified in E and*

$$\psi_v = \bigoplus_i \tau_{i,v}[d_i]$$

be an Arthur parameter for $U_{E/F}(N)(F_v)$. Then

$$\mathrm{tr}_{\psi_v} \mathbf{1}_{K_v} = \prod_i \mathrm{tr}_{\psi_{i,v}} \mathbf{1}_{K_{i,v}}$$

for appropriately chosen hyperspecial subgroups $K_{\star,v}$ in appropriate $U_{E/F}(N_\star)(F_v)$.

Proof. This follows from [76, Lem. 4.1.1] since ψ_v is unramified if and only if each of the $\tau_{i,v}$ is. Then, just apply that for any π_v, $\mathrm{tr}_{\pi_v} \bar{\mathbf{1}}_{K_{i,v}} = \mathbf{1}_{\pi_v \text{ is unram.}}$.

The $K_{\star,v}$ are chosen as in the fundamental lemma according to a choice of Whittaker datum. $\square$

We will eventually attempt to prove bounds in the style of [73], so we need a more general statement for any element of $\mathscr{H}^{\mathrm{ur}}(G_v)$.

6.2.2. *Truncated Hecke Algebras.* We first recall the notion of a bounded Hecke algebra from [73]. The elements $\tau_\lambda^G = \mathbf{1}_{K_v \lambda(\varpi) K_v}$ for a chosen unformizer ϖ and $\lambda \in X_*(A)^+$ generate $\mathscr{H}^{\mathrm{ur}}(G_v)$. Pick a basis $\mathcal{B}$ for $X_*(A)$ and define the norm

$$\|\lambda\|_{\mathcal{B}} = \max_{\omega \in \Omega}(\text{biggest } \mathcal{B}\text{-coordinate of } \omega\lambda).$$

For $\lambda \in X_*(A)$, define truncated Hecke algebra

$$\mathscr{H}(G,K)^{\leq \kappa, \mathcal{B}} = \langle \tau_\lambda^G : \|\lambda\|_{\mathcal{B}} \leq \kappa \rangle.$$

It turns out (see [73, §2]) that for any two $\mathcal{B}, \mathcal{B}'$, $\|\lambda\|_{\mathcal{B}}$ and $\|\lambda\|_{\mathcal{B}'}$ are proportional. All the bounds we use will depend on κ only up to an unspecified constant. Therefore we can suppress the $\mathcal{B}$.

6.2.3. *Basis of Characters.* Recall that the Satake transform gives an isomorphism

$$\mathscr{H}^{\mathrm{ur}}(G_v) \xrightarrow{\sim} \mathbb{C}[X_*(A)]^{\Omega_F},$$

where A is a maximally split maximal torus of G_v in good position with respect to K_v. The right side of this isomorphism has a basis χ_λ of trace characters of finite dimensional representations λ of the twisted group $\widehat{G} \rtimes \mathrm{Frob}_v$.

Any unramified parameter ψ_v determines an unramified L-parameter which determines a Satake parameter: a semisimple conjugacy class $\sigma_{\psi_v} \in \widehat{G} \rtimes \mathrm{Frob}_v$. Because of [76, Lem. 4.1.1], this satisfies that

$$\mathrm{tr}_{\psi_v} \chi_\lambda = \mathrm{tr}_\lambda(\sigma_{\psi_v}). \tag{6.2.1}$$

See [73, §2.2] for more detail.

The consistency of unramified packets constructed by Arthur/Mok and the Satake isomorphism is implicit in the isolation of the ψ-part I_ψ of I_{disc}—see §3.3 in [11] for example. It depends on the full fundamental lemma for all spherical functions.

6.2.4. *General Unramified Transfer.* Let v be a place unramified in E/F and

$$\psi_v = \bigoplus_i \tau_{i,v}[d_i] \in \Delta = (t_i, d_i, \lambda_i, \eta_i)_i$$

be an Arthur parameter for $U_{E/F}(N)(F_v)$. Let each $\tau_{i,v}$ be a parameter for U_i. Then there is an associated embedding

$$\iota_\Delta : \mathbf{H} := {}^L H_v \times \prod_i {}^L\mathrm{GL}_{d_i} := \prod_i ({}^L U(n_i) \times {}^L\mathrm{GL}_{d_i}) \hookrightarrow {}^L G_v. \tag{6.2.2}$$

Let ψ_n^I be the parameter of the trivial representation on GL_n. Then we can write the Langlands parameter φ_{ψ_v} corresponding to ψ_v as the pushforward of

$$\prod_i \tau_{i,v} \times \psi_{d_i}^I.$$

This gives the map on Satake parameters

$$\sigma_{\psi_v} = \mathcal{S}_\Delta((\sigma_{\tau_{i,v}})_i) := \iota_\Delta \left(\prod_i \sigma_{\tau_{i,v}} \times \sigma_{d_i}^I \right), \tag{6.2.3}$$

where $s_{d_i}^I$ is the Satake parameter of the d_i-dimensional trivial representation.

Restricting the embedding to $\star \rtimes \mathrm{Frob}$ cosets determines an unramified transfer map:

$$\mathcal{T}_\Delta : \mathscr{H}^{\mathrm{ur}}(G_v) \to \mathscr{H}^{\mathrm{ur}}(H_v) : \chi_\lambda \mapsto \chi_\lambda \circ \mathcal{S}_\Delta.$$

Equation (6.2.1) then gives:

Lemma 6.2.2. *With notation as above, let $f \in \mathscr{H}^{\mathrm{ur}}(G_v)$. Then*

$$\mathrm{tr}_{\psi_v}(f) = \boxplus_i \mathrm{tr}_{\tau_{i,v}}(\mathcal{T}_{\Delta,i} f),$$

where the $\mathcal{T}_{\Delta,i}$ are the factors of $\mathcal{T}_\Delta f$ at each $U(n_i)$.

The next lemma gives some control over the size of the $\mathcal{T}_{\Delta,i} f$.

Lemma 6.2.3. *Let $f \in \mathscr{H}^{\mathrm{ur}}(G_v)^{\leq \kappa}$ with $\|f\|_\infty \leq 1$. Then $\mathcal{T}_\Delta f \in \mathscr{H}^{\mathrm{ur}}(H_v)^{\leq \kappa}$ and $\|\mathcal{T}_\Delta f\|_\infty = O(q_v^{D\kappa} \kappa^E)$ for some constants D and E that only depend on G and Δ.*

Proof. This is a slightly more complicated version of the argument of lemma 5.5.4 in [26]. There is an additional step from bounding the trace of the Satake parameter of the trivial representation against the finite dimensional irreps of ${}^L\mathrm{GL}_{d_i}$ that appear as factors in restrictions from ${}^L G_v$ to $\mathbf{H}$. This can be seen to be $O(q_v^{D'\kappa})$ for some D' by the Weyl character formula. $\square$

We also can similarly define a partial unramified transfer map $\mathcal{T}_{\bar{\Delta}}$ such that

$$\mathrm{tr}_{\psi_v}(f) = \boxplus_i \mathrm{tr}_{\tau_{i,v}[d_i]}(\mathcal{T}_{\bar{\Delta},i} f).$$

By the same arguments, this map also satisfies Lemma 6.2.3. We will often suppress Δ or $\bar{\Delta}$ in notation—which version of $\mathcal{T}$ we are using should always be clear by the context of what the $\mathcal{T}_i$'s have image in. We note for intuition that since it acts on unramified functions, $\mathcal{T}_{\bar{\Delta}}$ can be thought of as hyperendoscopic transfer as in Theorem 5.2.1. See [26, §5.5] for more details.

Finally, if Δ is a simple shape of the form $(1, d, \lambda, \eta)$, then for $\psi \in \Delta$ the possible s_{ψ_v} from formula (6.2.3) are all Satake parameters of characters. In addition, $H_v = U_1(F_v) = (G_v)^{\mathrm{ab}} = (G^{\mathrm{ab}})_v$ (note that $G_{v,\mathrm{der}}$ is semisimple and simply-connected and therefore has trivial cohomology). We can therefore compute

$$\mathcal{T}_\Delta(f_v)(h) = \int_{G_{v,\mathrm{der}}} f_v(hg)\, dg. \tag{6.2.4}$$

6.3. **Transfers at General Places: Inequalities.** We can also say a little at general places:

6.3.1. *Twisted Bernstein Components.* We recall some facts from [62, §6] on Bernstein components for twisted groups. Let G_v be the F_v points of a connected reductive group and let $\widetilde{G}_v = G_v \rtimes \theta$ be a twisted group for some automorphism θ. We assume G_v has a minimal parabolic P_0 and Levi factor M_0 that are both θ-stable. Let $\mathcal{L}(G_v)$ be the set of standard Levis of G_v with respect to M_0.

Definition 6.3.1. The twisted cuspidal supports for $\widetilde{G}_v$ are pairs (M, σ) with $M \in \mathcal{L}(G_v)$ such that $(\theta M, \theta\sigma)$ is conjugate to (M, σ) in G_v.

Definition 6.3.2. We say that a θ-invariant irrep π of G_v has infinitesimal character (M, σ) if π is an irreducible subquotient of $\mathrm{Ind}_{MP_0}^{G_v} \sigma$.

Every θ-invariant representation has an infinitesimal character that is a twisted cuspidal support.

Definition 6.3.3. Let $\tilde{\pi}$ is an irrep of $\widetilde{G}_v$ with non-zero (twisted trace) character. Then the infinitesimal character of $\tilde{\pi}$ is that of $\tilde{\pi}|_{G_v}$.

Note that $\tilde{\pi}|_{G_v}$ is necessarily θ-invariant if $\tilde{\pi}$ has non-zero character. Furthermore, all extensions of $\tilde{\pi}|_{G_v}$ differ by a root of unity of order dividing that of θ.

Definition 6.3.4. Let (M, σ) be a twisted cuspidal support. Its twisted Bernstein component is the set of irreps $\tilde{\pi}$ with non-zero character on $\widetilde{G}_v$ such that their infinitesimal characters are of the form $(M, \sigma\chi)$ for χ an unramified character of M (and $\sigma\chi$ θ-invariant).

Part of the main result of [62] is the following key point:

Lemma 6.3.5. *Let f_v be a compactly supported, smooth function on $\widetilde{G}_v$. Then $\pi \mapsto \operatorname{tr}_\pi f_v$ is supported on a finite number of twisted Bernstein components.*

Now specialize to the case of $G = \tilde{G}(N)$:

Proposition 6.3.6. *Let $\mathfrak{s}$ be a twisted Bernstein component of $\widetilde{G}(N)_v$ and let $H \in \mathcal{E}_{\mathrm{ell}}(N)$. Then there is a finite list $\mathfrak{s}_1, \dots, \mathfrak{s}_n$ of Bernstein components of H_v such that the Arthur packets Π_{ψ_v} for $\psi \in \Psi_v^+(H)$ with $\widetilde{\pi}_\psi \in \mathfrak{s}$ only contain representations in the $\mathfrak{s}_i$.*

Proof. This follows from three facts: first, the compatibility of twisted endoscopic transfer of characters with Jacquet modules as in diagram (C.4) in [80], second, the finiteness of A-packets, and third, the compatibility of transfer of characters with unramified character twist as in proposition 4.4 of [61].

We thank Masao Oi for pointing this out to us. □

6.3.2. *Inequalities.* This lets us show:

Lemma 6.3.7. *Let v be a finite place of F and f_v be a test function on $U_{E/F}(N)(F_v)$. Let $dN_1 = N$. Then there exists test function φ_v on $U_{E/F}(N_1)(F_v)$ such that for all Arthur parameters $\psi_v = \psi_{1,v}[d]$ with $\psi_{\star,v}$ a parameter for $U_{E/F}(N_\star)(F_v)$ and $\psi_{1,v}$ cuspidal:*

$$|\operatorname{tr}_{\psi_v} f_v| \leq (\operatorname{tr}_{\psi_{1,v}} \varphi_v)^d.$$

Proof. First,

$$\operatorname{tr}_{\psi_v} f_v = \operatorname{tr}_{\tilde{\pi}_{\psi_v}} f_v^N.$$

By Bernstein's admissibility theorem (as used in [73, prop. 9.6]) there is C (without loss of generality, $C > 1$) such that for all unirreps $\tilde{\pi}'$ of $\widetilde{G}(N)_v$,

$$|\operatorname{tr}_{\tilde{\pi}'} f_v^N| \leq C.$$

Now consider the function from the unitary dual of $\widetilde{G}(N_1)_v$ to $\mathbb{C}$ given by

$$\Phi_0 : \tilde{\pi} \mapsto tr_{\tilde{\pi}[d]} f_v^N.$$

This is supported the on finitely many Bernstein components $\mathfrak{s}_1, \dots, \mathfrak{s}_i$ by Lemma 6.3.5 and since the cuspidal support of $\tilde{\pi}[d]$ determines that of $\tilde{\pi}$.

Therefore, the representations $\pi \in \Pi_{\psi_{1,v}}$ for $\tilde{\pi}_{\pi_{1,v}} \in \mathfrak{s}_i$ lie in a finite set of Bernstein components by Proposition 6.3.6. Since $U_{E/F}(N_1)$ is not twisted, one can find a function φ_v on $U_{E/F}(N_1)$ so that $\operatorname{tr}_\star \varphi_v \geq C$ on each of these components (e.g. a scalar multiple of the indicator function of a small enough maximal compact depending on the Bushnell-Kutzko types associated to the Bernstein components).

Finally, since $\psi_{1,v}$ is cuspidal and therefore simple, $\operatorname{tr}_{\psi_{1,v}} \varphi_v$ is a sum of various $\operatorname{tr}_\pi \varphi_v$ and in particular larger. Therefore, this choice of φ_v suffices. $\square$

Lemma 6.3.8. *Let v be a place of F and f_v be a test function on $U_{E/F}(N)(F_v)$. Let $N_1 + \cdots + N_n = N$. Then there exists test functions f_i on $U_{E/F}(N_i)(F_v)$ such that for all Arthur parameters $\psi_v = \psi_{1,v} \oplus \cdots \oplus \psi_{n,v}$ with each $\psi_{\star,v}$ a parameter for $U_{E/F}(N_\star)(F_v)$ and the $\psi_{i,v}$ simple:*

$$|\operatorname{tr}_{\psi_v} f_v| \leq \prod_i \operatorname{tr}_{\psi_{i,v}} f_{i,v}.$$

Proof. This is the same argument as Lemma 6.3.7. $\square$

6.4. **Transfers at General Places: Conjectural Equalities.** At the current moment, we do not have any exact equality results for transfers at general places. This is the main reason why our exact asymptotic result 11.4.1 is restricted to unitary groups for unramified E/F. We state the desired conjecture:

Conjecture 6.4.1 (Full Transfer)**.** *Let $\Delta = ((T_i, d_i, \eta_i, \lambda_i))_i$ be a shape and φ_v a trace-positive test function on G_v. Then there are trace-positive functions $\varphi_{i,v}$ on the appropriate quasisplit unitary groups of rank T_i such that for all $\Delta \ni \psi = \bigoplus_i \psi_i[d_i]$ with $\psi_i \in (T_i, 1, \eta_i, \lambda_i)$,*

$$\operatorname{tr}_\psi \varphi_v = \prod_i \operatorname{tr}_{\psi_i} \varphi_{i,v}.$$

Except for the positivity statement, Conjecture 6.4.1 would be implied by two local conjectures on the $\widetilde{G}(N)_v$, the first of which has been expected by experts for a while:

Conjecture 6.4.2 (Stable Transfer)**.** *Let $\Delta = (\Delta_i)_i = (T_i, d_i, \eta_i, \lambda_i)_i$ be a shape of rank N and φ_v a test function on $\widetilde{G}(N)$. Then there are test functions $\varphi_{i,v}$ on each $\widetilde{G}(T_i d_i)$ such that for all choices of $\psi_i \in \Delta_i$*

$$\operatorname{tr}_{\widetilde{\pi}(\psi)} \varphi_{i,v} = \prod_i \operatorname{tr}_{\widetilde{\pi}(\psi_i)} \varphi_{i,v}.$$

Note that $\pi(\psi)$ is just the parabolic induction of the product of the $\pi(\psi_i)$. The difficulty in constructing stable transfers is that the relation between $\widetilde{\pi}(\psi)$ and the $\widetilde{\pi}(\psi_i)$ is far more complicated because of the choice of extension in Section 2.2.3.

The second necessary local conjecture is:

Conjecture 6.4.3 (Speh Transfer)**.** *Let $\Delta = (T, d, \eta, \lambda)$ be a simple shape and φ_v a test function on $\widetilde{G}(Td)$. Then there is test function φ'_v on $\widetilde{G}(T)$ such that for all $\psi \in (T, 1, \eta, \lambda)$,*

$$\operatorname{tr}_{\widetilde{\pi}(\psi[d])} \varphi_v = \operatorname{tr}_{\widetilde{\pi}(\psi)} \varphi'_v.$$

For Speh transfer, even the relation between the untwisted representations $\pi(\psi[d])$ and $\pi(\psi)$ is difficult due to the reducibility of the relevant parabolic inductions.

7. Induction Step Details: General Shapes

In the following two sections, we prove the identities needed in the the third, inductive step (§5.4) of the strategy outlined in §5. While our goal in §5 was to produce exact formulas, in our applications we only need to find an asymptotic formula for $S^H_\Delta(\mathrm{EP}_\lambda f^\infty)$. Therefore, we will only need to solve two problems:

- In this section, prove an upper bound (Theorem 7.3.3) for a general shape to show that terms for shapes with non-dominant contribution are negligible.
- In Section 8, find an exact asymptotic for terms with the very special types of shapes that could be dominant in our application.

This section heavily uses our notations from 1.6.2 for non-factorizable functions.

7.1. **Preliminary Bound.** For any refined shape Δ of rank N, recall from §5.3 that

$$S^N_\Delta(f^\infty) := \sum_{\psi \in \Delta} S^N_\psi(f^\infty) = \sum_{\psi \in \Delta} \epsilon^{H(\Delta)}_\psi(s^{H(\Delta)}_\psi) m_\psi |S_\psi|^{-1} \operatorname{tr}_{\widetilde{\pi}(\psi)^\infty}(f^\infty).$$

Note that Proposition 5.3.1 still holds with subscripts of Δ added to both sides.

By removing the ϵ-sign, we also defined in 5.4:

$$S^{|H(\Delta)|}_\Delta(\eta_\lambda(f^\infty)^{H(\Delta)}) = S^{|N|}_\Delta(f^\infty) := \sum_{\psi \in \Delta} m_\psi |S_\psi|^{-1} \operatorname{tr}_{\widetilde{\pi}(\psi)^\infty}(f^\infty) \tag{7.1.1}$$

To approximate away ϵ_ψ difficulties, we first prove a technical bound relating S^N_Δ to $S^{|H'|}_{\Delta'}$ terms for a suitable choice of H', see [34] for a version of this argument with less notational overhead. It relies on a technical condition that we will assume henceforth for various test functions:

Definition 7.1.1. Let S be a (finite or infinite) set of places of F and f_S a function on G_S. We say that f_S is trace-positive on S if for all unirreps π_S of G_S, $\operatorname{tr}_{\pi_S} f_S \geq 0$.

Proposition 7.1.2. *Let $H = H(\Delta) \in \mathcal{E}_{\mathrm{ell}}(N)$. Assume that f^∞ is trace-positive. Then there is an elliptic endoscopic group $H' = H_1 \times H_2$ of H such that for λ' any infinitesimal character of H' conjugate to λ over H:*

$$|S^N_\Delta((f^\infty)^N)| \leq C \bigoplus_{i=1,2} S^{H_i}_{\Delta_i}(\eta_{\lambda'_i}(f^\infty)^i).$$

Here we use shorthand: the data Δ, λ', and $(f^\infty)^{H'}$ all factor into components for H_1 and H_2 denoted with appropriate sub/superscripts. The C is a constant depending only on H and Δ.

The H' further satisfies that

$$S^{H_1}_{\Delta_1}(\eta_{\lambda'_1}(f^\infty)^1) S^{H_2}_{\Delta_2}(\eta_{\lambda'_2}(f^\infty)^2) = S^{|H_1|}_{\Delta_1}(\eta_{\lambda'_1}(f^\infty)^1) S^{|H_2|}_{\Delta_2}(\eta_{\lambda'_2}(f^\infty)^2).$$

Proof. Moving the absolute value within the sum from an intermediate computation in 5.3.1, we know that

$$|S^N_\Delta((f^\infty)^N)| \leq \sum_{\psi \in \Delta} m_\psi |\mathcal{S}_\psi|^{-1} \left(\sum_{\pi \in \Pi^H_{\psi^\infty}} \operatorname{tr}_{\pi^\infty}(f^\infty) \right).$$

Consider $\psi \in \Delta$. All such ψ have the same s_ψ and therefore the same H' in the pair $(H', \psi^{H'})$ corresponding to s_ψ as in Proposition 2.6.1. The tuples (H', ψ') that come from (ψ, x) form an $\widehat{H}$-conjugacy class intersected with the parameters of $\widehat{H}'$. They can therefore be specified by their infinitesimal character at ∞. In particular, the choice of λ' uniformly determines a unique choice of ψ' for each $\psi \in \Delta$.

Then, by the stable multiplicity formula, iterations of the computations of Proposition 5.3.1, and the twisted character identity away from infinity:

$$S^{H'}_{\psi'}(\eta_{\lambda'}(f^\infty)') = \epsilon^{H'}_{\psi'}(s_{\psi'}) m_{\psi'} |\mathcal{S}_{\psi'}|^{-1} \left(\sum_{\pi \in \Pi^H_{\psi^\infty}} \eta^{\psi^\infty}_{\pi^\infty}(s_\psi x) \operatorname{tr}_{\pi^\infty}(f^\infty) \right)$$
$$= m_{\psi'} |\mathcal{S}_{\psi'}|^{-1} \left(\sum_{\pi \in \Pi^H_{\psi^\infty}} \operatorname{tr}_{\pi^\infty}(f^\infty) \right),$$

where the second equality uses the endoscopic sign lemma 2.6.4 that $\epsilon^{H'}_{\psi'}(s_{\psi'}) = \epsilon^G_\psi(s_\psi x)$ and that $s_\psi x = s_\psi^2 = 1$. We also use f' as shorthand for the transfer to H'.

Since $\epsilon^{H'}_{\psi'}(s_{\psi'}) = 1$, the stable multiplicity formula also gives

$$S^{H'}_{\psi'}(\eta_{\lambda'}(f^\infty)') = S^{|H'|}_{\psi'}(\eta_{\lambda'}(f^\infty)').$$

Summing over $\psi \in \Delta$ on one side and the corresponding ψ' determined by λ' gives

$$|S^N_\Delta((f^\infty)^N)| \le \frac{m_\psi |\mathcal{S}_\psi|^{-1}}{m_{\psi'} |\mathcal{S}_{\psi'}|^{-1}} S^{H'}_\Delta(\eta_{\lambda'}(f^\infty)')$$

where $S^{H'}_\Delta$ is defined by summing over parameters that pushforward to something of the right shape on H. Both desired expressions then follow from factoring the H' term into terms for H_1 and H_2 in these last two formula. □

7.2. **Reduction to Simple Shapes.** We can now attempt to bound S_Δ in terms $S_{\Delta'}$ for cuspidal Δ'. We will do this in two steps: in this section, reduce to simple shapes and in the next, from simple to cuspidal.

To apply results of §6, we present test functions in a particular form: choose $G \in \mathcal{E}_{\mathrm{ell}}(N)$ and let $S = \infty \sqcup S_s \sqcup S_b$ be finite sets of places such that S_s is split and S_b contains all the "bad" ramified places. Write test functions as:

$$f^\infty = \varphi_{S_b} f_{S_s} f^S$$

where φ_{S_b} and f_{S_s} are arbitrary and $f^S \in \mathscr{H}^{\mathrm{ur}}(G^S)$.

Proposition 7.2.1. *Consider the refined shape $\Delta = (T_i, d_i, \lambda_i, \eta_i)_i$. Assume that f_{S_s} and f^S are trace-positive. Then*

$$S^{|N|}_\Delta((f_{S_s} f_{S_b} f^S)^N) \le C_\Delta \bigoplus_i S^{T_i d_i}_{(T_i, d_i, \lambda_i, \eta_i)}((((f_{S_s})_{M,i} \varphi_{i,S_b} \mathcal{T}_i f^S)^{T_i d_i}).$$

for some constant C_Δ that only depends on Δ and some trace-positive φ_{i,S_b}.

Proof. We will first work with a single $\psi = \tau_1[d_1] \oplus \cdots \oplus \tau_k[d_k] \in \Delta$ and compute the terms in the stable multiplicity formula. Iteratively applying Lemma 6.1.2 on transfer for split groups, we realize

$$M = \prod_i \mathrm{GL}_{T_i d_i}(F_{S_s})$$

as a Levi of $G(F_{S_s}) \simeq \mathrm{GL}_N(F_{S_s})$ with

$$\mathrm{tr}_{\psi_{S_s}} f_{S_s} = \coprod_i \mathrm{tr}_{\tau_{i,S_s}[d_i]} (f_{S_s})_{M,i},$$

where $(\star)_{M,i}$ is the ith factor of the constant term map to M.

For the places not in S, Lemma 6.2.2 gives that

$$\mathrm{tr}_{\psi^S} f^S = \coprod_i \mathrm{tr}_{\tau_i^S[d_i]} \mathcal{T}_i f^S,$$

and for places in S_b, Lemma 6.3.8 constructs φ_{i,S_b} so that

$$|\,\mathrm{tr}_{\psi_{S_s}} \varphi_{S_b}| \le \prod_i \mathrm{tr}_{\tau_{i,S_b}[d_i]} \varphi_{i,S_b}.$$

Next, each $\epsilon_{\tau_i[d_i]}$ is trivial for each of the $\tau_i[d_i]$ since they are simple. Multiplying together the above trace identities, we get

$$\left| S_\psi^{|N|}((f_{S_s}\varphi_{S_b} f^S)^N) \right| \le C_\Delta \left| \coprod_i S^{T_i d_i}_{\tau_i[d_i]}((((f_{S_s})_{M,i}\varphi_{i,N}\mathcal{T}_i f^S)^{T_i d_i}) \right| \tag{7.2.1}$$

for some constant C_Δ that only depends on Δ.

By regularity of the infinitesimal character, none of the $(T_i, d_i, \lambda_i, \eta_i)$ appear with multiplicity more than 1. Therefore, summing (7.2.1) over all $\psi \in \Delta$ is the same as iteratively summing over each factor $\tau_i[d_i] \in (T_i, d_i, \lambda_i, \eta_i)$.

To remove the absolute values on the right side of the summed equation, $(f_{S_s})_{M,i}$ and $\mathcal{T}_i f^S$ satisfy the same positivity condition: positivity for $(f_{S_s})_{M,i}$ follows by "adjointness" of constant term and parabolic induction. That for $\mathcal{T}_i f^S$ comes from the "adjointness" between $\mathcal{T}_i$ and pushforward of Satake parameter. The $\varphi_{i,N}$ satisfy positivity by construction. In addition, the $(T_i, d_i, \lambda_i, \eta_i)$ are simple so the ϵ_ψ factors are positive. This guarantees that the absolute value can be removed on the right of (7.2.1). □

Doing the bookkeeping for what exactly Δ_1, Δ_2 are in Proposition 7.1.2 lets us bound S^N instead of $S^{|N|}$:

Corollary 7.2.2. *Let the notation be as in the above discussion and assume that f_{S_s}, φ_{S_b}, and f^S are all trace-positive. Then for some constant C_Δ only depending on Δ and some functions $\varphi'_{i,N}$,*

$$|S^N_\Delta((f_{S_s} f_{S_b} f^S)^N)| \le C_\Delta \coprod_i S^{T_i d_i}_{(T_i,d_i,\lambda_i,\eta_i)}((((f_{S_s})_{M,i}\varphi'_{i,S_b}\mathcal{T}_i f^S)^{T_i d_i}).$$

Furthermore, $(f_{S_s})_{M,i}$, φ'_{i,S_b} and $\mathcal{T}_i f^S$ satisfy the same positivity condition.

Proof. First, apply Proposition 7.1.2 to get a bound by $S^{|N|}$ terms and then apply Proposition 7.2.1 to the two terms on the right side of the bound in 7.1.2. Note that we need to use the full fundamental lemma (see e.g. Theorem 5.4.2 in [26]) to choose a transfer f^S to H' that is dual to the pushforward of Satake parameters. The φ' are constructed through the above discussion applied to to transfer to H' of f_{S_b}. Finally, the transfer of f_{S_s} to H' is given by taking constant terms since we are in a degenerate case of H'_{S_s} being a Levi subgroup.

