SPECTRAL ASPECT SUBCONVEX BOUNDS FOR $\mathrm{U}_{n+1} \times \mathrm{U}_n$

PAUL D. NELSON

ABSTRACT. Let (π, σ) traverse a sequence of pairs of cuspidal automorphic representations of a unitary Gan–Gross–Prasad pair $(\mathrm{U}_{n+1}, \mathrm{U}_n)$ over a number field, with U_n anisotropic. We assume that at some distinguished archimedean place, the pair stays away from the conductor dropping locus, while at every other place, the pair has bounded ramification and satisfies certain local conditions (in particular, temperedness). We prove that the subconvex bound

$$L(\pi \times \sigma, 1/2) \ll C(\pi \times \sigma)^{1/4-\delta}$$

holds for any fixed

$$\delta < \frac{1}{16n^5 + 56n^4 + 84n^3 + 72n^2 + 28n}.$$

Among other ingredients, the proof employs a refinement of the microlocal calculus for Lie group representations developed with A. Venkatesh and an observation of S. Marshall concerning the geometric side of the relative trace formula.

CONTENTS

Date: January 8, 2021.
2010 *Mathematics Subject Classification.* Primary 11F67; Secondary 11F70, 11M99, 35A27, 53D50.

1. Introduction

1.1. The refined GGP conjecture for unitary groups. Let F be a number field with adele ring $\mathbb{A}$. Let E/F be a quadratic extension, let V be an $(n+1)$-dimension hermitian space over E, and let W be an n-dimensional nondegenerate subspace of V. The pair of unitary groups $(G,H) := (\mathrm{U}(V), \mathrm{U}(W))$ will be referred to as a unitary Gan–Gross–Prasad (GGP) pair over F (see §3.4 for details).

Let S be a finite set of places of F, containing every archimedean place. We assume that S is sufficiently large in various technical senses enumerated in §3.6.

Let (π, σ) be a pair of cuspidal automorphic representations π of $G(\mathbb{A})$ and σ of $H(\mathbb{A})$, respectively, that are unramified outside S and tempered inside S. One may attach to this pair a branching coefficient $\mathcal{L}(\pi, \sigma)$, quantifying how automorphic forms in π correlate with those in σ (see §2.6 and §3.7).

Assume for the moment that π and σ are everywhere tempered, i.e., that each of their local components is tempered. As explained in [BCZ, §1.1.6] π and σ are known to admit base change lifts π_E and σ_E to $\mathrm{GL}_{n+1}(\mathbb{A}_E)$ and $\mathrm{GL}_n(\mathbb{A}_E)$, respectively. We write

$$L(\pi, \sigma, s) := L(\pi_E \otimes \sigma_E^\vee, s),$$

where a superscripted $\vee$ denotes the contragredient representation and the RHS is the finite part of the Rankin–Selberg L-function, given for s of large real part by a degree $2n(n+1)$ Euler product over the finite primes of F. It is known then that the central L-value $L(\pi, \sigma, 1/2)$ coincides up to mild factors with the branching coefficient $\mathcal{L}(\pi, \sigma)$ (see §3.8 for a more precise statement). This result, an affirmation of conjectures of Ichino–Ikeda [II] and N. Harris [Ha], was established in the stated generality recently by Beuzart-Plessis–Chaudouard–Zydor [BCZ, §1.1.6], generalizing earlier results of Wei Zhang [Zh2], Beuzart-Plessis [BP2, BP3] and Beuzart-Plessis–Liu–Zhang–Zhu [BLZZ].

1.2. Conductor dropping. We fix an archimedean place $\mathfrak{q} \in S$, which plays a privileged role in our analysis. According as the pair $(F_\mathfrak{q}, E_\mathfrak{q})$ of local components of our fields is $(\mathbb{R}, \mathbb{R} \times \mathbb{R})$ or $(\mathbb{R}, \mathbb{C})$ or $(\mathbb{C}, \mathbb{C} \times \mathbb{C})$, the pair $(G(F_\mathfrak{q}), H(F_\mathfrak{q}))$ of local components of our GGP pair is

$$(\mathrm{GL}_{n+1}(\mathbb{R}), \mathrm{GL}_n(\mathbb{R})), \quad (\mathrm{U}(p+1,q), \mathrm{U}(p,q)) \text{ with } p+q=n, \quad (\mathrm{GL}_{n+1}(\mathbb{C}), \mathrm{GL}_n(\mathbb{C})).$$

As explained in [NV, §15], the local factor of $L(\pi, \sigma, s)$ at the distinguished archimedean place $\mathfrak{q}$ is given in terms of the "archimedean Satake parameters" (see §7.4) $\lambda_{\pi,1}, \dots, \lambda_{\pi,n+1}$ of $\pi_\mathfrak{q}$ and $\lambda_{\sigma,1}, \dots, \lambda_{\sigma,n}$ of $\sigma_\mathfrak{q}$ by the formula

$$L_\mathfrak{q}(\pi, \sigma, s) = \begin{cases} \prod_{i,j,\pm} \Gamma_\mathbb{R}(s + (\pm(\lambda_{\pi,i} - \lambda_{\sigma,j}))^+ + a_{ij}) & \text{if } F = \mathbb{R}, \\ \prod_{i,j,\pm} \Gamma_\mathbb{C}(s + (\pm(\lambda_{\pi,i} - \lambda_{\sigma,j}))^+) & \text{if } F = \mathbb{C}, \end{cases} \tag{1.1}$$

for some $a_{ij} \in \{0,1\}$, where $(x+iy)^+ := |x| + iy$, $\Gamma_\mathbb{R}(s) := \pi^{-s/2}\Gamma(s/2)$ and $\Gamma_\mathbb{C}(s) := 2(2\pi)^{-s}\Gamma(s)$. The local analytic conductor at $\mathfrak{q}$ is thus

$$C_\mathfrak{q}(\pi, \sigma) \asymp \prod_{i,j} (1 + |\lambda_{\pi,i} - \lambda_{\sigma,j}|_{F_\mathfrak{q}})^2, \tag{1.2}$$

where $|x|_\mathbb{R} = |x|$ and $|x|_\mathbb{C} = |x|^2$.

Informally, we say that (π, σ) experiences *conductor dropping* at $\mathfrak{q}$ if the factors in (1.2) are not all of comparable size.

1.3. Main results. Our results address the subconvexity problem, a fundamental problem in the analytic theory of L-functions (see for instance [IS2, §2], [Mi, §4, §5], [MV1] and [MV2]). Informally, we show that if there is no conductor dropping at $\mathfrak{q}$, then the branching coefficient $\mathcal{L}(\pi, \sigma)$ enjoys a subconvex bound in the archimedean depth aspect at $\mathfrak{q}$. In the everywhere tempered case, this yields a subconvex bound for $L(\pi, \sigma, 1/2)$.

We now formulate our results more precisely.

The number field F and the GGP pair (G, H) are regarded as fixed, as is the finite set of places S taken large enough in the sense of §3.6.

We assume that H is anisotropic (so that $H(F)\backslash H(\mathbb{A})$ is compact) and that $G(F_\mathfrak{p})$ and $H(F_\mathfrak{p})$ are compact for all archimedean places $\mathfrak{p} \neq \mathfrak{q}$. The anisotropy of H excludes in particular the split case $E = F \times F$, corresponding to $(G, H) = (\mathrm{GL}_{n+1}(F), \mathrm{GL}_n(F))$.

For each "auxiliary" place $\mathfrak{p} \in S - \{\mathfrak{q}\}$, we fix "bounded subsets" $\Pi_{G,\mathfrak{p}}$ and $\Sigma_{H,\mathfrak{p}}$ of the respective unitary duals of $G(F_\mathfrak{p})$ and $H(F_\mathfrak{p})$. More precisely:

- For archimedean $\mathfrak{p} \neq \mathfrak{q}$ (so that $G(F_\mathfrak{p})$ and $H(F_\mathfrak{p})$ are compact), the sets $\Pi_{G,\mathfrak{p}}$ and $\Sigma_{H,\mathfrak{p}}$ are assumed to be finite.
- For non-archimedean $\mathfrak{p} \in S$, we require that $\Pi_{G,\mathfrak{p}}$ and $\Sigma_{H,\mathfrak{p}}$ have "bounded depth" in the sense that there is a compact open subgroup J of $G(F_\mathfrak{p})$ such that every element of $\Pi_{G,\mathfrak{p}}$ has a nonzero J-invariant vector, and similarly for $\Sigma_{H,\mathfrak{p}}$.

We fix a positive constant $c > 0$, which we use to quantify conductor dropping at the place $\mathfrak{q}$.

We denote by $\mathcal{F}$ the set of pairs (π, σ) of cuspidal automorphic representations of $G(\mathbb{A})$ and $H(\mathbb{A})$ with the following properties:

(i) π and σ have unitary central characters.

(ii) (π, σ) is locally distinguished: there is a nonzero $H(\mathbb{A})$-invariant functional $\pi \to \sigma$. Moreover, $(\pi_\mathfrak{q}, \sigma_\mathfrak{q})$ is orbit-distinguished (see §11.3).

(iii) For $\mathfrak{p} \notin S$, the local components $\pi_\mathfrak{p}$ and $\sigma_\mathfrak{p}$ are unramified, i.e., contain nonzero vectors invariant by $G(\mathbb{Z}_\mathfrak{p})$ and $H(\mathbb{Z}_\mathfrak{p})$ (see §3.6).

(iv) For $\mathfrak{p} \in S$, the local components $\pi_\mathfrak{p}$ and $\sigma_\mathfrak{p}$ are tempered.

(v) For $\mathfrak{p} \in S - \{\mathfrak{q}\}$, the local components $\pi_\mathfrak{p}$ and $\sigma_\mathfrak{p}$ belong to the bounded subsets $\Pi_{G,\mathfrak{p}}$ and $\Sigma_{H,\mathfrak{p}}$ of the respective unitary duals fixed above.

(vi) Let
$$T := \max(\{|\lambda_{\pi,i}|_{F_\mathfrak{q}}\} \cup \{|\lambda_{\sigma,j}|_{F_\mathfrak{q}}\})$$
denote the size of the largest archimedean Satake parameter of either $\pi_\mathfrak{q}$ or $\sigma_\mathfrak{q}$. Then for all i and j, we have
$$|\lambda_{\pi,i} - \lambda_{\sigma,j}|_{F_\mathfrak{q}} \geqslant cT. \tag{1.3}$$

We write $\mathcal{F}_T \subseteq \mathcal{F}$ for the subset corresponding to a given value of the parameter T defined above.

For $T \geqslant 1$ and $(\pi, \sigma) \in \mathcal{F}_T$, the ramification of π and σ at places $\mathfrak{p} \neq \mathfrak{q}$ is uniformly bounded, so the global analytic conductor $C(\pi, \sigma)$ is comparable to the local analytic conductor $C_\mathfrak{q}(\pi, \sigma)$. The content of the assumption (1.3) is thus that
$$C(\pi, \sigma) \asymp T^{2n(n+1)}.$$

We refer to §3.1 for our conventions on asymptotic notation and to §5.2.1 for the meaning of "ϑ-tempered."

Theorem 1.1. *Let T be sufficiently large. Let $(\pi, \sigma) \in \mathcal{F}_T$. Suppose for some fixed $\vartheta \in [0, 1/2)$ that σ is ϑ-tempered at every finite place $\mathfrak{p} \notin S$ that splits in E. Then the subconvex bound*

$$\mathcal{L}(\pi, \sigma) \ll T^{n(n+1)/2-\delta}$$

holds for each fixed

$$\delta < \frac{1-2\vartheta}{4(A+1-2\vartheta)}, \quad A := (2(n+1)^2 - n + 1/2)(n+1).$$

For example, if $\vartheta = 0$ and $n = 1, 2, \ldots, 10$, then the above bound may be written $\mathcal{L}(\pi, \sigma) \ll_\varepsilon T^{n(n+1)/2-\delta+\varepsilon}$, where δ is the reciprocal of

$$64, 202, 476, 934, 1624, 2594, 3892, 5566, 7664, 10234.$$

Corollary 1.2. *Let $(\pi, \sigma) \in \mathcal{F}$. Suppose that π and σ are everywhere tempered. Then the subconvex bound*

$$L(\pi, \sigma, 1/2) \ll C(\pi, \sigma)^{1/4-\delta}$$

holds for each fixed

$$\delta < \frac{1}{8n(n+1)((2(n+1)^2 - n + 1/2)(n+1) + 1)},$$

the RHS of which simplifies to the fraction indicated in the abstract.

For example, for $n = 1, 2, \ldots, 10$, we obtain $L(\pi, \sigma, 1/2) \ll_\varepsilon C(\pi, \sigma)^{1/4-\delta+\varepsilon}$, with δ the reciprocal of

$$256, 2424, 11424, 37360, 97440, 217896, 435904, 801504, 1379520, 2251480.$$

Proof of Corollary 1.2. We apply Theorem 1.1 with $\vartheta = 0$. By [BCZ, §1.1.6], we have $L(\pi, \sigma, 1/2) \asymp B_1 B_2 \mathcal{L}(\pi, \sigma)$, where B_1 denotes the value at $s = 1$ of a certain adjoint L-function and $B_2 := \prod_{\mathfrak{p} \in S} L_{\mathfrak{p}}(\pi, \sigma, 1/2)$. We bound $B_1 \ll T^\varepsilon$ cheaply by majorizing its Euler product on the line $\mathrm{Re}(s) = 1 + \varepsilon$, invoking the temperedness assumption, and applying Phragmen–Lindelöf. We bound $B_2 \ll 1$ using the temperedness assumption. □

Remark 1.3. There are many infinite families to which Corollary 1.2 provably applies. For instance, taking for E/F an imaginary quadratic extension of $\mathbb{Q}$ and starting with one pair $(\pi, \sigma) \in \mathcal{F}$ for which π and σ are known to be everywhere tempered (e.g., via the cohomology of Shimura varieties), we obtain an infinite family of such pairs by twisting either π or σ by any sequence of characters having fixed finite conductor and increasing analytic conductor at the archimedean place. Corollary 1.2 gives in such cases what one might call a "twisted t-aspect" subconvex bound on $\mathrm{U}_{n+1} \times \mathrm{U}_n$.

Remark 1.4. We have not optimized the numerical quality of our results. We have aimed instead for the simplest argument that yields an explicit exponent. We indicate throughout the text several places where, with additional effort, one should be able to improve that exponent (see Remarks 4.3, 6.2, 6.6, 14.7, 14.9). To indicate the potential for improvement, suppose $n = 1$. Corollary 1.2 then reads

$$L(\pi, \sigma, 1/2) \ll_\varepsilon C(\pi, \sigma)^{1/4-1/256+\varepsilon},$$

but an optimized version of the proof would be equivalent to the original argument of Iwaniec–Sarnak [IS1] (see Remark 2.4), which gives (at least for σ trivial) the much stronger bound

$$L(\pi, \sigma, 1/2) \ll_\varepsilon C(\pi, \sigma)^{1/4-1/24+\varepsilon}.$$

We expect similar potential for improvement in higher rank. We hope the pursuit of such improvement may serve as an interesting challenge.

Remark 1.5. In the special case of Theorem 1.1 in which σ is fixed, one can likely use Rankin–Selberg estimates to remove the dependence upon ϑ.

Remark 1.6. The interested reader will see that many of our local calculations apply over any local field. The restriction to the archimedean aspect comes primarily from invocation of the Kirillov formula, which is most readily available in the desired generality over archimedean local fields.

Remark 1.7. It would be an interesting challenge to adapt the proof to the split case $(G, H) = (\mathrm{GL}_{n+1}, \mathrm{GL}_n)$, possibly with π or σ an Eisenstein series. The non-compactness of the corresponding quotients presents significant difficulties.

1.4. Related results. There have been relatively few subconvex bounds in higher rank (i.e., for groups with a simple factor of rank $\geqslant 2$, such as GL_3). The first were X. Li's bounds [Li] for $\mathrm{GL}_3 \times \mathrm{GL}_1$ and $\mathrm{GL}_3 \times \mathrm{GL}_2$. In her setup, the GL_3 form is fixed (and self-dual) while the $\mathrm{GL}_2 / \mathrm{GL}_1$ form varies. Much recent work on the problem is similar in spirit (see for instance [Kh, Mu1, Mu2, HN, Ag, Mu3, Sh, LMS]).

The first subconvex bounds involving genuine variation of a form on a higher rank group were achieved in a seminal paper of Blomer–Buttcane [BB1], following their development of the GL_3 Kuznetsov formula and a deep study of the associated integral transforms [Blo, Bu1, Bu2]. They proved that $L(\pi, 1/2) \ll T^{3/4-1/120000}$ for full-level spherical Maass forms π on PGL_3 that are tempered at ∞ and whose parameter $\mu \in (i\mathbb{R})^3$ satisfies, with $T := \max(|\mu_1|, |\mu_2|, |\mu_3|)$, the following conditions for some fixed $c > 0$:

$$|\mu_k| \geqslant cT \quad (1 \leqslant k \leqslant 3) \quad \text{“no conductor dropping,”} \tag{1.4}$$

$$|\mu_i - \mu_j| \geqslant cT \quad (1 \leqslant i < j \leqslant 3) \quad \text{“avoidance of Weyl chamber walls.”} \tag{1.5}$$

Our assumption (1.3) is analogous to (1.4) (with the differences $\lambda_{\pi,i} - \lambda_{\sigma,j}$ playing the role of the μ_k), but we do not require any assumption like (1.5). Indeed, our method applies in the t-aspect, where $\mu_i - \mu_j \ll 1$.

Blomer–Buttcane generalized their result to the family of generalized principal series on GL_3 [BB2].

Simon Marshall (March 2018 informal IAS seminar) tentatively announced a subconvex bound on GGP pairs in the p-adic depth aspect, for principal series representations with parameters satisfying the analogues of (1.4) and (1.5), introducing important and fundamental ideas (see §1.5.3).

Kumar–Mallesham–Singh [KMS], by a different method than that of Blomer–Buttcane, recently established subconvex bounds on PGL_3 (with any fixed GL_2 twist) assuming (1.4) and a modified form of (1.5). P. Sharma [Sh] recently generalized Blomer–Buttcane's results to the case of fixed cuspidal GL_2 twists.

1.5. Proof sketch. We record here a very high-level overview of the proof, intended for experts. We give in §2 a more leisurely introduction to the main ideas of the paper.

1.5.1. *Setup and notation.* For simplicity of presentation, we pretend that

$$F = \mathbb{Q}, \quad S = \{\infty\}, \quad \mathfrak{q} = \infty$$

so that $F_{\mathfrak{q}} = \mathbb{R}$. We suppose further that

$$G(F_{\mathfrak{q}}) = G(\mathbb{R}) = \mathrm{GL}_{n+1}(\mathbb{R}), \quad H(F_{\mathfrak{q}}) = H(\mathbb{R}) = \mathrm{GL}_n(\mathbb{R}).$$

We replace the adelic quotients $G(F)\backslash G(\mathbb{A})$ and $H(F)\backslash H(\mathbb{A})$ with real quotients, and simplify our notation a bit:

$$G := \mathrm{GL}_{n+1}(\mathbb{R}), \quad H := \mathrm{GL}_n(\mathbb{R}), \quad [G] := \Gamma\backslash G, \quad [H] := \Gamma_H\backslash H.$$

Here Γ and Γ_H are lattices satisfying $\Gamma_H = \Gamma \cap H$, with Γ_H cocompact. Finally, for $(\pi, \sigma) \in \mathcal{F}_T$, we will confuse π and σ with their local components at ∞.

We write $\mathfrak{g}, \mathfrak{h}$ for the Lie algebras and $\mathfrak{g}^\wedge, \mathfrak{h}^\wedge$ for their imaginary duals. We write $Z \leqslant G$ and $Z_H \leqslant H$ for the centers and $\mathfrak{z}, \mathfrak{z}_H$ for their Lie algebras.

1.5.2. *Analytic test vectors.* Let $(\pi, \sigma) \in \mathcal{F}_T$ with T large. We construct test vectors in the manner described in [NV, §1.10]:

The distinction assumption implies that we may choose an element $\tau \in \mathfrak{g}^\wedge$, with restriction $\tau_H \in \mathfrak{h}^\wedge$, so that τ (resp. τ_H) lies in the coadjoint orbit $\mathcal{O}_\pi$ of π (resp. $\mathcal{O}_\sigma$ of σ) and $|\tau| \asymp T$.

We define a test function $f \in C_c^\infty(G)$, as follows. We take f supported near the identity. We may describe it by its pullback to the Lie algebra $\mathfrak{g}$, hence by the Fourier transform $a : \mathfrak{g}^\wedge \to \mathbb{C}$ of that pullback. We specify that a is a smooth bump concentrated on

$$\left\{ \tau + \xi : \xi' \ll T^{1/2+\varepsilon}, \xi'' \ll T^{\varepsilon} \right\},$$

where $\xi = (\xi', \xi'')$ denotes a coordinate system for which the ξ'-axis consists of all directions tangent to $\mathcal{O}_\pi$ at τ while the ξ''-axis consists of the orthogonal directions (see Figure 1). Then f is essentially supported on

$$\left\{ g \in G : g = 1 + \mathrm{O}(T^{-\varepsilon}),\ \mathrm{Ad}^*(g)\tau = \tau + \mathrm{O}(T^{1/2-\varepsilon}) \right\}. \tag{1.6}$$

The corresponding operator $\pi(f)$ is approximately a rank one projector with range spanned by a unit vector microlocalized at τ (in the sense of [NV], see §2.5).

We construct a unit vector $\Psi \in \sigma$, microlocalized at τ_H.

The definition of $\mathcal{L}(\pi, \sigma)$, combined with the local matrix coefficient integral asymptotics of [NV, §19], give the integral representation

$$\mathcal{L}(\pi, \sigma) \approx T^{n^2/2} \sum_{\varphi \in \mathcal{B}(\pi)} \left| \int_{[H]} \pi(f)\varphi \cdot \bar{\Psi} \right|^2. \tag{1.7}$$

A key ingredient in these constructions and proofs is a refinement of the microlocal calculus developed in [NV] (see §2.4).

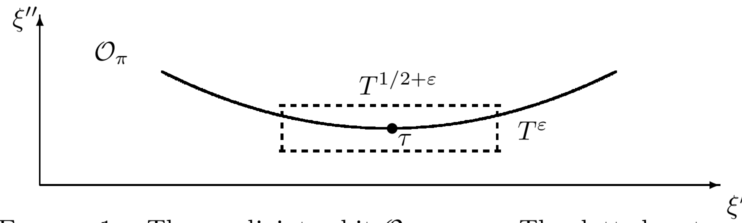

FIGURE 1. The coadjoint orbit $\mathcal{O}_\pi$ near τ. The dotted rectangle indicates the support of a.

1.5.3. *Relative trace formula.* The main point is to show that

$$T^{-n/2} \int_{x,y\in[H]} \sum_{\gamma\in\Gamma} \bar{\Psi}(x)\Psi(y) f(x^{-1}\gamma y)\, dx\, dy = (\text{main term}) + \mathrm{O}(T^{-\delta}). \tag{1.8}$$

Indeed, by combining (1.7) with the pretrace formula for $[G]$ applied to f, we may understand the LHS of (1.8) as an average of $\mathcal{L}(\pi,\sigma)$ over π in a family of size $T^{n(n+1)/2}$. In view of our "no conductor dropping" assumption, the family size is approximately $C(\pi,\sigma)^{1/4}$. Strictly speaking, we should replace f here by a convolution $f * f^*$, but that convolution has the same shape as f. Given a sufficiently robust proof of (1.8), the amplification method will thus deliver a subconvex bound. We refer to §2.7 for further discussion concerning the reduction to (1.8).

An apparent main term in (1.8) of size $\mathrm{O}(1)$ comes from those γ lying in HZ, so it suffices to show that the remaining terms contribute $\mathrm{O}(T^{-\delta})$. Since f is supported in a fixed compact set, the number of γ is $\mathrm{O}(1)$ (although after amplification, that number will instead be a small positive power of T). Fixing a fundamental domain $\mathcal{H}$ for $[H]$, our task is thus to show for each essentially fixed $\gamma \in \Gamma - HZ$ that

$$T^{-n/2} \int_{x,y\in\mathcal{H}} \bar{\Psi}(x)\Psi(y) f(x^{-1}\gamma y)\, dx\, dy \ll T^{-\delta}. \tag{1.9}$$

Such integrals were first considered by S. Marshall in a related p-adic depth aspect. He proposed that (1.9) should follow from consideration of the L^2-normalization of Ψ, the (approximate) equivariance of Ψ with respect to the centralizer H_{τ_H} of τ_H, and the size and support properties of f. More precisely, he proposed the nontrivial volume bound

$$\mathrm{vol}\left\{ z \in H_{\tau_H} \,\middle|\, \begin{array}{c} z = \mathrm{O}(1), \\ f(x^{-1}\gamma y z) \neq 0 \text{ for some } x \in \mathcal{H} \end{array} \right\} \ll T^{-\delta}, \tag{1.10}$$

which turns out to yield (1.9) (with a different value of δ) after using the approximate equivariance to replace $\Psi(y)$ by an average of $\Psi(yz)$ over small elements $z \in H_{\tau_H}$ and appealing to Cauchy–Schwarz. We emphasize here that $\gamma \in \Gamma - HZ$ is essentially fixed.

1.5.4. *Volume bounds.* In the setting of Theorem 1.1, we establish (1.10) indirectly, only as a consequence of the stronger estimate

$$\operatorname{vol}\left\{ z \in Z_H \,\middle|\, \begin{array}{c} z = \mathrm{O}(1), \\ f(x^{-1}\gamma y z) \neq 0 \text{ for some } x \in \mathcal{H} \end{array} \right\} \ll T^{-\delta} \tag{1.11}$$

obtained by specializing from the n-dimensional centralizer H_{τ_H} of τ_H to the 1-dimensional center Z_H of H.

1.5.5. *Coadjoint reformulation.* The crucial case in proving (1.11) (to which others ultimately reduce) is when $y = 1$ and $\mathrm{Ad}^*(\gamma)\tau = \tau$. In view of the support condition (1.6) for f, we reduce to verifying that

$$\operatorname{vol}\left\{ z \in Z_H \,\middle|\, \begin{array}{c} z = \mathrm{O}(1), \\ \mathrm{Ad}^*(\gamma z)\tau \in \mathrm{Ad}^*(H)\tau + \mathrm{O}(T^{1/2}) \end{array} \right\} \ll T^{-\delta}. \tag{1.12}$$

1.5.6. *Reduction to the Lie algebra.* The estimate (1.12) would fail most spectacularly if $\mathrm{Ad}^*(\gamma Z_H)\tau \subseteq \mathrm{Ad}^*(H)\tau$. Using that $\mathrm{Ad}^*(\gamma z)\tau = \mathrm{Ad}^*(\gamma z \gamma^{-1})\tau$ and taking Lie algebras, we would then have $\mathrm{ad}^*(\mathrm{Ad}(\gamma)\mathfrak{z}_H)\tau \subseteq \mathrm{ad}^*(\mathfrak{h})\tau$. If this last containment were to hold for all γ in some one-parameter subgroup, we would obtain

$$\mathrm{ad}^*(\mathrm{ad}(x)\mathfrak{z}_H)\tau \subseteq \mathrm{ad}^*(\mathfrak{h})\tau \text{ for some } x \in \mathfrak{g}_\tau - \mathfrak{z}, \tag{1.13}$$

where $\mathfrak{g}_\tau := \{x \in \mathfrak{g} : \mathrm{ad}^*(x)\tau = 0\}$.

In §15, we use the implicit function theorem and related ideas to reverse the above reasoning: we show that to establish the required volume bound (1.11), it suffices to exclude the possibility of the apparent "worst-case scenario" (1.13) at the level of the Lie algebra.

1.5.7. *Endgame.* The exclusion of (1.13) is the content of the following linear algebra result, proved in §16. Here the eigenvalue condition on τ corresponds precisely to the "no conductor dropping" assumption on (π, σ).

Theorem 1.8 (Theorem 15.3, formulated explicitly)**.** *Let M_n denote the space of complex $n \times n$ matrices, included in the space M_{n+1} of $(n+1) \times (n+1)$ matrices as the upper-left block, e.g., for $n = 2$, as*

$$\begin{pmatrix} * & * & 0 \\ * & * & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Let τ be an element of M_{n+1} with the property that no eigenvalue of τ is also an eigenvalue of the upper-left $n \times n$ submatrix τ_H of τ. Let $x \in M_{n+1}$ with $[x, \tau] = 0$, where $[a, b] := ab - ba$. Let z denote the image in M_{n+1} of the identity element of M_n, thus $z = \mathrm{diag}(1, \ldots, 1, 0)$ with n ones. Suppose that

$$[x, [z, \tau]] = [y, \tau]$$

for some $y \in M_n$. Then x is a scalar matrix.

1.6. Organization and reading suggestions. The paper is intended to be read linearly, but some general comments on its organization may be useful.

§3 introduces general notation and conventions, in effect throughout the paper.

The paper is organized around a main local result, Theorem 4.2. We encourage the interested reader to focus first on understanding the statement of that result. In the language of §2, it describes the essential properties of the test function f and the vector Ψ.

The remainder of the paper divides roughly into four parts, each of which may be read independently:

- In §5–§6, we assume our main local result and deduce our main global result, Theorem 1.1. These sections contain the expected global ingredients: amplification method, sums over rational points, and so on. The remainder of the paper is devoted to the proof of the main local result.
- §7–§14 develop the harmonic analysis needed to reduce the proof of the main local result to that of the volume estimates described in §1.5.4. The techniques developed here may be more broadly useful.
- §15 applies the implicit function theorem and related ideas to reduce the proof of the volume estimates to a problem in linear algebra.
- §16 solves the linear algebra problem.

We have attempted to make each section self-contained. We have included a short notational index at the end of the paper.

Acknowledgments. We would like to thank Valentin Blomer, Simon Marshall, Philippe Michel, Akshay Venkatesh and Wei Zhang for their helpful feedback on an earlier draft.

2. Overview

Expanding upon the proof sketch of §1.5, we now explain the main ideas of the paper in more detail, but still informally. The reader interested only in actual proofs may proceed directly to §3.

We retain throughout this section the setup and notation of §1.5.1.

2.1. Amplified moments. Our basic strategy may be phrased in a familiar way in terms of the amplification method applied to moments of families of L-functions. The implementation of that strategy relies heavily upon the method developed in [NV] for analyzing families of automorphic forms using coadjoint orbits and microlocalized vectors. We outline below how a refinement of that method applies in our setting and reduces our task to elementary problems in calculus and linear algebra.

We note the arguments in the body of this paper do not explicitly refer to families or microlocalized vectors, but instead work with them implicitly through their approximate projectors. We hope that by phrasing this overview in such terms, it may serve as a useful guide to those arguments.

Let $(\pi_0, \sigma) \in \mathcal{F}_T$ with T large. The branching coefficients $\mathcal{L}(\pi, \sigma)$ are nonnegative (see §3.7), so we may bound $\mathcal{L}(\pi_0, \sigma)$ by the first moment $\sum_{\pi \in \Pi} \mathcal{L}(\pi, \sigma)$ taken over any family Π containing π_0. The Lindelöf hypothesis predicts that each $\mathcal{L}(\pi, \sigma)$ has size $\mathrm{O}(T^{\varepsilon})$. We will chose Π to have cardinality

$$|\Pi| \approx T^{n(n+1)/2}. \tag{2.1}$$

In view of our "no conductor dropping" assumption, Lindelöf on average for the moment then recovers the convexity bound for the branching coefficient of interest. Experience suggests that if we can prove an asymptotic formula for the moment with enough room to spare, then the amplification method will deliver a subconvex bound. Our main task is thus to prove (in a sufficiently robust way) that

$$|\Pi|^{-1} \sum_{\pi \in \Pi} \mathcal{L}(\pi, \sigma) = (\text{constant}) + \mathrm{O}(T^{-\delta}). \tag{2.2}$$

Remark 2.1. If we were in the excluded case that $(G, H) = (\mathrm{GL}_3(\mathbb{R}), \mathrm{GL}_2(\mathbb{R}))$, $(\Gamma, \Gamma_H) = (\mathrm{GL}_3(\mathbb{Z}), \mathrm{GL}_2(\mathbb{Z}))$ and σ were the $1 \boxplus 1$ Eisenstein series, then $\mathcal{L}(\pi, \sigma)$ would be essentially the fourth power of the standard L-function $L(\pi, 1/2)$, and so the moment problem (2.2) would be of the same shape as that in [BB1].

2.2. Families and coadjoint orbits. Which family Π should we take? We explain our choice here, and then describe how the orbit method as advocated in [NV] provides a natural tool for its study.

The general shape of the family is perhaps unsurprising: we take those representations whose local components at ∞ belong to some carefully-chosen subset of the tempered dual of G, which we define in turn using the parameters $\lambda_{\pi,i}$ appearing in the formula (1.1) for the local L-factor at ∞. Roughly speaking, we take our family to consist of π for which each coefficient of the polynomial

$$\prod_i \left(X - \frac{\lambda_{\pi,i}}{T} \right) \in \mathbb{C}[X]$$

differs from the corresponding coefficient for π_0 by at most $\mathrm{O}(1/T)$. Strictly speaking, we relax this condition by a factor of T^{ε}, but to simplify exposition do not display such factors here.

It is natural to consider the coefficients of such polynomials. One reason is that they are asymptotic to the eigenvalues of π under a standard system of generators for the center of the universal enveloping algebra of

$$\mathfrak{g} = \mathrm{Lie}(G) = \mathfrak{gl}_{n+1}(\mathbb{R}).$$

Using that Plancherel measure for G is asymptotic to Lebesgue measure with respect to such coefficients (see for instance [NV, §17.5]) and assuming the Weyl law, one can check that Π has the cardinality promised in (2.1).

Another reason is that such coefficients are intimately related to harmonic analysis on the representation π. The relationship is encoded by the Kirillov formula (§7.5) for the distributional character $\chi_\pi : G \to \mathbb{C}$, which we now describe in the present example (simplifying a bit using special features of that example). Recall that $\mathfrak{g}^\wedge = \mathrm{Hom}(\mathfrak{g}, i\mathbb{R})$ denotes the imaginary dual of $\mathfrak{g}$. We identify $\mathfrak{g}^\wedge$ with the space $i\,\mathfrak{gl}_{n+1}(\mathbb{R})$ of imaginary matrices using the trace pairing. A standard theorem from linear algebra is that the set of matrices

$$\mathcal{O}_\pi := \left\{ \xi \in \mathfrak{g}^\wedge \text{ with minimal polynomial } \prod_i (X - \lambda_{\pi,i}) \right\}$$

forms a conjugacy class (one can check that the unitarity of π implies that $\mathcal{O}_\pi$ is nonempty). In other words, $\mathcal{O}_\pi$ is a *coadjoint orbit*, i.e., an orbit for the group G under its coadjoint action on $\mathfrak{g}^\wedge$. The Kirillov formula asserts roughly that for small enough $x \in \mathfrak{g}$,

$$\chi_\pi(\exp(x)) \approx \int_{\xi \in \mathcal{O}_\pi} e^{\langle x, \xi \rangle} \, d\omega(\xi), \tag{2.3}$$

where ω denotes the normalized symplectic measure on a coadjoint orbit. The measures ω are defined most directly using the Lie bracket on $\mathfrak{g}$, but may be characterized in the relevant examples as follows (see for instance [NV, (17.2)] for details): one may compute an integral over $\mathfrak{g}^\wedge$ with respect to Lebesgue measure by first integrating over the space of polynomials (with respect to Lebesgue measure in each coefficient) and then integrating over each corresponding (regular) coadjoint orbit with respect to its symplectic measure ω. The Kirillov formula and the geometry of coadjoint orbits form the cornerstone of our approach to harmonic analysis on π, so it is natural that we parametrize our families in terms of associated quantities.

Remark 2.2. If the parameters of π_0 are dyadically spaced in the sense that

$$\lambda_{\pi_0,i} - \lambda_{\pi_0,j} \gg T \quad \text{for } i \neq j, \tag{2.4}$$

then our family Π consists of π for which

$$\lambda_{\pi,i} = \lambda_{\pi_0,i} + \mathrm{O}(1), \tag{2.5}$$

as in [BB1]. On the other hand, if we suppose instead that, for instance, (2.4) holds for $(i,j) \neq (1,2)$ but $\lambda_{\pi_0,1} = \lambda_{\pi_0,2}$, then the condition (2.5) is unchanged for $i \geqslant 3$ but must be modified for $i = 1, 2$ to the pair of conditions

$$\lambda_{\pi,1} + \lambda_{\pi,2} = \lambda_{\pi_0,1} + \lambda_{\pi_0,2} + \mathrm{O}(1),$$

$$\lambda_{\pi,1} - \lambda_{\pi,2} = \lambda_{\pi_0,1} - \lambda_{\pi_0,2} + \mathrm{O}(T^{1/2}).$$

As this example already suggests, the description of our family in terms of individual parameters $\lambda_{\pi,i}$ becomes more complicated the more those parameters "collide." We do not make use of any such description in this paper.

2.3. Weyl's law. We pause to indicate one way to prove the Weyl law for the family Π (smoothened and T^ε-enlarged), i.e., for the moment as in (2.2), but with each branching coefficient $\mathcal{L}(\pi, \sigma)$ replaced by the simpler quantity 1. This problem is much simpler than the moment problem of interest, but discussing it allows us to describe some techniques that apply also to the latter problem.

By the trace formula for the compact quotient $[G]$, our main task is to produce a test function f on G, supported arbitrarily close to the identity element, for which $\mathrm{trace}(\pi(f))$ approximates the indicator function of Π. Such a function f may be described in terms of its pullback to the Lie algebra $\mathfrak{g}$ and then, by Fourier transform, in terms of a Schwartz function a on $\mathfrak{g}^\wedge$. By the Kirillov formula (2.3) and the Parseval relation, our task reduces to arranging that

- the inverse Fourier transform $a^\vee$ is supported close to the origin in $\mathfrak{g}$ (so that f is supported close to the identity in G), and
- the integral $\int_{\xi \in \mathcal{O}_\pi} a(\xi) \, d\omega(\xi)$ approximates the indicator function of Π.

The support condition is valid provided that a is essentially constant on balls of size $\mathrm{O}(1)$, so the main point is to achieve the required approximation property.

To that end, we choose some element π of our family Π (e.g., $\pi = \pi_0$, although the choice won't matter) and a regular element $\tau \in \mathfrak{g}^\wedge$ of Euclidean norm $|\tau| \asymp T$ which belongs to the corresponding coadjoint orbit $\mathcal{O}_\pi$, so that $\mathcal{O}_\pi = \mathrm{Ad}^*(G)\tau$.

For concreteness, we might construct τ using rational canonical form, e.g., for $G = \mathrm{GL}_3(\mathbb{R})$ by

$$\tau = \sqrt{-1}\begin{pmatrix} 0 & 0 & c_3 \\ T & 0 & c_2 \\ 0 & T & c_1 \end{pmatrix} \quad \text{if} \prod_{j=1}^{3}\left(X - \frac{\lambda_{\pi,j}}{\sqrt{-1}T}\right) = X^3 - c_1X^2 - c_2X - c_3. \tag{2.6}$$

Near such (regular) elements τ, the map $\xi \mapsto \det(X - \xi)$ equips the space $\mathfrak{g}^\wedge$ with the structure of a fibered manifold, with fibers the coadjoint orbits (lemma 13.2). It is convenient to introduce coordinates $\xi = (\xi', \xi'')$ on $\mathfrak{g}^\wedge$ as in §1.5.2 (see §9.4.1 for details), so that the ξ'-axis consists of all directions tangent to $\mathcal{O}_\pi$ at τ while the ξ''-axis consists of those transverse to the fibers. We then take

$$a = \text{ smooth bump on } \left\{\tau + \xi : |\xi'| \ll T^{1/2+\varepsilon}, |\xi''| \ll T^{\varepsilon}\right\}. \tag{2.7}$$

As depicted in Figure 1, the support of a is thus a thin "coin-shaped" neighborhood of τ, concentrated near the coadjoint orbit $\mathcal{O}_\pi$. One checks that the coadjoint orbits intersecting this neighborhood are those arising from our family Π, while the symplectic measures of the intersections are approximately one (see, e.g., the proof of lemma 14.5). The required approximation property follows.

The number of ξ' (resp. ξ'') directions is $n(n+1)$ (resp. $n+1$), so the total volume of the support of a is approximately the promised cardinality $T^{n(n+1)/2}$ of the family Π (as we could have predicted using the characterization of symplectic measures noted in §2.2). In particular,

$$\|f\|_{L^\infty(G)} \approx f(1) \approx T^{n(n+1)/2} \tag{2.8}$$

We note that the strategy suggested above is flexible with respect to the choice of τ: the same argument works if we replace τ by any element of $\mathcal{O}_\pi$ with similar properties, e.g., by $\mathrm{Ad}^*(g)\tau$ for any group element $g \in G$ close to the identity.

2.4. Microlocal calculus. A rigorous implementation of the strategy suggested in §2.3 might require working with positive-definite test functions, e.g., those obtained by convolving a function f as above against its adjoint. In §9.2, we study in detail the asymptotics of the convolutions of two functions f_1 and f_2 as above in terms of the corresponding Fourier transforms a_1 and a_2 on $\mathfrak{g}^\wedge$ (i.e., both a_1 and a_2 are smooth bumps as in (2.7) and Figure 1). We provide a general calculus for doing so, valid for any connected reductive group G. We hope this calculus will be more broadly useful. It is rooted in a basic estimate to the effect that the convolution $f_1 * f_2$ corresponds to a certain star product $a_1 \star a_2$ admitting an asymptotic expansion

$$a_1 \star a_2 \sim a_1 a_2 + a_1 \star^1 a_2 + a_1 \star^2 a_2 + \cdots, \tag{2.9}$$

with $\star^1$ a multiple of the Poisson bracket and $\star^j$ defined in general by a certain bidifferential operator. The derivative bounds enjoyed by bumps as in (2.7) turn out to force each successive term on the RHS of (2.9) to be smaller than the previous terms, so that the expansion is not merely formal. For example, using that the Poisson bracket is defined by vector fields tangent to the coadjoint orbits, we may check that $a_1 \star^1 a_2$ is of size $T^{-2\varepsilon}$. Our results concerning this calculus may be understood as further steps in a direction suggested by Rieffel [Ri].

A first version of that calculus, inspired by the pseudodifferential calculus, was given in [NV], but under hypotheses that would be prohibitively restrictive for our aims. Indeed, that calculus applies to smooth bumps a on regions roughly of the form $\{\tau + \xi : |\xi| \ll T^{1/2+\varepsilon}\}$, but not on somewhat rougher regions as in (2.7). Such bumps are inadequate for the analysis of "short" families Π required here. For instance, in the "dyadically spaced" case, they would suffice for studying families defined by conditions like (2.5), but with $\mathrm{O}(1)$ replaced by $\mathrm{O}(T^{1/2})$.

We indicate why bumps as in (2.7) represent a natural limit for our calculus:

(i) The coadjoint orbit $\mathcal{O}_\pi$ plays the role of the phase space for the representation π. Balls of the form $\mathcal{O}_\pi \cap (\tau + \mathrm{O}(T^{1/2}))$ have symplectic volume $\asymp 1$, corresponding to the Planck scale. By the uncertainty principle, we cannot hope for a meaningful calculus involving smooth bumps on much smaller balls. The resolution of our calculus is thus optimal (up to epsilons) in the ξ'-directions.

(ii) To understand the role of the ξ''-directions, suppose for instance that π belongs to the discrete series and τ is regular elliptic. The representations π' whose coadjoint orbits $\mathcal{O}_{\pi'}$ intersect the support of (2.7) are then discrete series representations whose parameters differ from those of π by $\mathrm{O}(T^\varepsilon)$. The number of such representations is $\mathrm{O}(T^\varepsilon)$ (for a different value of ε). The indicated ξ''-scale is thus the relevant one for projecting onto rather short families of representations.

2.5. Microlocalized vectors. The function a as in (2.7), being approximately the characteristic function of a set, satisfies $a^2 \approx a$. From our calculus, it follows that the corresponding test function f on G satisfies $f * f \approx f$, and so the corresponding operator $\pi(f)$ is approximately idempotent. The real-valuedness of a forces $\pi(f)$ to be self-adjoint, while the noted symplectic measure calculation shows that $\operatorname{trace}(\pi(f)) \approx 1$. Speaking informally, $\pi(f)$ is thus approximately a rank one orthogonal projection, and so corresponds approximately to some scaling class of unit vectors $v \in \pi$ for which $\pi(f)v \approx v$.

The relationship between v and τ may be quantified more directly in terms of the action of Lie algebra elements $x \in \mathfrak{g}$ with $x = \mathrm{O}(T^{-1/2-\varepsilon})$, under which v enjoys the approximate equivariance property

$$\exp(x)v \approx e^{\langle x,\tau\rangle}v.$$

This estimate persists for small x lying within $\mathrm{O}(T^{-1/2-\varepsilon})$ of the centralizer of τ. The test function f should in turn be regarded as a proxy for the matrix coefficient $\langle gv, v\rangle$ (conjugated, normalized and truncated).

We say that such a vector v is *microlocalized* at the parameter $\tau \in \mathfrak{g}^\wedge$. A key property of such vectors is that their matrix coefficients $\langle gv, v\rangle$, or equivalently, the corresponding test functions $f(g)$, concentrate near the centralizer of the parameter τ. Quantitatively, f is essentially supported on the set (1.6) consisting of small group elements that approximately centralize τ.

One may understand our calculus and the associated notion of microlocalized vectors as giving an analytic archimedean analogue of the theory of types for p-adic groups. The latter gives, among other things, many examples of open subgroups J of $\mathrm{GL}_n(\mathbb{Z}_p)$ and one-dimensional representations χ of J with the property that, up to twisting, there is at most one supercuspidal representation of $\mathrm{GL}_n(\mathbb{Q}_p)$ whose restriction to J contains χ; moreover, χ occurs in that restriction with multiplicity one. In other words, the function h on $\mathrm{GL}_n(\mathbb{Z}_p)$ given by χ^{-1} times the normalized characteristic function of J has the property that, for each irreducible representation ρ of $\mathrm{GL}_n(\mathbb{Q}_p)$, the operator $\rho(h)$ vanishes unless ρ is a twist of the given supercuspidal, in which case $\rho(h)$ is a rank one projector with range spanned by some (J, χ)-isotypic vector $u \in \rho$. By comparison, we have noted in §2.4 that our calculus produces functions f supported near the identity element of the real group G for which $\pi(f)$ is negligible unless π belongs to $\mathrm{O}(T^{\varepsilon})$-many discrete series representations, in which case it is approximately a projector of rank $\mathrm{O}(T^{\varepsilon})$. Thus f and h are analogous. The (informally and ambiguously defined) microlocalized vector $v \in \pi$ is analogous to the isotypic vector $u \in \rho$.

2.6. Period formulas and matrix coefficient integral asymptotics. Recall that our aim is to understand the moment (2.2) of branching coefficients $\mathcal{L}(\pi, \sigma)$ taken over π in some family Π. We have thus far discussed only the family, emphasizing the relevant local harmonic analysis on G. We turn now to the branching coefficients. They arise from a formula of the following shape: for automorphic forms $\varphi \in \pi$ and $\Psi \in \sigma$,

$$\left| \int_{[H]} \varphi \bar{\Psi} \right|^2 \approx \mathcal{L}(\pi, \sigma) \mathcal{Q}(\varphi \otimes \Psi), \quad \mathcal{Q}(\varphi \otimes \Psi) := \int_{h \in H} \langle h\varphi, \varphi \rangle \langle \Psi, h\Psi \rangle \, dh, \tag{2.10}$$

where $\approx$ means "up to unimportant inaccuracies" (leading constants, local factors at "auxiliary" places, etc). This formula amounts to the definition of $\mathcal{L}(\pi, \sigma)$ (see §3.7), and is all that enters into the proof of Theorem 1.1; the fact that $\mathcal{L}(\pi, \sigma)$ is conjectured (and in many cases known) to be given by a special value of an L-function is relevant only for interpreting our results, as in Corollary 1.2.

Bernstein–Reznikov [BR] initiated the systematic use of period formulas like (2.10) as a tool for estimating branching coefficients, emphasising spectral aspects. Venkatesh [Ve] introduced related ideas in the level aspect, as well as the use of arithmetic amplification in the spirit of Duke–Friedlander–Iwaniec (see [DFI] and references). We mention also the influential works of Sarnak [Sa] and Iwaniec–Sarnak [IS1]. A capstone application of this tool was the resolution of the subconvexity problem for GL_2 [MV2]. The art in each such application lies in choosing vectors φ and Ψ for which one may prove that

- $\mathcal{Q}(\varphi \otimes \Psi)$ is not too small, and
- $\int_{[H]} \varphi \bar{\Psi}$ is not too large.

The works mentioned above considered low rank examples. The paper [NV] studied the higher rank setting and described

- an approach based on the orbit method for choosing vectors φ to Ψ to which the above strategy may be profitably applied (see [NV, §1.10]), and
- a technique for asymptotically evaluating the local integrals $\mathcal{Q}(\varphi \otimes \Psi)$ (see [NV, §1.9, §19]).

We briefly recall the main points. One takes for φ and Ψ a pair of microlocalized unit vectors with corresponding parameters $\tau \in \mathcal{O}_\pi \subseteq \mathfrak{g}^\wedge$ and $\nu \in \mathcal{O}_\sigma \subseteq \mathfrak{h}^\wedge$, say. If the restriction τ_H of τ to $\mathfrak{h}^\wedge$ is not close to ν, then the approximate equivariance forces $\mathcal{Q}(\varphi \otimes \Psi)$ to be small. The crucial case to understand is thus when $\nu = \tau_H$. The "no conductor dropping" assumption now becomes relevant: it manifests at the level of the coadjoint orbits $\mathcal{O}_\pi \subseteq \mathfrak{g}^\wedge$ and $\mathcal{O}_\sigma \subseteq \mathfrak{h}^\wedge$ in that the "relative coadjoint orbit"

$$\mathcal{O}_{\pi,\sigma} := \{\xi \in \mathcal{O}_\pi : \xi_H \in \mathcal{O}_\sigma\} \tag{2.11}$$

is an H-torsor, i.e., transitive and free under the coadjoint action of H. This says both that any two elements τ are equivalent under H (so that little is lost in choosing a specific τ) and that the centralizer in H of τ is trivial. The matrix coefficient concentration property noted in §2.5 then forces the integrand in the definition of $\mathcal{Q}(\varphi \otimes \Psi)$ to be concentrated roughly in the neighborhood $1 + \mathrm{O}(T^{-1/2})$ of the identity element of H, where it may be estimated using the Kirillov formula. In particular, the integral has size $\approx (T^{-1/2})^{\dim(H)} = T^{-n^2/2}$. We have oversimplified in this discussion; in practice, it becomes necessary to average over short (i.e., size $\mathrm{O}(T^{\varepsilon})$) famillies of microlocalized vectors.

In this paper, we choose vectors as in [NV] and appeal to the matrix coefficient integral asymptotics given in [NV, §19] as a black box (see §12). The paper [NV] contains other ideas (e.g., related to ergodic theory) which do not enter here.

2.7. Relative trace formula. From the above discussion, we may construct

- a coadjoint element $\tau \in \mathfrak{g}^\wedge$ (the same choice of which works for all $\pi \in \Pi$, since the coadjoint orbits $\mathcal{O}_\pi$ are close to one another),
- a unit vector $\Psi \in \sigma$, microlocalized at $\tau_H \in \mathfrak{h}^\wedge$, and
- for each π in our family Π, a unit vector $\varphi_\pi \in \pi$, microlocalized at τ,

so that the integral representation

$$\mathcal{L}(\pi, \sigma) \approx T^{n^2/2} \left| \int_{[H]} \varphi_\pi \bar{\Psi} \right|^2 \tag{2.12}$$

holds. Since φ_π is microlocalized at the parameter τ, it satisfies $\pi(f)\varphi \approx \varphi$ for the test function f described in §2.3.

We have noted that the operator $\pi(f)$ is approximately a rank one idempotent, so that for an orthonormal basis $\mathcal{B}(\pi)$, we have $\sum_{\varphi \in \mathcal{B}(\pi)} \pi(f)\varphi(x)\overline{\varphi(y)} \approx \varphi_\pi(x)\overline{\varphi_\pi(y)}$. The pretrace formula for $[G]$ applied to f thus gives

$$\sum_{\pi \in \Pi} \varphi_\pi(x)\overline{\varphi_\pi(y)} \approx \sum_{\gamma \in \Gamma} f(x^{-1}\gamma y).$$

Integrating both sides against Ψ, appealing to the integral representation (2.12) and recalling the size (2.1) of Π, we obtain the basic formula

$$|\Pi|^{-1} \sum_{\pi \in \Pi} \mathcal{L}(\pi, \sigma) \approx T^{-n/2} \int_{x,y \in [H]} \sum_{\gamma \in \Gamma} \bar{\Psi}(x)\Psi(y) f(x^{-1}\gamma y) \, dx \, dy. \tag{2.13}$$

A comparison of (2.2) and (2.13) now explains the sufficiency of (1.8).

We emphasize that in the body of the paper, we work most directly with the test function f and the operators $\pi(f)$ rather than with the corresponding microlocalized vectors φ_π (see for instance §4.3 and §14). We hope that by having phrased this overview informally in terms of φ_π, it may usefully illustrate the rigorous arguments given in the body in terms of f.

Remark 2.3. The present setup differs from that in [NV], where an average of $\mathcal{L}(\pi, \sigma)$ for fixed π over a large family of σ is studied by spectrally decomposing the L^2-norm $\int_{[H]} |\varphi_\pi|^2$, averaging over a large family of microlocalized vectors φ_π, and appealing to Ratner theory.

Remark 2.4. When $n = 1$, the present setup is closely related to that of Iwaniec–Sarnak [IS1].

To explain, it is convenient to switch to the pair $(G, H) = (\mathrm{U}(1,1), \mathrm{U}(1))$, to which the above discussion applies without essential modification. Modding out by the center of G and ignoring issues of isogeny, we might as well pretend then that $(G, H) = (\mathrm{SL}_2(\mathbb{R}), \mathrm{SO}(2))$.

Suppose that σ is trivial, so that Ψ is constant, and that π belongs to the principal series, so that it contains an H-invariant unit vector $\varphi_{\pi,0}$. Let us identify G/H with the upper half-plane $\mathbb{H}$, so that $\varphi_{\pi,0}$ corresponds to some L^2-normalized Maass form on $\Gamma\backslash\mathbb{H}$.

The integral $\int_{[H]} \varphi_\pi \bar{\Psi}$ is then a multiple of $\varphi_{\pi,0}(z)$, with $z \in \mathbb{H}$ the point stabilized by H. For suitable π_0, our family Π consists of π with $r_\pi = T + \mathrm{O}(1)$, with $1/4 + r_\pi^2$ the eigenvalue of $\varphi_{\pi,0}$.

The LHS of (2.13) is thus a multiple of $\sum_{r_\pi = T + \mathrm{O}(1)} |\varphi_{\pi,0}(z)|^2$. These sums and their amplified variants were considered by Iwaniec–Sarnak [IS1] in their pioneering work on the sup norm problem. Since Ψ is constant, the RHS of (2.13) is unchanged by replacing f with its bi-H-invariant average, in which case it is essentially the same kernel function as considered by Iwaniec–Sarnak.

