QUANTUM VARIANCE ON QUATERNION ALGEBRAS, III

PAUL D. NELSON

Abstract. We determine the asymptotic quantum variance of microlocal lifts of Hecke–Maass cusp forms on the arithmetic compact hyperbolic surfaces attached to maximal orders in quaternion algebras. Our result extends those of Luo–Sarnak–Zhao concerning the non-compact modular surface. The results of this article’s prequel (which involved the theta correspondence, Rallis inner product formula and equidistribution of translates of elementary theta functions) reduce the present task to some local problems over the reals involving the construction and analysis of microlocal lifts via integral operators on the group. We address these here using an analytic incarnation of the method of coadjoint orbits.

Contents
1. Introduction 1
2. Reduction of the proof 5
3. Preliminaries 7
4. Characterizing microlocal lifts via their symmetry 14
5. Constructing microlocal lifts via integral operators 16
6. Quantum variance sums via integral operators 17
7. Main term estimates 18
8. Error estimates 21
9. Removing the arithmetic weights 23
10. Heuristics 24
Appendix A. Calculations with raising and lowering operators 28
References 31

1. Introduction

1.1. Overview. Let \( Y \) be an arithmetic hyperbolic surface attached to a maximal order in a quaternion algebra, and let \( X \) denote the unit cotangent bundle of \( Y \). The purpose of this article is to determine the asymptotic quantum variance of microlocal lifts to \( X \) of Hecke–Maass eigenforms on \( Y \) when these spaces are compact. The corresponding conclusion in the non-compact case is a theorem of Sarnak–Zhao [19], building on earlier work of Luo–Sarnak [11] and Zhao [31]. The method employed in those works relies upon parabolic Fourier expansions, which are unavailable in the compact case. We appeal here to a different method, based
on the theta correspondence, that was introduced in the prequels \cite{15, 16} to this article and applied there to simpler $p$-adic variants of the motivating problem treated here.

The global inputs to our argument were developed in the prequel. Those inputs will be applied here as a “black box,” so that the two articles can be read independently. The main purpose of this article is to supply the remaining local inputs at an archimedean place.

1.2. Setup and notation. We recall the parametrization of the spaces $X$ and $Y$ (see for instance \cite{20, §9} or \cite{22, §4} or \cite{23, §38} for further background). Let $M := M_2(\mathbb{R})$ denote the $2 \times 2$ matrix algebra and $G := \text{PGL}_2(\mathbb{R}) = M^\times / \mathbb{R}^\times$ its projectivized unit group. Let $F$ be a totally real number field. (Our results are new when $F = \mathbb{Q}$, but the generality does not introduce complication.) Let $B$ be a quaternion algebra over $F$ with the property that there is exactly one real place $q$ of $F$ such that $B_q$ is isomorphic to $M$; we fix such an isomorphism, together with a maximal order $R \subseteq B$, and denote by $\Gamma \leq G$ the image of $R^\times$ under the induced isomorphism $B_q^\times / \mathbb{R}_q^\times \cong G$. Then $\Gamma \leq G$ is a discrete cofinite subgroup; it is cocompact except when $B$ is split, in which case $F = \mathbb{Q}$ and $B \cong M_2(\mathbb{Q})$. We denote by $K \leq G$ the image of $O(2)$ and by $G^1, K^1, \Gamma^1$, the subgroups consisting of positive determinant elements.

We set $X := \Gamma \backslash G$ and $Y := X / K^1$. We equip $G$ with any Haar measure $dg$ and $X$ with any $G$-invariant measure, and write $\langle \varphi_1, \varphi_2 \rangle := \int_X \varphi_1 \overline{\varphi_2} \, dx$ for $\varphi_1, \varphi_2 \in L^2(X)$. The group $G$ acts unitarily on $L^2(X)$ by right translation: $g \varphi(x) := \varphi(xg)$.

We assume for simplicity of presentation that $F$ has odd narrow class number. The strong approximation theorem (see for instance \cite{16, §4.4.2}, \cite{23, §28}) then identifies $X$, rather than a finite disjoint union of similarly defined quotients, with the adelic quotient $G(F) \backslash G(\mathbb{A}) / J$, where

- $\mathbb{A}$ denotes the adele ring of $F$,
- $G$ denotes the $F$-algebraic group $R \mapsto (B \otimes_F R)^\times / R^\times$, and
- $J = J_\infty \prod_{p < \infty} J_p$, with $J_p$ the image of $R_p^\times$ and $J_\infty$ the points of $G$ over the product of the archimedean completions of $F$ other than $F_q$. If $p \notin S$, then $J_p$ is a maximal compact subgroup of $G(F_p)$; otherwise, $J_p$ has index 2 in the compact group $G(F_p)$. We have $J_\infty \cong \text{SO}(3)^{|F^\times| - 1}$.