The positivity condition follows for $(f_{S_s})_{M,i}$ by the "adjointness" of constant term and parabolic induction. That for $\mathcal{T}_i f^S$ comes from the "adjointness" between $\mathcal{T}_i$ and pushforward of Satake parameter. That for φ'_{i,S_b} is by construction. □

7.3. **Full Bound.** Now we reduce from simple shapes to cuspidal shapes. let $G \in \mathcal{E}_{\mathrm{ell}}(N)$ and $\Delta = (T, d, \lambda, \eta)$ a simple shape for G, so that all $\psi \in \Delta$ are of the form $\psi = \tau[d]$. As before, choose test function

$$f^\infty = f_{S_s} f_{S_b} f^S$$

on G where S_s is all split places, S_b contains all ramified places, and $f^S \in \mathscr{H}^{\mathrm{ur}}(G^S)$. Further assume:

- $f_{S_s} f_{S_b}$ is supported on the kernel of $|\det|_{S_s \sqcup S_b}$,
- f_{S_s} and f^S are trace-positive.

The argument is very similar to the previous section. First,

$$|S^M_\Delta((f_{S_s} f_N \mathbf{1}_{K^S})^M)| \leq \sum_{\psi \in \Delta} |S^M_\psi((f_{S_s} f_N \mathbf{1}_{K^S})^M)|$$
$$= \sum_{\psi \in \Delta} m_\psi |\mathcal{S}_\psi| (\operatorname{tr}_{\psi_{S_s}} f_{S_s})(\operatorname{tr}_{\psi^S} \mathbf{1}_{K^S})| \operatorname{tr}_{\psi_{S_b}} f_{S_b}|.$$

For all $\psi \in \Delta$, apply Lemma 6.1.3 to get a Levi $M_{S_s} \simeq {\mathrm{GL}_T}^d$ of $\mathrm{GL}_M(F_{S_s})$ so that

$$\operatorname{tr}_{\psi_{S_s}} f_{S_s} \leq (\operatorname{tr}_{\tau_{S_s}} (f_{S_s})_{M,1})^{d\oplus}.$$

We also apply 6.3.7 to construct the functions φ'_{S_b} satisfying

$$|\operatorname{tr}_{\psi_{S_b}} \varphi_{S_b}| \leq (\operatorname{tr}_{\tau_{S_b}} \varphi'_{S_b})^{d\oplus}$$

since the τ_{S_b} are necessarily cuspidal. Note that the φ'_{S_b} so constructed is trace-positive by definition. Applying Lemma 6.2.2 then gives

$$|\operatorname{tr}_{\psi^\infty}(f_{S_s} \varphi_{S_b} f^S)| \leq (\operatorname{tr}_{\tau^\infty}((f_{S_s})_{M,1} \varphi'_{S_b} \mathcal{T} f^S))^{d\oplus}. \tag{7.3.1}$$

Finally, summing over all $\tau[d] \in \Delta$ gives that:

Proposition 7.3.1. *With notation and conditions from the above discussion:*

$$|S^M_\Delta((f_{S_s} \varphi_{S_b} f^S)^M)| \leq C_\Delta (S^T_{\Sigma_{\lambda,\eta}}((((f_{S_s})_{M,1} \varphi'_{S_b} \mathcal{T} f^S))^T)^{d\oplus}$$

for some constant C that only depends on Δ. Recall that $(f_{S_s})_{M,1}$ and $\mathcal{T} f^S$ are trace positive and we can choose φ'_{S_b} to be so too.

Proof. Expanding out and using that $\Sigma_{\lambda,\eta}$ is simple:

$$(S^T_{\Sigma_{\lambda,\eta}}((((f_{S_s})_{M,1} \varphi'_{S_b} \mathcal{T} f^S))^T)^{d\oplus}$$
$$= \sum_{\tau \in \Sigma_{\lambda,\eta}} m^d_\psi |\mathcal{S}_\psi|^d (\operatorname{tr}_{\tau^\infty}((f_{S_s})_{M,1} \varphi'_{S_b} \mathcal{T} f^S))^{d\oplus} + \text{cross terms},$$

where trace-positivity comes from the conditions in the above discussion and arguments similar to Corollary 7.2.2. Since m_ψ and $|S_\psi|$ only depend on $\Sigma_{\lambda,\eta}$, equation (7.3.1) then proves the inequality since trace positivity gives that the cross terms are all positive. We use here that Δ is simple so $\epsilon_\psi(s_\psi)$ is identically 1. $\square$

We will need a slight technical variation of this:

Proposition 7.3.2. *Fix notation and conditions as in Proposition 7.3.1. Assume there is a constant B such that for all $\tau \in \Sigma_{\lambda,\eta}$:*

$$|\operatorname{tr}_{\tilde{\tau}^\infty}((((f_{S_s})_{M,1} \varphi'_{S_b} \mathcal{T} f^S)^T)|^{\oplus} \leq B$$

where the $\oplus$ represents a sum over factorizable summands. Then

$$|S^M_\Delta((f_{S_s}\varphi_{S_b}f^S)^M)| \leq CB^{d-1}S^T_{\Sigma_{\lambda,\eta}}((((f_{S_s})_{M,1}\varphi'_{S_b}\mathcal{T}f^S)^T)^\oplus$$

for some constant C that only depends on Δ.

Proof. This is the same argument as 7.3.1 except we bound

$$\sum_{\tau\in\Sigma_{\lambda,\eta}} m^d_\psi|\mathcal{S}_\psi|^d(\mathrm{tr}_{\tau^\infty}((f_{S_s})_{M,1}\varphi'_{S_b}\mathcal{T}f^S))^{d\oplus}$$

directly instead of adding in cross terms. □

Our final bound for a general shape then becomes:

Proposition 7.3.3. *Let $\Delta = (T_i, d_i, \eta_i, \lambda_i)_i$ and $G = H(\Delta) \in \mathcal{E}_{\mathrm{ell}}(N)$. Let*

$$f = f_{S_s}\varphi_{s_b}f^S$$

be a test function on G^∞ where S_s is all split places, S_b contains all the ramified places, and $f^S \in \mathscr{H}^{\mathrm{ur}}(G^S)$. We further require:

- *f_{S_s} is supported on the kernel of $|\det|_{S_s}$,*
- *f_{S_s}, φ_{S_b}, and f^S are trace-positive.*

Then there is a Levi subgroup

$$M \simeq \prod_i \mathrm{GL}_{T_i}(F_v)^{d_i}$$

of $G(F_{S_s}) \simeq \mathrm{GL}_N(F_{S_s})$ and functions φ'_{i,S_b} such that

$$|S^N_\Delta((f_{S_s}f_{S_b}f^S)^N)| \leq C_\Delta \prod_i^{\oplus}(S^{T_i}_{\Sigma_{\lambda_i,\eta_i}}((((f_{S_s})_{M,i}\varphi'_{i,S_b}\mathcal{T}_i f^S)^{T_i}))^{d_i\oplus}.$$

Here $(f_{S_s})_{M,i}$ is the restriction of the constant term to M to the first factor in $\mathrm{GL}_{T_i}(F_v)^{d_i}$, $\mathcal{T}_i$ is defined similarly, and the constant C_Δ only depends on Δ.

Finally, we may choose $\varphi'_{i,N}$ so that $(f_{S_s})_{M,i}$, $\varphi'_{i,N}$ and $\mathcal{T}_i f^S$ satisfy the same bulleted conditions.

Proof. The inequality comes from applying 7.2.2 and then 7.3.1. The first thing to check is that the necessary conditions for 7.3.1 still hold after applying 7.2.2. Positivity is guaranteed by the second implication of 7.2.2. Support follows from the integral formula for constant term.

The final φ', $(f_{S_s})_{M,i}$, and $\mathcal{T}_i f^s$ satisfy the bulleted conditions by the last implications of 7.3.1. □

8. Induction Step Details: Particular Shapes

We compute more detailed information for certain particular kinds of shapes with the goal of providing exact asymptotics:

8.1. **Shapes of Characters.** Let $\Delta = (1, d, \lambda, \eta)$ be a shape for $H = H(\Delta)$. If $\psi \in \Delta$, then the packet Π_{ψ_∞} constructed by Adams-Johnson, see §3.2.2, is a singleton containing a character ξ_∞ of H. Let $\mathrm{EP}_{\xi_\infty} = \mathrm{vol}(A_{G,\infty}\backslash G_\infty/G_{\mathrm{der},\infty})^{-1}\xi^{-1}_\infty$ be the corresponding Euler-Poincaré function (note that by our assumptions on E/F, H is cuspidal in the sense of [7] so ξ^{-1}_∞ is a valid test function factor for the global trace formula).

Proposition 8.1.1. *Let $\Delta = (1, d, \lambda, \eta)$ be a shape for H. Then,*

$$S^H_\Delta(\mathrm{EP}_{\xi_\infty} f^\infty) = \sum_{\substack{\chi \in \mathcal{AC}_{\mathrm{disc}}(H) \\ \chi_\infty = \xi_\infty}} \mathrm{tr}_{\chi^\infty} f^\infty,$$

where $\mathcal{AC}_{\mathrm{disc}}(H)$ is the set of one-dimensional representations in the discrete automorphic spectrum of H.

Proof. If $\psi \in \Delta$, any $\pi \in \Pi_\psi$ has $\pi_\infty = \xi_\infty$ as explained above. This implies π is one-dimensional by a well-known result (see [46, lem. 6.2] for example). Furthermore, ψ is simple, so as distributions,

$$S_\psi = \sum_{\pi \in \Pi_\psi} \mathrm{tr}_\pi = I_\psi.$$

Evaluating at our test function,

$$S_\psi((\mathrm{EP}_{\xi_\infty} f^\infty) = I_\psi(\mathrm{EP}_\xi f^\infty) = \sum_{\pi \in \Pi_\psi} \mathrm{tr}_{\pi^\infty}(f^\infty).$$

On the other hand, any $\chi \in \mathcal{AC}_{\mathrm{disc}}(H)$ appears in some A-packet Π_ψ. If $\chi_\infty = \xi_\infty$, then ψ_∞ has full Arthur-SL_2 and infinitesimal character λ_i, which forces $\pi \in \Delta$. Furthermore, we recall that characters appear with multiplicity at most one in the automorphic spectrum (realized as functions by evaluation). Therefore every such χ can only appear in one packet. In total, the union over Π_ψ for $\psi \in \Delta$ is exactly the subset of $\mathcal{AC}_{\mathrm{disc}}(H)$ with infinite component ξ_∞ and this union is disjoint.

Summing over $\psi \in \Delta$ then finishes the argument. $\square$

Next, $\mathcal{AC}_{\mathrm{disc}}(H)$ are the characters of

$$\Xi(H) := H(F)^{\mathrm{ab}} \backslash H(\mathbb{A})^{\mathrm{ab}}$$

so we can therefore write

$$S^H_\Delta(\mathrm{EP}_{\xi_\infty} f^\infty) = \frac{1}{\mathrm{vol}(H^{\mathrm{ab}}_\infty)} \sum_{\chi \in \Xi(H)^\vee} \widehat{\xi_\infty^{-1} f^{\infty,\mathrm{ab}}}(\chi),$$

where $f^{\infty,\mathrm{ab}}$ is the pushforward (by integration against $H(\mathbb{A}^\infty)_{\mathrm{der}}$) of f^∞ to $H(\mathbb{A}^\infty)^{\mathrm{ab}}$. Then:

Corollary 8.1.2. *Let $H = H(\Delta)$ for Δ be a shape of the form $(1, d, \lambda, \eta)$ corresponding to character ξ_∞ on H_∞. Then:*

$$S^H_\Delta(\mathrm{EP}_{\xi_\infty} f^\infty) = \frac{\mathrm{vol}(H(F)^{\mathrm{ab}} \backslash H(\mathbb{A})^{\mathrm{ab}})}{\mathrm{vol}(H^{\mathrm{ab}}_\infty)} \sum_{\gamma \in H(F)^{\mathrm{ab}}} \xi_\infty^{-1} f^{\infty,\mathrm{ab}}(\gamma),$$

where $f^{\infty,\mathrm{ab}}$ is as immediately above.

Proof. Since $H^{\mathrm{ab}}(F) \backslash H^{\mathrm{ab}}(\mathbb{A})$ is compact for our specific H, we get that $H(F)^{\mathrm{ab}}$ is co-compact in $H(\mathbb{A})^{\mathrm{ab}}$. Since $H(F)^{\mathrm{ab}}$ is a subgroup of $H^{\mathrm{ab}}(F)$ and $H(\mathbb{A})^{\mathrm{ab}}$ is a subgroup of $H^{\mathrm{ab}}(\mathbb{A})$, discreteness of $H^{\mathrm{ab}}(F)$ in $H^{\mathrm{ab}}(\mathbb{A})$ gives discreteness of $H(F)^{\mathrm{ab}}$ in $H(\mathbb{A})^{\mathrm{ab}}$.

Therefore Poisson summation gives the result. $\square$

Our $H \in \mathcal{E}_{\mathrm{ell}}(N)$ is necessarily isomorphic as a reductive group to $U(N)$ so we will usually have $H^{\mathrm{ab}} = U(1)$ in the above.

8.2. **Odd GSK Shapes.** We will eventually focus on shapes that are similar to the Saito-Kurokawa case (d) at the end of [9]:

Definition 8.2.1. We say a shape $\Delta = (T_i, d_i, \lambda_i, \eta_i)_{1 \le i \le k}$ is *GSK* or *generalized Saito-Kurokawa* if:

- $d_1 = 1$ and for all $i > 1$, $T_i = 1$,
- The d_i are distinct integers.

We furthermore say it is *odd GSK* if:

- The d_i are all odd.

We will get exact asymptotic bounds for such S_Δ since these are the ones where the only Speh representations that appear are characters. Understanding more general shapes would require the Speh Transfer Conjecture 6.4.3.

We recall the setup of Sections 7.2 and 7.3 and follow a similar argument: choose $G \in \mathcal{E}_{\mathrm{ell}}(N)$ and pick finite sets of places $S = \infty \sqcup S_s \sqcup S_b$ where S_s is all split and S_b contains all the ramified places. We will look at test functions of the form:

$$f^\infty = \varphi_{S_b} f_{S_s} f^S$$

where φ_{S_b} and f_{S_s} are arbitrary and $f^S \in \mathscr{H}^{\mathrm{ur}}(G^S)$. We will further assume that the pair (Δ, φ_{S_b}) satisfies the Stable Transfer Conjecture 6.4.1 and that Δ is odd GSK.

Consider $\psi = \tau_1[d_1] \oplus \cdots \oplus \tau_k[d_k] \in \Delta$. For all such ψ, Lemma 6.1.2 gives Levi

$$M = \prod_i \mathrm{GL}_{T_i d_i}(F_{S_s})$$

of $G(F_{S_s}) \simeq \mathrm{GL}_N(F_{S_s})$ so that

$$\mathrm{tr}_{\psi_{S_s}} f_{S_s} = \coprod_i \mathrm{tr}_{\tau_{i,S_s}[d_i]}(f_{S_s})_{M,i},$$

and Lemma 6.2.2 gives that

$$\mathrm{tr}_{\psi^S} f^S = \coprod_i \mathrm{tr}_{\tau_i^S[d_i]} \mathcal{T}_i f^S.$$

Note that the constant terms and $\mathcal{T}_i f^S$ satisfy the positivity condition by various "adjointesses" of trace and various transfers.

Finally, our assumption that φ_{S_b} satisfies Conjecture 6.4.1 gives φ_{i,S_b} such that

$$\mathrm{tr}_{\psi_{S_b}} \varphi_{S_b} = \coprod_i \mathrm{tr}_{\tau_{i,S_s}[d_i]} \varphi_{i,S_b}$$

where the φ_{i,S_b} are trace-positive.

This produces:

Proposition 8.2.2. *With notation and conditions as above (in particular, that the pair (Δ, φ_{S_b}) satisfies Conjecture 6.4.1 and Δ is odd GSK),*

$$S^H_\Delta(\mathrm{EP}_\lambda \varphi_{S_b} f_{S_s} f^S) = 2^{-k+1} \coprod_i S^{H_i}_{(1,d,\lambda_i,\eta_i)}(\mathrm{EP}_\lambda (f_{S_s})_{M,i} \varphi_{i,S_b} \mathcal{T}_i f^S),$$

where $H_i = H(1, d, \lambda_i, \eta_i)$.

Proof. As in Section 7.2, we multiply the together and sum the trace equalities above. Note that Δ having all odd d_i gives that $S^H_\Delta = S^{|H|}_\Delta$ by Lemma 2.4.3. Furthermore, since we are working with unitary groups $m_\psi = 1$ always. Finally, for $\psi \in \Delta$ as above, since the $\tau_i[d_i]$ are simple,

$$|\mathcal{S}_\psi|^{-1} \prod_i |\mathcal{S}_{\tau_i[d_i]}| = 2^{-k+1}.$$

This computes all the terms in the stable multiplicty formula. □

9. Level-Aspect Asymptotics

9.1. **Setup.** In this section, we use the strategy outlined in §5 to compute some asymptotics of $S^G_\Delta(\eta_\lambda f^\infty)$ for a specific sequence of f^∞.

Fix N and $G \in \mathcal{E}_{\mathrm{sim}}(N)$ such that $G = H(\Sigma_{\lambda,\eta})$. We use the setup from [73]. First, as some notation:

- $K := K^G_R$ is a choice of hyperspecial subgroup of G at a set of unramified places R. Similarly define $K^{G,R}$.
- $\dim \lambda$ for λ an integral regular infinitesimal character is the dimension of the associated finite-dimensional representation on $\mathrm{GL}_n\mathbb{C}$.
- $K(\mathfrak{n}) := K^G_R(\mathfrak{n})$ for $\mathfrak{n}$ and ideal of $\mathcal{O}_F$ supported on a set of unramified places R is the $\mathfrak{n}$th congruence subgroup of K^G_R:

$$K^G_R(\mathfrak{n}) = \prod_{v|R} K^G_v(q_v^{v(\mathfrak{n})}),$$

  where $K^G_v(q_v^{v(\mathfrak{n})})$ is the corresponding step in the Moy-Prasad filtration. Define $K^{G,R}(\mathfrak{n})$ similarly avoiding places in R.
- A choice of measure on G_∞ induces one on the compact form G^c_∞ in the standard way by relating to top forms on $G_{\infty,\mathbb{C}}$.

Then, fix:

- A finite set of finite places S_0 including all those where E/F is ramified,
- An arbitrary test function φ_{S_0} at S_0,
- A finite set of finite places S_1 disjoint from S_0,
- An $f_{S_1} \in \mathscr{H}^{\mathrm{ur}}(G_{S_1})^{\leq \kappa}$ for some κ,
- An ideal $\mathfrak{n}$ of F relatively prime to $2, 3, S_0, S_1$ and that is all split in E. We will let $\mathfrak{n}$ vary and consider asymptotics as $|\mathfrak{n}| \to \infty$.
- As notation, let S be the union of ∞, S_0, and S_1.

Then define the test function:

$$f^\infty_{\mathfrak{n}} = \varphi_{S_0} f_{S_1} \bar{\mathbf{1}}_{K^{G,S}(\mathfrak{n})}$$

where $\bar{\mathbf{1}}_{K^{G,S}(\mathfrak{n})}$ is the indicator function normalized by volume to have integral 1.

We will also need some constants related to the $\mathfrak{n}$:

- The norm

$$|\mathfrak{n}| := \prod_{v|\mathfrak{n}} q_v^{v(\mathfrak{n})}.$$

- Euler factors: for $n, n_i \in \mathbb{Z}^+$:

$$\Gamma_n(\mathfrak{n}) := \prod_{v|\mathfrak{n}} (1 - q_v^{-1})(1 - q_v^{-2}) \cdots (1 - q_v^{-n}),$$
$$\Gamma_{-n}(\mathfrak{n}) := \prod_{v|\mathfrak{n}} (1 + q_v^{-1})(1 + q_v^{-2}) \cdots (1 + q_v^{-n}),$$
$$\Gamma_{\pm n_1, \dots, \pm n_k}(\mathfrak{n}) := \Gamma_{\pm n_1}(\mathfrak{n}) \cdots \Gamma_{\pm n_k}(\mathfrak{n}).$$

Our initial input will be the level-aspect bounds of [73].

Theorem 9.1.1 (Special case of [73, Thm. 9.16])**.** *With notation defined as above, there are constants A, B, C, D, E with $C \geq 1$ and depending only on G such that whenever $|\mathfrak{n}| \geq D q_{S_1}^{E\kappa}$,*

$$|\mathfrak{n}|^{-\dim G} \Gamma_N(\mathfrak{n})^{-1} I^G(\mathrm{EP}_\lambda f_\mathfrak{n}^\infty) = \Lambda + O(|\mathfrak{n}|^{-C} q_{S_1}^{A+B\kappa}),$$

where we define the mass:

$$\Lambda = \Lambda(G, f, \varphi) = \varphi_{S_0}(1) f_{S_1}(1) \frac{\dim \lambda}{|\Pi_{\mathrm{disc}}(\lambda)|} \frac{\mathrm{vol}(G(F) \backslash G(\mathbb{A}_F))}{\mathrm{vol}(K^S) \mathrm{vol}(G_\infty^c)}.$$

Proof. Since only split primes divide $\mathfrak{n}$, we have

$$[K : K(\mathfrak{n})] = |\mathfrak{n}|^{\dim G} \Gamma_N(\mathfrak{n})$$

by a standard formula for the sizes of the $\mathrm{GL}_n(\mathcal{O}_v / q_v^{v(\mathfrak{n})} \mathcal{O}_v)$. □

When $S_0 = S_1 = \emptyset$, the quotient of volumes is the modified Tamagawa number computed in [73] corollary 6.14:

$$\frac{\mathrm{vol}(G(F) \backslash G(\mathbb{A}_F))}{\mathrm{vol}(K^\infty) \mathrm{vol}(G_\infty^c)} = \tau'(G) = 2^{-(N-1)\deg F} \tau(G) L(\mathrm{Mot}_G) |\Omega_G| |\Omega_{G_\infty}^c|^{-1} \tag{9.1.1}$$

where $\tau(G)$ is the Tamagawa number, $L(\mathrm{Mot}_G)$ is the L-value of the motive from [36], and $\Omega_{G_\infty}^c$ is the Weyl group of the maximal compact at ∞.

Since we are not tracking explicit values for A, B, C, D, E, we will allow them to change throughout the following argument.

9.2. **Bounds on Stable Trace.** To extend 9.1.1 to S^G, we next recall a standard formula:

Lemma 9.2.1. *Let G be an unramified reductive group over F, $\mathfrak{n}_i$ an ideal relatively prime to all places where G is ramified, and M a Levi component of parabolic subgroup P. Then we have identity of indicator functions normalized by volume:*

$$(\bar{\mathbf{1}}_{K^G(\mathfrak{n})})_M = I(\mathfrak{n}) \bar{\mathbf{1}}_{K^M(\mathfrak{n})},$$

where we recall that $(\star)_M$ denotes the constant term and

$$I(\mathfrak{n}) = [K : K \cap K(\mathfrak{n}) P].$$

Furthermore, if $G = \mathrm{GL}_N$ then

$$I(\mathfrak{n}_i) = (1 + O(|\mathfrak{n}|^{-2})) |\mathfrak{n}|^{\dim G/P} \Gamma_{-1}(\mathfrak{n})^{\sigma(M)},$$

where

$$\sigma(M) = \mathrm{rank}_{\mathrm{ss}} G - \mathrm{rank}_{\mathrm{ss}} M.$$

Proof. This is well known—see for example the proof of lemma 5.2 in [52]. □

Note for intuition later that $\dim G/P = 1/2(\dim G - \dim M)$. Using this:

Proposition 9.2.2. *For the $f_{\mathfrak{n}}^\infty$ and Λ as in Section 9.1, there are constants A, B, C, D, E depending only on G with $C \geq 1$ such that whenever $|\mathfrak{n}| \geq D q_{S_1}^{E\kappa}$,*

$$|\mathfrak{n}|^{-\dim G} \Gamma_N(\mathfrak{n})^{-1} S^G(\mathrm{EP}_\lambda f_{\mathfrak{n}}^\infty) = \Lambda + O(|\mathfrak{n}|^{-C} q_{S_1}^{A+B\kappa}).$$

Proof. We apply Theorem 5.2.1 and apply Theorem 9.1.1 to each term. This argument is a much less general and much simpler version of the main result of [26] so we present it tersely.

We need only show that all the non-I^G summands can be put into the error term:

- In this case of unitary groups, there are a finite number of such terms.
- We can ignore dependence on φ_{S_0}.
- Lemma 5.5.4 in [26] allows us to bound the value at 1 and the support of the $f_{S_1}^{\mathcal{H}}$.
- Lemma 6.1.2 lets us iteratively use Lemma 9.2.1 to bound the $(\bar{\mathbf{1}}_{K^{G,S}(\mathfrak{n}_i)})^{\mathcal{H}}$.
- Averaging Corollary 5.1.6 in [26] shows that the transfer of the EP-function is a linear combination of a number of EP-functions uniformly bounded over $\mathcal{H}$.

Putting all this together, 9.1.1 shows that the sum of the non-I^G terms are upper-bounded by the claimed error as long as we extremize A, B, C, D, E appropriately over all hyperendoscopic groups. ☐

9.3. **The Induction.** Now we induct to bound the limit multiplicities restricted to specific shapes. Their are two pieces to this argument—first, as a consequence of Proposition 5.3.1,

$$S^G_{\Sigma_{\lambda,\eta}}(\eta_\lambda f_{\mathfrak{n}}^\infty) = S^G(\mathrm{EP}_\lambda f_{\mathfrak{n}}^\infty) - \sum_{\substack{H(\Delta)=H(\Sigma_{\lambda,\eta}) \\ \text{inf. char}(\Delta)=\lambda \\ \Delta \neq \Sigma_{\lambda,\eta}}} S^G_\Delta(\mathrm{EP}_\lambda f_{\mathfrak{n}}^\infty). \tag{9.3.1}$$

Second, the bound in Proposition 7.3.3 lets us show that non-Σ terms S^G_Δ have limit multiplicities controlled by the groups they are lifts from.

We start with a technical trick that allows us to get the trace-positivity conditions needed to apply the results of the previous section:

Lemma 9.3.1. *Let f_v be a test function on G_v. Then there are trace-positive functions $f_1, \dots, f_k$ such that*

$$f_v = \lambda_1 f_1 + \cdots + \lambda_k f_k.$$

Furthermore, if $f_v \in \mathscr{H}^{\mathrm{ur}}(G_v)^{\leq\kappa}$, then so are the f_i and

$$\sum_i |\lambda_i| \leq C, \qquad \sum_i \|f_i\|_\infty \leq C\|f_v\|_\infty$$

for some uniform constant C.

Proof. The first statement works in great generality by the Dixmier-Malliavin decomposition theorem as used [69, Lem. 3.5].

Concretely, we can without loss of generality assume $f_v^* = f_v$ by symmetrizing:

$$f_v = \left(\frac{f_v + f_v^*}{2}\right) - i\left(\frac{f_v - f_v^*}{2i}\right).$$

By smoothness, there is a open compact subgroup U such that $f_v = f_v \star \bar{\mathbf{1}}_U = \bar{\mathbf{1}}_U \star f_v$. Then

$$f_v = \frac{1}{4}(f_v + \bar{\mathbf{1}}_U) \star (f_v + \bar{\mathbf{1}}_U) - \frac{1}{4}(f_v - \bar{\mathbf{1}}_U) \star (f_v - \bar{\mathbf{1}}_U),$$

a linear combination of functions of the form $g \star g^*$, each necessarily trace-positive.

Using $U = K_v$ shows the required bounds for the second statement. □

We can now state the first important technical bound resulting from our methods:

Theorem 9.3.2. *Fix the* $f_{\mathfrak{n}}^\infty$, $G \in \mathcal{E}_{\mathrm{ell}}(N)$, *and* Λ *as in Section 9.1. Then there are constants* A, B, C, D, E *with* $C \geq 1$ *such that whenever* $|\mathfrak{n}| \geq D q_{S_1}^{E\kappa}$,

$$|\mathfrak{n}|^{-\dim G} \Gamma_N(\mathfrak{n})^{-1} S^G_{\Sigma_{\lambda,\eta}}(\mathrm{EP}_\lambda f_{\mathfrak{n}}^\infty) = \Lambda + O(|\mathfrak{n}|^{-C} q_{S_1}^{A+B\kappa}) \tag{9.3.2}$$

and for $\Delta \neq \Sigma_{\lambda,\eta}$,

$$S^G_\Delta(\mathrm{EP}_\lambda f_{\mathfrak{n}}^\infty) = O(|\mathfrak{n}|^{\bar{R}(\Delta)} q_{S_1}^{A+B\kappa}), \tag{9.3.3}$$

where

$$\bar{R}((T_i, d_i, \lambda_i, \eta_i)_{1 \leq i \leq k}) = \frac{1}{2}\left(N^2 + \sum_i T_i^2 d_i\right). \tag{9.3.4}$$

Proof. We induct on the N such that $G \in \mathcal{E}_{\mathrm{ell}}(N)$. For $N = 1$, $\Sigma_{\lambda,\eta}$ is the only possible shape so (by extreme overkill) this follows from Theorem 9.1.1.