Thus the general method described here specializes to that of Iwaniec–Sarnak (although, as noted in Remark 1.4, we have optimized our arguments to a much lesser extent).

2.8. Reduction to volume bounds: remarks. We record here some miscellaneous remarks concerning the reduction to (1.9) and (1.11) described in §1.5.

Remark 2.5. The exploitation of invariance in the central direction is reminiscent of the conductor lowering trick applied by Munshi and others to the subconvexity problem on GL_3 (see [Mu1, Mu3, Sh, LMS]). It might be interesting to understand any relation more precisely.

Remark 2.6. There is another way to view the reduction to (1.10) or (1.11). Recall the integral representation (2.12). In view of the H_{τ_H}-equivariance of Ψ, it is natural to replace the microlocalized vector φ_π with a modified vector $\varphi_{\pi,0}$, given by a weighted average of φ_π under small elements of H_{τ_H}, for which $\int_{[H]} \varphi_{\pi,0}\bar{\Psi} \approx \int_{[H]} \varphi_\pi \bar{\Psi}$. We regard $\varphi_{\pi,0}$ as the more natural vector for the problem. For example, in the $n = 1$ case discussed in Remark 2.4, the modified vector $\varphi_{\pi,0}$ is a multiple of the normalized Maass form, which is clearly the natural vector for that case. The microlocalized vectors φ_π may be regarded as footholds into our understanding of the harmonic analytic difficulties and matrix coefficient integral asymptotics in higher rank.

We apply the pretrace formula as before, but with φ_π replaced by $\varphi_{\pi,0}$. We reduce to estimating for $\gamma \in \Gamma - HZ$ the integrals

$$\int_{x,y \in \mathcal{H}} \bar{\Psi}(x)\Psi(y) f_0(x^{-1}\gamma y) \, dx \, dy,$$

where f_0 is given by a weighted average of f on the left and right under small elements of H_{τ_H}. We may understand f_0 as a proxy for the matrix coefficient of $\varphi_{\pi,0}$. Each factor in the integrand now transforms approximately under small elements of H_{τ_H} by a unitary character, so taking absolute values of the integrand leads to an integral that morally takes place on $(H/H_{\tau_H})^2$.

Recall now from Perron–Frobenius theory that the spectral radius of a matrix with nonnegative entries is bounded by the largest row sum, and that for a symmetric matrix, the spectral radius bounds the operator norm. If we formally apply this fact to the above integral with absolute value signs inserted (more precisely, a finite sum of such integrals having the required symmetry), then we are led to consider

$$\max_{y \in \mathcal{H}} \int_{x \in \mathcal{H}} |f_0(x^{-1}\gamma y)| \, dx \lessapprox \max_{y \in \mathcal{H}} \int_{x \in \mathcal{H}} \int_{\substack{z \in H_{\tau_H}: \\ z = \mathrm{O}(1)}} |f(x^{-1}\gamma y z)| \, dz \, dx.$$

Nontrivial bounds for this last expression are essentially equivalent to (1.10), but this approach to the reduction may render it less mysterious: we are "just" estimating the matrix coefficients of the natural vector $\varphi_{\pi,0}$ and applying Perron–Frobenius.

2.9. Coadjoint reformulation: further discussion. We explain here in more detail the content of the reformulation (1.12) of the required volume bounds in terms of coadjoint orbits.

We may similarly reformulate (1.10) in terms of the coadjoint action; it translates to an estimate like (1.12), but with Z_H replaced by H_{τ_H}. We have noted already in §2.6 that the orbit $\mathrm{Ad}^*(H)\tau$ is the relative coadjoint orbit $\mathcal{O}_{\pi,\sigma}$ defined in (2.11). It follows from the H-torsor property of $\mathcal{O}_{\pi,\sigma}$ that $\mathrm{Ad}^*(H_{\tau_H})\tau$ is the fiber over τ_H in $\mathcal{O}_{\pi,\sigma}$:

$$\mathrm{Ad}^*(H_{\tau_H})\tau = \mathcal{O}_{\pi,\sigma}(\tau_H) := \{\xi \in \mathcal{O}_{\pi,\sigma} : \xi_H = \tau_H\}.$$

The estimate (1.10) thus boils down to exhibiting some approximate transversality between the varieties

$$\mathrm{Ad}^*(\gamma)\mathcal{O}_{\pi,\sigma}(\tau_H) \quad \text{and} \quad \mathcal{O}_{\pi,\sigma}. \tag{2.14}$$

These varieties have respective dimensions n and n^2. Both are contained in the (n^2+n)-dimensional G-orbit $\mathcal{O}_\pi$. It thus seems reasonable to expect that for "generic" γ (in particular, $\gamma \notin HZ$), they should be literally transverse. The point of (1.12) is to exhibit an approximate form of such transversality using only the central direction in H_{τ_H}, i.e., replacing the n-dimensional variety $\mathcal{O}_{\pi,\sigma}(\tau_H)$ with its one-dimensional subvariety $\mathrm{Ad}^*(Z_H)\tau$.

We can visualize (2.14) directly only in the special case $n = 1$. That case is overly simplistic (e.g., because $Z_H = H_{\tau_H} = H$ and $\mathcal{O}_{\pi,\sigma}(\tau_H) = \mathcal{O}_{\pi,\sigma}$), but may nevertheless convey the flavor of the problem. The visualization becomes slightly simpler to describe if we switch to the setting of compact unitary groups $(G,H) = (\mathrm{U}(2), \mathrm{U}(1))$. Suppose then that π is an irreducible representation of $\mathrm{U}(2)$ with trivial central character and highest weight $r \asymp T$ and that σ is the trivial representation of $\mathrm{U}(1)$. The varieties $\mathcal{O}_\pi$ and $\mathcal{O}_{\pi,\sigma}$ are then contained in the trace zero subspace $(\mathfrak{g}^\wedge)^0$ of $\mathfrak{g}^\wedge$, which we may identify with $\mathbb{R}^3$ in such a way that $\mathcal{O}_\pi$ is the Euclidean sphere

$$\mathcal{O}_\pi = \{(x,y,z) : x^2 + y^2 + z^2 = (r+1/2)^2\}$$

and so that the projection $(\mathfrak{g}^\wedge)^0 \to \mathfrak{h}^\wedge \cong \mathbb{R}$ is the vertical coordinate map $(x,y,z) \mapsto z$. Then $\mathcal{O}_{\pi,\sigma} \subseteq \mathcal{O}_\pi$ is the equator

$$\mathcal{O}_{\pi,\sigma} = \{(x,y,0) : x^2 + y^2 = (r+1/2)^2\}.$$

We may take for $\tau \in \mathcal{O}_{\pi,\sigma}$ any point along that equator. The basic content of the required transversality is that the image of the equator under any nontrivial rotation of the sphere that fixes τ is transverse to the equator (see Figure 2).

Remark 2.7. There is another (closely related) way to arrive at the problem of controlling the transversality between the varieties (2.14). For this remark, we assume familiarity with [NV, §1]. The microlocal support of the vector $\varphi_{\pi,0}$ considered in

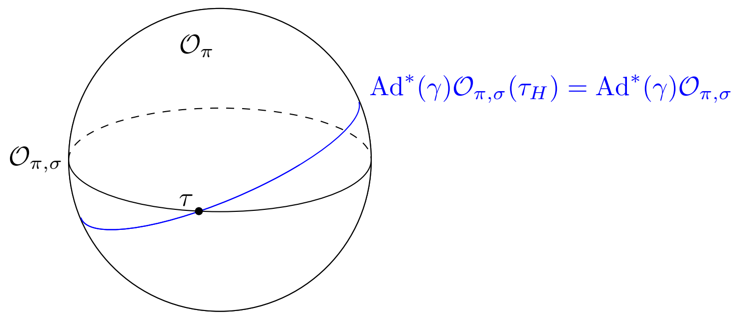

FIGURE 2. The required transversality for $(G, H) = (\mathrm{U}(2), \mathrm{U}(1))$.

Remark 2.6 is concentrated along the H_{τ_H}-orbit $\mathcal{O}_{\pi,\sigma}(\tau_H)$. The matrix coefficient of that vector at γ should thus be related to the overlap between small neighborhoods of that orbit and its image under $\mathrm{Ad}^*(\gamma)$. We noted in Remark 2.6 that the main point is to control the matrix coefficient of that vector along left H-cosets. Since $\mathcal{O}_{\pi,\sigma}$ is an H-torsor, we arrive again at the problem of establishing some transversality between the varieties (2.14).

2.10. Endgame: discussion. Recall from §1.5 that – having reduced the required moment estimate ((1.8) or (2.2)) to volume bounds, and reduced the latter to a Lie algebra problem – we must eventually confront the linear algebra problem addressed by Theorem 1.8. Here we outline the solution to that problem and record some miscellaneous remarks.

Theorem 1.8 is equivalent to the exclusion of the condition (1.13). Assuming for the sake of contradiction that that condition holds, we apply an infinitesimal form of the H_{τ_H}-torsor property of $\mathcal{O}_{\pi,\sigma}(\tau_H)$ to deduce that

$$\langle \mathfrak{h}_{\tau_H}, \mathrm{ad}^*(\mathrm{ad}(x)\mathfrak{z}_H)\tau\rangle = 0 \text{ for some } x \in \mathfrak{g}_\tau - \mathfrak{z}, \tag{2.15}$$

where $\mathfrak{h}_{\tau_H} := \mathrm{Lie}(H_{\tau_H})$. We may understand (2.15) as an infinitesimal form of the failure of transversality in (2.14), specialized to central directions. We interpret (2.15) as the degeneracy of a certain bilinear form on $\mathfrak{h}_{\tau_H} \times \mathfrak{g}_\tau/\mathfrak{z}$. We evaluate the determinant of that bilinear form with respect to a suitable basis. That determinant turns out to be a nonzero multiple of the resultant of the characteristic polynomials of τ and τ_H, hence vanishes precisely in the conductor dropping case.

The proof of Theorem 1.8 is then complete, modulo the verification of the determinantal identity. There seem to be a few ways to do this (see Remark 16.4). The given proof mimicks one of the standard "divisibility and homogeneity" calculations of the Vandermonde determinant, but with each step more complicated. We appeal repeatedly to the results of [NV, §14]. We eventually reduce to showing that there is *some* τ for which the determinant in question does not vanish. We conclude by exhibiting an explicit τ for which we may evaluate the determinant directly. (We take for τ a regular nilpotent element for which τ_H is a regular semisimple element having the nth roots of unity as eigenvalues.)

Remark 2.8. This last step is the only part of the proof of Theorem 1.1 specific to *unitary* GGP pairs. We leave to the reader the challenge of finding an analogous argument in the orthogonal case. One should presumably focus there not on the (zero-dimensional) center Z_H, but instead on the unique one-parameter subgroup of the centralizer H_{τ_H} of the regular element τ_H for which some maximal isotropic subspace is an eigenspace. The calculus arguments of §15 apply, so the main challenge is to find a substitute for the linear algebra arguments of §16.

Remark 2.9. We have noted already that the eigenvalue condition on τ in Theorem 1.8 comes from our "no conductor dropping" assumption. Indeed, the eigenvalues of τ (resp. τ_H) correspond to the parameters of π (resp. σ) after rescaling by T^{-1} and taking a limit. It may be instructive to note moreover that any special case of Theorem 1.8 yields a corresponding special case of Theorem 1.1. For example, the case in which τ and τ_H are both regular semisimple corresponds to that in which the parameters of π and σ are "dyadically spaced" in the sense of (2.4), so that the analogue of the "avoidance of Weyl chamber walls" condition (1.5) holds. On the other hand, the case in which τ and τ_H are both regular nilpotent (modulo the center) corresponds to the (twisted) t-aspect, as in Remark 1.3.

Remark 2.10. It would be interesting to understand whether ideas related to the case $n = 2$ of Theorem 1.8 are implicit in existing proofs of spectral aspect subconvex bounds on GL_3 [BB1, BB2, KMS, Sh].

3. General preliminaries

3.1. Asymptotic notation.

3.1.1. *Motivation.* Analytic number theory papers often employ asymptotic notation in the following informal way:

- *If $A \ll B$ and $B \ll C$, then $A \ll C$.*

Experienced readers know how to spell out such implications precisely:

- *For each $c_1, c_2 \geqslant 0$ there exists $c_3 \geqslant 0$ so that for all A, B, C with $|A| \leqslant c_1|B|$ and $|B| \leqslant c_2|C|$, we have $|A| \leqslant c_3|C|$.*

This paper will involve the iterated application of many such implications. To manage these coherently, it will be useful to take some care with foundational issues.

3.1.2. *Fixed elements.* We adopt E. Nelson's approach to nonstandard analysis [Ne], but with the terminological substitution of "fixed" for "standard." This choice of foundations should require little adjustment of habit for the reader familiar with analytic number theory papers.

Informally, a "fixed" quantity is an "absolute constant," i.e., an element of some finite collection of objects in our mathematical universe, taken sufficiently large for the purposes of the paper. Formally, "fixed" is an undefined predicate equipped with some axioms for manipulating it, detailed in [Ne], which yield a relatively consistent extension of ZFC.

This formulation of nonstandard analysis is well-suited to our aims. It does not require constructing new objects, such as the hyperreals; one starts instead with the "usual" field of real numbers and introduces an undefined predicate "fixed" (or "standard") that applies to some of them. We do not use any of the more sophisticated constructions of nonstandard analysis, but do require a coherent way to organize and iterate certain estimates.

The usual asymptotic notation and terminology is defined accordingly. If $|A| \leqslant C|B|$ for some fixed $C \geqslant 0$, then we write $A \ll B$ or $B \gg A$ or $A = \mathrm{O}(B)$ ("A is not much larger than B"); otherwise, we write $B \lll A$ or $A \ggg B$ or $B = o(A)$ ("A is much larger than B"). We write $A \asymp B$ as shorthand for $A \ll B \ll A$. We often introduce absolute value signs, as in $|A| \asymp |B|$, for cosmetic reasons. The notation $A \simeq B$ signifies that $A = B + o(1)$. When an infinite exponent appears in an estimate, the meaning is that the indicated estimate holds with that exponent replaced by any fixed integer (e.g., $A \ll \mathrm{h}^{\infty} B$ means that $A \ll \mathrm{h}^{N} B$ for each fixed N). Notation such as $A \ll_j B$ signifies that $A \ll B$ holds for all fixed j. We write ε for a positive quantity, typically fixed and small enough. "Fix n" carries the same meaning as "Let n be fixed."

While the notion "fixed," and hence any asymptotic notation defined in terms of that notion, does not carry any intrinsic meaning, any statement that refers to that notion may be algorithmically translated to one that does not. The algorithm is detailed in [Ne, §2], but may be inferred from the informal equivalence between "fixed" and "absolute constant." An example was given in §3.1.1.

Experience suggests that the notation $A \lll B$ (or equivalently, $A = o(B)$) is more problematic than the more customary notation $\ll$ and $\gg$. We emphasize that it carries the same content as "not $A \gg B$." By introducing negations and taking contrapositives, any statement that involves $\lll$ or $\ggg$ may be translated to one that does not. For example, the following statements are formally equivalent to each other (and true):

- *If $A \lll B$ and $B \ll C$, then $A \lll C$.*
- *For each $c_2, c_3 \geqslant 0$ there exists $c_1 \geqslant 0$ so that for all A, B, C with $|A| < c_1|B|$ and $|B| \leqslant c_2|C|$, we have $|A| < c_3|C|$.*
- *If $A \gg C$ and $B \ll C$, then $A \gg B$.*
- *For each $c_2, c_3 \geqslant 0$ there exists $c_1 \geqslant 0$ so that for all A, B, C with $|A| \geqslant c_3|C|$ and $|B| \leqslant c_2|C|$, we have $|A| \geqslant c_1|B|$.*

We give a slightly more involved example of such translation in Remark 9.4.

3.1.3. *Classes.* The discussion here becomes relevant only starting in §9.3, so the reader may wish to refer back as needed.

"Illegal set formation" is the forbidden practice of defining a set using "fixed" (or anything defined in terms of that notion, such as the above asymptotic notation). For instance, there does not exist a subset $S \subseteq \mathbb{Z}$ consisting of precisely the integers n with $n = \mathrm{O}(1)$; note that no axiom of ZFC allows one to construct such a set. The customary way to work algebraically with predicates such as "$n = \mathrm{O}(1)$" (e.g., to assert that elements satisfying that predicate satisfy the ring axioms) is to introduce a dichotomy between "internal" and "external" sets, as in [Ne, §3], [Ka] or [Hr, §3]. For our limited purposes, a cheaper approach suffices.

Definition 3.1. By a *class*, we mean a pair (X, P), where X is a set and P is a boolean formula that takes an element $x \in X$ as its argument and whose formulation may involve "fixed."

We write "S is the class of all $x \in X$ satisfying $P(x)$" or simply $S = \{x \in X : P(x)\}$ as longhand for $S = (X, P)$. (In the language of the references noted above, classes define external subsets of internal sets.)

Classes play a purely syntactic role for us, giving a way to define symbolic expressions similar to "$\mathrm{O}(B)$." For instance, given a scalar domain X (e.g., $X = \mathbb{C}$) and a scalar $B \in X$, we may define the scalar class $\mathrm{O}(B)$ to be the pair (X, P), where $P(A)$ is the predicate "there is a fixed C so that $|A| \leqslant C|B|$." This particular class will not play a role in this paper, but illustrates the basic idea.

We define only a handful of classes (Definitions 9.2, 9.8, 10.1 and mild derivatives thereof), but the notion provides a useful way to organize our estimates.

We may regard sets as classes (take for P the predicate "true") and work with classes in much the same way we work with sets. An element x belongs to the class (X, P) if $x \in X$ and if $P(x)$ is true. One class (X, P) is contained in another class (Y, Q) if for every $x \in X$ satisfying $P(x)$, we have $x \in Y$ and $Q(x)$. Two classes are equal if they contain the same elements, or equivalently, if each is contained in the other. We may define intersections or unions of classes, class maps $f : (X, P) \to (Y, Q)$ between classes (i.e., subclasses of the product class satisfying the map axioms), and so on.

3.2. Algebraic groups. Let F be a field of characteristic zero, and let G be a linear algebraic group over F. When G is connected reductive, we often identify G with its set of F-points: $G := G(F)$. When F is a global field, we write $[G] := G(F)\backslash G(\mathbb{A})$ for the corresponding adelic quotient, equip $G(\mathbb{A})$ with Tamagawa measure as in [BCZ, §2.3.2], and equip $[G]$ with the corresponding quotient measure.

3.3. Unitary groups. Let F be a field of characteristic zero. Let E/F be quadratic étale F-algebra, thus either $E = F \times F$ or E/F is a quadratic field extension. Let ι denote the involution of E fixing F. Let $(V, \langle\rangle)$ be an E-vector space equipped with a nondegenerate ι-hermitian form. We write $\mathrm{GL}(V/E)$ for the group of E-linear automorphisms of V. The subgroup

$$G := \mathrm{Aut}(V/E, \langle\rangle)$$

consisting of automorphisms that preserve the given hermitian form is then an algebraic group. It comes with a standard embedding $G \hookrightarrow \mathrm{GL}(V/E)$. By a *unitary group* over F, we mean a group arising in this way, together with its standard embedding.

In the split case $E = F \times F$, we may decompose $V = V^+ \oplus V^-$ as a sum of two F-vector spaces, with E acting on the summands via the two projections to F. We may identify V^- with the dual of V^+ and

$$G \cong \mathrm{GL}(V^+)$$

with the set of pairs $(g, {}^t g^{-1}) \in \mathrm{GL}(V^+) \times \mathrm{GL}(V^-)$ (see [NV, §13.3] for details).

3.4. GGP pairs. Retaining the setup of §3.3, let e be an element of V for which Ee is a free rank one E-module on which the form $\langle\rangle$ is nondegenerate. Writing $V_H \subseteq V$ for the orthogonal complement of Ee, we then have

$$V = V_H \oplus Ee.$$

The group

$$H := \mathrm{Aut}(V_H/E, \langle\rangle)$$

embeds in G. By a *unitary GGP pair* over F, we mean a pair (G, H) arising in this way, together with its standard embeddings $H \hookrightarrow G \hookrightarrow \mathrm{GL}(V/E)$ and choice of $e \in E$.

We write Z for the center of G and $\bar{G} := G/Z$ for the adjoint group. We note that the intersection $Z \cap H$ is trivial, so we may regard H also as a subgroup of $\bar{G}$.

In the split case $E = F \times F$, we may decompose $V = V^+ \oplus V^-$ and $V_H = V_H^+ \oplus V_H^-$ as in §3.3, and thereby identify

$$H \cong \mathrm{GL}(V_H^+) \hookrightarrow G \cong \mathrm{GL}(V^+).$$

We may analogously define an *orthogonal group* G over F and an *orthogonal GGP pair* (G, H) by taking $E := F$ in the above discussion (see [NV, §13] for details). By a *GGP pair*, we mean a unitary or orthogonal GGP pair. (We focus in this paper on the unitary case, but many of our results apply just as well in the orthogonal case, so we state them in their natural generality.)

3.5. Local distinction and matrix coefficient integrals. Let (G, H) be a GGP pair over a local field F. Let π and σ be irreducible unitary representations of G and H. More precisely, we write π and σ for the spaces of smooth vectors in the given Hilbert spaces.

We say that the pair (π, σ) is *distinguished* if there is a nonzero H-invariant functional $\ell : \pi \to \sigma$. It is known that the space of such functionals is then one-dimensional (see [SZ]).

Suppose now that π and σ are tempered and that we are given a Haar measure dh on H. Then for $v \in \pi$ and $u \in \sigma$, the integral

$$\int_{h \in H} \langle hv, v \rangle \langle u, hu \rangle \, dh$$

converges absolutely and defines a quadratic form $\mathcal{Q}$ on $\pi \otimes \sigma$ (see [NV, §18], which refers to [II, Prop 1.1] and [Ha, §2]). It is expected that

(i) $\mathcal{Q}$ is not identically zero if and only if (π, σ) is distinguished, and
(ii) $\mathcal{Q}$ is nonnegative-valued (rather than merely real-valued).

Expectation (i) is known to hold in the unitary case [BP1], so in that case, we may normalize the nonzero invariant functional $\ell : \pi \to \sigma$ for distinguished (π, σ) by requiring that the identity

$$\mathcal{Q}(v \otimes u) = \pm |\langle \ell(v), u \rangle|^2 \tag{3.1}$$

hold for some sign ± 1 depending only upon (π, σ); in particular, $\mathcal{Q}$ is (positive or negative) definite. Both expectations are known to hold in (at least) the following cases, in which we may provably normalize so that (3.1) holds with the expected sign $+1$:

- F is non-archimedean, in which see [BP1, Thm 5], [Wa, Prop 5.7], [SV].
- F is archimedean and G and H are compact, in which the expectation follows readily from the Schur orthogonality relations for compact groups.
- F is archimedean and π, σ arise by taking local components at the distinguished place $\mathfrak{q}$ of some constituent of the family $\mathcal{F}_T$ defined in §1.3, with T large enough. In that case, the positivity follows from the relative character asymptotics established in [NV, §19] and recalled below in §12.

3.6. Global notation and assumptions. Let F be a fixed number field. We denote by $\mathbb{Z}_F$ its ring of integers and by $\mathbb{A}$ its adele ring. For a place $\mathfrak{p}$ of F, we write $F_\mathfrak{p}$ for the completion. If $\mathfrak{p}$ is finite, we write $\mathbb{Z}_\mathfrak{p} \subseteq F_\mathfrak{p}$ for the ring of integers and $q_\mathfrak{p}$ for the absolute norm. For a finite set S of places of F, we write $F_S := \prod_{\mathfrak{p} \in S} F_\mathfrak{p}$.

Let (G, H) be a fixed unitary GGP pair over F, with standard representation $G \hookrightarrow \mathrm{GL}(V/E)$. We take for S a fixed finite set of places of F that is sufficiently large in the following senses.

- S contains every archimedean place.
- G and H admit smooth models over $\mathbb{Z}_F[1/S]$ and the inclusion $H \hookrightarrow G$ extends to a closed immersion. We continue to denote simply by G and H these smooth models. For each $\mathfrak{p} \notin R$, the groups $G(\mathbb{Z}_\mathfrak{p}) \leqslant G_\mathfrak{p}$ and $H(\mathbb{Z}_\mathfrak{p}) \leqslant H_\mathfrak{p}$ are then hyperspecial maximal compact subgroups.
- $\mathbb{Z}_F[1/S]$ and $\mathbb{Z}_E[1/S]$ are principal ideal domains.
- $G(F)G(F_S)\prod_{\mathfrak{p}\notin S} G(\mathbb{Z}_\mathfrak{p}) = G(\mathbb{A})$, and similarly for H.

The last two assumptions are likely technical, but convenient.

Under the above assumptions, we fix factorizations of the Tamagawa measures on $G(\mathbb{A})$ and $H(\mathbb{A})$ that assign volume one to $G(\mathbb{Z}_\mathfrak{p}) \subseteq G(F_\mathfrak{p})$ and $H(\mathbb{Z}_\mathfrak{p}) \subseteq H(F_\mathfrak{p})$ for all $\mathfrak{p} \notin S$.

3.7. Branching coefficients. We retain the setting of §3.6. Recall the family $\mathcal{F}$ defined in §1. Let $(\pi, \sigma) \in \mathcal{F}$. We equip π (resp. σ) with the Petersson inner product defined by integrating over $[G/Z]$ (resp. $[H/Z_H]$). As in §3.5, we write π and σ for the spaces of smooth vectors in the given Hilbert spaces.

Writing $\pi_S \subseteq \pi$ and $\sigma_S \subseteq \sigma$ for the subspaces of vectors invariant by $\prod_{\mathfrak{p}\notin S} G(\mathbb{Z}_\mathfrak{p})$ and $\prod_{\mathfrak{p}\notin S} H(\mathbb{Z}_\mathfrak{p})$, respectively, we define quadratic forms $\mathcal{P}$ and $\mathcal{Q}$ on $\pi_S \otimes \sigma_S$ by the formulas

$$\mathcal{P}(v \otimes u) := \left| \int_{[H]} v\bar{u} \right|^2$$

and, as in §3.5,

$$\mathcal{Q}(v \otimes u) := \int_{h \in H(F_S)} \langle hv, v\rangle \langle u, hu\rangle \, dh.$$

As recalled in §3.5, the local distinction of (π, σ) implies that $\mathcal{Q}$ is not identically zero. By multiplicity one, we may thus define a scalar $\mathcal{L}(\pi, \sigma) \in \mathbb{R}_{\geqslant 0}$ by the relation

$$\mathcal{P} = \mathcal{L}(\pi, \sigma) \cdot |\mathcal{Q}|. \tag{3.2}$$

As discussed in §3.5, it is expected that $\mathcal{Q}$ is positive-definite, so that the absolute value signs above are not necessary. This expectation is likely provable in general, but not written down to the best of our knowledge. It follows for $(\pi, \sigma) \in \mathcal{F}_T$ with T large enough from the discussion of §3.5, and we will apply (3.2) only to such pairs (π, σ). For our proofs, we require only the obvious positivity of $\mathcal{P}$ (rather than that of $\mathcal{L}(\pi, \sigma)$, contrary to what the simplified presentation of §2 might suggest).

3.8. Period formulas. The conjectures of Ichino–Ikeda and N. Harris (see [II, Ha] and [NV, §25.5]) assert that, with notation and assumptions as above, and with $L^{(S)}$ denoting the partial L-function obtained by omitting Euler factors in S,

$$\mathcal{L}(\pi, \sigma) = 2^{-\beta} \frac{L^{(S)}(\pi_E \otimes \sigma_E^\vee, 1/2)}{L^{(S)}(\mathrm{Ad}, \pi \boxtimes \sigma^\vee, 1)} \Delta_G^{(S)}$$

for some specific $\beta = \mathrm{O}(1)$ and $\Delta_G^{(S)} \asymp 1$. A proof is given in [BCZ, Thm 1.1.6.1] under the assumption that π and σ are everywhere tempered. (In fact, an inspection of the proof suggests that it should suffice to assume that $\pi_\mathfrak{p}$ and $\sigma_\mathfrak{p}$ are tempered for all $\mathfrak{p} \in S$, with S sufficiently large.)

4. Statement of main local result

In §4, we formulate our main local result, Theorem 4.2. In §6, we explain how Theorem 4.2 implies our main global result, Theorem 1.1. The remainder of the paper will then be devoted to the proof of Theorem 4.2.

4.1. Norms.

4.1.1. Let F be a local field. We denote by $|.|_F : F \to \mathbb{R}_{\geqslant 0}$ the modulus. Let V be a finite-dimensional vector space over F. By an F-*norm* on V, we mean a continuous function $|.|_F : V \to \mathbb{R}_{\geqslant 0}$ such that

- $|v|_F = 0$ only if $v = 0$, and
- $|tv|_F = |t|_F |v|_F$ for all $t \in F, v \in V$.

For example, given an F-basis $e_1, \dots, e_n$ of V, we obtain a "standard" F-norm $|.|_F$ by

$$|\sum c_j e_j|_F = \begin{cases} \max_j |c_j|_F & \text{if } F \text{ is non-archimedean,} \\ (\sum_j |c_j|_F^2)^{1/2} & \text{if } F = \mathbb{R}, \\ \sum_j |c_j|_F & \text{if } F = \mathbb{C}. \end{cases}$$

Any two F-norms are comparable in the sense that each is bounded from above and below by a positive real multiple of the other, so the precise choice is unimportant when F and V are fixed; in that case, we often write $|.|_F$ for some fixed but unspecified F-norm.

When $F = \mathbb{R}$, we often write simply $|.|$ rather than $|.|_{\mathbb{R}}$.

4.1.2. Let V be a finite-dimensional vector space over a global field F. By choosing a basis $e_1, \dots, e_n$ of V, we obtain from the above construction for each place $\mathfrak{p}$ of F an $F_{\mathfrak{p}}$-norm $|.|_{F_{\mathfrak{p}}}$ on the completion $V_{\mathfrak{p}}$.

4.1.3. If E/F is an extension of local fields and V is an E-vector space, then we may also regard V as an F-vector space, hence we may speak of E-norms $|.|_E$ and F-norms $|.|_F$ on V. Any such norms are related by the estimate (for fixed $E, F, V, |.|_F, |.|_E$)

$$|v|_E \asymp |v|_F^{[E:F]}. \tag{4.1}$$

4.1.4. Suppose now that $E = F \times F$ is the split quadratic extension of a local field F and that V is an E-vector space. We may then decompose

$$V = V^+ \oplus V^-,$$

where E acts on the summands via the two projections to F. Writing $v = v^+ + v^-$ for the corresponding decomposition of a vector $v \in V$, we have (for fixed $F, V, |.|_F$)

$$|v|_F \asymp |v^+|_F + |v^-|_F \gg \sqrt{|v^+|_F |v^-|_F}. \tag{4.2}$$

4.1.5. Let G be a linear algebraic group over a local field F. Suppose given some faithful embedding $G \hookrightarrow \mathrm{SL}_N(F)$. By embedding $\mathrm{SL}_N(F)$ further in the space $M_N(F)$ of $N \times N$ matrices and choosing an F-norm $|.|_F$ on $M_N(F)$, we may define $|g - 1|_F$ for $g \in G$.

Suppose that F and G are fixed. Let $|.|_F$ and $|.|_F'$ be any fixed pair of norms arising as above. For $g \in G$, we then have $|g-1|_F \lll 1$ if and only if $|g-1|_F' \lll 1$, in which case $|g-1|_F \asymp |g-1|_F'$. Thus $|g-1|_F$ gives a well-defined notion of the distance from small elements $g \in G$ to the identity element.

4.2. Quantifying distance to the subgroup H. Let (G, H) be a fixed unitary GGP pair over a fixed local field F, with standard representation $G \hookrightarrow \mathrm{GL}(V/E)$ and accompanying decomposition

$$V = V_H \oplus Ee. \tag{4.3}$$

Recall that we denote by Z the center of G and by $\bar{G} = G/Z$ the adjoint group. The purpose of this section is to record a quantification of the "distance" between an element $g \in \bar{G}$ and the subgroup $H \hookrightarrow \bar{G}$. Since we will eventually need to perform explicit calculations involving this quantification, we will be explicit here.

The basic idea can be seen in the example $H = \mathrm{GL}_1(F) \hookrightarrow \bar{G} = \mathrm{PGL}_2(F)$. An element of $\bar{G}$ is described by a scaling class of 2×2 matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Such an element lies in H precisely when the ratios b/d and c/d both vanish. The sizes of those ratios thus quantify the distance from that element to H.

Turning to details, we may regard V as a vector space over E (which need not be a field) or over F. We fix an F-norm $|.|_F : V \to \mathbb{R}_{\geqslant 0}$. We then define $\tilde{d}_H : \mathrm{PGL}(V/E) \to [0, \infty]$ as follows. Temporarily lifting g to an element of $\mathrm{GL}(V/E)$, the decomposition of the vector $ge \in V$ with respect to (4.3) is given by

$$ge = \left(ge - \frac{\langle ge, e \rangle}{\langle e, e \rangle} e \right) + \frac{\langle ge, e \rangle}{\langle e, e \rangle} e.$$

We define $\tilde{d}_H(g)$ to be the size of the vector obtained by dividing the V_H-component of ge by the coefficient of $e/\langle e, e\rangle$ in the Ee-component of ge:

$$\tilde{d}_H(g) := \left| \frac{ge}{\langle ge, e \rangle} - \frac{e}{\langle e, e \rangle} \right|_F,$$

with the convention that $\tilde{d}_H(g) := \infty$ if $\langle ge, e \rangle \notin E^\times$. We then define

$$d_H : \mathrm{PGL}(V/E) \to [0, 1],$$
$$d_H(g) := \min(1, \tilde{d}_H(g) + \tilde{d}_H(g^{-1})).$$

We will be interested primarily in the restrictions of $\tilde{d}_H$ and d_H to $\bar{G}$. Since $H \hookrightarrow \bar{G}$ is the stabilizer of the line Ee, it is clear that

$$H = \{ g \in \bar{G} : d_H(g) = 0 \},$$

so we may regard d_H as quantifying how far $g \in \bar{G}$ is from H.

We may explicate these definitions in terms of matrix entries. We choose an orthogonal E-basis $e_1, \dots, e_{n+1}$ of V with $e_{n+1} = e$. We accordingly identify $\mathrm{GL}(V/E) = \mathrm{GL}_{n+1}(E)$ and $\mathrm{GL}(V_H) = \mathrm{GL}_n(E)$, with the latter included in the former as the subgroup of upper-left block matrices. For $g \in \mathrm{PGL}(V/E) = \mathrm{PGL}_{n+1}(E)$, we may write

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{4.4}$$

with a, b, c, d matrices of respective dimensions $n \times n, n \times 1, 1 \times n, 1 \times 1$, well-defined up to simultaneous scaling. We similarly write

$$g^{-1} = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}. \tag{4.5}$$

Then

$$\tilde{d}_H(g) \asymp |b/d|_F, \quad d_H(g) \asymp \min(1, |b/d|_F + |c'/d'|_F),$$

where the implied constants quantify the ambiguity in defining F-norms.

To illustrate the definition of d_H further, we explicate its restriction to $\bar{G}$ in the split case $E = F \times F$. As explained in §3.4, we may identify

$$H = \mathrm{GL}_n(F) \hookrightarrow G = \mathrm{GL}_{n+1}(F) \hookrightarrow \mathrm{GL}(V/E) = \mathrm{GL}_{n+1}(F) \times \mathrm{GL}_{n+1}(F),$$

with the final inclusion given by $g \mapsto (g, {}^t g^{-1})$. For $g \in \bar{G} = \mathrm{PGL}_{n+1}(F)$, we have with the notation (4.4) and (4.5) that

$$\tilde{d}_H(g) \asymp |b/d|_F + |c'/d'|_F,$$
$$d_H(g) \asymp \min(1, |b/d|_F + |b'/d'|_F + |c/d|_F + |c'/d'|_F).$$

The essential properties of d_H are summarized in the following lemma.

Lemma 4.1. *Let $g \in \bar{G}$.*

(i) We have
$$d_H(g) \ll |\operatorname{Ad}(g) - 1|_F. \tag{4.6}$$

(ii) Let $u_1, u_2 \in \bar{G}$. Suppose that g belongs to a fixed compact neighborhood of the identity and that
$$|\operatorname{Ad}(u_j) - 1|_F \lll d_H(g) \quad (j = 1, 2).$$
Then
$$d_H(u_1 g u_2) \asymp d_H(g). \tag{4.7}$$

(iii) Let $h_1, h_2 \in H$. Suppose that g, h_1, h_2 each belong to some fixed compact neighborhood of the identity. Then
$$d_H(h_1 g h_2) \asymp d_H(g). \tag{4.8}$$

Proof. We abbreviate $|.| := |.|_F$. For $x \in \operatorname{End}(V/E)$, we set

$$\nu(x) := \langle xe, e \rangle / \langle e, e \rangle$$

and

$$\mu(x) := xe - \nu(x)e$$

so that $xe = \mu(x) + \nu(x)e$ with $\mu(x) \in V_H$ and $\nu(x) \in E$. Then $\tilde{d}_H(g) \asymp |\mu(g)/\nu(g)|$, with the RHS well-defined.

Suppose that $g \in G$ lies in a fixed compact set. We may define $\tilde{d}_H(g)$ and $d_H(g)$ by composing with the quotient map $G \to \bar{G}$. We claim that

$$d_H(g) \asymp |\mu(g)|. \tag{4.9}$$

To see this, note first that if $|\mu(g)| \asymp 1$, then $|\mu(g)/\nu(g)| \gg 1$, so $d_H(g) \asymp 1 \asymp |\mu(g)|$. We may thus suppose that $|\mu(g)| \lll 1$. Since $\mu(g)$ and $\nu(g)$ describe the rightmost column of $g \in \mathrm{GL}_{n+1}(E)$, we see then that $\nu(g)$ lies in a fixed compact subset of $E^\times$. Since each entry of g as well as $1/\det(g)$ is $\mathrm{O}(1)$, we deduce by Cramer's rule that $\nu(g^{-1})$ likewise lies in a fixed compact subset of $E^\times$ and that $|\mu(g^{-1})| \asymp |\mu(g)|$. Thus $\tilde{d}_H(g) \asymp |\mu(g)|$ and $\tilde{d}_H(g^{-1}) \asymp |\mu(g^{-1})| \asymp |\mu(g)|$. The estimate (4.9) follows.

We now let $g \in \bar{G}$ and prove each assertion from the lemma in turn.

(i) Since $d_H(g) \leqslant 1$, we may assume that $|\operatorname{Ad}(g) - 1| \lll 1$, as the required estimate is otherwise trivial. We may then lift g to an element of G with $|g - 1| \asymp |\operatorname{Ad}(g) - 1|$; in particular, $g \simeq 1$. Writing $\mu(g) = (g-1)e - \nu(g-1)e$, we deduce that
$$d_H(g) \asymp |\mu(g)| \ll |g - 1| \ll |\operatorname{Ad}(g) - 1|,$$
as required.

(ii) Since g lies in a fixed compact neighborhood of the identity, we may lift g to an element of G with the same property. We may likewise lift u_j to an element of G with $|u_j - 1| \lll d_H(g) \asymp \mu(g)$ (by (4.9)). Then $\mu(u_1 g u_2) = \mu(g) + o(\mu(g))$, hence $|\mu(u_1 g u_2)| \asymp |\mu(g)|$. We conclude via (4.9).

(iii) Since H stabilizes e and h_1 lies in a fixed compact subset of H, we see that

$$|\mu(h_1 g h_2)| = \left| h_1 \left(\frac{ge}{\langle ge, e \rangle} - \frac{e}{\langle e, e \rangle} \right) \right| \asymp |\mu(g)|.$$

We conclude once again via (4.9).

□

4.3. Construction of analytic test vectors: statements. Fix a GGP pair (G, H) over an archimedean local field F, with standard representation $G \hookrightarrow \mathrm{GL}(V/E)$. Set $n+1 := \dim_E(V)$, and write $\bar{G} = G/Z$ as usual. We assume given fixed Haar measures on each group. The following is the main local result of this paper.

Theorem 4.2. *Let T be a positive real with $T \ggg 1$. Let π and σ be tempered irreducible unitary representations of G and H, respectively. Assume that (π, σ) satisfies the following conditions (i.e., the local conditions imposed at the distinguished place "$\mathfrak{q}$" in the definition of "$\mathcal{F}_T$" from §1):*

- *(π, σ) is orbit-distinguished (§11.3),*
- *$T = \max(\{|\lambda_{\pi,i}|_F\} \cup \{|\lambda_{\sigma,j}|_F\})$ is the size of the largest archimedean Satake parameter of π or σ (§7.4), and*
- *we have $|\lambda_{\pi,i} - \lambda_{\sigma,j}|_F \gg T$ for all i and j, or equivalently, $C(\pi, \sigma) \asymp T^{2n(n+1)}$.*

Then for each fixed $\kappa > 0$ and fixed open neighborhood U of the identity element in G, there is a test function $f \in C_c^\infty(U)$ and a smooth unit vector $u \in \sigma$ with the following properties.

Let $\omega_\pi : Z \to \mathrm{U}(1)$ denote the central character of π. Define $f^\sharp \in C_c^\infty(\bar{G}, \omega_\pi^{-1})$ by the formula

$$f^\sharp(g) := \int_{z \in Z} \omega_\pi(z)(f * f^*)(zg)\, dz, \tag{4.10}$$

where $f^*(g) := \overline{f(g^{-1})}$ and $f * f^*$ denotes the convolution product.

(i) With $\mathcal{Q}$ as in §3.5, we have

$$\sum_{v \in \mathcal{B}(\pi)} \mathcal{Q}(\pi(f)v \otimes u) \gg T^{-n^2/2-\kappa}. \tag{4.11}$$

where $\mathcal{B}(\pi)$ denotes any orthonormal basis.

(ii) We have

$$\int_H |f^\sharp| \ll T^{n/2+\kappa}.$$

(iii) Let Ω, Ω' be fixed compact neighborhoods of the identity element of H, with Ω' large enough in terms of Ω. Let $\Psi_1, \Psi_2 : H \to \mathbb{C}$ be measurable functions satisfying $\Psi_j(gz) = \omega(z)\Psi_j(g)$ for some unitary character $\omega : Z_H \to \mathrm{U}(1)$ and all $g \in H, z \in Z_H$. Let $\gamma \in \bar{G} - H$. Then

$$\int_{x,y \in \Omega} \left| \overline{\Psi_1(x)} \Psi_2(y) f^\sharp(x^{-1}\gamma y) \right| dx\, dy \ll \frac{T^{n/2-1/4+\kappa}}{d_H(\gamma)^{1/2}} \|\Psi_1\|_{L^2(\Omega')} \|\Psi_2\|_{L^2(\Omega')}.$$

Proof. See §14. ☐

For details concerning the convergence of sums as in (4.11), we refer to [NV, §18] and §12.

Remark 4.3. There is potential to improve part (iii) by employing Perron–Frobenius theory in place of Cauchy–Schwarz (see Remarks 2.6 and 14.9). For instance, this works when $n = 1$. Such a strengthening should lead to an improvement in the numerical exponent of Theorem 1.1. One could also attempt a stronger estimate for the integral without absolute value signs.

5. Preliminaries for the proof of the reduction

In this section we record some preliminaries for our deduction, given in §6, of Theorem 1.1 from Theorem 4.2. Some of these results may be more broadly useful, but they will not be otherwise applied in this paper.

5.1. Uniform distinction at auxiliary places. Let (G, H) be a GGP pair over a local field F. We assume given a Haar measure dh on F. We define quadratic forms $\mathcal{Q}$ as in §3.5. All representations considered here are assumed smooth. For a smooth unitary representation V of some group, we write $\mathcal{B}(V)$ for an orthonormal basis consisting of vectors isotypic under some chosen maximal compact subgroup.

Lemma 5.1. *Suppose that F is archimedean and that G and H are compact. For each irreducible unitary representation π of G and open neighborhood Ω of the identity element in G, there is a positive real c and a test function $f \in C_c^\infty(\Omega)$ with the following property. Let σ be an irreducible unitary representation H for which (π, σ) is distinguished. There is a unit vector $u \in \sigma$ with*

$$\sum_{v \in \mathcal{B}(\pi)} \mathcal{Q}(\pi(f)v, u) \geqslant c.$$

Proof. We take for f a smooth bump with $\int_G f = 1$ that is supported close enough to the identity, where the meaning of "close enough" depends upon of π and Ω. Since G is compact, the representation π is finite-dimensional. The operator $\pi(f)$ thus approximates the identity operator on π, and so the above sum is at least

$$\frac{1}{2} \sum_{v \in \mathcal{B}(\pi)} \mathcal{Q}(v, u). \tag{5.1}$$

Let c_σ denote the maximum of (5.1) over all unit vectors $u \in \sigma$. By our distinction assumption, we have $c_\sigma > 0$. Since $\pi|_H$ is finite-dimensional, there are only finitely many possibilities for the isomorphism class of σ. We conclude taking for c the minimum of c_σ over σ. ☐

Lemma 5.2. *Suppose that F is non-archimedean. For each pair of compact open subgroups $J_G \leqslant G$ and $J_H \leqslant H$, there is a positive real c and a pair of arbitrarily small compact open subgroups $U_G \leqslant G$ and $U_H \leqslant H$ with the following property. Let π and σ be tempered irreducible unitary representations of G and H. Assume that (π, σ) is distinguished and that the invariant subspaces π^{J_G} and σ^{J_H} are nontrivial. Then there are unit vectors $v \in \pi^{U_G}$ and $u \in \sigma^{U_H}$ so that $\mathcal{Q}(v, u) \geqslant c$.*

Proof. By uniform admissibility (see [MV2, §2.6.3]), the dimensions of the spaces π^{U_G} and σ^{U_H} are bounded by a constant depending at most upon U_G and U_H. It is thus equivalent to produce $c > 0$ and U_G, U_H so that

$$\sum_{v \in \mathcal{B}(\pi^{U_G})} \sum_{u \in \mathcal{B}(\sigma^{U_H})} \mathcal{Q}(v,u) \geqslant c.$$

To conclude, we apply the following result to the representation $\pi \otimes \sigma^\vee$ of $G \times H$. □

Lemma 5.3. *Suppose that F is non-archimedean. For each compact open subgroup J of $G \times H$, there is a positive real c and an arbitrarily small compact open subgroup U of $G \times H$ so that for every tempered irreducible representation π of $G \times H$ for which*

- *π^J is nontrivial, and*
- *π is distinguished, i.e., there is a nonzero H-invariant functional $\ell : \pi \to \mathbb{C}$,*

we have

$$\sum_{v \in \mathcal{B}(\pi^U)} \int_{h \in H} \langle hv, v \rangle \, dh \geqslant c. \tag{5.2}$$

Proof. The proof is similar to that of [NV, §24.5.1].

Let π satisfy the stated hypotheses. It follows from the facts noted in §3.5 that the LHS of (5.2) converges absolutely, is nonnegative for all U, and is positive for some U.

We recall (see [NV, §23.4.3]) that every tempered irreducible representation π of $G \times H$ arises as a submodule of the normalized induction $i_M \tau := \mathrm{Ind}_P^{G \times H} \tau$ for some parabolic subgroup $P = MN$ of $G \times H$ and some irreducible representation τ of M that is *square-integrable*, i.e., has unitary central character and matrix coefficients square-integrable modulo the center. The conjugacy class of the pair (M, τ) is moreover unique. The unitary representation $i_M \tau$ decomposes as a finite direct sum of tempered irreducible representations of $G \times H$, one of which is π. By strong multiplicity one (see [NV, §24.1] and references), π is the only distinguished summand of $i_M \tau$. It follows that

$$\sum_{v \in \mathcal{B}(\pi^U)} \int_{h \in H} \langle hv, v \rangle \, dh = \sum_{v \in \mathcal{B}((i_M \tau)^U)} \int_{h \in H} \langle hv, v \rangle \, dh.$$

Given a pair (M, τ) as above, we obtain another such pair by replacing τ with its twist τ_t by any unramified character t of M. As explained in [NV, §23.8], we may find a finite list of pairs (M, τ) so that whenever $\pi^J \neq 0$, there is an element (M, τ) from this list so that $\pi \hookrightarrow i_M(\tau_t)$ for some t. Since π is distinguished, we know by the proof of [NV, §24.1, Lemma 2] that each of the representations $i_M(\tau_t)$ is distinguished. We may thus find for each t_0 a compact open subgroup U so that

$$\sum_{v \in \mathcal{B}((i_M \tau_t)^U)} \int_{h \in H} \langle hv, v \rangle \, dh > 0 \tag{5.3}$$

when $t = t_0$. The LHS of (5.3) varies continuously in t, so the same choice of U works uniformly in a small neighborhood of t_0. Since the group of unramified characters of M is compact, we may choose U small enough that (5.3) holds for all t. By continuity and compactness, the LHS of (5.3) is thus bounded from below by some $c > 0$. This completes the proof. □

5.2. Constructing an amplifier. An amplifier on general semisimple groups is constructed by Silberman–Venkatesh [SV, Lemma A.1]. The output of their construction is formulated qualitatively, so invoking it here as a black box would not lead to an effective numerical exponent δ as in Theorem 1.1. Blomer–Maga [BM1, §4] give an explicit and quantitative amplifier in the setting of GL_n. We will recall their construction (in local language) and then carry out some estimates concerning the restriction to GL_n of the "square" of an amplifier on GL_{n+1}.

Let F be a non-archimedean local field with ring of integers $\mathfrak{o}$, maximal ideal $\mathfrak{p}$, uniformizer $\varpi \in \mathfrak{p}$, and $q := \#\mathfrak{o}/\mathfrak{p}$, so that q is a power of some rational prime p.

5.2.1. *Preliminaries on the Hecke algebra.* Let n be a natural number. Set $G := \mathrm{GL}_n(F)$, $K := \mathrm{GL}_n(\mathfrak{o})$. We equip G with the Haar measure assigning volume one to K. The Hecke algebra $\mathcal{H}_G$ is the space of compactly-supported functions $t : K\backslash G/K \to \mathbb{C}$ equipped with the convolution product $*$. A basis for $\mathcal{H}_G$ is indexed by n-tuples of integers $a = (a_1, \dots, a_n) \in \mathbb{Z}^n$ modulo permutations, with the corresponding basis element

$$T(a) \in \mathcal{H}_G$$

given by the characteristic function of the double coset $K \operatorname{diag}(\varpi^{a_1}, \dots, \varpi^{a_n})K$. Rather than working modulo permutations, one can restrict to a satisfying the dominance condition $a_1 \geqslant \dots \geqslant a_n$. We introduce the abbreviation

$$T[j] := T((j, 0, \dots, 0))$$

(here the implicit integer n will always be clear from context).

For each Hecke algebra element $t \in \mathcal{H}_G$, we define the adjoint element $t^* \in \mathcal{H}_G$ by the formula $t^*(g) := \overline{t(g^{-1})}$. Then

$$T(a_1, \dots, a_n)^* = T(-a_n, \dots, -a_1).$$

A *Satake parameter* of G is an element of $(\mathbb{C}^\times)^n/S(n)$, i.e., a multiset $\alpha = \{\alpha_1, \dots, \alpha_n\}$ of nonzero complex numbers. We write π_α for the corresponding unramified induced representation of G, normalized so that π_α is tempered when each $|\alpha_j| = 1$. The space π_α^K of K-fixed vectors is one-dimensional, so $\mathcal{H}_G$ acts on it via a character (i.e., algebra homomorphism)

$$\lambda_\alpha : \mathcal{H}_G \to \mathbb{C},$$

i.e., for $v \in \pi_\alpha^K$ and $t \in \mathcal{H}_G$, we have $\pi_\alpha(t)v = \lambda_\alpha(t)v$. The Satake isomorphism says that $\alpha \to \lambda_\alpha$ defines a bijection between the set of Satake parameters of G and the set of characters of $\mathcal{H}_G$. We note in passing that every irreducible representation of G having a nontrivial K-fixed vector is isomorphic to some π_α.

Let α be a Satake parameter. For $\vartheta \geqslant 0$, we say that π_α is ϑ-*tempered* if

$$|\alpha_1 \cdots \alpha_n| = 1 \tag{5.4}$$

and

$$q^{-\vartheta} \leqslant |\alpha_j| \leqslant q^{\vartheta} \text{ for } j = 1, \dots, n. \tag{5.5}$$

The condition (5.4) says that π_α has unitary central character. We note that π_α is 0-tempered if and only if π_α is tempered in the usual sense.

5.2.2. *Construction and properties.* Set

$$G := \mathrm{GL}_{n+1}(F), \quad H := \mathrm{GL}_n(F),$$

$$K := \mathrm{GL}_{n+1}(\mathfrak{o}), \quad K_H := \mathrm{GL}_n(\mathfrak{o}),$$

with H embedded in G in the usual way as the upper-left block. Thus, as discussed in §3.4, (G, H) is a unitary GGP pair over F attached to the split quadratic extension $E = F \times F$. We equip G and H with measures as above, and define the corresponding Hecke algebras $\mathcal{H}_G$ and $\mathcal{H}_H$. As usual, we write Z for the center of G. We equip $Z \cong F^\times$ with the Haar measure dz assigning volume one to its maximal compact subgroup $\cong \mathfrak{o}^\times$.

Let ω be an unramified unitary character of Z. We define a linear map

$$\mathrm{res}_\omega : \mathcal{H}_G \to \mathcal{H}_H$$

by the formula: for $g \in H$,

$$(\mathrm{res}_\omega t)(g) := \int_{z \in Z} t(zg)\omega(z)\, dz.$$

Lemma 5.4. *Assume that q is sufficiently large and that n is fixed.*

The elements

$$t_j := q^{-\frac{nj}{2}} T[j] \in \mathcal{H}_G \quad (j \geqslant 0)$$

have the following properties.

(i) For each Satake parameter α of G for which π_α has unitary central character, we have

$$\max_{1 \leqslant j \leqslant n+1} |\lambda_\alpha(t_j)| \gg 1. \tag{5.6}$$

(ii) We have

$$t_j * t_j^* = \sum_{i=0}^{j} c_{ij} q^{-ni} T(i, 0, \dots, 0, -i) \tag{5.7}$$

for some coefficients $c_{ij} \ll_j 1$.