As a consequence of this identification, we obtain for each finite prime $p$ of $F$ a Hecke operator $T_p$ acting on $L^2(X)$ (see for instance \cite{16, §4.4.3}); these operators commute with one another and also with $G$. Strong approximation also implies that the group $\Gamma$ contains elements of negative determinant. We may thus identify $X$ with $\Gamma^1 \backslash G^1$ and $Y$ with $\Gamma^1 \backslash \mathbb{H}$, where $\mathbb{H} \cong G^1 / K^1$ is the hyperbolic plane. This identification may be useful for interpreting our results, but is not used directly in the proofs.

By an eigenfunction $\Psi : X \to \mathbb{C}$, we mean a smooth $K$-finite function that generates an irreducible representation of $G$ and is a $T_p$-eigenfunction for each $p$.

We assume that $B$ is non-split, so that $X$ is compact. We denote by $L^2_0(X) \subseteq L^2(X)$ the subspace of mean zero functions and by $A_q$ the set of subspaces $\pi \subseteq L^2_q(X)$ that are irreducible under $G$ and that are eigenspaces for each $T_p$. The smooth $K$-finite vectors in $\pi$ are eigenfunctions in the above sense, and each nonzero eigenfunction generates one such $\pi$. The multiplicity one theorem implies
that the space $L^2_0(X)$ is the Hilbert direct sum of the $\pi \in A_0$. Under strong approximation in the sense noted above, $A_0$ identifies with the set of generic automorphic representations of $G$ containing a nonzero $J$-invariant vector. Each $\pi \in A_0$ has an infinitesimal character $\lambda_\pi \in \mathbb{R}$, describing the action of the center of the universal enveloping algebra of $G$ (see §3.3). If $\lambda_\pi < 0$, then $\pi$ contains a one-dimensional space of $K^1$-invariant vectors $\varphi_\pi$; these descend to functions on $Y$ of Laplace eigenvalue $1/4 - \lambda_\pi$, giving a bijection

$$\{\pi \in A_0 : \lambda_\pi < 0\} \leftrightarrow \{\text{Hecke–Maass eigenforms } \varphi_\pi \text{ on } Y \text{ of eigenvalue } 1/4\}.$$ 

We normalize $\varphi_\pi$ so that $\langle \varphi_\pi, \varphi_\pi \rangle = 1$. We denote by $\mu_\pi$ the representation-theoretic microlocal lift of $\varphi_\pi$ constructed by Zelditch (see [29, 28, 30, 26, 25, 9, 21]) and as studied in the related work of Sarnak–Zhao [19]. We recall the precise construction of $\mu_\pi$ in §3.5; we mention for now only that

$$\mu_\pi : \{K\text{-finite smooth } \Psi : X \to \mathbb{C}\} \to \mathbb{C}$$ (1.1)

is a functional with the following properties:

- If $\Psi$ is $K^1$-invariant, so that it comes from a function on $Y$, then $\mu_\pi(\Psi) = \langle \varphi_\pi, \Psi \rangle$.
- $\mu_\pi$ is asymptotically invariant by the diagonal subgroup $H$ of $G$ in the sense that for each fixed $h \in H$ and $\Psi$ as above, the difference $\mu_\pi(h\Psi) - \mu_\pi(\Psi)$ tends to zero as $|\lambda_\pi| \to \infty$.

A theorem of Lindenstrauss [10], resolving a case of the arithmetic quantum unique ergodicity conjecture of Rudnick–Sarnak [18], implies that $\mu_\pi(\Psi) \to \langle \Psi, 1 \rangle / \text{vol}(X)$ for each fixed continuous function $\Psi : X \to \mathbb{C}$ as $|\lambda_\pi| \to 0$. Equivalently, $\mu_\pi(\Psi) \to 0$ when $\Psi$ has mean zero.

Let $S$ denote the set of finite primes of $F$ at which $B$ ramifies. For $p \in S$, the operator $T_p$ is an involution, so we may speak of the parity of an eigenfunction with respect to $T_p$. The local root number of $\sigma \in A_0$ at the distinguished real place $q$ is an element of $\{\pm 1\}$; it is $+1$ precisely when $\pi$ admits a nonzero functional invariant by the normalizer in $G$ of $H$. We say that $\sigma \in A_0$ is even if

- for each $p \in S$, the $T_p$-eigenvalue of $\sigma$ is $+1$, and
- the local root number of $\sigma$ at $q$ is $+1$.

If $\sigma$ is not even, then $\mu_\pi(\Psi) = 0$ for all $\Psi \in \sigma$ (see §3.6); for this reason we focus primarily on even $\sigma$. We say also that an eigenfunction $\Psi$ is even if it belongs to an even $\sigma \in A_0$. (We note that this terminology is not directly related to the customary distinction between, for instance, “even and odd Maass forms on $\text{SL}_2(\mathbb{Z})\backslash \mathbb{H}$.”)

For a function $\Psi : X \to \mathbb{C}$, we write $\Psi^w := \frac{1}{2} (\Psi + w \Psi)$ for its symmetrization with respect to the Weyl element $w := \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \in G$.