For the inductive step, we first argue that it suffices to show (9.3.3): take D and E to be the maximum values over those for all smaller shapes appearing. Then (9.3.3) will show that all the terms in the sum (9.3.1) are lower order in $|\mathfrak{n}|$ than the S^G-asymptotic from Proposition 9.2.2. This produces the exact asymptotic (9.3.2).

Therefore, let $\Delta = (T_i, d_i, \lambda_i, \eta_i)_{1 \leq i \leq k} \neq \Sigma_{\lambda,\eta}$ have rank N. First, apply Lemma 9.3.1 on each of the finitely many unramified factors of φ_{i,S_0} and $\mathcal{T}_i f_{S_i}$ to write

$$S^G_\Delta(\mathrm{EP}_\lambda f_{\mathfrak{n}}^\infty) = \sum_j \lambda_j S^G_\Delta(\mathrm{EP}_\lambda f_{\mathfrak{n},j}^\infty)$$

where the sum of the $|\lambda_j|$ is $O(C^{|S_1|}) = O(q_{S_1})$ and each $f_{\mathfrak{n},j}^\infty$ is trace-positive.

We then apply Proposition 7.3.3 with S_s the primes that divide $\mathfrak{n}$ and $S_b = S_0$. Over S_s, let $M = \prod_i \mathrm{GL}_{T_i}(F_{S_s})^{d_i}$ be the Levi from 7.3.3 and P the parabolic defining the constant term. Then (using notation from 7.3.3 to decompose $f_{\mathfrak{n},j}^\infty$),

$$\begin{aligned} |S^G_\Delta(\mathrm{EP}_\lambda f_{\mathfrak{n},j}^\infty)| &\leq C_\Delta \coprod\!\!\!\!\!\!\bigoplus \left(S^{T_i}_{\Sigma_{\lambda_i,\eta_i}}((((f_{S_s})_{M,i} \varphi'_{i,S_0} \mathcal{T}_i f_{S_1} \mathbf{1}_{K^S})^{T_i})\right)^{d_i \oplus} \\ &= C_\Delta I(\mathfrak{n}) \coprod\!\!\!\!\!\!\bigoplus \left(S^{H_i}_{\Sigma_{\lambda_i,\eta_i}}(\mathrm{EP}_{\lambda_i} \bar{\mathbf{1}}_{K^{H_i}(\mathfrak{n})} \varphi'_{i,S_0} \mathcal{T}_i f_{S_1})\right)^{d_i \oplus}, \end{aligned} \tag{9.3.5}$$

where each $H_i = H(\Sigma^{T_i}_{\lambda_i,\eta_i})$ and the second equality uses Lemma 9.2.1 together with factoring $\mathrm{vol}(K^M(\mathfrak{n}))$ over places in S_s.

From Lemma 6.2.3, we know that $\mathcal{T} f_{S_1} \in \mathscr{H}^{\mathrm{ur}}(M_{S_1})^{\leq \kappa}$ and

$$\|\mathcal{T} f_{S_1}\|_\infty = \|f_{S_1}\|_\infty O(q_v^{\kappa F} \kappa^G).$$

Therefore, we can use the inductive hypothesis on the $\Sigma_{\lambda_i,\eta_i}$, summing the asymptotics of the Λ and the error to get that each term in the product (9.3.5) is

$$O\left(|\mathfrak{n}|^{\dim H_i} \Gamma_{T_i^{(d_i)}}(\mathfrak{n}) q_{S_1}^{A'+B'\kappa}\right)$$

for some constants A', B' depending on Δ. Note that the positivity condition is guaranteed by the last implication of Proposition 7.3.3. Summing over j, this gives

$$|S^G_\Delta(\mathrm{EP}_\lambda f^\infty_{\mathfrak{n}})| = O\left(\Gamma_{T_1^{(d_1)},\dots,T_k^{(d_k)}}(\mathfrak{n})I(\mathfrak{n})|\mathfrak{n}|^{\dim M} q_{S_1}^{A'+B'\kappa}\right),$$

ignoring all dependence on the φ'.

Recalling that $G_{S_s} \simeq \mathrm{GL}_{N,S_s}$, we can use the asymptotic for $I(\mathfrak{n})$ from 9.2.1:

$$I(\mathfrak{n})|\mathfrak{n}|^{\dim M} = O\left(|\mathfrak{n}|^{\dim P}\Gamma_{-1}(\mathfrak{n})^{\sigma(M)}\right).$$

We finally note that $\Gamma_{T_1^{d_1},\dots,T_k^{d_k}}(\mathfrak{n})\Gamma_{-1}(\mathfrak{n})^{\sigma(M)} \leq 1$ and that

$$\dim P = \bar{R}(\Delta) = \frac{1}{2}\left(N^2 + \sum_i T_i^2 d_i\right).$$

This finishes the bound for Δ and therefore the induction. $\square$

We emphasize that the value $\bar{R}(\Delta)$, which is the dimension of the parabolic attached to the partition

$$(T_1^{(d_1)}, \dots, T_k^{(d_k)}),$$

is an approximate upper bound to what we expect to be the true growth rate attached to Δ. It is maximized at $\bar{R}(\Sigma_{\lambda,\eta}) = \dim G$ where it is exact.

9.4. **Improving the Bound.** We can prove a slightly tighter bound $R(\Delta)$. Instead of applying 7.3.3 directly as in the proof of 9.3.2, we will apply 7.2.2 and then use a different method to bound the factors for simple blocks $(T_i, d_i, \lambda_i, \eta_i)$ with small T_i. For large T_i, we will proceed as normal using the arguments of §7.3. This argument is nothing more than a rephrasing in our context of the key technical trick that makes the bounds in [52] work.

9.4.1. *Terms with $T_i = 1$.* In this case we actually prove a stronger exact formula:

Lemma 9.4.1. *When $T_i = 1$, the terms for the factors $S_{(T_i,d_i,\lambda_i,\eta_i)}$ coming from the use of Corollary 7.2.2 implicit in* (9.3.5) *satisfy the bound:*

$$\begin{aligned}
&S^{H_i'}_{(1,d_i,\lambda_i,\eta_i)}(EP_{\lambda_i'}\bar{\mathbf{1}}_{K^{H_i'}(\mathfrak{n})}\varphi_{i,S_0}\mathcal{T}_i f_{S_1}) \\
&\qquad = |\mathfrak{n}|\Gamma_1(\mathfrak{n}_i)\frac{\mathrm{vol}(H_i'(F)^{\mathrm{ab}}\backslash H_i'(\mathbb{A})^{\mathrm{ab}})}{\mathrm{vol}(K^{(H_i')^{\mathrm{ab}},S})\,\mathrm{vol}((H'_{i,\infty})^{\mathrm{ab}})}\int_{H'_{i,\mathrm{der},S_1,S_0}} \mathcal{T}_i f_{S_1}\varphi_{i,S_0}(h)\,dh \\
&\qquad = O(|\mathfrak{n}|\Gamma_1(\mathfrak{n})q_{S_1}^{A+B\kappa})
\end{aligned}$$

as long as $|\mathfrak{n}| > Dq_{S_1}^{E\kappa}$.

Proof. By Corollary 8.1.2,

$$\begin{aligned}
&S^{H_i'}_{(1,d_i,\lambda_i,\eta_i)}(EP_{\lambda_i'}\bar{\mathbf{1}}_{K^{H_i'}(\mathfrak{n})}\varphi_{i,S_0}\mathcal{T}_i f_{S_1}) \\
&\quad = \frac{\mathrm{vol}(H_i'(F)^{\mathrm{ab}}\backslash H_i'(\mathbb{A})^{\mathrm{ab}})}{\mathrm{vol}((H'_{i,\infty})^{\mathrm{ab}})}\sum_{\gamma\in H_i'(F)^{\mathrm{ab}}}(\bar{\mathbf{1}}_{K^{H_i'}(\mathfrak{n})})^{\mathrm{ab}}(\gamma)(\mathcal{T}_i f_{S_1})^{\mathrm{ab}}(\gamma)\varphi^{\mathrm{ab}}_{i,S_0}(\gamma)\xi_i^{-1}(\gamma)
\end{aligned}$$

where ξ_i is the character of H_i' associated to λ_i'.

Since at places dividing $\mathfrak{n}_i$ we know H_i is a general linear group, we can compute

$$(\bar{\mathbf{1}}_{K^{H_i'}(\mathfrak{n})})^{\mathrm{ab}} = |\mathfrak{n}|\Gamma_1(\mathfrak{n})\bar{\mathbf{1}}_{K^{(H_i')^{\mathrm{ab}}}(\mathfrak{n})}$$

using standard formulas for $|\mathrm{SL}_n(\mathcal{O}_F/\mathfrak{p}_v^n)|$. Next, $H_i'(F)^{\mathrm{ab}} \subseteq (H')_i^{\mathrm{ab}}(F)$ so a trivial case of Lemma 8.4 of [73] applied to $(H_i')^{\mathrm{ab}}$ gives D_i and E_i such that whenever $|\mathfrak{n}| \geq D_i q_{S_1}^{E_i\kappa}$, all terms in the sum vanish except for $\gamma = 1$. This finishes the argument. □

9.4.2. *Terms with $T_i = 2$.*

Lemma 9.4.2. *When $T_i = 2$, the terms for factors $S_{(T_i,d_i,\lambda_i,\eta_i)}$ coming from the use of Corollary 7.2.2 implicit in* (9.3.5) *satisfy the bound:*

$$S^{H_i'}_{(2,d_i,\lambda_i,\eta_i)}(EP_{\lambda_i'}\bar{\mathbf{1}}_{K^{H_i'}(\mathfrak{n})}\varphi_{i,S_0}\mathcal{T}_i f_{S_1}) = O(|\mathfrak{n}|^{2d_i(d_i-1)+d_i+3}\Gamma_{-1^{(2d_i-2)},2}(\mathfrak{n})q_{S_1}^{A+B\kappa}).$$

Proof. We use Proposition 7.3.2 instead of Proposition 7.3.1. Bound $\mathrm{tr}_{\pi_{S_0}}(\varphi_{i,S_0})$ by a constant through Bernstein admissibility as in Lemma 6.3.7. Next, the Satake eigenvalues of unirreps of GL_N always have their $|\log_{q_v}(\cdot)|$ bounded by $(N-1)/2$ by the main result of [75] (this is the value achieved by the trivial representation). Therefore, arguments as in [26, §5.5] show that

$$\mathrm{tr}_{\pi_{S_1}}(\mathcal{T}_i f_{S_1}) = O(q_{S_1}^{A+B\kappa})$$

for all unirreps π_{S_1} of GL_2 at S_1 (the number of factorizable summands of $\mathcal{T} f_{S_1}$ is bounded similarly).

Finally, for any irrep π^S we have

$$\mathrm{tr}_{\pi^S}\,\bar{\mathbf{1}}_{K^{\mathrm{GL}_2}(\mathfrak{n})} \leq \Gamma_{-1}(\mathfrak{n})|\mathfrak{n}|$$

as in the proof of lemma 5.2 in [52]. Applying 7.3.2 by putting it all together (and changing A, B):

$$\begin{aligned} S^{H_i'}_{(2,d_i,\lambda_i,\eta_i)}&(EP_{\lambda_i'}\bar{\mathbf{1}}_{K^{H_i'}(\mathfrak{n})}\varphi_{i,S_0}\mathcal{T}_i f_{S_1}) \\ &= O(|\mathfrak{n}|^{d-1}\Gamma_{-1}(\mathfrak{n})^{d_i-1}q_{S_1}^{A+B\kappa}) \\ &\quad\times |\mathfrak{n}|^{2d_i(d_i-1)}\Gamma_{-1}(\mathfrak{n})^{d_i-1}S^{H_i}_{\Sigma_{\lambda_i,\eta_i}}(EP_{\lambda_i}\bar{\mathbf{1}}_{K^{H_i}(\mathfrak{n})}\varphi_{i,S_0}\mathcal{T}_i f_{S_1}). \end{aligned}$$

Where the factors on the second line outside the big-O come from taking the constant term of $\bar{\mathbf{1}}_{K^H(\mathfrak{n}_i)}$. Substituting in Theorem 9.3.2 produces the result noting that $H_i = H(2,1,\lambda_i,\eta_i) \in \mathcal{E}_{\mathrm{ell}}(2)$. □

9.4.3. *Terms with $T_i = 3$.*

Lemma 9.4.3. *When $T_i = 3$, the terms for factors $S_{(T_i,d_i,\lambda_i,\eta_i)}$ coming from the use of Corollary 7.2.2 implicit in* (9.3.5) *satisfy the bound:*

$$S^{H_i'}_{(3,d_i,\lambda_i,\eta_i)}(EP_{\lambda_i'}\bar{\mathbf{1}}_{K^{H_i'}(\mathfrak{n})}\varphi_{i,S_0}\mathcal{T}_i f_{S_1}) = O_\epsilon(|\mathfrak{n}|^{\frac{9}{2}d_i(d_i-1)+(4+\epsilon)d_i+5}\Gamma_{-1^{(d)},3}(\mathfrak{n}_i)q_{S_1}^{A+B\kappa}).$$

for all $\epsilon > 0$.

Proof. This is the same argument as Lemma 9.4.2 except we use

$$\mathrm{tr}_{\pi^S}\,\bar{\mathbf{1}}_{K^{\mathrm{GL}_3}(\mathfrak{n})} \leq C(\epsilon)|\mathfrak{n}|^{4+\epsilon}$$

from corollary 9.2 in [52]. Our $C(\epsilon)$ here is the product of Marshall-Shin's $C(\epsilon, q_v)$ for $q_v \leq q(\epsilon)$. □

9.4.4. *The full bound.* Applying the previous results with $T_i = 1, 2, 3$ instead of directly applying Proposition 7.3.3:

Corollary 9.4.4. *The bound for*

$$\Delta = (T_i, d_i, \lambda_i, \eta_i)_{1 \le i \le k} \neq \Sigma_{\lambda,\eta}$$

in Theorem 9.3.2 may be tightened to

$$S^G_\Delta(\mathrm{EP}_\lambda f^\infty_{\mathfrak{n}}) = O(|\mathfrak{n}|^{R(\Delta)} q_{S_1}^{A+B\kappa})$$

under all the same conditions and where

$$R(\Delta) = \bar{R}(\Delta) - \sum_{i:T_i=1} \left(\frac{1}{2}(d_i^2 + d_i) - 1\right) - \sum_{i:T_i=2} (4d_i - (d_i+3)) - \sum_{\substack{i:T_i=3 \\ d_i>1}} \left(9d_i - ((4 + 10^{-100})d_i + 5)\right). \tag{9.4.1}$$

Proof. For each summand of Δ with $T_i = 1$, let $H'_i = H((1, d_i, \lambda_i, \eta_i))$. Then the method of proof of Theorem 9.3.2 implicitly bounds terms

$$S^{H'_i}_{(1,d_i,\lambda_i,\eta_i)}(\mathrm{EP}_{\lambda'_i} \bar{\mathbf{1}}_{K^{H'_i}(\mathfrak{n})} \varphi'_{i,S_0} \mathcal{T}_i f_{S_1}) = O(|\mathfrak{n}|^{\frac{1}{2}(d_i^2+d_i)} \Gamma_{-1^{(d_i-1)},1^{(d_i)}}(\mathfrak{n}) q_{S_1}^{A+B\kappa}).$$

We instead use Lemma 9.4.1 to replace them by

$$S^{H'_i}_{(1,d_i,\lambda_i,\eta_i)}(EP_{\lambda'_i} f^\infty) = O(|\mathfrak{n}|\Gamma_1(\mathfrak{n}) q_{S_1}^{A+B\kappa}).$$

For summands of Δ with $T_i = 2$, we similarly use Lemma 9.4.2 to replace

$$O(|\mathfrak{n}|^{2(d_i^2+d_i)} \Gamma_{-1^{(d_i-1)},2^{(d_i)}}(\mathfrak{n}) q_{S_1}^{A+B\kappa}) \mapsto O(|\mathfrak{n}|^{2d_i(d_i-1)+d_i+3} \Gamma_{-1^{(2d_i-2)},2}(\mathfrak{n}) q_{S_1}^{A+B\kappa}).$$

For $T_i = 3$, we use 9.4.3 to replace

$$O(|\mathfrak{n}|^{\frac{9}{2}(d_i^2+d_i)} \Gamma_{-1^{(d_i-1)},3^{(d_i)}}(\mathfrak{n}) q_{S_1}^{A+B\kappa}) \mapsto O_\epsilon(|\mathfrak{n}|^{\frac{9}{2}d_i(d_i-1)+(4+10^{-100})d_i+5} \Gamma_{-1^{(d)},3}(\mathfrak{n}) q_{S_1}^{A+B\kappa}).$$

Substituting in these stronger bounds produces the result. [4] □

Remark. The $R(\Delta)$ is a better upper bound of the true growth rate than $\bar{R}(\Delta)$. It is obtained by making three modifications to the dimension count of the parabolic:

- When $T_i = 1$, replace the dimension of the Borel in the GL_{d_i}-block corresponding to that summand with 1,
- When $T_i = 2$, replace the dimension of the Levi $\mathrm{GL}_2^{d_i}$ in the GL_{2d_i}-block corresponding to that summand with $d_i + 3$,
- When $T_i = 3$, replace the dimension of the Levi $\mathrm{GL}_3^{d_i}$ in the GL_{3d_i}-block corresponding to that summand with $(4 + \epsilon)d_i + 5$.

The next section describe a case where $R(\Delta)$ is the optimal growth rate.

[4] Here, we were not very careful with the φ_{S_0} terms since we are not making claims about how the error term depends on them.

9.5. **Odd GSK Shapes.** Now that we understand $S_{\Sigma_{\lambda,\eta}}$, we can use Proposition 8.2.2 to compute the limiting asymptotics for odd Generalized Saito-Kurokawa (GSK) shapes (recall: these are shapes $\Delta = (\Delta_i)_i = (T_i, d_i, \lambda_i, \eta_i)_{1 \le i \le k}$ such that $d_1 = 1$, $T_i = 1$ for $i \ge 2$, and the d_i are odd and distinct). Keep the same setup as Section 9.1. Additionally, let Δ as above be odd GSK with $H(\Delta) = G$. Let $H_i = H(\Delta_i)$ and $\lambda'_\star$ the total infinitesimal character of $\Delta_\star$ as in (4.1.1).

We will now prove our second key technical bound, namely that $R(\Delta)$ is the exact growth rate for S^G_Δ.

Theorem 9.5.1. *Fix $f^\infty_{\mathfrak{n}}$ and $G \in \mathcal{E}_{\mathrm{ell}}(N)$ as in Section 9.1. Assume Δ is odd GSK as in Definition 8.2.1 and that the pair (Δ, φ_{S_0}) satisfies Conjecture 6.4.1.*

Then there are constants A, B, C, D, E with $C \ge 1$ depending only on G and Δ such that whenever $|\mathfrak{n}| \ge D q_{S_1}^{E\kappa}$,

$$\begin{aligned}
&|\mathfrak{n}|^{-R(\Delta)} \Gamma_{L(\Delta)}(\mathfrak{n})^{-1} S^G_\Delta(\mathrm{EP}_{\lambda'} f^\infty_{\mathfrak{n}}) \\
&\qquad = 2^{-k+1} \Lambda(H_1, \mathcal{T}_1 f_{S_1}, \varphi_{1,S_0}) \times \coprod_{i \ge 2} \Lambda^{\mathrm{ab}}(H_i, \mathcal{T}_i f_{S_1}, \varphi_{i,S_0}) + O(|\mathfrak{n}|^{-C} q_{S_1}^{A+B\kappa}),
\end{aligned}$$

with growth rate from Corollary 9.4.4

$$R(\Delta) = \frac{1}{2}\dim H_1 + (k-1) + \frac{1}{2}\left(\dim G - \sum_{i \ge 2} \dim H_i,\right),$$

indexing list

$$L(\Delta) = T_1, 1^{(k-1)}, -1^{(k-1)},$$

and masses

$$\begin{aligned}
\Lambda(H_1, \mathcal{T}_1 f_{S_1}, \varphi_{1,S_0}) &= \varphi_{1,S_0}(1)(\mathcal{T}_1 f_{S_1})(1) \frac{\dim \lambda_1}{|\Pi_{\mathrm{disc}}(\lambda_1)|} \frac{\mathrm{vol}(H_1(F)\backslash H_1(\mathbb{A}_F))}{\mathrm{vol}(K^S_{H_1}) \mathrm{vol}(H^c_{1,\infty})}, \\
\Lambda^{\mathrm{ab}}(H_i, \mathcal{T}_i f_{S_1}, \varphi_{i,S_0}) &= \frac{\mathrm{vol}(H_i(F)^{\mathrm{ab}}\backslash H_i(\mathbb{A})^{\mathrm{ab}})}{\mathrm{vol}(K^S_{H_i^{\mathrm{ab}}}) \mathrm{vol}((H_{i,\infty})^{\mathrm{ab}})} \int_{H_{i,\mathrm{der},S_0,S_1}} \mathcal{T}_i f_{S_1} \varphi_{i,S_0}(h)\, dh.
\end{aligned}$$

(Recall the $\mathcal{T}_i f_{S_1}$ and φ_{i,S_0} are defined as in Lemma 6.2.2 and Conjecture 6.4.1.)

Proof. We first apply Proposition 8.2.2 with $S_b = S_0$ and S_s the places dividing the $\mathfrak{n}$: whenever $|\mathfrak{n}|$ is big enough,

$$S^G_\Delta(\mathrm{EP}_\lambda f^\infty_{\mathfrak{n}}) = 2^{-k+1} I(\mathfrak{n}) \coprod_i S^{H_i}_{(T_i, d_i, \lambda_i, \eta_i)}(EP_{\lambda'_i} \bar{\mathbf{1}}_{K^{H_i}(\mathfrak{n})} \varphi_{i,S_0} \mathcal{T}_i f_{S_1}),$$

where we applied Lemma 9.2.1 to compute constant terms.

For $i = 1$, $d_i = 1$ so from Theorem 9.3.2 we get (for each summand in the $\coprod$):

$$\begin{aligned}
&S^{H_i}_{(T_1,1,\lambda_1,\eta_1)}(EP_{\lambda'_1} \bar{\mathbf{1}}_{K^{H_1}(\mathfrak{n})} \varphi_{1,S_0} \mathcal{T}_1 f_{S_1}) \\
&\qquad\qquad = |\mathfrak{n}|^{\dim H_1} \Gamma_{T_1}(\mathfrak{n}) \Lambda(H_1, \mathcal{T}_1 f_{S_1}, \varphi_{1,S_0}) + O(|\mathfrak{n}|^{\dim H_1 - C} q_{S_1}^{A+B\kappa}).
\end{aligned}$$

For $i > 1$, $T_i = 1$, so we apply Lemma 9.4.1 and get:

$$\begin{aligned}
&S^{H_i}_{(1,d_i,\lambda_i,\eta_i)}(EP_{\lambda'_i} f^\infty_{\mathfrak{n}}) \\
&\qquad = |\mathfrak{n}| \Gamma_1(\mathfrak{n}) \frac{\mathrm{vol}(H_i(F)^{\mathrm{ab}}\backslash H_i(\mathbb{A})^{\mathrm{ab}})}{\mathrm{vol}(K^{H_i^{\mathrm{ab}},S}) \mathrm{vol}((H_{i,\infty})^{\mathrm{ab}})} \int_{H_{i,\mathrm{der},S_1,S_0}} \mathcal{T}_i f_{S_1} \varphi_{i,S_0}(h)\, dh.
\end{aligned}$$

The result follows from multiplying and summing over factorizable summands in the $\prod\!\!\!\!\!\!\!\bigoplus$ and taking the maximum over the various A's through E's above. We use the second part of Lemma 9.2.1 to estimate $I(\mathfrak{n}_i)$ and note that the $\dim G/P$ there is $1/2(\dim G - \dim M)$. $\square$

Note. We write the scaling factor in the theorem statement as it is to emphasize the exact growth in $\mathfrak{n}$. It comes from the more conceptual formula:

$$\begin{aligned}&|\mathfrak{n}|^{-R(\Delta)}\Gamma_{L(\Delta)}(\mathfrak{n})\\&\qquad= \left(|\mathfrak{n}|^{\frac{1}{2}(\dim G - \sum_i \dim H_i)}\Gamma_{-1}(\mathfrak{n})^{k-1}\right)\left(\prod_i [K^{\mathrm{GL}_{T_i}} : K^{\mathrm{GL}_{T_i}}(\mathfrak{n})]\right).\end{aligned}$$

where the first factor come from the parabolic descent of the functions $\bar{\mathbf{1}}_{K(\mathfrak{n})}$ and the second from the expected growth rates of counts on the groups H_i.

We give two examples with $S_0 = \emptyset$, removing dependence on Conjecture 6.4.1.

Example. Consider the case when $S_1 = S_0 = \emptyset$. Then the theorem reduces to

$$\begin{aligned}(9.5.1)\quad &|\mathfrak{n}|^{-R(\Delta)}\Gamma_{L(\Delta)}(\mathfrak{n})S^G_\Delta(\mathrm{EP}_\lambda f^\infty_{\mathfrak{n}})\\&\qquad= 2^{-k+1}\frac{\dim\lambda_1}{|\Pi_{\mathrm{disc}}(\lambda_1)|}\frac{\mathrm{vol}(H(F)\backslash H(\mathbb{A}_F))}{\mathrm{vol}(K^\infty_H)\,\mathrm{vol}(H^c_\infty)} + O(|\mathfrak{n}|^{-C})\end{aligned}$$

where

$$H = H_1 \times \prod_{i>1} H_i^{\mathrm{ab}}.$$

Example. Consider the case when $S_0 = \emptyset$ and S_1 is a singleton $\{v\}$. Let $\mu^{\mathrm{pl,ur}}(H_v)$ be the Plancherel measure on the unramified spectrum of $\widehat{H}_v$ where

$$H = H_1 \times \prod_{i>1} H_i^{\mathrm{ab}}.$$

as before. Then H is dual to the group associated to Δ in (6.2.2) and comes with a map on space of Satake parameters as in (6.2.3):

$$\mathcal{S}_\Delta : \widehat{H}^{\mathrm{ur,temp}}_v \to \widehat{G}^{\mathrm{ur}}_v.$$

Define the pushforward

$$\mu^{\mathrm{pl}(\Delta),\mathrm{ur}}_v := \mu^{\mathrm{pl}(\Delta),\mathrm{ur}}(G_v) := (\mathcal{S}_\Delta)_*(\mu^{\mathrm{pl,ur}}(H_v)).$$

By (6.2.4), the factors related to f_{S_1} in the Λ's from Theorem 9.5.1 multiply to

$$(\mathcal{T}_\Delta f_v)(1) = \mu^\Delta_v(\widehat{f}_v)$$

by Fourier inversion. Substituting this into 9.5.1, we get

$$\begin{aligned}(9.5.2)\quad &|\mathfrak{n}|^{-R(\Delta)}\Gamma_{L(\Delta)}(\mathfrak{n})S^G_\Delta(\mathrm{EP}_\lambda f^\infty_{\mathfrak{n}})\\&= \left(2^{-k+1}\frac{\dim\lambda_1}{|\Pi_{\mathrm{disc}}(\lambda_1)|}\frac{\mathrm{vol}(H(F)\backslash H(\mathbb{A}_F))}{\mathrm{vol}(K^\infty_H)\,\mathrm{vol}(H^c_\infty)}\right)\mu^{\mathrm{pl}(\Delta),\mathrm{ur}}_v(\widehat{f}_v) + O(|\mathfrak{n}|^{-C}q_{S_1}^{A+B\kappa}).\end{aligned}$$

This will be related to an interpretation of our main result Theorem 11.4.1 as Plancherel equidistribution.

9.6. **General Shapes: Conjectural Optimal Bound.** Considerations of the notion of GK-dimension give us a heuristic for an optimal growth rate for any Δ.

For π_v a representation of a p-adic group G_v, the Harish-Chandra-Howe local character expansion gives an expression

$$\Theta_{\pi_v}(\exp g) = \sum_{O \in N} c_O(\pi) \widehat{\mu}_O(g)$$

where N is the set of nilpotent orbits of G_v acting on $\operatorname{Lie} G$, $\mu_O(G)$ is the Fourier transform of the δ-measure on the orbit O, $c_O(\pi)$ are constants, and $g \in \operatorname{Lie} G$ is in a small enough open compact at the identity.