(iii) Fix $0 \leqslant \vartheta < 1/2$. For each unramified unitary character ω of Z and each Satake parameter α of H for which π_α is ϑ-tempered, we have

$$\lambda_\alpha(\mathrm{res}_\omega t_j) \ll_j q^{-(\frac{1}{2} - \vartheta)j} \tag{5.8}$$

and for $i \geqslant 0$

$$\lambda_\alpha(\mathrm{res}_\omega q^{-ni} T(i, 0, \dots, 0, -i)) \ll_i q^{-(1-2\vartheta))i} \leqslant 1. \tag{5.9}$$

In particular,

$$\lambda_\alpha(\mathrm{res}_\omega(t_j * t_j^*)) \ll_j 1. \tag{5.10}$$

Proof. The estimate (5.6) is essentially given by Blomer–Maga [BM1, Cor 4.3]. We outline their proof, noting the minor differences required here. Let $\Pi(n+1)$ denote the set of all $a = (a_1, \dots, a_{n+1}) \in \mathbb{Z}_{\geqslant 0}^{n+1}$ with $a_1 \geqslant \cdots \geqslant a_{n+1}$. It is shown in [BM1, Lem 4.2] that for some $y_a \in \mathbb{Q}$ ($a \in \Pi(n+1)$) with $|y_a| \ll 1$, we have for large enough q the identity in $\mathcal{H}_G$

$$q^n \sum_{a \in \Pi(n+1)} y_a \prod_{j=1}^{n+1} T[a_j] = q^{\frac{(n+1)(n+2)}{2}} T(1, \dots, 1).$$

Strictly speaking, the cited reference addresses the case $F = \mathbb{Q}_p$ in classical language, but the same argument applies in our setting. Since π_α is assumed to have unitary central character, we have $|\lambda_\alpha(T(1,\dots,1))| = 1$. The stronger condition $\lambda_\alpha(T(1,\dots,1)) = 1$ is assumed in [BM1, §4], but the weaker condition $|\lambda_\alpha(T(1,\dots,1))| = 1$ suffices to apply the proof of [BM1, Lem 4.3] to see that at least one of the quantities $|\lambda_\alpha(t_j)|$ is bounded from below by $(|\Pi(n+1)| \max_{a \in \Pi(n+1)} |y_a|)^{-1} \gg 1$, as required.

The identity (5.7) is again due to Blomer–Maga [BM1, Lemma 4.4] (up to the cosmetic substitution $i \mapsto j - i$).

The proofs of (5.8) and (5.9) require a couple preliminaries to which we now turn.

Let $a = (a_1, \dots, a_{n+1}) \in \mathbb{Z}^{n+1}$ and $b \in \mathbb{Z}$. We observe that for $z = \varpi^{-b} \in Z \cong F^\times$, the double coset

$$zK \operatorname{diag}(\varpi^{a_1}, \dots, \varpi^{a_n})K = K \operatorname{diag}(\varpi^{a_1 - b}, \dots, \varpi^{a_n - b})K$$

has nontrivial intersection with H if and only if $b = a_j$ for some j, in which case that intersection is the double coset

$$K_H \operatorname{diag}(\varpi^{a_1 - b}, \dots, \varpi^{a_{j-1} - b}, \varpi^{a_{j+1} - b}, \dots, \varpi^{a_n - b})K_H.$$

We deduce the identity

$$\operatorname{res}_\omega T(a) = \sum_{j=1,\dots,n+1}^{\sharp} \omega(\varpi^{-a_j}) T(a_1 - a_j, \dots, a_{j-1} - a_j, a_{j+1} - a_j, \dots, a_{n+1}), \tag{5.11}$$

where $\sharp$ signifies that the sum runs over indices j such that the a_j run over the distinct elements of the multiset $\alpha = \{\alpha_1, \dots, \alpha_n\}$.

We pause to record the estimate: for any Satake parameter α of H for which π_α is ϑ-tempered, any integers $a_1 \geqslant \dots \geqslant a_n$ and any fixed $\varepsilon > 0$, we have

$$\lambda_\alpha(T(a)) \ll e^{\varepsilon \sum_{j=1}^n |a_j|} q^{-v(a) + \frac{n+1}{2}|a|} \max_{\sigma \in W} |\alpha_{\sigma(1)}^{a_1} \cdots \alpha_{\sigma(n)}^{a_n}| \tag{5.12}$$

where $v(a) := \sum_{k=1}^n k a_k$. This is likely well-known (in sharper forms), but since we could not quickly locate a suitable reference, we sketch a proof. We use the formula (see [BM1, (4.3)] or [An, (1.7)])

$$\lambda_\alpha(T(a_1, \dots, a_n)) = q^{-v(a) + \frac{n+1}{2}|a|} P_a(\alpha), \tag{5.13}$$

where

$$P_a(\alpha) = c(a) \sum_{\sigma \in S(n)} \alpha_{\sigma(1)}^{a_1} \cdots \alpha_{\sigma(n)}^{a_n} \prod_{i > j} \frac{1 - q^{-1} \alpha_{\sigma(j)} / \alpha_{\sigma(i)}}{1 - \alpha_{\sigma(j)} / \alpha_{\sigma(i)}} \tag{5.14}$$

for some leading constant $c(a) \asymp 1$. The ϑ-temperedness assumption implies $\alpha_j / \alpha_i \ll q^{2\vartheta} \ll q$, so the numerators of the fractions in (5.14) are $\mathrm{O}(1)$. The only issue in deducing (5.12) is that the denominator has singularities along the hyperplanes $\alpha_i = \alpha_j$ $(i \neq j)$. For $\delta > 0$, let $\mathcal{D}(\delta)$ denote the set of all Satake parameters α that avoid these hyperplanes in the sense that $|1 - \alpha_i / \alpha_j| > \delta$. For $\alpha \in \mathcal{D}(\delta)$, the denominators in (5.14) are bounded away from zero, so the required estimate (5.12) holds but with the implied constant depending upon δ. We deduce (5.12) in general using Cauchy's theorem, as follows. For each α, define a function $f_\alpha : \mathbb{C}^{n-1} \to \mathbb{C}$ by $f_\alpha(s) := \lambda_{\alpha(s)}(T(a))$, where $\alpha(s) := \{\alpha_1 e^{s_1}, \dots, \alpha_n e^{s_n}\}$ and $s_n := -(s_1 + \dots + s_{n-1})$. By a pigeonhole argument, we see that for each $r > 0$ there exists $r_0 > 0$ and $\delta > 0$ (independent of the local field F) so that for all α, there exists tuple of radii $\rho = (\rho_1, \dots, \rho_{n-1}) \in (r_0, r)^{n-1}$ for which $\alpha(s)$ belongs to $\mathcal{D}(\delta)$ whenever s lies in the polycircle $\mathcal{C}(\rho) := \{s : |s_j| = \rho_j\}$. By Cauchy's theorem, it follows that

$$\lambda_\alpha(T(a)) = f(0) \ll (\rho_1 \cdots \rho_{n-1})^{-1} \sup_{s \in \mathcal{C}(\rho)} |\lambda_{\alpha(s)}(T(a))|. \tag{5.15}$$

We now apply the bound (5.12) to the RHS of (5.15) and observe that

$$|\alpha(s)_{\sigma(1)}^{a_1} \cdots \alpha(s)_{\sigma(n)}^{a_n}| \leqslant e^{r \sum_{j=1}^n |a_j|} |\alpha_{\sigma(1)}^{a_1} \cdots \alpha_{\sigma(n)}^{a_n}|.$$

We deduce (5.13) by fixing r sufficiently small.

Below, we will frequently apply the identity

$$\lambda_\alpha(T(a_1 + b, \dots, a_n + b)) = (\alpha_1 \cdots \alpha_n)^b \lambda_\alpha(T(a_1, \dots, a_n))$$

and the hypothesis (5.4).

We now establish (5.8). We observe first using (5.11) that

$$\operatorname{res}_\omega T[j] = \omega(\varpi^{-j}) T(0, -j, \dots, -j) + T(j, 0, \dots, 0). \tag{5.16}$$

On the other hand, we verify readily using (5.12) that

$$\lambda_\alpha(T(j, 0, \dots, 0)) \ll_j q^{\frac{nj}{2} - (\frac{1}{2} - \vartheta) j}$$

and

$$\lambda_\alpha(T(0, -j, \dots, -j)) \asymp \lambda_\alpha(T(j, 0, \dots, 0)) \ll_j q^{\frac{nj}{2} - (\frac{1}{2} - \vartheta) j}.$$

This completes the proof of (5.8).

We now verify (5.9). For $i = 0$, we have the adequate estimate

$$\lambda_\alpha(\operatorname{res}_\omega(T(0, \dots, 0))) = \lambda_\alpha(T(0, \dots, 0)) = 1 \ll 1.$$

For $0 < i \leqslant j$, we have by (5.11) that

$$\begin{aligned} \operatorname{res}_\omega T(i, 0, \dots, 0, -i) = {} & \omega(\varpi^{-i}) T(-i, \dots, -i, -2i) \\ & + T(i, 0, \dots, 0, i) \\ & + \omega(\varpi^i) T(2i, i, \dots, i). \end{aligned} \tag{5.17}$$

By (5.12), we have

$$\lambda_\alpha(T(-i, \dots, -i, -2i)) \asymp \lambda_\alpha(T(0, \dots, 0, -i)) \ll_i q^{i(\frac{n}{2} - (\frac{1}{2} - \vartheta))},$$

$$\lambda_\alpha(T(i, 0, \dots, 0, -i)) \ll_i q^{i(n - (1 - 2\vartheta))}$$

and

$$\lambda_\alpha(T(2i, i, \dots, i)) \asymp \lambda_\alpha(T(i, \dots, 0, 0)) \ll_i q^{i(\frac{n}{2} - (\frac{1}{2} - \vartheta))}.$$

These identities and estimates combine to give (5.9). □

5.3. Pretrace inequality. Let (G, H) be a GGP pair over a number field F. As usual, we write Z for the center of G and $\bar{G} = G/Z$ for the adjoint group. Let π be a cuspidal automorphic representation of $G(\mathbb{A})$ with unitary central character $\omega : Z(\mathbb{A}) \to \mathrm{U}(1)$. We equip π with the Petersson inner product defined by integrating over $[\bar{G}]$. We write $\mathcal{B}(\pi)$ for an orthonormal basis consisting of vectors isotypic under some chosen maximal compact subgroup of $G(\mathbb{A})$.

For $f_1, f_2 \in C_c^\infty(G(\mathbb{A}))$, we write $f_1 * f_2$ for their convolution. For $f \in C_c^\infty(G(\mathbb{A}))$, we set $f^*(g) := \overline{f(g^{-1})}$ and $f^\omega(g) := \int_{z \in Z(\mathbb{A})} \omega(z) f(zg) \, dz$.

Lemma 5.5. *Let $f \in C_c^\infty(G(\mathbb{A}))$. Let $\Psi : [H] \to \mathbb{C}$ be a measurable function of rapid decay. Then*

$$\left| \sum_{\varphi \in \mathcal{B}(\pi)} \int_{[H]} \pi(f)\varphi \cdot \bar{\Psi} \right|^2 \leqslant \int_{x,y \in [H]} \bar{\Psi}(x)\Psi(y) \sum_{\gamma \in \bar{G}(F)} (f * f^*)^\omega (x^{-1}\gamma y)\, dx\, dy,$$

where ω is the central character of π. Each of the above iterated sums/integrals converges absolutely.

Proof. Morally, we may conclude via the pretrace formula, which says that the RHS of the required inequality is the integral of the LHS over all π occurring in the spectral decomposition of $L^2([G], \omega)$ – including the continuous spectrum, as we have not assumed that $[G]$ is compact. While such an argument could likely be implemented (following a preliminary discussion on spectral decomposition), it seems simpler to verify the required conclusion directly, as follows. Using the rapid decay of Ψ and the smoothness of f, we see that $\varphi \mapsto \int_{[H]} \pi(f)\varphi \cdot \bar{\Psi}$ defines a bounded linear functional on the Hilbert space π. We may represent that functional by a unique vector $\varphi_0 \in \pi$. By Parseval, we see that $\langle \varphi_0, \varphi_0 \rangle$ expands to the LHS of the required inequality, so our task is to bound $\langle \varphi_0, \varphi_0 \rangle$ by the RHS. By combining the invariance of f under an open subgroup of the finite adelic points of G with an application of the Dixmier–Malliavin lemma to the archimedean points of G, we see that φ_0 is smooth; since π is cuspidal, it follows that φ_0 is of rapid decay (a feature which simplifies the details of the proof, but is not ultimately necessary). By construction, $\langle \varphi_0, \varphi_0 \rangle = \int_{[H]} \pi(f)\varphi_0 \cdot \bar{\Psi}$. By folding up this last integral, we obtain

$$\langle \varphi_0, \varphi_0 \rangle = \int_{x \in [H]} \bar{\Psi}(x) \int_{y \in [\bar{G}]} \varphi_0(y) \sum_{\gamma \in \bar{G}(F)} f^\omega(x^{-1}\gamma y).$$

Using the rapid decay of Ψ and φ_0 and crudely bounding the sum over γ, we see that this iterated integral/sum converges absolutely. We may thus rearrange it and apply Cauchy–Schwarz to obtain

$$\langle \varphi_0, \varphi_0 \rangle \leqslant \int_{y \in [\bar{G}]} \left| \int_{x \in [H]} \bar{\Psi}(x) \sum_{\gamma \in \bar{G}(F)} f^\omega(x^{-1}\gamma y)\, dx \right|^2 dy.$$

Opening the square, executing the y-integral and unfolding, we readily obtain the RHS of the required inequality. □

6. Reduction of the proof of the main theorem

In this section we prove Theorem 1.1, assuming Theorem 4.2.

6.1. Setting. We recall the setting of §1.3. Let F be a fixed number field. Let (G, H) be a fixed unitary GGP pair over F with standard representation $G \hookrightarrow \mathrm{GL}(V/E)$. We choose a distinguished archimedean place $\mathfrak{q}$ of F. We recall from §1.3 the following further assumptions:

- H is anisotropic, so that $[H]$ is compact.
- $G(F_\mathfrak{p})$ and $H(F_\mathfrak{p})$ are compact for all non-archimedean places $\mathfrak{p}$ other than the distinguished archimedean place $\mathfrak{q}$.

We fix a finite set S of places of F, taken large enough in the sense of §3.6.

As usual, we write Z for the center of G and $\bar{G} = G/Z$. We write $n+1$ for the E-dimension of the hermitian space V, so that (G, H) is a form of $(\mathrm{U}_{n+1}, \mathrm{U}_n)$. We denote by ε some fixed positive quantity, not necessarily the same in each occurrence.

Recall the families $\mathcal{F}_T$ defined in §1.3. Let T be a large positive real, and let $(\pi, \sigma) \in \mathcal{F}_T$. We assume that σ is ϑ-tempered for some fixed $\vartheta \in [0, 1/2)$. To prove Theorem 1.1, we must show that

$$\mathcal{L}(\pi, \sigma) \ll T^{n(n+1)/2-\delta} \tag{6.1}$$

for suitable fixed $\delta > 0$.

6.2. Reduction to bounds for global periods. Recall from §3.7 that $\sigma_S \subseteq \sigma$ denotes the unramified-outside-S subspace. We choose an isometric factorization $\sigma_S = \otimes_{\mathfrak{p} \in S} \sigma_{\mathfrak{p}}$. Let $\Psi = \otimes_{\mathfrak{p} \in S} \Psi_{\mathfrak{p}} \in \sigma_S$ be a factorizable vector. Let $f_0 = f_{0,S} \otimes f_0^S \in C_c^\infty(G(\mathbb{A}))$, with $f_{0,S} = \otimes_{\mathfrak{p} \in S} f_{0,\mathfrak{p}}$ and $f_0^S = \otimes_{\mathfrak{p} \notin S} 1_{G(\mathbb{Z}_{\mathfrak{p}})}$, be a factorizable test function, unramified outside S. The period formula (3.2) implies that

$$\sum_{\varphi \in \mathcal{B}(\pi)} \left| \int_{[H]} \pi(f_0)\varphi \cdot \overline{\Psi} \right|^2 = \mathcal{L}(\pi, \sigma) \prod_{\mathfrak{p} \in S} I_{\mathfrak{p}},$$

where $\mathcal{B}(\cdot)$ is an orthonormal basis chosen as in §5.1 and

$$I_{\mathfrak{p}} := \sum_{v \in \mathcal{B}(\pi_{0,\mathfrak{p}})} \mathcal{Q}_{\mathfrak{p}}(\pi(f_{0,\mathfrak{p}})v \otimes \Psi_{\mathfrak{p}}),$$

with $\mathcal{Q}_{\mathfrak{p}}$ the quadratic form defined by integration over $H(F_{\mathfrak{p}})$ as in §3.5.

For $\mathfrak{p} \in S - \{\mathfrak{q}\}$, we fix $f_{0,\mathfrak{p}} \in C_c^\infty(G(F_{\mathfrak{p}}))$ as follows.

- If $\mathfrak{p}$ is archimedean, we take for $f_{0,\mathfrak{p}}$ an "approximate identity:" we require that $\int_{G(F_{\mathfrak{p}})} f_{0,\mathfrak{p}} = 1$ and that $f_{0,\mathfrak{p}}$ be supported in a small enough neighborhood of the identity element of the compact Lie group $G(F_{\mathfrak{p}})$.
- If $\mathfrak{p}$ is non-archimedean, we take for $f_{0,\mathfrak{p}}$ the normalized characteristic function of some small enough compact open subgroup $U_{\mathfrak{p}}$ of the p-adic group $G(F_{\mathfrak{p}})$,

Using Lemmas 5.1 and 5.2, we may find for each $\mathfrak{p} \in S - \{\mathfrak{q}\}$ a unit vector $\Psi_{\mathfrak{p}} \in \sigma_{\mathfrak{p}}$ so that $I_{\mathfrak{p}} \gg 1$.

At the distinguished place $\mathfrak{q}$, we choose $f_{0,\mathfrak{q}}$ and $\Psi_{\mathfrak{q}} \in \sigma_{\mathfrak{q}}$ according to Theorem 4.2, applied to $(\pi_{0,\mathfrak{q}}, \sigma_{\mathfrak{q}})$ and with the exponent κ taken sufficiently small (in particular, $\kappa \leqslant \varepsilon$). We may and shall assume that $f_{0,\mathfrak{q}}$ is supported sufficiently close to the identity element.

With these choices, we obtain

$$\prod_{\mathfrak{p} \in S} I_{\mathfrak{p}} \gg T^{-n^2/2-\varepsilon}.$$

Our task (6.1) thereby reduces to showing that

$$\sum_{\varphi \in \mathcal{B}(\pi)} \left| \int_{[H]} \pi(f_0)\varphi \cdot \overline{\Psi} \right|^2 \ll T^{n/2-\delta}. \tag{6.2}$$

6.3. Application of amplified pretrace inequality. Let L be a positive parameter with $\log L \asymp \log T$. We eventually take for L a small positive power of T (see (6.23) below).

Let $\mathbb{P}$ denote the set of finite primes $\mathfrak{p}$ of F with the following properties:

- The absolute norm $q_{\mathfrak{p}}$ of $\mathfrak{p}$ lies in the interval $[L, 2L]$.
- $\mathfrak{p}$ splits the field extension E/F, so that $E_{\mathfrak{p}} \cong F_{\mathfrak{p}} \times F_{\mathfrak{p}}$.

By a weak form of the prime number theorem in arithmetic progressions for F, we have

$$T^{-\varepsilon} L \ll |\mathbb{P}| \ll L. \tag{6.3}$$

For $\mathfrak{p} \notin S$, we write $\mathcal{H}_{G(F_{\mathfrak{p}})}$ for the corresponding spherical Hecke algebra with respect to $G(\mathbb{Z}_{\mathfrak{p}})$, and $\mathcal{H}_G^S := \otimes'_{\mathfrak{p} \notin S} \mathcal{H}_{G(F_{\mathfrak{p}})}$ for the restricted tensor product of these algebras.

For $\mathfrak{p} \in \mathbb{P}$, we have $G(F_{\mathfrak{p}}) \cong \mathrm{GL}_n(F_{\mathfrak{p}})$ (see §3.3). We write $T_{\mathfrak{p}}(a)$ $(a \in \mathbb{Z}^n)$ and $t_{\mathfrak{p},j}$ $(j \geqslant 0)$ for the elements of $\mathcal{H}_{G(F_{\mathfrak{p}})}$ defined in §5.2. By abuse of notation, we identify $t_{\mathfrak{p},j}$ with its image in $\mathcal{H}_G^S$.

We write λ_π for the character of $\mathcal{H}_G^S$ attached to π. By (5.6), we have $\sum_{j=1}^{n+1} |\lambda_\pi(t_{\mathfrak{p},j})| \gg 1$ for each $\mathfrak{p} \in \mathbb{P}$. By the pigeonhole principle, we may thus find $j \in \{1, \dots, n+1\}$ so that

$$\sum_{\mathfrak{p} \in \mathbb{P}} |\lambda_\pi(t_{\mathfrak{p},j})| \gg |\mathbb{P}|.$$

For $\mathfrak{p} \in \mathbb{P}$, we set $x_{\mathfrak{p}} := \overline{\mathrm{sgn}(\lambda_\pi(t_{\mathfrak{p},j}))}$, where $\mathrm{sgn}(0) := 0$ and $\mathrm{sgn}(z) := z/|z|$ for $z \neq 0$. In particular, $|x_{\mathfrak{p}}| \leqslant 1$. For each $\varphi \in \pi$, we then have

$$\pi \left(f_{0,S} \otimes \sum_{\mathfrak{p} \in \mathbb{P}} x_{\mathfrak{p}} t_{\mathfrak{p},j} \right) \varphi = \left(\sum_{\mathfrak{p} \in \mathbb{P}} |\lambda_\pi(t_{\mathfrak{p},j})| \right) \pi(f_0) \varphi.$$

Setting

$$\mathcal{R}(j) := \sum_{\varphi \in \mathcal{B}(\pi)} \left| \int_{[H]} \pi \left(f_{0,S} \otimes \sum_{\mathfrak{p} \in \mathbb{P}} x_{\mathfrak{p}} t_{\mathfrak{p},j} \right) \varphi \cdot \bar{\Psi} \right|^2,$$

the proof of (6.2) reduces to that of the estimate $\mathcal{R}(j) \ll T^{n/2-\delta} L^2$.

Let ω denote the central character of π. By lemma 5.5, we have $\mathcal{R}(j) \leqslant \sum_{\mathfrak{p}_1, \mathfrak{p}_2 \in \mathbb{P}} x_{\mathfrak{p}_1} \bar{x}_{\mathfrak{p}_2} \mathcal{R}(\mathfrak{p}_1, \mathfrak{p}_2, j)$, where

$$\mathcal{R}(\mathfrak{p}_1, \mathfrak{p}_2, j) := \int_{x, y \in [H]} \bar{\Psi}(x) \Psi(y) \sum_{\gamma \in \bar{G}(F)} f^{(\mathfrak{p}_1, \mathfrak{p}_2, j), \omega}(x^{-1} \gamma y)$$

with $f^{(\mathfrak{p}_1, \mathfrak{p}_2, j)} := (f_{0,S} \otimes t_{\mathfrak{p}_1,j}) * (f_{0,S} \otimes t_{\mathfrak{p}_2,j})^*$. In the special case $\mathfrak{p}_1 = \mathfrak{p}_2$, we recall from (5.7) that

$$t_{\mathfrak{p}_1,j} * t_{\mathfrak{p}_1,j}^* = \sum_{i=0}^{j} c_{ij} q_{\mathfrak{p}_1}^{-ni} T_{\mathfrak{p}_1}(i, 0, \dots, 0, -i)$$

with $c_{ij} \ll 1$. This yields a decomposition $\mathcal{R}(\mathfrak{p}_1, \mathfrak{p}_1, j) = \sum_{i=0}^{j} c_{ij} \mathcal{R}'(\mathfrak{p}_1, i)$ indexed by $0 \leqslant i \leqslant j \leqslant n+1$. Since $|x_{\mathfrak{p}}| \leqslant 1$, our task reduces to verifying that

$$\sum_{\mathfrak{p}_1 \neq \mathfrak{p}_2 \in \mathbb{P}} \sum_{j=1}^{n+1} |\mathcal{R}(\mathfrak{p}_1, \mathfrak{p}_2, j)| + \sum_{\mathfrak{p}_1 = \mathfrak{p}_2 \in \mathbb{P}} \sum_{j=0}^{n+1} |\mathcal{R}'(\mathfrak{p}_1, j)| \ll T^{n/2-\delta} L^2. \tag{6.4}$$

We henceforth focus on an individual term on the LHS of (6.4), corresponding to some specific values of $j, \mathfrak{p}_1, \mathfrak{p}_2$. We define $f = \otimes f_{\mathfrak{p}} \in C_c^\infty(G(\mathbb{A}))$, as follows.

- For $\mathfrak{p} \in S$, we set $f_{\mathfrak{p}} := f_{0,\mathfrak{p}} * f_{0,\mathfrak{p}}^*$.
- For $\mathfrak{p} \notin S \cup \{\mathfrak{p}_1, \mathfrak{p}_2\}$, we take $f_{\mathfrak{p}} := 1_{G(\mathcal{O}_{\mathfrak{p}})}$.
- For $\mathfrak{p}_1 \neq \mathfrak{p}_2$, we set
$$f_{\mathfrak{p}_1} = t_{\mathfrak{p}_1} = q_{\mathfrak{p}_1}^{-\frac{nj}{2}} T_{\mathfrak{p}_1}(j, 0, \dots, 0),$$
$$f_{\mathfrak{p}_2} = t_{\mathfrak{p}_2}^* = q_{\mathfrak{p}_2}^{-\frac{nj}{2}} T_{\mathfrak{p}_2}(0, \dots, 0, -j).$$
- For $\mathfrak{p}_1 = \mathfrak{p}_2$, we set
$$f_{\mathfrak{p}_1} = q_{\mathfrak{p}_1}^{-nj} T_{\mathfrak{p}}(j, 0, \dots, 0, -j).$$

We set
$$f^\sharp := \otimes f_{\mathfrak{p}}^\sharp, \quad f_{\mathfrak{p}}^\sharp(g) := \int_{Z_{\mathfrak{p}}} \omega_{\mathfrak{p}}(z) f_{\mathfrak{p}}(zg) \, dz,$$
so that $f^\sharp = f^\omega$ in the notation of §5.3. Then for $\mathfrak{p}_1 \neq \mathfrak{p}_2$ (resp. $\mathfrak{p}_1 = \mathfrak{p}_2$), the quantity $\mathcal{R}(\mathfrak{p}_1, \mathfrak{p}_2, j)$ (resp. $\mathcal{R}'(\mathfrak{p}_1, j)$) is given by
$$\mathcal{R}(f) := \int_{x,y \in [H]} \overline{\Psi}(x) \Psi(y) \sum_{\gamma \in \bar{G}(F)} f^\sharp(x^{-1}\gamma y) \, dx \, dy = \mathcal{M} + \mathcal{E}, \tag{6.5}$$
where $\mathcal{M}$ denotes the contribution from $\gamma \in H(F) \hookrightarrow \bar{G}(F)$ and $\mathcal{E}$ denotes the remaining contribution.

We note for future reference that
$$\prod_{\mathfrak{p} \notin S} \|f_{\mathfrak{p}}^\sharp\|_\infty \ll L^{-nj}, \tag{6.6}$$
which follows from the definition of $f_{\mathfrak{p}}$ for $\mathfrak{p} \in \{\mathfrak{p}_1, \mathfrak{p}_2\}$.

We estimate $\mathcal{M}$ in §6.4 and $\mathcal{E}$ in §6.5. We then combine these estimates and sum them in §6.6 to deduce the required bound (6.4).

6.4. Main term estimates. Since the quotient $[H]$ is compact, the Petersson inner product on σ (defined in general by integration over $[H/Z_H]$) may be written $\langle \Psi_1, \Psi_2 \rangle = c \int_{[H]} \Psi_1 \bar{\Psi}_2$ for some fixed $c > 0$. To simplify notation, let us renormalize the inner product on σ so that this last identity holds with $c = 1$; doing so has no effect on the estimates to be proved. We may then unfold $\mathcal{M}$ to
$$\begin{aligned} \mathcal{M} &= \int_{x \in [H]} \int_{y \in H(\mathbb{A})} \bar{\Psi}(x) \Psi(y) f^\sharp(x^{-1}y) \, dx \, dy \\ &= \int_{g \in H(\mathbb{A})} f^\sharp(g) \langle g\Psi, \Psi \rangle \, dg \\ &= \prod_{\mathfrak{p} \in S \cup \{\mathfrak{p}_1, \mathfrak{p}_2\}} I_{\mathfrak{p}}, \quad I_{\mathfrak{p}} := \int_{g \in H(F_{\mathfrak{p}})} f_{\mathfrak{p}}^\sharp(g) \langle g\Psi_{\mathfrak{p}}, \Psi_{\mathfrak{p}} \rangle \, dg. \end{aligned}$$
We estimate each of these local integrals separately:

- Since $\Psi_{\mathfrak{q}}$ is a unit vector, we may bound $I_{\mathfrak{q}}$ by $\int_{H(F_{\mathfrak{q}})} |f_{\mathfrak{q}}^\sharp|$, which is $\ll T^{n/2+\varepsilon}$ thanks to part (ii) of Theorem 4.2.
- For $\mathfrak{p} \in S - \{\mathfrak{q}\}$, we have $\int_{H(F_{\mathfrak{p}})} |f_{\mathfrak{p}}^\sharp| \ll 1$ and $\|\Psi_{\mathfrak{p}}\| = 1$, so $I_{\mathfrak{p}} \ll 1$.
- In the case $\mathfrak{p}_1 \neq \mathfrak{p}_2$, we see from (5.8) that $I_{\mathfrak{p}_j} \ll q_{\mathfrak{p}_j}^{-(1/2-\vartheta)j} \ll L^{-(1/2-\vartheta)j}$ for $j = 1, 2$.

• In the case $\mathfrak{p}_1 = \mathfrak{p}_2$, we see from (5.9) that $I_{\mathfrak{p}_1} \ll q_{\mathfrak{p}_1}^{-(1-2\vartheta)j} \ll L^{-(1-2\vartheta)j}$.

Therefore in all cases,

$$\mathcal{M} \ll T^{n/2+\varepsilon} L^{-(1-2\vartheta)j}.$$

6.5. Error term estimates. The estimation of $\mathcal{E}$ consists of a few steps.

6.5.1. *Preliminary reductions.* We observe that the integrand in (6.5) is invariant in both variables under $\prod_{\mathfrak{p} \notin S} H(\mathbb{Z}_{\mathfrak{p}})$. By our hypotheses concerning S (see §3.6) and strong approximation, we may thus replace the integrals over $[H]$ with integrals over the S-arithmetic quotient $\Gamma \backslash H_S$. This quotient is compact thanks to our hypothesis that H is anisotropic. We fix a nice compact fundamental domain $\mathcal{D}$ for this quotient. We fix a compact factorizable neighborhood $\Omega_S = \prod_{\mathfrak{p} \in S} \Omega_{\mathfrak{p}}$ of the identity element of H_S with $\mathcal{D} \subseteq \Omega_S$. Then by our normalization of measures,

$$\mathcal{E} \leqslant \int_{x,y \in \Omega_S} \left| \Psi(x)\Psi(y) \sum_{\gamma \in \bar{G}(F) - H(F)} f^{\sharp}(x^{-1}\gamma y) \right| dx\, dy.$$

Setting

$$\Sigma := \{\gamma \in \bar{G}(F) - H(F) : f^{\sharp}(x^{-1}\gamma y) \neq 0 \text{ for some } x, y \in \Omega_S\},$$

we obtain

$$\mathcal{E} \leqslant |\Sigma| \max_{\gamma \in \Sigma} I(\gamma), \quad I(\gamma) := \int_{x,y \in \Omega_S} \left| \Psi(x)\Psi(y) \sum_{\gamma \in \bar{G}(F) - H(F)} f^{\sharp}(x^{-1}\gamma y) \right| dx\, dy.$$

We estimate Σ in §6.5.2 and $I(\gamma)$ in §6.5.3. The resulting bound for $\mathcal{E}$ is then recorded in §6.5.4.

In what follows, we write

$$d_{H_{\mathfrak{q}}} : \bar{G}(F_{\mathfrak{q}}) \to [0, 1]$$

for the function attached in §4.2 to the GGP pair $(G(F_{\mathfrak{q}}), H(F_{\mathfrak{q}}))$.

6.5.2. *Bounds for Σ.*

Lemma 6.1. *We have*

$$|\Sigma| \ll L^{2(n+1)^2 j} \tag{6.7}$$

Moreover, for each $\gamma \in \Sigma$, we have

$$d_{H_{\mathfrak{q}}}(\gamma) \gg L^{-j} \tag{6.8}$$

Remark 6.2. The estimate (6.7) is likely far from optimal (see the sentence following (6.21) for details). The factor of 2 in the exponent could likely be removed with a bit more work, and the exponent could likely be reduced further for large n by counting more carefully.

Remark 6.3. The content of Lemma 6.1 is best illustrated by passing to split unitary groups, i.e., general linear groups, where it amounts in the simplest case to (a weakened form of) the following easy exercise:

For a natural number ℓ, the number of matrices $\gamma \in \mathrm{GL}_{n+1}(\mathbb{Z})$ for which $\det(\gamma) = \ell^{n+1}$ and $|\ell^{-1}\gamma - 1| \leqslant \varepsilon$ is $\mathrm{O}(\ell^{(n+1)^2})$. Moreover, if such a matrix does not lie in $\mathrm{GL}_n(\mathbb{R})$ modulo the center, then for some $j \in \{1, \ldots, n\}$, one of the entries $\gamma_{j,n+1}$ or $\gamma_{n+1,j}$ is a nonzero integer, hence $\gg 1$, while the entry $\gamma_{n+1,n+1}$ is $\asymp \ell$.

The proof of Lemma 6.1 is mildly more complicated due to the overhead coming from working S-arithmetically over a number field, but requires no significant new ideas. The reader might thus wish to skip ahead to §6.5.3 on a first reading.

The proof of Lemma 6.1 invokes the following preliminary result. We write $\mathfrak{P}$ for a place of E and, if $\mathfrak{P}$ is finite, $\mathcal{O}_{\mathfrak{P}}$ for the ring of integers of the completion $E_{\mathfrak{P}}$.

Lemma 6.4. *Fix a finite set S_E of places of the number field E, containing every archimedean place. For $\mathfrak{P} \in S_E$, suppose given a fixed compact neighborhood $\Omega_{\mathfrak{P}}$ of the identity element in $\mathrm{PGL}_{n+1}(E_{\mathfrak{P}})$. For $\mathfrak{P} \notin S_E$, suppose given an element $D_{\mathfrak{P}} \geqslant 1$ of the value group of $E_{\mathfrak{P}}^{\times}$, with $D_{\mathfrak{P}} = 1$ for almost all $\mathfrak{P}$, and set*

$$\Omega_{\mathfrak{P}} := \text{ image in } \mathrm{PGL}_{n+1}(E_{\mathfrak{P}}) \text{ of } \{g \in M_{n+1}(\mathcal{O}_{\mathfrak{P}}) : |\det g|_{\mathfrak{P}} = 1/D_{\mathfrak{P}}\},$$

so that $\Omega_{\mathfrak{P}} = \mathrm{PGL}_{n+1}(\mathcal{O}_{\mathfrak{P}})$ for almost all $\mathfrak{P}$. Set

$$\Omega := \prod_{\mathfrak{P}} \Omega_{\mathfrak{P}} \subseteq \mathrm{PGL}_{n+1}(\mathbb{A}_E), \quad D := \prod_{\mathfrak{P} \notin S_E} D_{\mathfrak{P}}.$$

(i) We have

$$|\mathrm{PGL}_{n+1}(E) \cap \Omega| \ll D^{n+1}.$$

(ii) Suppose moreover that the following conditions holds.

- *For each $\mathfrak{P} \in S_E$, the identity neighborhood $\Omega_{\mathfrak{P}}$ is of the form*

$$\Omega_{\mathfrak{P}} = \Omega_{\mathfrak{P}}^1 \Omega_{\mathfrak{P}}^0 \Omega_{\mathfrak{P}}^1, \tag{6.9}$$

  where
  - *$\Omega_{\mathfrak{P}}^1$ is a fixed compact neighborhood of the identity element in $\mathrm{GL}_n(E_{\mathfrak{P}})$, and*
  - *$\Omega_{\mathfrak{P}}^0$ is a fixed sufficiently small compact neighborhood of the identity element in $\mathrm{PGL}_{n+1}(E_{\mathfrak{P}})$.*
- *The set of places S_E is large enough that $\mathcal{O}[1/S_E]$ is a principal ideal domain.*

Let $\gamma \in \mathrm{PGL}_{n+1}(E) \cap \Omega$. Write

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{6.10}$$

with a, b, c, d matrices of respective dimensions $n \times n, n \times 1, 1 \times n, 1 \times 1$, well-defined up to simultaneous scaling. Then $d \neq 0$. Moreover, for each subset T of S_E, we have

$$b \neq 0 \implies \prod_{\mathfrak{P} \in T} |b/d|_{E_{\mathfrak{P}}} \gg D^{-\frac{1}{n+1}}, \tag{6.11}$$

$$c \neq 0 \implies \prod_{\mathfrak{P} \in T} |c/d|_{E_{\mathfrak{P}}} \gg D^{-\frac{1}{n+1}}. \tag{6.12}$$

where $|.|_{E_{\mathfrak{P}}}$ denotes any fixed $E_{\mathfrak{P}}$-norm.

Proof. We start with (i). We fix a Haar measure dz on $\mathbb{A}_E^{\times} \hookrightarrow \mathrm{GL}_{n+1}(\mathbb{A}_E)$ and a factorization $dz = \prod dz_{\mathfrak{P}}$ so that $dz_{\mathfrak{P}}$ assigns volume one to the unit group for all finite primes $\mathfrak{P}$. For each $\mathfrak{P} \in S_E$, we may find a small neighborhood $\tilde{\Omega}_{\mathfrak{P}}$ of the identity element of $\mathrm{GL}_{n+1}(E_{\mathfrak{P}})$ so that for all $g \in \mathrm{GL}_{n+1}(E_{\mathfrak{P}})$ with image $[g] \in \mathrm{PGL}_{n+1}(E_{\mathfrak{P}})$, we have

$$1_{\Omega_{\mathfrak{P}}}([g]) \ll \int_{z \in E_{\mathfrak{P}}^{\times}} 1_{\tilde{\Omega}_{\mathfrak{P}}}(zg)\, dz. \tag{6.13}$$

For each $\mathfrak{P} \notin S_E$, we set

$$\tilde{\Omega}_{\mathfrak{P}} := \{g \in M_{n+1}(\mathcal{O}_{\mathfrak{P}}) : |\det g|_{\mathfrak{P}} = 1/D_{\mathfrak{P}}\},$$

so that (6.13) holds. Setting $\tilde{\Omega} := \prod \tilde{\Omega}_{\mathfrak{P}} \subseteq \mathrm{GL}_{n+1}(\mathbb{A}_E)$, we obtain

$$\begin{aligned} |\mathrm{PGL}_{n+1}(E) \cap \Omega| &= \sum_{\gamma \in \mathrm{PGL}_{n+1}(E)} 1_{\Omega}(\gamma) \\ &\ll \sum_{\gamma \in \mathrm{PGL}_{n+1}(E)} \int_{z \in \mathbb{A}_E^{\times}} 1_{\tilde{\Omega}}(z\gamma)\, dz \\ &= \int_{z \in \mathbb{A}_E^{\times}/E^{\times}} \sum_{\gamma \in \mathrm{GL}_{n+1}(E)} 1_{\tilde{\Omega}}(z\gamma)\, dz. \end{aligned}$$

Suppose that z, γ as above satisfy $z\gamma \in \tilde{\Omega}$. Since the sets $\tilde{\Omega}_{\mathfrak{P}}$ ($\mathfrak{P} \in S_E$) are compact neighborhoods of the identity, we have for some fixed $B > 1$ that

$$1/B < \prod_{\mathfrak{P} \in S_E} |\det(z_{\mathfrak{P}}\gamma)|_{\mathfrak{P}} < B.$$

On the other hand, we have $\det(z) = z^{n+1}$, $\prod_{\mathfrak{P}} |\det(\gamma)|_{\mathfrak{P}} = 1$ (the product rule) and $\prod_{\mathfrak{P} \notin S_E} |\det(\gamma)|_{\mathfrak{P}} = D$. It follows that the adelic absolute value $|z| = \prod_{\mathfrak{P}} |z_{\mathfrak{P}}|_{\mathfrak{P}}$ of z satisfies

$$1/B < |z|^{n+1} D < B, \tag{6.14}$$

Since the idelic absolute value $|.| : \mathbb{A}_E^{\times}/E^{\times} \to \mathbb{R}_+^{\times}$ is measure-preserving up to a normalizing scalar, we have

$$\mathrm{vol}\{z \in \mathbb{A}_E^{\times}/E^{\times} : 1/B < |z|^{n+1} D < B\} \ll 1.$$

We thereby reduce to verifying that for each $z \in \mathbb{A}_E^{\times}$ satisfying (6.14), we have

$$|\mathrm{GL}_{n+1}(E) \cap z^{-1}\tilde{\Omega}| \ll D^{n+1}. \tag{6.15}$$

Let $\gamma \in \mathrm{GL}_{n+1}(E) \cap z^{-1}\tilde{\Omega}$. Then for all $i, j \in \{1, \dots, n+1\}$, the matrix entry γ_{ij} satisfies $|\gamma_{ij}|_{\mathfrak{P}} \leqslant C_{\mathfrak{P}}$, where $C_{\mathfrak{P}} \ll |z|_{\mathfrak{P}}^{-1}$ for $\mathfrak{P} \in S_E$ and $C_{\mathfrak{P}} = |z|_{\mathfrak{P}}^{-1}$ for $\mathfrak{P} \notin S_E$. By bounding separately the number of possibilities for each matrix entry of γ, it follows that

$$|\mathrm{GL}_{n+1}(E) \cap z^{-1}\tilde{\Omega}| \ll N^{(n+1)^2}, \quad N := |\{x \in E : |x|_{\mathfrak{P}} \leqslant C_{\mathfrak{P}} \text{ for all } \mathfrak{P}\}.$$

The required estimate (6.15) follows from the adelic Minkowski-type estimate

$$N \ll 1 + \prod_{\mathfrak{P}} C_{\mathfrak{P}} \ll 1 + |z|^{-1} \ll D^{\frac{1}{n+1}}.$$

We turn to (ii). For notational simplicity, we abbreviate $|.|_{\mathfrak{P}} := |.|_{E_{\mathfrak{P}}}$. By our hypothesis (6.9) concerning the $\Omega_{\mathfrak{P}}$ for $\mathfrak{P} \in S_E$, we have $d \neq 0$; moreover, $|b/d|_{\mathfrak{P}}, |c/d|_{\mathfrak{P}} \ll 1$. (The point here is that the stated estimates hold if γ is sufficiently close to the identity element of $\mathrm{PGL}_{n+1}(E_{\mathfrak{P}})$ and are unaffected by multiplying γ on the left and right by elements of a fixed compact subset of $\mathrm{GL}_n(E_{\mathfrak{P}})$, exactly as in the split case of (4.8).) By the product formula, the required implications (6.11) and (6.12) will follow if we can verify that

$$\prod_{\mathfrak{P} \notin S_E} |b/d|_{\mathfrak{P}}, \prod_{\mathfrak{P} \notin S_E} |c/d|_{\mathfrak{P}} \ll D^{\frac{1}{n+1}}. \tag{6.16}$$

Here b/d and c/d are regarded as elements of the vector space E^{n+1} or its dual, and we suppose that the various $E_{\mathfrak{P}}$-norms have been chosen as in §4.1.2 using the standard basis. We note for future reference that for $\mathfrak{P} \notin S_E$ and v in E^{n+1} or its dual, we have

$$|v|_{\mathfrak{P}} \leqslant 1 \text{ whenever each entry of } v \text{ is } \mathfrak{P}\text{-integral}. \tag{6.17}$$

To verify (6.16), let us lift γ to an element of $\mathrm{GL}_{n+1}(E)$. By our (simplifying) assumption that $\mathcal{O}[1/S_E]$ is a principal ideal domain, we may choose the lift in such a way that $\gamma \in \tilde{\Omega}_{\mathfrak{P}}$ for all $\mathfrak{P} \notin S_E$. By the proof of (i), we may find $z \in \mathbb{A}_E^\times$ with $|z| \asymp D^{-\frac{1}{n+1}}$ so that $z_{\mathfrak{P}}\gamma \in \tilde{\Omega}_{\mathfrak{P}}$ for all primes $\mathfrak{P}$.

For $\mathfrak{P} \notin S_E$, both γ and $z_{\mathfrak{P}}\gamma$ belong to $\tilde{\Omega}_{\mathfrak{P}}$; by comparing determinants, it follows that $|z_{\mathfrak{P}}|_{\mathfrak{P}} = 1$. Thus $\prod_{\mathfrak{P} \in S_E} |z_{\mathfrak{P}}|_{\mathfrak{P}} = |z| \asymp D^{-\frac{1}{n+1}}$.

For $\mathfrak{P} \in S_E$, we have $z_{\mathfrak{P}}\gamma \in \tilde{\Omega}_{\mathfrak{P}}$. We may assume that the lift $\tilde{\Omega}_{\mathfrak{P}}$ is of the form $\Omega^1_{\mathfrak{P}}\tilde{\Omega}^0_{\mathfrak{P}}\Omega^1_{\mathfrak{P}}$ with $\tilde{\Omega}^0_{\mathfrak{P}} \subseteq \mathrm{GL}_{n+1}(E_{\mathfrak{P}})$ sufficiently concentrated near the identity element. It follows then that $|z_{\mathfrak{P}} d|_{\mathfrak{P}} \asymp 1$.

Thus

$$\prod_{\mathfrak{P} \notin S_E} |d|_{\mathfrak{P}} = \prod_{\mathfrak{P} \in S_E} |d|_{\mathfrak{P}}^{-1} \asymp D^{-\frac{1}{n+1}}.$$

For $\mathfrak{P} \notin S_E$, we see from the condition $\gamma \in \tilde{\Omega}_{\mathfrak{P}}$ and (6.17) that $|b|_{\mathfrak{P}} \leqslant 1$. Therefore $\prod_{\mathfrak{P} \notin S_E} |b/d|_{\mathfrak{P}} \ll D^{\frac{1}{n+1}}$, as required. The same argument applies to c/d. □

Proof of Lemma 6.1. We first prove (6.7). The aim is to apply Lemma 6.4, taking for S_E the set of places of E lying above some place of S. To that end, we will construct $\Omega_{\mathfrak{P}}$ so that

$$\Sigma \subseteq \Omega := \prod_{\mathfrak{P}} \Omega_{\mathfrak{P}}.$$

For each $\mathfrak{p} \in S$, we see from the compactness of the sets $\Omega_{\mathfrak{p}}$ and $\mathrm{supp}(f_{\mathfrak{p}}^\sharp)$ that Σ is contained in some fixed compact neighborhood of the identity element of $\bar{G}(F_{\mathfrak{p}})$. Thus for the places $\mathfrak{P}$ of E that lie over S, we may take for $\Omega_{\mathfrak{P}}$ a fixed compact neighborhood of the identity element. Since $\Omega_{\mathfrak{p}}$ is contained in $H(F_{\mathfrak{p}})$ while $f_{\mathfrak{p}}^\sharp$ is supported near the identity element, we see moreover that $\Omega_{\mathfrak{P}}$ may be chosen to satisfy the hypothesis (6.9).

We recall from §6.3 that the primes $\mathfrak{p}_1, \mathfrak{p}_2 \in \mathbb{P}$ were chosen to split the extension E/F. Using this feature of our choice, we could "simplify" some of the discussion below by noting that all inertial degrees must be 1. However, we prefer to carry out the analysis in full and ignore this feature of our choice.

For places $\mathfrak{P}$ not lying over S, we define $\Omega_{\mathfrak{P}}$ as in Lemma 6.4 in terms of an element $D_{\mathfrak{P}}$ of the value group of $E_{\mathfrak{P}}^\times$, which we now describe case-by-case.

- If $\mathfrak{P}$ lies over some $\mathfrak{p} \notin S \cup \{\mathfrak{p}_1, \mathfrak{p}_2\}$, then we take $D_{\mathfrak{P}} := 1$. We have

$$\mathrm{supp}(f_{\mathfrak{p}_k}^\sharp) = G(\mathbb{Z}_{\mathfrak{p}_k}) \subseteq \mathrm{PGL}_{n+1}(\mathcal{O}_{\mathfrak{P}}) = \Omega_{\mathfrak{P}}.$$

- If $\mathfrak{p}_1 \neq \mathfrak{p}_2$ and $\mathfrak{P}$ lies over $\mathfrak{p}_k$ for some $k \in \{1, 2\}$, then we take
$$D_{\mathfrak{P}} := \begin{cases} q_{\mathfrak{p}_k}^{ej} & \text{if } k = 1, \\ q_{\mathfrak{p}_k}^{nej} & \text{if } k = 2, \end{cases}$$
where $e \in \{1, 2\}$ denotes the inertial degree of $\mathfrak{P}$ over $\mathfrak{p}_k$. Since $\operatorname{supp}(f_{\mathfrak{p}_1}^{\sharp})$ is the image of
$$\begin{aligned} \operatorname{supp}(f_{\mathfrak{p}_1}) &= \operatorname{supp}(T_{\mathfrak{p}_1}(j, 0, \ldots, 0)) \\ &\subseteq \{g \in M_{n+1}(\mathbb{Z}_{\mathfrak{p}}) : |\det g|_{\mathfrak{p}} = 1/q_{\mathfrak{p}_k}^{j}\} \end{aligned} \tag{6.18}$$
and $\operatorname{supp}(f_{\mathfrak{p}_2}^{\sharp})$ is the image of
$$\begin{aligned} \operatorname{supp}(f_{\mathfrak{p}_2}) &= \operatorname{supp}(T_{\mathfrak{p}_2}(j, \ldots, j, 0)) \\ &\subseteq \{g \in M_{n+1}(\mathbb{Z}_{\mathfrak{p}}) : |\det g|_{\mathfrak{p}} = 1/q_{\mathfrak{p}_k}^{nj}\} \end{aligned} \tag{6.19}$$
and $|\det g|_{\mathfrak{P}} = |\det g|_{\mathfrak{p}}^{e}$, we see that $\operatorname{supp}(f_{\mathfrak{p}_k}^{\sharp}) \subseteq \Omega_{\mathfrak{P}}$.
- If $\mathfrak{p}_1 = \mathfrak{p}_2$ and $\mathfrak{P}$ lies over $\mathfrak{p}_1$, then we take
$$D_{\mathfrak{P}} := q_{\mathfrak{p}_1}^{e(n+1)i}.$$
Arguing as in the previous case and noting that
$$\begin{aligned} \operatorname{supp}(f_{\mathfrak{p}_1}) &= \operatorname{supp}(T_{\mathfrak{p}_1}(i, 0, \ldots, -i)) \\ &= \operatorname{supp}(T_{\mathfrak{p}_1}(2i, i, \ldots, i, 0)) \\ &\subseteq \{g \in M_{n+1}(\mathbb{Z}_{\mathfrak{p}_1}) : |\det g|_{\mathfrak{p}_1} = 1/q_{\mathfrak{p}_1}^{(n+1)i}\}, \end{aligned} \tag{6.20}$$
we see again that $\operatorname{supp}(f_{\mathfrak{p}_k}^{\sharp}) \subseteq \Omega_{\mathfrak{P}}$.

Let D denote the product of $D_{\mathfrak{P}}$ taken over all such $\mathfrak{P}$. Then
$$D \asymp L^{2(n+1)j}.$$
For instance, suppose that $\mathfrak{p}_1 = \mathfrak{p}_2$ and that $\mathfrak{p}_1$ splits in E, say into two primes $\mathfrak{P}, \mathfrak{P}'$. Then $D = D_{\mathfrak{P}} D_{\mathfrak{P}'}$. Each of the factors $D_{\mathfrak{P}}, D_{\mathfrak{P}'}$ is $\asymp q_{\mathfrak{p}_1}^{(n+1)j} \asymp L^{(n+1)j}$, so their product satisfies the required estimate. We argue similarly in the other cases.

By Lemma 6.4, we deduce that
$$|\Sigma| \leqslant |\operatorname{PGL}_{n+1}(E) \cap \Omega| \ll D^{n+1}. \tag{6.21}$$
The proof of (6.7) is thus complete.

We remark that the estimate (6.21) could likely be improved by taking into account that $\Sigma \subseteq \bar{G}(F) \subsetneq \operatorname{PGL}_{n+1}(E)$ and that the containments (6.18), (6.19) and (6.20) are unlikely to be sharp when $n + 1 \geqslant 3$ (compare with [BM1, Lem 6.1] and [BM2]).

To prove the required lower bound (6.8) for $d_{H_{\mathfrak{q}}}(\gamma)$, we write γ as in (6.10) and see from the assumption $\gamma \notin H(F)$ that at least one of the conditions $b \neq 0$ or $c \neq 0$ holds. We then apply (6.11) or (6.12), taking for T the set of primes $\mathfrak{Q}$ of E lying over $\mathfrak{q}$. Supposing for instance that $b \neq 0$, we obtain
$$\prod_{\mathfrak{Q} | \mathfrak{q}} |b/d|_{E_{\mathfrak{Q}}} \gg D^{-\frac{1}{n+1}}.$$
We then apply (4.1) (for $E_{\mathfrak{q}}$ a field) or (4.2) (for $E_{\mathfrak{q}} \cong F_{\mathfrak{q}} \times F_{\mathfrak{q}}$) to deduce that
$$|b/d|_{F_{\mathfrak{q}}} \gg D^{-\frac{1}{2(n+1)}}.$$

Since $D \geqslant 1$, this yields the lower bound

$$d_{H_{\mathfrak{q}}}(\gamma) \gg D^{-\frac{1}{2(n+1)}},$$

which translates to (6.8). □

6.5.3. *Bounds for $I(\gamma)$.*

Lemma 6.5. *Let $\gamma \in \bar{G}(F) - H(F)$. Then*

$$I(\gamma) \ll T^{n/2-1/4+\varepsilon} L^{-nj} d_{H_{\mathfrak{q}}}(\gamma)^{-1/2}.$$

Proof. Setting $S_0 := S - \{\mathfrak{q}\}$ and

$$\Omega_0 := \prod_{\mathfrak{p} \in S_0} \Omega_{\mathfrak{p}}, \quad f_0^\sharp := \otimes_{\mathfrak{p} \in S_0} f_{\mathfrak{p}}^\sharp,$$

we see by (6.6) that

$$I(\gamma) \ll L^{-nj} \int_{x_0, y_0 \in \Omega_0} \left| f_0^\sharp(x_0^{-1} \gamma y_0) \right| I(\gamma; x_0, y_0) \, dx_0 \, dy_0,$$

where

$$I(\gamma; x_0, y_0) := \int_{x_{\mathfrak{q}}, y_{\mathfrak{q}} \in \Omega_{\mathfrak{q}}} \left| \Psi(x_0 x_{\mathfrak{q}}) \Psi(y_0 y_{\mathfrak{q}}) f_{\mathfrak{q}}^\sharp(x_{\mathfrak{q}}^{-1} \gamma y_{\mathfrak{q}}) \right| \, dx_{\mathfrak{q}} \, dy_{\mathfrak{q}}.$$

Theorem 4.2, part (iii) implies that

$$I(\gamma; x_0, y_0) \ll T^{n/2-1/4+\varepsilon} d_{H_{\mathfrak{q}}}(\gamma)^{-1/2} \mathcal{N}(x_0) \mathcal{N}(y_0),$$

where, with $\Omega_{\mathfrak{q}}'$ a fixed compact subset of $H(F_{\mathfrak{q}})$ large enough in temsr of $\Omega_{\mathfrak{q}}$,

$$\mathcal{N}(x_0) := \left(\int_{x_{\mathfrak{q}} \in \Omega_{\mathfrak{q}}'} |\Psi(x_0 x_{\mathfrak{q}})|^2 \, dx_{\mathfrak{q}} \right)^{1/2}.$$

Since $f_0^\sharp(x_0^{-1} \gamma y_0) \ll 1$, we obtain

$$I(\gamma) \ll T^{n/2-1/4+\varepsilon} L^{-nj} d_{H_{\mathfrak{q}}}(\gamma)^{-1/2} \left(\int_{x_0 \in \Omega_0} \mathcal{N}(x_0) \, dx_0 \right)^2.$$

By Cauchy–Schwarz and the estimate $\mathrm{vol}(\Omega_0) \ll 1$, we have

$$\left(\int_{x_0 \in \Omega_0} \mathcal{N}(x_0) \, dx_0 \right)^2 \ll \int_{x_0 \in \Omega_0} \mathcal{N}(x_0)^2 \, dx_0 = \|\Psi\|_{L^2(\Omega_S')}^2, \quad \Omega_S' := \Omega_{\mathfrak{q}}' \times \Omega_0.$$

By covering Ω_S' by finitely many translates of a fundamental domain for $\Gamma_S \backslash H_S$, we see that $\|\Psi\|_{L^2(\Omega_S')}^2 \ll \|\Psi\|^2 \ll 1$. The required estimate follows. □

6.5.4. *Summary.* Combining Lemmas 6.1 and 6.5, we obtain

$$\begin{aligned} \mathcal{E} &\ll T^{n/2-1/4+\varepsilon} L^{-nj} |\Sigma| \max_{\gamma \in \Sigma} d_{H_{\mathfrak{q}}}(\gamma)^{-1/2} \\ &\ll T^{n/2-1/4+\varepsilon} L^{(2(n+1)^2-n+1/2)j}. \end{aligned}$$

Remark 6.6. One could try to improve this estimate by taking into account that not all $\gamma \in \Sigma$ satisfy the worst case lower bound (6.8) for $d_H(\gamma)$.