Definition 9.6.1. With the notation as above, let the GK-dimension of π_v

$$d_{GK}(\pi_v) := \frac{1}{2} \max\{\dim O : c_O(\pi) \neq 0\}.$$

We can then compute:

Lemma 9.6.2. *Assume G_v is unramified. Then*

$$\dim \left(\pi_v^{K^G(q_v^n)}\right) = \operatorname{tr}_{\pi_v} \bar{\mathbf{1}}_{K^G(q_v^n)} \asymp q_v^{n d_{GK}(n)}.$$

Mœglin and Walspurger in [55] associate to each $O \in N$ a particular "degenerate Whittaker model" W_O. They prove that the maximal O such that $c_O(\pi) \neq 0$ are exactly the same as the maximal O such that $\operatorname{Hom}(\pi_v, W_O) \neq 0$.

Now specialize to $G_v = \mathrm{GL}_n(F_v)$. For π_v a tempered representation of $\mathrm{GL}_t(F_v)$ define $\pi_v[d]$ to be the Langlands quotient of the parabolic induction

$$\operatorname{Ind}_P^{G_v}(\pi_v|\det|^{(d-1)/2} \boxtimes \pi_v|\det|^{(d-3)/2} \boxtimes \cdots \boxtimes \pi_v|\det|^{-(d-1)/2}).$$

This is the local component of the Speh representations $\pi_{\tau[d]}$ in §2.2.2.

Any tempered π_v on $\mathrm{GL}_t(F_v)$ is generic and therefore satisfies

$$GK(\pi_v) = \frac{1}{2}t(t-1).$$

On the other hand, [55] compute (see [54] for a summary) that the maximal $O \in N$ such that $\operatorname{Hom}(\pi_v[d], W_O) \neq 0$ is the principal nilpotent orbit corresponding to the partition $(t^{(d)})$ through Jordan normal form. We can therefore compute

$$GK(\pi_v[d]) = \frac{1}{2}(t^2d^2 - td^2) = \frac{1}{2}d^2 t(t-1)$$

and get

$$\dim \left(\pi_v[d]^{K(q_v^n)}\right) \asymp q_v^{\frac{1}{2}t(t-1)(d^2-1)n} \dim \left(\pi_v^{K(q_v^n)}\right). \tag{9.6.1}$$

By the Ramanujan conjecture, for any simple parameter $\psi[d]$, we expect all the ψ_v to correspond to tempered representations on the GL side (this is in fact known for our case, see lemma 6.1 in [52]). Therefore we can use the heuristic (9.6.1) instead of Lemmas 9.4.1, 9.4.2,and 9.4.3 in Theorem 9.4.4 and get:

Conjecture 9.6.3. *The bound for any*

$$\Delta = (T_i, d_i, \lambda_i, \eta_i)_{1 \leq i \leq k} \neq \Sigma_{\lambda,\eta}$$

in Theorem 9.3.2 may be tightened to

$$S_\Delta^{|G|}(\mathrm{EP}_\lambda f_{\mathfrak{n}}^\infty) = O(|\mathfrak{n}|^{R_0(\Delta)} q_{S_1}^{A+B\kappa})$$

under all the same conditions and where:

$$\begin{aligned} R_0(\Delta) &:= \frac{1}{2}\left(N^2 - \sum_i T_i^2 d_i^2\right) + \sum_i \left(T_i^2 + \frac{1}{2}T_i(T_i-1)(d_i^2-1)\right) \\ &= \bar{R}(\Delta) - \sum_i \left(\frac{1}{2}T_i^2 d_i(d_i+1) - \left(T_i^2 + \frac{1}{2}T_i(T_i-1)(d_i^2-1)\right)\right). \end{aligned}$$

We think of $R_0(\Delta)$ again as making a modification to the dimension count of the the parabolic that gives $\bar{R}(\Delta)$: replace the dimension of the parabolic corresponding to partition $(d_i^{(t_i)})$ in the $\mathrm{GL}_{t_i d_i}$-block on the diagonal with $T_i^2 + \frac{1}{2}T_i(T_i-1)(d_i^2-1)$.

The main obstacle to proving Conjecture 9.6.3 is showing that the asymptotic (9.6.1) is uniform enough in π_v. This would require uniform upper bounds on the coefficients $c_O(\pi_v)$ for tempered π_v and for all O. In particular, we would need bounds on c_O for *non-maximal* O in the wavefront set of π_v so the techniques of [55] don't apply.

Of course, Conjecture 9.6.3 can only be exact for $S_\Delta^{|G|}$ instead of S_Δ^G. In the case where all the d_i have the same parity, $s_\psi = 1$ for all $\psi \in \Delta$ so these are the same. Otherwise, the terms S_ψ for different $\psi \in \Delta$ are attached to a varying sign $\epsilon_\psi(s_\psi)$. If we naïvely assume some non-trivial cancellation, we get the following:

Conjecture 9.6.4. *Recall the setup and conditions for Theorems 9.3.2, 9.4.4, and Conjecture 9.6.3. Then, if all the d_i have the same parity,*

$$C_{\epsilon,1}|\mathfrak{n}|^{R_0(\Delta)-\epsilon} \leq S_\Delta^G(\mathrm{EP}_\lambda f_\mathfrak{n}^\infty) \leq C_{\epsilon,2}|\mathfrak{n}|^{R_0(\Delta)+\epsilon}$$

for all $\epsilon > 0$ and some constants $C_{\epsilon,1}, C_{\epsilon,2}$ such that $C_{\epsilon,2} = O_\epsilon(q_{S_1}^{A+B\kappa})$.

If the d_i's have different parities, then

$$S_\Delta^G(\mathrm{EP}_\lambda f_\mathfrak{n}^\infty) = o(|\mathfrak{n}|^{R_0(\Delta)} q_{S_1}^{A+B\kappa}).$$

10. Application to Limit Multiplicities

Let G be an extended pure inner form of some $G^* \in \mathcal{E}_{\mathrm{ell}}(N)$ and fix a cohomological representation π_0 of G_∞. Fix f^∞ that is unramified outside of a finite set of finite places S. In this section, we use the stabilization of the trace formula to estimate

$$m^G(\pi_0, f^\infty) := \sum_{\pi \in \mathcal{AR}_{\mathrm{disc}}(G)} m_\pi \mathbf{1}_{\pi_\infty = \pi_0} \operatorname{tr}_{\pi^\infty}(f^\infty)$$

in terms of the bound on terms $S_{\Delta'}^H(\eta_\lambda(f^\infty)')$'s from §9.

10.1. **Preliminaries.** Let $\psi_\infty(\pi_0)$ be the Arthur parameters ψ_∞ at infinity such that $\pi_0 \in \Pi_{\psi_\infty}^G$. This is a finite (possibly empty) set since it is a subset of the set of Arthur parameters with a particular infinitesimal character. It can be seen that for each $\psi \in \psi_\infty(\pi_0)$, there is a finite number of choices of Δ such that ψ_∞ is the unique infinite component of global $\psi \in \Delta$ as in Lemma 4.3.4—intuitively, these are parameterized by ways to group together simple factors of ψ_∞ that share the same Arthur-SL_2.

In total, we get a finite (possibly empty) set of refined shapes $\Delta(\pi_0)$ such that

$$\pi_0 \otimes \pi^\infty \in \Pi_\psi \implies \psi \in \Delta \text{ for some } \Delta \in \Delta(\pi_0). \tag{10.1.1}$$

The spectral decomposition then produces

$$m^G(\pi_0, f^\infty) = \sum_{\Delta \in \Delta(\pi_0)} \sum_{\psi \in \Delta} \sum_{\pi \in \Pi_\psi} m_\pi^\psi \mathbf{1}_{\pi_\infty = \pi_0} \operatorname{tr}_{\pi^\infty}(f^\infty). \tag{10.1.2}$$

Specializing to a single $\Delta \in \Delta(\pi_0)$, consider for any test function, f_∞ on G_∞:

$$I_\Delta^G(f_\infty f^\infty) := \sum_{\psi \in \Delta} I_\psi^G(f_\infty f^\infty) = \sum_{\psi \in \Delta} \sum_{\pi \in \Pi_\psi} m_\pi^\psi \operatorname{tr}_{\pi_\infty}(f_\infty) \operatorname{tr}_{\pi^\infty}(f^\infty). \tag{10.1.3}$$

Applying Lemma 3.4.3 part (1) to (10.1.3) and comparing to (10.1.2) then gives:

Corollary 10.1.1. *Let π_d be a discrete series representation appearing in the character formula for π_0 with sign σ. Then:*

$$m^G(\pi_0, f^\infty) = \sigma \sum_{\Delta \in \Delta(\pi_0)} I_\Delta^G(\varphi_{\pi_d} f^\infty).$$

10.2. **Stabilization.** Now we compute the I_Δ^G by stabilizing. We start with a more conceptual formula:

Let $H \in \mathcal{E}_{\mathrm{ell}}(G)$. Then $H = H_1 \times H_2$ for H_i quasisplit unitary groups. Therefore, we can abuse notation and use each H_i to also denote a representative in some $\mathcal{E}_{\mathrm{ell}}(N)$ that is isomorphic as an algebraic group. Any refined shape Δ on H corresponds to a finite and possibly empty set of shapes $\Delta_1 \times \Delta_2$ on $H_1 \times H_2$ that push forward to Δ. Stabilization of each I_ψ^G (through Theorem 2.6.5) gives:

Proposition 10.2.1. *Let π_d be a discrete series representation appearing in the character formula for π_0 with sign σ. Then:*

$$m^G(\pi_0, f^\infty) = \sigma \sum_{\Delta \in \Delta(\pi_0)} \sum_{H_1 \times H_2 \in \mathcal{E}_{\mathrm{ell}}(G)} \sum_{\Delta_1 \times \Delta_2} \iota(G, H_1 \times H_2) \coprod_{i=1,2} S_{\Delta_i}^{H_i}((\varphi_{\pi_d})^{H_i} (f^\infty)^{H_i}),$$

where the $\star^{H_i}$ terms represent the corresponding factors of transfers to H.

Each $(\varphi_{\pi_d})^{H_i}$ can be chosen to be a linear combination of EP-functions by standard formulas for transfers of pseudocoefficients. Explicitly, consider a pair $(H(s), \Delta'(s))$ that pushes forward to Δ and the corresponding $s \in \mathcal{S}_\Delta$. Then,

$$\operatorname{tr}_{\psi_\infty^{\Delta'}}(\varphi_{\pi_d}^H) = \eta_{\pi_0}^{\psi_\infty^\Delta}(s' s_\Delta) = \sigma \eta_{\pi_0}^{\psi_\infty^\Delta}(s')$$

by the endoscopic character identity 2.6.2 where s' is the lift of s to $S_\Delta^\natural$ therein. For the second equality, we use $\eta_{\pi_0}^{\psi_\infty^\Delta}(s_\Delta) = \sigma$ as in the proof of part 2 of Lemma 3.4.3. If Δ' has total infinitesimal character λ, then this gives

$$\operatorname{tr}_{\psi_\infty^{\Delta'}}(\varphi_{\pi_d}^H) = \sigma \eta_{\pi_0}^{\psi_\infty^\Delta}(s') \operatorname{tr}_{\psi_\infty^\Delta}(\mathrm{EP}_\lambda),$$

since the EP-function trace is just 1. Changing notation a bit produces:

Corollary 10.2.2. *Let $G \in \mathcal{E}_{\mathrm{ell}}(N)$ and fix a cohomological representation π_0 of G_∞. Then*

$$m^G(\pi_0, f^\infty) = \sum_{\Delta \in \Delta(\pi_0)} m^G(\pi_0, \Delta, f^\infty)$$
$$:= \sum_{\Delta \in \Delta(\pi_0)} \sum_{s \in \mathcal{S}_\Delta} \iota(G, H(s)) \eta_{\pi_0}^{\psi_\infty^\Delta}(s') \sum_{\Delta_1 \times \Delta_2} \coprod_{i=1,2} S_{\Delta_i}^{H_i(s)}(\mathrm{EP}_{\lambda_i}(f^\infty)^{H_i}),$$

where $(H(s), \Delta'(s))$ is the pair of group and shape corresponding to $s \in \mathcal{S}_\Delta$ which is lifted to $s' \in S_\Delta^\natural$ as in Theorem 2.6.2, $\Delta_1 \times \Delta_2$ ranges over refined shapes on $H_1(s) \times H_2(s)$ such that the pair $(H(s), \Delta_1 \times \Delta_2)$ is equivalent to $(H(s), \Delta'(s))$, and λ_i is the total infinitesimal character of Δ_i.

Note. The sum over $\Delta_1 \times \Delta_2$ will be a singleton unless $H_1 \cong H_2$ in which case there will be two terms that differ by transposing the factors.

Note. We briefly discuss the relation between this and the formulas in [49] for transfers of pseudocoefficients. For simplicity, assume we are in a case where we never have $H_1 \cong H_2$.

Then, for each Δ there is only ever one possible choice $\Delta_1 \times \Delta_2$ due to the fixing of infinitesimal characters at infinity in shapes. The sum over EP-functions in Labesse's formulas then comes from the sum over $\Delta(\pi_0)$—these then correspond to different $\Delta_1 \times \Delta_2$ with non-conjugate infinitesimal characters.

Next, while different $\Delta \in \Delta(\pi_0)$ may have the same $\psi_\infty(\Delta)$ or even $\mathcal{S}_\Delta$, the embeddings $\mathcal{S}_\Delta \hookrightarrow \mathcal{S}_{\psi_\infty^\Delta}$ will differ. This accounts for unexpected differences in signs of coefficients of EP-functions in Labesse's formulas.

10.3. **Transfer Factors.** We eventually want to apply Proposition 10.2.2 to f^∞ as in §9.1. In the most important case where $S_0 = \emptyset$, we will need to compute explicit endoscopic transfers at all places so we need to choose explicit local transfer factors that are consistent globally.

First, pick a global Whittaker datum ω on the quasisplit form G^* of G. Since $S_0 = \emptyset$, we necessarily need that G is unramified at all finite places. Therefore $G_v^* = G_v$ for all finite v and G_v^* is in particular also unramified for all v. This implies that G^* can be defined over $\mathcal{O}_F$ so we can choose ω so that the the induced local data ω_v are unramified/admissible everywhere as in [38, §7]. This allows us to use the fundamental lemma for each G_v.

Next, [43, §4.4] shows that the choice of ω also gives us compatible local transfer factors on G itself (we note that G has simply connected derived subgroup to make the extra term in Theorem 4.4.1 disappear). At finite places G_v, the local factors stay the same as for the G_v^*.

10.4. **Limit Multiplicities.** As a preliminary/example computation, we work out what the summand $m^G(\pi_0, \Delta, f_\infty)$ from Proposition 10.2.2 is for

$$\Delta = (T_1, 1, \lambda_1, \eta_1), (1, d_2, \lambda_2, \eta_2)$$

with d_2 odd so that Δ is odd GSK. In our eventual application to unitary groups, this will be the dominant term in the sum.

Then, $\mathcal{S}_\Delta \cong \mathbb{Z}/2$ and the non-identity element s corresponds to

$$H(s) = H_1 \times H_2 = U(T_1) \times U(d_2).$$

If $T_1 \neq d_2$, there is a unique choice

$$\Delta_1 \times \Delta_2 = (T_1, 1, \lambda_1, \eta_1) \times (1, d_2, \lambda_2, \eta_2)$$

and $\iota(G, H(s)) = 1/2$. If $T_2 = d_2$, there are two choices for $\Delta_1 \times \Delta_2$ that correspond to the exact same product of S-terms and $\iota(G, H(s)) = 1/4$. Either way, 10.2.2 reduces to

$$\begin{aligned}(10.4.1)\quad m^G(\pi_0, \Delta, f_\infty) &= S^{G^*}_\Delta(\mathrm{EP}_\lambda (f^\infty)^{G^*}) \\ &+ \frac{1}{2}\eta^{\psi^\Delta_\infty}_{\pi_0}(s') S^{U(T_1)}_{(T_1,1,\lambda_1,\eta_2)}(\mathrm{EP}_{\lambda_1}(f^\infty)^{U(T_1)}) \times S^{U(d_2)}_{(1,d_2,\lambda_2,\eta_2)}(\mathrm{EP}_{\lambda_2}(f^\infty)^{U(d_2)}),\end{aligned}$$

where G^* is the quasisplit form of G and where the product implicitly includes a sum over factorizable summands of the transfer to $H(s)$.

Now assume

$$f^\infty := f^\infty_{\mathfrak{n}} = \varphi_{S_0} f_{S_1} \bar{\mathbf{1}}_{K^{G,S}(\mathfrak{n})}$$

is of the form in Section 9.1. Furthermore, assume that the chosen transfers of φ_{S_0} satisfy Conjecture 6.4.1 for Δ.

Then, by Theorem 9.5.1, the first summand has main term:

$$\begin{aligned}\frac{1}{2}|\mathfrak{n}_i|^{\frac{1}{2}(N^2+T_1^2-d_2^2)+1}\Gamma_{T_1,-1,1}(\mathfrak{n}_i)\frac{\dim\lambda_1}{|\Pi_{\mathrm{disc}}(\lambda_1)|}\frac{\mathrm{vol}(H'(F)\backslash H'(\mathbb{A}_f))}{\mathrm{vol}(K^S_{H'})\,\mathrm{vol}((H'_\infty)^c)} \\ \varphi_{1,S_0}(1) f^{H_1}_{S_1}(1)\int_{H_{2,\mathrm{der},S_0,S_1}} f^{H_2}_{S_1}\varphi_{2,S_0}(h)\,dh,\end{aligned}$$

where $H' = H_1 \times H_2^{\mathrm{ab}}$. We use here that being unramified makes $\mathcal{T}_i f_{S_1} = f^{H_i}_{S_1}$.

Next, considerations as in Lemma 6.1.2 give that the transfer of the $\bar{\mathbf{1}}_{K^{G,S}(\mathfrak{n})}$ term is a constant term to a Levi, so Lemma 9.2.1 gives

$$\bar{\mathbf{1}}^H_{K^{G,S}(\mathfrak{n})} = I(\mathfrak{n})\bar{\mathbf{1}}_{K^{H_1,S}(\mathfrak{n})} \times \bar{\mathbf{1}}_{K^{H_2,S}(\mathfrak{n})}.$$

We can therefore use Lemma 9.2.1 to get that the second summand has main term:

$$\begin{aligned}&\frac{1}{2}|\mathfrak{n}|^{\frac{1}{2}(N^2-T_1^2-d_2^2)}\Gamma_{-1}(\mathfrak{n})\eta^{\psi^\Delta_\infty}_{\pi_0}(s') \\ &\quad S^{H_1}_{(T_1,1,\lambda_1,\eta_2)}(\mathrm{EP}_{\lambda_1}\varphi^{H_1}_{S_0} f^{H_1}_{S_1}\bar{\mathbf{1}}_{K^{H_1,S}(\mathfrak{n})}) \times S^{H_2}_{(1,d_2,\lambda_2,\eta_2)}(\mathrm{EP}_{\lambda_2}\varphi^{H_2}_{S_0} f^{H_2}_{S_1}\bar{\mathbf{1}}_{K^{H_2,S}(\mathfrak{n})}).\end{aligned}$$

Theorem 9.3.2 gives that the first factor has main term

$$|\mathfrak{n}|^{T_1^2}\Gamma_{T_i}(\mathfrak{n})\varphi^{H_1}_{S_0}(1) f^{H_1}_{S_1}(1)\frac{\mathrm{vol}(H_1(F)\backslash H_1(\mathbb{A}_f))}{\mathrm{vol}(K^S_{H_1})\,\mathrm{vol}(H^c_{1,\infty})}$$

and by Proposition 8.1.2, the second factor is eventually

$$|\mathfrak{n}|\Gamma_1(\mathfrak{n})\frac{\mathrm{vol}(H_2^{\mathrm{ab}}(F)\backslash H_2^{\mathrm{ab}}(\mathbb{A}_f))}{\mathrm{vol}(K^S_{H_2^{\mathrm{ab}}})\,\mathrm{vol}(H^{\mathrm{ab}}_{2,\infty})}\int_{H_{2,\mathrm{der},S_0,S_1}} f^{H_2}_{S_1}\varphi^{H_2}_{S_0}(h)\,dh$$

after using an argument as in Corollary 9.4.4 to remove all the terms in the sum except for 1. After multiplying everything together, this shows that the summands for G and $H_1 \times H_2$ in (10.4.1) have the exact same asymptotic dependence on $\mathfrak{n}_i$.

When $S_0 = \emptyset$, we no longer need Conjecture 6.4.1 and can collect everything reasonably cleanly: for some A, B, C, D, E with $C \geq 1$, as long as $|\mathfrak{n}| \geq D q_{S_1}^{E\kappa}$:

$$
\begin{aligned}
(10.4.2)\quad & |\mathfrak{n}|^{-\frac{1}{2}(N^2+T_1^2-d_2^2)-1}\Gamma_{T_1,-1,1}(\mathfrak{n}_i)^{-1} m^G(\pi_0, \Delta, f_{S_1}\bar{\mathbf{1}}_{K^{G,S}(\mathfrak{n})}) \\
&= \mathbf{1}_{\eta_{\pi_0}^{\psi_\infty^\Delta}(\mathcal{S}_\Delta)=1} \frac{\dim \lambda_1}{|\Pi_{\mathrm{disc}}(\lambda_1)|} \frac{\mathrm{vol}(H'(F)\backslash H'(\mathbb{A}_f))}{\mathrm{vol}(K_{H'}^S)\,\mathrm{vol}((H'_\infty)^c)} \left(f_{S_1}^{H_1}(1) \int_{H_{2,\mathrm{der},S_1}} f_{S_1}^{H_2}(h)\, dh \right) \\
&\qquad + O(|\mathfrak{n}|^{-C} q_{S_1}^{A+B\kappa}).
\end{aligned}
$$

If $S_0 = \emptyset$, then G is unramified at all finite places so we necessarily have that $\eta_{\pi_0}^{\psi_\infty^\Delta}$ factors through $\mathcal{S}_\Delta$ because of the conditions on the χ_{G_v} in Theorem 2.5.1 to glue together to a global group as in Lemma 2.1.1.

The case when $\eta_{\pi_0}^{\psi_\infty^\Delta}(\mathcal{S}_\Delta)$ isn't trivial can actually be understood more simply. If $S_0 = \emptyset$, then each factor f_v of the test function is either on $\mathrm{GL}_{N,v}$ or unramified. Therefore, its traces vanish on all $\pi_v \in \Pi_{\psi_v}$ such that $\eta_{\pi_v}^{\psi_v} \neq 1$. By the multiplicity formula 2.5.4, this implies that

$$m_\pi^\psi \operatorname{tr}_{\pi^\infty}(f^\infty) = 0$$

whenever $\eta_{\pi_\infty}^{\psi_\infty} \neq 1$. In total,

$$\eta_{\pi_0}^{\psi_\infty^\Delta}(\mathcal{S}_\Delta) \neq 1 \implies m^G(\pi_0, \Delta, f^\infty) = 0.$$

An extension of this argument to all odd GSK shapes gives:

Theorem 10.4.1. *Let G be an extended pure inner form of $G \in \mathcal{E}_{\mathrm{ell}}(N)$ that is unramified at all finite places and let π_0 be a cohomological representation of G_∞. Choose odd GSK $\Delta = (T_i, d_i, \lambda_i, \eta_i)_{1 \leq i \leq k} \in \Delta(\pi_0)$ and let $f_\mathfrak{n}^\infty = f_{S_1}\bar{\mathbf{1}}_{K^{G,S}(\mathfrak{n})}$ be as in Section 9.1 with $S_0 = \emptyset$.*

Then, if $\eta_{\pi_0}^{\psi_\infty^\Delta}(\mathcal{S}_\Delta) \neq 1$

$$m^G(\pi_0, \Delta, f_\mathfrak{n}^\infty) = 0.$$

Otherwise, there are A, B, C, D, E with $C \geq 1$ such that as long as $|\mathfrak{n}| \geq D q_{S_1}^{E\kappa}$:

$$
\begin{aligned}
& |\mathfrak{n}|^{R(\Delta)}\Gamma_{L(\Delta)}(\mathfrak{n})^{-1} m^G(\pi_0, \Delta, f_\mathfrak{n}^\infty) \\
&\quad = \frac{\dim \lambda_1}{|\Pi_{\mathrm{disc}}(\lambda_1)|} \frac{\mathrm{vol}(H'(F)\backslash H'(\mathbb{A}_f))}{\mathrm{vol}(K_{H'}^S)\,\mathrm{vol}((H'_\infty)^c)} \left(f_{S_1}^{H_1}(1) \coprod_{i>1} \int_{H_{i,\mathrm{der},S_1}} f_{S_1}^{H_i}(h)\, dh \right) \\
&\qquad + O(|\mathfrak{n}|^{-C} q_{S_1}^{A+B\kappa})
\end{aligned}
$$

where $H_i = H(T_i, d_i, \lambda_i, \eta_i)$, $H' = H_1 \times \prod_{i>1} H_i^{\mathrm{ab}}$, and $R(\Delta)$ is as in (9.4.1).

For our upper bound we can allow $S_0 \neq \emptyset$:

Theorem 10.4.2. *Let G be an extended pure inner form of $G \in \mathcal{E}_{\mathrm{ell}}(N)$ that may or may not be unramified and let π_0 be a cohomological representation of G_∞. Let $\Delta \in \Delta(\pi_0)$ and let $f_\mathfrak{n}^\infty = \varphi_{S_0} f_{S_1}\bar{\mathbf{1}}_{K^{G,S}(\mathfrak{n})}$ be as in Section 9.1.*

Then there are A, B, D, E such that as long as $|\mathfrak{n}| \geq D q_{S_1}^{E\kappa}$

$$m^G(\pi_0, \Delta, f_\mathfrak{n}^\infty) = O(|\mathfrak{n}|^{R(\Delta)} q_{S_1}^{A+B\kappa})$$

where $R(\Delta)$ is as in Corollary 9.4.4.

Proof. Apply Proposition 10.2.2 and then 9.4.4 to each term. $\square$

Note that Theorem 10.4.2 does *not* depend on Conjecture 6.4.1 since Corollary 9.4.4 doesn't.

11. Explicit Computations on Unitary Groups

In this section, we recall the explicit combinatorial parameterization of cohomological representations of $U(p,q)$ and their A-packets and Adams-Johnson parameters following [56]; see also [78, 77, 15]. This allows us to compute the sets $\Delta(\pi_0)$ and work out explicit limit multiplicity statements from Proposition 10.2.1 together with the bounds in Theorems 10.4.1 and 10.4.2.

11.1. Cohomological Representations of $U(p,q)$.

11.1.1. *Setup:* We recall some general facts about the construction of cohomological representations. Let G be a reductive group with Lie algebra $\mathfrak{g}_0$ and $\mathfrak{g} = \mathfrak{g}_0 \otimes \mathbb{C}$. Let $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ be the Cartan decomposition corresponding to a choice of maximal compact subgroup K with Cartan involution ι[5]. Assume that G has a compact Cartan subgroup T with Lie algebra $\mathfrak{t} \subset \mathfrak{k}$. Let $\Delta(\mathfrak{t}, \mathfrak{g})$ be the root system for $\mathfrak{t}$ in $\mathfrak{g}$.

In [78], Vogan-Zuckerman introduce the notion of a ι-stable parabolic subalgebras $\mathfrak{q}$ of $\mathfrak{g}$, henceforth referred to as VZ subalgebras. To construct such $\mathfrak{q}$, choose $x \in i\mathfrak{t}_0$ and define $\lambda_0 \in \mathfrak{t}^*$ by $[x, \mathfrak{g}^\alpha] = \langle \lambda_0, \alpha \rangle \mathfrak{g}^\alpha$ for $\alpha \in \Delta(\mathfrak{g}, \mathfrak{t})$. Then

$$\mathfrak{q} := \mathfrak{l} \oplus \mathfrak{u}, \quad \mathfrak{l} := \mathfrak{t} \oplus \bigoplus_{\langle \lambda_0, \alpha \rangle = 0} \mathfrak{g}^\alpha, \quad \mathfrak{u} := \oplus \bigoplus_{\langle \lambda_0, \alpha \rangle \geq 0} \mathfrak{g}^\alpha. \tag{11.1.1}$$

Let $L = Z_G(\lambda_0)$ be the Levi subgroup of G with Lie algebra $\mathfrak{l}$. Let $\Delta(\mathfrak{t}, \mathfrak{u})$ be the roots of $\mathfrak{t}$ in $\mathfrak{u}$, and let $\lambda : \mathfrak{l} \to \mathbb{C}$ be a character such that:

(i) λ is the differential of a one-dimensional representation of L, and
(ii) $\langle \alpha, \lambda \rangle \geq 0$ for $\alpha \in \Delta(\mathfrak{t}, \mathfrak{u})$.

Then Vogan-Zuckerman define a representation $A_\mathfrak{q}(\lambda)$ and show that if F_λ is the irreducible finite-dimensional G-representation of highest weight λ, then $A_\mathfrak{q}(\lambda)$ and F_λ have the same infinitesimal character $\lambda + \rho_G$ and

$$H^*(\mathfrak{g}, K; A_\mathfrak{q}(\lambda) \otimes F_\lambda) \neq 0.$$

Moreover, any representation with nontrivial $(\mathfrak{g}, K)$-cohomology is isomorphic to $A_\mathfrak{q}(\lambda)$ for some pair $(\mathfrak{q}, \lambda)$.