6.6. Optimization. The estimates obtained above for $\mathcal{M}$ and $\mathcal{E}$ combine to give

$$\mathcal{M} + \mathcal{E} \ll T^{n/2+\varepsilon} \left(L^{-(1-2\vartheta)j} + T^{-1/4} L^{(2(n+1)^2-n+1/2)j} \right). \tag{6.22}$$

Recall our goal bound (6.4). By substituting (6.22) into the decomposition (6.4), we deduce that

$$\begin{aligned}\frac{\mathcal{R}}{L^2 T^{n/2+\varepsilon}} &\ll \sum_{j=1}^{n+1} \left(L^{-(1-2\vartheta)j} + T^{-1/4} L^{(2(n+1)^2-n+1/2)j} \right) \\ &\quad + L^{-1} \sum_{j=0}^{n+1} \left(L^{-(1-2\vartheta)j} + T^{-1/4} L^{(2(n+1)^2-n+1/2)j} \right) \\ &\ll L^{-(1-2\vartheta)} + T^{-1/4} L^{A},\end{aligned}$$

where

$$A := (2(n+1)^2 - n + 1/2)(n+1).$$

The optimal choice for L is the solution to

$$L^{-(1-2\vartheta)} = T^{-1/4} L^{A},$$

i.e.,

$$L = T^{\frac{1}{4(A+1-2\vartheta)}}, \tag{6.23}$$

which gives

$$\frac{\mathcal{R}}{L^2 T^{n/2+\varepsilon}} \ll T^{-\delta}, \quad \delta := \frac{1-2\vartheta}{4(A+1-2\vartheta)},$$

as required.

This completes the proof that Theorem 4.2 implies Theorem 1.1. The remainder of the paper is devoted to the proof of Theorem 4.2.

7. Representation-theoretic preliminaries

In §7–§12, we recall parts of the microlocal calculus developed in [NV, Parts 1,2,3], introducing some crucial refinements (§9.4, §9.5) along the way.

7.1. Lie groups. Let G be a Lie group. We write $\mathfrak{g}$ for its Lie algebra and $\mathfrak{g}^\wedge := i\mathfrak{g}^* = \operatorname{Hom}(\mathfrak{g}, i\mathbb{R})$ for the imaginary dual of the Lie algebra. We denote typical elements of $\mathfrak{g}$ (resp. $\mathfrak{g}^\wedge$) by x, y, z (resp. ξ, η, ζ, τ). We write $\langle x, \xi \rangle \in i\mathbb{R}$ for the result of the canonical pairing. We often abbreviate $x\xi := \langle x, \xi \rangle$. The $\mathrm{U}(1)$-valued pairing $(x, \xi) \mapsto e^{x\xi}$ identifies $\mathfrak{g}^\wedge$ with the Pontryagin dual of $\mathfrak{g}$.

The group G acts on $\mathfrak{g}$ via the adjoint representation $x \mapsto \operatorname{Ad}(g)x$ and on $\mathfrak{g}^\wedge$ via the coadjoint representation $\xi \mapsto \operatorname{Ad}^*(g)\xi$. We abbreviate these actions by $g \cdot x$ and $g \cdot \xi$ or simply $g\xi$. We similarly denote the action of G on functions on either space. For instance, given $a : \mathfrak{g}^\wedge \to \mathbb{C}$, we set $(g \cdot a)(\xi) := a(\operatorname{Ad}^*(g)^{-1}\xi)$. The Lie algebra $\mathfrak{g}$ acts on $\mathfrak{g}$ via the adjoint representation $\operatorname{ad}_x y = [x, y]$ and on $\mathfrak{g}^\wedge$ via the coadjoint representation $\operatorname{ad}_x^* \xi$. We often abbreviate the latter by $[x, \xi]$.

For a Lie subgroup H of G, we obtain a natural inclusion $\mathfrak{h} \leqslant \mathfrak{g}$ and restriction map $\mathfrak{g}^\wedge \twoheadrightarrow \mathfrak{h}^\wedge$. For $\xi \in \mathfrak{g}^\wedge$, we write $\xi_H \in \mathfrak{h}^\wedge$ for its restriction.

7.2. Reductive groups. Suppose now that G is a connected reductive group over an archimedean local field F. By regarding G as a real Lie group, we may apply the above discussion. We write $\mathfrak{g}_{\mathbb{C}} := \mathfrak{g} \otimes_{\mathbb{R}} \mathbb{C}$ for the complexification and $\mathfrak{g}_{\mathbb{C}}^*$ for its complex dual. We regard $\mathfrak{g}^\wedge$ as a real form of $\mathfrak{g}_{\mathbb{C}}^*$; it is the subspace fixed by the involution $\xi \mapsto -\overline{\xi}$.

We write $[\mathfrak{g}_{\mathbb{C}}^*]$ for the geometric invariant theory quotient $\mathfrak{g}_{\mathbb{C}}^* /\!\!/ G$, i.e., the spectrum of the ring of G-invariant complex-valued polynomials on the complex vector space $\mathfrak{g}_{\mathbb{C}}^*$. The real form $[\mathfrak{g}^\wedge]$ of $[\mathfrak{g}_{\mathbb{C}}^*]$ corresponds to those polynomials taking real values on $\mathfrak{g}^\wedge$. It is the fixed point set of the map $\lambda \mapsto -\overline{\lambda}$ on $[\mathfrak{g}_{\mathbb{C}}^*]$ descended from $\mathfrak{g}_{\mathbb{C}}^*$. For details on these points, we refer to [NV, §9.2].

We use a subscripted reg, as in $\mathfrak{g}_{\mathrm{reg}}$ or $\mathfrak{g}_{\mathrm{reg}}^\wedge$, to denote the subset of *regular* elements, i.e., those whose centralizer in $\mathfrak{g}$ has minimal dimension. We record in §13 the essential properties of such elements.

For a vector space V over F, we write $\dim_F(V)$ for its F-dimension and $\dim(V)$ for its $\mathbb{R}$-dimension, so that $\dim(V) = [F : \mathbb{R}] \dim_F(V)$. This notation applies in particular to $V = \mathfrak{g}$. We write $\operatorname{rank}_F(\mathfrak{g})$ (resp. $\operatorname{rank}(\mathfrak{g})$) for the F-dimension (resp. $\mathbb{R}$-dimension) of any Cartan subalgebra of $\mathfrak{g}$ (e.g., $\operatorname{rank}_F(\mathfrak{g}) = n$ if $G = \mathrm{GL}_n(F)$).

7.3. Infinitesimal characters and Langlands parameters. We write $\mathfrak{U}(\cdot)$ for the universal enveloping algebra and $\mathfrak{Z}(\cdot)$ for its center. On any irreducible representation π of G, the algebra $\mathfrak{Z}(\mathfrak{g}_{\mathbb{C}})$ acts via scalars. The infinitesimal character λ_π is the character of $\mathfrak{Z}(\mathfrak{g}_{\mathbb{C}})$ describing that action. The Harish–Chandra isomorphism (see [NV, §9.4]) identifies $\mathfrak{Z}(\mathfrak{g}_{\mathbb{C}})$ with the ring of regular functions on $[\mathfrak{g}_{\mathbb{C}}^*]$. In this way, we may regard λ_π as an element of $[\mathfrak{g}_{\mathbb{C}}^*]$. In the special case that π is unitary, we have in fact $\lambda_\pi \in [\mathfrak{g}^\wedge]$ (see [NV, §9.5]). There is a natural scaling action of $\mathbb{R}^\times$ (resp. $\mathbb{C}^\times$) on $[\mathfrak{g}^\wedge]$ (resp. $[\mathfrak{g}_{\mathbb{C}}^*]$), descended from the corresponding actions on $\mathfrak{g}^\wedge$ (resp. $\mathfrak{g}_{\mathbb{C}}^*$).

Let $G^\vee$ denote the complex dual group of G – regarding G as an algebraic group over F – and $\mathfrak{g}^\vee$ its complex Lie algebra. We may canonically identify

$$\mathfrak{g}^\vee /\!\!/ G^\vee \cong \mathfrak{t}^\vee / W \cong \begin{cases} \mathfrak{t}_{\mathbb{C}}^* / W \cong [\mathfrak{g}_{\mathbb{C}}^*] & \text{if } F = \mathbb{R}, \\ \mathfrak{t}^\vee / W \cong [\mathfrak{g}^\wedge] & \text{if } F = \mathbb{C}, \end{cases} \tag{7.1}$$

where $\mathfrak{t}_{\mathbb{C}} \leqslant \mathfrak{g}_{\mathbb{C}}$ (resp. $\mathfrak{t} \leqslant \mathfrak{g}$) and $\mathfrak{t}^\vee \leqslant \mathfrak{g}^\vee$ are Cartan subalgebras and W denotes the Weyl group. The case $F = \mathbb{R}$ is discussed in [NV, §15.1]; for the case $F = \mathbb{C}$, we first identity $\mathfrak{t}^\vee$ canonically-modulo-W with $\operatorname{Hom}(\mathfrak{t}, \mathbb{C})$ and then identify the latter with $\mathfrak{t}^\vee$ via the map $\mathbb{C} \to i\mathbb{R}$ sending a complex number to its imaginary component.

The identification (7.1) is compatible with the local Langlands correspondence in the following sense (see [NV, §15.1] and references). Let π be a tempered irreducible unitary representation of G. It corresponds to a conjugacy class of representations $\phi_\pi : W_F \to {}^L G$ having bounded image. Here W_F denotes the Weil group of F and ${}^L G = G^\vee \rtimes \operatorname{Gal}(\mathbb{C}/F)$ the Langlands dual group. Using (7.1), we may identify λ_π with a semisimple $G^\vee$-conjugacy class in $\mathfrak{g}^\vee$. The restriction of ϕ_π to the Weil group $W_{\mathbb{C}} = \mathbb{C}^\times \subseteq W_F$ of the complex numbers is then given for $t \in \mathbb{C}$ by the formula $\phi_\pi(\exp(t)) = \exp(2i \operatorname{Im}(t\lambda_\pi))$, where $t\lambda_\pi = \operatorname{Re}(t\lambda_\pi) + i \operatorname{Im}(t\lambda_\pi)$.

7.4. Satake parameters for archimedean classical groups. Let G be a unitary or orthogonal group over an archimedean local field F. Then $G^\vee$ comes equipped with a standard representation $G^\vee \hookrightarrow \mathrm{GL}_n(\mathbb{C})$. In the unitary case,

$n = \dim_E V$; in the orthogonal case, n is even and $\dim_E V$ is either n or $n+1$. For $\lambda \in [\mathfrak{g}^\wedge]$, we write $\mathrm{ev}(\lambda) = \{\lambda_1, \dots, \lambda_n\}$ for the multiset of eigenvalues for the semisimple conjugacy class in $\mathfrak{g}^\vee \hookrightarrow \mathfrak{gl}_n(\mathbb{C})$ corresponding to λ via (7.1). This definition is consistent with that of [NV, §13.4.1].

In particular, we may associate to each tempered irreducible representation π of G a multiset $\mathrm{ev}(\lambda_\pi) = \{\lambda_{\pi,1}, \dots, \lambda_{\pi,n}\}$ of complex numbers. We refer to these as the archimedean Satake parameters of π.

7.5. Kirillov formula. Let π be an irreducible unitary representation of the connected reductive group G over $\mathbb{R}$. Set $2d := \dim(\mathfrak{g}) - \mathrm{rank}(\mathfrak{g})$. We recall the statement of Rossmann's theorem [Ro3, Ro2, Ro1] as summarized in [NV, §6], which asserts the validity of the Kirillov formula [Ki] for such G.

Theorem 7.1. *Assume that π is tempered. There is a (unique) nonempty G-invariant subset $\mathcal{O}_\pi \subseteq \mathfrak{g}^\wedge_{\mathrm{reg}}$ with the following properties.*

(i) The distributional character χ_π of G is described for x in a small enough neighborhood of the identity of $\mathfrak{g}$ by the formula

$$\chi_\pi(\exp(x))\sqrt{\mathrm{jac}(x)} = \int_{\xi \in \mathcal{O}_\pi} e^{x\xi},$$

where $\mathrm{jac}(x)$ denotes the Jacobian of the exponential map as described in [NV, §2.1] and the integral is taken with respect to the symplectic measure on $\mathcal{O}_\pi$, normalized as in [NV, §6.1].

(ii) Let $\lambda_\pi \in [\mathfrak{g}^\wedge]$ denote the infinitesimal character of π, as defined in §7.3. Then $\mathcal{O}_\pi$ is contained in the preimage of λ_π.

We refer to $\mathcal{O}_\pi$ as the *coadjoint multiorbit* attached to π. The terminology reflects that in general, $\mathcal{O}_\pi$ is a finite union of coadjoint orbits. We refer to [NV, §6] for further discussion of this point, as well as the precise definition and normalization of the symplectic measures.

8. Basic operator assignment

We fix a Lie group G and retain the notation of §7.

8.1. Measures and Fourier transforms. We fix Haar measures dx on $\mathfrak{g}$ and $d\xi$ on $\mathfrak{g}^\wedge$ that are Fourier dual in the sense that for Schwartz functions a on $\mathfrak{g}^\wedge$ and ϕ on $\mathfrak{g}$, the Fourier transforms given by

$$a^\vee(x) := \int_{\xi \in \mathfrak{g}^\wedge} a(\xi) e^{-x\xi}\, d\xi, \quad \phi^\wedge(\xi) = \int_{x \in \mathfrak{g}} \phi(x) e^{x\xi}\, dx$$

define mutually inverse transforms. We note that the Schwartz measure $a^\vee(x)\, dx$ on $\mathfrak{g}$ depends only upon a, not upon the choice of Haar measure.

8.2. Nice cutoffs. By a *nice cutoff*, we mean a smooth function $\chi : \mathfrak{g} \to [0,1]$ that is supported in a sufficiently small neighborhood of the origin, satisfies $\chi(x) = \chi(-x)$, and is identically 1 in some neighborhood of the origin.

8.3. The wavelength parameter h. Many of the definitions below depend upon the choice of a positive "wavelength parameter" $\mathrm{h} \in (0,1]$. One can take, e.g., $\mathrm{h} = 1$, but we will be interested primarily in the case that $\mathrm{h} \lll 1$. Since we will work primarily with one value of h at a time, we will often suppress the dependence of our definitions upon h.

8.4. From Schwartz functions on $\mathfrak{g}^\wedge$ to distributions on G. Let a be a Schwartz function on $\mathfrak{g}^\wedge$ and let χ be a nice cutoff. We define the rescaling a_{h} of a by $a_{\mathrm{h}}(\xi) := a(\mathrm{h}\,\xi)$. We obtain the rescaled inverse Fourier transform $a_{\mathrm{h}}^\vee$, given explicitly by $a_{\mathrm{h}}^\vee(x) = \mathrm{h}^{-\dim(\mathfrak{g})}\, a^\vee(x/\mathrm{h})$. The cutoff $\chi a_{\mathrm{h}}^\vee$ of the latter is then a smooth compactly-supported function on $\mathfrak{g}$. We denote by

$$\widetilde{\mathrm{Op}}_{\mathrm{h}}(a:\chi) := \exp_*(\chi a_{\mathrm{h}}^\vee\, dx)$$

the pushforward under the exponential map of the smooth compactly-supported distribution $\chi(x) a_{\mathrm{h}}^\vee(x)\, dx$ on $\mathfrak{g}$. Thus $\widetilde{\mathrm{Op}}_{\mathrm{h}}(a:\chi)$ is a smooth distribution on G, supported in a small neighborhood of the identity element and given there by a smooth multiple of any (left or right) Haar measure dg on G. Explicitly,

$$\int_{g\in G} \widetilde{\mathrm{Op}}_{\mathrm{h}}(a:\chi)(g)\phi(g) = \int_{x\in\mathfrak{g}} \chi(x) a_{\mathrm{h}}^\vee(x) \phi(\exp(x))\, dx \qquad \text{for } \phi : G \to \mathbb{C}.$$

When G is unimodular, we identify $\widetilde{\mathrm{Op}}_{\mathrm{h}}(a,\chi)$ with an element of $C_c^\infty(G)$ by dividing by some fixed (left and right) Haar measure dg.

8.5. Definition of operator assignment. Given a representation π of G (on, e.g., a locally convex space), we define an operator $\mathrm{Op}_{\mathrm{h}}(a:\pi,\chi) \in \mathrm{End}(\pi)$ by integrating: for $v \in \pi$,

$$\begin{aligned}\mathrm{Op}_{\mathrm{h}}(a:\pi,\chi)v &= \int_{g\in G} \widetilde{\mathrm{Op}}_{\mathrm{h}}(a:\chi)(g)\pi(g)v \\ &= \int_{x\in\mathfrak{g}} \chi(x) a_{\mathrm{h}}^\vee(x)\pi(\exp(x))v\, dx.\end{aligned}$$

8.6. Adjoints. The significance of the evenness condition $\chi(x) = \chi(-x)$ imposed in §8.2 is that if π is a unitary representation, then the adjoint of such an operator is given by

$$\mathrm{Op}_{\mathrm{h}}(a:\pi,\chi)^* = \mathrm{Op}_{\mathrm{h}}(\bar{a}:\pi,\chi). \tag{8.1}$$

8.7. Star products and composition. We recall the star product considered in [NV] (and earlier, in [Ri]). Let $\phi_1, \phi_2 \in C_c^\infty(\mathfrak{g})$ be supported close enough to the origin. The distribution $\phi_j(x)\, dx$ pushes forward under the exponential map to a smooth distribution f_j on G supported near the identity. The convolution $f_1 * f_2$ is another smooth distribution on G supported near the identity. We define $\phi_1 \star \phi_2 \in C_c^\infty(\mathfrak{g})$ to be the function supported near the origin for which $(\phi_1 \star \phi_2)(x)\, dx$ pushes forward to $f_1 * f_2$. Thus for any $\Psi \in C_c^\infty(G)$, we have

$$\int_{x_1,x_2\in\mathfrak{g}} \phi_1(x_1)\phi_2(x_2)\Psi(\exp(x_1)\exp(x_2))\, dx_1\, dx_2 = \int_{x\in\mathfrak{g}} (\phi_1 \star \phi_2)(x)\Psi(\exp(x))\, dx.$$

Given a nice cutoff χ and Schwartz functions a, b on $\mathfrak{g}^\wedge$, we define the star product $a \star b$ to be the Schwartz function on $\mathfrak{g}^\wedge$ given by the formula

$$a \star b := (\chi a^\vee \star \chi b^\vee)^\wedge. \tag{8.2}$$

The definition depends implicitly upon χ. The rescaled star product $a \star_{\mathrm{h}} b$ is characterized by

$$(a \star_{\mathrm{h}} b)_{\mathrm{h}} = a_{\mathrm{h}} \star b_{\mathrm{h}}. \tag{8.3}$$

For small elements $x, y \in \mathfrak{g}$, we write $x * y := \log(\exp(x)\exp(y))$ for the small element of $\mathfrak{g}$ for which $\exp(x * y) = \exp(x)\exp(y)$. If χ and χ' are nice cutoffs with the property that $\chi'(x * y) = 1$ whenever $\chi(x)\chi(y) \neq 0$, then

$$\widetilde{\mathrm{Op}}_{\mathrm{h}}(a : \chi) * \widetilde{\mathrm{Op}}_{\mathrm{h}}(b : \chi) = \widetilde{\mathrm{Op}}_{\mathrm{h}}(a \star_{\mathrm{h}} b : \chi'),$$

hence for a representation π of G,

$$\mathrm{Op}_{\mathrm{h}}(a : \pi, \chi)\mathrm{Op}_{\mathrm{h}}(b : \pi, \chi) = \mathrm{Op}_{\mathrm{h}}(a \star_{\mathrm{h}} b : \pi, \chi'). \tag{8.4}$$

8.8. Abbreviations. We abbreviate $\mathrm{Op}_{\mathrm{h}}(a : \pi, \chi)$ by $\mathrm{Op}_{\mathrm{h}}(a : \pi)$ or simply $\mathrm{Op}_{\mathrm{h}}(a)$ when χ and π are clear by context, and similarly for $\widetilde{\mathrm{Op}}_{\mathrm{h}}$. We note that the precise choice of χ is often unimportant (see for instance [NV, §5.4]). We abbreviate further by simply

$$\mathrm{Op}(a)$$

when the wavelength parameter h is understood.[1]

9. Symbols and star product asymptotics

We retain the setting and notation of §8.

9.1. Spaces of symbols. We fix a basis of $\mathfrak{g}^\wedge$. With respect to this choice, we define for each multi-index

$$\alpha = (\alpha_1, \dots, \alpha_{\dim(\mathfrak{g})}) \in \mathbb{Z}_{\geqslant 0}^{\dim(\mathfrak{g})}$$

a differential operator ∂^α on $C^\infty(\mathfrak{g}^\wedge)$ (see [NV, §4.1] for details).

For each integer m and multi-index α, we define a seminorm $\nu_{m,\alpha}$ on $C^\infty(\mathfrak{g}^\wedge)$ (valued in the extended nonnegative reals) by the formula

$$\nu_{m,\alpha}(a) := \sup_{\xi \in \mathfrak{g}^\wedge} \langle \xi \rangle^{|\alpha| - m} \left| \partial^\alpha a(\xi) \right|,$$

where $|\alpha| := \sum \alpha_j$ denotes the order of α and $\langle \xi \rangle := (1 + |\xi|^2)^{1/2}$. We set[2]

$$\underline{S}^m := \{a \in C^\infty(\mathfrak{g}^\wedge) : \nu_{m,\alpha}(a) < \infty \text{ for all } \alpha\}.$$

We extend the definition of $\underline{S}^m$ to $m = \infty$ (resp. $m = -\infty$) by taking the union (resp. intersection) over all integers m. We write more verbosely $\underline{S}^m(\mathfrak{g}^\wedge)$ when we wish to indicate which group is being considered. For $m < \infty$, the space $\underline{S}^m$ is a Frechet space equipped with a distinguished family of seminorms, while $\underline{S}^\infty$ is an inductive limit of such spaces. We note that $\underline{S}^{-\infty}$ is the Schwartz space on $\mathfrak{g}^\wedge$.

9.2. Star product: formal expansion and extension to symbols. Recall from §8.7 that the star product $\star$ depends upon the choice of a nice cutoff χ. For small $x, y \in \mathfrak{g}$, write $\{x, y\} := x * y - x - y$, with $x * y = \log(\exp(x)\exp(y))$ as in §8.7. For the moment, let a and b be Schwartz functions on $\mathfrak{g}^\wedge$. The rescaled star product admits the integral representation (see [NV, (4.2)])

$$a \star_{\mathrm{h}} b(\zeta) = \int_{x,y \in \mathfrak{g}} a_{\mathrm{h}}^\vee(x) b_{\mathrm{h}}^\vee(y) e^{\langle x, \zeta/h \rangle} e^{\langle y, \zeta/\mathrm{h} \rangle} e^{\langle \{x,y\}, \zeta/\mathrm{h} \rangle} \chi(x)\chi(y) \, dx \, dy. \tag{9.1}$$

[1] This convention differs from that in [NV], where Op refers to the "unscaled" operator assignment, i.e., the case $\mathrm{h} = 1$. We refer here to the latter more verbosely as Op_1.

[2] The space $\underline{S}^m$ was denoted in [NV, §4.3] by S^m. We do not adopt the latter notation here.

The analytic function $e^{\{x,y\}\zeta} = e^{\langle \{x,y\},\zeta\rangle}$ admits the Taylor expansion

$$e^{\{x,y\}\zeta} = \sum_{\alpha,\beta,\gamma} c_{\alpha\beta\gamma} x^\alpha y^\beta \zeta^\gamma, \tag{9.2}$$

where the multi-indices α, β, γ range over $\mathbb{Z}_{\geqslant 0}^{\dim(\mathfrak{g})}$ and we write

$$x^\alpha = \prod x_j^{\alpha_j}, \quad y^\beta = \prod y_j^{\beta_j} \text{ and } \zeta^\gamma = \prod \zeta_j^{\gamma_j}$$

for the corresponding monomials, defined using coordinates $\mathfrak{g} \cong \mathbb{R}^{\dim(\mathfrak{g})}$ and $\mathfrak{g}^\wedge \cong i\mathbb{R}^{\dim(\mathfrak{g})}$ with respect to which the natural pairing $\mathfrak{g} \otimes \mathfrak{g}^\wedge \to i\mathbb{R}$ is given by $x\xi = \sum_{j=1}^{\dim(\mathfrak{g})} x_j \xi_j$.

Using the estimate

$$\{x, y\} \ll |x| \cdot |y| \quad (\text{for } x, y \lll 1),$$

we may see as in [NV, §4.2] that the coefficient $c_{\alpha\beta\gamma}$ in the Taylor expansion (9.2) of $e^{\{x,y\},\zeta}$ vanishes unless

$$|\gamma| \leqslant \min(|\alpha|, |\beta|). \tag{9.3}$$

Grouping the terms in (9.2) by their homogeneity degree $j = |\alpha| + |\beta| - |\gamma|$ (which is nonnegative in view of (9.3)) and applying Fourier inversion to (9.1) suggests the formal asymptotic expansion

$$a \star_{\mathrm{h}} b \sim \sum_{j \geqslant 0} \mathrm{h}^j \, a \star^j b, \tag{9.4}$$

where $\star^j$ denotes the finite bidifferential operator defined by

$$a \star^j b(\zeta) = \sum_{|\alpha|+|\beta|-|\gamma|=j} c_{\alpha\beta\gamma} \zeta^\gamma \partial^\alpha a(\zeta) \partial^\beta b(\zeta).$$

Various rigorous forms of (9.4) were verified in [NV]. For instance:

Theorem 9.1. *The star product $\star$ and its rescaled variant $\star_{\mathrm{h}}$, defined initially for Schwartz functions, extend uniquely to a compatible family of continuous bilinear maps*

$$\star_{\mathrm{h}} : \underline{S}^{m_1} \times \underline{S}^{m_2} \to \underline{S}^{m_1+m_2} \tag{9.5}$$

with the convention $\infty + (-\infty) := -\infty$. For each $(a, b) \in \underline{S}^{m_1} \times \underline{S}^{m_2}$ and $J \in \mathbb{Z}_{\geqslant 0}$, we have

$$a \star_{\mathrm{h}} b \equiv \sum_{0 \leqslant j < J} \mathrm{h}^j \, a \star^j b \mod \underline{S}^{m_1+m_2-J},$$

where the remainder term defines a continuous map $\underline{S}^{m_1} \times \underline{S}^{m_2} \to \underline{S}^{m_1+m_2-J}$.

Proof. See [NV, Thm 1]. □

9.3. Basic symbol classes. As motivation, let us explicate the continuity of the map (9.5), restricting for simplicity to the case that m_1 and m_2 are finite.

- *Let G be a Lie group. Let $m_1, m_2 \in \mathbb{Z}$. Let $\mathrm{h} \in (0, 1]$. Let χ be a nice cutoff, so that $\star$ and $\star_{\mathrm{h}}$ are defined. For each multi-index γ, there exists $C \geqslant 0$ and a finite family $\mathcal{N}$ of pairs of multi-indices (α, β) so that for all $(a, b) \in \underline{S}^{m_1} \times \underline{S}^{m_2}$, we have*

$$\nu_{m_1+m_2,\gamma}(a \star_{\mathrm{h}} b) \leqslant C \max_{(\alpha,\beta)\in\mathcal{N}} \nu_{m_1,\alpha}(a)\nu_{m_2,\beta}(b).$$

In this statement, $\mathcal{N}$ and C may depend freely upon h. The purpose of this section is to quantify that dependence under additional hypotheses on a and b.

Definition 9.2. For fixed $0 \leqslant \delta < 1$ and $m \in \mathbb{Z}$, we define the following class (§3.1.3) of symbols:

$$S^m_\delta := \left\{ a \in \underline{S}^m : \nu_{m,\alpha}(a) \ll \mathrm{h}^{-\delta|\alpha|} \text{ for each fixed } \alpha \right\}.$$

Thus S^m_δ consists of those $a \in \underline{S}^m$ satisfying

$$\partial^\alpha a(\xi) \ll \mathrm{h}^{-\delta|\alpha|} \langle \xi \rangle^{m-|\alpha|} \tag{9.6}$$

for each fixed multi-index α and all $\xi \in \mathfrak{g}^\wedge$. (More pedantically, S^m_δ is the pair $(\underline{S}^m, P)$, where $P(a)$ is the predicate "for each fixed α there is a fixed C so that $|\partial^\alpha a(\xi)| \leqslant C\, \mathrm{h}^{-\delta|\alpha|} \langle \xi \rangle^{m-|\alpha|}$ for all ξ.")

We extend the definition to $m = \pm\infty$ by taking for S^∞_δ the union of S^m_δ over all fixed m and for $S^{-\infty}_\delta$ the intersection of $\underline{S}^{-\infty} \cap S^m_\delta$ over all fixed m. Thus in all cases, $S^m_\delta \subseteq \underline{S}^m$. We write more verbosely $S^m_\delta(\mathfrak{g}^\wedge)$ or $S^m_\delta[\mathrm{h}]$ or $S^m_\delta(\mathfrak{g}^\wedge)[\mathrm{h}]$ when we wish to indicate which group and/or wavelength parameter we are considering.

Remark 9.3. This definition is a mild reformulation of that given in [NV, §4.4]. There, we worked instead with functions $a(\xi, \mathrm{h})$ of two variables $(\xi, \mathrm{h}) \in \mathfrak{g}^\wedge \times (0, 1]$, smooth in the first variable, satisfying $|\partial^\alpha a(\xi, \mathrm{h})| \leqslant C_\alpha\, \mathrm{h}^{-\delta|\alpha|} \langle \xi \rangle^{m-|\alpha|}$. The two definitions have similar content, and the reader will lose little by working with whichever feels more comfortable. The practical disadvantage of the definition of [NV, §4.4] is that its user must require many quantities to be "h-dependent." This disadvantage would be more severe for the applications pursued in this paper. We discuss how to translate between the two formulations in the proof of Theorem 9.5.

Remark 9.4. We emphasize, following §3.1, that "let $a \in S^m_\delta$" itself carries no semantic content, but any complete statement involving S^m_δ does. For example, statement (i) below, formulated in terms of symbol classes, is formally equivalent to statement (ii), formulated explicitly. Both statements are moreover equivalent to (iii), an apparent strengthening of (ii) that follows by linearity. (Each statement also happens to be true and readily provable using integration by parts.)

(i) *Fix a Lie group G. Let $\mathrm{h} \in (0, 1]$. Fix $\delta \in [0, 1)$. Let $a \in S^{-\infty}_\delta$. Then the integral defining $a^\vee$ converges absolutely, and we have $a^\vee(x) \ll |\mathrm{h}^\delta x|^{-N}$ for all fixed $N \geqslant 0$ and all x.*
(ii) *Let G be a Lie group. Let $\delta \in [0, 1)$. Let $(C_{m,\alpha})$ be a collection of nonnegative reals indexed by $m \in \mathbb{Z}$ and $\alpha \in \mathbb{Z}^{\dim(\mathfrak{g})}_{\geqslant 0}$. For each $N \geqslant 0$, there is a finite subset $\mathcal{N}$ of $\mathbb{Z} \times \mathbb{Z}^{\dim(\mathfrak{g})}_{\geqslant 0}$ and $C \geqslant 0$ with the following property. Let $\mathrm{h} \in (0, 1]$. Let $a \in \underline{S}^{-\infty}$ with $|\partial^\alpha a(\xi)| \leqslant C_{m,\alpha}\, \mathrm{h}^{-\delta|\alpha|} \langle \xi \rangle^{m-|\alpha|}$ for $(m, \alpha) \in \mathcal{N}$ and all ξ. Then the integral defining $a^\vee$ converges absolutely, and we have $|a^\vee(x)| \leqslant C|\mathrm{h}^\delta x|^{-N}$ for all x.*
(iii) *Let G be a Lie group. Let $\delta \in [0, 1)$. For each $N \geqslant 0$, there is a finite subset $\mathcal{N}$ of $\mathbb{Z} \times \mathbb{Z}^{\dim(\mathfrak{g})}_{\geqslant 0}$ and $C \geqslant 0$ with the following property. Let $\mathrm{h} \in (0, 1]$. Let $a \in \underline{S}^{-\infty}$ with*

$$\nu(a) := \max_{(m,\alpha) \in \mathcal{N}} \sup_\xi \mathrm{h}^{\delta|\alpha|} \langle \xi \rangle^{|\alpha|-m} |\partial^\alpha a(\xi)| < \infty.$$

Then the integral defining $a^\vee$ *converges absolutely, and we have* $|a^\vee(x)| \leqslant C\nu(a)|\mathrm{h}^\delta x|^{-N}$ *for all* x.

Every theorem that we state involving such classes admits a similar translation. We hope the passage from the first statement to the second is intuitively clear thanks to the informal association between "fixed" and "absolute constant," and refer again to [Ne, §2] for details on how to carry out such translations algorithmically.

For a positive real c, we write cS^m_δ for the class of symbols of the form cv, with $v \in S^m_\delta$. We write $\mathrm{h}^\infty S^m_\delta$ for the intersection over all fixed N of the classes $\mathrm{h}^N S^m_\delta$. The class $\mathrm{h}^\infty S^{-\infty}_\delta$ is independent of δ; it consists of Schwartz functions $a : \mathfrak{g}^\wedge \to \mathbb{C}$ satisfying $\partial^\alpha a(\xi) \ll \mathrm{h}^N \langle \xi \rangle^{-N}$ for all fixed α and N. This class, which should be regarded as consisting of "negligible" symbols, will be denoted simply by

$$\mathrm{h}^\infty S^{-\infty}.$$

Theorem 9.5. *Fix a nice cutoff* χ. *Fix* $\delta \in [0, 1/2)$ *and* $m_1, m_2 \in \mathbb{Z} \cup \{\pm\infty\}$. *The star product, extended as in Theorem 9.1, induces a class map*

$$\star_{\mathrm{h}} : S^{m_1}_\delta \times S^{m_2}_\delta \to S^{m_1+m_2}_\delta. \tag{9.7}$$

For $(a, b) \in S^{m_1}_\delta \times S^{m_2}_\delta$ *and any fixed* $j \in \mathbb{Z}_{\geqslant 0}$, *we have*

$$a \star^j b \in \mathrm{h}^{-2\delta j} S^{m_1+m_2-j}_\delta. \tag{9.8}$$

Moreover, for any fixed $J \in \mathbb{Z}_{\geqslant 0}$,

$$a \star_{\mathrm{h}} b \equiv \sum_{0 \leqslant j < J} \mathrm{h}^j a \star^j b \mod \mathrm{h}^{(1-2\delta)J} S^{m_1+m_2-J}_\delta. \tag{9.9}$$

Proof. As we will explain, this follows formally from [NV, Thm 1].

Recall from §9.1 the spaces $\underline{S}^m$, defined in [NV, §4.3], consisting of smooth functions $a : \mathfrak{g}^\wedge \to \mathbb{C}$ such that for each α, there exists $C_\alpha(a) \geqslant 0$ so that $|\partial^\alpha a(\xi)| \leqslant C_\alpha(a) \langle \xi \rangle^{m-|\alpha|}$ for all ξ. We now record another definition, this time from [NV, §4.4], which we do not otherwise use in this paper. Given a subset $\mathcal{H}$ of $(0, 1]$ having 0 as an accumulation point, let $\underline{S}^m_\delta$ denote the space of functions $a : \mathcal{H} \times \mathfrak{g}^\wedge \to \mathbb{C}$, smooth in the second variable, so that for each α, there exists $C_\alpha(a) \geqslant 0$ so that for each $\mathrm{h} \in \mathcal{H}$, the specialization $a[\mathrm{h}] : \mathfrak{g}^\wedge \to \mathbb{C}, \xi \mapsto a(\mathrm{h}, \xi)$ satisfies $|\partial^\alpha a[\mathrm{h}](\xi)| \leqslant C_\alpha(a) \, \mathrm{h}^{-\delta|\alpha|} \langle \xi \rangle^{m-|\alpha|}$ for all ξ. For $(a, b) \in \underline{S}^{m_1}_\delta \times \underline{S}^{m_2}_\delta$, we define $a \star_{\mathrm{h}} b : \mathcal{H} \times \mathfrak{g}^\wedge \to \mathbb{C}$ by requiring that $(a \star_{\mathrm{h}} b)[\mathrm{h}] = a[\mathrm{h}] \star_{\mathrm{h}} b[\mathrm{h}]$ for each $\mathrm{h} \in \mathcal{H}$. We extend these definitions to infinite exponents $m = \pm\infty$ by taking unions or intersections. Each of these spaces comes equipped with an "evident topology" defined by a family of seminorms (detailed in *loc. cit.*).[3]

We verify (9.7) in detail; we may similarly deduce (9.8) and (9.9) from the corresponding assertions in [NV, Thm 1]. We consider the case $m_1, m_2 \in \mathbb{Z}$; the cases in which some $m_j = \pm\infty$ follow similarly. As shown in [NV, Thm 1] (and noted partially in Theorem 9.1), the star product and its rescaling define compatible families of continuous bilinear maps

$$\star : \underline{S}^{m_1} \times \underline{S}^{m_2} \to \underline{S}^{m_1+m_2}. \tag{9.10}$$

$$\star_{\mathrm{h}} : \underline{S}^{m_1}_\delta \times \underline{S}^{m_2}_\delta \to \underline{S}^{m_1+m_2}_\delta. \tag{9.11}$$

We will see that (9.7) is a formal consequence of these mapping properties.

[3]The space $\underline{S}^m_\delta$ was denoted "S^m_δ" in [NV, §4.4]. The latter notation is used differently here.

For clarity, we employ here the more verbose notation $S_\delta^m = S_\delta^m[\mathrm{h}]$ for our symbol classes. Let $\mathrm{h} \in (0,1]$ and $(a,b) \in S_\delta^{m_1}[\mathrm{h}] \times S_\delta^{m_2}[\mathrm{h}]$. By Theorem 9.1, the rescaled star product $a \star_{\mathrm{h}} b$ lies in the space $\underline{S}^{m_1+m_2}$. We must check that it in fact lies in the class $S_\delta^{m_1+m_2}[\mathrm{h}]$. In other words, we must verify the following:

- *Fix a Lie group G and a nice cutoff χ. Let $\mathrm{h} \in (0,1]$. Fix $\delta \in [0,1/2)$. Let $a \in S_\delta^{m_1}[\mathrm{h}]$ and $b \in S_\delta^{m_2}[\mathrm{h}]$. Then $a \star_{\mathrm{h}} b \in S_\delta^{m_1+m_2}[\mathrm{h}]$.*

As in Remark 9.4, this is formally equivalent to:

- *Let G be a Lie group. Let χ be a nice cutoff. Let $\delta \in [0,1/2)$. For each $m \in \mathbb{Z}$ and $\alpha \in \mathbb{Z}_{\geqslant 0}^{\dim(\mathfrak{g})}$ and $\mathrm{h} \in (0,1]$, we define a seminorm $\mu_{m,\alpha}[\mathrm{h}]$ on $C^\infty(\mathfrak{g}^\wedge)$ by the formula*

$$\mu_{m,\alpha}[\mathrm{h}](a) := \sup_\xi \mathrm{h}^{\delta|\alpha|} \langle \xi \rangle^{|\alpha|-m} |\partial^\alpha a(\xi)|.$$

  Write $\mu_{m,\alpha}$ for the map $\mathrm{h} \mapsto \mu_{m,\alpha}[\mathrm{h}]$. Set $\mathcal{N}(m) := \{\mu_{m,\alpha} : \alpha \in \mathbb{Z}_{\geqslant 0}^{\dim(\mathfrak{g})}\}$.

  Let $m_1, m_2 \in \mathbb{Z}$. For each $\mu \in \mathcal{N}(m_1+m_2)$, there exists $C \geqslant 0$ and finite subsets $\mathcal{N}_1 \subseteq \mathcal{N}(m_1)$ and $\mathcal{N}_2 \subseteq \mathcal{N}(m_2)$ with the following property. For each $(a,b) \in \underline{S}^{m_1} \times \underline{S}^{m_2}$ and $\mathrm{h} \in (0,1]$ for which the quantities

$$\nu_{\mathcal{N}_1}[\mathrm{h}](a) := \max_{\nu \in \mathcal{N}_1} \nu[\mathrm{h}](a) \quad \text{and} \quad \nu_{\mathcal{N}_2}[\mathrm{h}](b) := \max_{\nu \in \mathcal{N}_2} \nu[\mathrm{h}](b)$$

  are finite, the rescaled star product $a \star_{\mathrm{h}} b$ satisfies

$$\mu[\mathrm{h}](a \star_{\mathrm{h}} b) \leqslant C \nu_{\mathcal{N}_1}[\mathrm{h}](a) \nu_{\mathcal{N}_2}[\mathrm{h}](b).$$

Suppose this fails. We may then find a Lie group G, a nice cutoff χ, and elements $\delta \in [0,1/2)$, $m_1, m_2 \in \mathbb{Z}$, $\mu \in \mathcal{N}(m_1+m_2)$ with the following property: for each $C \geqslant 0$ and all finite subsets $\mathcal{N}_1 \subseteq \mathcal{N}(m_1)$ and $\mathcal{N}_2 \subseteq \mathcal{N}(m_2)$, there exists $(a,b,\mathrm{h}) \in \underline{S}^{m_1} \times \underline{S}^{m_2} \times (0,1]$ so that $\mu(a \star_{\mathrm{h}} b) > C \nu_{\mathcal{N}_1}[\mathrm{h}](a) \nu_{\mathcal{N}_2}[\mathrm{h}](a)$. (We incorporate into this inequality the assertion that each factor on the RHS is finite.) By a diagonalization argument and linearity, we may find a sequence of elements $(a_j, b_j, \mathrm{h}_j) \in \underline{S}^{m_1} \times \underline{S}^{m_2} \times (0,1]$ so that for all finite subsets $\mathcal{N}_1 \subseteq \mathcal{N}(m_1)$ and $\mathcal{N}_2 \subseteq \mathcal{N}(m_2)$, we have

$$\sup_j \nu_{\mathcal{N}_1}[\mathrm{h}_j](a_j) < \infty, \quad \sup_j \nu_{\mathcal{N}_2}[\mathrm{h}_j](b_j) < \infty, \quad \lim_{j \to \infty} \mu[\mathrm{h}_j](a_j \star_{\mathrm{h}_j} b_j) = \infty. \tag{9.12}$$

By passing to a subsequence, we may reduce to the following cases.

(i) $\inf_j \mathrm{h}_j > 0$. In this case, $\star_{\mathrm{h}}$ and $\star$ differ mildly, and our supposition (9.12) contradicts the continuity of (9.10).

(ii) $\mathrm{h}_j \to 0$. By passing to a further subsequence, we may assume that the sequence h_j is decreasing. Set $\mathcal{H} := \{h_j\}$, and define $a, b : \mathcal{H} \times \mathfrak{g}^\wedge \to \mathbb{C}$ by requiring that $a[\mathrm{h}_j] = a_j$ and $b[\mathrm{h}_j] = b_j$. By (9.12), we then have $a \in \underline{S}_\delta^{m_1}$ and $b \in \underline{S}_\delta^{m_2}$, but $a \star_{\mathrm{h}} b \notin \underline{S}_\delta^{m_1+m_2}$, contrary to (9.11). $\square$

9.4. Refined symbol classes. Supposing now that the Lie group G is a connected reductive group over $\mathbb{R}$ (so that centralizers of regular elements behave as nicely as possible), we introduce some refinements of the basic symbol classes described above. For motivation concerning the introduction of such refinements, we refer to §2.4.

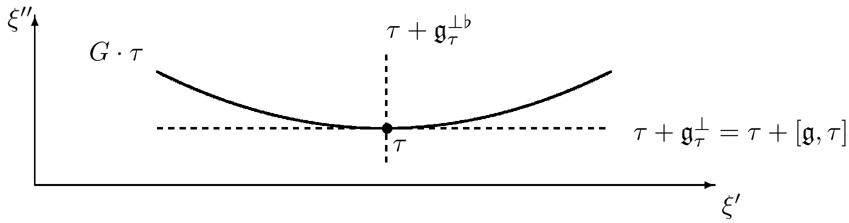

FIGURE 3. The coadjoint orbit $G \cdot \tau$ near τ in τ-coordinates.

9.4.1. *Coordinates tailored to regular coadjoint elements.* Recall from §7.2 that $\mathfrak{g}^\wedge_{\mathrm{reg}}$ denotes the set of regular elements in $\mathfrak{g}^\wedge$. Let $\tau \in \mathfrak{g}^\wedge_{\mathrm{reg}}$. The centralizer $\mathfrak{g}_\tau$ in $\mathfrak{g}$ is then an abelian subalgebra of dimension equal to the rank of $\mathfrak{g}$ (see [Ku]). As τ varies in $\mathfrak{g}^\wedge_{\mathrm{reg}}$, the centralizer $\mathfrak{g}_\tau$ varies smoothly in the Grassmannian of rank($\mathfrak{g}$)-dimensional subspaces of $\mathfrak{g}$.

We may form the complement $\mathfrak{g}_\tau^\perp \subseteq \mathfrak{g}^\wedge$ of $\mathfrak{g}_\tau$ with respect to the canonical duality $\mathfrak{g} \otimes \mathfrak{g}^\wedge \to i\mathbb{R}$. It has codimension equal to the rank of $\mathfrak{g}$.

The tangent plane in $\mathfrak{g}^\wedge$ to the coadjoint orbit $G \cdot \tau$ at τ is the coset $\tau + \mathfrak{g}_\tau^\perp$. This fact follows from the identities

$$\langle [x, y], \tau \rangle = \langle x, [y, \tau] \rangle = -\langle y, [x, \tau] \rangle \quad \text{for } x, y \in \mathfrak{g} \tag{9.13}$$

and a comparison of dimensions.

We fix an inner product on $\mathfrak{g}^\wedge$. We use this inner product to define an orthogonal complement $\mathfrak{g}_\tau^{\perp\flat} \subseteq \mathfrak{g}^\wedge$ to $\mathfrak{g}_\tau^\perp$. We have the direct sum decomposition

$$\mathfrak{g}^\wedge = \mathfrak{g}_\tau^\perp \oplus \mathfrak{g}_\tau^{\perp\flat}. \tag{9.14}$$

Remark 9.6. It is convenient, but not essential, to take for $\mathfrak{g}_\tau^{\perp\flat}$ the orthogonal complement of $\mathfrak{g}_\tau^\perp$ with respect to an inner product; we could take instead any linear complement arising from a smooth assignment $\tau \mapsto \mathfrak{g}_\tau^{\perp\flat}$.

Let $\mathfrak{g}_\tau^\flat \subseteq \mathfrak{g}$ denote the complement of $\mathfrak{g}_\tau^{\perp\flat}$ with respect to the canonical duality between $\mathfrak{g}$ and $\mathfrak{g}^\wedge$. We obtain the following decomposition, dual to (9.14):

$$\mathfrak{g} = \mathfrak{g}_\tau^\flat \oplus \mathfrak{g}_\tau. \tag{9.15}$$

We equip $\mathfrak{g}$ with the inner product dual to that on $\mathfrak{g}^\wedge$. Then both decompositions (9.14) and (9.15) are orthogonal. In particular, $\mathfrak{g}_\tau^\flat$ is the orthogonal complement of $\mathfrak{g}_\tau$.

For $x \in \mathfrak{g}$ and $\xi' \in \mathfrak{g}^\wedge$, we write $x = (x', x'')$ and $\xi = (\xi', \xi'')$ for the coordinates induced by the above decompositions, so that

$$\xi' \in \mathfrak{g}_\tau^\perp, \qquad \xi'' \in \mathfrak{g}_\tau^{\perp\flat},$$
$$x' \in \mathfrak{g}_\tau^\flat, \qquad x'' \in \mathfrak{g}_\tau.$$

We refer to these as *τ-coordinates*. We note that $\langle x, \xi \rangle = \langle x', \xi' \rangle + \langle x'', \xi'' \rangle$.

Example 9.7. Suppose that $G = \mathrm{GL}_3(\mathbb{R})$. Let us identify $\mathfrak{g}^\wedge$ with $\mathfrak{g}$ via the trace pairing divided by the imaginary unit i, and equip $\mathfrak{g}^\wedge$ with the inner product defined by taking the standard Euclidean inner product of the matrix entries. We consider two representative examples:

(i) Suppose first that τ is regular semisimple, say a diagonal matrix with distinct entries. Then

$$\mathfrak{g}_\tau^\perp = \begin{pmatrix} & * & * \\ * & & * \\ * & * & \end{pmatrix}, \quad \mathfrak{g}_\tau^{\perp\flat} = \begin{pmatrix} * & & \\ & * & \\ & & * \end{pmatrix},$$

$$\mathfrak{g}_\tau^\flat = \begin{pmatrix} & * & * \\ * & & * \\ * & * & \end{pmatrix}, \quad \mathfrak{g}_\tau = \begin{pmatrix} * & & \\ & * & \\ & & * \end{pmatrix}.$$

(ii) Suppose next that τ is regular nilpotent, say

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

Then

$$\mathfrak{g}_\tau^\perp = \left\{ \begin{pmatrix} a & b & 0 \\ c & d & -b \\ e & f & -a-d \end{pmatrix} \right\}, \quad \mathfrak{g}_\tau^{\perp\flat} = \left\{ \begin{pmatrix} a & b & c \\ & a & b \\ & & a \end{pmatrix} \right\},$$

$$\mathfrak{g}_\tau^\flat = \left\{ \begin{pmatrix} a & b & c \\ d & e & f \\ 0 & -d & -a-e \end{pmatrix} \right\}, \quad \mathfrak{g}_\tau = \left\{ \begin{pmatrix} a & & \\ b & a & \\ c & b & a \end{pmatrix} \right\}.$$

We note that τ does not belong to a consistent summand of (9.14): we have $\tau \in \mathfrak{g}_\tau^{\perp\flat}$ in case (i), $\tau \in \mathfrak{g}_\tau^\perp$ in case (ii).

9.4.2. *Euclidean coordinates.* By choosing orthonormal bases for each of the spaces occurring in the decompositions (9.14) and (9.15), we equip those spaces with Euclidean coordinates. We assume our bases chosen so that the coordinates on the mutually dual spaces in the indicated decompositions are compatible in the sense that

$$\langle x', \xi' \rangle = \sum_j x_j' \xi_j', \quad \langle x'', \xi'' \rangle = \sum_j x_j'' \xi_j''. \tag{9.16}$$

9.4.3. *Definition and informal discussion of refined symbol classes.*

Definition 9.8. Assume that the wavelength parameter satisfies $\mathrm{h} \lll 1$. Let δ', δ'' be fixed quantities satisfying

$$0 < \delta' < \delta'' < 2\delta' < 1. \tag{9.17}$$

Let τ be an element of $\mathfrak{g}^\wedge$ that belongs to some fixed compact subset of $\mathfrak{g}^\wedge_{\mathrm{reg}}$. We then define the symbol class (§3.1.3)

$$S^\tau_{\delta',\delta''}$$

to consist of all $a \in \underline{S}^{-\infty}$ (i.e., Schwartz functions) with the following properties.

(i) For some fixed $R > 0$, we have $\mathrm{supp}(a) \subseteq \{\xi \in \mathfrak{g}^\wedge : |\xi - \tau| \leqslant R\,\mathrm{h}^{\delta''-\delta'}\}$.

(ii) For all fixed multi-indices $\alpha \in \mathbb{Z}_{\geqslant 0}^{\dim(\mathfrak{g})-\operatorname{rank}(\mathfrak{g})}$ and $\beta \in \mathbb{Z}_{\geqslant 0}^{\operatorname{rank}(\mathfrak{g})}$ and all $\xi \in \mathfrak{g}^\wedge$, we have

$$\partial_{\xi'}^{\alpha} \partial_{\xi''}^{\beta} a(\xi) \ll \mathrm{h}^{-\delta'|\alpha|-\delta''|\beta|}, \tag{9.18}$$

where $\xi = \xi' + \xi''$ are the τ-coordinates and the differential operators $\partial_{\xi'}^{\alpha}, \partial_{\xi''}^{\beta}$ are defined using Euclidean coordinates as in §9.4.2. (The definition does not depend upon the choice of orthonormal bases used to define those coordinates: any two choices are related via an orthogonal change of variables, the partial derivatives of which are $\mathrm{O}(1)$.)

We write more verbosely $S^{\tau}_{\delta',\delta''}(\mathfrak{g}^\wedge)$ or $S^{\tau}_{\delta',\delta''}[\mathrm{h}]$ or $S^{\tau}_{\delta',\delta''}(\mathfrak{g}^\wedge)[\mathrm{h}]$ when we wish to indicate which group and/or wavelength parameter we are considering.

Notation as at the end of §9.3 may be applied to the refined classes. For instance, we may define $cS^{\tau}_{\delta',\delta''}$ for a positive real c.

Informally, $S^{\tau}_{\delta',\delta''}$ consists of symbols that are supported quite close to τ and that oscillate with wavelength

- at least $\mathrm{h}^{\delta'}$ in the ξ'-directions tangent to the coadjoint orbit $G \cdot \tau$ and
- at least $\mathrm{h}^{\delta''}$ in the ξ''-directions transverse to that coadjoint orbit.

Since δ'' is larger than δ', less regularity is required in the transverse directions than in the tangent directions. In applications, we will take δ' close to $1/2$ and δ'' close to 1.