11.1.2. *The Parameterization.* The parameterization of cohomological representations of $U(p,q)$ will be given in terms of the following combinatorial data:

Definition 11.1.1. For a pair of non-negative integers p, q with $p + q = N$, let:

- $\mathcal{P}(N)$ to be the set of *ordered* partitions of N, i.e. tuples $(N_1, ..., N_r)$ where r is arbitrary, each N_i is positive, and $\sum_i N_i = N$.
- $\mathcal{P}(p,q)$ be the set of ordered bipartitions of (p,q), i.e. the set of tuples of pairs $((p_1,q_1), ..., (p_r,q_r))$ where the p_i, q_i are a non-negative integers with $\max p_i, q_i > 0$ and $\sum_i p_i = p$, $\sum_i q_i = q$.
- $\mathcal{P}_1(p,q) \subset \mathcal{P}(p,q)$ be the subset consisting of expressions $((p_1,q_1), ..., (p_r,q_r))$ where if $p_i q_i = 0$, then $\max(p_i, q_i) = 1$.

[5] In lieu of the traditional notation θ, which we reserve for the involution defining our unitary group.

There are natural surjective maps:

$$\beta : \mathcal{P}(p,q) \to \mathcal{P}(N), \quad ((p_1,q_1),...,(p_r,q_r)) \mapsto (p_1+q_1,...,p_r+q_r), \tag{11.1.2}$$

$$\gamma : \mathcal{P}(p,q) \to \mathcal{P}_1(p,q) \tag{11.1.3}$$

where γ replaces any term of the form $(n,0)$ (resp. $(0,m)$) by n copies of $(1,0)$ (resp. m copies of $(0,1)$.)

The bipartitions $B \in \mathcal{P}(p,q)$ parameterize VZ subalgebras following [77, §1-3], who proves:

Proposition 11.1.2. *Let $G = U(p,q)$.*

(i) *The K-conjugacy classes of VZ subalgebras of $\mathfrak{g}$ are in bijection with $\mathcal{P}(p,q)$.*
(ii) *Let $\mathfrak{q}$, $\mathfrak{q}'$ be VZ subalgebras corresponding to $B_{\mathfrak{q}}$, $B_{\mathfrak{q}'} \in \mathcal{P}(p,q)$. Then $A_{\mathfrak{q}}(0) \simeq A_{\mathfrak{q}'}(0)$ if and only if $\gamma(B_{\mathfrak{q}}) = \gamma(B_{\mathfrak{q}'})$.*

We write the Levi subgroup L_B associated to $B \in \mathcal{P}(p,q)$ as

$$L_B = U(p_1,q_1) \times ... \times U(p_r,q_r).$$

To realize the bijection, embed $K = U(p,0) \times U(0,q)$ in $U(p,q)$, write

$$\mathfrak{t}_0 = \mathfrak{t}_0 \cap \mathfrak{u}_0(p,0) \oplus \mathfrak{t}_0 \cap \mathfrak{u}_0(0,q) \simeq \mathbb{R}^p \times \mathbb{R}^q,$$

and associate to $B = ((p_1,q_1),...,(p_r,q_r))$ the algebra constructed from the element ix_B where

$$x_B = (\overbrace{r,...,r}^{p_1}, \overbrace{r-1,...,r-1}^{p_2}, ..., \overbrace{1,...,1}^{p_r}, \overbrace{r,...,r}^{q_1}, ..., \overbrace{1,...,1}^{q_r}) \in \mathfrak{t}_0 \subset \mathfrak{k}_0.$$

Let λ be an infinitesimal character. Recall the notion of P-parts of λ from §1.6.3.

Definition 11.1.3. We say that a regular integral infinitesimal character

$$\lambda = \xi_1 > \cdots > \xi_n$$

is adapted to partition P if the P-parts of λ are all of the form

$$X^r \sum_{i=1}^{n} X^{(n-2i+1)/2}$$

for some integer or half-integer r and some integer n.

For example, the infinitesimal character ρ_G of the trivial representation is adapted to all partitions of N. The following is also deduced from [77]:

Proposition 11.1.4. *Let $G = U(p,q)$. The cohomological representations with regular integral infinitesimal character λ are all of the form $A_{\mathfrak{q}}(\lambda - \rho_G)$ with $\mathfrak{q}$ corresponding to bipartition $B \in \mathcal{P}(p,q)$ such that λ is adapted to $\beta(B)$. In particular, they are in bijection with the bipartitions $B \in \mathcal{P}_1(p,q)$ with λ adapted to $\beta(B)$.*

To compute the cohomology associated to the representations, one makes a choice of complex structure on $\mathfrak{p}$ i.e. on the quotient G/K. To do this, fix a Shimura datum for G to be the conjugacy class of

$$h_K : \mathbb{S} \to U(p,q), \quad h_K(z) = \left(\frac{z}{\bar{z}} I_p, I_q\right) \in U(p,0) \times U(0,q) \subset U(p,q).$$

This induces a decomposition $\mathfrak{p} = \mathfrak{p}^+ \oplus \mathfrak{p}^-$ where $\mathrm{Ad}(h(z))$ acts on $\mathfrak{p}^+$ by $z/\bar{z}$ and on $\mathfrak{p}^-$ by its inverse.

Lemma 11.1.5. *Let $G = U(p,q)$ with a choice of Shimura datum as above. Let F_λ be a finite-dimensional representation with highest weight λ. Let $B = ((p_1,q_1),...,(p_r,q_r)) \in \mathcal{P}(p,q)$ be a partition such that $\lambda+\rho_G$ is adapted to $\beta(B)$, and let*

$$R = pq - \sum_i p_i q_i, \quad R^+ = \sum_{i<j} p_i q_j, \quad R^- = \sum_{i>j} p_i q_j.$$

Then:

(i) the smallest value of i such that $H^i(\mathfrak{g},K;A_{\mathfrak{q}_B}(\lambda)\otimes F_\lambda) \neq 0$ is $i = R$;
(ii) $H^j(\mathfrak{g},K;A_{\mathfrak{q}_B}(\lambda)\otimes F_\lambda) \neq 0$ if and only if $j = R+2p$ for $0 \le p \le \sum_i p_i q_i$;
(iii) the Hodge weights of $A_{\mathfrak{q}_B}(\lambda)$ in degree $R+2p$ are $(R^+ + p, R^- + p)$.

Proof. This follows from [78, Thm. 3.3, 5.4, 6.19]: the first nonzero degree of cohomology of $A_{\mathfrak{q}}(\lambda)$ is $R = \dim \mathfrak{u}\cap\mathfrak{p} = \dim\mathfrak{p} - \dim \mathfrak{l}\cap\mathfrak{p} = pq - \sum_i p_i q_i$. More precisely, the cohomology of $A_{\mathfrak{q}_B}(\lambda)$ in degree $\dim \mathfrak{u}\cap\mathfrak{p}$ appears in weight $(R^+,R^-) = (\mathfrak{u}\cap\mathfrak{p}^+, \mathfrak{u}\cap\mathfrak{p}^-)$. All other weights (a,b) for which $A_{\mathfrak{q}}(\lambda)$ has cohomology are of the form $(R^+ + p, R^- + p)$. One computes R^+ and R^- from x_B and the choice of Shimura datum; see e.g. [15, §5] for more details. □

11.1.3. *Packets of cohomological representations.* Recall form §3.2 that any parameter associated to a packet of cohomological representations is an Adams-Johnson parameter. In particular, there is a Levi subgroup

$$\widehat{L} \simeq GL_{n_1} \times ... \times GL_{n_r} \subset \widehat{G}$$

such that $\psi(W_{\mathbb{C}} \times I) \subset Z(\widehat{L})$, and such that $\psi(1 \times SL_2)$ is a principal SL_2 in $\widehat{L}$. Thus we have

$$\psi\,|_{W_{\mathbb{C}}\times SL_2} = \bigoplus_{i=1}^r \chi_{t_i} \otimes [n_i]$$

where $[n_i]$ is the irreducible n_i-dimensional representation of $SL_2(\mathbb{C})$ and

$$\chi_{t_i}(z) = \left(\frac{z}{\bar{z}}\right)^{t_i/2}, \quad z \in W_{\mathbb{C}} \simeq \mathbb{C}^\times.$$

In this case, the total infinitesimal character of ψ is

$$\lambda_\psi = \sum_{i=1}^r X^{\frac{t_i}{2}} \sum_{j=1}^{n_i} X^{\frac{n_i-2j+1}{2}}.$$

Since we assume that λ_ψ is the infinitesimal character of a finite-dimensional representation, we have for each i that $t_i - n_i \equiv N \mod 2$, i.e. we are in the good parity case in the sense of [56]. To a good parity parameter, Mœglin-Renard attach the ordered partition $P = (n_1,...,n_r) \in \mathcal{P}(N)$: specifically, the unordered multiset of n_i come from restricting to the Arthur-SL_2 and the ordering is such that the infinitesimal character of the summand $\chi_{t_i}\otimes[n_i]$ of ψ is the ith P-part of λ_ψ (in particular, P is adapted to λ_ψ). They also show that the corresponding packet is

$$\Pi_\psi = \{A_{\mathfrak{q}_B}(\lambda_B) \mid \beta(B) = P\},$$

where $\lambda_B = \boxtimes_i \det^{(t_i+a_i-N)/2-a_{<i}}$ for $a_i = \sum_{j<i} a_j$. In short, the representations in the packet π_ψ are in bijection with the bipartitions $((p_1,q_1),...,(p_r,q_r)) \in \mathcal{P}(p,q)$ such that $p_i + q_i = n_i$.

The article [56] does more. First they fix a Whittaker datum for a choice of quasisplit form G^* of G (see [56, Remarque 4.5]) which without loss of generality we can have match the one from Section 10.3. Then, they write down explicitly the character $\eta_\pi : S_\psi^\natural \to \pm 1$ attached by Arthur to each representation in the packet Π_ψ in terms of the bipartition B such that $\pi = \pi_B$. These constructions are recalled in §11.3, where they are used.

11.2. **Understanding $\Delta(\pi_0)$ and $R(\Delta)$.** Fix an extended pure inner form G of $G^* \in \mathcal{E}_{\mathrm{sim}}(N)$. Let

$$\pi_0 = \bigotimes_{v \in \infty} \pi_{0,v}$$

be a cohomological representation of

$$G_\infty = \prod_{v \in \infty} U(p_v, q_v).$$

We will study the set $\Delta(\pi_0)$ introduced in §10.1: for each $\Delta \in \Delta(\pi_0)$, we compute the invariant $R(\Delta)$ from Theorem 9.4.4 and the set of shapes that realize it.

Definition 11.2.1. Let $\Delta^{\max}(\pi_0)$ be the set of $\Delta \in \Delta(\pi_0)$ with maximal $R(\Delta)$.

Definition 11.2.2. Let $R(\pi_0)$ be the common value of $R(\Delta)$ for $\Delta \in \Delta^{\max}(\pi_0)$.

11.2.1. *Ignoring η_i.* First, since each $\Delta = (T_i, d_i, \lambda_i, \eta_i)_i \in \Delta(\pi_0)$ satisfies $H(\Delta) = G^*$, the η_i are completely determined by T_i and d_i according to Section 2.2.4. Therefore, their information is redundant so we will ignore it in this section.

11.2.2. *Arthur-*SL_2*'s.* We first study Δ according to their Arthur-SL_2. These N-dimensional representations of SL_2 correspond to unordered partitions of N via their decomposition into irreducibles. It is easy to see that:

Lemma 11.2.3. *Among the shapes Δ with Arthur-*SL_2 *given by unordered partition*

$$Q = (a_1^{(r_1)}, \dots, a_k^{(r_k)})$$

with a_i distinct, the value of the invariant $R(\Delta)$ introduced in Theorem 9.4.4 is maximized for shapes $\Delta = (r_i, a_i, \lambda_i, \eta_i)_{1 \le i \le k}$, in which each distinct integer a_j appears once. Denote by $R(Q)$ this maximized value of $R(\Delta)$.

Furthermore, for cohomological representations π_0, if there is $\Delta \in \Delta(\pi_0)$ whose Arthur-SL_2 *matches Q, then there is also $\Delta' \in \Delta(\pi_0)$ such that $R(\Delta') = R(Q)$.*

Proof. Constructing any other shape with the same Arthur-SL_2 would require splitting up some $(r_i, a_i, \lambda_i, \eta_i)$ into smaller blocks, which would decrease $R(\Delta)$.

For the second part, we may always merge blocks in Δ with the same a_i by concatenating their infinitesimal characters and leaving ψ_Δ^∞ unchanged by the construction in Lemma 4.3.4. ☐

This reduces the study of $R(\pi_0)$ to understanding the Arthur-SL_2's for $\Delta \in \Delta(\pi_0)$:

Definition 11.2.4. Let $Q^{\max}(\pi_0)$ be the unordered partitions representing the Arthur-SL_2's of $\Delta \in \Delta(\pi_0)$ with maximal $R(\Delta)$. Equivalently by Lemma 11.2.3, it is the set of Arthur-SL_2's of $\Delta \in \Delta^{\max}(\pi_0)$.

By Lemma 4.3.4, the possible Arthur-SL_2's for $\Delta \in \Delta(\pi_0)$ are the possible Arthur-SL_2's for ψ_∞ with $\pi_0 \in \Pi_{\psi_\infty}$. We can therefore enumerate them by our classification of cohomological representations.

Fix a place v and let $\pi_{0,v}$ correspond to $B_v = (p_{i,v}, q_{i,v})_i \in \mathcal{P}_1(p_v, q_v)$ and infinitesimal character λ_v. We next study $\Delta(\pi_{0,v})$: the union of $\Delta(\pi')$ over all

π' with $\pi'_v = \pi_{0,v}$. Recall from (11.1.2) that $\beta(B_v)$ is the ordered partition of N associated to B_v. We define some combinatorial objects:

- $\beta_+(B)$ is the unordered subpartition of $\beta(B)$ corresponding to parts with size bigger than 1.
- $Q_p(\pi_0)$ is the unordered partition $(n_j)_{j\in J}$ where the j correspond to runs of length n_j of consecutive (p_i, q_i) of the form $(1,0)$ such that the corresponding piece of λ is of the form
$$X^r \sum_{i=1}^{n_j} X^{(n-2i+1)/2}.$$
- $Q_q(\pi_0)$ is similarly defined for parts of the form $(0,1)$.

Next, if $Q_1 = (n_i)_{i\in I}$ and $Q_2 = (m_j)_{j\in J}$ are two unordered partitions, then we say that Q_2 refines Q_1 if there is a map $J \to I$ such that the sum of m_j over the fiber at i is n_i.

Lemma 11.2.5. *The possible Arthur-*SL_2*'s for* $\Delta \in \Delta(\pi_{0,v})$ *correspond exactly to unordered partitions*
$$(X, Y, \beta_+(B_v)),$$
where X *refines* $Q_p(\pi_{0,v})$ *and* Y *refines* $Q_q(\pi_{0,v})$*.*

Proof. Lemma 11.1.5 tells us that the $P \in \mathcal{P}(p+q)$ that correspond to ψ_∞ with $\pi_0 \in \Pi_{\psi_\infty}$ are produced by merging runs of consecutive 1's in $\beta(B_v)$ that correspond to parts in B_v all of the form $(1,0)$ or all of the form $(0,1)$. Furthermore, we require that the coarsened partition thereby produced is still adapted to λ_v. These conditions together show that P is an ordering of something of the form $(X, Y, \beta_+(B_v))$.

Next, consider
$$P = (n_1, \dots, n_k)$$
of the claimed form. We will show that there is $\Delta \in \Delta(\pi_{0,v})$ with ψ_v^Δ corresponding to P. Let I_1 be subset of indices such that $n_i = 1$ and I_+ its complement. Let λ_v have P-parts $\lambda_1^P, \dots, \lambda_l^P$. Since λ_v is adapted to P, for each $i \in I_+$, there exists $\lambda_{i,v}$ such that shape $(1, n_i, \lambda_{i,v})$ has total infinitesimal character λ_i^P at v. Next, let λ'_v be the concatenation of λ_i^P for $i \in I_1$. Finally, choose the other components for $w \neq v$ of λ_i and λ' arbitrarily. Consider
$$\Delta = (|I_1|, 1, \lambda', \eta), ((1, n_i, \lambda_i, \eta_i))_{i\in I_+}.$$
Then, by the constructions in §11.1.3, ψ_v^Δ corresponds to P. Note that $\Delta \in \Delta(\pi_{0,v})$ by the form of P. ☐

Next, we define
$$\beta_+(\pi_0) := \bigcup_{v\in\infty} \beta_+(B_v)$$
where the union is interpreted as of non-disjoint multisets (the multiplicity of an element is exactly equal to the maximum of its multiplicities in the $\beta_+(B_v)$).

Lemma 11.2.6. *The possible Arthur-*SL_2*'s for* $\Delta \in \Delta(\pi_0)$ *correspond exactly to unordered partitions* Q *such that, for each* $v \in \infty$*,* Q *can be written in the form:*
$$Q = (X_v, Y_v, \beta_+(B_v))$$
where X_v *refines* $Q_p(\pi_{0,v})$ *and* Y_v *refines* $Q_q(\pi_{0,v})$

In particular:

- *All Arthur-*SL_2*'s for* $\Delta \in \Delta(\pi_0)$ *contain* $\beta_+(\pi_0)$ *as a subpartition,*
- *if* $\Delta(\pi_0)$ *isn't empty, there is* $\Delta \in \Delta(\pi_0)$ *with Arthur-*SL_2

$$(1, \dots, 1, \beta_+(\pi_0)).$$

Proof. That the Arthur-SL_2's must be contained in this set is an elementary combinatorial extension of the argument in 11.2.5.

Existence of a $\Delta = (T_i, d_i, \lambda_i)_i$ with a particular SL_2 follows from fixing each component of the λ_i as in the argument of Lemma 11.2.5 instead of just the v-component.

Finally, the two bullet points are also elementary combinatorial properties of the set of such simultaneous $(X_v, Y_v, \beta_+(B_v))$. □

Note. We summarize this section as a three step algorithm for computing $\Delta^{\max}(\pi_0)$:

(1) Find the possible Arthur-SL_2's for $\Delta \in \Delta(\pi_0)$ by Lemma 11.2.6.
(2) Compute $R(Q)$ for each of these Arthur-SL_2's to compute $Q^{\max}(\pi_0)$.
(3) $\Delta^{\max}(\pi_0)$ is then partitioned into non-empty parts corresponding to the $Q \in Q^{\max}(\pi_0)$. Each part can be determined by Lemma 11.2.3.

It turns out that the second step becomes much easier for GSK-shapes. Showing this will make up the remainder of our combinatorial work.

11.2.3. *The Key Bound.* We next prove an elementary combinatorial bound that is basically a reformulation of Lemma 7.1 in [52]. Recall the numerical invariants $\bar{R}(\Delta)$ and $R(\Delta)$, introduced respectively in (9.3.4) and (9.4.1); $R(Q)$ is the maximum of $R(\Delta)$ over the Δ such that ψ_∞^Δ corresponds to Q. Part of the complexity of this argument is an artifact of only being able to prove the suboptimal bound $R(\Delta)$ from Corollary 9.4.4 instead of the optimal $R_0(\Delta)$ from Conjecture 9.6.3.

Lemma 11.2.7. *Let* Q_0 *be an unordered partition that has distinct parts and no parts of size* 1*. Then the maximum value of* $R(Q)$ *over all unordered partitions* Q *of* N *that have subpartition* Q_0 *is achieved by*

$$Q_{\mathrm{can}} = (1^{(r)}, Q_0).$$

Furthermore, if either $r \neq 2$ *or* Q_0 *has no parts of size* 2*, this is the unique such* Q *that achieves this maximum.*

Proof. Let

$$Q_{\mathrm{can}} = (1, \dots, 1, Q_0) = (1^{(r)}, (a_i^{(r_i)})_i)$$

for a_i distinct and $r_i = 1$. The other possible Q containing Q_0 are produced by decreasing the number of 1's and increasing one of the r_i's. We will therefore show that $R(Q)$ decreases if we increase any of the r_i.

Recall from Remark 9.4.4 that $R(Q)$ is obtained by starting from $\bar{R}(Q)$, equal to the dimension of a certain parabolic block matrix, and then replacing the dimensions of certain blocks on the diagonal with modified counts. Along the diagonal, the blocks are indexed by the number 1 and the distinct a_i. Changing an r_i changes three blocks of the parabolic: the two on the diagonal associated to 1 and a_i and one corresponding to this pair above the diagonal. In particular, we can treat changes in each r_i independently. Furthermore, changing an r_i from 0 (i.e. creating a new block) can be easily seen to decrease $R(Q)$.

In general, going from $r_i = 1$ to $r_i = k$ changes the modified summand associated to these three blocks from:

$$r^2 + ra_i + 1 \mapsto$$
$$(r - (k-1)a_i)^2 + ka_i(r - (k-1)a_i) + (\text{modified count for } a_i \text{ with } r_i = k).$$

Expanding out, the change in $R(Q)$ is

$$(\text{modified count for } a_i \text{ with } r_i = k) - a_i(a_i + r)(k-1) - 1,$$

where the modified counts are the counts for the part of formula for $R(Q)$ associated to the a_i-block on the diagonal. The modified count defining $R(Q)$ is defined differently for $r_i = 1, 2, 3$ versus everything else. Therefore we have to look at cases:

- If an r_i is increased to 2, the modified count is
$$2a_i(a_i - 1) + (a_i + 3)$$
making the total difference
$$2a_i(a_i - 1) + (a_i + 3) - a_i(a_i + r) - 1.$$
Using $r \geq a_i$, this is bounded above by
$$-a_i + 2$$
and is always negative unless $r = a_i = 2$.
- If an r_i is increased to 3, the modified count is
$$\frac{9}{2}a_i(a_i - 1) + ((4 + \epsilon)a_i + 5)$$
making the total difference
$$\frac{9}{2}a_i(a_i - 1) + ((4 + \epsilon)a_i + 5) - 2a_i(a_i + r) - 1.$$
Using $r \geq 2a_i$, this is bounded above by
$$-\frac{3}{2}a_i^2 - \left(\frac{1}{2} - \epsilon\right) a_i + 4,$$
which is always negative for in particular $\epsilon < 1/2$ since $a_i \geq 2$.
- If an r_i is increased to $k \geq 4$, then the change in modified counts is
$$\frac{k^2}{2}a_i(a_i + 1),$$
making the total difference
$$\frac{k^2}{2}a_i(a_i + 1) - a_i(a_i + r)(k-1) - 1.$$
Using $r \geq (k-1)a_i$, this is bounded above by
$$-\left(\frac{k^2}{2} - k\right) a_i^2 + \frac{k^2}{2}a_i - 1$$
and using $a_i \geq 2$ and $k^2/2 - k > 0$ gives an upper bound by
$$-\left(\frac{k^2}{2} - 2k\right) a_i - 1,$$
which is always negative when $k \geq 4$.

In total, if we increase any of the r_i, then $R(Q)$ decreases. □

11.2.4. *Summary of combinatorial work.* To conclude:

Definition 11.2.8. Let π_0 factor into places $\pi_{0,v}$ for each v. Let each $\pi_{0,v}$ correspond to $B_v = (p_{i,v}, q_{i,v})_v \in \mathcal{P}_1(p_v, q_v)$. Let $\beta_+(B_v)$ be the unordered subpartiton of pieces of size greater than 1 in $\beta(B_v)$. Let $\beta_+(\pi_0)$ be the union of all the $\beta_+(B_v)$ as multisets. Finally, define

$$Q_{\mathrm{can}}(\pi_0) := (1, \dots, 1, \beta_+(\pi_0))$$

as an unordered partition of N (if it exists).

Then:

Proposition 11.2.9. *Let $Q_{\mathrm{can}}(\pi_0)$ be of the form*

$$Q_{\mathrm{can}}(\pi_0) = (1^{(r)}, a_1, \dots, a_k)$$

for the a_i distinct (i.e. $\beta_+(\pi_0)$ has distinct parts). Further assume that either $r \neq 2$ or there is no i such that $a_i = 2$. Then $\Delta^{\max}(\pi_0)$ consists of all shapes of the form

$$(r, 1, \lambda), (1, a_i, \lambda_i)_{1 \le i \le k}$$

in $\Delta(\pi_0)$ and $\Delta^{\max}(\pi_0) \neq \emptyset$ provided that $\Delta(\pi_0) \neq \emptyset$.

Proof. This is the result of applying algorithm 11.2.2 keeping Lemma 11.2.7 in mind for step (2) to get that $Q^{\max}(\pi_0) = \{Q_{\mathrm{can}}(\pi_0)\}$. □

We warn that Proposition 11.2.9 does not neccesarily hold if $\beta_+(\pi_0)$ has repeated elements. For example, consider $\beta_+(\pi_0) = (2, 2, 2, 2)$. Then we have

$$R(2^{(4)}, 1, 1) = 67 < R(2^{(5)}) = 74.$$

In fact, this is a counterexample to even the analogous statement with the conjectural optimal bound R_0. Whether any $\Delta \in \Delta(\pi_0)$ can have Arthur-SL_2 given by $(2^{(5)})$ depends on what exactly the $Q_p(\pi_{0,v})$ and $Q_q(\pi_{0,v})$ are. Therefore, a general description of $Q^{\max}(\pi_0)$ is much more complicated.

11.2.5. *GSK-maxed representations.* Now we restrict to the special class of representations we can study:

Definition 11.2.10. Let π_0 be a cohomological representation of some $U(p, q)$ corresponding to bipartition $(p_i, q_i)_i \in \mathcal{P}_1(p + q)$. We say π_0 is GSK-maxed if the only value of $p_i + q_i$ that appears with multiplicity is 1.

We say π_0 is odd GSK-maxed if in addition all the $p_i + q_i$ are odd.

Definition 11.2.11. Let π_0 factor into $\pi_{0,v}$ that each corresponding to bipartition $(p_{i,v}, q_{i,v})$. Then we say that π_0 is GSK-maxed if $\Delta(\pi_0) \neq \emptyset$ and the only number that appears with multiplicity among the $p_{i,v} + q_{i,v}$ is 1.

We say π_0 is odd GSK-maxed if in addition all the $p_{i,v} + q_{i,v}$ are odd. Equivalently, $\beta_+(\pi_0)$ is a partition into (odd) distinct parts.

These definitions are justified by the following corollary of Proposition 11.2.9:

Corollary 11.2.12. *Fix an extended pure inner form G of $G^* \in \mathcal{E}_{\mathrm{ell}}(N)$ and let π_0 be a cohomological representation of G_∞ that is (odd) GSK-maxed. Further assume that if $\beta_+(\pi_0) = (1^{(r)}, a_1, \dots, a_k)$, then either $r \neq 2$ or none of the $a_i = 2$ (this is automatically satisfied if π_0 is odd GSK-maxed).*

Then all elements of $\Delta^{\max}(\pi_0)$ are (odd) GSK.

We warn that in general, $\Delta^{\max}(\pi_0)$ isn't a singleton. The different possibilities differ by different assignments of infinitesimal characters $\lambda_{i,v}$ to each block (T_i, d_i).

Example. Consider F with two infinite places v, w and $G_v \cong G_w \cong U(6,1)$. Let

$$\pi_{0,v} = (1,1),(1,0)^{(5)}, \quad \text{and} \quad \pi_{0,w} = (2,1),(1,0)^{(4)}$$

have the infinitesimal character of the trivial representation:

$$\lambda = (3,2,1,0,-1,-2,-3).$$

If $\Delta^{\max}(\pi_0) \neq \emptyset$, any $\Delta \in \Delta^{\max}(\pi_0)$ is of the form

$$\Delta = (2,1,(\lambda_v^1, \lambda_w^1)),(1,2,(\lambda_v^2,\lambda_w^2)),(1,3,(\lambda_v^3,\lambda_w^3)),$$

and the unordered partition $Q^{\max}(\pi_0)$ is $(3,2,1,1)$. We are forced to choose

$$\lambda_v^2 = (3,2), \qquad \lambda_w^3 = (3,2,1).$$

However, we still need to pick λ_v^3 and λ_w^2. This will correspond to a choice of ordering of $Q^{\max}(\pi_0)$ at each of v and w.

There are three choices

$$\lambda_v^3 = (1,0,-1),(0,-1,-2), \text{ or } (-1,-2,-3)$$

corresponding to three choices

$$\lambda_v^1 = (-2,-3),(1,-3), \text{ or } (1,0)$$

and three orderings of $Q^{\max}(\pi_0)$:

$$(2,3,1,1),(2,1,3,1), \text{ or } (2,1,1,3).$$

Similarly, there are three choices

$$\lambda_w^2 = (0,-1),(-1,-2), \text{ or } (-2,-3)$$

corresponding to three choices

$$\lambda_w^1 = (-2,-3),(0,-3), \text{ or } (0,-1)$$

and three orderings of $Q^{\max}(\pi_0)$:

$$(3,2,1,1),(3,1,2,1), \text{ or } (3,1,1,2).$$

Thus in total $\Delta^{\max}(\pi_0)$ can contain up to nine elements. Note that the different possibilities for λ_v^1 and λ_w^2 aren't necessarily even character twists of each other.