For the sake of visualization, a good example to keep in mind is when $G = \mathrm{SO}(3)$, so that $\mathfrak{g}^\wedge$ identifies with $\mathbb{R}^3$ in such a way that the coadjoint orbits are the Euclidean spheres, with G acting via rotation. If $\tau = (0,0,1)$ is the "north pole" of the unit sphere and $\langle , \rangle$ is the standard Euclidean inner product, then the τ-coordinates of $\xi = (\xi_1, \xi_2, \xi_3)$ are $\xi' = (\xi_1, \xi_2)$ and $\xi'' = x_3$. For $\delta' \approx 1/2$ and $\delta'' \approx 1$, a typical element of $S^{\tau}_{\delta',\delta''}$ is roughly a smooth bump on the coin-shaped domain

$$\{\xi \in \mathbb{R}^3 : 1 - \mathrm{h} < |\xi| < 1 + \mathrm{h}, |\xi - \tau| < \mathrm{h}^{1/2}\}.$$

We note that $S^{\tau}_{\delta',\delta''} \subseteq S^{-\infty}_{\delta''}$.

9.4.4. *Insensitivity to basepoint.* The significance of the exponent $\delta'' - \delta'$ in the support condition (i) is to ensure that $S^{\tau}_{\delta',\delta''}$ is insensitive to the precise choice of "basepoint" τ, in the following sense:

Lemma 9.9. *Let $\mathrm{h}, \delta', \delta''$ be as in Definition 9.8. Let τ_1, τ_2 be elements of $\mathfrak{g}^\wedge$ that belong to some fixed compact subset of $\mathfrak{g}^\wedge_{\mathrm{reg}}$. If*

$$\tau_1 = \tau_2 + \mathrm{O}(\mathrm{h}^{\delta''-\delta'}), \tag{9.19}$$

then the symbol classes coincide:

$$S^{\tau_1}_{\delta',\delta''} = S^{\tau_2}_{\delta',\delta''};$$

otherwise, their intersection is trivial:

$$\cap_{j=1,2} S^{\tau_j}_{\delta',\delta''} = \{0\}.$$

More precisely, if (9.19) holds, then for each smooth function $a : \mathfrak{g}^\wedge \to \mathbb{C}$ and each $\xi = \tau_j + \mathrm{O}(\mathrm{h}^{\delta''-\delta'})$, the estimate (9.18) holds with respect to τ_1-coordinates if and only if it holds with respect to τ_2-coordinates.

A useful consequence of this result is that to check whether a smooth function a supported on $\tau_1 + \mathrm{O}(\mathrm{h}^{\delta''-\delta'})$ belongs to the class $S^{\tau_1}_{\delta',\delta''}$, it suffices to verify for each $\tau_2 = \tau_1 + \mathrm{O}(\mathrm{h}^{\delta''-\delta'})$ that the derivative bounds (9.18) hold at $\xi = \tau_2$ with respect to τ_2-coordinates.

Proof. The support condition (i) and our assumption $\mathrm{h} \lll 1$ imply that the intersection is trivial if (9.19) is not satisfied, so suppose (9.19) holds. The support condition (i) is then the same for both symbol classes, so it is enough to verify that if a smooth function $a : \mathfrak{g}^\wedge \to \mathbb{C}$ satisfies (9.18) with respect to the coordinates defined by τ_1, then it satisfies the same with respect to the coordinates defined by τ_2. Set

$$m := \dim(\mathfrak{g}) - \mathrm{rank}(\mathfrak{g}), \quad n := \mathrm{rank}(\mathfrak{g}).$$

Let

$$\mathfrak{g}^{\perp}_{\tau_j} \cong \mathbb{R}^m, \quad \mathfrak{g}^{\perp\flat}_{\tau_j} \cong \mathbb{R}^n \tag{9.20}$$

denote the isomorphisms arising from the Euclidean coordinates defined in §9.4.2. Write $\xi = \xi_1' + \xi_1'' = \xi_2' + \xi_2''$ for the respective τ_1- and τ_2-coordinates. With respect to (9.20), these coordinates are related via an invertible linear change of variables

$$\begin{pmatrix} \xi_1' \\ \xi_1'' \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} \xi_2' \\ \xi_2'' \end{pmatrix}$$

with A, B, C, D matrices of respective dimensions $m \times m, m \times n, n \times m, n \times n$. Since the assignment $\tau \mapsto \mathfrak{g}^{\perp}_{\tau}$ is smooth and in particular Lipschitz, we deduce from (9.19) the operator norm estimates

$$A, D = 1 + \mathrm{O}(\mathrm{h}^{\delta''-\delta'}) = \mathrm{O}(1), \quad B, C = \mathrm{O}(\mathrm{h}^{\delta''-\delta'}). \tag{9.21}$$

Let $f_1 : \mathbb{R}^m \oplus \mathbb{R}^n \to \mathbb{C}$ be the function describing a with respect to the coordinates $\xi \mapsto (\xi_1', \xi_1'')$. Then $f_2(x, y) := f_1(Ax + By, Cx + Dy)$ describes a with respect to the coordinates $\xi \mapsto (\xi_2', \xi_2'')$, and our task is to show that if the estimates

$$\partial_x^\alpha \partial_y^\beta f_j(x, y) \ll \mathrm{h}^{-\delta'|\alpha| - \delta''|\beta|} \quad (\alpha \in \mathbb{Z}^m_{\geqslant 0}, \beta \in \mathbb{Z}^n_{\geqslant 0} \text{ fixed})$$

hold for $j = 1$, then they hold also for $j = 2$.

We explain the required implication in the notationally simplest case $m = n = 1$, in which A, B, C, D are scalars and α, β are integers. (This case does not actually arise for any $\mathfrak{g}$, but the calculus exercise we are solving makes sense for general m and n.) In that case, $\partial_x^\alpha \partial_y^\beta f_2(x, y)$ evaluates via the chain rule to

$$\sum_{k=0}^{\alpha} \sum_{l=0}^{\beta} \binom{\alpha}{k} \binom{\beta}{l} A^{\alpha-k} C^k B^l D^{\beta-l} \partial_1^{\alpha-k+l} \partial_2^{\beta-l+k} f_1(Ax + By, Cx + Dy),$$

where for clarity we write ∂_1, ∂_2 for derivatives in the first and second variables. Invoking our hypotheses concerning f_1, we reduce the required bound for f_2 to the estimate

$$A^{\alpha-k} C^k B^l D^{\beta-l} \, \mathrm{h}^{-\delta'(-k+l) - \delta''(-l+k)} \ll 1.$$

To verify this, we invoke (9.21) in the weaker form

$$A, B, D = \mathrm{O}(1), \quad C = \mathrm{O}(\mathrm{h}^{\delta''-\delta'}).$$

and recall that $\delta'' \geqslant \delta'$. The case of general (m, n) is treated similarly by inserting indices throughout the above argument. ∎

9.4.5. *Basic properties.* We record some basic mapping properties concerning the symbol classes just introduced, to be applied repeatedly in what follows. Each property follows readily from the definition.

Lemma 9.10. *Let* $\mathrm{h}, \delta', \delta'', \tau$ *be as in Definition 9.8. Abbreviate* $S := S^{\tau}_{\delta', \delta''}$.

(i) Let $P : \mathfrak{g}^{\wedge} \to \mathbb{C}$ *be a fixed polynomial. Then multiplication by* P *defines a class map*

$$P : S \to S.$$

(ii) Set $m := \dim(\mathfrak{g}) - \operatorname{rank}(\mathfrak{g})$, $n := \operatorname{rank}(\mathfrak{g})$. *Fix* $\alpha' \in \mathbb{Z}^m_{\geqslant 0}$, $\alpha'' \in \mathbb{Z}^n_{\geqslant 0}$. *Define the partial differential operators* $\partial^{\alpha'}_{\xi'}, \partial^{\alpha''}_{\xi''}$ *as in (9.18), using the coordinates defined by* τ. *Then*

$$\partial^{\alpha'}_{\xi'} : S \to \mathrm{h}^{-\delta'|\alpha'|} S,$$

$$\partial^{\alpha''}_{\xi'} : S \to \mathrm{h}^{-\delta''|\alpha''|} S.$$

9.4.6. *"Completing" the refined symbol classes.* By

$$S^{\tau}_{\delta', \delta''} + \mathrm{h}^{\infty} S^{-\infty}$$

we mean the class of symbols of the form $a + b$ with $a \in S^{\tau}_{\delta', \delta''}$ and $b \in \mathrm{h}^{\infty} S^{-\infty}$. It may be regarded informally as a "completion" of $S^{\tau}_{\delta', \delta''}$ obtained by adjoining the "negligible" symbols. This "completed" class admits a more direct characterization:

Lemma 9.11. *Let* $a \in \underline{S}^{-\infty}$. *Then* a *belongs to* $S^{\tau}_{\delta', \delta''} + \mathrm{h}^{\infty} S^{-\infty}$ *if and only if there is a fixed* $R > 0$ *so that the following conditions hold.*

- *For all fixed* α, N *and all* ξ *with* $|\xi - \tau| > R\,\mathrm{h}^{\delta'' - \delta'}$, *we have*

$$\partial^{\alpha} a(\xi) \ll \mathrm{h}^N \langle \xi \rangle^{-N},$$

where as before $\langle \xi \rangle = (1 + |\xi|^2)^{1/2}$.

- *For all fixed* α, β *and all* ξ *with* $|\xi - \tau| < 2R\,\mathrm{h}^{\delta'' - \delta'}$, *we have in* τ*-coordinates*

$$\partial^{\alpha}_{\xi'} \partial^{\beta}_{\xi''} \alpha(\xi) \ll \mathrm{h}^{-\delta'|\alpha| - \delta''|\beta|}.$$

Proof. The necessity of these conditions is clear from the definitions. Conversely, suppose these conditions hold. We may construct an element p of the basic symbol class $S^{-\infty}_{\delta'' - \delta'}$ satisfying

- $p(\xi) = 1$ for $|\xi - \tau| < 3R\,\mathrm{h}^{\delta'' - \delta'}$,
- $p(\xi) = 0$ for $|\xi - \tau| > 4R\,\mathrm{h}^{\delta'' - \delta'}$.

We then verify readily that pa belongs to $S^{\tau}_{\delta', \delta''}$ and $(1 - p)a$ belongs to $\mathrm{h}^{\infty} S^{-\infty}$. We conclude by noting that $a = pa + (1 - p)a$. □

9.4.7. *Star product asymptotics.* We will establish analogues of (9.8) and (9.9) for the refined symbol classes. The possibility of doing so boils down to the fact that the BCHD formula consists of Lie polynomials; by contrast, the proof of Theorem 9.5 made direct use only of the analyticity of that formula and the fact that $x * y = x + y + \mathrm{O}(|x| \cdot |y|)$.

Theorem 9.12. *Fix a nice cutoff* χ. *Assume that* $\mathrm{h} \lll 1$. *Fix* δ', δ'' *satisfying (9.17). Write* j, J *for fixed elements of* $\mathbb{Z}_{\geqslant 0}$ *and* τ, τ_1, τ_2 *for elements of some fixed compact subset of* $\mathfrak{g}^{\wedge}_{\mathrm{reg}}$.

(i) $\star_{\mathrm{h}}$ *enjoys the mapping property*

$$\star_{\mathrm{h}} : S^{\tau_1}_{\delta',\delta''} \times S^{\tau_2}_{\delta',\delta''} \to S^{\tau_1}_{\delta',\delta''} \cap S^{\tau_2}_{\delta',\delta''} + \mathrm{h}^{\infty} S^{-\infty}. \tag{9.22}$$

(ii) If $\tau_1 - \tau_2 \ggg \mathrm{h}^{\delta''-\delta'}$*, then* $a \star^j b = 0$ *for all* $a \in S^{\tau_1}_{\delta',\delta''}$*,* $b \in S^{\tau_2}_{\delta',\delta''}$*.*

(iii) $\star^j$ *enjoys the mapping property*

$$\star^j : S^{\tau}_{\delta',\delta''} \times S^{\tau}_{\delta',\delta''} \to \mathrm{h}^{-2\delta' j} S^{\tau}_{\delta',\delta''}. \tag{9.23}$$

For any $a, b \in S^{\tau}_{\delta',\delta''}$*, we have the asymptotic expansion*

$$a \star_{\mathrm{h}} b \equiv \sum_{0 \leqslant j < J} \mathrm{h}^j a \star^j b \mod \mathrm{h}^{(1-2\delta')J} S^{\tau}_{\delta',\delta''} + \mathrm{h}^{\infty} S^{-\infty}. \tag{9.24}$$

(iv) Fix $\delta \in [0, \delta']$*. We have the mapping properties*

$$\star_{\mathrm{h}} : S^{\infty}_{\delta} \times S^{\tau}_{\delta',\delta''} \to S^{\tau}_{\delta',\delta''} + \mathrm{h}^{\infty} S^{-\infty},$$

$$\star^j : S^{\infty}_{\delta} \times S^{\tau}_{\delta',\delta''} \to \mathrm{h}^{-(\delta+\delta')j} S^{\tau}_{\delta',\delta''}$$

and, for $(a, b) \in S^{\infty}_{\delta} \times S^{\tau}_{\delta',\delta''}$*, the asymptotic expansion*

$$a \star_{\mathrm{h}} b \equiv \sum_{0 \leqslant j < J} \mathrm{h}^j a \star^j b \mod \mathrm{h}^{(1-\delta-\delta')J} S^{\tau}_{\delta',\delta''} + \mathrm{h}^{\infty} S^{-\infty}.$$

Analogous results hold with $S^{\infty}_{\delta} \times S^{\tau}_{\delta',\delta''}$ *replaced by* $S^{\tau}_{\delta',\delta''} \times S^{\infty}_{\delta}$*.*

Proof. We indicate here the overall structure of the proof, the details of which are given below.

Assertion (ii) follows immediately from the support condition (i) in Definition 9.8.

As for (iii), we establish in §9.5.2 the mapping property (9.23) for $\star^j$, then in §9.5.3 the asymptotic expansion (9.24). The mapping property (9.22) for $\star_{\mathrm{h}}$ follows in the special case $\tau_1 = \tau_2$ by specializing the asymptotic expansion (9.24) to $J = 0$.

By Lemma 9.9, we may suppose in proving the general case of (9.22) that $\tau_1 - \tau_2 \ggg \mathrm{h}^{\delta'-\delta''}$. The proof in that case is implicit in the proof of (9.24), but for expository purposes it seems best treated separately in §9.5.5.

We discuss part (iv) in §9.5.6. The proof of amounts to an "interpolation" between the proofs given here and the proof of Theorem 9.5 given in [NV, §7.3]. □

9.5. Proofs of refined star product asymptotics. Here we complete the proof of Theorem 9.12 following the outline indicated above. On a first reading, we recommend skipping this section and proceeding to §10.

9.5.1. *Taylor coefficient support conditions.* We retain the notation and setting of §9.2. The support condition (9.3) was sufficient for the purposes of [NV]. We require here a stronger support condition. Let $\tau \in \mathfrak{g}^{\wedge}_{\mathrm{reg}}$. Recall from §9.2 the definition of $\{x, y\}$. Let

$$e^{\{x,y\}\tau} = \sum_{\alpha',\alpha'',\beta',\beta'',\gamma} c_{\alpha'\alpha''\beta'\beta''\gamma} (x')^{\alpha'} (x'')^{\alpha''} (y')^{\beta'} (y'')^{\beta''} \tau^{\gamma} \tag{9.25}$$

denote the Taylor expansion in τ-coordinates, where monomials such as $(x')^{\alpha'} = \prod_j (x'_j)^{\alpha'_j}$ are defined with respect to the Euclidean coordinates of §9.4.2.

Lemma 9.13. *Suppose that* $c_{\alpha'\alpha''\beta'\beta''\gamma} \neq 0$. *Then*

$$|\alpha'| + |\beta'| \geqslant 2|\gamma|,$$

hence, with $j := |\alpha'| + |\beta'| + |\alpha''| + |\beta''| - |\gamma|$,

$$|\alpha'| + |\beta'| + 2|\alpha''| + 2|\beta''| \leqslant 2j \tag{9.26}$$

Proof. The two required inequalities are restatements of one another, so we focus on the first. We use the fact (BCHD formula) that each homogeneous component of the Taylor expansion of $\{x, y\}$ is a finite iterated Lie polynomial in x and y, i.e., a linear combination of expressions $L(x, y) = \mathrm{ad}_{z_1} \cdots \mathrm{ad}_{z_{n-1}} z_n$ with each z_j equal to either x or to y. By expanding the exponential series, we reduce to verifying for each such expression that if we write

$$L(x,y)\tau = \sum_{\alpha',\alpha'',\beta',\beta'',\gamma} C_{\alpha'\alpha''\beta'\beta''\gamma} (x')^{\alpha'} (x'')^{\alpha''} (y')^{\beta'} (y'')^{\beta''} \tau^{\gamma},$$

then

$$C_{\alpha'\alpha''\beta'\beta''\gamma} \neq 0 \implies |\alpha'| + |\beta'| \geqslant 2|\gamma|.$$

Suppose $C_{\alpha'\alpha''\beta'\beta''\gamma} \neq 0$. Set $n := |\alpha'| + |\alpha''| + |\beta'| + |\beta''|$. Then n is the degree of the Lie monomial L (e.g., $n = 3$ if $L(x, y) = [x, [x, y]]$), and we may find nonzero elements $x_1, \dots, x_n$ of $\mathfrak{g}_\tau^\flat$ or of $\mathfrak{g}_\tau$ so that

- the number of x_j lying in $\mathfrak{g}_\tau^\flat$ is $|\alpha'| + |\beta'|$,
- the number of x_j lying in $\mathfrak{g}_\tau$ is $|\alpha''| + |\beta''|$, and
- $\langle [x_1, [x_2, \dots, [x_{n-1}, x_n]]], \tau \rangle \neq 0$.

We have $|\gamma| = 1$, so our task is to verify that $|\alpha'| + |\beta'| \geqslant 2$. Suppose otherwise that $|\alpha'| + |\beta'| \leqslant 1$. Then $|\alpha''| + |\beta''| \geqslant n - 1$, i.e., at least $n - 1$ of the x_j centralize τ. Since $\mathfrak{g}_\tau$ is a subalgebra, it follows that at least one of the elements $x := x_1$ or $y := [x_2, \dots, [x_{n-1}, x_n]]$ centralizes τ, and yet $\langle [x, y], \tau \rangle = 0$. The required contradiction then follows from (9.13). ☐

9.5.2. *Mapping properties of homogeneous components.*

Proof of (9.23). Let $a, b \in S^{\tau}_{\delta',\delta''}$. We must verify that $a \star^j b$ belongs to the symbol class $\mathrm{h}^{-2\delta' j} S^{\tau}_{\delta',\delta''}$. Since $\star^j$ is a finite order differential operator, the support condition in the definition of that symbol class clearly holds. It is also clear that $a \star^j b \in \underline{S}^{-\infty}$. Setting $m := \dim(\mathfrak{g}) - \mathrm{rank}(\mathfrak{g})$ and $n := \mathrm{rank}(\mathfrak{g})$, we must thus check that for all fixed multi-indices $\eta' \in \mathbb{Z}^m_{\geqslant 0}, \eta'' \in \mathbb{Z}^n_{\geqslant 0}$ and all $\xi = \tau + \mathrm{O}(\mathrm{h}^{\delta'' - \delta'})$, we have

$$\partial^{\eta'}_{\xi'} \partial^{\eta''}_{\xi''} (a \star^j b)(\xi) \ll \mathrm{h}^{-\delta'|\eta'| - \delta''|\eta''| - 2\delta' j}. \tag{9.27}$$

By Lemma 9.9, we may and shall assume that $\xi = \tau$.

We begin by verifying (9.27) in the special case $\eta' = 0$, $\eta'' = 0$. By expanding the definition of $\star^j$ and appealing to Lemma 9.13, we reduce to checking for all fixed $\alpha', \alpha'', \beta', \beta'', \gamma$ satisfying (9.26) that in τ-coordinates, we have

$$\left(\partial^{\alpha'}_{\xi'} \partial^{\alpha''}_{\xi''} a(\tau) \right) \left(\partial^{\beta'}_{\xi'} \partial^{\beta''}_{\xi''} b(\tau) \right) \tau^{\gamma} \ll \mathrm{h}^{-2\delta' j}. \tag{9.28}$$

By our hypothesis that $\tau = \mathrm{O}(1)$, we have in particular $\tau^\gamma \ll 1$. By the definition of the symbol class $S^{\tau}_{\delta',\delta''}$, we see that

$$\partial^{\alpha'}_{\xi'} \partial^{\alpha''}_{\xi''} a(\tau) \ll \mathrm{h}^{-\delta'|\alpha'| - \delta''|\alpha''|}, \quad \partial^{\beta'}_{\xi'} \partial^{\beta''}_{\xi''} b(\tau) \ll \mathrm{h}^{-\delta'|\beta'| - \delta''|\beta''|}.$$

It will thus suffice to verify under the stated conditions that

$$\delta'|\alpha'| + \delta''|\alpha''| + \delta'|\beta'| + \delta''|\beta''| \leqslant 2\delta' j. \tag{9.29}$$

To see this, we use that $\delta'' \leqslant 2\delta'$ and apply (9.26).

The proof is complete in the special case $\eta' = 0, \eta'' = 0$. The general case is deduced similarly, using the product rule to evaluate the application of the differential operator $\partial_{\xi'}^{\eta'} \partial_{\xi''}^{\eta''}$ to the LHS of (9.28) and recalling that $\tau = \mathrm{O}(1)$. ☐

9.5.3. *Estimates for the remainder.* We now prove (9.24). The proof is similar in overall structure to [NV, Thm 6]. The present proof is simpler because our symbols have uniformly bounded support, but more complicated because of the more complicated nature of the symbol class.

Reduction to tail estimates. Set $r := a \star_{\mathrm{h}} b - \sum_{0 \leqslant j < J} \mathrm{h}^j \, a \star^j b$. By Theorem 9.1, we have $r \in \underline{S}^{-\infty}$. By the criterion of Lemma 9.11, we must verify that for some fixed (large enough) $R > 0$,

- for all $\zeta \in \mathfrak{g}^\wedge$ with $|\zeta - \tau| > R\,\mathrm{h}^{\delta'' - \delta'}$, all fixed multi-indices γ and each fixed N, we have (with $\langle \zeta \rangle = (1 = |\zeta|^2)^{1/2}$, as before)

$$\partial_\zeta^\gamma r(\zeta) \ll \mathrm{h}^N \langle \zeta \rangle^{-N}, \tag{9.30}$$

  and
- for all $\zeta \in \mathfrak{g}^\wedge$ with $|\zeta - \tau| < 2R\,\mathrm{h}^{\delta'' - \delta'}$ and all fixed multi-indices γ', γ'', we have

$$\partial_{\xi'}^{\gamma'} \partial_{\xi''}^{\gamma''} r(\zeta) \ll \mathrm{h}^{(1 - 2\delta')J - \delta'|\gamma'| - \delta''|\gamma''|}. \tag{9.31}$$

In fact, since $1 - 2\delta' < 0$, the factors $\mathrm{h}^{(1-2\delta')J}$ decrease as J increases. Thus, by the mapping properties (9.23) established above, we see that the terms $\mathrm{h}^j \, a \star^j b$ ($j \leqslant J$) satisfy the analogue of the estimate (9.31) required by r. These terms are supported on the intersection of the supports of a and b. Choosing R large enough that $\{\xi : |\xi - \tau| < R\}$ contains those supports, we reduce to showing that for each fixed γ and N, the estimate (9.30) holds provided that J is large enough but fixed. We record the details below in the notationally simplest case $\gamma = 0$. The general case may be verified by differentiating each step of the proof, noting that our inputs from [NV] apply to derivatives (see [NV, §7.7, Proof of Prop] for further details).

Reduction to the case of localized symbols. We may assume – after smoothly decomposing a and b into $\mathrm{h}^{-\mathrm{O}(1)}$ pieces – that for some $\omega_1, \omega_2 \in \mathfrak{g}^\wedge$ satisfying the estimates

$$\omega_1, \omega_2 = \tau + \mathrm{O}(\mathrm{h}^{\delta'' - \delta'}),$$

the symbols a and b satisfy the support conditions

$$\begin{aligned} \operatorname{supp}(a) &\subseteq \{\xi \in \mathfrak{g}^\wedge : |\xi' - \omega_1'| \leqslant \mathrm{h}^{\delta'}, |\xi'' - \omega_1''| \leqslant \mathrm{h}^{\delta''}\}, \\ \operatorname{supp}(b) &\subseteq \{\xi \in \mathfrak{g}^\wedge : |\xi' - \omega_2'| \leqslant \mathrm{h}^{\delta'}, |\xi'' - \omega_2''| \leqslant \mathrm{h}^{\delta''}\}, \end{aligned} \tag{9.32}$$

with τ-coordinates $\xi = (\xi', \xi'')$ as usual. In view of our assumptions that $\mathrm{h} \lll 1$ and $\delta', \delta'' > 0$, these support conditions constrain the symbols a and b to small neighborhoods of their respective "basepoints" ω_1 and ω_2.

We pause to describe the Fourier transforms $a_{\mathrm{h}}^\vee, b_{\mathrm{h}}^\vee$. Set

$$A' := \mathrm{h}^{-1+\delta'}, \quad A'' := \mathrm{h}^{-1+\delta''}, \quad m := \dim(\mathfrak{g}) - \operatorname{rank}(\mathfrak{g}), \quad n := \operatorname{rank}(\mathfrak{g}). \tag{9.33}$$

It follows then readily from the definition of $S^{\tau}_{\delta',\delta''}$, the support conditions (9.32) of a and b and elementary Fourier analysis that with respect to τ-coordinates $x = (x', x'')$ on $\mathfrak{g}$, we may write

$$a_{\mathrm{h}}^{\vee}(x) = e^{-\langle x, \omega_1/\mathrm{h}\rangle}(A')^m (A'')^n \phi_1(A'x', A''x''), \tag{9.34}$$
$$b_{\mathrm{h}}^{\vee}(x) = e^{-\langle x, \omega_2/\mathrm{h}\rangle}(A')^m (A'')^n \phi_2(A'x', A''x''),$$

where ϕ_1, ϕ_2 are Schwartz functions on $\mathfrak{g}$ with each fixed Schwartz seminorm of size $\mathrm{O}(1)$. From this description we see that for any fixed nonnegative integers k', k'', l', l'',

$$\int_{x,y\in\mathfrak{g}} |a_{\mathrm{h}}^{\vee}(x) b_{\mathrm{h}}^{\vee}(y)|\, |x'|^{k'} |x''|^{k''} |y'|^{l'} |y''|^{l''}\, dx\, dy \ll (A')^{-k'-l'}(A'')^{-k''-l''}. \tag{9.35}$$

Set

$$Q := \mathrm{h}^{-1}\langle\zeta\rangle \in \mathbb{R}_{\geqslant 1},$$

so that our goal bound (specialized as indicated to $\gamma' = 0, \gamma'' = 0$) may be written

$$r(\zeta) \ll Q^{-N}. \tag{9.36}$$

Since $\mathrm{h} \lll 1$, we have $Q \ggg 1$.

Reduction to the case of controlled arguments. We fix $\varepsilon > 0$ sufficiently small in terms of δ' and δ''. We consider first the case that $|\zeta| \geqslant Q^{\varepsilon}$. In that case – in view of the estimate $Q \ggg 1$ and the assumed support conditions (9.32) on a and b – we have $a \star^j b(\zeta) = 0$ for all j. We thereby reduce to showing that $a \star_{\mathrm{h}} b(\zeta) \ll Q^{-N}$. This estimate may be verified by a crude application of partial integration to the integral representation (9.1), repeating the argument around [NV, (7.15), (7.16)] verbatim. We thereby reduce to the case $|\zeta| < Q^{\varepsilon}$.

Reduction to the case of nearby arguments. We next to reduce to the critical case

$$\zeta = \tau + O(\mathrm{h}^{\delta''-\delta'}). \tag{9.37}$$

Suppose otherwise that $\zeta - \tau \ggg \mathrm{h}^{\delta''-\delta'}$. Since a and b are both supported on $\tau + \mathrm{O}(\mathrm{h}^{\delta''-\delta'})$, we then have $a \star^j b(\zeta) = 0$ for all j, so our task is to check that

$$a \star_{\mathrm{h}} b(\zeta) \ll \mathrm{h}^N.$$

We start with the integral representation (9.1). The estimate (9.35) says informally that $a_{\mathrm{h}}^{\vee}(x) b_{\mathrm{h}}^{\vee}(y)$ is concentrated on $x', y' \ll \mathrm{h}^{1-\delta'}$ and $x'', y'' \ll \mathrm{h}^{1-\delta''}$. Quantitatively, let us define smooth cutoffs χ' (resp. χ'') on $\mathfrak{g}_{\tau}^{\flat}$ (resp. $\mathfrak{g}_{\tau}$) by using the coordinates defined in §9.4.2 to transport some fixed smooth cutoffs on the corresponding Euclidean spaces. Setting

$$B' := \mathrm{h}^{-\varepsilon}(A')^{-1}, \quad B'' := \mathrm{h}^{-\varepsilon}(A'')^{-1}.$$

we deduce from (9.35) that the error incurred in the integral representation (9.1) by replacing our original cutoff $\chi(x)\chi(y)$ with the reduced cutoff

$$f(x,y) := \chi'(x'/B')\chi''(x''/B'')\chi'(y'/B')\chi''(y''/B'')$$

is $\mathrm{O}(\mathrm{h}^{\infty})$. Having shrunk the cutoff in this way, we open the Fourier integrals defining $a_{\mathrm{h}}^{\vee}(x)$ and $b_{\mathrm{h}}^{\vee}(y)$ to express $a \star_{\mathrm{h}} b(\zeta)$ up to negligible error as

$$\mathrm{h}^{-2\dim(\mathfrak{g})} \int_{\xi,\eta} a(\xi) b(\eta) \left(\int_{x,y} f(x,y) e^{(x(\zeta-\xi)+y(\zeta-\eta)+\{x,y\}\zeta)/\mathrm{h}}\, dx\, dy \right) d\xi\, d\eta$$

Since $a(\xi)b(\eta)$ vanishes unless

$$\xi, \eta = \tau + \mathrm{O}(\mathrm{h}^{\delta''-\delta'}) \tag{9.38}$$

in which case $a(\xi)b(\eta) \ll 1$, it is enough to show that if (9.38) holds and $|\zeta - \tau| \ggg \mathrm{h}^{\delta''-\delta'}$, then

$$\int_{x,y} f(x,y) e^{(x(\zeta-\xi)+y(\zeta-\eta)+\{x,y\}\zeta)/\mathrm{h}}\, dx\, dy \ll \mathrm{h}^N.$$

To see this, we integrate by parts, but in a more refined way than in [NV, §7]. We record the details below in Lemma 9.14.

Taylor expansion. Having reduced to the critical case (9.37), we introduce the notation

$$\Omega(x,y,\zeta) := e^{\{x,y\}\zeta} = \sum_{\substack{\alpha,\beta,\gamma: \\ |\gamma| \leqslant \min(|\alpha|,|\beta|)}} c_{\alpha\beta\gamma} x^\alpha y^\beta \zeta^\gamma = \sum_{j \geqslant 0} \Omega_j(x,y,\zeta),$$

where Ω_j denotes the homogeneous component obtained by restricting the summation to $|\alpha| + |\beta| - |\gamma| = j$, so that

$$a \star^j b(\zeta) = \int_{x,y} a^\vee(x) b^\vee(y) e^{x\zeta} e^{y\zeta} \Omega_j(x,y,\zeta)\, dx\, dy.$$

We write $\Omega^{(J)} := \Omega - \sum_{0 \leqslant j < J} \Omega_j$ for the remainder obtained by subtracting from Ω its first J homogeneous components. Using the integral representation (9.1), we may then split $r(\zeta) = r'(\zeta) + r''(\zeta)$, where

$$r'(\zeta) := \int_{x,y} a_{\mathrm{h}}^\vee(x) b_{\mathrm{h}}^\vee(y) \Omega^{(J)}(x,y,\zeta/\mathrm{h})\, dx\, dy,$$

$$r''(\zeta) := \int_{x,y} a_{\mathrm{h}}^\vee(x) b_{\mathrm{h}}^\vee(y) (\chi(x)\chi(y) - 1) \Omega(x,y,\zeta/\mathrm{h})\, dx\, dy.$$

Since $|\Omega(x,y,\zeta)| = 1$ and χ is identically 1 near the origin, we obtain from (9.35) the estimate $r''(\zeta) \ll (A'A'')^{-M}$ for any fixed M. Since $\max(\delta', \delta'') < 1$, it follows that $r''(\zeta) \ll \mathrm{h}^M$ for any fixed M. Our restriction $|\zeta| \leqslant Q^\varepsilon$ implies that, say, $Q \ll \mathrm{h}^{-2}$, so this last estimate for $r''(\zeta)$ suffices.

It remains to estimate $r'(\zeta)$. For this we Taylor expand $\Omega^{(J)}$ as in [NV, §7.4], but incorporating the refined Taylor coefficient support condition afforded by Lemma 9.13. Thanks to the condition (9.37) and Lemma 9.9, we may and shall assume that $\zeta = \tau$. We apply Lemma 9.13 to write $\Omega_j(x,y,\tau/\mathrm{h})$ as

$$\sum_{\substack{\alpha',\alpha'',\beta',\beta'',\gamma: \\ |\alpha'|+|\alpha''|+|\beta'|+|\beta''|-|\gamma|=j, \\ |\gamma| \leqslant \min(|\alpha'|+|\alpha''|, |\beta'|+|\beta''|), \\ |\alpha'|+|\beta'| \geqslant 2|\gamma|}} c_{\alpha'\alpha''\beta'\beta''\gamma} (x')^{\alpha'} (x'')^{\alpha''} (y')^{\beta'} (y'')^{\beta''} (\tau/\mathrm{h})^\gamma,$$

with coordinates as in the formulation of that Lemma. Using the analyticity of $\{x,y\}$ and the estimate $\{x,y\} \ll |x| \cdot |y|$, we see that $c_{\alpha'\alpha''\beta'\beta''\gamma} \ll R^j$ for some fixed $R > 0$ (with the implied constant independent of j). We claim that for each term as in the above sum and all $x, y = \mathrm{O}(1)$, we have

$$(x')^{\alpha'} (x'')^{\alpha''} (y')^{\beta'} (y'')^{\beta''} (\tau/\mathrm{h})^\gamma \ll \rho^j \tag{9.39}$$

where we abbreviate

$$\rho := \max(|x'|, |y'|, |x''|, |y''|, |x'|^2/\mathrm{h}, |y'|^2/\mathrm{h}).$$

Indeed, since $|\tau| \asymp 1$, the LHS of (9.39) is majorized by

$$\max(|x'|,|y'|)^{|\alpha'|+|\beta'|-2|\gamma|}\max(|x''|,|y''|)^{|\alpha''|+|\beta''|}(\max(|x'|,|y'|)^2/\mathrm{h})^{|\gamma|}$$

which is bounded in turn, thanks to the inequality $|\alpha'|+|\beta'|-2|\gamma| \geqslant 0$, by

$$\rho^{|\alpha'|+|\beta'|-2|\gamma|}\rho^{|\alpha''|+|\beta''|}\rho^{|\gamma|} = \rho^j,$$

as required. Bounding the number of relevant multi-indices crudely by $(1+j)^{\mathrm{O}(1)}$, we deduce that

$$\Omega_j(x,y,\tau/\mathrm{h}) \ll (1+j)^{\mathrm{O}(1)}(R\rho)^j.$$

We now argue as in [NV, §7.4] – using the trivial bounds $\Omega(x,y,\tau/\mathrm{h}) \ll 1$, $\Omega_j(x,y,\tau/\mathrm{h}) \ll_j 1$ when $\rho \geqslant 1/2R$ and summing over $j \geqslant J$ when $\rho \leqslant 1/2R$ – to see that

$$\Omega^{(J)}(x,y,\tau/\mathrm{h}) \ll \rho^J.$$

Our task thereby reduces to verifying that if J is fixed but large enough in terms of N, then

$$\int_{x,y} |a_{\mathrm{h}}^{\vee}(x)b_{\mathrm{h}}^{\vee}(y)|\rho^J\,dx\,dy \ll \mathrm{h}^N. \tag{9.40}$$

The moment bound (9.35) gives

$$\int_{x,y} |a_{\mathrm{h}}^{\vee}(x)b_{\mathrm{h}}^{\vee}(y)|\rho^J\,dx\,dy \ll \max\left((A')^{-1},(A'')^{-1},(A')^{-2}/\mathrm{h}\right)^J$$
$$= \max\left(\mathrm{h}^{1-\delta'},\mathrm{h}^{1-\delta''},\mathrm{h}^{1-2\delta'}\right)^J.$$

Since $\delta' < \delta'' < 2\delta' < 1$, this last expression is indeed $\leqslant \mathrm{h}^N$ for large enough J. The proof of (9.24) is thus complete. □

9.5.4. *Integration by parts.* We postponed in the above proof the following application of integration by parts.

Lemma 9.14. *Let* $\mathrm{h} \lll 1$, *let* δ', δ'' *be fixed quantities satisfying* (9.17), *let* $\varepsilon > 0$ *be fixed but small enough in terms of* δ', δ'', *let* τ *belong to a fixed compact subset of* $\mathfrak{g}^{\wedge}_{\mathrm{reg}}$, *let* $\xi, \eta \in \mathfrak{g}^{\wedge}$ *with* $\xi = \tau + \mathrm{O}(\mathrm{h}^{\delta''-\delta'})$, *and let* $\zeta \in \mathfrak{g}^{\wedge}$ *be of the form* $\mathrm{O}(\mathrm{h}^{-\varepsilon})$. *Suppose moreover that*

$$\zeta - \tau \ggg \mathrm{h}^{\delta''-\delta'}.$$

Set

$$B' := \mathrm{h}^{1-\delta'-\varepsilon} \leqslant B'' := \mathrm{h}^{1-\delta''-\varepsilon} \lll 1.$$

Let $f \in C_c^{\infty}(\mathfrak{g} \times \mathfrak{g})$ *satisfy the support condition (in* τ*-coordinates)*

$$f(x,y) \neq 0 \quad \Longrightarrow \quad x', y' \ll B', \quad x'', y'' \ll B''$$

and the derivative bounds

$$\partial_{x'}^{\alpha'}\partial_{x''}^{\alpha''}\partial_{y'}^{\beta'}\partial_{y''}^{\beta''} f(x,y) \ll (B')^{-|\alpha'|-|\beta'|}(B'')^{-|\alpha''|-|\beta''|}$$

for all fixed multi-indices $\alpha', \alpha'', \beta', \beta''$. *Define the phase function*

$$\phi(x,y,\zeta) := x(\zeta-\xi) + y(\zeta-\eta) + \{x,y\}\zeta \in i\mathbb{R}.$$

Then for each fixed $\gamma \in \mathbb{Z}_{\geqslant 0}^{\dim(\mathfrak{g})}$ *and* $N \geqslant 0$, *we have*

$$\partial_{\zeta}^{\gamma} \int_{x,y} f(x,y)e^{\phi(x,y,\zeta)/\mathrm{h}}\,dx\,dy \ll \mathrm{h}^N \tag{9.41}$$

The proof of Lemma 9.14 invokes the following useful lemma from [BKY, §8].

Lemma 9.15. *Let $Y \geqslant 1$, $X, Q, U, R > 0$, and suppose that w is a smooth function with support on $[\alpha, \beta]$, satisfying*

$$w^{(j)}(t) \ll_j XU^{-j}.$$

Suppose Φ is a smooth real-valued function on $[\alpha, \beta]$ such that

$$|\Phi'(t)| \geqslant R$$

for some $R > 0$, and

$$\Phi^{(j)}(t) \ll_j YQ^{-j}, \qquad \text{for } j = 2, 3, \dots.$$

Then

$$\int_{t \in \mathbb{R}} w(t) e^{i\Phi(t)} \, dt \ll_A (\beta - \alpha) X[(QR/\sqrt{Y})^{-A} + (RU)^{-A}].$$

Proof of Lemma 9.14. We may and shall reduce to the case $\gamma = 0$ by absorbing factors such as $\{x, y\}^\gamma$ into the weight function f.

At least one of the following possibilities occurs:

(i) $|\zeta' - \tau'| \gg |\zeta - \tau|$.
(ii) $|\zeta'' - \tau''| \gg |\zeta - \tau|$.

Consider first case (i). We may then choose a unit speed one-parameter subgroup $t \mapsto z_t'$ of $\mathfrak{g}_\tau^\flat$ so that $\partial_t z_t'(\zeta - \tau) \gg |\zeta - \tau|$. We extend the unit vector $e_1 := \partial_t z_t'$ to an orthonormal basis $e_1, \dots, e_{\dim(\mathfrak{g})}$ of $\mathfrak{g}$ and set $W := \oplus_{j \geqslant 2} \mathbb{R} e_j$. The LHS of (9.41) may then be written as the iterated integral

$$\int_{x \in W, y \in \mathfrak{g}} \int_{t \in \mathbb{R}} f(x + z_t', y) e^{\phi(x + z_t', y, \zeta)/\mathrm{h}} \, dt \, dx \, dy.$$

It suffices to show that the inner integral over t is $\ll \mathrm{h}^N$ for all x, y in the support of the outer integral. We accordingly focus on individual $x, y \in \mathfrak{g}$ satisfying

$$x', y' \ll B' \text{ and } x'', y'' \ll B'' \tag{9.42}$$

and set

$$w(t) := f(x + z_t', y), \quad \Phi(t) := \phi(x + z_t', y, \zeta)/i\,\mathrm{h}\,.$$

We aim to show that

$$\int_{t \in \mathbb{R}} w(t) e^{i\Phi(t)} \, dt \ll \mathrm{h}^N. \tag{9.43}$$

We will do so via Lemma 9.15. To apply that lemma, we must estimate the derivatives of Φ. To that end, we claim first that for $t \ll B'$ (equivalently, $z_t' \ll B'$), we have

$$\partial_t \{x + z_t', y\} \zeta \ll B' |\zeta| + B'' |\zeta - \tau|. \tag{9.44}$$

To verify (9.44), we expand the analytic function $\{,\}$ as a Taylor series. From our assumed upper bounds on x, y, ζ, we see that the error incurred in $\partial_t \{x + z_t', y\} \zeta$ by replacing $\{,\}$ with the first J homogeneous components of its Taylor series is $\mathrm{O}(\mathrm{h}^N)$, for any fixed N, provided that J is fixed but large enough in terms of N. Since our hypotheses imply that the RHS of (9.44) is bounded from below by some fixed power of h, it is thus enough to verify for each fixed iterated Lie monomial L of degree at least two that

$$\partial_t L(z_t', y) \zeta \ll B' |\zeta| + B'' |\zeta - \tau|. \tag{9.45}$$

We verify this first for L the commutator bracket $[,]$. By linearity, we may consider separately the cases $y = y'$ and $y = y''$. In the first case, we have $\partial_t [z_t', y'] \ll B'$, which gives the adequate estimate $\partial_t[z_t', y]\zeta \ll B'|\zeta|$. In the second case, the element y'' centralizes τ, hence

$$\begin{aligned}\langle [z_t', y''], \zeta\rangle &= \langle z_t', [y'', \zeta]\rangle \\ &= \langle z_t', [y'', \zeta - \tau]\rangle \\ &= \langle [z_t', y''], \zeta - \tau\rangle,\end{aligned}$$

giving again the adequate estimate $\partial_t[z_t', y'']\zeta \ll B''|\zeta - \tau|$. This completes the proof of (9.45) for L of degree two. We argue similarly for L of higher degree. For instance, in the case of degree three, we obtain the adequate estimates

- $\partial_t[z_t', [z_t', y']]\zeta \ll (B')^2|\zeta|$,
- $\partial_t[y_1', [z_t', y_2']]\zeta \ll (B')^2|\zeta|$,
- $\partial_t[z_t', [z_t', y'']]\zeta \ll B'B''|\zeta|$, and
- $\partial_t[y_1'', [z_t', y_2'']]\zeta \ll (B'')^2|\zeta - \tau|$
- $\partial_t[y'', [z_t', y']]\zeta \ll B'B''|\zeta|$.

for $y', y_1', y_2' \ll B'$ in $\mathfrak{g}_\tau^\flat$ and $y'', y_1'', y_2'' \ll B''$ in $\mathfrak{g}_\tau$. In general, if our monomial features at least one y_j' or at least two z_t', then we obtain an adequate estimate by taking into account the sizes of z_t', y_j' and $|\zeta|$. Otherwise, our monomial features one z_t', many y_j'' and no y_j'. We then use that each y_j'' centralizes τ to replace ζ by $\zeta - \tau$ and argue as in the case $L(x, y) = [x, y]$. This completes the proof of (9.45), hence that of (9.44).

We observe next that

$$\Phi'(t) \gg |\zeta - \tau| / \mathrm{h}. \tag{9.46}$$

To see this, we first rewrite the definition in the form

$$i\,\mathrm{h}\,\Phi'(t) = \partial_t z_t'(\zeta - \tau) + \partial_t z_t'(\tau - \xi) + \partial_t\{x + z_t', y\}\zeta. \tag{9.47}$$

By the construction of z_t', the first term on the RHS of (9.47) is $\gg |\zeta - \tau|$. The second term is

$$\ll \mathrm{h}^{\delta'' - \delta'} \lll |\zeta - \tau|.$$

The third term is

$$\ll B'|\zeta| + B''|\zeta - \tau| \lll |\zeta - \tau|$$

by (9.44) and our hypothesis $|\zeta| \ll \mathrm{h}^{-\varepsilon}$. This completes the verification of (9.46).

We now apply Lemma 9.15 with

$$X = 1, U = B', \quad R \asymp |\zeta - \tau| / \mathrm{h}, \quad Y = \frac{B''|\zeta|}{\mathrm{h}} \ll B''\,\mathrm{h}^{-1+\varepsilon}, Q = 1. \tag{9.48}$$

By (9.46), we may find such an R with $|\Phi'(t)| \geqslant R$. We note that for fixed $j \geqslant 2$, we have $\Phi^{(j)}(t) \ll Y$, or equivalently, $\partial_t^j\{x + z_t', y\}\zeta \ll B''|\zeta|$. To see this, we expand $\{,\}$ as above in a Taylor series consisting of iterated Lie monomials and use that $y \ll B' + B'' \asymp B'' \leqslant 1$. We note also that w is a smooth function on $\mathbb{R}$ that is supported on an interval $[\alpha, \beta]$ with $\alpha, \beta \ll B'$ and satisfies the derivative bounds $w^{(j)}(t) \ll_j (B')^{-j}$. The hypotheses of Lemma 9.15 are thus satisfied. We observe that

$$RU \gg \mathrm{h}^{\delta'' - 2\delta' - \varepsilon} \geqslant \mathrm{h}^{-\varepsilon}$$

and

$$\frac{(QR)^2}{Y} \gg \frac{(\mathrm{h}^{\delta'' - \delta'} / \mathrm{h})^2}{B''\,\mathrm{h}^{-1-\varepsilon}} = \mathrm{h}^{2\varepsilon - 2(1-\delta') - (2\delta' - \delta'')} \geqslant \mathrm{h}^{-\varepsilon}.$$

Thus the required estimate (9.43) follows from the conclusion of Lemma 9.15 upon taking A sufficiently large in terms of ε and N. This completes the proof of Lemma 9.14 in case (i).

The proof in case (ii) is similar but slightly simpler. We choose a unit speed one-parameter subgroup $t \mapsto z''_t$ of $\mathfrak{g}_\tau$ so that $\partial_t z''_t(\zeta - \tau) \gg |\zeta - \tau|$. The estimate (9.44) remains valid for $t \ll B''$, with essentially the same proof: we reduce to estimates for monomials as before and argue separately according as some y'_j appears (in which case we get the bound $\ll B'|\zeta|$) or none appears (in which case we obtain $\ll B''|\zeta - \tau|$). (The second case could even be eliminated altogether by exploiting the fact that $\mathfrak{g}_\tau$ is abelian in our setup.) We make the same choices (9.48) as before (improving U to B'' if desired) and conclude once again via Lemma 9.15. $\square$

9.5.5. *The case of disjoint supports.* We have noted following the statement of Theorem 9.12 that in order to complete the proof of (9.22), it suffices to show for

$$\tau_1 - \tau_2 \ggg \mathrm{h}^{\delta'' - \delta'} \tag{9.49}$$

that $a \star_{\mathrm{h}} b \in \mathrm{h}^\infty S^{-\infty}$ for $(a, b) \in S^{\tau_1}_{\delta',\delta''} \times S^{\tau_2}_{\delta',\delta''}$. We must check that $\partial_\zeta^\gamma (a \star_{\mathrm{h}} b)(\zeta) \ll \mathrm{h}^N \langle \zeta \rangle^{-N}$ for fixed γ, N and all $\zeta \in \mathfrak{g}^\wedge$. To see this, we reduce as in §9.5.3 to the case that a, b satisfy the support conditions (9.32) for some $\omega_j = \tau_j + \mathrm{O}(\mathrm{h}^{\delta''-\delta'})$ and that $|\zeta| < Q^\varepsilon$. We see then by (9.49) that $|\zeta - \tau_j| \ggg \mathrm{h}^{\delta''-\delta'}$ for some $j = 1, 2$. If $j = 1$, then we conclude by applying the argument following (9.37) verbatim. If $j = 2$, then we apply the same argument but with the roles of a and b reversed.

9.5.6. *The mixed case.* Here we discuss the proof of part (iv) of Theorem 9.12. We have noted already that the proof amounts to an interpolation between the proofs of part (iii) and of [NV, §7.3], so we will be brief.

As in the proof of part (iii), it is enough to verify

- the claimed mapping properties for $\star^j$, and
- the claimed asymptotic expansion for $a \star_{\mathrm{h}} b$, where (say) $a \in S^\tau_{\delta',\delta''}$ and $b \in S^m_\delta$ (the same argument applies with the roles of a and b reversed).

The former may be verified exactly as in §9.5.2. For the latter, we reduce as in §9.5.3 to verifying that the remainder $r := a \star_{\mathrm{h}} b - \sum_{0 \leqslant j < J} \mathrm{h}^j \, a \star^j b$ enjoys the estimate $\partial_\zeta^\gamma r(\zeta) \ll \mathrm{h}^N \langle \zeta \rangle^{-N}$, as in (9.30), provided that J is fixed large enough in terms of N. By decomposing our symbols into localized pieces, we may reduce as in [NV, §7.7] and §9.5.3 to the case that

- a satisfies the support condition (9.32) for some $\omega_1 = \tau + \mathrm{O}(\mathrm{h}^{\delta''-\delta'})$, while
- b is supported on $\omega_2 + \mathrm{O}(\mathrm{h}^\delta \langle \omega_2 \rangle)$ for some $\omega_2 \in \mathfrak{g}^\wedge$.

We then have with (A', A'', m, n) as in (9.33) and $A := \mathrm{h}^{-1+\delta} \langle \omega_2 \rangle$ the moment bound

$$\int_{x,y \in \mathfrak{g}} |a_{\mathrm{h}}^\vee(x) b_{\mathrm{h}}^\vee(y)| \, |x'|^{k'} |x''|^{k''} |y'|^{l'} |y''|^{l''} \, dx \, dy \ll (A')^{-k'} (A'')^{-k''} A^{-l'-l''}, \tag{9.50}$$

with notation as in (9.35). We set $Q := \mathrm{h}^{-1} \langle \zeta \rangle \langle \omega_2 \rangle$. As in [NV, §7.7], the required estimate follows by trivially estimating the integral representation (9.1) unless $|\omega_2| \leqslant Q^\varepsilon$. We reduce further to the case $|\zeta| \geqslant Q^{2\varepsilon}$ by integrating by parts (crudely) as in [NV, §7.7]. We reduce further to the case $\zeta = \tau + \mathrm{O}(\mathrm{h}^{\delta''-\delta'})$ by the same argument as in the reduction to (9.37), noting that Lemma 9.14 imposes no constraints on the variable η. Arguing exactly as in §9.5.3, we reduce finally to verifying the estimate (9.40). For this we appeal to the moment bound (9.50) and the fact that $\max(2\delta, 2\delta', \delta'') < 1$.

10. Operators

Let π be a unitary representation of the fixed Lie group G.

10.1. Spaces of operators. We set $\Delta := 1 - \sum_{x \in \mathcal{B}(\mathfrak{g})} x^2 \in \mathfrak{U}(\mathfrak{g})$ for some fixed basis $\mathcal{B}(\mathfrak{g})$ of $\mathfrak{g}$. We write more verbosely Δ_G when we wish to indicate which group is being considered. As discussed in [NV, §3.1], $\pi(\Delta)$ is a densely-defined self-adjoint positive operator, with bounded inverse having operator norm $\leqslant 1$.

We write π^∞ for the space of smooth vectors. For $s \in \mathbb{Z}$, we denote by π^s the Hilbert space completion of π^∞ with respect to the inner product $\langle v_1, v_2 \rangle_{\pi^s} := \langle \pi(\Delta)^s v_1, v_2 \rangle$. Up to natural identifications,

$$\pi^\infty = \cap \pi^s \leqslant \cdots \leqslant \pi^{s+1} \leqslant \pi^s \leqslant \pi^{s-1} \leqslant \cdots .$$

The space $\pi^{-\infty}$ of distributional vectors is defined to be the union of the spaces π^s. For future reference, we record [NV, (3.1)]: for fixed $s \in \mathbb{Z}_{\geqslant 0}$,

$$\|v\|_{\pi^s}^2 \asymp \sum_{r=0}^{s} \sum_{x_1, \ldots, x_r \in \mathcal{B}(\mathfrak{g})} \|\pi(x_1 \cdots x_r) v\|^2. \tag{10.1}$$

By an *operator* on π, we mean a linear map $T : \pi^\infty \to \pi^{-\infty}$. We denote by $\underline{\Psi}^m$ the space of operators T such that for each $s \in \mathbb{Z}$, $n \in \mathbb{Z}_{\geqslant 0}$ and $x_1, \ldots, x_n \in \mathfrak{g}$, the following iterated commutator defines a bounded map between the indicated Hilbert spaces:

$$[\pi(x_1), [\pi(x_2), [\ldots, [\pi(x_n), T] \cdots]]] : \pi^s \to \pi^{s-m}. \tag{10.2}$$

We extend the definition to $m = \pm\infty$ by taking the union or intersection over all integers m. We write more verbosely $\underline{\Psi}^m(\pi)$ when we wish to indicate the representation under consideration. As in §9.1, for $m < \infty$, the space $\underline{\Psi}^m$ is a Frechet space equipped with a distinguished family of seminorms, while $\underline{\Psi}^\infty$ is an inductive limit of such spaces. We refer to [NV, §3] for general discussion of the spaces $\underline{\Psi}^m$, parts of which will be recalled below as needed.

10.2. Operator classes.

Definition 10.1. For fixed $m \in \mathbb{Z}$ and $\delta \in [0, 1)$, we write[4]

$$\Psi_\delta^m$$

for the class (§3.1.3) of all $T \in \underline{\Psi}^m$ such that for each fixed s, n and $x_1, \ldots, x_n$, the map (10.2) has operator norm $\mathrm{O}(\mathrm{h}^{-\delta n})$. We extend the definition to $m = \pm\infty$ as in Definition 9.2, thus Ψ_δ^∞ (resp. $\Psi_\delta^{-\infty}$) is the union (resp. intersection) of Ψ_δ^m (resp. $\underline{\Psi}^{-\infty} \cap \Psi_\delta^m$) over fixed integers m. We write more verbosely $\Psi_\delta^m(\pi)$ when we wish to indicate the representation under consideration.

[4]The notation here differs from that in [NV, §3]. There, we used "Ψ^m" in place of $\underline{\Psi}^m$ and "Ψ_δ^m" for the space of h-dependent operators whose norms are bounded in the indicated manner. The class Ψ_δ^m defined here plays a similar role. The relationship between these definitions is as described in the proof of Theorem 9.5.

As in §9.3, we may define for a positive real c the class $c\Psi_\delta^m$, and we denote by $\mathrm{h}^\infty \Psi_\delta^m$ the intersection of $\mathrm{h}^N \Psi_\delta^m$ over all fixed N. The classes $\mathrm{h}^\infty \Psi_\delta^m$ are independent of δ and will be denoted simply by $\mathrm{h}^\infty \Psi^m$.

Lemma 10.2. *Composition of operators induces, for each fixed m_1, m_2,*

$$\Psi_\delta^{m_1} \times \Psi_\delta^{m_2} \to \Psi_\delta^{m_1+m_2},$$

with the convention $\infty + (-\infty) := -\infty$.

Proof. This follows from the proof of [NV, §3.4]. Indeed, for $(T_1, T_2) \in \Psi_\delta^{m_1} \times \Psi_\delta^{m_2}$ and fixed $x_1, \dots, x_n \in \mathfrak{g}$, we may write $[\pi(x_1), \dots, [\pi(x_n), T_1 T_2]]$ as a linear combination of compositions $T_1' T_2'$, where $T_1' = [\pi(y_1), \dots, [\pi(y_{n_1}), T_1]]$ and $T_2' = [\pi(z_1), \dots, [\pi(z_{n_1}), T_2]]$ with $n_1 + n_2 = n$ and $y_i, z_j \in \mathfrak{g}$ fixed. We must check that the composition $T_1' T_2' : \pi^s \to \pi^{s-m}$ has operator norm $\mathrm{O}(\mathrm{h}^{-\delta n})$. To do so, we factor that composition as $\pi^s \xrightarrow{T_2'} \pi^{s-m_2} \xrightarrow{T_1'} \pi^{s-m}$ and apply our hypotheses to each factor. □

Lemma 10.3. *For fixed $m \in \mathbb{Z}$,*

$$\pi(\Delta)^m \in \Psi_0^{2m}. \tag{10.3}$$

For fixed $x_1, \dots, x_m \in \mathfrak{g}$,

$$\pi(x_1 \cdots x_m) \in \Psi_0^m. \tag{10.4}$$

Proof. This follows from [NV, §3.5]. □

Using the self-adjointness of $\pi(\Delta)$, it is straightforward to check that Ψ^m is preserved under taking adjoints.

10.3. Operators attached to symbols. Recall that $\underline{S}^{-\infty}$ is the Schwartz space on $\mathfrak{g}^\wedge$. As explained in [NV, §5.1], the assignment

$$\begin{aligned} \mathrm{Op} : \underline{S}^{-\infty} &\to \{\text{operators on } \pi\} \\ a &\mapsto \mathrm{Op}(a) := \mathrm{Op}_{\mathrm{h}}(a : \pi, \chi), \end{aligned} \tag{10.5}$$

defined in §8 with respect to some nice cutoff χ, extends naturally to the space of symbols $\underline{S}^\infty$. In particular, it extends to every symbol class defined in §9. The extension may be characterized in terms of matrix coefficients: for smooth vectors $u, v \in \pi$,

$$\langle \mathrm{Op}(a) u, v \rangle = \int_{\xi \in \mathfrak{g}^\wedge} a(\mathrm{h}\, \xi) \left(\int_{x \in \mathfrak{g}} e^{-x\xi} \chi(x) \langle \pi(\exp(x)) u, v \rangle \, dx \right) d\xi.$$

Note that the parenthetical integral over x defines a Schwartz function of ξ, so the remaining integral over ξ converges absolutely.

Each polynomial function $p : \mathfrak{g}^\wedge \to \mathbb{C}$ (equivalently, element p of the symmetric algebra $\mathrm{Sym}(\mathfrak{g}_\mathbb{C})$) defines an element of $\underline{S}^\infty$. The image of such an element under Op is described as follows.

Lemma 10.4. *For each $p \in \mathrm{Sym}(\mathfrak{g}_\mathbb{C}) \subseteq \underline{S}^\infty$, we have*

$$\mathrm{Op}(p) = \pi(\mathrm{sym}(p_{\mathrm{h}})),$$

where p_{h} denotes the rescaling $p_{\mathrm{h}}(\xi) := p(\mathrm{h}\, \xi)$ and $\mathrm{sym} : \mathrm{Sym}(\mathfrak{g}_\mathbb{C}) \to \mathfrak{U}(\mathfrak{g}_\mathbb{C})$ denotes the symmetrization map, i.e., the linear isomorphism sending each monomial to the average of its permutations.

Proof. See [NV, §5.2, §8.1]. □

For applications of Lemma, it is useful to note 10.4 that $\mathrm{sym}(p^k) = \mathrm{sym}(p)^k$ for $k \in \mathbb{Z}_{\geqslant 0}$. Moreover, for $x \in \mathfrak{g}$ – regarded as a linear function $x : \mathfrak{g}^\wedge \to i\mathbb{R}$, or equivalently, as a degree one element of $\mathrm{Sym}(\mathfrak{g}_\mathbb{C})$ – we have $\mathrm{sym}(x_\mathrm{h}) = \mathrm{h}\, x$ (regarded as an element of $\mathfrak{U}(\mathfrak{g})$), hence $\mathrm{Op}(x) = \mathrm{h}\, \pi(x)$.

Throughout this section, we retain the abbreviation (10.5). Our aim is to recall from [NV] the effect of Op on the basic symbol classes $\underline{S}^m$ and S^m_δ and to describe its effect on the refined classes $S^\tau_{\delta',\delta''}$.

Lemma 10.5. *For any nice cutoffs χ_1, χ_2 and any $a \in \underline{S}^\infty$, we have*

$$\mathrm{Op}_\mathrm{h}(a; \pi, \chi_1) - \mathrm{Op}_\mathrm{h}(a; \pi, \chi_2) \in \underline{\Psi}^{-\infty}.$$

This difference defines a continuous map $\underline{S}^\infty \to \underline{\Psi}^{-\infty}$.

Theorem 10.6. *We have*

$$\mathrm{Op}(\underline{S}^m) \subseteq \underline{\Psi}^m,$$

and the induced map is continuous. In particular, elements of $\mathrm{Op}(\underline{S}^\infty)$ act on π^∞, and so may be composed. For $(a, b) \in \underline{S}^{m_1} \times \underline{S}^{m_2}$, the composition formula (8.4) remains valid, and we have

$$\mathrm{Op}(a)\,\mathrm{Op}(b) \equiv \mathrm{Op}(a \star_\mathrm{h} b) \mod \underline{\Psi}^{-\infty}.$$

Proof of Lemma 10.5 and Theorem 10.6. The rescaling map $a \mapsto a_\mathrm{h}$ defines a topological automorphism of $\underline{S}^m$, so it suffices to consider the case $\mathrm{h} = 1$. The required conclusions are given then by [NV, §5.4] and [NV, Thm 2]. □

Lemma 10.7. *For fixed nice cutoffs χ_1 and χ_2, fixed $\delta \in [0, 1)$ and any $a \in S^\infty_\delta$, we have*

$$\mathrm{Op}_\mathrm{h}(a; \pi, \chi_1) - \mathrm{Op}_\mathrm{h}(a; \pi, \chi_2) \in \mathrm{h}^\infty\, \Psi^{-\infty}. \tag{10.6}$$

Theorem 10.8. *Fix a nice cutoff χ. Fix $\delta \in [0, 1/2)$. For $m \in \mathbb{Z}$, we have*

$$\mathrm{Op}(S^m_\delta) \subseteq \mathrm{h}^{\min(m,0)}\, \Psi^m_\delta. \tag{10.7}$$

For $a, b \in S^\infty_\delta$, we have

$$\mathrm{Op}(a)\,\mathrm{Op}(b) \equiv \mathrm{Op}(a \star_\mathrm{h} b) \mod \mathrm{h}^\infty\, \Psi^{-\infty} \tag{10.8}$$

For each fixed $m_1, m_2 < \infty$ and $N \in \mathbb{Z}_{\geqslant 0}$ there is a fixed $J \in \mathbb{Z}_{\geqslant 0}$ so that

$$\mathrm{Op}(a)\,\mathrm{Op}(b) \equiv \sum_{0 \leqslant j < J} \mathrm{h}^j\, \mathrm{Op}(a \star^j b) \mod \mathrm{h}^N\, \Psi^{-N}_\delta. \tag{10.9}$$

Proof of Lemma 10.7 and Theorem 10.8. We appeal to [NV, §5.4] and [NV, §5.6], translating to the present formulation as in the proof of Theorem 9.5. □

We turn to the main new result of this section.

Theorem 10.9. *Fix a connected real reductive group G and a nice cutoff χ. Let $\mathrm{h}, \delta', \delta'', \tau$ be as in Definition 9.8. Then for each fixed $m \in \mathbb{Z}$,*

$$\mathrm{Op}(S^\tau_{\delta',\delta''}) \subseteq \mathrm{h}^m\, \Psi^m_{\delta'}. \tag{10.10}$$

In particular, $\mathrm{Op}(S^\tau_{\delta',\delta''}) \subseteq \Psi^0_{\delta'}$. Moreover, the composition formulas (10.8) and (10.9) remain valid for $a, b \in S^\infty_{\delta'} \cup S^\tau_{\delta',\delta''}$.

Proof. Using the refined star product asymptotics afforded by Theorem 9.12, we may complete the proof of Theorem 10.9 exactly as in [NV, §8.6]. For the sake of completeness, we record the details here. The proof occupies the remainder of §10.3.

We verify first that the composition formula (10.8) remains valid for $a, b \in S^{\infty}_{\delta'} \cup S^{\tau}_{\delta',\delta''}$. Since in particular $a, b \in \underline{S}^{\infty}$, we have by Theorem 10.6 that the composition formula (8.4) remains valid. On the other hand, we see from the star product asymptotics given in Theorem 9.12 and the crude inclusion $S^{\tau}_{\delta',\delta''} \subseteq S^{-\infty}_{\delta''}$ that $a \star_{\mathrm{h}} b \in S^{\infty}_{\delta''}$. By Lemma 10.7, it follows that

$$\mathrm{Op}_{\mathrm{h}}(a \star_{\mathrm{h}} b : \pi, \chi') \equiv \mathrm{Op}(a \star_{\mathrm{h}} b) \mod \mathrm{h}^{\infty} \Psi^{-\infty}.$$

The required formula (10.8) then follows from (8.4).

For the proof of the remaining assertions, a key step is to verify the following consequence of (10.10).

Lemma 10.10. *Let $\|.\|$ denote the operator norm on the space of linear maps $\pi \to \pi$. For $a \in S^{\tau}_{\delta',\delta''}$, we have*

$$\| \mathrm{Op}(a)\| \ll 1.$$

Proof of Lemma 10.10. Let $\mathcal{N}$ denote the norm on the Schwartz space $\mathcal{S}(\mathfrak{g}^{\wedge})$ given by the L^1-norm of the Fourier transform: $\mathcal{N}(a) := \|a^{\vee}\|_{L^1(\mathfrak{g})}$. This norm is dilation-invariant: $\mathcal{N}(a) = \mathcal{N}(a_{\mathrm{h}})$. It follows readily from the definition of Op that $\| \mathrm{Op}(a)\| \leqslant \mathcal{N}(a)$. We have the crude bound

$$\mathcal{N}(a) \ll \mathrm{h}^{-\mathrm{O}(1)} \tag{10.11}$$

where the implied constant depends at most upon $\dim(\mathfrak{g})$ and (δ', δ''). While $\mathcal{N}(a)$ can be quite large (e.g., a fixed positive power of h^{-1}) for elements a of $S^{\tau}_{\delta',\delta''}$, we will see below that any such element may be decomposed into "almost-orthogonal" pieces on which $\mathcal{N}$ is $\mathrm{O}(1)$. We will then conclude by the Cotlar–Stein lemma [Hör, Lem 18.6.5], which we recall here:

- *Let V_1, V_2 be Hilbert spaces. Let $T_j : V_1 \to V_2$ be a sequence of bounded linear operators. Assume that*

$$\sup_j \sum_k \|T_j^* T_k\|^{1/2} \leqslant C, \quad \sup_j \sum_k \|T_j T_k^*\|^{1/2} \leqslant C, \tag{10.12}$$

  Then the series $T := \sum T_j$ converges in the Banach space of bounded linear operators from V_1 to V_2, and has operator norm $\|T\| \leqslant C$.

Let $a \in S^{\tau}_{\delta',\delta''}$. Recall that a is supported on elements of the form $\tau + \mathrm{O}(\mathrm{h}^{\delta''-\delta'})$. For each element ω of that form, denote by $\mathcal{D}(\omega)$ the rectangular domain

$$\mathcal{D}(\omega) := \left\{ \xi \in \mathfrak{g}^{\wedge} : |\xi' - \omega'| \leqslant \mathrm{h}^{\delta'}, |\xi'' - \omega''| \leqslant \mathrm{h}^{\delta''} \right\}.$$

Note that, by the definition of $S^{\tau}_{\delta',\delta''}$ (Definition 9.8), the variation of a on each such domain is mild. We say that a is *localized at* ω if it is supported on $\mathcal{D}(\omega)$. This notion appeared implicitly in §9.5.3, before the estimate (9.5.3). By the one-dimensional analogue of that estimate, we see that if a is localized at ω, then $\mathcal{N}(a) \ll 1$.

By taking a suitable smooth partition of unity, we may decompose any element $a \in S^{\tau}_{\delta',\delta''}$ as a finite sum $a = \sum_{\omega \in \Omega} a_{\omega}$, where

- Ω is a finite subset of $\mathfrak{g}^\wedge$, consisting of elements ω of the form $\tau + \mathrm{O}(\mathrm{h}^{\delta'' - \delta'})$, of cardinality $\#\Omega \ll \mathrm{h}^{-\mathrm{O}(1)}$,
- the domains $\mathcal{D}(\omega)$ $(\omega \in \Omega)$ have uniformly bounded overlaps in the sense that
$$\sup_{\omega_1 \in \Omega} \# \{\omega_2 \in \Omega : \mathcal{D}(\omega_1) \cap \mathcal{D}(\omega_2) \neq \emptyset\} \ll 1, \tag{10.13}$$
and
- each summand a_ω lies in $S^\tau_{\delta', \delta''}$ and is localized at the corresponding element ω.

Turning to the proof of the lemma, let $a \in S^\tau_{\delta',\delta''}$. We decompose $a = \sum a_\omega$ as above. Then $\mathrm{Op}(a) = \sum \mathrm{Op}(a_\omega)$ and $\mathrm{Op}(a_\omega)^* = \mathrm{Op}(\bar{a}_\omega)$ (cf. §8.6). By the composition formula (8.4) for Schwartz functions, we see that
$$\| \mathrm{Op}(a_{\omega_1})^* \mathrm{Op}(a_{\omega_2})\| \leqslant \mathcal{N}(\bar{a}_{\omega_1} \star_{\mathrm{h}} a_{\omega_2}).$$
By the Cotlar–Stein lemma, we reduce to verifying that
$$\sup_{\omega_1 \in \Omega} \sum_{\omega_2 \in \Omega} \mathcal{N}(\bar{a}_{\omega_1} \star_{\mathrm{h}} a_{\omega_2})^{1/2} \ll 1$$
(together with the analogous estimate in which complex conjugation is applied instead to a_{ω_2}). To see this, we fix $N \in \mathbb{Z}_{\geqslant 0}$ sufficiently large and evaluate $\bar{a}_{\omega_1} \star_{\mathrm{h}} a_{\omega_2}$ using the star product asymptotics (9.24) as the sum $b_{\omega_1,\omega_2} + r_{\omega_1,\omega_2}$, where
$$b_{\omega_1,\omega_2} := \sum_{0 \leqslant j < J} \mathrm{h}^j \, \bar{a}_{\omega_j} \star^j a_{\omega_2}$$
denotes the result of truncating the asymptotic expansion up to some fixed index J, taken large enough in terms of N, and
$$r_{\omega_1,\omega_2} \in \mathrm{h}^{(1-2\delta')J} \, S^\tau_{\delta',\delta''} + \mathrm{h}^\infty \, S^{-\infty}$$
denotes the remainder. By the triangle inequality for $\mathcal{N}$ and the bound $\sqrt{x+y} \ll \sqrt{x} + \sqrt{y}$ for $x, y \geqslant 0$, we reduce to verifying that each of the quantities
$$\sup_{\omega_1 \in \Omega} \sum_{\omega_2 \in \Omega} \mathcal{N}(b_{\omega_1,\omega_2})^{1/2}, \quad \sup_{\omega_1 \in \Omega} \sum_{\omega_2 \in \Omega} \mathcal{N}(r_{\omega_1,\omega_2})^{1/2}$$
is $\mathrm{O}(1)$. Since $1 - 2\delta' > 0$, the crude bound (10.11) and our choice of J yield the strong estimate $\mathcal{N}(r_{\omega_1,\omega_2}) \ll \mathrm{h}^N$, which gives an acceptable contribution after summing over ω_2. On the other hand, we have $b \in S^\tau_{\delta',\delta''}$ and $\mathrm{supp}(b_{\omega_1,\omega_2}) \subseteq \mathcal{D}(\omega_1) \cap \mathcal{D}(\omega_2)$. In particular, b_{ω_1,ω_2} is localized at both ω_1 and ω_2. We noted above that these conditions imply that $\mathcal{N}(b_{\omega_1,\omega_2}) \ll 1$ and that for each ω_1, we have $b_{\omega_1,\omega_2} = 0$ for all ω_2 outside some set of cardinality $\mathrm{O}(1)$. The required bound follows. The proof of Lemma 10.10 is thus complete. $\square$

We now resume the proof of Theorem 10.9, turning our attention to the estimate (10.10). Let $a \in S^\tau_{\delta',\delta''}$. By Theorem 10.6, we have $\mathrm{Op}(a) \in \underline{\Psi}^{-\infty}$. We must verify that for all fixed $m, s \in \mathbb{Z}$ and $x_1, \dots, x_n \in \mathfrak{g}$, the operator norm of
$$[\pi(x_1), \dots, [\pi(x_n), \mathrm{Op}(a)]] : \pi^s \to \pi^{s-m}$$
is $\mathrm{O}(\mathrm{h}^{m - n\delta'})$. We first reduce to the case $n = 0$. The star product asymptotics imply that for $x \in \mathfrak{g}$ and $y \in S^\tau_{\delta',\delta''}$, the star product commutator $\{x, y\} := x \star_{\mathrm{h}} y - y \star_{\mathrm{h}} x$ lies in $\mathrm{h}^{1-\delta'} \, S^\tau_{\delta',\delta''} + \mathrm{h}^\infty \, S^{-\infty}$. Thus the iterated star

product commutator $b := \mathrm{h}^{-n+n\delta'}\{x_1,\dots,\{x_n,a\}\}$ lies in $S^{\tau}_{\delta',\delta''} + \mathrm{h}^{\infty} S^{-\infty}$. Each fixed element $x \in \mathfrak{g}$ defines a linear function $\mathfrak{g}^{\wedge} \to i\mathbb{R}$, hence a symbol $x \in S^1_0$, with $\pi(x) = \mathrm{h}^{-1}\operatorname{Op}(x)$. We see by iterated application of the composition formula (10.8) that $[\pi(x_1),\dots,[\pi(x_n),\operatorname{Op}(a)]] \equiv \operatorname{Op}(b) \mod \mathrm{h}^{\infty}\Psi^{-\infty}$. Theorem 10.8 gives $\operatorname{Op}(\mathrm{h}^{\infty} S^{-\infty}) \subseteq \mathrm{h}^{\infty}\Psi^{-\infty}$. It follows that the $n = 0$ case of the required operator norm bound implies the general case.

It remains to show that $\operatorname{Op}(a) : \pi^s \to \pi^{s-m}$ has operator norm $\ll \mathrm{h}^m$, i.e., that for all smooth vectors $v \in \pi$, we have

$$\mathrm{h}^{-m}\|\operatorname{Op}(a)v\|_{\pi^{s-m}} \ll \|v\|_{\pi^s}. \tag{10.14}$$

We consider first the special case $s = m \geqslant 0$, in which our task is to show that

$$\mathrm{h}^{-m}\|\operatorname{Op}(a)v\| \ll \|v\|_{\pi^m}. \tag{10.15}$$

Let us fix an element $z \in \mathfrak{g}$ with $z(\tau) \asymp 1$. Then $z(\xi) \asymp 1$ for all ξ in the support of a. We may construct a symbol $q \in S^{\tau}_{\delta',\delta''}$ that is an approximate quotient for a with respect to z^m in the sense that

$$q \star_{\mathrm{h}} z^m \equiv a \mod \mathrm{h}^N S^{\tau}_{\delta',\delta''} + \mathrm{h}^{\infty} S^{-\infty},$$

where N is fixed large enough in terms of m. To do so, we take for q the series $\sum_{0\leqslant j<J} \mathrm{h}^j q_j$, where the terms q_j are the components of the formal solution:

$$q_0 := \frac{a}{z^m}, \quad q_1 := \frac{-a\star^1 q_0}{z^m}, \quad q_2 := \frac{-a\star^2 q_0 + a\star^1 q_1}{z^m}.$$

and so on. We verify readily – by induction on j, using the quotient rule for derivatives and the fact that $z(\xi) \asymp 1$ for $\xi \in \operatorname{supp}(a)$ – that $q_j \in \mathrm{h}^{-\delta' j} S^{\tau}_{\delta',\delta''}$ and that q has the claimed properties. The operator norm for $\pi \to \pi$, hence also for $\pi^m \to \pi^0$, of any element of $\mathrm{h}^N S^{\tau}_{\delta',\delta''} + \mathrm{h}^{\infty} S^{-\infty}$ is $\operatorname{O}(\mathrm{h}^N)$, as follows from Theorem 10.8 and (a weak form of) Lemma 10.10. It will thus suffice to verify the following modification of (10.15), obtained by replacing $\operatorname{Op}(a)$ by its approximation $\operatorname{Op}(q)\operatorname{Op}(z^m) = \mathrm{h}^m \operatorname{Op}(q)\pi(z)^m$:

$$\|\operatorname{Op}(a)z^m v\| \ll \|v\|_{\pi^m}.$$

By Lemma 10.10, we have $\|\operatorname{Op}(a)z^m v\| \ll \|z^m v\|$. By (10.1), we have $\|z^m v\| \ll \|v\|_{\pi^m}$. The proof of (10.15) is thus complete.

Turning to the general case of (10.14), let us fix $k \in \mathbb{Z}_{\geqslant 0}$ large enough in terms of s and m. It will be enough then to verify the modification of (10.14) obtained by replacing v with its image under the invertible operator Δ^k. The required estimate expands out to

$$\mathrm{h}^{2m}\langle \Delta^{s-m}\operatorname{Op}(a)\Delta^k v, \operatorname{Op}(a)\Delta^k v\rangle \ll \|v\|^2_{\pi^{s+2k}}.$$

We consider first the case $s - m \geqslant 0$. We expand Δ as a sum of monomials, appeal to the skew-adjointness of the action of $\mathfrak{g}$ on π, and apply Cauchy–Schwarz. We reduce in this way to verifying that for all fixed $x_1,\dots,x_{s-m} \in \{1\}\cup\mathfrak{g}$,

$$\mathrm{h}^{-m}\|x_1\cdots x_{s-m}\operatorname{Op}(a)\Delta^k v\| \ll \|v\|_{\pi^{s+2k}}.$$

To that end, we assume our coordinates and basis $\mathcal{B}(\mathfrak{g})$ as in §10.1 chosen compatibly, so that by Lemma 10.4, we have $\Delta = \operatorname{Op}(p)$ with $p(\xi) = \mathrm{h}^2 + |\xi|^2$. We then apply the composition formula to write

$$x_1\cdots x_{s-m}\operatorname{Op}(a)\Delta^k \equiv \mathrm{h}^{-(s-m+2k)}\operatorname{Op}(x_1\star_{\mathrm{h}}\cdots\star_{\mathrm{h}} x_{s-m}\star_{\mathrm{h}} a\star_{\mathrm{h}} p^k), \tag{10.16}$$

where $\equiv$ denotes congruence modulo $\mathrm{h}^{\infty} \Psi^{-\infty}$. (Strictly speaking, the star product $\star_{\mathrm{h}}$ is not quite associative due to the cutoffs, so we should specify that this last expression is evaluated from left to right, say. The formal expansion is associative, so the asymptotic expansion is independent of the order of evaluation.) The remainder term in this congruence contributes acceptably, since the $\pi^{s+k} \to \pi^0$ operator norm of any element of $\mathrm{h}^{\infty} \Psi^{-\infty}$ is $\mathrm{O}(\mathrm{h}^{\infty})$. We likewise incur an acceptable error in replacing the iterated star product $x_1 \star_{\mathrm{h}} \cdots \star_{\mathrm{h}} p^k$ with its approximation $b \in S^{\tau}_{\delta',\delta''}$ obtained by approximating each star product with its asymptotic expansion taken to order J, provided that J is fixed sufficiently large in terms of m, s, δ'. Our task thereby reduces to verifying for all $b \in S^{\tau}_{\delta',\delta''}$ that $\mathrm{h}^{-(s+2k)} \| \mathrm{Op}(b) v\| \ll \|v\|_{\pi^{s+2k}}$, which follows from the special case (10.15) treated above.

We turn to the case $s - m \leqslant 0$. The argument is quite similar, so we will be brief. Using the composition formula, we may replace $\mathrm{Op}(a)\Delta^k$ with $\mathrm{h}^{-2k} \mathrm{Op}(b)$, where $b \in S^{\tau}_{\delta',\delta''}$ is a suitable truncation of the asymptotic expansion of $a \star_{\mathrm{h}} p^k$ with $p(\xi) = \mathrm{h}^2 + |\xi|^2$ as before. We fix $z \in \mathfrak{g}$ as above. We approximately divide b on the left by z^{m-s} and on the right by z^{s+2k}, yielding $q \in S^{\tau}_{\delta',\delta''}$ for which $\mathrm{Op}(b)$ is approximated by $\mathrm{Op}(z^{m-s}) \mathrm{Op}(q) \mathrm{Op}(z^{s+2k}) = \mathrm{h}^{m+2k} \pi(z^{m-s}) \mathrm{Op}(q) \pi(z^{s+2k})$. We thereby reduce to verifying that

$$\langle z^{m-s} \Delta^{s-m} z^{m-s} \mathrm{Op}(q) z^{s+2k} v, \mathrm{Op}(q) z^{s+2k} v \rangle \ll \|v\|^2_{\pi^{s+2k}}.$$

By Lemma 10.3, the operator $\pi(z^{m-s} \Delta^{s-m} z^{m-s})$ lies in Ψ^0_0, and in particular has $\pi \to \pi$ operator norm $\mathrm{O}(1)$. By Cauchy–Schwarz, we reduce to verifying that $\| \mathrm{Op}(q) z^{s+2k} v\| \ll \|v\|_{\pi^{s+2k}}$. This last estimate follows (again) from Lemma 10.10 and (10.1).

The proof of (10.10) is now complete. We may deduce the remaining assertion (10.9) from (10.8), (9.24) and (10.10) by choosing J large enough that $(1 - 2\delta')J \geqslant N$. ☐

10.4. Trace estimates. With notation and assumptions as in §7.5, we recall the rescaled asymptotic form of the Kirillov formula given in [NV, (12.4)]. Throughout this section, the nice cutoff χ is fixed.

Theorem 10.11. *Assume that π is tempered. Fix $\delta \in [0, 1)$ and $J, N \in \mathbb{Z}_{\geqslant 0}$. For all $a \in S^{-\infty}_{\delta}$, we have*

$$\mathrm{h}^d \operatorname{trace}(\mathrm{Op}(a)) = \sum_{0 \leqslant j < J} \mathrm{h}^j \int_{\mathrm{h}\, \mathcal{O}_\pi} \mathcal{D}_j a + \mathrm{O}(\mathrm{h}^{(1-\delta)J} \langle \mathrm{h}\, \lambda_\pi \rangle^{-N}),$$

where

- *d is as in §7.5,*
- *$\mathcal{D}_j$ is a fixed constant coefficient differential operator on $\mathfrak{g}^\wedge$ of pure degree j,*
- *$\int_{\mathrm{h}\, \mathcal{O}_\pi}$ denotes the integral over the rescaled coadjoint multiorbit $\mathrm{h}\, \mathcal{O}_\pi$ with respect to its symplectic measure, normalized as in [NV, §6.1], and*
- *$\langle \mathrm{h}\, \lambda_\pi \rangle = (1 + |\mathrm{h}\, \lambda_\pi|^2)^{1/2} \geqslant 1$ denotes the norm of the rescaled infinitesimal character, as in [NV, §9.8].*

As a consequence, we obtain strong trace estimates when the support of the symbol is disjoint from the rescaled coadjoint multiorbit.

Corollary 10.12. *Assume that π is tempered. Let $a \in S_{\delta}^{-\infty}$, with $0 \leqslant \delta < 1$ fixed. If $\operatorname{supp}(a) \cap \mathrm{h}\, \mathcal{O}_\pi = \emptyset$, then*

$$\operatorname{trace}(\operatorname{Op}(a)) \ll \mathrm{h}^{\infty} \langle \mathrm{h}\, \lambda_\pi \rangle^{-\infty} \ll \mathrm{h}^{\infty}.$$

Since the refined symbol classes $S^{\tau}_{\delta',\delta''}$ are contained in $S^{-\infty}_{\delta''}$, Theorem 10.11 and Corollary 10.12 apply to elements of those classes. It is crucial here that these results apply to $\delta < 1$ (rather than merely to $\delta < 1/2$, as in Theorem 10.8).

The following result, given in [NV, Thm 9 (iii)], is particularly useful for "clean-up." It does not require any temperedness assumption on π (although we will only apply it here in tempered cases).

Theorem 10.13. *Fix $N \in \mathbb{Z}_{\geqslant 0}$ large enough in terms of $\dim(\mathfrak{g})$. Then each operator T on π with $\|T\|_{\pi^0 \to \pi^N} \ll 1$ is trace class, with trace norm*

$$\|T\|_1 \ll \langle \lambda_\pi \rangle^{d-N} \leqslant 1.$$

11. Stability

Let (G, H) be a GGP pair over an archimedean local field F. (A similar discussion applies over any field of characteristic zero; see [NV, §14] for details.)

We retain the notation of §7.

11.1. Definition and characterization.

Definition 11.1. Following [NV, §14], we say that $(\lambda, \mu) \in [\mathfrak{g}^\wedge] \times [\mathfrak{h}^\wedge]$ is *stable* if

$$\operatorname{ev}(\lambda) \cap \operatorname{ev}(\mu) = \emptyset,$$

where $\operatorname{ev}(\cdot)$ denotes the eigenvalue multiset defined in §7.4.

We may reformulate the local assumptions at $\mathfrak{q}$ in the definition of $\mathcal{F}_T$ (§1.3), or equivalently, those concerning the archimedean Satake parameters of π and σ in Theorem 4.2, using the following equivalence, implicit in [NV, §15].

Lemma 11.2. *Let T and h be positive reals with $T \ggg 1$ and $\mathrm{h} \asymp T^{-[F:\mathbb{R}]}$. Let (π, σ) be tempered irreducible unitary representations of a fixed GGP pair (G, H) over F. Let $\lambda_\pi \in [\mathfrak{g}^\wedge]$ and $\lambda_\sigma \in [\mathfrak{h}^\wedge]$ denote the infinitesimal characters (§7.3). Let $\lambda_{\pi,i}$ and $\lambda_{\sigma,j}$ denote the archimedean Satake parameters (§7.4). The following conditions are equivalent:*

- *(i)* $\max(\{|\lambda_{\pi,i}|_F\} \cup \{|\lambda_{\sigma,j}|_F\}) \asymp T$ *and* $|\lambda_{\pi,i} - \lambda_{\sigma,j}|_F \gg T$ *for all* i, j.
- *(ii)* $(\mathrm{h}\, \lambda_\pi, \mathrm{h}\, \lambda_\sigma)$ *lies in some fixed compact subset of* $\{$*stable* $(\lambda, \mu) \in [\mathfrak{g}^\wedge] \times [\mathfrak{h}^\wedge]\}$.

We next recall a standard definition from geometric invariant theory.

Definition 11.3. We write $\mathfrak{g}^\wedge_{\mathrm{stab}}$ for the set of all $\xi \in \mathfrak{g}^\wedge$ that are *stable* with respect to H, i.e., have finite H-stabilizer and (Zariski) closed H-orbit.

We summarize some results from [NV, §14, §17.3] relating the above definitions.

Lemma 11.4. *The set $\mathfrak{g}^\wedge_{\mathrm{stab}}$ consists of all $\xi \in \mathfrak{g}^\wedge$ with $\operatorname{ev}(\xi) \cap \operatorname{ev}(\xi_H) = \emptyset$, hence coincides with the preimage of $\{$stable $(\lambda, \mu) \in [\mathfrak{g}^\wedge] \times [\mathfrak{h}^\wedge]\}$ under the map $\xi \mapsto ([\xi], [\xi_H])$. In particular, it is a dense open subset of $\mathfrak{g}^\wedge$.*

11.2. Basic consequences. We recall from [NV, §14, §17.3] that the map $\mathfrak{g}^{\wedge}_{\mathrm{stab}} \to \{\text{stable } (\lambda, \mu)\}$ is a principal H-bundle over its image. In particular, if (λ, μ) is stable, then the fiber

$$\mathcal{O}^{\lambda,\mu} := \{\xi \in \mathfrak{g}^{\wedge} : [\xi] = \lambda, [\xi_H] = \mu\} \tag{11.1}$$

is either empty, or is an H-*torsor*, i.e., a closed H-invariant subset of $\mathfrak{g}^{\wedge}$ on which H acts simply-transitively.

We assume given a Haar measure on H. We equip $\mathcal{O}^{\lambda,\mu}$ with the measure $d\,\mathrm{Haar}_H$ given by pushforward of that Haar measure via the orbit map, thus

$$\int_{\mathcal{O}^{\lambda,\mu}} a := \int_{\xi \in \mathcal{O}^{\lambda,\mu}} a(\xi)\, d\,\mathrm{Haar}_H(\xi) := \int_{s \in H} a(s \cdot \tau)\, ds.$$

We recall from [NV, §17.3.2] that integration defines a continuous map

$$\{\text{stable } (\lambda, \mu) \in [\mathfrak{g}^{\wedge}] \times [\mathfrak{h}^{\wedge}]\} \times \mathcal{S}(\mathfrak{g}^{\wedge}) \to \mathbb{C},$$

$$(\lambda, \mu, a) \mapsto \int_{\mathcal{O}^{\lambda,\mu}} a,$$

where $\mathcal{S}(\mathfrak{g}^{\wedge})$ denotes the Schwartz space.

11.3. Relative coadjoint orbits. Let π and σ be tempered irreducible unitary representations of G and H, respectively. Let $\mathcal{O}_\pi \subseteq \mathfrak{g}^{\wedge}_{\mathrm{reg}}$ and $\mathcal{O}_\sigma \subseteq \mathfrak{h}^{\wedge}_{\mathrm{reg}}$ denote their coadjoint multiorbits, as described in Theorem 7.1. We set

$$\mathcal{O}_{\pi,\sigma} := \{\xi \in \mathcal{O}_\pi : \xi_H \in \mathcal{O}_\sigma\}.$$

We say that (π, σ) is *orbit-distinguished* if $\mathcal{O}_{\pi,\sigma}$ is nonempty.

Remark 11.5. As remarked in [NV, §22.2], it is likely provable that distinction in the traditional sense is asymptotically equivalent to orbit-distinction, e.g., in the setting of Theorem 4.2 with T large. (That orbit-distinction implies distinction follows from Theorem 12.2; the converse should then follow by combining Theorem 12.2 with strong multiplicity one for archimedean L-packets [BP1, Lu].) Since it is straightforward to check orbit-distinction in examples, we are content to leave it in the hypotheses of our main results. We note that for complex groups, orbit-distinction always holds, while for compact unitary groups, it is an asymptotic form of the interlacing condition describing the classical branching laws.

If $(\lambda_\pi, \lambda_\sigma)$ is stable and $\mathcal{O}_{\pi,\sigma}$ is nonempty, then it follows from the discussion of §11.2 that $\mathcal{O}_{\pi,\sigma} = \mathcal{O}^{\lambda_\pi,\lambda_\sigma}$ is an H-torsor, and so we may integrate over it as above; otherwise, we define any integral over $\mathcal{O}_{\pi,\sigma}$ to be zero. We define in the same way integrals over the rescalings $\mathrm{h}\,\mathcal{O}_{\pi,\sigma} = \mathcal{O}^{\mathrm{h}\,\lambda_\pi,\mathrm{h}\,\lambda_\sigma}$ of such sets.

12. Relative character asymptotics

Here we recall the main results of [NV, §19].

12.1. Setup. Let (G, H) be a fixed GGP pair over an archimedean local field F. We assume given a fixed Haar measure on H. Let π and σ be tempered irreducible unitary representations of G and H, respectively. Let $\sigma^\vee$ denote the contragredient. We obtain the (tempered irreducible unitary) representation $\pi \otimes \sigma^\vee$ of

$$M := G \times H.$$

12.2. Formal definitions and identities. We write $\mathcal{B}(\pi \otimes \sigma^\vee), \mathcal{B}(\pi), \mathcal{B}(\sigma)$ for orthonormal bases consisting of eigenvectors for $\Delta_M, \Delta_G, \Delta_H$, with Δ as in §10.1. Such eigenvectors are in particular smooth.

For an operator T on $\pi \otimes \sigma^\vee$, we define

$$\mathcal{H}(T) := \sum_{v \in \mathcal{B}(\pi \otimes \sigma^\vee)} \int_{h \in H} \langle hTv, v \rangle \, dh$$

provided that the sum converges absolutely.

Working formally for the moment, if T_1 (resp. T_2) is an operator on π (resp. σ), then we may naturally define operators $T_2^\vee$ on $\sigma^\vee$ and $T_1 \otimes T_2^\vee$ on $\pi \otimes \sigma^\vee$, and we have

$$\mathcal{H}(T_1 \otimes T_2^\vee) = \sum_{v \in \mathcal{B}(\pi)} \sum_{u \in \mathcal{B}(\sigma)} \int_{h \in H} \langle hT_1 v, v \rangle \langle T_2 u, hu \rangle \, dh.$$

Moreover, writing T_j^* for the adjoint operator,

$$\begin{aligned} \mathcal{H}(T_1 T_1^* \otimes (T_2 T_2^*)^\vee) &= \sum_{v \in \mathcal{B}(\pi)} \sum_{u \in \mathcal{B}(\sigma)} \int_{h \in H} \langle hT_1 v, T_1 v \rangle \langle T_2 u, hT_2 u \rangle \, dh \\ &= \sum_{v \in \mathcal{B}(\pi)} \sum_{u \in \mathcal{B}(\sigma)} \mathcal{Q}(T_1 v \otimes T_2 u), \end{aligned} \tag{12.1}$$

where the quadratic form $\mathcal{Q}$ is defined as in §3.5. In particular, writing 1 for the identity operator,

$$\mathcal{H}(T_1 T_1^* \otimes 1) = \sum_{v \in \mathcal{B}(\pi)} \sum_{u \in \mathcal{B}(\sigma)} \mathcal{Q}(T_1 v \otimes u). \tag{12.2}$$

12.3. Convergence criteria and *a priori* bounds. In this section, for an operator T on one of the representations $\pi \otimes \sigma^\vee, \pi$ or σ and a nonnegative integer N, we write $\nu_N(T) \in [0, \infty]$ for the operator norm of $\Delta^N T \Delta^N$, where Δ is respectively Δ_M, Δ_G or Δ_H.

Suppose for instance that T is an operator on π. Then $\nu_N(T)$ is the operator norm of the map $\pi^{-N} \to \pi^N$; in particular, $\nu_N(T)$ is finite and $\mathrm{O}(1)$ whenever T belongs to the operator class Ψ_δ^{-2N} (for any fixed δ).

Lemma 12.1. *Fix a natural number N large enough in terms of (G, H).*

(i) Let T be an operator on $\pi \otimes \sigma^\vee$ for which $\nu_N(T) = \mathrm{O}(1)$. Then $\mathcal{H}(T)$ is defined and $\mathrm{O}(1)$.

(ii) Let T be an operator on π for which $\nu_N(T) = \mathrm{O}(1)$. Then $\mathcal{H}(T \otimes 1)$ is defined and $\mathrm{O}(1)$.

(iii) Let T_1 and T_2 be operators on π and σ, respectively, for which $\nu_N(T_1), \nu_N(T_2) = \mathrm{O}(1)$. Then $\mathcal{H}(T_1 \otimes T_2^\vee)$ is defined and $\mathrm{O}(1)$; moreover, the formal identities stated in §12.2 are valid.

(iv) The quadratic form $\mathcal{Q}$ extends continuously to the tensor product of Sobolev spaces $\pi^N \otimes \sigma^N$.

(v) The identities (12.1) and (12.2) hold more generally taking for $\mathcal{B}(\pi)$ and $\mathcal{B}(\sigma)$ any orthonormal bases, provided that for (12.2), the basis $\mathcal{B}(\sigma)$ is contained in σ^N.

Proof. For (i), see just after [NV, (19.14)]. For (ii), see just before [NV, (18.5)]. The translation to the present formulation is as in the proof of Theorem 9.5. For (iii), we observe first, by writing Δ_M in terms of Δ_G and Δ_H for suitably chosen bases, that $\nu_N(T_1 \otimes T_2^\vee) = \mathrm{O}(1)$; we then apply (i) to obtain the first assertion. It remains to verify the formal identities. We start with (12.2). The RHS of (12.2) may be written

$$\sum_{v_1 \in \mathcal{B}(\pi)} \sum_{u \in \mathcal{B}(\sigma)} \int_{h \in H} \langle hT_1 v_1, T_1 v_1 \rangle \langle u, hu \rangle \, dh.$$

We expand the second copy of $T_1 v_1$ as $\overline{T_1 v_1} = \sum_{v_2 \in \mathcal{B}(\pi)} \langle v_2, T_1 v_1 \rangle \overline{v_2}$ and rewrite the inner product as $\langle T_1^* v_2, v_1 \rangle$. By arguments as in [NV, §18] (which boil down to the uniform trace class property of Δ^{-N} given, e.g., in Theorem 10.13), we have

$$\sum_{v_2 \in \mathcal{B}(\pi)} \int_{h \in H} |\langle T_1^* v_2, v_1 \rangle \langle hT_1 v_1, v_2 \rangle \langle u, hu \rangle| \, dh < \infty.$$

We may thus swap the sum over v_2 with the integral over h, giving

$$\sum_{v_1 \in \mathcal{B}(\pi)} \sum_{u \in \mathcal{B}(\sigma)} \left(\sum_{v_2 \in \mathcal{B}(\pi)} \langle T_1^* v_2, v_1 \rangle \int_{h \in H} \langle hT_1 v_1, v_2 \rangle \langle u, hu \rangle \, dh \right).$$

Again applying arguments as in [NV, §18], we see that

$$\sum_{v_1 \in \mathcal{B}(\pi)} \sum_{u \in \mathcal{B}(\sigma)} \sum_{v_2 \in \mathcal{B}(\pi)} \left| \langle T_1^* v_2, v_1 \rangle \int_{h \in H} \langle hT_1 v_1, v_2 \rangle \langle u, hu \rangle \, dh \right| < \infty.$$

We may thus rearrange the outer summations, giving

$$\sum_{v_2 \in \mathcal{B}(\pi)} \sum_{u \in \mathcal{B}(\sigma)} \left(\sum_{v_1 \in \mathcal{B}(\pi)} \langle T_1^* v_2, v_1 \rangle \int_{h \in H} \langle hT_1 v_1, v_2 \rangle \langle u, hu \rangle \, dh \right).$$

By reversing the reasoning applied in the previous step, we see that

$$\sum_{v_1 \in \mathcal{B}(\pi)} \langle T_1^* v_2, v_1 \rangle \int_{h \in H} \langle hT_1 v_1, v_2 \rangle \langle u, hu \rangle \, dh = \int_{h \in H} \langle hT_1 T_1^* v_2, v_2 \rangle \langle u, hu \rangle \, dh,$$

which completes the proof. The proof of (12.1) is similar, using, e.g., that

$$\sum_{v_1, v_2 \in \mathcal{B}(\pi)} \sum_{u_1, u_2 \in \mathcal{B}(\sigma)} \left| \int_{h \in H} \langle hT_1 v_1, T_1 v_2 \rangle \langle T_2 u_2, hT_2 u_1 \rangle \right| < \infty.$$

For (iv), combine [NV, A.2.2] and [NV, (18.1)]. The proof of (v) is similar to that of (iii), cf. [Kn, Lem 10.1] or [BM, Appendix 4]. □

12.4. Stability in the product setting. We say that $\xi \in \mathfrak{m}^\wedge$ is *stable* if under the diagonal action of H, it has finite stabilizer and closed orbit. We note that $\xi \in \mathfrak{g}^\wedge$ is stable in the sense of §11.1 if and only if the element $(\xi, -\xi_H) \in \mathfrak{m}^\wedge$ is stable in the present sense (see [NV, §19.4] for further discussion).

12.5. Main estimates.

Theorem 12.2. *Let $0 \leqslant \delta < 1/2$ be fixed. Let π and σ be tempered irreducible unitary representations of G and H, as above.*

(i) *Let $a \in S_\delta^{-\infty}(\mathfrak{m}^\wedge)$ be supported in some fixed compact collection of stable elements of $\mathfrak{m}^\wedge$. Set $A := \mathrm{Op}_{\mathrm{h}}(a : \pi \otimes \sigma^\vee)$. Then for each fixed $J \geqslant 0$, we have*

$$\mathcal{H}(A) = \sum_{0 \leqslant j < J} \mathrm{h}^j \int_{\xi \in \mathrm{h}\, \mathcal{O}_{\pi,\sigma}} \mathcal{D}_j a(\xi, -\xi_H) \, d\, \mathrm{Haar}_H(\xi) + \mathrm{O}(\mathrm{h}^{(1-2\delta)J}), \tag{12.3}$$

where the $\mathcal{D}_j$ are fixed differential operators on $\mathfrak{m}^\wedge$ with the following properties:

- *$\mathcal{D}_0 a = a$ for all smooth $a : \mathfrak{m}^\wedge \to \mathbb{C}$.*
- *$\mathcal{D}_j$ has order $\leqslant 2j$ and homogeneous degree j: $\mathcal{D}_j(a_{\mathrm{h}}) = \mathrm{h}^j (\mathcal{D}_j a)_{\mathrm{h}}$.*

(ii) *Let $a \in S_\delta^{-\infty}(\mathfrak{g}^\wedge)$ be supported in some fixed compact collection of stable elements of $\mathfrak{g}^\wedge$. Set $A := \mathrm{Op}_{\mathrm{h}}(a : \pi)$, so that $A \otimes 1$ is an operator on $\pi \otimes \sigma^\vee$. Then*

$$\mathcal{H}(A \otimes 1) = \sum_{0 \leqslant j < J} \mathrm{h}^j \int_{\xi \in \mathrm{h}\, \mathcal{O}_{\pi,\sigma}} \mathcal{D}_j a(\xi) \, d\, \mathrm{Haar}_H(\xi) + \mathrm{O}(\mathrm{h}^{(1-2\delta)J}), \tag{12.4}$$

where the $\mathcal{D}_j$ are fixed differential operators on $\mathfrak{g}^\wedge$ with properties analogous to those stated in (i).

As explained in [NV, §19.3], such estimates apply readily to compositions: one just substitutes the composition formula (10.9) into (12.3) or (12.4), using Lemma 12.1 as an *a priori* bound to clean up the remainder terms.

13. Interlude on regular elements

We record some technical estimates to be applied below. Fix a connected reductive group G over $\mathbb{R}$. Recall from §7.2 that $\mathfrak{g}^\wedge_{\mathrm{reg}} \subseteq \mathfrak{g}^\wedge$ denotes the subset of regular elements, i.e., those whose centralizer has minimal dimension (equal to $\mathrm{rank}(\mathfrak{g})$, the dimension of any Cartan subalgebra). It is an open subset. For our purposes, the most important property of this subset is the following result of Kostant (see [Ko, Thm 0.1]).

Theorem 13.1. *For $\tau \in \mathfrak{g}^\wedge$, the derivative of the map $\mathfrak{g}^\wedge \to [\mathfrak{g}^\wedge]$ is surjective at τ if and only if $\tau \in \mathfrak{g}^\wedge_{\mathrm{reg}}$.*

We will frequently apply the facts summarized in the following Lemma.

Lemma 13.2. *Let $\tau \in \mathfrak{g}^\wedge_{\mathrm{reg}}$. Let U be a small enough open neighborhood of τ. Let γ denote the restriction to U of the map $\mathfrak{g} \to [\mathfrak{g}^\wedge]$. Then $\gamma : U \to \gamma(U)$ is a fibered manifold, i.e., a smooth surjection whose derivative is surjective at each point. The fibers of γ are precisely the intersections with U of the coadjoint orbits.*

Proof. Each assertion follows readily from Theorem 13.1. □

We estimate the symplectic measure of a small ball based at a regular element:

Lemma 13.3. *Let τ belong to a fixed compact subset of $\mathfrak{g}^\wedge_{\mathrm{reg}}$. Let $G \cdot \tau$ denote the coadjoint orbit containing τ and $\omega_{G\cdot\tau}$ the corresponding symplectic measure, normalized as in [NV, §6.1]. For $r > 0$, set $B_\tau(r) := \{\xi \in \mathfrak{g}^\wedge : |\tau - \xi| < r\}$. Then for $r \lll 1$, we have*

$$\omega_{G\cdot\tau}(B_\tau(r)) \asymp r^{\dim(\mathfrak{g}) - \mathrm{rank}(\mathfrak{g})}.$$

Proof. Set $n := \operatorname{rank}(\mathfrak{g})$ and $m := \dim(\mathfrak{g}) - \operatorname{rank}(\mathfrak{g})$. Let $\tau \in \mathfrak{g}^\wedge_{\mathrm{reg}}$. We choose a small enough neighborhood U of τ. We have noted (lemma 13.2) that the fibers of $U \to [\mathfrak{g}^\wedge]$ are then the intersections with U of coadjoint orbits. By shrinking U if necessary, we may suppose given a trivialization $\kappa : U \hookrightarrow \mathbb{R}^m \times \mathbb{R}^n$ of the fibered manifold $U \to \operatorname{image}(U) \subseteq [\mathfrak{g}^\wedge]$, i.e., a coordinate chart under which the map $\mathfrak{g} \to [\mathfrak{g}^\wedge]$ corresponds to the projection onto $\mathbb{R}^n$. We may find a smooth function $\mu : \kappa(U) \to \mathbb{R}_{>0}$ so that the symplectic measures on those fibers are given by μ times the Lebesgue measure on $\mathbb{R}^m$. (We use here that if $2d = \dim(\mathfrak{g}) - \operatorname{rank}(\mathfrak{g})$, then the symplectic measure on a regular coadjoint orbit is induced by a constant multiple of the dth power of the canonical symplectic form, which is nonvanishing on $\mathfrak{g}^\wedge_{\mathrm{reg}}$.) By shrinking U if necessary, we may assume that μ is bounded from above and below by positive scalars. The required estimates follow in the special case that Ω is a small enough neighborhood of τ, then in general by compactness. $\square$

We record another simple consequence of Lemma 13.2:

Lemma 13.4. *Let τ belong to a fixed compact subset of $\mathfrak{g}^\wedge_{\mathrm{reg}}$. Write $\mathcal{O}_\tau := G \cdot \tau$ for its coadjoint orbit. Let $\xi \in \mathfrak{g}^\wedge$ with $\xi \lll 1$ and $\tau + \xi \in \mathcal{O}_\tau$. Then there exists $x \in \mathfrak{g}^\flat_\tau$ with*

$$\exp(x)\tau = \tau + \xi, \quad x \asymp \xi.$$

Proof. Consider first the special case in which τ is fixed (and regular). The derivative at the identity of the orbit map $G \to \mathcal{O}_\tau$, $g \mapsto g\tau$ is the map

$$\mathfrak{g} \to T_\tau(\mathcal{O}_\tau) = [\mathfrak{g}, \tau], \quad x \mapsto [x, \tau].$$

That derivative has kernel $\mathfrak{g}_\tau$, so its restriction to the complementary subspace $\mathfrak{g}^\flat_\tau$ is an isomorphism. The inverse function theorem tells us that the map $f : \mathfrak{g}^\flat_\tau \to \mathcal{O}_\tau$ given by $x \mapsto \exp(x)\tau$ induces a diffeomorphism between some fixed neighborhoods $0 \in U \subseteq \mathfrak{g}^\flat_\tau$ and $\tau \in V \subseteq \mathcal{O}_\tau$. Shrinking those neighborhoods if needed and appealing to $\mathrm{O}(1)$ bounds for the second partial derivatives of f, we may arrange that $(df)_x$ is close in operator norm to the isomorphism $(df)_0$ for all $x \in U$. We may assume also that U is convex. Let $x_0, x_1 \in U$. By the fundamental theorem of calculus applied to the linear map

$$\gamma : [0, 1] \to U, \quad \gamma(0) = x_0, \quad \gamma(1) = x_1$$

as well as the composition $f \circ \gamma$, we deduce that $|f(x_1) - f(x_0)| \asymp |x_1 - x_0|$. (We refer to Lemma 15.5, below, for complete details concerning a closely-related argument.) In particular, for ξ as in the hypotheses of the Lemma, we have $\tau + \xi \in V = \exp(U)\tau = f(U)$, so we may write $\tau + \xi = f(x)$ for some $x \in U$. Then

$$|\xi| = |(\tau + \xi) - \tau| = |f(x) - f(0)| \asymp |x - 0| = |x|,$$

as required.

In the general case that τ merely belongs to a fixed compact collection of regular elements, we may obtain the required uniformity by considering the product spaces

$$X := \{(\tau, x) : \xi \in \mathfrak{g}^\wedge_{\mathrm{reg}}, x \in \mathfrak{g}^\flat_\tau\}, \quad Y = \{(\tau, \xi) : \tau \in \mathfrak{g}^\wedge_{\mathrm{reg}}, \tau + \xi \in \mathcal{O}_\tau\}.$$

In view of Lemma 13.2, X and Y are naturally manifolds, and the map $X \to Y$ given by $(\tau, x) \mapsto (\tau, \exp(x)\tau - \tau)$ is smooth. We now argue as in the previous paragraph, but with the inverse function theorem applied in the product setting. $\square$

We record a convenient relationship between τ-coordinates (§9.4.1) and the extent to which a Lie group element centralizes τ:

Lemma 13.5. *Let τ belong to a fixed compact subset of $\mathfrak{g}^{\wedge}_{\mathrm{reg}}$. Let $x = (x', x'')$ denote τ-coordinates. Let $x \in \mathfrak{g}$ with $x \lll 1$. Then*

$$x' \asymp [x, \tau] \asymp \exp(x)\tau - \tau.$$

Proof. We show first that

$$x' \asymp [x, \tau].$$

We have $[x, \tau] = [x', \tau]$, so we may assume that $x = x' \in \mathfrak{g}^{\flat}_{\tau}$. We then argue as in the proof of Lemma 13.4, using that the map $\mathfrak{g}^{\flat}_{\tau} \to [\mathfrak{g}, \tau]$, $x \mapsto [x, \tau]$ is injective.