Example. As an simple example where this difficulty doesn't appear, assume

$$G_\infty = U(N-1,1)^r$$

and let $\pi_0 \cong \pi_{0,v}^r$ be diagonal. There is only one non-1 entry in any element of $\mathcal{P}_1(N-1,1)$ so π_0 is necessarily GSK-maxed. The arguments in Lemmas 11.2.5 and 11.2.6 show that since $\beta_+(\pi_0) = \beta_+(B_v)$ for all v and is a singleton, there is exactly one way to assign infinitesimal characters $\lambda_{i,v}$ to the blocks (T_i, d_i). This implies that $\Delta^{\max}(\pi_0)$ is a singleton.

11.3. **Characters on the Component Group.** As a last piece of the puzzle to derive explicit limit multiplicities for individual representations from Theorem 10.4.1, we attach characters of $S^\natural_{\psi_\infty}$ to elements of the AJ packet Π_{ψ_∞}. Let $P = (a_1, \dots, a_k) \in \mathcal{P}(N)$ correspond to an A-parameter ψ_∞ at infinity for $U(p,q)$ considered as an extended pure inner form. Recall the group $S^\natural_{\psi_\infty}$ from §2.4.2. As mentioned after [56, (1.3)], the equation in (3.2.1) reduces to a canonical isomorphism

$$S^\natural_{\psi_\infty} = \bigoplus_{1 \le i \le k} \mathbb{Z}/2 = \langle \epsilon_i \rangle_{1 \le i \le k},$$

where each index i corresponds to part a_i of P. In addition, the subgroup $\mathbb{Z}(\widehat{G})^\Gamma = \mathbb{Z}/2$ embedded diagonally.

Though the character attached to a π is not explicit in general, Mœglin-Renard in [56] manage to do so for AJ packets on unitary groups. Let

$$\pi_0 = (p_i, q_i)_i \in \beta^{-1}(P) \subseteq \mathcal{P}(p,q)$$

and define $a_{<i} = \sum_{j=1}^{i-1} a_i$. Then,

$$\eta^{\psi_\infty}_{\pi_0}(\epsilon_i) = (-1)^{p_i a_{<i} + q_i(a_{<i}+1) + a_i(a_i-1)/2}.$$

We can simplify this to

$$\eta^{\psi_\infty}_{\pi_0}(\epsilon_i) = (-1)^{a_i a_{<i} + q_i + \chi_4(a_i)}, \tag{11.3.1}$$

where

$$\chi_4(a_i) := \begin{cases} 0 & a_i \equiv 0, 1 \pmod 4 \\ 1 & a_i \equiv 2, 3 \pmod 4 \end{cases}.$$

If the a_i are all odd, this further simplifies to

$$\eta^{\psi_\infty}_{\pi_0}(\epsilon_i) = (-1)^{(i-1) + q_i + \chi_4(a_i)}. \tag{11.3.2}$$

Finally, if $\psi_\infty = \psi^\Delta_\infty$ for Δ of the type in Lemma 11.2.3, then

$$\mathcal{S}^\natural_\Delta = \left\langle s_d := \sum_{i : a_i = d} \epsilon_i \right\rangle_{d \in \mathbb{Z}^+}$$

is a subgroup of $S^\natural_{\psi_\infty}$. As a result, we can characterize representations such that $\eta^{\psi^\Delta_\infty}_{\pi_0}(S^\natural_\Delta) = 1$; this will be used to give asymptotics for multiplicities of individual representations in the odd GSK-maxed case.

Lemma 11.3.1. *Let $G^* \in \mathcal{E}_{\mathrm{ell}}(N)$ and G an extended pure inner form of G^*. Let $\pi_0 = \bigotimes_v \pi_{0,v}$ be an odd GSK-maxed representation of $G_\infty = \prod_v U(p_v, q_v)$.*

Let $Q^{\max}(\pi_0) = (1^{(r)}, d_1, \dots, d_k)$ with d_i odd and distinct. The $\Delta \in \Delta^{\max}(\pi_0)$ each determine orderings

$$P^\Delta_v := (a_{1,v}, \dots, a_{k+r,v})$$

of $Q^{\max}(\pi_0)$ for each v such that $\psi^\Delta_v = P^\Delta_v$. Let

$$\pi_{0,v} = (p^\Delta_{i,v}, q^\Delta_{i,v})_i \in \beta^{-1}(P^\Delta_v) \subseteq \mathcal{P}(p_v, q_v)$$

and for each d_j, let $i^\Delta_v(d_j)$ be the index i such that $d_j = a_{i,v}$.

Then, for $\Delta \in \Delta^{\max}(\pi_0)$, $\eta^{\psi^\Delta_\infty}_{\pi_0}(S^\natural_\Delta) = 1$ if and only if

$$t_j := \sum_{v \in \infty} (i^\Delta_v(d_j) - 1 + q^\Delta_{i^\Delta_v(d_j),v} + \chi_4(d_j))$$

is even for each $1 \le j \le k$ *and* G_∞ *satisfies the parity conditions from* 2.1.1.

Proof. The condition on t_j comes from checking that $\eta_{\pi_0}^{\psi_\infty^\Delta}(s_d) = 1$ for $d > 1$. The second comes from checking that $\eta_{\pi_0}^{\psi_\infty^\Delta}$ factors through $\mathcal{S}_\Delta$. □

Beware that it is important to keep track of whether $G_v = U(p,q)$ or $U(q,p)$ as an extended pure inner form to compute the characters $\eta_{\pi_0}^{\psi_\infty}$.

11.4. **Explicit Limit Multiplicities.** Now that we understand $\Delta(\pi_0)$, we can compute our main result: exact limit multiplicities for odd GSK-maxed π_0.

Theorem 11.4.1. *Let G be a pure inner form of $G^* \in \mathcal{E}_{\mathrm{ell}}(N)$. Choose*

$$f_{\mathfrak{n}}^\infty = f_{S_1} \bar{\mathbf{1}}_{K^{G,S}(\mathfrak{n})}$$

as in Section 9.1 with $S_0 = \emptyset$ (in particular, G and therefore E/F is unramified at all finite places and as in Lemma 2.1.1).

Pick a cohomological representation π_0 of G_∞ that is odd GSK-maxed and for

$$\Delta = (T_1, 1, \lambda_1, \eta_1), (1, d_i, \lambda_i, \eta_i)_{2 \le i \le k} \in \Delta^{\max}(\pi_0),$$

let $\lambda_1(\Delta) = \lambda_1$. Let T_1, d_i, R and L be the common values of T_1, d_i, $R(\Delta)$ and $L(\Delta)$ over $\Delta^{\max}(\pi_0)$. Then there are A, B, C, D, E with $C \ge 1$ such that as long as $|\mathfrak{n}| \ge D q_{S_1}^{E\kappa}$:

$$\begin{aligned}
|\mathfrak{n}|^{-R}\Gamma_L(\mathfrak{n})^{-1} &\sum_{\pi \in \mathcal{AR}_{\mathrm{disc}}(G)} \mathbf{1}_{\pi_\infty = \pi_0} \operatorname{tr}_{\pi^\infty}(f_{\mathfrak{n}}^\infty) \\
&= \frac{\operatorname{vol}(H(F)\backslash H(\mathbb{A}_f))}{\operatorname{vol}(K_H^S)} \left(\sum_{\Delta \in \Delta^{\max}(\pi_0)} \mathbf{1}_{\eta_{\pi_0}^{\psi_\infty^\Delta}(\mathcal{S}_\Delta) = 1} \frac{\dim \lambda_1(\Delta)}{|\Pi_{\mathrm{disc}}(\lambda_1(\Delta))|} \right. \\
&\qquad \left. \times \left(f_{S_1}^{H_1}(1) \coprod_{i>1} \int_{H_{i,\mathrm{der},S_1}} f_{S_1}^{H_i}(h)\, dh \right) + O(|\mathfrak{n}|^{-C} q_{S_1}^{A+B\kappa}\right),
\end{aligned}$$

where $H_i = H(T_i, d_i, \lambda_i, \eta_i)$ and $H = H_1 \times \prod_{i>1} H_i^{\mathrm{ab}}$ are both constant over $\Delta \in \Delta^{\max}(\pi_0)$.

Finally, recall that the condition on $\eta_{\pi_0}^{\psi_\infty^\Delta}$ can be checked as in Lemma 11.3.1, that

$$R = (k-1) + \frac{1}{2}\left(N^2 + T_1^2 - \sum_{i \ge 2} d_i^2, \right),$$

and that

$$L = T_1, 1^{(k-1)}, -1^{(k-1)}.$$

Proof. Apply Proposition 10.2.2. Then Corollary 11.2.12 allows us to apply Theorem 10.4.1 to compute the main terms $m(\pi_0, \Delta, f^\infty)$ for $\Delta \in \Delta^{\max}(\pi_0)$. We compute $R = R(\pi_0)$ by Corollary 11.2.9. Theorem 10.4.2 bounds all the terms $m(\pi_0, \Delta', f^\infty)$ for $\Delta' \notin \Delta^{\max}(\pi_0)$. □

Note. Since the different $\Delta = (T_i, d_i, \lambda_i, \eta_i)_i \in \Delta^{\max}(\pi_0)$ only differ in their λ_i-coordinates, they correspond to the same map

$$\mathcal{S}_\Delta : \widehat{H}_v^{\mathrm{ur,temp}} \to \widehat{G}_v^{\mathrm{ur}}$$

as in (6.2.3). Let $\mu^{\mathrm{pl,ur}}(H_{S_1})$ be the Plancherel measure on the unramified spectrum of H_{S_1} and define the pushforward

$$\mu^{\mathrm{pl}(\pi_0),\mathrm{ur}}(G_{S_1}) := \mu_{S_1}^{\mathrm{pl}(\pi_0),\mathrm{ur}} := (\mathcal{S}_\Delta)_*(\mu^{\mathrm{pl,ur}}(H_{S_1})).$$

Then, as in (9.5.2), we can compute

$$f_{S_1}^{H_1}(1) \coprod_{i>1}^{\oplus} \int_{H_{i,\mathrm{der},S_1}} f_{S_1}^{H_i}(h)\, dh = \mu_{S_1}^{\mathrm{pl}(\pi_0),\mathrm{ur}}(\widehat{f}_{S_1}).$$

This interprets Theorem 11.4.1 as an unramified "Plancherel" equidistribution theorem for the local component π_S like in [73]. Beware that $\mu^{\mathrm{pl}(\pi_0),\mathrm{ur}}$ can have support on the non-tempered spectrum.

Theorem 11.4.1 only computes exact asymptotics for odd GSK-maxed representations. For general representations, we have an upper bound:

Theorem 11.4.2. *Let G be an extended pure inner form of $G^* \in \mathcal{E}_{\mathrm{ell}}(N)$. Choose*

$$f_{\mathfrak{n}}^{\infty} = \varphi_{S_0} f_{S_1} \overline{\mathbf{1}}_{K^{G,S}(\mathfrak{n})}$$

as in Section 9.1 with f_{S_1} and φ_{S_0} is arbitrary.

Pick a cohomological representation π_0 and let $R_0(\pi)$ be defined as in the end of §11.2.2 of G_∞. Then if $\Delta(\pi_0) = \emptyset$,

$$\sum_{\pi \in \mathcal{AR}_{\mathrm{disc}}(G)} \mathbf{1}_{\pi_\infty = \pi_0} = 0.$$

Otherwise, there are A, B, D, E such that as long as $|\mathfrak{n}| \geq D q_{S_1}^{E\kappa}$:

$$\sum_{\pi \in \mathcal{AR}_{\mathrm{disc}}(G)} \mathbf{1}_{\pi_\infty = \pi_0} \operatorname{tr}_{\pi^\infty}(f_{\mathfrak{n}}^{\infty}) = O(|\mathfrak{n}|^{R(\pi_0)} q_{S_1}^{A+B\kappa}).$$

Proof. Apply Proposition 10.2.2 and Theorem 10.4.2. □

The above result applies to any CM extension E/F, any extended pure inner form G, and any φ_{S_0}. Assuming Conjecture 9.6.3, the exponent $R(\pi_0)$ can be improved to an analogous $R_0(\pi_0)$. Furthermore, in the sum from Theorem 10.2.1, at least one of the $\Delta_1 \times \Delta_2$ for each $\Delta \in \Delta^{\max}(\pi_0)$ satisfies that the d_i assigned to each Δ_i all have the same parity. Therefore, if we assume the even stronger Conjecture 9.6.4 and don't have an obstruction from the multiplicity formula (like the $\eta_{\pi_0}^{\Delta}$ condition from Theorem 11.4.2), we should expect $R(\pi_0)$ to be optimal.

12. Examples and Corollaries

This section contains examples and corollaries of the main Theorems 11.4.1 and 11.4.2, including applications to Sato-Tate equidistribution in families, the Sarnak-Xue density hypothesis, and studying the cohomology of locally symmetric spaces. We first recall all our various growth rates for reader's convenience:

Let π_0 be on a rank-N group. Each $\star(\pi_0)$ is a maximum of $\star(\Delta) = \star((T_i, d_i, \lambda_i, \eta_i)_i)$ for $\Delta \in \Delta(\pi_0)$. Then there are two formulas:

First, the provable growth rate from Proposition 9.4.4, which will be used in all theorem statements:

$$R(\pi_0) = \max_{\Delta\in\Delta(\pi_0)} \frac{1}{2}\left(N^2 + \sum_i T_i^2 d_i\right) - \sum_{i:T_i=1}\left(\frac{1}{2}(d_i^2+d_i) - 1\right) - \sum_{i:T_i=2}(3d_i-3) - \sum_{\substack{i:T_i=3\\ d_i>1}}\left((5-10^{-100})d_i+5)\right). \tag{12.0.1}$$

Second, we have the conjecturally sharp growth rate from 9.6.3 which we will use as a comparison:

$$R_0(\pi_0) = \max_{\Delta\in\Delta(\pi_0)} \frac{1}{2}\left(N^2 - \sum_i T_i^2 d_i\right) + \sum_i\left(T_i^2 + \frac{1}{2}T_i(T_i-1)(d_i^2-1)\right).$$

12.1. **Examples.** First, we work out what Theorem 11.4.1 says in some simple cases. To discuss infinitesimal characters, we let $\lambda_1,\dots,\lambda_{n-1}$ be the fundamental weights of GL_n,

$$\lambda_i = (1^{(i)}, 0^{(N-i)}),$$

in the standard basis of $X^*(T)$ corresponding to the entries of a diagonal matrix. Define λ_0, λ_n similarly for indexing purposes. Two weights are character twists of each other if they differ by a multiple of λ_n, and the half-sum of positive roots is

$$\rho_n = \left(\frac{n-1}{2}, \frac{n-3}{2}, \dots, \frac{1-n}{2}\right).$$

12.1.1. *Example 1: parallel case for $U(N-1,1)$.* For the simplest example with non-tempered representations at infinity, assume:

- $\deg F/\mathbb{Q} = d$ is even,
- $G_\infty \cong U(N-1,1)^d$ as is allowed by §2.1,
- $\pi_\infty \cong \pi_0^d$ with $\pi_0 = ((1,0)^{(r)}, (k-1,1), (1,0)^{(N-k-r)})$ for $k>1$ odd and $r+k\le N$,
- $S_1 = \emptyset$.

Then $\Delta^{\max}(\pi_0)$ is a singleton

$$(N-k, 1, (\lambda_{1,v})_v, \eta_1), (1, k, (\lambda_{2,v})_v, \eta_2),$$

with $\lambda_{1,v}$ a character twist of $k\lambda_r + \rho_{n-k}$ on GL_{n-k} and $\lambda_{2,v}$ the infinitesimal character of a 1-dimensional irrep. We can also recall the the discrete L-packet at infinitesimal character $\lambda_{2,v}$ has size N. Finally, d is even, so $\eta_{\pi_\infty}^{\psi_\infty} = (\eta_{\pi_0}^{\psi_v})^d$ is is trivial.

Denoting by $\pi_{k\lambda_r}$ the finite dimensional representation of $\mathrm{GL}_{N-k}\mathbb{C}$ with highest weight $k\lambda_r$, we compute

$$|\mathfrak{n}|^{-(N(N-k)+1)} L_{k,1,-1}(\mathfrak{n})^{-1} \sum_{\substack{\pi\in\mathcal{AR}_{\mathrm{disc}}(G)\\ \pi_\infty=\pi_0}} \dim((\pi^\infty)^{K(\mathfrak{n})}) = \frac{1}{N^d}\dim(\pi_{k\lambda_r})^d \tau'(U_{E/F}(N-k)\times U(1)) + O(|\mathfrak{n}|^{-C} q_{S_1}^{A+B\kappa}). \tag{12.1.1}$$

as an asymptotic count of automorphic *forms* of level $\mathfrak{n}$ corresponding to π_∞. Recall that τ' is the modified Tamagawa number from (9.1.1).

Note that the "masses" (i.e. relative abundances in the automorphic spectrum)

$$\frac{1}{N^d}\dim(\pi_{k\lambda_r})^d$$

depend on r even though the corresponding π have the same infinitesimal character and come from the same Levi in the Langlands classification. This is because how the infinitesimal character divides up between the blocks of this Levi also matters.

More specifically, $\dim(\pi_{k\lambda_r})$ is largest for r close to $(N-k)/2$ and decreases to 1 towards the extreme cases of $r = 0$ or $N-k$. For another prespective, all representations considered are cohomological of degree $d(N-k)$. The representations with Hodge weights closer to $\left(\frac{d}{2}(N-k), \frac{d}{2}(N-k)\right)$ have larger masses, whereas the ones whose weights are closer to $(0, d(N-k))$ and $(d(N-k), 0)$ are rarer.

12.1.2. *Example 2.* As a slight complication, now assume:

- $\deg F/\mathbb{Q} = d$ is odd,
- $N \not\equiv 0 \pmod 4$ and
$$G_\infty \cong \begin{cases} U(N-1,1)^d & N \equiv 2,3 \pmod 4 \\ U(1,N-1)^d & N \equiv 1 \pmod 4 \end{cases}$$
to satisfy the conditions in §2.1,
- $\pi_\infty = \pi_0^d$ with
$$\pi_0 = \begin{cases} ((1,0)^{(r)}, (k-1,1), (1,0)^{(N-k-r)}) & N \equiv 2,3 \pmod 4 \\ ((0,1)^{(r)}, (1,k-1), (0,1)^{(N-k-r)}) & N \equiv 1 \pmod 4 \end{cases}$$
for $k > 1$ odd and $r + k \le N$,
- $S_1 = \emptyset$.

This situation is similar to the first example, but no longer necessarily have $\eta^\Delta_{\pi_\infty} = 1$.

Using the test from Lemma 11.3.1, $\eta^\Delta_{\pi_\infty} = 1$ if and only if

$$r + \begin{cases} 1 & N \equiv 2,3 \pmod 4 \\ N-1 & N \equiv 1 \pmod 4 \end{cases} + \begin{cases} 1 & k \equiv 0,1 \pmod 4 \\ 0 & k \equiv 2,3 \pmod 4 \end{cases} \equiv 0 \pmod 2$$

is even. We write this condition as

$$r + 1 + \chi_4(N) + \chi_4(k) \equiv 0 \pmod 2. \tag{12.1.2}$$

in terms of the character χ_4 from Lemma 11.3.1 and think of it as a parity condition on r.

If (12.1.2) holds, then we have the same result (12.1.1) as in the previous example. Otherwise, we have that

(12.1.3)
$$|\mathfrak{n}|^{-(N(N-k)+1)} L_{k,1,-1}(\mathfrak{n})^{-1} \sum_{\substack{\pi \in \mathcal{AR}_{\mathrm{disc}}(G) \\ \pi_\infty = \pi_0}} \dim((\pi^\infty)^{K(\mathfrak{n})}) = O(|\mathfrak{n}|^{-C} q_{S_1}^{A+B\kappa}).$$

There are two consequences. First, we only have exact asymptotics in the case where r satisfies the parity condition (12.1.2). Second, the asymptotic growth rate of counts of forms can be different for representations coming from the same Levi in the Langlands classification.

These phenomena are caused by an obstruction from the multiplicity formula in our specific setup: all automorphic representations counted have only unramified local components at non-split, non-Archimedean places. In particular, they all corresponded to trivial character on the component group. Furthermore, all those representations came from parameters ψ with $\epsilon_\psi = 1$. In total, the multiplicity formula requires $\eta^\psi_{\pi_0} = 1$ for the packet Π_ψ to contribute to the multiplicity of π_0.

Even more surprisingly, growth rates can be different for different members of the same L-packet in cases beyond $U(N,1)$. If $\pi = ((p_i, q_i))_i$ on $U(p,q)$, then it can be seen from the description in §6 of [78] that some other members of its (pseudo- and therefore true) L-packet can be produced by reversing some number of pairs (p_i, q_i) such that we remain in $\mathcal{P}(p,q)$. For an L-packet like

$$\{((2,1),(0,1)),((1,2),(1,0))\}$$

on $U(2,2)$, only one member satisfies the parity condition from Lemma 11.3.1. This is starkly different from the discrete-at-infinity case in [26] and caused by our dominant contribution to growth rates coming from shapes Δ with non-trivial $\mathcal{S}_\Delta$.

12.1.3. *Example 3.* We will also consider an example where there is only one non-compact place. Assume:

- $\deg F/\mathbb{Q} = d$ with a fixed place $v_0 \in \infty$.
- $G_\infty \simeq U(p,q) \times U(N,0)^{d-1}$ where
$$q \equiv \begin{cases} 0 \pmod 2 & d \text{ even or } N \equiv 0,1 \pmod 4 \\ 1 \pmod 2 & d \text{ odd and } N \equiv 2,3 \pmod 4 \end{cases}$$
to satisfy the conditions of §2.1. The $U(p,q)$ factor is at v_0.
- $\pi_\infty = \pi_0 \times \mathbf{1}^{d-1}$ where
$$\pi_0 = ((p_i, q_i))_i, \qquad n_i = p_i + q_i.$$
is odd GSK-maxed with the infinitesimal character of the trivial representation, and $\mathbf{1}$ is the trivial representation.
- $S_1 = \emptyset$.

We need some more combinatorial parameters

- d_j for $1 \le i \le k$ are the distinct non-1 values among the n_i,
- If $d_j = n_{i(d_j)}$, let $r_j = i(d_j) - \#\{i(d_{j'}) < i(d_j)\}$
- $M = N - \sum_j d_j$.

The $\Delta \in \Delta^{\max}(\pi_\infty)$ are then all of the form

$$(M, 1, (\lambda_v)_v, \eta), (1, d_j, (\lambda_{j,v})_j, \eta_j)_j,$$

where all the $\lambda_{j,v}$ are infinitesimal characters of 1-dimensional irreps and λ_{v_0} is a character twist of

$$\rho_M + \sum_j d_j \lambda_{r_j}.$$

The possible choices for each λ_v with $v \neq v_0$ are in bijection with reorderings $(n_{v,i})_i$ of $(n_i)_i$. We can define $i_v(d_j)$ and $r_{v,j}$ from $(n_{v,i})_i$ similar to the original definitions of $i(d_j)$ and r_j. Then λ_v is equal to a character twist of

$$\rho_M + \sum_j d_j \lambda_{r_{v,j}}.$$

The condition on the character from Lemma 11.3.1 then reduces to

$$i(d_j) + \sum_{v \neq v_0} i_v(d_j) \equiv d(\chi_4(d_j) - 1) + q_{i(d_j)} \pmod 2 \tag{12.1.4}$$

for all $1 \le j \le k$. There is always a set of reorderings $(n_{v,i})_i$ that satisfy (12.1.4): the only way there couldn't be is if the d_j are all the n_i and a parity condition from summing (12.1.4) over all j fails [6]. However, the parity condition on q implies that the condition from summing always holds.

Therefore, this example is analogous to Example 1 where we do not need to check the character condition form Lemma 11.3.2 to get lower bounds. In particular the asymptotic (12.1.1) holds with a much more complicated factor replacing the $N^{-d}\dim(\pi_{r\lambda})^d$.

12.2. **Growth of Cohomology.** Our limit multiplicities can be translated into computations of upper and lower bounds for the growth cohomology of arithmetic lattices in G_∞. Recall that for a cocompact lattice Γ in a Lie group G_∞ with maximal compact K_∞ and Lie algebra $\mathfrak{g}$, Matsushima [53] computed the cohomology of Γ with coefficients in the finite-dimensional representation F of G:

$$H^*(\Gamma, F) = \sum_{\pi \in G_\infty^\vee} m(\pi, \Gamma) H^*(\mathfrak{g}, K_\infty; \pi \otimes F^*). \tag{12.2.1}$$

Here $G_\infty^\vee$ is the unitary dual of G_∞, the integer $m(\pi,\Gamma)$ is the multiplicity of π in $L^2(\Gamma\backslash G_\infty)$, and $H^*(\mathfrak{g}, K_\infty)$ is the $(\mathfrak{g}, K)$-*cohomology* of π with coefficients in F; see [17]. We say π is *cohomological with coefficients in* F if $H^*(\mathfrak{g}, K_\infty; \pi \otimes F^*)$ is nontrivial.

We will restrict our groups G so that $G_\infty = U(p,q) \times U(N,0)^a \times U(0,N)^b$. Following Lemma 2.1.1, such G exist for all values of N, though possibly not over all unramified extensions E/F. We will abuse notation and say that a degree i of cohomology appears in an A-packet Π_ψ if Π_ψ contains a cohomological representation with non-vanishing cohomology in degree i.

12.2.1. *Lower Bounds.* We give a sample result of the kind of lower bounds on growth of cohomology produced by Theorem 11.4.1. First, as a direct consequence of Lemma 11.1.5 and of Section 11.1.3, we find:

Lemma 12.2.1. *Let $G_v = U(p,q)$ with $p+q = N$ and $\min(p,q) = r$. Let $1 < d \le N$ be odd.*

(i) *Let $Q \in \mathcal{P}(N)$ be an ordered partition with one entry equal to d and all others equal to 1. Then the lowest degree of cohomology appearing in the packet $\Pi_{\psi_Q,v}$ is:*
$$i = i(d,N,r) = \begin{cases} r(N-r) - \frac{d^2-1}{4} & r \ge \frac{d-1}{2} \\ r(N-d) & r \le \frac{d-1}{2} \end{cases}.$$

(ii) *If $r \le \frac{d-1}{2}$, the degree i is achieved by a unique $\pi_i \in \Pi_{\psi_Q,v}$. If*
$$Q = (\overbrace{1,...,1}^{s}, d, 1, ..., 1) \in \mathcal{P}(N),$$
then the only Hodge weights of π_i in degree $i = r(N-d)$ are:
$$(a,b) = \begin{cases} (rs, r(N-d-s)) & G = U(N-r, r) \\ (r(N-d-s), rs) & G = U(r, N-r). \end{cases}$$

[6] this statement boils down to a combinatorial puzzle about filling in a $(\#\{n_i\}-1) \times k$ grid of squares black or white such that each row has $\lfloor(\#\{n_i\}-1)/2\rfloor$ black squares and the first $d-1$ columns have a fixed parity of black squares

From this we deduce the following:

Theorem 12.2.2. *Let E/F be unramified at all finite places, and let G be an extended pure inner form of G^* unramified at all finite places, isomorphic to $U(p,q)$ at one infinite place, and compact at all the others. For $\mathfrak{n}$ divisible only by primes that split in E, let $\Gamma(\mathfrak{n}) = G(F) \cap K^G(\mathfrak{n})$ be a lattice in $U(p,q)$ of level $\mathfrak{n}$. Let $N = p+q$ and $r = \min(p,q)$. Let $j \not\equiv N \mod 2$ be such that $j \le |p-q|-1$. Then*

$$\dim H^{rj}(\Gamma(\mathfrak{n}), \mathbb{C}) \gg |\mathfrak{n}|^{Nj}.$$

The exact same bound holds for each $H^{rk,r(j-k)}(\Gamma(\mathfrak{n}), \mathbb{C})$ for $0 \le k \le j$.

Proof. Let v_0 be the infinite place at which G is not compact. By Lemma 12.2.1, it suffices to give lower bounds on multiplicities for the representation $\pi_{rj} = \pi_{rj,v_0} \otimes \mathbf{1}^{[F:\mathbb{Q}]-1}$. Since π_{rj} is odd GSK-maxed by the congruence condition on j, Theorem 11.4.1 gives exact multiplicities with $R = Nj+1$ provided that

$$\eta_{\pi_{rj}}^{\psi_\infty^\Delta}(\mathcal{S}_\Delta) \equiv 1.$$

Since we are in the same setup as Example 12.1.3, there automatically exists a shape Δ satisfying this condition.

This shows the result for the cohomology of the disconnected locally symmetric spaces $Y(\mathfrak{n}) = G(F)\backslash G(\mathbb{A})/K(\mathfrak{n})K_\infty$, of which $\Gamma(\mathfrak{n})\backslash G_\infty/K_\infty$ is one connected component. Though the different connected components of $Y(\mathfrak{n})$ are not necessarily isomorphic, their cohomology has the same dimension by Lemma 12.2.3 below. It then follows from [28, §2] that

$$\dim H^i(Y(\mathfrak{n}), \mathbb{C}) = |\pi_0(T(F)\backslash T(\mathbb{A})/\nu(K(\mathfrak{n})K_\infty))| \dim H^i(X(\mathfrak{n}), \mathbb{C}),$$

for $T = G/G^{\mathrm{der}}$. Since $|\pi_0(T(\mathbb{A})/T(F)\nu(K_\infty \times K(\mathfrak{n})))| \gg |\mathfrak{n}|^{1-\epsilon}$, we conclude. $\square$

Lemma 12.2.3. *Let G^* be an inner form of $G = U_{E/F}(N)$ and K be a compact open subgroup of $G(\mathbb{A}_f)$. Let*

$$Y_K = G(F)\backslash G(\mathbb{A})/K_\infty \times K$$

be the locally symmetric space of level K. Let $X_{K,\gamma}$ denote the connected components of Y_K, indexed by representatives of $G(F)\backslash G(\mathbb{A}^\infty)/K$. Then for any γ, γ' we have

$$\dim H^i(X_{K,\gamma}, \mathbb{C}) = \dim H^i(X_{K,\gamma'}, \mathbb{C}).$$

Proof. Let G' be the derived subgroup of G with $T = G/G'$ and quotient map ν. We have

$$Y_K = \bigsqcup X_{K,\gamma} := \bigsqcup_\gamma G'(F)\backslash G'(\mathbb{A})/\gamma K \gamma^{-1}(K_\infty \cap G'(F_\infty))$$

for a finite set of representatives γ of $G(F)\backslash G(\mathbb{A})/K$. Consider the space $S_G = G(F)\backslash G(\mathbb{A})/K_\infty = \varprojlim_K Y_K$: it carries an action of $G(\mathbb{A}_f)$ by translation which, following [28, §2], is transitive on connected components. Moreover, since G' is simply connected, $G'(\mathbb{A}_f)$ is the stabilizer of a given connected component. In fact, the space S_G is the induction of the connected component

$$S_{G'} = G'(F)\backslash G'(\mathbb{A})/(K_\infty \cap G'_\infty),$$

in a precise sense laid out in [29, §2.7]. In [60], Newton shows that this translates actual induction when passing to cohomology:

$$H^*(S_G, \mathbb{C}) = \mathrm{Ind}_{G'(\mathbb{A}_f)}^{G(\mathbb{A}_f)} H^*(S_{G'}, \mathbb{C}),$$

where the cohomology is a smooth representation realized as direct limit over the cohomology of the S_K (viewed as K-fixed vectors) in the natural way. Fixing a set of representatives γ for $G(\mathbb{A}_f)/G'(\mathbb{A}_f)$, we deduce that

$$H^*(S_G,\mathbb{C})\,|_{G'(\mathbb{A}_f)} = \bigoplus_\gamma H^*(S_{G'},\mathbb{C})^\gamma,$$

where the superscript γ denotes conjugating the representation. In particular, for the subgroup K of $G'(\mathbb{A}_f)$, we have

$$\dim\left((H^*(S_{G'},\mathbb{C})^\gamma)^K\right) = \dim(H^*(S_{G'},\mathbb{C})^{\gamma K\gamma^{-1}}) = \dim H^*(X_{\gamma K\gamma^{-1}},\mathbb{C}),$$

where the last equality follows in particular from Matsushima's formula. □

We compare these lower bounds with previously known results: in [52] Marshall-Shin showed that, when $d < N-1$, $\dim H^d(\Gamma(\mathfrak{n}),\mathbb{C}) \ll_\epsilon |\mathfrak{n}|^{Nd+\epsilon}$ for split-level lattices $\Gamma(\mathfrak{n})$ in $U(N-1,1)$ and conjectured that this bound was sharp. Theorem 12.2.2 specialized to $(p,q)=(N-1,1)$ then gives:

Corollary 12.2.4. *If $d \not\equiv N \mod 2$, then the upper bounds obtained by Marshall-Shin are sharp.*

12.2.2. *Upper Bounds.* As an example of the upper bounds, fix:

- a CM extension E/F,
- an extended pure inner form G of $G^* \in \mathcal{E}_{\mathrm{ell}}(N)$ that is isomorphic to $U(p,q)$ at one infinite place v_0 and compact at all other infinite places,
- a finite set of finite places S at which G is split,
- an open compact $U^{S,\infty}$ away from S and ∞,
- $\mathfrak{n}$ an ideal supported over S. We compute asymptotics as $|\mathfrak{n}| \to \infty$,
- $\Gamma(\mathfrak{n}) = G(F) \cap U^{S,\infty}K_S^G(\mathfrak{n})$ a lattice.

As before, we are interested in $H^{a,b}(\Gamma_i,\mathbb{C})$. To compute this:

(1) We can use Lemma 11.1.5 to enumerate all the $\pi_0 \in \mathcal{P}_1(p,q)$ that contribute to $H^{a,b}$.
(2) We use the algorithm at the end of §11.2.2 and the formula (12.0.1) to compute all the $R(\pi_0)$. Let the maximum value be $R(a,b)$.

Then:

Proposition 12.2.5. *For all $\epsilon > 0$,*

$$\dim H^{a,b}(\Gamma(\mathfrak{n}),\mathbb{C}) \ll_\epsilon |\mathfrak{n}|^{R(a,b)-1+\epsilon}.$$

Proof. This follows from applying Theorem 11.4.2 to each of the π_0 contributing, applying Matushima's formula, and then using the bounds on the number of connected components to reduce to a single connected component of the adelic quotient. □

Example (Lowest Degree). If the non-compact factor of G is isomorphic to $U(p,q)$, let $r=\min(p,q)$. Then r is the lowest degree of cohomology that is not guaranteed to vanish for local reasons. There are only two nontrivial cohomological representations in degree r, and they have weights $(r,0)$ and $(0,r)$ respectively. Then $R(0,r) = R(r,0) = p+q$.

Example (Upper and Lower Bounds for $U(N-2,2)$). Consider the case case where the non-compact factor is $U(N-2,2)$ and choose degree of cohomology $0 < i < 2(N-2)$. We also assume $N > 6$ for simplicity.

Let $j = 2(N-2) - i$. The possible π_0 that contribute correspond to the elements

$$(j/2, 2), (1, 0)^{(N-2-j/2)} \text{ or } (a, 1), (j-a, 1), (1, 0)^{(N-2-j)}$$

of $\mathcal{P}_1(N-2, 2)$ up to reordering. This corresponds to smallest Arthur-SL_2's

$$[j/2+2] \oplus [1]^{(N-2-j/2)} \qquad \text{or} \qquad [a+1] \oplus [j-a+1] \oplus [1]^{(N-2-j)}.$$

The possible shapes (ignoring infinitesimal character components) in $\Delta^{\max}(\pi_0)$ for π_0 contributing to H^i are therefore of the form

$$\begin{cases} (1, j/2+2), (N-j/2-2, 1) \\ (1, a+1), (1, a-j+1), (N-j-2, 1) \\ (2, j/2+1), (N-j-2, 1) \end{cases}$$

The first two cases come from applying Lemma 11.2.7 while the last can be checked by hand with our assumption that $N > 6$ (the only relevant failure of 11.2.7 on non-GSK shapes is Arthur-SL_2 $(2, 2, 1, 1)$ vs $(2, 2, 2)$).

The second case is only maximizing when $a = (j-1)/2$ since this maximizes R. Therefore, breaking into cases mod 2, the possible dominant shapes contributing to H^i are

$$\begin{cases} (1, j/2+2), (N-j/2-2, 1) & \text{i even} \\ (2, j/2+1), (N-j-2, 1) & \text{i even, } i \geq N-2 \\ (1, (j-1)/2+1), (1, (j+1)/2+1), (N-j-2, 1) & \text{i odd, } i \geq N-2 \end{cases} \tag{12.2.2}$$

with no shapes contributing when none of these conditions hold. We can then compute that $H^i(\Gamma(\mathfrak{n}), \mathbb{C}) \ll_\epsilon |\mathfrak{n}|^{R_i} - 1$ where

$$R_i = \max \begin{cases} Ni/2 + 1 & \text{i even,} \\ 1/2(i+5/2)^2 + N(N-i-3) + 23/8 & \text{i even, } i \geq N-2 \\ (i/2+1)^2 + 7/4 & \text{i odd, } i \geq N-2 \\ 0 \end{cases}$$

As in the example §12.1.3, for any π_0 contributing to H^i from a given shape, we are always able to add infinitesimal characters to the shapes in (12.2.2) to produce refined shapes such that the character condition from Lemma 11.3.1 is satisfied. Therefore, R_i gives an exact asymptotic whenever the dominant shape is odd-GSK—i.e. whenever $i \equiv 2N \pmod 4$ and the first case achieves the maximum.

In addition, assuming Conjecture 9.6.3, the exact asymptotic should actually be

$$R_{0,i} = \max \begin{cases} Ni/2 + 1 & \text{i even,} \\ (i/2+1)^2 + 3 & \text{i even, } i \geq N-2 \\ (i/2+1)^2 + 7/4 & \text{i odd, } i \geq N-2 \\ 0 \end{cases}$$

and no longer requires the $N > 6$ condition.

12.2.3. *A Note on Vanishing of Cohomology.* Historically, another case of interest has been to compute examples where certain degrees of cohomology actually vanish; see for example the final result in [63] for rank three unitary groups and [22] for general rank.

This can be achieved by choosing G that is a division algebra at some place v_0 where E/F is split. Then various shapes could potentially never contain parameters that are relevant on G (as in [44, §1.3.7]) and therefore never contribute to $\mathcal{AR}_{\mathrm{disc}}$. In particular, certain cohomological π_0 can simply not appear at infinity.

To reconcile with our bounds, this mechanism is hidden within the computation of endoscopic transfers in Proposition 10.2.2. It manifests through transfers at v_0 either vanishing, immediately zeroing out terms in 10.2.2, or having certain constant terms vanish, zeroing out terms after further split-place transfers computed by Lemma 6.1.2. We do not take into account this potential tightening in Theorem 11.4.2 for two reasons: first, it is already known and second, it is better thought of as a corollary of [44] (after plugging some computations with the parameterization of cohomological representations on unitary groups recalled in §11.1) instead of a consequence of the new techniques here.

12.3. **Sato-Tate Equidistribution in Families.** We prove an averaged Sato-Tate result similar to Theorem 9.26 in [73] since our main theorem 11.4.1 has error bounds of the same strength in f_{S_1}.

We consider families of automorphic representations with infinite component equal to an odd GSK-maxed π_0. Their Satake parameters will not equidistribute with respect to the Sato-Tate measure on G, but rather with respect to the pushforward of the Sato-Tate measure on a smaller group related to the $\Delta \in \Delta^{\max}(\pi_0)$. Otherwise, this section will follow [73] extremely closely.

12.3.1. *Sato-Tate Measures.* First we recall the definition of Sato-Tate measures from [73, §§3,5] (the full details can be found there). Choose a place v of F and unramified reductive group G_v over F_v. Let $A \subseteq T$ be a maximally split torus of G and maximal torus containing it. Let A_c, T_c be their maximal compact subgroups. By looking at Satake parameters, we get a parameterization of the tempered, unramified dual of G_v:

$$\widehat{G}_v^{\mathrm{ur,temp}} \simeq \Omega_{F_v} \backslash \widehat{A}_c \simeq \Omega_{F_v} \backslash \widehat{T}_c/(\mathrm{id} - \mathrm{Frob}_v)\widehat{T}_c.$$

This space is also the same as $\widehat{G}$-conjugacy classes in $\widehat{K}_{\mathrm{Frob}_v} \rtimes \mathrm{Frob}_v$ where $\widehat{K}_{\mathrm{Frob}_v}$ is the maximal Frob_v-invariant compact subgroup of $\widehat{G}$.

Of course, not every v has the same Frobenius. To deal with this, if G splits over F_1, let $\Gamma = \mathrm{Gal}(F_1/F)$. Then for each $\theta \in \Gamma$, let

$$\widehat{T}_{c,\theta} := \Omega_{F_v} \backslash \widehat{T}_c/(\mathrm{id} - \theta)\widehat{T}_c.$$

For $\gamma \in \Gamma$, $t \mapsto \gamma t$ canonically identifies $T_{c,\theta}$ with $T_{c,\gamma\theta\gamma^{-1}}$ so $\widehat{T}_{c,\theta}$ can be considered to only depend on the conjugacy class of θ. This is therefore a uniform description of $\widehat{G}_v^{\mathrm{ur,temp}}$ whenever $\mathrm{Frob}_v = \theta$.

Definition 12.3.1. The Sato-Tate measure $\mu_\theta^{\mathrm{ST}} := \mu_\theta^{\mathrm{ST}}(G)$ on $\widehat{T}_{c,\theta}$ is the quotient under $\widehat{G}$-conjugation of the Haar measure on $\widehat{K}_\theta \rtimes \theta$ with total volume 1.

This should be thought of as the "most canonical" possible measure to put on $T_{c,\theta}$.

Now, let $\mathcal{V}_F(\theta)$ be the set of places v of F such that $\mathrm{Frob}_v = \theta$. For $v \in \mathcal{V}_F(\theta)$, the Plancherel measure on $\widehat{G}_v^{\mathrm{ur,temp}}$ (normalized so that a maximal compact of G_v has volume 1) gives another measure $\mu_v^{\mathrm{pl,ur}} := \mu^{\mathrm{pl,ur}}(G_v)$ on $\widehat{T}_{c,\theta}$.

Lemma 12.3.2. *Let sequence $v \in \mathcal{V}_F(\theta)$ such that $q_v \to \infty$. Then there is weak convergence $\mu_v^{\mathrm{pl,ur}} \to \mu_\theta^{\mathrm{ST}}$.*

Proof. This follows by explicit formulas [73, Prop 3.3] and [73, Lem 5.2]. □

12.3.2. *Equidistribution.* We can now state and prove the equidistribution result. Fix G an unramified extended pure inner form of some $G^* \in \mathcal{E}_{\mathrm{ell}}(N)$. Note that at all finite places, G splits over E. We fix:

- an odd GSK-maxed cohomological representation π_0 of G_∞,
- $\theta \in \mathrm{Gal}(E/F)$,
- a sequence $v_i \in \mathcal{V}_F(\theta)$ (i.e. either all split or all non-split),
- a sequence of ideals $\mathfrak{n}_i$ of $\mathcal{O}_F$ relatively prime to v_i.

The different $\Delta = (T_i, d_i, \lambda_i, \eta_i)_i \in \Delta^{\max}(\pi_0)$ differ only in their λ_i-coordinates and therefore correspond to the same map

$$\mathcal{S}_\Delta : \widehat{H}_v^{\mathrm{ur,temp}} \hookrightarrow \widehat{G}_v^{\mathrm{ur}}$$

as in formula (6.2.3). Furthermore, the common group H_v as in (6.2.2) is the same as from Theorem 11.4.1. For $\theta \in \mathrm{Gal}(E/F)$, define the pushforward

$$\mu_\theta^{\mathrm{ST}(\pi_0)} := \mu_\theta^{\mathrm{ST}(\pi_0)}(G) := (\mathcal{S}_\Delta)_*(\mu_\theta^{\mathrm{ST}}(H)).$$

Beware that this is a measure on the full unramified dual $\widehat{G}_v^{\mathrm{ur}}$ for $v \in \mathcal{V}_F(\theta)$ instead of just the tempered part.

Finally, for each i, define the empirical distribution on $\widehat{G}_\theta^{\mathrm{ur}}$:

$$\mu_{\mathfrak{n}_i, v_i}^{\pi_0} := \sum_{\pi \in \mathcal{AR}_{\mathrm{disc}}(G)} \mathbf{1}_{\pi_\infty = \pi_0} \dim((\pi^\infty)^{K^G(\mathfrak{n}_i)}) \delta(\sigma_{\pi_{v_i}}).$$

Here, $\delta(\sigma_{\pi_{v_i}})$ is the delta-measure at Satake parameter $\sigma_{\pi_{v_i}}$.

Then:

Theorem 12.3.3 (Sato-Tate Equidistribution in Families)**.** *Recall the notation for constants in the statement of the main theorem 11.4.1. Assume that $|\mathfrak{n}_i|$ grows faster than any power of q_{v_i}. Then for all continuous $\widehat{f}$ on $\widehat{G}_\theta^{\mathrm{ur}}$,*

$$|\mathfrak{n}_i|^{-R} \Gamma_L(\mathfrak{n}_i)^{-1} \mu_{\mathfrak{n}_i, v_i}^{\pi_0}(\widehat{f}) \to C(\pi_0) \mu_\theta^{\mathrm{ST}(\pi_0)}(\widehat{f})$$

as $i \to \infty$ and where normalizing constant

$$C(\pi_0) = \frac{\mathrm{vol}(H(F)\backslash H(\mathbb{A}_f))}{\mathrm{vol}(K_H^S)} \sum_{\Delta \in \Delta^{\max}(\pi_0)} \mathbf{1}_{\eta_{\pi_0}^{\psi_\infty^\Delta}(\mathcal{S}_\Delta)=1} \frac{\dim \lambda_1(\Delta)}{|\Pi_{\mathrm{disc}}(\lambda_1(\Delta))|}.$$

Proof. By the Weierstrass approximation argument in Remark 9.5 of [73], it suffices to show that

$$|\mathfrak{n}_i|^{-R} \Gamma_L(\mathfrak{n}_i)^{-1} \mu_{\mathfrak{n}_i, v_i}^{\pi_0}(\widehat{f}_{v_i}) \to C(\pi_0) \mu_\theta^{\mathrm{ST}(\pi_0)}(\widehat{f}_{v_i})$$

for $f_{v_i} \in \mathscr{H}^{\mathrm{ur}}(G_{v_i})$ (note that this Hecke algebra is constant on $\mathcal{V}_F(\theta)$). We do this by applying Theorem 11.4.1 with $S_1 = \{v_i\}$. Note that the growth condition on $|\mathfrak{n}_i|$ shows that we will eventually have $|\mathfrak{n}_i| \geq D q_{v_1}^{E\kappa(f_{v_i})}$.

After using formula (6.2.4) and Fourier inversion to get that

$$f_{S_1}^{H_1}(1) \coprod_{i>1} \int_{H_{i,\mathrm{der},S_1}} f_{S_1}^{H_i}(h)\, dh = \mathcal{T}_\Delta f_{S_1}(1)$$
$$= \mu^{\mathrm{pl,ur}}(H_v)(\widehat{\mathcal{T}_\Delta f_{S_1}}) = \mu^{\mathrm{pl,ur}}(H_v)(\widehat{f}_{S_1} \circ \mathcal{S}_\Delta),$$

the argument follows exactly as that for 9.26 in [73]. □

We repeat an interpretation from the introduction: recall that as part of the conjectures surrounding Langlands functoriality, every automorphic representation π on some G/F should correspond to a group H_π/F that is the smallest group it is a functorial transfer from. The Satake parameters σ_{π_v} for v ranging over a particular $\mathcal{V}_F(\theta)$ are then expected to equidistribute according to a Sato-Tate distribution coming from H.

At the current time, actually finding H appears out of reach. However, in reasonable families of automorphic representations, most π should correspond to some fixed "maximal" H that is computable. Therefore, if we look at Satake parameters over the entire family, we can hopefully prove an equidistribution-on-average result towards the Sato-Tate measure for this maximal H.

This is conceptually what is happening here: most automorphic representations with π_0 at infinity come from group $H_{\pi_0} = H'$. Therefore the σ_{π_v} ranging over both v and a reasonable family of such π should equidistribute according to Sato-Tate measures from H'. Unlike previous cases built off of [73], we are in a more complicated situation where this maximal H_{π_0} isn't actually G itself.

12.4. **Sarnak-Xue Conjecture.** As an application of Theorem 11.4.2, we prove certain cases of the Sarnak-Xue conjecture of [68] for unitary groups. This conjecture is stated in terms of classical symmetric spaces instead of adelic quotients. Consider a reductive G/F, open compact $U \subseteq G^\infty$, and π_0 a unirrep of G_∞. Let $\Gamma(U) = U \cap G(F)$ and let

$$m(\pi_0, \Gamma(U)) := \dim \operatorname{Hom}(\pi_0, L^2(\Gamma(U)\backslash G_\infty)).$$

Note that $\Gamma(U)\backslash G_\infty$ is a connected component of the adelic quotient

$$Y(U) := G(F)\backslash G(\mathbb{A})/U.$$

Then:

Conjecture 12.4.1 (Cohomological Sarnak-Xue density hypothesis)**.** *Let U_i be an sequence of open compacts of G^∞ decreasing to the identity. Then for all cohomological unirreps π_0 of G_∞:*

$$m(\pi_0, \Gamma(U_i)) \ll_\epsilon \operatorname{vol}(\Gamma(U_i)\backslash G_\infty)^{\frac{2}{p(\pi_0)}+\epsilon},$$

where $p(\pi_0)$ is the infimum over p such that the K-finite matrix coefficients of π_0 are in $L^p(G_\infty)$.

In this section, we will study this conjecture for the case of unitary groups and U_i decreasing through only increasing principal-congruence levels at split places. We will prove it for all π_0 except for those that have a single particular representation on $U(2,2)$ as a factor.

It is important to mention that the work [52] already had strong enough bounds to achieve the Sarnak-Xue threshold. We did not need to improve these in any way, only use them as input (through Lemmas 9.4.2 and 9.4.3) for our more general framework.

12.4.1. *Computing $p(\pi)$.* Before we can check this for unitary groups, we first need extend the computations of [34] to find a way to understand $p(\pi)$ in terms of our parameterization of cohomological representations.

First, given (possibly ordered) bipartition $Q = ((p_1, q_1), \dots, (p_r, q_r))$, let

$$\Xi(B) = (\chi_j(B))_j$$

be the list of numbers obtained by setting $m_i = \min\{p_i, q_i\}$ and $n_i = p_i + q_i$, concatenating the lists

$$\bigsqcup_{i:m_i \neq 0} (n_i - 1, n_i - 3, \dots, n_i - 2m_i + 1)$$

and reordering it to be decreasing. For indexing purposes, we define $\chi_j(B) = 0$ for j out of bounds. Also define

$$\sigma_j(B) = \sum_{k \leq j} \chi_k(B) \tag{12.4.1}$$

for all j.

Proposition 12.4.2. *Let $G = U(p,q)$ and π_0 be the cohomological representation of G associated to $B = ((p_1, q_1), ..., (p_r, q_r))$. Then,*

$$\frac{2}{p(\pi_0)} \geq 1 - \max_i \left\{ \frac{\sigma_i(B)}{i(N-i)} \right\}.$$

Proof. First, we recall how to obtain, from results in [47, §§7-8], a formula to compute $p(\pi)$, when $\pi = J(S, \sigma, \nu)$ is a Langlands quotient. To describe such a quotient we need:

- S_0, a minimal parabolic of G, with Langlands decomposition $S_0 = M_0 A_0 N_0$, whose respective subgroups have Lie algebras $\mathfrak{m}_0$, $\mathfrak{a}_0$, and $\mathfrak{n}_0$,
- $\alpha_1, ..., \alpha_{\dim \mathfrak{a}_0}$ the simple roots of $\mathfrak{a}_0$ in $\mathfrak{g}$, and $\omega_1, ..., \omega_{\dim \mathfrak{a}_0}$ the basis of $\mathfrak{a}_0$ dual to the α_i,
- ρ_0 the corresponding half-sum of positive roots of $\mathfrak{a}_0$ in $\mathfrak{g}$,
- $S = MAN$, a parabolic subgroup of G standard with respect to S_0, with Lie algebras $\mathfrak{m}$, $\mathfrak{a}$, and $\mathfrak{n}$,
- a discrete series representation σ of M,
- a weight $\nu \in \mathfrak{a}^*$ such that $\langle \nu, \alpha \rangle > 0$ for all roots α of $\mathfrak{a}$ in $\mathfrak{n}$.

Then the parabolic induction $I(S, \sigma, \nu)$ has a unique Langlands quotient $J(S, \sigma, \nu)$. We have a direct sum decomposition

$$\mathfrak{a}_0 = \mathfrak{a} \oplus \mathfrak{a}_M,$$

where $\mathfrak{a}_M$ is the Lie algebra of the maximal split torus of M. Define $\nu_0 \in \mathfrak{a}_0^*$ by extending ν by zero on $\mathfrak{a}_M$. Proposition 5.13 of [34] then deduces from [47, §§7-8] the inequality

$$p(J(S, \sigma, \nu)) \leq \inf \left\{ p \geq 2 \mid p > \frac{2\langle \rho_0, \omega_j \rangle}{\langle \rho_0 - \nu_0, \omega_j \rangle} \text{ for all } \omega_j \right\}. \tag{12.4.2}$$

We now compute the appropriate pairings for π_0 corresponding to the bipartition $B = ((p_1, q_1), ..., (p_r, q_r))$, by first writing it as a Langlands quotient following [78, §6]. Begin with the Levi subgroup

$$L = U(p_1, q_1) \times ... \times U(p_r, q_r)$$

associated to π_0. Let K be the maximal compact of G and let $(K \cap L)A_L N_L$ the Iwasawa decomposition of L. Define ν_L to be the the half-sum of positive roots of A_L acting on N_L, and let $Z = \{\alpha \text{ a root of } \mathfrak{a}_L \text{ in } \mathfrak{g} \mid \langle \alpha, \nu_L \rangle = 0\}$, and let

$$A = \bigcap\nolimits_{\alpha \in \mathbb{Z}} \ker \alpha \subseteq A_L.$$

Then M is defined to be the centralizer of A in G. Define also M_L to be the centralizer of A_L in G, and let $S_L = M_L A_L N_L$ be any parabolic subgroup of G with respect to which ν_L is dominant. Then by construction, there is a unique parabolic subgroup S with Levi MA and containing S_L. Let $\nu = \nu_L \mid_A$. Then following [78, Theorem 6.16], there exists a discrete series representation σ of M such that $\pi_0 = J(S, \sigma, \nu)$.

We now use this construction to compute the pairings in (12.4.2); to do so, we write everything in coordinates. To begin, we choose a posteriori a minimal parabolic subgroup S_0 for which S is standard. Define numbers $m_\star = \min\{p_\star, q_\star\}$ and $n_\star = p_\star + q_\star$. Each $U(p_\star, q_\star)$ has minimal parabolic corresponding to the partition

$$(1^{(m_\star)}, n_\star - 2m_\star, 1^{(m_\star)})$$

and maximal split torus isomorphic to $\mathbb{R}^{m_\star}$, which we embed in a maximal torus $T_\star$. Fix coordinates in the standard way: i.e. so that we can write

$$X^*(T_\star) = \mathbb{Z}\langle e_1, \dots, e_{n_\star} \rangle$$

with simple roots $\alpha_i = e_i - e_{i+1}$. We can therefore realize as a vector:

$$\rho_0 = \left(\frac{n-1}{2}, \frac{n-3}{2}, \dots, \frac{n-2m+1}{2}, 0^{(n-2m)}, \frac{-n+2m-1}{2}, \dots, \frac{-n+3}{2}, \frac{-n+1}{2} \right)$$

in $X^*(T)$. In $X^*(A_0)$, this becomes

$$\rho_0 = \rho_{p,q} = (n-1, n-3, \dots, n-2m+1).$$

Similarly, $\nu_L \in X^*(A_L)$ is the concatenation of sequences

$$\bigsqcup_{k=1}^{r} \rho_{p_i, q_i} \tag{12.4.3}$$

reordered to be decreasing (the reordering comes from choosing the S_0 with respect to which which S_L is standard). The subtorus $A \subset A_L$ is then chosen so that $\langle \mathrm{Re}(\nu_L), \alpha \rangle > 0$ for all simple roots α of A. We have further direct sum decompositions

$$A_0 = A' \oplus A_L = A' \oplus A'' \oplus A$$

where A' is the maximal split torus of L and $A' \oplus A''$ is that of M. The extension of ν by 0 to A_L is then just ν_L again. Let ν_0 be the common extension by 0 to A_0, obtained in coordinates by adding a string of zeros to (12.4.3). Following (12.4.2), we have

$$\frac{2}{p(\pi_0)} \geq 1 - \max_i \left\{ \frac{\langle \nu_0, \omega_i \rangle}{\langle \rho_0, \omega_i \rangle} \right\}.$$

By symmetry of the ν_0 and ρ_0, the maximum value is achieved for some $i \leq m$. In this case, we check that $2\langle \rho_0, \omega_i \rangle = i(N-i)$ and $2\langle \nu_0, \omega_i \rangle = \sigma_i(B)$. $\square$

12.4.2. *Some Combinatorial Lemmas.* We next need some more involved but elementary combinatorial bounds, this time for the $\sigma_i(B)$ defined in (12.4.1). Once again, part of the complexity of this section is due to our use the suboptimal bound $R(\Delta)$ from Corollary 9.4.4 instead of the conjectural optimal bound $R_0(\Delta)$ from 9.6.3.