To complete the proof, we open the exponential series, giving

$$\exp(x)\tau - \tau = [x, \tau] + \sum_{j \geqslant 2} \frac{1}{j!} (\mathrm{ad}^*_x)^{j-1} [x, \tau].$$

We use now that $x \lll 1$ and $\tau \ll 1$ to see that the sum over j is much smaller than $[x, \tau]$. It follows that the RHS is $\asymp [x, \tau]$, as required. ☐

We compare the coadjoint orbit $G \cdot \tau$ containing a regular element τ to its tangent plane $\tau + \mathfrak{g}^{\perp}_{\tau}$ at that element:

Lemma 13.6. *Let τ belong to a fixed compact subset of $\mathfrak{g}^{\wedge}_{\mathrm{reg}}$. Let $\xi, \eta \in \mathfrak{g}^{\wedge}$. Suppose $\xi \lll 1$ and $\tau + \xi \in G \cdot \tau$. Then, with τ-coordinates $\xi = (\xi', \xi'')$, we have $|\xi''| \ll |\xi'|^2$.*

Proof. The reader is encouraged to review Figure 3, which suggests a convincing "proof by picture" of the required estimate. We record a written verification for completeness. By Lemma 13.4, we may write $\tau + \xi = \exp(x)\tau$ with $x \in \mathfrak{g}^{\flat}_{\tau}$ and $x \asymp \xi \lll 1$. Expanding the exponential series, we obtain

$$\xi = [\tau, x] + \sum_{j \geqslant 2} \frac{1}{j!} (\mathrm{ad}^*_x)^{j-1} [\tau, x].$$

By Lemma 13.5, we have $[\tau, x] \asymp x \asymp \xi$. Since $x \lll 1$, the sum over $j \geqslant 2$ is thus of size $\mathrm{O}(|\xi|^2)$. Since $[\tau, x] \in [\mathfrak{g}, \tau] = \mathfrak{g}^{\perp}_{\tau}$, it follows from the definition of τ-coordinates that $\xi'' = \mathrm{O}(|\xi|^2)$. We conclude via the relation $|\xi| \asymp |\xi'| + |\xi''|$ and the hypothesis $\xi \lll 1$. ☐

14. Construction of analytic test vectors

The purpose of this section is to prove our main local result, Theorem 4.2, modulo a technical estimate which we postpone.

14.1. Setup. We adopt the setting of Theorem 4.2: (G, H) is a fixed GGP pair over an archimedean local field F, $T \ggg 1$ is a positive real, and (π, σ) is a pair of tempered irreducible unitary representations satisfying certain assumptions. In view of Lemma 11.2, those assumptions imply the following:

- Setting

  $$\mathrm{h} := T^{-1/[F:\mathbb{R}]} \lll 1, \quad \text{so that } T = \mathrm{h}^{-[F:\mathbb{R}]},$$

  the pair $(\mathrm{h}\, \lambda_\pi, \mathrm{h}\, \lambda_\sigma)$ of rescaled infinitesimal characters lies in a fixed compact subset of $\{\text{stable } (\lambda, \mu) \in [\mathfrak{g}^{\wedge}] \times [\mathfrak{h}^{\wedge}]\}$.
- $\mathcal{O}_{\pi,\sigma}$ is nonempty.

We have noted in §11.3 that $\mathcal{O}_{\pi,\sigma}$ is then an H-torsor.

We must show that for each fixed $\kappa > 0$, there is a test function $f \in C_c^\infty(G)$ and a smooth unit vector $u \in \sigma$, with f supported arbitrarily close to the identity, so that the three properties enunciated in Theorem 4.2 are satisfied. We construct f in §14.6 and u in §14.7. We verify the three properties in §14.8, §14.9, §14.10.

We fix $0 < \delta = \delta' < 1/2$ and $\delta' < \delta'' < 2\delta'$ (so that the condition (9.17) is satisfied), with both $1/2 - \delta'$ and $1 - \delta''$ taken small enough in terms of κ.

Throughout this section, N is a fixed sufficiently large positive integer.

14.2. Choice of τ. Let $\tau \in \mathfrak{g}^\wedge$ be an element of $\mathcal{O}_{\pi,\sigma}$ of (say) minimal Euclidean norm. By the condition on $(\mathrm{h}\,\lambda_\pi, \mathrm{h}\,\lambda_\sigma)$ stated above and the "principal bundle" consequence of stability, τ belongs to a fixed compact subset of $\mathfrak{g}^\wedge_{\mathrm{stab}}$. In view of the following lemma, τ (resp. its restriction τ_H) belongs to a fixed compact subset of $\mathfrak{g}^\wedge_{\mathrm{reg}}$ (resp. $\mathfrak{h}^\wedge_{\mathrm{reg}}$).

Lemma 14.1. *Let $\tau \in \mathfrak{g}^\wedge_{\mathrm{stab}}$. Then $\tau \in \mathfrak{g}^\wedge_{\mathrm{reg}}$ and $\tau_H \in \mathfrak{h}^\wedge_{\mathrm{reg}}$.*

Proof. This is [NV, §14.3, Lem 2]. □

14.3. Construction of basic symbols. We may find symbols $a \in S_\delta^{-\infty}(\mathfrak{g}^\wedge)$ and $b \in S_\delta^{-\infty}(\mathfrak{h}^\wedge)$, valued in the unit interval $[0,1]$, that are smooth bumps on $\tau + \mathrm{O}(\mathrm{h}^\delta)$ and $\tau_H + \mathrm{O}(\mathrm{h}^\delta)$, respectively, in the sense that

$$a(\xi) \neq 1 \implies |\xi - \tau| \gg \mathrm{h}^\delta, \quad a(\xi) \neq 0 \implies |\xi - \tau| \ll \mathrm{h}^\delta,$$
$$b(\eta) \neq 1 \implies |\eta - \tau_H| \gg \mathrm{h}^\delta, \quad b(\eta) \neq 0 \implies |\eta - \tau_H| \ll \mathrm{h}^\delta.$$

14.4. Application of relative character estimates. Set

$$\begin{aligned} A &:= \mathrm{Op}_{\mathrm{h}}(a : \pi) \in \Psi_\delta^0(\pi) \cap \mathrm{h}^{-N}\,\Psi_\delta^{-N}(\pi),\\ B &:= \mathrm{Op}_{\mathrm{h}}(b : \sigma) \in \Psi_\delta^0(\sigma) \cap \mathrm{h}^{-N}\,\Psi_\delta^{-N}(\sigma), \end{aligned} \tag{14.1}$$

where the memberships follow from (10.7). Recall from §3.5 the quadratic form $\mathcal{Q}$ on the smooth subspace of $\pi \otimes \sigma$ given by

$$\mathcal{Q}(v \otimes u) = \int_{s \in H} \langle s v, v\rangle \langle u, s u\rangle \, ds.$$

As noted in Lemma 12.1, $\mathcal{Q}$ extends continuously to the tensor product of Sobolev spaces $\pi^N \otimes \sigma^N$. We write $\mathcal{B}(\pi)$ and $\mathcal{B}(\sigma)$ for any orthonormal bases.

Lemma 14.2. *We have*

$$\sum_{v \in \mathcal{B}(\pi)} \sum_{u \in \mathcal{B}(\sigma)} \mathcal{Q}(Av \otimes Bu) \asymp \mathrm{h}^{\delta \dim(H)}, \tag{14.2}$$

with the LHS absolutely convergent.

Proof. Setting $M := G \times H$, we apply the results of §12 to the symbol $c \in S_\delta^{-\infty}(\mathfrak{m}^\wedge)$ given by $c(\xi, \eta) := a(\xi) b(-\eta)$. We choose the nice cutoff on $\mathfrak{h}$ implicit in the definition of Op to be the product of those on $\mathfrak{g}$ and $\mathfrak{h}$. We see then that

$$C := \mathrm{Op}_{\mathrm{h}}(c : \pi \otimes \sigma^\vee) = A \otimes B^\vee$$

and that the LHS of (14.2) is $\mathcal{H}(CC^*)$. The required absolute convergence follows from Lemma 12.1. We note that c is supported in a small neighborhood of $(\tau, -\tau_H)$,

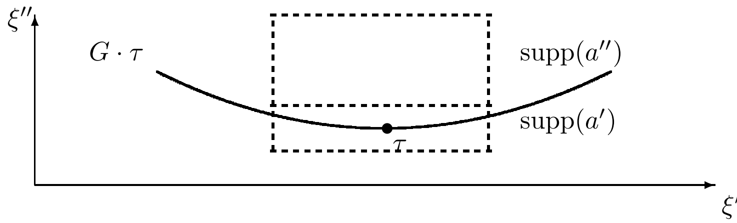

FIGURE 4. A schematic of the decomposition $a = a' + a''$.

hence is supported in some fixed compact collection of stable elements. The results of §12.5 give (for fixed N and J, with J large enough in terms of N)

$$\mathcal{H}(CC^*) = \sum_{0 \leqslant j_1, j_2 < J} \mathrm{h}^{j_1 + j_2} \int_{\xi \in \mathrm{h}\, \mathcal{O}_{\pi,\sigma}} \mathcal{D}_{j_1}(c \star_{j_2} \bar{c})(\xi, -\xi_H)\, d\, \mathrm{Haar}_H(\xi) + \mathrm{O}(\mathrm{h}^N).$$

The $(j_1, j_2) = (0,0)$ term is

$$\int_{\xi \in \mathrm{h}\, \mathcal{O}_{\pi,\sigma}} |a(\xi)|^2 |b(\xi_H)|^2 \, d\, \mathrm{Haar}_H(\xi) = \int_{s \in H} |a(s \cdot \tau)|^2 |b(s \cdot \tau_H)|^2 \, ds$$

By the construction of a and b and the "principal bundle" consequence of stability, this term is $\asymp \mathrm{h}^{\delta \dim(H)}$. The general (j_1, j_2) term is an integral of a function of size $\mathrm{h}^{(1-2\delta)(j_1+j_2)}$ over a domain of volume $\ll \mathrm{h}^{\delta \dim(H)}$, and is thus

$$\ll \mathrm{h}^{(1-2\delta)(j_1+j_2) + \delta \dim(H)}.$$

Therefore the $(0,0)$ term dominates and the required estimate (14.2) follows. □

14.5. Construction of refined symbols. Recall from the end of §14.1 that we have chosen some parameters δ', δ''. Using that $\delta' > \delta'' - \delta'$, we see that the symbol a belongs to $S^{\tau}_{\delta',\delta''}$. We may smoothly decompose $a = a' + a''$, where $a', a'' \in S^{\tau}_{\delta',\delta''}$ satisfy the following support conditions given in terms of τ-coordinates (§9.4.1):

$$\begin{aligned} a'(\tau + \xi) \neq 0 \quad &\implies \quad \xi' \ll \mathrm{h}^{\delta'}, \quad \xi'' \ll \mathrm{h}^{\delta''}, \\ a''(\tau + \xi) \neq 0 \quad &\implies \quad \xi' \ll \mathrm{h}^{\delta'}, \quad \mathrm{h}^{\delta''} \ll \xi'' \ll \mathrm{h}^{\delta'}. \end{aligned} \tag{14.3}$$

Then

$$A = A' + A'', \quad A' := \mathrm{Op}_{\mathrm{h}}(a' : \pi), \quad A'' := \mathrm{Op}_{\mathrm{h}}(a'' : \pi).$$

Lemma 14.3. *$A'' \in \mathrm{h}^{\infty}\, \Psi^{-\infty}(\pi)$.*

Proof. The key observation is that, as suggested by Figure 4, the support of a'' is disjoint from the rescaled coadjoint orbit $\mathrm{h}\, \mathcal{O}_{\pi}$. To see this, suppose $\tau + \xi$ lies in the support of a'' and also in $\mathrm{h}\, \mathcal{O}_{\pi}$. We see then from (14.3) that $\xi \lll 1$. By Lemma 13.2 and the fact that $\mathcal{O}_{\pi}$ is contained in the fiber of λ_{π} (Theorem 7.1), it follows that $\tau + \xi \in G \cdot \tau$. By Lemma 13.6, we obtain $|\xi''| \ll |\xi'|^2$. From (14.3), we deduce that

$$\mathrm{h}^{\delta''} \ll |\xi''| \ll |\xi|^2 \ll \mathrm{h}^{2\delta'}.$$

We obtain a contradiction in view of the inequality $2\delta' > \delta''$ and our assumption $\mathrm{h} \lll 1$.

Recall from §10.1 the notation $\Delta = 1 - \sum_{x \in \mathcal{B}(\mathfrak{g})} x^2 \in \mathfrak{U}(\mathfrak{g})$. By Theorem 10.6, we have $A'' \in \underline{\Psi}^{-\infty}(\pi)$, so to deduce the required conclusion that A'' belongs to $\mathrm{h}^\infty \Psi^{-\infty}(\pi)$, it suffices to show that for each fixed $m \in \mathbb{Z}_{\geqslant 0}$, the bounded operator $\pi \to \pi$ defined by $S := \pi(\Delta)^m A'' \pi(\Delta)^m$ has operator norm of size $\mathrm{O}(\mathrm{h}^\infty)$. Since the operator norm is bounded by the Hilbert–Schmidt norm, it is enough to check that

$$\operatorname{trace}(SS^*) \ll \mathrm{h}^\infty .$$

To that end, let $p \in S_0^2(\mathfrak{g}^\wedge)$ denote the symbol given by the polynomial function

$$p(\xi) = 1 - \sum_{x \in \mathcal{B}(\mathfrak{g})} \langle x, \xi \rangle^2 \geqslant 0.$$

Since Δ is the symmetrization of p, we see that Δ^m is likewise the symmetrization of $p^m \in S_0^{2m}(\mathfrak{g}^\wedge)$, hence by Lemma 10.4 that

$$\pi(\Delta)^m = \pi(\Delta^m) = \mathrm{Op}_1(p^m) = \mathrm{h}^{-2m} \mathrm{Op}_{\mathrm{h}}(p^m).$$

Fix $N \in \mathbb{Z}_{\geqslant 0}$ large enough in terms of m. By the composition formula (10.9), we may write

$$SS^* = \mathrm{h}^{-8m} \mathrm{Op}_{\mathrm{h}}(q) + \mathcal{E}$$

with (recall that $\delta' = \delta$)

$$\mathcal{E} \in \mathrm{h}^N \Psi_\delta^{-N}(\pi)$$

and with $q \in S^{\tau}_{\delta', \delta''} \subseteq S^{-\infty}_{\delta''}$ obtained by truncating the asymptotic expansion of

$$p^m \star_h a'' \star_{\mathrm{h}} p^m \star_{\mathrm{h}} p^m \star_{\mathrm{h}} \overline{a''} \star_{\mathrm{h}} p^m$$

(with the same conventions concerning order of evaluation as in (10.16)). Then $\operatorname{supp}(q) \subseteq \operatorname{supp}(a'')$, hence

$$\operatorname{supp}(q) \cap \mathrm{h}\, \mathcal{O}_\pi = \emptyset.$$

By Corollary 10.12, we have $\operatorname{trace}(\mathrm{Op}_{\mathrm{h}}(q)) \ll \mathrm{h}^\infty$. By Theorem 10.13, we have $\operatorname{trace}(\mathcal{E}) \ll \mathrm{h}^N$. Since N was arbitrary, the proof is now complete. □

Lemma 14.4. *We have*

$$\sum_{v \in \mathcal{B}(\pi)} \sum_{u \in \mathcal{B}(\sigma)} \mathcal{Q}(A'v \otimes Bu) \asymp \mathrm{h}^{\delta \dim(H)} . \tag{14.4}$$

Proof. We will show that the left hand sides of (14.2) and (14.4) differ by $\mathrm{O}(\mathrm{h}^\infty)$. The difference in question is

$$\sum_{u \in \mathcal{B}(\sigma)} \sum_{v \in \mathcal{B}(\pi)} (\mathcal{Q}(Av \otimes Bu) - \mathcal{Q}(A'v \otimes Bu)) = \mathcal{H}(D \otimes E^\vee),$$

where

$$D := A'(A'')^* + A''(A')^* + A''(A'')^*, \quad E := BB^*.$$

By Theorem 10.9, we have $A' \in \Psi_\delta^0(\pi)$ (recall that $\delta = \delta'$). By Lemma 14.3, we have $A'' \in \mathrm{h}^\infty \Psi^{-\infty}(\pi)$. As noted in (14.1), we have $B \in \mathrm{h}^{-N} \Psi_\delta^{-N}(\sigma)$. By Lemma 10.2, it follows that $D \in \mathrm{h}^\infty \Psi^{-\infty}(\pi)$ and $E \in \mathrm{h}^{-2N} \Psi_\delta^{-2N}(\sigma)$. In particular, with the notation ν_N as in Lemma 12.1, we obtain $\nu_N(D) \ll \mathrm{h}^\infty$ and $\nu_N(E) \ll \mathrm{h}^{-2N}$. Applying Lemma 12.1 (part (iii)), we conclude that $\mathcal{H}(D \otimes E^\vee) \ll \mathrm{h}^\infty$. □

14.6. Construction of the test function f. By integrating by parts in the Fourier integral defining $(a'_{\mathrm{h}})^\vee$ and taking into account the smoothness and support properties of a', we see that in τ-coordinates $x = (x', x'')$,

$$\partial_x^\alpha (a'_{\mathrm{h}})^\vee(x) \ll \frac{\mathrm{h}^{-|\alpha|}(\mathrm{h}^{-1+\delta'})^{\dim(\mathfrak{g})-\operatorname{rank}(\mathfrak{g})}(\mathrm{h}^{-1+\delta''})^{\operatorname{rank}(\mathfrak{g})}}{\langle \mathrm{h}^{-1+\delta'} x'\rangle^N \langle \mathrm{h}^{-1+\delta''} x''\rangle^N} \tag{14.5}$$

for all fixed α and N, where as before $\langle z\rangle := (1+|z|^2)^{1/2}$. For now, the significance of the numerator in (14.5) is just that it is of size $\mathrm{h}^{-\operatorname{O}(1)}$.

We now truncate $(a'_{\mathrm{h}})^\vee$ to its essential support, as follows. We fix $\nu > 0$ and $\varepsilon > 0$ so that

$$1 - \delta' - \varepsilon > 1/2 + \nu, \quad 1 - \delta'' - \varepsilon > \nu, \tag{14.6}$$

as we may. For $r > 0$, we set

$$\mathcal{D}(r) := \left\{ x \in \mathfrak{g} : |x'| \leqslant r\,\mathrm{h}^{1-\delta'-\varepsilon}, |x''| \leqslant r\,\mathrm{h}^{1-\delta''-\varepsilon} \right\}.$$

Let $\chi \in C_c^\infty(\mathfrak{g})$ be the nice cutoff (§8.2) implicit in the definition of A', thus $A' = \operatorname{Op}_{\mathrm{h}}(a : \pi, \chi)$. We choose another nice cutoff $\tilde{\chi} \in C_c^\infty(\mathfrak{g})$ with

- $\tilde{\chi}(x) = 1$ for $x \in \mathcal{D}(1)$,
- $\tilde{\chi}(x) = 0$ for $x \notin \mathcal{D}(2)$, and
- $\partial^\alpha \tilde{\chi}(x) \ll \mathrm{h}^{-\operatorname{O}(1)}$ for all fixed multi-indices α and all $x \in \mathfrak{g}$.

Informally, $\tilde{\chi}$ is an envelope for the essential support of $(a'_{\mathrm{h}})^\vee$ as quantified by (14.5). Quantitatively, the difference $\chi(a'_{\mathrm{h}})^\vee - \tilde{\chi}(a'_{\mathrm{h}})^\vee$ is negligible in the sense that it is supported on a fixed compact set and has each fixed partial derivative of size $\operatorname{O}(\mathrm{h}^\infty)$. The difference $\widetilde{\operatorname{Op}}_{\mathrm{h}}(a' : \chi) - \widetilde{\operatorname{Op}}_{\mathrm{h}}(a' : \tilde{\chi})$ (see §8 for notation) is then negligible in the analogous sense. Setting

$$\tilde{A}' := \operatorname{Op}_{\mathrm{h}}(a' : \pi, \tilde{\chi})$$

and using that $f \mapsto \pi(f)$ maps fixed bounded subsets of $C_c^\infty(G)$ to $\Psi^{-\infty}(\pi)$ (see [NV, §3.6, Lem 2]), we deduce that the difference $A' - \tilde{A}'$ is negligible in that it lies in $\mathrm{h}^\infty \Psi^{-\infty}(\pi)$. By the proof of lemma 14.4, we deduce that replacing A' with $\tilde{A}'$ modifies the LHS of (14.4) by $\operatorname{O}(\mathrm{h}^\infty)$; in particular, that estimate remains valid for $\tilde{A}'$.

We now construct $f \in C_c^\infty(G)$ by setting

$$f := \widetilde{\operatorname{Op}}_{\mathrm{h}}(a' : \tilde{\chi}), \quad \text{so that} \quad \pi(f) = \tilde{A}'.$$

Strictly speaking, this defines f as a smooth compactly-supported distribution on G; we identify it with an element of $C_c^\infty(G)$ by dividing by our choice of Haar measure dg on G.

Since $\mathcal{D}(2)$ is concentrated near the origin in $\mathfrak{g}$, we see that f is supported near the identity element of G.

14.7. Passage to an individual vector u.

Lemma 14.5. *There exists a smooth unit vector $u \in \sigma$ so that*

$$\sum_{v \in \mathcal{B}(\pi)} \mathcal{Q}(\tilde{A}'v \otimes Bu) \gg \mathrm{h}^{\delta \dim(H) + \kappa/10}. \tag{14.7}$$

Remark 14.6. The passage from (14.4) to (14.7) serves primarily to improve the cosmetics of the arguments of §6. It would be fine to retain the sum over u throughout the whole argument. We would then eventually require an estimate for the Hilbert–Schmidt norm of B, which is anyway the main input in the proof of Lemma 14.5. Thus while Lemma 14.5 is, strictly speaking, unnecessary, it allows us to simplify our presentation with negligible net cost.

The proof appeals to results of §13, which are applicable in view of Lemma 14.1.

Proof of Lemma 14.5. We observe first that the estimate (14.4) holds with B replaced by the identity operator:

$$\sum_{v \in \mathcal{B}(\pi)} \sum_{u \in \mathcal{B}(\sigma)} \mathcal{Q}(\tilde{A}' v \otimes u) \asymp \mathrm{h}^{\delta \dim(H)}. \tag{14.8}$$

To see this, we may reduce as in the proof of (14.4) to verifying the same estimate but with $\tilde{A}'$ replaced by A. We then apply the proof of (14.2), but using (12.4) in place of (12.3). The main term is given by $\int_{s \in H} |a(s \cdot \tau)|^2 \, ds \asymp \mathrm{h}^{\delta \dim(H)}$, while the remaining terms contribute much less.

We observe next that, since b is real-valued, the operator B is self-adjoint (see (8.1)). We now estimate the squared Hilbert–Schmidt norm $\|B\|_2^2$ of B, i.e., the trace of $BB^* = B^2$. To do so, we apply the composition formula (10.9) to write $B^2 = \mathrm{Op}_{\mathrm{h}}(q : \sigma) + \mathcal{E}$, where

- $q \in S_{\delta}^{-\infty}(\mathfrak{h}^\wedge)$ is supported on $\tau_H + \mathrm{O}(\mathrm{h}^{\delta})$, and
- $\mathcal{E} \in \mathrm{h}^N \, \Psi_{\delta}^{-N}(\sigma)$.

By Theorem 10.13, we have $\mathrm{trace}(\mathcal{E}) \ll \mathrm{h}^N$. By the asymptotic form of the Kirillov formula (Theorem 10.11), we have

$$\mathrm{trace}(\mathrm{Op}_{\mathrm{h}}(q : \sigma)) \ll \mathrm{h}^{-(\dim(\mathfrak{h}) - \mathrm{rank}(\mathfrak{h}))/2} \, \mu + \mathrm{h}^N,$$

where μ denotes the symplectic volume of $\mathrm{h}\,\mathcal{O}_\sigma \cap \mathrm{supp}(b)$. To estimate μ, we recall the support conditions stated in §14.3 and apply Lemma 13.3. We obtain

$$\|B\|_2^2 \ll \mathrm{h}^{2(\delta - 1/2)(\dim(\mathfrak{h}) - \mathrm{rank}(\mathfrak{h}))/2} \ll \mathrm{h}^{-\kappa/10},$$

using in the final step that $1/2 - \delta$ is small enough in terms of κ.

We may take for $\mathcal{B}(\sigma)$ a basis of eigenvectors u for B, with eigenvalues λ_u (cf. Lemma 12.1, part (v)). The LHS of (14.2) then reads

$$\sum_{u \in \mathcal{B}(\sigma)} |\lambda_u|^2 \sum_{v \in \mathcal{B}(\pi)} \mathcal{Q}(\tilde{A}' v \otimes u).$$

We have

$$\sum_{u \in \mathcal{B}(\sigma)} |\lambda_u|^2 = \|B\|_2^2 \ll \mathrm{h}^{-\kappa/10}. \tag{14.9}$$

Recall from (3.1) that $\mathcal{Q}$ is definite. For small enough fixed $\varepsilon > 0$, we see by comparing (14.4) and (14.8) that

$$\sum_{\substack{u \in \mathcal{B}(\sigma): \\ |\lambda_u| \geqslant \varepsilon}} |\lambda_u|^2 \sum_{v \in \mathcal{B}(\pi)} \mathcal{Q}(\tilde{A}' v \otimes u) \asymp \mathrm{h}^{\delta \dim(H)}. \tag{14.10}$$

On the other hand, by (14.9), we have $\#\{u \in \mathcal{B}(\sigma) : |\lambda_u| \geqslant \varepsilon\} \ll \mathrm{h}^{-\kappa/10}$. There is thus a unit eigenvector u of B for which (14.7) holds (e.g., any eigenvector whose eigenvalue has maximal size). Since $B \in \underline{\Psi}^{-\infty}(\sigma)$, the vector $u = \lambda_u^{-1} B u$ is smooth (cf. §10.1). The proof is now complete. $\square$

As noted in (14.1), we have $B \in \Psi_\delta^0(\sigma)$, so the operator norm of B is $\mathrm{O}(1)$. We deduce that there is a smooth unit vector $u \in \sigma$ so that

$$\sum_{v \in \mathcal{B}(\pi)} \mathcal{Q}(\tilde{A}' v \otimes u) \gg \mathrm{h}^{\delta \dim(H) + \kappa/10}. \tag{14.11}$$

Remark 14.7. The estimate (14.11) suffices for our purposes, but is less precise than what is given by Lemma 14.5 in that it "forgets" the microlocalization of u at τ_H imposed by B. It may be possible to exploit that microlocalization property to refine the numerical exponent of Theorem 1.1. In the setting of §2.9, this would amount to exploiting more than just the central directions in H_{τ_H}.

14.8. Proof of part (i). We write n for the natural number for which (G, H) is a form of $(\mathrm{U}_{n+1}, \mathrm{U}_n)$, so that $\dim H = [F : \mathbb{R}] n^2$ and $\dim_F H = n^2$. Then $\mathrm{h}^{\dim(H)/2} = T^{-n^2/2}$. The required bound

$$\sum_{v \in \mathcal{B}(\pi)} \mathcal{Q}(\pi(f) v \otimes u) \gg T^{-n^2/2 - \kappa} \tag{4.11}$$

follows from (14.11) upon recalling that δ is close enough to $1/2$ in terms of κ.

14.9. Proof of part (ii). We recall (4.10):

$$f^\sharp(g) := \int_{z \in Z} \omega_\pi(g) (f * f^*)(zg) \, dz. \tag{4.10}$$

Our task is to check that $\int_H |f^\sharp| \ll T^{n/2 + \kappa}$.

By construction, f is supported near the identity element of G and satisfies the support condition that $f(\exp(x))$ vanishes unless $x' \ll \mathrm{h}^{1-\delta'-\varepsilon}$ and $x'' \ll \mathrm{h}^{1-\delta''-\varepsilon}$. These conditions imply in particular that $x \ll \mathrm{h}^\nu$, hence $\exp(x) = 1 + \mathrm{O}(\mathrm{h}^\nu)$ (recall from (14.6) the defining properties of ν). By Lemma 13.5, we deduce further that $\exp(x)\tau = \tau + \mathrm{O}(\mathrm{h}^{1/2+\nu})$. Thus every element g of the support of f satisfies

$$g = 1 + \mathrm{O}(\mathrm{h}^\nu), \quad g\tau = \tau + \mathrm{O}(\mathrm{h}^{1/2+\nu}). \tag{14.12}$$

If g satisfies (14.12), then so does g^{-1}, and if two group elements satisfy (14.12), then so does their product. It follows that every element of the support of $f * f^*$ satisfies (14.12). The same is then true for every element of the support of $|f^\sharp| : \bar{G} \to \mathbb{R}_{\geqslant 0}$, where we now interpret the first estimate in (14.12) as taking place inside the adjoint group $\bar{G}$.

Recall from §7.2 the notation $\dim, \dim_F, \mathrm{rank}, \mathrm{rank}_F$. It follows from the basic Fourier estimate (14.5) that

$$\|f\|_{L^1(G)} \ll \|(a'_{\mathrm{h}})^\vee\|_{L^1(\mathfrak{g})} \ll 1$$

and that

$$\|f\|_{L^\infty(G)} \ll \mathrm{h}^{(-1+\delta')(\dim(\mathfrak{g}) - \mathrm{rank}(\mathfrak{g}))} \mathrm{h}^{(-1+\delta'') \mathrm{rank}(\mathfrak{g})} \ll T^{n(n+1)/2 + \kappa},$$

using here that $\dim_F(\mathfrak{g}) - \mathrm{rank}_F(\mathfrak{g}) = n(n+1)$ and that (δ', δ'') is close enough to $(1/2, 1)$. It follows that

$$\|f * f^*\|_{L^\infty(G)} \leqslant \|f\|_{L^1(G)} \|f\|_{L^\infty(G)} \ll T^{n(n+1)/2+\kappa}$$

and hence also that

$$\|f^\sharp\|_\infty \ll T^{n(n+1)/2+\kappa}, \tag{14.13}$$

using for this last estimate that for a fixed compact subset U of G and all $g \in G$, we have $\mathrm{vol}\{z \in Z : gz \in U\} \ll 1$.

It follows from (14.12) and (14.13) that

$$\int_H |f^\sharp| \ll T^{n(n+1)/2+\kappa} \,\mathrm{vol}\{g \in H : |g\tau - \tau| \leqslant \mathrm{h}^{1/2}\},$$

say. Using the "principal bundle" consequence of stability (§11.2), we see that

$$g \in H,\, |g\tau - \tau| \leqslant \mathrm{h}^{1/2} \quad \Longrightarrow \quad |g-1| \ll \mathrm{h}^{1/2},$$

so the volume in question is $\ll \mathrm{h}^{(\dim H)/2} = T^{-n^2/2}$. The required bound $\int_H |f^\sharp| \ll T^{n/2+\kappa}$ follows.

14.10. Reduction of the proof of part (iii). First, some notation. For $\xi, \eta \in \mathfrak{g}^\wedge$, we set $\mathrm{dist}(\xi,\eta) := |\xi - \eta|$ and $\mathrm{dist}_F(\xi,\eta) := |\xi-\eta|_F$, where as usual $|z|_F = |z|^{[F:\mathbb{R}]}$. For a subset S of $\mathfrak{g}^\wedge$, we set $\mathrm{dist}(\xi, S) := \inf_{\eta \in S} \mathrm{dist}(\xi,\eta)$ and define $\mathrm{dist}_F(\xi,S)$ similarly.

We begin by partitioning the domain of integration. Recall that Ω and Ω' are fixed compact subsets of H, with Ω' large enough in terms of Ω. Let U be a fixed small enough neighborhood of the identity in H. We may cover Ω by $\mathrm{O}(1)$ many translates sU with $s \in \Omega$. We thereby reduce to showing for all $s, t \in \Omega$ that, with $g := s^{-1}\gamma t$,

$$\int_{x,y \in U} \left| \overline{\Psi_1(sx)} \Psi_2(ty) f^\sharp(x^{-1}gy) \right| \, dx\, dy \ll \frac{T^{n/2-1/4+\kappa}}{d_H(\gamma)^{1/2}} \|\Psi_1\|_{L^2(\Omega')} \|\Psi_2\|_{L^2(\Omega')}.$$

We see from (14.12) that $f^\sharp(x^{-1}gy)$ vanishes unless

$$x^{-1}gy = 1 + \mathrm{O}(\mathrm{h}^\nu), \quad \mathrm{dist}(x^{-1}gy\tau, \tau) \ll \mathrm{h}^{1/2+\nu}.$$

Since U is sufficiently small, it follows then that g is close to the identity element of $\bar{G}$ and that $\mathrm{dist}(gy\tau, x\tau) \leqslant \mathrm{h}^{1/2}$, say. Then $\mathrm{dist}_F(gy\tau, x\tau) \leqslant T^{-1/2}$. By (4.8), we have $d_H(\gamma) \asymp d_H(g)$. Setting

$$u_1(x) := |\Psi_1(sx)|, \quad u_2(y) := |\Psi_2(ty)|$$

and invoking the L^∞-bound (14.13), we reduce to showing that

$$\int_{\substack{x,y \in U: \\ \mathrm{dist}(gy\tau, x\tau) \leqslant T^{-1/2}}} u_1(x) u_2(y)\, dx\, dy \ll \frac{T^{-\dim_F(H)/2-1/4}}{d_H(g)^{1/2}} \|\Psi_1\|_{L^2(\Omega')} \|\Psi_2\|_{L^2(\Omega')}.$$

This last estimate is a consequence of the following lemma, applied with $r = T^{-1/2}$.

Lemma 14.8. *Let τ belong to a fixed compact subset of $\mathfrak{g}^\wedge_{\mathrm{stab}}$. Let*

$$u_1, u_2 : H \to \mathbb{R}_{\geqslant 0}$$

be measurable nonnegative-valued functions, with u_2 invariant under translation by Z_H. Let U and U'' be small neighborhoods of the identity in H, with U small enough in terms of U''. Let $g \in \bar{G}$ be close enough to the identity element, and let $r > 0$ be sufficiently small. Then the integral

$$I := \int_{\substack{x,y \in U: \\ \mathrm{dist}_F(gy\tau, x\tau) \leqslant r}} u_1(x) u_2(y)\, dx\, dy$$

satisfies the estimate

$$I \ll \frac{r^{\dim_F(H)+1/2}}{d_H(g)^{1/2}} \|u_1\|_{L^2(U'')} \|u_2\|_{L^2(U'')}$$

Remark 14.9. We give a direct proof involving a few applications of Cauchy–Schwarz. Morally, the lemma should be a consequence of Perron–Frobenius theory as sketched in Remark 2.6. Implementing that sketch rigorously here may lead to a stronger estimate, perhaps with $r^{1/2}/d_H(g)^{1/2}$ replaced with $r/d_H(g)$.

Proof. We may assume given three small neighborhoods U, U', U'' of the identity in H, with U (resp. U') small enough in terms of U' (resp. U'').

We multiply the integrand by the redundant factor $1_{\mathrm{dist}_F(gy\tau, H\tau) \leqslant r}$ and apply Cauchy–Schwarz to see that $I \leqslant \sqrt{I_1 I_2}$, where

$$I_1 := \int_{y \in U} \left(\int_{x \in U} u_1(x) 1_{\mathrm{dist}_F(gy\tau, x\tau) \leqslant r} \, dx \right)^2 \, dy,$$

$$I_2 := \int_{y \in U} u_2(y)^2 1_{\mathrm{dist}_F(gy\tau, H\tau) \leqslant r} \, dy.$$

The claim then follows from the separate estimates given below for I_1 and I_2.

We show first that

$$I_1 \ll r^{2 \dim_F(H)} \|u_1\|^2_{L^2(U')}.$$

To see this, we expand the square, giving

$$I_1 = \int_{x_1, x_2 \in U} u_1(x_1) u_1(x_2) J(x_1, x_2) \, dx_1 \, dx_2$$

with

$$J(x_1, x_2) := \int_{y \in U} 1_{\mathrm{dist}_F(gy\tau, x_1\tau) \leqslant r} 1_{\mathrm{dist}_F(gy\tau, x_2\tau) \leqslant r} \, dy.$$

The integrand in the definition of J is nonzero only if $\mathrm{dist}_F(x_1\tau, x_2\tau) \ll r$ and $x_1, x_2 \ll 1$ and $\mathrm{dist}_F(y\tau, g^{-1}x_1\tau) \ll r$. By the "principal bundle" consequence of stability (§11.2), it follows that $|x_1 - x_2|_F \ll r$ and that y is constrained to a set of volume $\ll r^{\dim_F(H)}$ (depending upon g, x_1, x_2). Thus for some $r' \asymp r$, we have $J(x_1, x_2) \ll r^{\dim_F(H)} 1_{|x_1 - x_2|_F \leqslant r'}$, and so

$$I_1 \ll r^{\dim_F(H)} \int_{\substack{x_1, x_2 \in U: \\ |x_1 - x_2|_F \leqslant r'}} u_1(x_1) u_1(x_2) \, dx_1 \, dx_2.$$

We apply the change of variables $x_1 \mapsto x_2 x_1$ and rename (x_1, x_2) to (y, x), giving

$$\begin{aligned} I_1 &\ll r^{\dim_F(H)} \int_{x, y \in H} 1_U(xy) 1_U(x) 1_{|xy - x|_F \leqslant r'} u_1(xy) u_1(x) \, dx \, dy \\ &\leqslant r^{\dim_F(H)} \int_{x, y \in U'} 1_{|y-1|_F \leqslant r''} u_1(xy) u_1(x) \, dx \, dy \end{aligned}$$

for some $r'' \asymp r'$. We then pull the y-integral outside and apply Cauchy–Schwarz to the inner x-integral, giving

$$I_1 \ll r^{\dim_F(H)} \|u_1\|^2_{L^2(U'')} \int_{y \in U'} 1_{|y-1|_F \leqslant r''} \, dy \ll r^{2 \dim_F(H)} \|u_1\|^2_{L^2(U'')},$$

as required.

We show next that

$$I_2 \ll \frac{r}{d_H(g)} \|u_2\|_{L^2(U'')}^2. \tag{14.14}$$

To see this, we fix a sufficiently small neighborhood $\mathcal{Z}$ of the identity in Z_H and invoke the Z_H-invariance of u_2 to write

$$I_2 = \frac{1}{\mathrm{vol}(\mathcal{Z})} \int_{z \in \mathcal{Z}} \int_{y \in U} u_2(yz^{-1})^2 1_{\mathrm{dist}_F(gy\tau, H\tau) \leqslant r} \, dy \, dz.$$

We now apply the change of variables $y \mapsto yz$. Since U is small in terms of U', the latter contains all products yz^{-1} with $(y, z) \in U \times \mathcal{Z}$. We obtain

$$I_2 \ll \int_{y \in U'} u_2(y)^2 \left(\int_{z \in \mathcal{Z}} 1_{\mathrm{dist}_F(gyz\tau, H\tau) \leqslant r} \, dz \right) dy.$$

Since g, y and hence gy are close to the identity element, we may apply Theorem 15.1 (stated and proved below, independently of what we are proving now) to majorize the parenthetical integral over z by $r/d_H(gy)$. By (4.8), we have $d_H(gy) \asymp d_H(g)$, so the required bound (14.14) follows. $\square$

In summary, we have reduced the proof of our main local result (Theorem 4.2), hence that of our main global result (Theorem 1.1), to that of the volume estimate given below in Theorem 15.1.

15. Volume estimates

15.1. Statement of result. Let (G, H) be a fixed unitary GGP pair over an archimedean local field F. Write, as usual, Z (resp. Z_H) for the center of G (resp. H), $\mathfrak{z}$ (resp. $\mathfrak{z}_H$) for the center of $\mathfrak{g}$ (resp. $\mathfrak{h}$), and $\bar{G} = G/Z$ for the adjoint group. We assume given a fixed Haar measure on Z_H. We note that each of these centers is one-dimensional over F.

Recall from §11.1 the definition of the stable subset $\mathfrak{g}^\wedge_{\mathrm{stab}} \subseteq \mathfrak{g}^\wedge$, as well as the properties of that subset summarized in §11.2. Recall from §4.2 the definition of $d_H : \bar{G} \to [0, 1]$.

We retain the notation dist and dist_F from §14.10.

Theorem 15.1. *Let τ belong to a fixed compact subset of $\mathfrak{g}^\wedge_{\mathrm{stab}}$. Let $\mathcal{Z}$ be a sufficiently small neighborhood of the identity element in Z_H. Let $g \in \bar{G}$ be sufficiently close to the identity element, and let $r > 0$ be sufficiently small. Then*

$$\mathrm{vol}\{z \in \mathcal{Z} : \mathrm{dist}_F(gz\tau, H\tau) \leqslant r\} \ll \frac{r}{d_H(g)}. \tag{15.1}$$

We note that Theorem 15.1 may be formulated over any local field of characteristic zero, with a nearly identical proof.

15.2. Reduction to a size estimate. We may reduce the proof of Theorem 15.1 to that of the following.

Theorem 15.2. *Let τ belong to a fixed compact subset of $\mathfrak{g}^\wedge_{\mathrm{stab}}$. Let $g \in \bar{G}_\tau$ with $g \simeq 1$. Let $z \in Z_H$ with $z \simeq 1$. Then*

$$|z - 1| \cdot |\mathrm{Ad}(g) - 1| \ll \mathrm{dist}(gz\tau, H\tau). \tag{15.2}$$

Proof of Theorem 15.1, assuming Theorem 15.2. By our hypotheses, we may assume that $z \simeq 1$ for all $z \in \mathcal{Z}$, that $g \simeq 1$, and that $r \lll 1$.

We first reduce to showing that the conclusion holds under the additional hypothesis that

$$\operatorname{dist}_F(g\tau, \tau) \ll r. \tag{15.3}$$

Supposing for the moment that it does, let us verify (15.1). In doing so, we may assume that there exists $z_0 \in \mathcal{Z}$ and $h_0 \in H$ with $\operatorname{dist}_F(gz_0\tau, h_0\tau) \leqslant r$, since otherwise the set whose volume we must estimate is empty. Since $g \simeq 1$ and $z_0 \simeq 1$, we then have $h_0\tau \simeq gz_0\tau \simeq \tau$. By stability, it follows that $h_0 \simeq 1$. Setting $g' := h_0^{-1} g z_0 \in \bar{G}$, we see by construction that

$$\operatorname{dist}_F(g'\tau, \tau) = \operatorname{dist}_F(h_0^{-1} g z_0 \tau, \tau) \asymp \operatorname{dist}_F(g z_0 \tau, h_0 \tau) \leqslant r,$$

so g' satisfies our hypothesis (15.3). We see from (4.8) that $d_H(g') \asymp d_H(g)$. Suppose now that $z \in \mathcal{Z}$ satisfies $\operatorname{dist}_F(gz\tau, H\tau) \leqslant r$. Then for some $r' \asymp r$, the element $z' := z_0^{-1} z$ satisfies $\operatorname{dist}_F(g'z'\tau, H\tau) \leqslant r'$. Setting $\mathcal{Z}' := \{z_0^{-1} z : z \in \mathcal{Z}\}$, we see that the hypotheses for (15.1) and (15.3) are satisfied by $(g', \mathcal{Z}', r')$, hence

$$\begin{aligned} \operatorname{vol}\{z \in \mathcal{Z} : \operatorname{dist}_F(gz\tau, H\tau) \leqslant r\} &\leqslant \operatorname{vol}\{z' \in \mathcal{Z}' : \operatorname{dist}_F(g'z'\tau, H\tau) \leqslant r'\} \\ &\ll \frac{r'}{d_H(g')} \asymp \frac{r}{d_H(g)}, \end{aligned}$$

as required.

We now reduce further to the case that

$$g\tau = \tau, \tag{15.4}$$

i.e., $g \in \bar{G}_\tau$. Suppose that $g \in \bar{G}$ satisfies $g \simeq 1$ and (15.3). By Lemma 13.4, we may find $u \in \bar{G}$ so that

$$ug\tau = \tau, \quad |\operatorname{Ad}(u) - 1|_F \ll r.$$

Set $g' := ug$, so that g' satisfies (15.4). For $z \in \mathcal{Z}$, we have

$$\operatorname{dist}_F(g'\tau, g\tau) \ll |\operatorname{Ad}(u) - 1|_F \ll r,$$

hence for some $r' \asymp r$ we have the implication

$$\operatorname{dist}(gz\tau, H\tau) \leqslant r \quad \Longrightarrow \quad \operatorname{dist}(g'z\tau, H\tau) \leqslant r'.$$

We may assume that $r \lll d_H(g)$, because otherwise the required estimate (15.1) follows from the fact that $\operatorname{vol} \mathcal{Z} \ll 1$. Then $|\operatorname{Ad}(u) - 1|_F \lll d_H(g)$, so (4.7) implies that $d_H(g') \asymp d_H(g)$. The estimate (15.1) for $(g', \mathcal{Z}, r')$ thus implies the same estimate for $(g, \mathcal{Z}, r)$. This completes the proof of the claimed reduction.

It remains to consider the case $g \in \bar{G}_\tau$. It follows then from (4.6) and (15.2) that

$$|z - 1|_F \cdot d_H(g) \ll |z - 1|_F \cdot |\operatorname{Ad}(g) - 1|_F \ll \operatorname{dist}_F(gz\tau, H\tau).$$

Since $\dim_F \mathfrak{z}_H = 1$, we have for $r' \lll 1$ that

$$\operatorname{vol}\{z \in \mathcal{Z} : |z - 1|_F \leqslant r'\} \ll r'.$$

We may thus majorize the LHS of (15.1) by $r / d_H(g)$, as required. This completes the proof. □

15.3. Orbit maps for partial actions on some bundles. The proof of Theorem 15.2 requires a bit of setup.

As in §9.4.1, we fix an inner product on $\mathfrak{g}^\wedge$ and equip $\mathfrak{g}$ with the dual inner product, so that the decompositions defined by τ-coordinates ($\tau \in \mathfrak{g}^\wedge_{\mathrm{reg}}$, §9.4.1) are orthogonal.

For $\tau \in \mathfrak{g}^\wedge$, we write

$$\mathcal{O}_\tau := G \cdot \tau$$

for the corresponding coadjoint orbit. We note that

$$T_\tau(\mathcal{O}_\tau) = [\mathfrak{g}, \tau] \subseteq \mathfrak{g}^\wedge. \tag{15.5}$$

By Lemma 13.2, the set of pairs (τ, v) with $\tau \in \mathfrak{g}^\wedge_{\mathrm{reg}}$ and $v \in T_\tau(\mathcal{O}_\tau)$ forms a vector bundle $\tilde{Q}$ over $\mathfrak{g}^\wedge_{\mathrm{reg}}$. The restriction of $\tilde{Q}$ to each coadjoint orbit $\mathcal{O}$ is the tangent bundle of $\mathcal{O}$.

For $\tau \in \mathfrak{g}^\wedge_{\mathrm{stab}}$, we may identify $[\mathfrak{h}, \tau]$ with the tangent space at τ to the subvariety $H \cdot \tau$ of $\mathcal{O}_\tau$. We denote by Q the vector bundle over $\mathfrak{g}^\wedge_{\mathrm{stab}}$ whose fiber at τ is the corresponding quotient, i.e.,

$$Q_\tau := T_\tau(\mathcal{O}_\tau)/[\mathfrak{h}, \tau].$$

Thus Q consists of pairs (τ, v) with $\tau \in \mathfrak{g}^\wedge_{\mathrm{stab}}$ and $v \in Q_\tau$.

In other words, the restriction of Q to the fiber of $\mathfrak{g}^\wedge_{\mathrm{stab}} \to [\mathfrak{g}^\wedge] \times [\mathfrak{h}^\wedge]$ over some stable pair (λ, μ) is the normal bundle to $\mathcal{O}^{\lambda,\mu} = \{\xi : ([\xi], [\xi_H]) = (\lambda, \mu)\}$ inside $\mathcal{O}^\lambda = \{\xi : [\xi] = \lambda\}$. Equivalently, Q_τ is the normal bundle to $H \cdot \tau$ inside $\mathcal{O}_\tau = G \cdot \tau$. The reader might wish to review Figure 2; in that depiction, $T_\tau(\mathcal{O}_\tau)$ (resp. $[\mathfrak{h}, \tau]$) is the tangent space at τ of the sphere (resp. equator), so the fiber Q_τ may be regarded as the one-dimensional space of "vertical" tangent vectors.

We note that if (G, H) is a form of $(\mathrm{U}_{n+1}, \mathrm{U}_n)$, then

$$\dim_F [\mathfrak{h}, \tau] = \dim_F \mathfrak{h} = n^2,$$

$$\dim_F T_\tau(\mathcal{O}_\tau) = \dim_F \mathfrak{g} - \operatorname{rank}_F \mathfrak{g} = n^2 + n,$$

so $\dim_F Q_\tau = n$.

We may use the inner product on $\mathfrak{g}^\wedge$ to identify Q_τ with the orthogonal complement of $[\mathfrak{h}, \tau]$ inside $T_\tau(\mathcal{O}_\tau) = [\mathfrak{g}, \tau]$. We obtain in this way an embedding

$$Q \hookrightarrow \mathfrak{g}^\wedge_{\mathrm{stab}} \times \mathfrak{g}^\wedge.$$

In particular, Q and its fibers Q_τ come with norms defined by the inner product on $\mathfrak{g}^\wedge$.

The restriction of $\tilde{Q}$ to $\mathfrak{g}^\wedge_{\mathrm{stab}}$ surjects onto Q. We obtain for each $\tau \in \mathfrak{g}^\wedge_{\mathrm{stab}}$ a natural quotient map

$$\rho_\tau : T_\tau(\mathcal{O}_\tau) \to Q_\tau.$$

Having identified Q_τ with a subspace of $T_\tau(\mathcal{O}_\tau)$ as above, the map ρ_τ identifies with the orthogonal projection parallel to $[\mathfrak{h}, \tau]$.

We write in what follows $f : A \dashrightarrow B$ for a locally-defined map, i.e., a map $f : U \to B$ defined on some open subset U of A. For instance, the quotient $\tilde{Q} \dashrightarrow Q$ is defined on the open set $\{(\tau, v) : \tau \in \mathfrak{g}^\wedge_{\mathrm{stab}}\} \subseteq \tilde{Q}$.

The coadjoint action of G or its adjoint group $\bar{G}$ on $\mathfrak{g}^\wedge_{\mathrm{reg}}$ extends to the bundle $\tilde{Q}$. The group $\bar{G}$ does not quite act on $\mathfrak{g}^\wedge_{\mathrm{stab}}$, but we have a locally-defined map

$$\bar{G} \times \mathfrak{g}^\wedge_{\mathrm{stab}} \dashrightarrow \mathfrak{g}^\wedge_{\mathrm{stab}} \tag{15.6}$$

$$(g, \tau) \mapsto g \cdot \tau,$$

defined and smooth on the open set

$$\{(g,\tau) : g \cdot \tau \in \mathfrak{g}^{\wedge}_{\text{stab}}\}. \tag{15.7}$$

We now fix a nonzero element $z \in \mathfrak{z}_H$. For each $\tau \in \mathfrak{g}^{\wedge}_{\text{stab}}$, we obtain a tangent vector $[z,\tau] \in [\mathfrak{g},\tau] = T_\tau(\mathcal{O}_\tau)$, hence a point $(\tau, [z,\tau]) \in \tilde{Q}$. By composing the orbit map $\bar{G} \to \tilde{Q}$ at this point with the locally-defined quotient map $\tilde{Q} \dashrightarrow Q$, we obtain a family of locally-defined maps

$$\kappa_\tau : \bar{G} \dashrightarrow Q$$

which fit into a locally-defined product map

$$\kappa : \bar{G} \times \mathfrak{g}^{\wedge}_{\text{stab}} \dashrightarrow Q,$$

defined and smooth over (15.7). Explicitly,

$$\kappa_\tau(g) = \kappa(g,\tau) = (g \cdot \tau, \rho_{g\cdot\tau}(g \cdot [z,\tau])).$$

15.4. A Lie-algebraic result. We recall that $[x,\xi]$ denotes the coadjoint action of $x \in \mathfrak{g}$ on $\xi \in \mathfrak{g}^\wedge$. As usual, $\mathfrak{g}_\xi := \{x \in \mathfrak{g} : [x,\xi] = 0\}$ denotes the centralizer of $\xi \in \mathfrak{g}^\wedge$.

Theorem 15.3. *Let $\tau \in \mathfrak{g}^{\wedge}_{\text{stab}}$, $x \in \mathfrak{g}_\tau - \mathfrak{z}$ and $0 \neq z \in \mathfrak{z}_H$. Then*

$$[x,[z,\tau]] \notin [\mathfrak{h},\tau]. \tag{15.8}$$

We remark that, via the Lefschetz principle, an analogous result holds for any unitary GGP pair over any field of characteristic zero.

The proof of Theorem 15.3 occupies §16. That proof is independent of the remainder of §15.

Corollary 15.4. *For each $\tau \in \mathfrak{g}^{\wedge}_{\text{stab}}$, the derivative $(d\kappa_\tau)_1$ of κ_τ at the identity element of $\bar{G}$ is injective.*

Proof. Let $x \in \mathfrak{g}/\mathfrak{z} = \operatorname{Lie}(\bar{G})$, and suppose that $(d\kappa_\tau)_1(x) = 0$. Then

- $[x,\tau] = 0$, i.e., $x \in \mathfrak{g}_\tau/\mathfrak{z}$, and
- the element $[x,[z,\tau]] = [[x,z],\tau]$ lies in the kernel $[\mathfrak{h},\tau]$ of ρ_τ.

Theorem 15.3 now forces $x = 0$, as required. □

15.5. The implicit function theorem.

Lemma 15.5. *Let $X \subseteq \mathbb{R}^m$ and $Z \subseteq \mathbb{R}^{m'}$ be submanifolds of Euclidean spaces. Let Y be a manifold, and let $f : X \times Y \to Z$ be a smooth map. Let $x_0 \in X$. For $y \in Y$, let $f_y : X \to Z$ denote the map $x \mapsto f(x,y)$. Suppose that for each $y \in Y$, the derivative df_y is injective at x_0. Let Ω be a compact subset of Y. Then there is an open neighborhood U of x_0 and there exist $c_2 > c_1 > 0$ so that*

$$c_1|x_1 - x_2| \leqslant |f(x_1,y) - f(x_2,y)| \leqslant c_2|x_1 - x_2| \tag{15.9}$$

for all $x_1, x_2 \in U$ and $y \in \Omega$, where $|.|$ denotes the restriction of the Euclidean norm.