First, for a (possibly ordered) partition Q, define $\sigma_i(Q) = \sigma_i(B)$ for the $B = ((p_i, q_i))_i \in \beta^{-1}(Q)$ such that $|p_i - q_i| \leq 1$ so that for all i, $\sigma_i(B) \leq \sigma_i(\beta(B))$. Next:

Lemma 12.4.3. *Let $d < N \in \mathbb{Z}^+$.*

(1) If
$$Q_d := \left(d^{(\lfloor N/d \rfloor)}, N - d\lfloor N/d \rfloor\right),$$
then for all $Q = (n_1, \dots, n_r)$ a partition of N with each part of size $\leq d$, $\sigma_i(Q) \leq \sigma_i(Q_d)$ for all i.

(2) Asssume $N \geq 2d$. If
$$Q'_d := \begin{cases} \left(d^{(\lfloor N/d \rfloor - 1)}, d-1, N - d\lfloor N/d \rfloor + 1\right) & N \not\equiv -1 \pmod{d} \\ \left(d^{(\lfloor N/d \rfloor - 1)}, d-1, d-1, 1\right) & N \equiv -1 \pmod{d} \end{cases},$$
then for all $Q = (n_1, \dots, n_r)$ a partition of N with each part of size $\leq d$ and at most $\lfloor n/d \rfloor - 1$ parts of size d (i.e. not equal to Q_d), $\sigma_i(Q) \leq \sigma_i(Q'_d)$ for all i.

Proof. For the first claim, choose such $Q = (n_1, \dots, n_r) \neq Q_d$ without loss of generality in decreasing order. Then $n_r, n_{r-1} < d$ so $Q' = (n_1, \dots, n_{r-2}, n_{r-1} + 1, n_r - 1)$ also satisfies the conditions. Switching any pair (a, b) with $a \geq b$ to $(a+r, b-r)$ in $\Xi(Q)$ will increase or keep equal every $\sigma_i(Q)$. With this in mind, $\Xi(Q')$ differs from $\Xi(Q)$ by removing the numbers
$$(n_r - 1, n_r - 3, \dots,), (n_{r-1} - 1, n_{r-1} - 3, \dots,)$$
and adding in the numbers
$$(n_r - 2, n_r - 4, \dots,), (n_{r-1}, n_{r-1} - 2, \dots,).$$
Since $n_{r-1} \geq n_r$, this is a sequence of pair-switches as above together with some strict increases of coordinates. Therefore $\sigma_i(Q') \geq \sigma_i(Q)$ for all i.

Repeating this process until producing Q_d proves the first part. The second part follows by similar argument. $\square$

Lemma 12.4.4. *With notation from Lemma 12.4.3:*
$$\max_i \left\{ \frac{\sigma_i(Q_d)}{i(N-i)} \right\} = \frac{d-1}{N - \lfloor N/d \rfloor}.$$
Furthermore, if $N \geq 2d$,
$$\max_i \left\{ \frac{\sigma_i(Q'_d)}{i(N-i)} \right\} = \frac{d-1}{N - \lfloor N/d \rfloor + 1}.$$
Proof. This is by computer check. $\square$

This gives our final results:

Proposition 12.4.5. *Let $B \in \mathcal{P}_1(p,q)$ with $p+q = N$ and let π_0 be the cohomological representation of $U(p,q)$ corresponding to β. Let $Q \in Q^{\max}(B)$ and assume $\beta(B) \neq (2,2)$. Then*

$$(N^2-1)\left(1 - \max_i \left\{ \frac{\sigma_i(B)}{i(N-i)} \right\}\right) \geq R(Q) - 1$$

with equality only if Q has a single element. (Recall the definition of $R(Q)$ from Corollary 9.4.4.)

Proof. It suffices to prove the bound with

$$\max_i \left\{ \frac{\sigma_i(\beta(B))}{i(N-i)} \right\} \geq \max_i \left\{ \frac{\sigma_i(B)}{i(N-i)} \right\}$$

instead. Let d be the maximal element of $\beta(B)$. First, if $N < 2d$, then $\beta(B) = (d, (a_i)_i)$ for $\sum_i a_i < d$. Then by Lemma 12.4.3,

$$\max_i \left\{ \frac{\sigma_i(\beta(B))}{i(N-i)} \right\} \leq \frac{d-1}{N-1}.$$

Furthermore, by Lemma 11.2.5, Q has a part of size d so by Lemma 11.2.7,

$$R(Q) \leq R(d, 1^{(N-d)}) = N(N-d) + 1.$$

The result then follows from

$$1 - \frac{d-1}{N-1} = \frac{N-d}{N-1} > \frac{N(N-d)}{N^2-1}.$$

Therefore assume $N \geq 2d$. If $\beta(B) \neq Q_d$, then by Lemmas 12.4.3 and 12.4.4, we have that

$$\max_i \left\{ \frac{\sigma_i(\beta(B))}{i(N-i)} \right\} \leq \max_i \left\{ \frac{\sigma_i(\beta(Q_d'))}{i(N-i)} \right\} = \frac{d-1}{N - \lfloor N/d \rfloor + 1}.$$

Using again that $R(Q) \leq N(N-d)+1$, the result then follows since

$$1 - \frac{d-1}{N - \lfloor N/d \rfloor + 1} > \frac{N(N-d)}{N^2-1}$$

always.

If on the other hand $\beta(B) = Q_d$, then

$$\max_i \left\{ \frac{\sigma_i(\beta(B))}{i(N-i)} \right\} = \frac{d-1}{N - \lfloor N/d \rfloor}.$$

In addition, by Lemma 11.2.5, $Q = Q_d$ so

$$R(\beta(B)) \leq \bar{R}(Q) = \frac{1}{2}\left(N^2 + \left\lfloor \frac{N}{d} \right\rfloor^2 d + N - d \left\lfloor \frac{N}{d} \right\rfloor\right).$$

By a computer check, the desired inequality

$$1 - \frac{d-1}{N - \lfloor N/d \rfloor} > \frac{1}{N^2-1}\left(\frac{1}{2}\left(N^2 + \left\lfloor \frac{N}{d} \right\rfloor^2 d + N - d \left\lfloor \frac{N}{d} \right\rfloor\right) - 1\right)$$

is true except for the cases

$$\beta(B) = (d,d), \quad \beta(B) = (d,d,1), \quad \beta(B) = (2,2,2).$$

All these cases except $(2,2)$ can be checked by using the tighter bound $R(Q)$ instead of $\bar{R}(Q)$. $\square$

Extending to all number fields F:

Corollary 12.4.6. *Let G be an extended pure inner form of some $G^* \in \mathcal{E}_{\mathrm{sim}}(N)$ and let $\pi_0 = \prod_v \pi_v$ be a cohomological representation of G_∞ such that $\Delta^{\max}(\pi_0) \neq \emptyset$ and where each π_v corresponds to bipartition $B_v \in \mathcal{P}_1(p_v, q_v)$.*

Then, for all B_v such that $\beta(B_v) \neq (2,2)$,

$$(N^2-1)\left(1 - \max_i \left\{ \frac{\sigma_i(B_v)}{i(N-i)} \right\}\right) \geq R(\pi_0) - 1$$

with equality only if π_∞ is a character.

Proof. For all v, $R(\pi_v) \geq R(\pi_0)$ since it is a maximum over a larger set by Lemma 11.2.6. The result then follows from Proposition 12.4.5. □

12.4.3. *Sarnak-Xue Density.* We can now return to the original setup and specialize to unitary groups. Let G be an extended pure inner form of some $G^* \in \mathcal{E}_{\mathrm{sim}}(N)$. We choose:

- a cohomological representation $\pi_0 = \prod_v \pi_v$ of G_∞,
- a finite set of places S_0 containing all places where G is ramified,
- an ideal $\mathfrak{n}$ relatively prime to S_0 (we will compute asymptotics as $\mathfrak{n} \to \infty$),
- an open compact $U_{S_0} \subseteq G_{S_0}$.

We then define

$$U_{\mathfrak{n}} = U_{S_0} K^G(\mathfrak{n})$$

using the principle congruence subgroups associated to $\mathfrak{n}$.

Theorem 12.4.7 (Cohomological Split-level Sarnak-Xue Density for Unitary Groups)**.** *With setup as above, assume:*

- *$\mathfrak{n}$ is only divisible by places of F that split in E.*
- *If $N = 4$: for each v with $G_v = U(2,2)$, $\pi_v \neq ((1,1),(1,1))$.*

Then,

$$m(\pi_0, \Gamma(U)) \ll_\epsilon \mathrm{vol}(\Gamma(U_{\mathfrak{n}}) \backslash G_\infty)^{\frac{2}{p(\pi)}+\epsilon}.$$

(The ϵ may be removed if π_0 isn't a character).

Proof. As in §1.1 of [52], $Y(U_{\mathfrak{n}})$ contains $\gg_\epsilon |\mathfrak{n}|^{1-\epsilon}$ copies of $\Gamma(U_{\mathfrak{n}})\backslash G_\infty$ and

$$\mathrm{vol}(\Gamma(U_{\mathfrak{n}})\backslash G_\infty) \gg_\epsilon |\mathfrak{n}|^{N^2-1+\epsilon}.$$

The bound on connected components gives us that

$$\begin{aligned} m(\pi_0, \Gamma(U_{\mathfrak{n}})) &\ll_\epsilon |\mathfrak{n}|^{-1+\epsilon} \dim \mathrm{Hom}(\pi_0, L^2(Y(U_{\mathfrak{n}}))) \\ &= |\mathfrak{n}|^{-1+\epsilon} \sum_{\pi \in \mathcal{AR}_{\mathrm{disc}}(G)} \mathbf{1}_{\pi_\infty = \pi_0} \dim((\pi^\infty)^{U_{\mathfrak{n}}}). \end{aligned}$$

We can bound the sum by Theorem 11.4.2 with $S_1 = \emptyset$ and $\varphi_{S_0} = \bar{\mathbf{1}}_{U_{S_0}}$ to get

$$m(\pi_0, \Gamma(U_{\mathfrak{n}})) \ll_\epsilon |\mathfrak{n}|^{R(\pi_0)-1+\epsilon}.$$

Next, Proposition 12.4.2 computes

$$\frac{2}{p(\pi)} = \min_v \frac{2}{p(\pi_v)} \geq \min_v \left(1 - \max_i \left\{ \frac{\sigma_i(B_v)}{i(N-i)} \right\}\right),$$

where each π_v corresponds to bipartition B_v and with a strict inequality if π_0 isn't a character. The result follows from applying the bound in Corollary 12.4.6 and the volume estimate. □

Of course, either by varying φ_{S_0} beyond just the indicator function of U_{S_0} or by using non-trivial S_1, we can prove similar results for very general weighted counts of representations. These don't have as clean a statement in terms of the classical symmetric spaces $\Gamma(U_i)\backslash G_\infty$ however.

The leftover $\pi_0 = ((1,1),(1,1))$ is likely an artifact of $R(\Delta)$ not being optimal. However, provably improving it enough seems to be hard—see [51] (the conjectural optimal bound $R_0(\Delta)$ of Conjecture 9.6.3 would of course be enough).

12.4.4. *Examples.* We compute some small cases where π_0 on a group of rank N is the same value π_v at all infinite places v so that $R(\pi_0) = R(\pi_v)$. Let $\pi_v = B \in \mathcal{P}_1(p_v, q_v)$. Let Q be the unordered partition corresponding to $\beta(B)$. If π_v is GSK-maxed, let

$$Q = (d_1, \dots, d_r, 1^{(k)})$$

with $d_1 > d_2 > \dots > d_r > 1$. Then we can compute

$$R(\pi_v) - 1 = R_0(\pi_v) - 1 = \frac{1}{2}\left(N^2 + k^2 - \sum_i d_i^2\right) + (r-1),$$

so we get

$$m(\pi_0, \Gamma(U_\mathfrak{n})) \ll_\epsilon |\mathfrak{n}|^{\frac{1}{2}\left(N^2+k^2-\sum_i d_i^2\right)+(r-1)+\epsilon}.$$

This is conjecturally the exact exponent and provably so in the odd GSK-case for unramified E/F and U_i as in Theorem 11.4.1. We also have

$$\frac{2}{p(\pi_0)} \geq \frac{N-d}{N-1}$$

so the Sarnak-Xue bound asks for $m(\pi_0, \Gamma(U_{/mfn}))$ to be asymptotically less than

$$\operatorname{vol}(\Gamma(U_\mathfrak{n})\backslash G_\infty)^{\frac{2}{p(\pi_0)}} \gg_\epsilon |\mathfrak{n}|^{(N+1)(N-d)-\epsilon},$$

which is always true. If $r = 1$, we save a factor of $|\mathfrak{n}|^{N-d}$ over the Sarnak-Xue bound. If we keep n and k fixed but increase r, the saving is even larger.

When π_v isn't GSK-maxed, the formulas are much more complicated and $R(\pi_v) \neq R_0(\pi_v)$. Table 12.4.1 lists some values based on $Q = \beta(\pi_v)$. For each Q, we list the maximum possible values of $R(\pi_0) - 1$ and $R_0(\pi_0) - 1$, which in our setup only depends on Q. These are the provable and conjectural exponents on $|\mathfrak{n}|$ in the growth rate of $m(\pi_0, \Gamma(\mathfrak{n}))$ respectively. We also list the target exponent from the Sarnak-Xue density bound and the "trivial bound" growth rate $N^2 - 1$ when π_0 is discrete series. Finally, we italicize cases where a failure of Lemma 11.2.7 occurs due to Q not being GSK.

Our $R(Q)$ beats target growth rate in every case except the bolded number when $Q = (2,2)$. The improvement is often large, though not in some cases like $(3,3), (4,4)$ and $(2,2,2,2)$. The conjectural optimum $R_0(Q)$ is usually much smaller still.

We in particular want to point out the case of $(2,2,2,2,1,1)$ vs $(2,2,2,2,2)$ where 11.2.7 fails even for the conjectural bound R_0. This suggests very strange behavior—for example, different asymptotic growth rates for $H^{a,b}$ with the same sum of $a+b$, or where the average Sato-Tate distribution for the family may differ depending on whether one weights the average by counts of automorphic forms or representations.

TABLE 12.4.1. Comparison of Growth Rates

Arthur-SL_2 Q	provable: $\max R(\pi_0)-1$	conjectural: $\max R_0(\pi_0)-1$	SX goal: $2(N^2-1)p(\pi_v)^{-1}$	trivial: N^2-1
(2,2)	**8**	6	**7.5**	15
(2,2,1)	13	11	16	24
(2,2,2)	21	17	23.33	35
(2,2,1,1)	*21*	18	26.25	35
(3,3)	17	11	17.5	35
(2,2,2,1)	28	24	36	48
(3,3,1)	24	18	28.8	48
(3,2,2)	21	19	32	48
(2,2,2,2)	47	33	47.25	63
(2,2,2,1,1)	*47*	33	50.4	63
(4,4)	30	18	31.5	63
(3,3,3)	43	32	53.33	80
(3,2,2,2)	40	36	60	80
(5,5)	47	27	49.5	99
(2,2,2,2,2)	74	54	79.2	99
(2,2,2,2,1,1)	*74*	*54*	82.5	99

REFERENCES

[1] Miklos Abert, Nicolas Bergeron, Ian Biringer, Tsachik Gelander, Nikolay Nikolav, Jean Raimbault, and Iddo Samet. On the growth of l^2-invariants for sequences of lattices in lie groups. *Annals of Mathematics*, 185(3):711–790, 2017.

[2] Jeffrey Adams, Dan Barbasch, and David A. Vogan, Jr. *The Langlands classification and irreducible characters for real reductive groups*, volume 104 of *Progress in Mathematics*. Birkhäuser Boston, Inc., Boston, MA, 1992.

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

[4] Nicolás Arancibia, Colette Mœglin, and David Renard. Paquets d'arthur des groupes classiques et unitaires. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 27, pages 1023–1105, 2018.

[5] Nicolás Arancibia Robert. Characteristic cycles, micro local packets and packets with cohomology. *Trans. Amer. Math. Soc.*, 375(2):997–1049, 2022.

[6] Nicolás Arancibia Robert and P Mezo. Equivalent definitions of arthur packets for real quasisplit unitary groups. *arXiv preprint arXiv:2204.10715*, 2022.

[7] James Arthur. The L^2-Lefschetz numbers of Hecke operators. *Invent. Math.*, 97(2):257–290, 1989.

[8] James Arthur. Unipotent automorphic representations : conjectures. In *Orbites unipotentes et représentations - II. Groupes p-adiques et réels*, number 171-172 in Astérisque. Société mathématique de France, 1989.

[9] James Arthur. Automorphic representations of GSp(4). In *Contributions to automorphic forms, geometry, and number theory*, pages 65–81. Johns Hopkins Univ. Press, Baltimore, MD, 2004.

[10] James Arthur. An introduction to the trace formula. In *Harmonic analysis, the trace formula, and Shimura varieties*, volume 4 of *Clay Math. Proc.*, pages 1–263. Amer. Math. Soc., Providence, RI, 2005.

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

[12] Avner Ash, Paul E Gunnells, and Mark McConnell. Cohomology of congruence subgroups of sl (4, z) ii. *Journal of Number Theory*, 128(8):2263–2274, 2008.

[13] Thomas Barnet-Lamb, Toby Gee, and David Geraghty. The Sato-Tate conjecture for Hilbert modular forms. *J. Amer. Math. Soc.*, 24(2):411–469, 2011.

[14] Tom Barnet-Lamb, David Geraghty, Michael Harris, and Richard Taylor. A family of Calabi-Yau varieties and potential automorphy II. *Publ. Res. Inst. Math. Sci.*, 47(1):29–98, 2011.

[15] Nicolas Bergeron and Laurent Clozel. *Spectre automorphe des variétés hyperboliques et applications topologiques*. Société mathématique de France, 2005.

[16] Valentin Blomer. Density theorems for $${\textrm{gl}}(n)$$. *Inventiones mathematicae*, 2022.

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

[18] Frank Calegari and Nathan M Dunfield. Automorphic forms and rational homology 3–spheres. *Geometry & Topology*, 10(1):295–329, 2006.

[19] Frank Calegari and Matthew Emerton. Bounds for multiplicities of unitary representations of cohomological type in spaces of cusp forms. *Annals of mathematics*, pages 1437–1446, 2009.

[20] Gaëtan Chenevier and David Renard. Level one algebraic cusp forms of classical groups of small rank. *Mem. Amer. Math. Soc.*, 237(1121):v+122, 2015.

[21] Laurent Clozel. On limit multiplicites of discrete series representations in spaces of automorphic forms. 1986.

[22] Laurent Clozel. On the cohomology of kottwitz's arithmetic varieties. *Duke Mathematical Journal*, 72(3):757–795, 1993.

[23] Laurent Clozel and Patrick Delorme. Le théorème de Paley-Wiener invariant pour les groupes de Lie réductifs. II. *Ann. Sci. École Norm. Sup. (4)*, 23(2):193–228, 1990.

[24] Mathieu Cossutta. Asymptotique des nombres de Betti des variétés arithmétiques. *Duke Mathematical Journal*, 150(3):443 – 488, 2009.

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

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

[27] David L DeGeorge and Nolan R Wallach. Limit formulas for multiplicities in. *Annals of Mathematics*, pages 133–150, 1978.

[28] Pierre Deligne. Travaux de shimura. In *Séminaire Bourbaki vol. 1970/71 Exposés 382–399*, pages 123–165. Springer, 1971.

[29] Pierre Deligne. Variétés de shimura: interprétation modulaire, et techniques de construction de modeles canoniques. In *Automorphic forms, representations and L-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part*, volume 2, pages 247–289, 1979.

[30] Shai Evra and Ori Parzanchevski. Ramanujan complexes and golden gates in $PU(3)$. *Geom. Funct. Anal.*, 32(2):193–235, 2022.

[31] Axel Ferrari. Théorème de l'indice et formule des traces. *Manuscripta Math.*, 124(3):363–390, 2007.

[32] Tobias Finis, Erez Lapid, and Werner Müller. Limit multiplicities for principal congruence subgroups of $\mathrm{GL}(n)$ and $\mathrm{SL}(n)$. *J. Inst. Math. Jussieu*, 14(3):589–638, 2015.

[33] Mikolaj Fraczyk, Gergely Harcos, Peter Maga, and Djordje Milicevic. The density hypothesis for horizontal families of lattices, 2020.

[34] Mathilde Gerbelli-Gauthier. Limit multiplicity for unitary groups and the stable trace formula, 2021.

[35] Konstantin Golubev and Amitay Kamber. On sarnak's density conjecture and its applications, 2020.

[36] Benedict H. Gross. On the motive of a reductive group. *Invent. Math.*, 130(2):287–313, 1997.

[37] Paul E Gunnells, Mark McConnell, and Dan Yasaki. On the cohomology of congruence subgroups of gl3 over the eisenstein integers. *Experimental Mathematics*, 30(4):499–512, 2021.

[38] Thomas C. Hales. A simple definition of transfer factors for unramified groups. In *Representation theory of groups and algebras*, volume 145 of *Contemp. Math.*, pages 109–134. Amer. Math. Soc., Providence, RI, 1993.

[39] R. Howe and I. I. Piatetski-Shapiro. A counterexample to the "generalized Ramanujan conjecture" for (quasi-) split groups. In *Automorphic forms, representations and L-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part 1*, Proc. Sympos. Pure Math., XXXIII, pages 315–322. Amer. Math. Soc., Providence, R.I., 1979.

[40] Dihua Jiang, Dongwen Liu, and Lei Zhang. Arithmetic wavefront sets and generic l-packets, 2022.

[41] Tasho Kaletha. The local Langlands conjectures for non-quasi-split groups. In *Families of automorphic forms and the trace formula*, Simons Symp., pages 217–257. Springer, [Cham], 2016.

[42] Tasho Kaletha. Rigid inner forms of real and p-adic groups. *Annals of Mathematics*, 184(2):559–632, 2016.

[43] Tasho Kaletha. Global rigid inner forms and multiplicities of discrete automorphic representations. *Invent. Math.*, 213(1):271–369, 2018.

[44] Tasho Kaletha, Alberto Minguez, Sug Woo Shin, and Paul-James White. Endoscopic classification of representations: inner forms of unitary groups. *arXiv preprint arXiv:1409.3731*, 2014.

[45] Henry H. Kim, Satoshi Wakatsuki, and Takuya Yamauchi. Equidistribution theorems for holomorphic siegel cusp forms of general degree: the level aspect, 2021.

[46] Ju-Lee Kim, Sug Woo Shin, and Nicolas Templier. Asymptotic behavior of supercuspidal representations and sato-tate equidistribution for families, 2016.

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

[48] Robert E Kottwitz. Shimura varieties and λ-adic representations. *Automorphic forms, Shimura varieties, and L-functions*, 1:161–210, 1990.

[49] Jean-Pierre Labesse. Introduction to endoscopy: Snowbird lectures, revised version, May 2010 [revision of mr2454335]. In *On the stabilization of the trace formula*, volume 1 of *Stab. Trace Formula Shimura Var. Arith. Appl.*, pages 49–91. Int. Press, Somerville, MA, 2011.

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

[51] Simon Marshall. Endoscopy and cohomology of a quasi-split $U(4)$. In *Families of automorphic forms and the trace formula*, Simons Symp., pages 297–325. Springer, [Cham], 2016.

[52] Simon Marshall and Sug Woo Shin. Endoscopy and cohomology of u(n,1), 2018.

[53] Yozô Matsushima. A formula for the betti numbers of compact locally symmetric riemannian manifolds. *Journal of Differential Geometry*, 1(1-2):99–109, 1967.

[54] Arnab Mitra. A note on degenerate Whittaker models for general linear groups. *J. Number Theory*, 209:212–224, 2020.

[55] C. Mœglin and J.-L. Waldspurger. Modèles de Whittaker dégénérés pour des groupes p-adiques. *Math. Z.*, 196(3):427–452, 1987.

[56] Colette Mœglin and David Renard. Sur les paquets d'arthur des groupes unitaires et quelques conséquences pour les groupes classiques. *Pacific Journal of Mathematics*, 299(1):53–88, 2019.

[57] Colette Mœglin and J-L Waldspurger. Le spectre residuel de $gl(n)$. In *Annales scientifiques de l'École normale superieure*, volume 22, pages 605–674, 1989.

[58] Chung Pang Mok. *Endoscopic classification of representations of quasi-split unitary groups*. American Mathematical Soc., 2015.

[59] Arvind N Nair and Dipendra Prasad. Cohomological representations for real reductive groups. *Journal of the London Mathematical Society*, 104(4):1515–1571, 2021.

[60] James Newton. Serre weights and Shimura curves. *Proceedings of the London Mathematical Society*, 108(6):1471–1500, 12 2013.

[61] Masao Oi. Depth preserving property of the local langlands correspondence for non-quasi-split unitary groups. *Mathematical Research Letters*, 28(1):175–211, 2021.

[62] J. D. Rogawski. Trace Paley-Wiener theorem in the twisted case. *Trans. Amer. Math. Soc.*, 309(1):215–229, 1988.

[63] Jonathan D. Rogawski. *Automorphic representations of unitary groups in three variables*, volume 123 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, NJ, 1990.

[64] Manami Roy, Ralf Schmidt, and Shaoyun Yi. Dimension formulas for siegel modular forms of level 4, 2022.

[65] Susana A. Salamanca-Riba. On the unitary dual of real reductive Lie groups and the $A_{\mathfrak{g}}(\lambda)$ modules: the strongly regular case. *Duke Math. J.*, 96(3):521–546, 1999.

[66] Peter Sarnak. Notes on the generalized Ramanujan conjectures. In *Harmonic analysis, the trace formula, and Shimura varieties*, volume 4 of *Clay Math. Proc.*, pages 659–685. Amer. Math. Soc., Providence, RI, 2005.

[67] Peter Sarnak, Sug Woo Shin, and Nicolas Templier. Families of L-functions and their symmetry. In *Families of automorphic forms and the trace formula*, Simons Symp., pages 531–578. Springer, [Cham], 2016.

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

[69] François Sauvageot. Principe de densité pour les groupes réductifs. *Compositio Math.*, 108(2):151–184, 1997.

[70] Gordan Savin. Limit multiplicities of cusp forms. 1989.

[71] Joachim Schwermer and Steffen Kionke. On the growth of the first betti number of arithmetic hyperbolic 3-manifolds. *Groups, Geometry, and Dynamics*, 9(2):531–565, 2015.

[72] Jean-Pierre Serre. Propriétés conjecturales des groupes de Galois motiviques et des représentations l-adiques. In *Motives (Seattle, WA, 1991)*, volume 55 of *Proc. Sympos. Pure Math.*, pages 377–400. Amer. Math. Soc., Providence, RI, 1994.

[73] Sug Woo Shin and Nicolas Templier. Sato-Tate theorem for families and low-lying zeros of automorphic L-functions. *Invent. Math.*, 203(1):1–177, 2016. Appendix A by Robert Kottwitz, and Appendix B by Raf Cluckers, Julia Gordon and Immanuel Halupczok.

[74] Daniel Studenmund and Bena Tshishiku. Counting flat cycles in the homology of locally symmetric spaces, 2022.

[75] Marko Tadić. Spherical unitary dual of general linear group over non-Archimedean local field. *Ann. Inst. Fourier (Grenoble)*, 36(2):47–55, 1986.

[76] Olivier Taïbi. Dimensions of spaces of level one automorphic forms for split classical groups using the trace formula. *Ann. Sci. Éc. Norm. Supér. (4)*, 50(2):269–344, 2017.

[77] Peter E Trapa. Annihilators and associated varieties of a (λ) modules for u (p, q). *Compositio Mathematica*, 129(1):1–45, 2001.

[78] David A. Vogan, Jr. and Gregg J. Zuckerman. Unitary representations with nonzero cohomology. *Compositio Math.*, 53(1):51–90, 1984.

[79] J.-L. Waldspurger. Les facteurs de transfert pour les groupes classiques: un formulaire. *Manuscripta Math.*, 133(1-2):41–82, 2010.

[80] Bin Xu. On the cuspidal support of discrete series for p-adic quasisplit $Sp(N)$ and $SO(N)$. *Manuscripta Math.*, 154(3-4):441–502, 2017.