Proof. Let $y_0 \in \Omega$. Since $(df_{y_0})_{x_0}$ is an injective linear map, we may find $C_2 > C_1 > 0$ so that

$$C_1|v| \leqslant |(df_{y_0})_{x_0}(v)| \leqslant C_2|v| \tag{15.10}$$

for all $v \in \mathbb{R}^m$. Since the product map $(x,y) \mapsto (f(x,y),y)$ is continuously differentiable, we may find open neighborhoods $x_0 \in U \subseteq X$ and $y_0 \in V \subseteq Y$, with U

convex, so that for all $(x, y) \in U \times V$, the operator norm of $(df_y)_x$ differs from that of $(df_{y_0})_{x_0}$ by at most $C_1/10$, say. For $x_1, x_2 \in U$, we define the linear map

$$\gamma : [0,1] \to U, \quad \gamma(0) = x_1, \quad \gamma(1) = x_2.$$

Then $f(x_2, y) - f(x_1, y) = \int_{t=0}^{1} (f_y \circ \gamma)'(t)\, dt$. By the chain rule, $(f_y \circ \gamma)'(t) = (df_y)_{\gamma(t)}(\gamma'(t))$. By construction, $\gamma'(t) = x_2 - x_1$. For $y \in V$, the operator norms of $(df_y)_{\gamma(t)}$ and $(df_{y_0})_{x_0}$ differ by at most $C_1/10$. It follows that

$$|f(x_2, y) - f(x_1, y) - (df_{y_0})_{x_0}(x_2 - x_1)| \leqslant \frac{C_1}{10} |x_2 - x_1|.$$

By the triangle inequality and (15.10), we deduce that (15.9) holds with $c_2 := C_2 + C_1/10$ and $c_1 := 9C_1/10$. We conclude via the compactness of Ω. $\square$

15.6. Application of the implicit function theorem. We retain the notation of §15.3.

Lemma 15.6. *Let Ω be a fixed compact subset of $\mathfrak{g}^{\wedge}_{\mathrm{stab}}$, fix $0 \neq z \in \mathfrak{z}_H$, and let $U \subseteq \bar{G}$ be a fixed sufficiently small neighborhood of the identity element. For $\tau \in \Omega$ and $g \in U \cap \bar{G}_\tau$, we have*

$$|\operatorname{Ad}(g) - 1| \asymp |\rho_\tau([\operatorname{Ad}(g) z, \tau])| \tag{15.11}$$

Proof. It is enough to show that for all $(g, \tau) \in U \times \Omega$, we have

$$|\operatorname{Ad}(g) - 1| \asymp |g \cdot \tau - \tau| + |\rho_{g \cdot \tau}(g \cdot [z, \tau])|. \tag{15.12}$$

The required estimate then follows by noting that for $g \in \bar{G}_\tau$, we have $g \cdot \tau = \tau$ and $g \cdot [z, \tau] = [\operatorname{Ad}(g) z, \tau]$.

To verify (15.12), we apply Lemma 15.5 with

- X any small enough fixed neighborhood of $x_0 := 1 \in \bar{G}$,
- Y any fixed neighborhood of Ω inside $\mathfrak{g}^{\wedge}_{\mathrm{stab}}$,
- $Z := Q$,

and $f := \kappa$, using Corollary (15.4) to verify its hypotheses. The LHS of (15.12) is the Euclidean distance between $g \in \bar{G}$ and $1 \in \bar{G}$ with respect to the Euclidean embedding of $\bar{G}$ defined by the adjoint representation, while the RHS is the Euclidean distance between $\kappa(g, \tau)$ and $\kappa(1, \tau)$ with respect to the embedding $Q \hookrightarrow \mathfrak{g}^{\wedge}_{\mathrm{stab}} \times \mathfrak{g}^{\wedge}$, so the required estimate (15.12) is the content of (15.9). $\square$

15.7. Quantitative Lie-algebraic transversality.

Lemma 15.7. *Retain the hypotheses of Lemma 15.6. For $\tau \in \Omega$ and $g \in U \cap \bar{G}_\tau$, the subspace*

$$W(g) := \mathfrak{h} + \operatorname{Ad}(g) \mathfrak{z}_H \subseteq \mathfrak{g} \tag{15.13}$$

satisfies

$$W(g) \cap \mathfrak{g}_\tau = \{0\}. \tag{15.14}$$

Quantitatively, we have for $x \in W(g)$ with respect to τ-coordinates $x = (x', x'')$ as in §9.4.1 the estimate

$$x' \asymp x. \tag{15.15}$$

Proof. We first verify (15.14). If $g = 1$ (in $\bar{G}$), then $\operatorname{Ad}(g) \mathfrak{z}_H = \mathfrak{z}_H \subseteq \mathfrak{h}$, so $W(g) = \mathfrak{h}$. The claim then follows from the following consequence of the definition of $\mathfrak{g}^{\wedge}_{\mathrm{stab}}$:

$$\mathfrak{h} \cap \mathfrak{g}_\tau = \{0\}. \tag{15.16}$$

If $g \neq 1$ (in $\bar{G}$), then we see from Lemma 15.6 that $\rho_\tau([\mathrm{Ad}(g)z, \tau]) \neq 0$, i.e., that $[\mathrm{Ad}(g)z, \tau]$ is not contained in $[\mathfrak{h}, \tau]$, hence that $\mathrm{Ad}(g)z$ is not contained in $\mathfrak{h} + \mathfrak{g}_\tau$. Since z spans the one-dimensional F-vector space $\mathfrak{z}_H$, we deduce that $\mathrm{Ad}(g)\mathfrak{z}_H \cap (\mathfrak{h} + \mathfrak{g}_\tau) = \{0\}$. We may now conclude (15.14) via (15.16).

We deduce the quantitative assertion (15.15) from a compactness argument, as follows. Let n be such that (G, H) is a form of $(\mathrm{U}_{n+1}, \mathrm{U}_n)$, so that, e.g., $\dim_F \mathfrak{g}_\tau = n + 1$. Let $\mathcal{Y}$ denote the manifold consisting of pairs (Y_1, Y_2) of F-subspaces Y_1, Y_2 of $\mathfrak{g}$ with $\dim_F(Y_1) = n + 1$ and $\dim_F(Y_2) = n^2 + 1$. Thus $\mathcal{Y}$ is a product of two Grassmannians. In particular, $\mathcal{Y}$ is compact. Let $\mathcal{Y}_0 \subseteq \mathcal{Y}$ denote the open submanifold consisting of pairs (Y_1, Y_2) for which $Y_1 \cap Y_2 = \{0\}$. Set

$$\mathcal{X} := \{(\mathfrak{g}_\tau, W(g)) : \tau \in \Omega, 1 \neq g \in \bar{G}_\tau\}.$$

A priori, $\mathcal{X}$ is a subset of $\mathcal{Y}$, but we saw in the proof of (15.14) that in fact $\mathcal{X} \subseteq \mathcal{Y}_0$.

We claim that the closure $\bar{\mathcal{X}}$ of $\mathcal{X}$ inside the compact space $\mathcal{Y}$ is actually contained in $\mathcal{Y}_0$:

$$\bar{\mathcal{X}} \subseteq \mathcal{Y}_0. \tag{15.17}$$

The claim implies the required estimate (15.15). Indeed, for $(Y_1, Y_2) \in \mathcal{Y}_0$, write $Y_1^\flat \subseteq \mathfrak{g}$ for the orthogonal complement of Y_1. For $x \in \mathfrak{g}$, write $x = x' + x''$ with $x' \in Y_1^\flat, x'' \in Y_1$ (thus if $Y_1 = \mathfrak{g}_\tau$, then these are the τ-coordinates). Since $Y_1 \cap Y_2 = \{0\}$, we see that $x' \neq 0$ for all $0 \neq x \notin Y_2$. The minimum over $0 \neq x \in Y_2$ of $|x'|/|x|$ defines a continuous function $\mathcal{Y}_0 \to \mathbb{R}_{>0}$. The claim (15.17) implies that this function is defined and bounded from below on the compact subset $\bar{\mathcal{X}}$ by some fixed $c > 0$. The estimate (15.15) follows.

It remains to verify the claim (15.17). Let (τ_j, g_j) be a sequence of pairs with $\tau_j \in \Omega$ and $g_j \in U \cap (G_{\tau_j}/Z - \{1\})$. It is enough to check that any subsequential limit in $\mathcal{Y}$ of $(g_{\tau_j}, W(g_j))$ belongs to $\mathcal{Y}_0$. To that end, we may assume that τ_j and g_j tend to respective limits $\tau \in \Omega$ and $g \in \bar{U}$. If $g \neq 1$, then the pair $(\mathfrak{g}_{\tau_j}, W(g_j) \in \mathcal{X}$ tends to $(\mathfrak{g}_\tau, W(g)) \in \mathcal{X}$. (Note that since we have taken U small enough, the identity (15.14) persists on $\bar{U} \times \Omega$.) It remains to consider the case $g = 1$. Let S denote the unit sphere in $\mathfrak{g}/\mathfrak{z}$ with respect to some fixed norm. For j large enough, we may write $g_j = \exp(t_j x_j)$, where $t_j \in F^\times$ and $x_j \in S$. Thus $t_j \to 0$. We may assume that x_j tends to some limit $x \in S$. Since $x_j \in \mathfrak{g}_{\tau_j}/\mathfrak{z}$, we have $x \in \mathfrak{g}_\tau/\mathfrak{z}$. Since $x \neq 0$, we know from Theorem 15.3 that $[[x, z], \tau] \notin [\mathfrak{h}, \tau]$. In particular, $[x, z] \notin \mathfrak{h}$. Since z spans the one-dimensional F-vector space $\mathfrak{z}_H \subseteq \mathfrak{h}$, we have

$$W(g_j) = \mathfrak{h} + F \,\mathrm{Ad}(g_j)z = \mathfrak{h} + F(\mathrm{Ad}(g_j)z - z).$$

By the Taylor approximation $t_j^{-1}(\mathrm{Ad}(g_j)z - z) = [x_j, z] + \mathrm{O}(t_j)$, we deduce that the limit of $W(g_j)$ exists and equals $\mathfrak{h} + [x, \mathfrak{z}_H]$. By another application of Theorem 15.3, we see that this last space intersects $\mathfrak{g}_\tau$ trivially. Thus the limit of $(g_{\tau_j}, W(g_j))$ in $\mathcal{Y}$ exists and belongs to $\mathcal{Y}_0$, as required. $\square$

15.8. Exponential coordinates on a coadjoint orbit. Let $\tau \in \mathfrak{g}^\wedge_{\mathrm{reg}}$. Recall that the tangent space to the coadjoint orbit $\mathcal{O}_\tau$ at τ may be described as

$$T_\tau(\mathcal{O}_\tau) = [\mathfrak{g}, \tau] = \mathfrak{g}_\tau^\perp.$$

We have $[\mathfrak{g}, \tau] \cong \mathfrak{g}/\mathfrak{g}_\tau$, so the choice of subspace W of $\mathfrak{g}$ complementary to $\mathfrak{g}_\tau$ (e.g., $W = \mathfrak{g}_\tau^\flat$) induces a map

$$\exp_{(\tau, W)} : \mathfrak{g}_\tau^\perp \to \mathcal{O}_\tau$$

as follows. Each element of $\mathfrak{g}_\tau^\perp$ may be written uniquely in the form $[x, \tau]$ for some x in our chosen complement W. We then set

$$\exp_{(\tau,W)}[x, \tau] := \exp(x)\tau \quad (x \in W).$$

Informally, the following estimate says that $\exp_{(\tau,W)}$ approximately preserves distances near the origin provided that W is "separated enough" from $\mathfrak{g}_\tau$. (Conversely, one can check that such an assumption on W is necessary.)

Lemma 15.8. *Let τ belong to a fixed compact subset of $\mathfrak{g}^\wedge_{\mathrm{reg}}$. Let W be a subspace of $\mathfrak{g}$ complementary to $\mathfrak{g}_\tau$ such that for all $x \in W$, we have with respect to τ-coordinates as in §9.4.1 the estimate (15.15). Then for all $\xi, \eta \in \mathfrak{g}_\tau^\perp$ with $\xi, \eta \lll 1$, we have*

$$|\xi - \eta| \asymp |\exp_{(\tau,W)} \xi - \exp_{(\tau,W)} \eta|. \tag{15.18}$$

Proof. We may write $\exp_{(\tau,W)}(\xi) = \exp(c(\xi))\tau$, where c denotes the composition

$$c : \mathfrak{g}_\tau^\perp = [\mathfrak{g}, \tau] \cong \mathfrak{g}/\mathfrak{g}_\tau \cong W \hookrightarrow \mathfrak{g}.$$

Using coordinates as in §9.4.2, we may identify $\mathfrak{g}_\tau^\perp$, $\mathfrak{g}$ and $\mathfrak{g}^\wedge$ with Euclidean spaces; in particular, we may speak of sizes of partial derivatives. In view of (15.15), the linear map c satisfies $c(\xi) \asymp \xi$, and so each fixed partial derivative of that map is $\mathrm{O}(1)$. Since $\tau \ll 1$, we deduce that on arguments $\xi \lll 1$, each fixed partial derivative of $\exp_{(\tau,W)}$ is $\mathrm{O}(1)$. We see from the Taylor expansion

$$\exp_{(\tau,W)}[x, \tau] = \tau + [x, \tau] + \sum_{j \geqslant 2} \frac{1}{j!} (\mathrm{ad}_x^*)^{j-1}[x, \tau]$$

that the derivative $(d\exp_{(\tau,W)})_0$ of $\exp_{(\tau,W)}$ at the origin is the identity map $1 : \mathfrak{g}_\tau^\perp \hookrightarrow \mathfrak{g}^\wedge$. By the fundamental theorem of calculus and the noted $\mathrm{O}(1)$ bound for the second partial derivatives of $\exp_{(\tau,W)}$ near the origin, we deduce that

$$(d\exp_{(\tau,W)})_\xi = 1 + o(1) \tag{15.19}$$

for all $\xi \lll 1$. The required estimate now follows from the fundamental theorem of calculus applied to the linear map

$$\gamma : [0, 1] \to \mathfrak{g}_\tau^\perp, \quad \gamma(0) = \xi, \quad \gamma(1) = \eta$$

and its composition with $\exp_{(\tau,W)}$. $\square$

15.9. Proof of Theorem 15.2. We recall the setting of Theorem 15.2: τ belongs to some fixed compact subset Ω of $\mathfrak{g}^\wedge_{\mathrm{stab}}$, $g \in \bar{G}_\tau$ with $g \simeq 1$, and $z \in Z_H$ with $z \simeq 1$.

It is enough to show for each $h \in H$ that

$$|z - 1| \cdot |\operatorname{Ad}(g) - 1| \ll \operatorname{dist}(gz\tau, h\tau).$$

Since $z \simeq 1$ and $g \simeq 1$, the required estimate is trivial unless $h\tau \simeq gz\tau$. Since $gz\tau \simeq \tau$, it follows that $h\tau \simeq \tau$, hence by the "principal bundle" consequence of stability (see §11.2) that $h \simeq 1$. In particular, we may write $z = \exp(\log(z))$ and $h = \exp(\log(h))$ for some $\log(z) \in \mathfrak{z}_H$ and $\log(h) \in \mathfrak{h}$ with $\log(z), \log(h) \simeq 0$. Then

$$gz\tau = gzg^{-1}\tau = \exp(\operatorname{Ad}(g)\log(z))\tau, \quad h\tau = \exp(\log(h))\tau.$$

Let U be as in Lemma 15.7. Since $g \simeq 1$, we have in particular $g \in U$. Set $W_0 := W(g)$ as in (15.13), so that W_0 is a subspace of $\mathfrak{g}$ that satisfies

$$W_0 \supseteq \mathfrak{h} \cup \operatorname{Ad}(g)\mathfrak{z}_H$$

and, by (15.14),

$$W_0 \cap \mathfrak{g}_\tau = \{0\}.$$

Let $W_1 \subseteq \mathfrak{g}$ denote the orthogonal complement of $W_0 \oplus \mathfrak{g}_\tau$. Set $W := W_0 \oplus W_1$. Then W is a subspace of $\mathfrak{g}$ complementary to $\mathfrak{g}_\tau$. By (15.15), we have with respect to τ-coordinates that the estimate (15.15) holds for all $x \in W_0$. For $x \in W_1$, we have $x' = x$, so the estimate (15.15) persists for $x \in W$.

Since $g, z, h \simeq 1$ and $\mathrm{Ad}(g)\log(z), \log(h) \in W$, we deduce from lemma 15.8 that

$$|gz\tau - h\tau| \asymp |[\mathrm{Ad}(g)\log(z), \tau] - [\log(h)\tau]|.$$

Since $|\rho_\tau(\xi)|$ is the *minimum* over $x \in \mathfrak{h}$ of $|\xi - [x,\tau]|$, it follows that

$$|gz\tau - h\tau| \gg |\rho_\tau([\mathrm{Ad}(g)\log(z), \tau])| \overset{(15.11)}{\asymp} |z - 1| \cdot |\mathrm{Ad}(g) - 1|,$$

which is the required estimate.

16. Lie-algebraic considerations

In this section we prove Theorem 15.3, which is all that remains. We retain the notation and setting of that theorem; in particular, (G, H) is a unitary GGP pair over an archimedean local field F.

16.1. Reduction to a stability characterization. We may reduce the proof of Theorem 15.3 to that of the following.

Theorem 16.1. *Let $\tau \in \mathfrak{g}^\wedge$ and $0 \neq z \in \mathfrak{z}_H$. The following are equivalent.*

(i) τ is stable.

(ii) Both $\tau \in \mathfrak{g}^\wedge$ and its restriction $\tau_H \in \mathfrak{h}^\wedge$ are regular, and the bilinear form

$$\begin{aligned} \mathfrak{g}_\tau/\mathfrak{z} \otimes \mathfrak{h}_{\tau_H} &\to F \\ (x,y) &\mapsto \langle [z,[x,y]], \tau\rangle \end{aligned} \tag{16.1}$$

is nondegenerate.

Proof of Theorem 15.3, assuming Theorem 16.1. We first verify (using that $\tau \in \mathfrak{g}^\wedge_{\mathrm{stab}}$) the identity

$$[\mathfrak{h}, \tau] = \{\xi \in \mathfrak{g}_\tau^\perp : \xi|_{\mathfrak{h}_{\tau_H}} = 0\} =: \mathfrak{g}_\tau^\perp \cap \mathfrak{h}_{\tau_H}^\perp, \tag{16.2}$$

where as usual $\mathfrak{g}_\tau^\perp := \{\xi \in \mathfrak{g}^\wedge : \langle x, \xi\rangle = 0 \text{ for all } x \in \mathfrak{g}_\tau\}$. This identity is implicit in [NV, §16]; we collect the proof here for convenience. We recall from §11.2 that

- the map $\mathfrak{g}^\wedge_{\mathrm{stab}} \to \{\text{stable } (\lambda,\mu) \in [\mathfrak{g}^\wedge] \times [\mathfrak{h}^\wedge]\}$ is a principal H-bundle over its image, and that
- the fiber of that map over (λ, μ), if nonempty, is the H-torsor $\mathcal{O}^{\lambda,\mu}$ defined as in (11.1).

It follows that

- $[\mathfrak{h}, \tau]$ is the tangent space $T_\tau(H \cdot \tau)$ at τ to the subvariety $H \cdot \tau$ of $\mathfrak{g}^\wedge$, and
- $H \cdot \tau = \{\xi \in \mathfrak{g}^\wedge : [\xi] = [\tau], [\xi_H] = [\tau_H]\}$.

Since τ is stable, we see from the "principal bundle" assertion that the map $\mathfrak{g}^\wedge \to [\mathfrak{g}^\wedge] \times [\mathfrak{h}^\wedge]$ has surjective derivative at τ. Thus $T_\tau(H \cdot \tau)$ identifies with the kernel of that derivative. Since τ and τ_H are regular (lemma 14.1), the maps $\mathfrak{g}^\wedge \to [\mathfrak{g}^\wedge]$ and $\mathfrak{h}^\wedge \to [\mathfrak{h}^\wedge]$ have surjective derivatives at τ and τ_H, respectively (theorem 13.1). The kernels of those derivatives are thus $T_\tau(G \cdot \tau) = \mathfrak{g}_\tau^\perp$ and $T_{\tau_H}(H \cdot \tau_H) = \mathfrak{h}_{\tau_H}^\perp$. The required identity follows.

We now deduce (15.8). Let $x \in \mathfrak{g}_\tau - \mathfrak{z}$ and $0 \neq z \in \mathfrak{z}_H$. Suppose that $[x,[z,\tau]]$ lies in $[\mathfrak{h},\tau]$. By (16.2), $[x,[z,\tau]]$ is in particular orthogonal to $\mathfrak{h}_{\tau_H}$, and so for any $y \in \mathfrak{h}_{\tau_H}$, we have

$$0 = \langle y, [x,[z,\tau]]\rangle = -\langle [x,y],[z,\tau]\rangle = \langle [z,[x,y]],\tau\rangle.$$

Thus the image of x in $\mathfrak{g}_\tau/\mathfrak{z}$ is a nonzero element of the kernel of the pairing (16.1). But since τ is stable, we know by Theorem 16.1 that that pairing is nondegenerate. We thereby obtain the required contradiction. $\square$

16.2. Reduction to a determinantal identity. Condition (ii) of Theorem 16.1 asserts the insolvability of a system $\mathcal{S}$ of linear equations over F. The association $(G,H) \mapsto \mathcal{S}$ is compatible with base extension, i.e., replacing (G,H) with the corresponding GGP pair over an extension field of F. In proving Theorem 16.1, there is thus no loss of generality in assuming that F is algebraically closed (i.e., $F = \mathbb{C}$). As explained in §3.4, we may assume then that

$$(G,H) = (\mathrm{GL}_{n+1}(F), \mathrm{GL}_n(F)), \quad (\mathfrak{g},\mathfrak{h}) = (\mathfrak{gl}_{n+1}(F), \mathfrak{gl}_n(F)), \tag{16.3}$$

with GL_n included in GL_{n+1} in the usual way as the upper-left $n \times n$ block. We denote by $\langle,\rangle$ the trace pairing on $\mathfrak{gl}_{n+1}(F)$. Using the trace pairing multiplied by some fixed imaginary unit $\pm i$, we may G-equivariantly identify $\mathfrak{g}^\wedge = \mathfrak{g}$ and H-equivariantly identify $\mathfrak{h}^\wedge = \mathfrak{h}$. The restriction map $\mathfrak{g}^\wedge \to \mathfrak{h}^\wedge$, $\xi \mapsto \xi_H$ then identifies with the map $\mathfrak{g} \to \mathfrak{h}, x \mapsto x_H$ given by

$$x_H := \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix} \text{ if } x = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \tag{16.4}$$

where a,b,c,d have respective dimensions $n \times n, n \times 1, 1 \times n, 1 \times 1$.

We define what it means for an element of $\mathfrak{g}$ to be stable exactly as we did for $\mathfrak{g}^\wedge$. Since the identification $\mathfrak{g} = \mathfrak{g}^\wedge$ is equivariant under G, hence under H, this definition is compatible with our earlier definition. Our task is then equivalent to showing that holds (15.8) for stable $\tau \in \mathfrak{g}$. The advantage of switching from $\mathfrak{g}^\wedge$ to $\mathfrak{g}$ is that matrix multiplication equips $\mathfrak{g}$ with the structure of an associative algebra.

We write $1 \in \mathfrak{g}$ for the identity matrix and $1_H \in \mathfrak{h}$ for its image under the map $\mathfrak{g} \to \mathfrak{h}$ described in (16.4). Explicitly,

$$1_H = \mathrm{diag}(1,\dots,1,0).$$

We note that 1_H defines a basis element for the center $\mathfrak{z}_H$ of $\mathfrak{h}$.

We introduce the abbreviation

$$\tau_H^j := 1_H(\tau_H)^j.$$

The factor 1_H is significant only when $j = 0$, in which case it ensures that $\tau_H^0 = 1_H$ (i.e., that the zeroth power is taken inside the algebra $\mathfrak{h}$ rather than in $\mathfrak{g}$). We will never refer to the other potential interpretation $(\tau^j)_H$ of the LHS.

For $\tau \in \mathfrak{g}$, let $\Delta(\tau)$ denote the determinant of the $n \times n$ matrix $D(\tau)$ with (i,j) entry $(1 \leqslant i,j \leqslant n)$ given by

$$D(\tau)_{ij} := \langle [1_H, [\tau^i, \tau_H^{j-1}]], \tau\rangle.$$

Let $\Delta_0(\tau)$ denote the following normalization of the resultant of the characteristic polynomials of $\tau \in \mathfrak{g}$ and $\tau_H \in \mathfrak{h}$:

$$\Delta_0(\tau) = \prod_{i=1}^{n+1}\prod_{j=1}^{n}(\lambda_i - \mu_j), \tag{16.5}$$

where $\{\lambda_1, \dots, \lambda_{n+1}\}$ and $\{\mu_1, \dots, \mu_n, 0\}$ denote the multisets of eigenvalues of $\tau, \tau_H \in \mathfrak{g}$.

Theorem 16.2. *We have $\Delta(\tau) = 2^n (-1)^{n(n-1)/2} \Delta_0(\tau)$ for all $\tau \in \mathfrak{g}$.*

Proof of Theorem 16.1, assuming Theorem 16.2. Let $\tau \in \mathfrak{g}$. It is clear from the stability characterization noted in Lemma 11.4 that $\Delta_0(\tau) \neq 0$ if and only if τ is stable. We conclude via the following lemma. $\square$

Lemma 16.3. *We have $\Delta(\tau) \neq 0$ if and only if condition (ii) of Theorem 16.1 holds.*

Proof. Suppose first that $\Delta(\tau) \neq 0$. If τ is irregular, then its minimal polynomial has degree strictly less than $n+1$, so we may find coefficients $c_i \in F$ (not all zero) so that $x := \sum_{i=1}^n c_i \tau^i$ belongs to $\mathfrak{z}$. The row vector $(c_1, \dots, c_n)$ then lies in the left kernel of the matrix $D(\tau)$. Similarly, if τ_H is irregular, then we may find coefficients $c_j \in F$ (not all zero) so that $\sum_{j=1}^n c_j \tau_H^{j-1} = 0$, in which case the column vector $(c_1, \dots, c_n)^t$ lies in the right kernel of $D(\tau)$. In either case, we obtain the contradiction that $\Delta(\tau) = 0$. Thus τ and τ_H are both regular. In that case,

$$\{\tau^i : 1 \leqslant i \leqslant n\} \quad \text{and} \quad \{\tau_H^{j-1} : 1 \leqslant j \leqslant n\}$$

give bases for $\mathfrak{g}_\tau / \mathfrak{z}$ and $\mathfrak{h}_{\tau_H}$, respectively. Since z and 1_H are nonzero multiples of one another, we deduce that the nondegeneracy of the indicated bilinear form is equivalent to the nonvanishing of $\Delta(\tau)$.

Conversely, if $\Delta(\tau) = 0$, then we see from the previous sentence that, even if τ and τ_H are regular, the indicated bilinear form is degenerate. Thus condition (ii) of Theorem 16.1 fails. $\square$

Remark 16.4. Theorem 16.2 yields the determinantal characterization

$$\mathfrak{g}_{\mathrm{stab}} = \{\tau \in \mathfrak{g} : \Delta(\tau) \neq 0\}$$

of the stable subset. Rallis–Schiffmann [RS, Prop 6.1] obtained a different determinantal characterization of a set closely related to $\mathfrak{g}_{\mathrm{stab}}$. Their arguments give

$$\mathfrak{g}_{\mathrm{stab}} = \{\tau \in \mathfrak{g} : \Delta_{\mathrm{RS}}(\tau) \neq 0\},$$

where $\Delta_{\mathrm{RS}}(\tau)$ denotes the determinant of the matrix $D_{\mathrm{RS}}(\tau)$ with (i,j) entry

$$D_{\mathrm{RS}}(\tau)_{ij} = \langle e_{n+1}^*, \tau^{i+j-2} e_{n+1} \rangle \quad (1 \leqslant i, j \leqslant n),$$

with $e_1, \dots, e_{n+1}$ and $e_1^*, \dots, e_{n+1}^*$ the standard basis and dual basis for F^{n+1}. (In the notation used below, $D_{\mathrm{RS}}(\tau)_{ij} = q\tau^{i+j-2}p$.) Wei Zhang has pointed out to us that one can use the identity (16.19) stated below and some elementary row/column operations to show that $\Delta_{\mathrm{RS}}(\tau) = \pm 2^n \Delta(\tau)$. This observation yields a shorter overall proof of Theorem 16.1, the result of primary interest. One can show in turn that $\Delta_{\mathrm{RS}}(\tau) = \pm \Delta_0(\tau)$ by explicit calculation in the case that τ_H is regular semisimple, yielding an alternative proof of Theorem 16.2. We suspect that it may be possible to give further proofs using the many standard determinantal interpretations of resultants (Sylvester matrix, Bézout matrix, etc.). We retain below our original treatment, which we emphasize is one among many possibilities. Finally, it may comfort the skeptical reader to note that the statement of Theorem 16.2 was discovered via computer algebra and has been thoroughly tested.

16.3. Verification of the determinantal identity. In this section we prove Theorem 16.2.

Before diving into the details, we record an overall roadmap. Recall the Vandermonde determinantal identity:

$$P(x) := \det\left((x_i^{j-1})_{i,j=1}^n\right) = P_0(x) := \prod_{1 \leqslant i < j \leqslant n} (x_j - x_i).$$

One of the standard proofs consists of three steps:

(1) *Divisibility.* Viewing both sides as polynomials in the variables $x_1, \dots, x_n$, show that P_0 divides P. The key observation is that $P(x)$ vanishes whenever some $x_i = x_j$ $(i \neq j)$.
(2) *Homogeneity.* Both sides are polynomials of the same degree, so the divisibility forces $P(x) = cP_0(x)$ for some constant c.
(3) *Evaluation of the constant.* Comparing the coefficients of some well-chosen monomial, deduce that $c = 1$. One can formulate this step in other ways, e.g., as a comparison of asymptotics as $x_{i+1}/x_i \to \infty$.

The proof given below of Theorem 16.2 follows the same overall strategy, with Lemmas 16.7, 16.8 and 16.10 fulfilling the three indicated steps.

Turning to details, we retain the setup of §16.2, but replacing (16.3) with

$$G = \mathrm{GL}(V), \quad H = \mathrm{GL}(V_H), \quad \mathfrak{g} = \mathfrak{gl}(V) = \mathrm{End}(V), \quad \mathfrak{h} = \mathfrak{gl}(V_H) = \mathrm{End}(V_H),$$

where V is an $(n+1)$-dimensional vector space equipped with a decomposition $V = V_H \oplus Fe$ into an n-dimensional subspace V_H and a one-dimensional subspace Fe spanned by some nonzero vector $e \in V$. Thus (16.3) corresponds to the case

$$V = F^{n+1}, \quad V_H = F^n = \{(*, \dots, *, 0)\} \subseteq V, \quad e = (0, \dots, 0, 1). \tag{16.6}$$

The reason for introducing this abstraction is that it will be convenient for us to work with multiple choices of coordinates throughout the proof.

We denote as usual by $[\mathfrak{g}]$ and $[\mathfrak{h}]$ the geometric invariant theory quotients. The maps $\mathfrak{g} \to [\mathfrak{g}]$ and $\mathfrak{h} \to [\mathfrak{h}]$ may then be regarded as sending a matrix to the coefficients of its characteristic polynomial (see, e.g., [NV, §13.4.1]). Let R denote the coordinate ring of $[\mathfrak{g}] \times [\mathfrak{h}]$. It is a free polynomial ring in $2n+1$ variables. We may identify R with

$$F[\lambda_1, \dots, \lambda_{n+1}, \mu_1, \dots, \mu_n]^{S(n+1) \times S(n)},$$

where the λ_i and μ_j are algebraically independent variables, corresponding to the generalized eigenvalues of a matrix, while $S(n+1)$ and $S(n)$ are symmetric groups acting by permutations on $\{\lambda_i\}$ and $\{\mu_j\}$, respectively. Let

$$\iota : \mathfrak{g} \to [\mathfrak{g}] \times [\mathfrak{h}]$$

denote the map assigning to $\tau \in \mathfrak{g}$ the invariants of the pair (τ, τ_H). The map ι is surjective (see for instance [NV, §14.3, Lemma 1]), so we may use the pullback ι^* to identify R with a space of functions $\mathfrak{g} \to F$. For instance, it is clear from (16.5) that $\Delta_0 \in R$.

Lemma 16.5. *We have $\Delta \in R$.*

Proof. We work with the standard coordinates (16.6). We regard F^{n+1} as a space of column vectors. We set $p := e$ (a column vector) and $q := e^t$ (a row vector).

Then, multiplying matrices and vectors in the usual way, we have

$$\mathfrak{h} = \{x \in \mathfrak{g} : xp = 0, qx = 0\}, \tag{16.7}$$

$$1_H = 1 - pq. \tag{16.8}$$

We observe that for each $j \in \mathbb{Z}_{\geqslant 0}$, we have

$$[\tau \mapsto q\tau^j p] \in R. \tag{16.9}$$

This fact was observed (implicitly) in [NV, §14] to follow from the identity of formal power series

$$\sum_{j \in \mathbb{Z}_{\geqslant 0}} t^j q \tau^j p = q(1 - t\tau)^{-1} p = \frac{\det(1 - t\tau_H)}{\det(1 - t\tau)},$$

which is itself a consequence of Cramer's rule.

We thereby reduce to verifying that each matrix entry $D(\tau)_{ij}$ is a polynomial combination of the quantities $q\tau^j p$. To that end, we apply (16.8) and the fact that 1 commutes with everything to see that for any $x \in \mathfrak{g}$, we have

$$\begin{aligned} \langle [1_H, x], \tau \rangle &= -\langle [pq, x], \tau \rangle \\ &= -\langle pq, [x, \tau] \rangle \\ &= q[\tau, x]p. \end{aligned} \tag{16.10}$$

We then specialize this observation to $x = [\tau^i, \tau_H^{j-1}]$, so that $D_{ij}(\tau) = \langle [1_H, x], \tau \rangle$, and expand out $\tau_H = 1_H \tau 1_H$ and $1_H = 1 - pq$ and all commutators. $\square$

The following result is not necessary for our purposes, but seems worth recording.

Lemma 16.6. *The ring R consists of all H-invariant polynomial functions $\mathfrak{g} \to F$, i.e., ι defines the geometric invariant theory quotient of $\mathfrak{g}$ by H.*

Proof. Let S denote the ring of H-invariant polynomial functions $\mathfrak{g} \to F$. We must check that $R = S$. Clearly $R \subseteq S$. Conversely, it follows from [Zh1, Lem 3.1] that S is generated by

- the coefficients of the characteristic polynomial of τ, and
- the quantities $q\tau^j p$ with $j \in \{1, \dots, n\}$ as in (16.9).

By (16.9), we conclude that $S \subseteq R$. $\square$

We next verify "one divisibility" in the required identity relating Δ and Δ_0.

Lemma 16.7. *Δ_0 divides Δ in the ring R.*

Proof. The ideal (Δ_0) is generated by a squarefree element of the unique factorization domain R, hence is radical, so by the Nullstellensatz, it is enough to verify that

$$\Delta_0(\tau) = 0 \implies \Delta(\tau) = 0.$$

Suppose that $\Delta_0(\tau) = 0$, i.e., that $\tau \in \mathfrak{g}$ is not stable. We must check that $\Delta(\tau) = 0$. In view of Lemma 16.3, it is enough to check that condition (ii) of Theorem 16.1 fails. Suppose otherwise, for the sake of contradiction, that it holds. In particular, τ and τ_H are regular.

Since τ and τ_H are not stable, we see from [NV, §14.2, Lem] that either

- τ and τ_H admit a common eigenvector, or
- their transposes admit a common eigenvector.

We consider below the first case. The second case may be treated similarly by applying the same argument to the transposes.

Since τ_H is regular, we may decompose V_H into generalized eigenspaces $W_1, \dots, W_m$ for τ_H, with distinct eigenvalues $\mu_1, \dots, \mu_m$, and with the action of τ_H on W_j given with respect to some basis by a standard Jordan block, given in the case $\dim(W_j) = 3$ (say) by

$$\tau_H|_{W_j} \sim \begin{pmatrix} \mu_j & 1 & 0 \\ & \mu_j & 1 \\ & & \mu_j \end{pmatrix}. \tag{16.11}$$

The operator τ_H has exactly m eigenspaces, each one-dimensional, one for each summand W_j.

By relabeling, we may suppose that τ has an eigenvector in W_1. Let $e_1, \dots, e_{\dim(W_1)}$ be a basis of W_1 with respect to which $\tau_H|_{W_1}$ is described by a matrix as in (16.11). Then e_1 is an eigenvector of τ. We extend to a basis $e_1, \dots, e_n$ of V_H and then further to a basis $e_1, \dots, e_{n+1}$ of V.

The centralizer $\mathfrak{h}_{\tau_H}$ is the direct sum over $j \in \{1, \dots, m\}$ of the $\mathfrak{gl}(W_j)$-centralizer of $\tau_H|_{W_j}$, the latter of which consists in the case $\dim(W_j) = 3$ (say) of all matrices of the form

$$\begin{pmatrix} a & b & c \\ 0 & a & b \\ 0 & 0 & a \end{pmatrix}.$$

There is thus a unique (nonzero) element $y \in \mathfrak{h}_{\tau_H}$ for which

- $y|_{W_j} = 0$ for $j \neq 1$, and
- $y|_{W_1} e_i = 0$ for $1 \leqslant i < \dim(W_1)$ and $y|_{W_1} e_{\dim(W_1)} = e_1$.

For instance, in the case $\dim(W_1) = 3$ (say),

$$y|_{W_1} \sim \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Let x be any element of g_τ. We claim that

$$[1_H, [x, y]]_{ij} = 0 \text{ unless } (i, j) = (1, n+1). \tag{16.12}$$

Indeed, since τ is regular, we know that the element $x \in \mathfrak{g}_\tau$ is a linear combination of powers of τ, and so e_1 is an eigenvector of x. Thus $x_{i,1} \neq 0$ only if $i = 1$. By the construction of y and some matrix multiplication, we deduce that $[x, y]_{ij} \neq 0$ only if $i = 1$. For any $z \in \mathfrak{g}$, we have

$$[1_H, z]_{ij} = \begin{cases} 0 & \text{if } i \leqslant n, j \leqslant n, \\ 0 & \text{if } i = j = n+1, \\ z_{ij} & \text{if } i \leqslant n, j = n+1, \\ -z_{ij} & \text{if } i = n+1, j \leqslant n. \end{cases}$$

Applying this observation to $z = [x, y]$ yields the claim.

We deduce from (16.12) that $\langle [1_H, [x, y]], \tau \rangle$ is a multiple of $\tau_{n+1,1}$, which vanishes in view of our assumption that e_1 is a τ-eigenvector. Since y is nonzero and this last property holds for all $x \in \mathfrak{g}_\tau$, we deduce that the bilinear form in condition (ii) of Theorem 16.1 is degenerate, contrary to hypothesis. This gives the required contradiction and completes the proof of the lemma. □

We may upgrade lemma 16.7 as follows:

Lemma 16.8. *There exists $c \in F$ so that $\Delta(\tau) = c\Delta_0(\tau)$ for all $\tau \in \mathfrak{g}$.*

Proof. We consider the behavior of Δ and Δ_0 under homothety. For $t \in F^\times$, we have

$$D(t\tau)_{ij} = t^{i+j} D(\tau)_{ij}.$$

Since $1 + 2 + \cdots + n = n(n+1)/2$, we obtain

$$\Delta(t\tau) = t^{n(n+1)}\Delta(\tau). \tag{16.13}$$

On the other hand, if $\lambda_1, \dots, \lambda_{n+1}$ and $\mu_1, \dots, \mu_n$ are the eigenvalues of τ and τ_H, then

$$\Delta_0(t\tau) = \prod_{i=1}^{n+1}\prod_{j=1}^{n}(t\lambda_i - t\mu_j) = t^{n(n+1)}\Delta_0(\tau). \tag{16.14}$$

Thus Δ and Δ_0 are homogeneous elements of R of the same degree. We have seen in Lemma 16.7 that Δ_0 divides Δ. Since Δ_0 is not identically zero, the conclusion follows. $\square$

Remark 16.9. To complete the proofs of Theorems 15.3 and 16.1, it suffices at this point to show that Δ is not identically zero: lemma 16.8 then forces Δ to be a nonzero constant multiple of Δ_0.

The following calculation shows that the constant c in Lemma 16.8 equals $2^n(-1)^{n(n-1)/2}$, thereby completing the proof of Theorem 16.2.

Lemma 16.10. *There exists an element $\tau \in \mathfrak{g}$ for which*

$$\Delta(\tau) = 2^n(-1)^{n(n-1)/2} \tag{16.15}$$

and

$$\Delta_0(\tau) = 1. \tag{16.16}$$

Proof. We will exhibit τ explicitly. To do so, it is convenient to choose our coordinates carefully.

First, some notation. For $v \in V, m \in \mathrm{End}(V)$ and $u \in V^*$, we may canonically define

$$mv \in V, \quad um \in V^*, \quad umv \in F, \quad uv \in F, \quad vu \in \mathrm{End}(F)$$

via "matrix multiplication," regarding V (resp. V^*) as a space of column (resp. row) vectors.

Recall that the $(n+1)$-dimensional space V is the sum of the n-dimensional space V_H and the line Fe. Set $p := e \in V$, and let $q \in V^*$ denote the unique element with $q|_{V_H} = 0$ and $qp = 1$. Then $\mathfrak{h}$ is described as in (16.7), and $1_H = 1 - pq$.

We choose an isomorphism $V \cong F^{n+1}$ so that, with the respect to the corresponding basis $e_1, \dots, e_{n+1}$ of V and dual basis $e_1^*, \dots, e_{n+1}^*$ of V^*, we have

$$p = e_1, \quad q = e_1^* + e_{n+1}^*.$$

In these coordinates, we take for τ the standard lower-triangular nilpotent Jordan block, given for (say) $n = 3$ by

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

We note that both p and q are τ-cyclic, so the element $\tau \in \mathfrak{g}$ is stable by the stability characterization recorded in [NV, §14.3, Lem 3] (see also [RS, Thm 6.1]). We write $e_{ij} := e_{i,j} := e_i e_j^* \in \mathfrak{g}$ for the elementary matrices with respect to the chosen coordinates. The matrix pq is given by

$$pq = e_{1,1} + e_{1,n+1},$$

and so

$$1_H = 1 - pq = \sum_{i=2}^{n+1} e_{i,i} - e_{1,n+1}.$$

We deduce that

$$\tau_H = \sum_{i=2}^{n} e_{i+1,i} - e_{1,n} - e_{2,n+1}$$

and then, by induction on $\ell \geqslant 0$, that

$$\tau_H^\ell = \sum_{i,j: i \geqslant 2, i-j=\ell} e_{i,j} - \sum_{i,j: j-i=n-\ell} e_{i,j}. \tag{16.17}$$

For example, when $n = 3$, the elements τ_H^ℓ are given for $\ell = 0, 1, 2$ by the matrices

$$\begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{16.18}$$

The identity (16.16) is now simple to verify. Since τ is nilpotent, every eigenvalue λ_i of τ is zero. By computing the characteristic polynomial of τ_H (e.g., the second matrix in (16.18)), we see that the eigenvalues $\{\mu_1, \dots, \mu_n\}$ of τ_H on V_H are the solutions μ to $\mu^n = -1$. The product of $-\mu$ over such μ is 1, whence (16.16).

It remains to verify (16.15). Let $k, \ell \in \{1, \dots, n\}$. We use (16.10) to write

$$D(\tau)_{k,l} = q[\tau, [\tau^k, \tau_H^{\ell-1}]]p.$$

Expanding both commutators, we obtain four terms. Two terms vanish because $\tau_H^{l-1} p = 0$ and $q \tau_H^{l-1} = 0$. We obtain

$$D(\tau)_{k,l} = -q\tau\tau_H^{\ell-1}\tau^k p - q\tau^k\tau_H^{\ell-1}\tau p. \tag{16.19}$$

Using (16.17) and the definition of τ, we see that

$$q\tau\tau_H^{\ell-1} = 1_{l<n} e_{n-l+1}^* - 1_{l=n} e_{n+1}^*, \quad \tau^k p = e_{k+1}, \tag{16.20}$$

and

$$q\tau^k = e_{n-k+1}^*, \quad \tau_H^{\ell-1}\tau p = e_{\ell+1} - 1_{\ell=n} e_1. \tag{16.21}$$

Combining (16.19), (16.20) and (16.21), we arrive at

$$D(\tau)_{k,l} = \begin{cases} -2 & \text{if } k + \ell = n, \\ 2 & \text{if } (k, \ell) = (n, n), \\ 0 & \text{otherwise.} \end{cases}$$

For instance, in the case $n = 3$, we obtain

$$D(\tau) = \begin{pmatrix} 0 & -2 & 0 \\ -2 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Thus only one permutation contributes to the sum formula for the determinant of $D(\tau)$. The required identity (16.15) follows after computing the sign of that permutation, which we leave to the reader. $\square$

Index

References

[Ag] Keshav Aggarwal. A new subconvex bound for GL(3) L-functions in the t-aspect. *arXiv e-prints*, page arXiv:1903.09638, March 2019.

[An] A. N. Andrianov. Spherical functions for GL_n over local fields, and the summation of Hecke series. *Mat. Sb. (N.S.)*, 83 (125):429–451, 1970.

[BM] Ehud Moshe Baruch and Zhengyu Mao. Bessel identities in the Waldspurger correspondence over the real numbers. *Israel J. Math.*, 145:1–81, 2005.

[BR] Joseph Bernstein and Andre Reznikov. Subconvexity bounds for triple L-functions and representation theory. *Ann. of Math. (2)*, 172(3):1679–1718, 2010.

[BP1] R. Beuzart-Plessis. A local trace formula for the Gan-Gross-Prasad conjecture for unitary groups: the archimedean case. *ArXiv e-prints*, June 2015.

[BP2] R. Beuzart-Plessis. Comparison of local spherical characters and the Ichino-Ikeda conjecture for unitary groups. *ArXiv e-prints*, February 2016.

[BP3] Raphaël Beuzart-Plessis. Plancherel formula for $\mathrm{GL}_n(F)\backslash\mathrm{GL}_n(E)$ and applications to the Ichino-Ikeda and formal degree conjectures for unitary groups. *arXiv e-prints*, page arXiv:1812.00047, November 2018.

[BCZ] Raphaël Beuzart-Plessis, Pierre-Henri Chaudouard, and Michał Zydor. The global Gan-Gross-Prasad conjecture for unitary groups: the endoscopic case. *arXiv e-prints*, page arXiv:2007.05601, July 2020.

[BLZZ] Raphaël Beuzart-Plessis, Yifeng Liu, Wei Zhang, and Xinwen Zhu. Isolation of cuspidal spectrum, with application to the Gan–Gross–Prasad conjecture. *arXiv e-prints*, page arXiv:1912.07169, December 2019.

[BB1] V. Blomer and J. Buttcane. On the subconvexity problem for L-functions on GL(3). *to appear in Ann. Sci. ENS.*

[Blo] Valentin Blomer. Applications of the Kuznetsov formula on $GL(3)$. *Invent. Math.*, 194(3):673–729, 2013.

[BB2] Valentin Blomer and Jack Buttcane. Subconvexity for L-functions of non-spherical cusp forms on GL(3). *Acta Arith.*, 192(1):31–62, 2020.

[BKY] Valentin Blomer, Rizwanur Khan, and Matthew Young. Distribution of mass of holomorphic cusp forms. *Duke Math. J.*, 162(14):2609–2644, 2013.

[BM1] Valentin Blomer and Péter Maga. The sup-norm problem for PGL(4). *Int. Math. Res. Not. IMRN*, (14):5311–5332, 2015.

[BM2] Valentin Blomer and Péter Maga. Subconvexity for sup-norms of cusp forms on PGL(n). *Selecta Math. (N.S.)*, 22(3):1269–1287, 2016.

[Bu1] Jack Buttcane. On sums of $SL(3,\mathbb{Z})$ Kloosterman sums. *Ramanujan J.*, 32(3):371–419, 2013.

[Bu2] Jack Buttcane. The spectral Kuznetsov formula on $SL(3)$. *Trans. Amer. Math. Soc.*, 368(9):6683–6714, 2016.

[DFI] W. Duke, J. B. Friedlander, and H. Iwaniec. The subconvexity problem for Artin L-functions. *Invent. Math.*, 149(3):489–577, 2002.

[Ha] R. Neal Harris. The refined Gross-Prasad conjecture for unitary groups. *Int. Math. Res. Not. IMRN*, (2):303–389, 2014.

[HN] Roman Holowinsky and Paul D. Nelson. Subconvex bounds on gl(3) via degeneration to frequency zero. *Mathematische Annalen*, 01 2018.

[Hör] Lars Hörmander. *The analysis of linear partial differential operators. III.* Classics in Mathematics. Springer, Berlin, 2007. Pseudo-differential operators, Reprint of the 1994 edition.

[Hr] Karel Hrbacek. Nonstandard set theory. *Amer. Math. Monthly*, 86(8):659–677, 1979.

[II] Atsushi Ichino and Tamutsu Ikeda. On the periods of automorphic forms on special orthogonal groups and the Gross-Prasad conjecture. *Geom. Funct. Anal.*, 19(5):1378–1425, 2010.

[IS1] Henryk Iwaniec and Peter Sarnak. L^∞ norms of eigenfunctions of arithmetic surfaces. *Ann. of Math. (2)*, 141(2):301–320, 1995.

[IS2] Henryk Iwaniec and Peter Sarnak. Perspectives on the analytic theory of L-functions. *Geom. Funct. Anal.*, (Special Volume, Part II):705–741, 2000. GAFA 2000 (Tel Aviv, 1999).

[Ka] Toru Kawai. Nonstandard analysis by axiomatic method. In *Southeast Asian conference on logic (Singapore, 1981)*, volume 111 of *Stud. Logic Found. Math.*, pages 55–76. North-Holland, Amsterdam, 1983.

[Kh] Rizwanur Khan. On the subconvexity problem for $\mathrm{GL}(3)\times\mathrm{GL}(2)$ L-functions. *Forum Math.*, 27(2):897–913, 2015.

[Ki] A. A. Kirillov. Merits and demerits of the orbit method. *Bull. Amer. Math. Soc. (N.S.)*, 36(4):433–488, 1999.

[Kn] Anthony W. Knapp. *Representation theory of semisimple groups*, volume 36 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1986. An overview based on examples.

[Ko] Bertram Kostant. Lie group representations on polynomial rings. *Amer. J. Math.*, 85:327–404, 1963.

[KMS] Sumit Kumar, Kummari Mallesham, and Saurabh Kumar Singh. Sub-convexity bound for $GL(3) \times GL(2)$ L-functions: $GL(3)$-spectral aspect. *arXiv e-prints*, page arXiv:2006.07819, June 2020.

[Ku] John F. Kurtzke, Jr. Centralizers of irregular elements in reductive algebraic groups. *Pacific J. Math.*, 104(1):133–154, 1983.

[Li] Xiaoqing Li. Bounds for $\mathrm{GL}(3) \times \mathrm{GL}(2)$ L-functions and $\mathrm{GL}(3)$ L-functions. *Ann. of Math. (2)*, 173(1):301–336, 2011.

[LMS] Yongxiao Lin, Philippe Michel, and Will Sawin. Algebraic twists of $\mathrm{GL}_3\times\mathrm{GL}_2$ L-functions. *arXiv e-prints*, page arXiv:1912.09473, December 2019.

[Lu] Zhilin Luo. A Local Trace Formula for the Local Gan-Gross-Prasad Conjecture for Special Orthogonal Groups. *arXiv e-prints*, page arXiv:2009.13947, September 2020.

[Mi] Philippe Michel. Analytic number theory and families of automorphic L-functions. In *Automorphic forms and applications*, volume 12 of *IAS/Park City Math. Ser.*, pages 181–295. Amer. Math. Soc., Providence, RI, 2007.

[MV1] Philippe Michel and Akshay Venkatesh. Equidistribution, L-functions and ergodic theory: on some problems of Yu. Linnik. In *International Congress of Mathematicians. Vol. II*, pages 421–457. Eur. Math. Soc., Zürich, 2006.

[MV2] Philippe Michel and Akshay Venkatesh. The subconvexity problem for GL_2. *Publ. Math. Inst. Hautes Études Sci.*, (111):171–271, 2010.

[Mu1] Ritabrata Munshi. The circle method and bounds for L-functions—III: t-aspect subconvexity for $GL(3)$ L-functions. *J. Amer. Math. Soc.*, 28(4):913–938, 2015.

[Mu2] Ritabrata Munshi. The circle method and bounds for L-functions—IV: subconvexity for twists of GL(3) L-functions. *Ann. of Math. (2)*, 182(2):617–672, 2015.

[Mu3] Ritabrata Munshi. Subconvexity for $GL(3)\times GL(2)$ L-functions in t-aspect. *arXiv e-prints*, page arXiv:1810.00539, October 2018.

[Ne] Edward Nelson. Internal set theory: a new approach to nonstandard analysis. *Bull. Amer. Math. Soc.*, 83(6):1165–1198, 1977.

[NV] Paul D. Nelson and A. Venkatesh. The orbit method and analysis of automorphic forms. *Acta Math.*, to appear.

[RS] S. Rallis and G. Schiffmann. Multiplicity one Conjectures. *ArXiv e-prints*, May 2007.

[Ri] Marc A. Rieffel. Lie group convolution algebras as deformation quantizations of linear Poisson structures. *Amer. J. Math.*, 112(4):657–685, 1990.

[Ro1] W. Rossmann. Limit characters of reductive Lie groups. *Invent. Math.*, 61(1):53–66, 1980.

[Ro2] W. Rossmann. Tempered representations and orbits. *Duke Math. J.*, 49(1):231–247, 1982.

[Ro3] Wulf Rossmann. Kirillov's character formula for reductive Lie groups. *Invent. Math.*, 48(3):207–220, 1978.

[SV] Yiannis Sakellaridis and Akshay Venkatesh. Periods and harmonic analysis on spherical varieties. *Astérisque*, (396):viii+360, 2017.

[Sa] Peter Sarnak. Fourth moments of Grössencharakteren zeta functions. *Comm. Pure Appl. Math.*, 38(2):167–178, 1985.

[Sh] Prahlad Sharma. Subconvexity for $GL(3) \times GL(2)$ L-functions in $GL(3)$ spectral aspect. *arXiv e-prints*, page arXiv:2010.10153, October 2020.

[SV] Lior Silberman and Akshay Venkatesh. Quantum unique ergodicity for locally symmetric spaces II.

[SZ] Binyong Sun and Chen-Bo Zhu. Multiplicity one theorems: the Archimedean case. *Ann. of Math. (2)*, 175(1):23–44, 2012.

[Ve] Akshay Venkatesh. Sparse equidistribution problems, period bounds and subconvexity. *Ann. of Math. (2)*, 172(2):989–1094, 2010.

[Wa] Jean-Loup Waldspurger. Une formule intégrale reliée à la conjecture locale de Gross-Prasad, 2e partie: extension aux représentations tempérées. *Astérisque*, (346):171–312, 2012. Sur les conjectures de Gross et Prasad. I.

[Zh1] Wei Zhang. Fourier transform and the global Gan-Gross-Prasad conjecture for unitary groups. *Ann. of Math. (2)*, 180(3):971–1049, 2014.

[Zh2] Wei Zhang. Automorphic period and the central value of Rankin-Selberg L-function. *J. Amer. Math. Soc.*, 27(2):541–612, 2014.

ETH Zürich, Department of Mathematics, Rämistrasse 101, CH-8092, Zürich, Switzerland

Email address: `nelson.paul.david@gmail.com`