We equip the diagonal subgroup $H \leq G$ with the Haar measure given by

$$\int_H f := \int_{y \in \mathbb{R}^k} f \left( \begin{array}{c} y \\ 1 \end{array} \right) \frac{dy}{|y|},$$

where $dy$ denotes Lebesgue measure.

We denote in what follows by $L^{(S)}(\cdots, s)$ the finite part of an $L$-function, omitting Euler factors in $S$. For $\pi \in A_0$, we abbreviate $\iota_\pi := L^{(S)}(\text{ad } \pi, 1)$. 


1.3. Main result. We henceforth fix a pair of nonzero mean zero even eigenfunctions $\Psi_1, \Psi_2$. We denote by $\sigma_1, \sigma_2 \in A_0$ the representations that they generate.

**Theorem 1.** The limit
\[
\lim_{h \to 0} \frac{1}{h} \sum_{\pi \in A_0: 0 < h^2 \lambda_\pi < 1} \tau_{\pi} \mu_{\pi}(\Psi_1) \mu_{\pi}(\Psi_2)
\]
exists. If $\sigma_1 \neq \sigma_2$, then that limit is zero. If $\sigma_1 = \sigma_2 =: \sigma$, then it is given by
\[
\frac{c_B}{2\pi} L_S^{(\sigma, \frac{1}{2})} \int_{s \in H} (s\Psi_1^w, \Psi_2^w),
\]
where $c_B := 2\#S F(2) / \text{vol}(X)$.

As discussed (among other places) in [19] or [16, §2.1.6], the integral on the RHS of (1.3) converges absolutely.

**Remark 1.** Each of the expressions (1.2) and (1.3) is independent of the choice of Haar measure on $X$. If we equip $X$ with the pullback of the standard hyperbolic measure $\frac{dx dy}{y^2}$ on $Y$, then we may verify as in [24, §1] or [16, §4.4.2] that
\[
\frac{c_B}{2\pi} = 2\#S \prod_{p \in S} \left(1 + 1/|p|\right)
\]
where $\Delta_F$ and $\Delta_B$ denote the absolute discriminant and absolute reduced discriminant, respectively, and $|p|$ denotes the absolute norm of the finite prime $p$. For instance, if $F = \mathbb{Q}$, then
\[
\frac{c_B}{2\pi} = 2\#S \prod_{p \in S} \left(1 + 1/p\right)
\]
The factor $2\#S$ may be understood (see §10) as coming from the nontrivial normalizer of $\Gamma$, corresponding to the involutory Hecke operators $T_p$ ($p \in S$). If one instead sums over only those $\pi$ having eigenvalue $+1$ under such operators, then this factor disappears.

**Remark 2.** For the sake of comparison with [19], we note that
\[
\int_{s \in H} (s\Psi_1^w, \Psi_2^w) = 2 \int_{u \in \mathbb{R}} \langle e^{u/2}, e^{-u/2} \rangle \Psi_1^\text{sym}, \Psi_2^\text{sym} \rangle du,
\]
where $\Psi^\text{sym}$ denotes the average of $\Psi$ over its translates by the four-element subgroup of $G$ generated by $\text{diag}(-1, 1)$ and $w$.

**Remark 3.** The “arithmetic weights” $\tau_{\pi}$ arise in our method for reasons illustrated best by [15, §2.8, §7]. They have mild size ($O(h^{-\varepsilon})$ for any fixed $\varepsilon > 0$) and mean 1. Sarnak–Zhao [19] showed in the non-compact case that if one modifies the sums (1.2) by omitting the weights $\tau_{\pi}$, then the conclusion remains valid after multiplying the main term (1.3) by a certain explicit factor $c_\sigma > 0$. To do this, they used zero density estimates for families of $L$-functions to approximate $\tau_{\pi}^{-1}$ for most $\pi$ by a short Dirichlet polynomial and then appealed to estimates for Hecke-twisted variants of (1.2). Their method applies in our setting with the (analogous) constant
\[
c_\sigma := \frac{1}{\zeta_F^{(S)}(2)} \prod_{p \notin S} \left(1 - \frac{\lambda_\sigma(p)}{|p|^{3/2} + |p|^{1/2}}\right),
\]
where \( \lambda_\sigma(p) \) denotes the Hecke eigenvalue normalized so that the Ramanujan conjecture says \( |\lambda_\sigma(p)| \leq 2 \). We do not replicate here the details of their argument, but explain in \S 9 how the factors \( c_\sigma \) arise naturally from the perspective of our method.

\textbf{Remark 4.} In \S 10, we extend the semiclassical heuristics for variance asymptotics from the generic non-arithmetic setting (see e.g. [27, \S 15.6], [17, \S 4.1.3]) to the setting of Theorem 1. The resulting predictions are consistent with our results.

\section{2. Reduction of the proof}

We first recall the main result of the prequel, which will be seen below to reduce the proof of Theorem 1 to several local problems. This requires some notation. We identify finite-rank operators \( T \) on \( \pi \) with finite-rank tensors \( T = \sum_i v_i \otimes v'_i \in \pi \otimes \pi \).

Given any such \( T \) and any bounded measurable \( \Psi : X \to \mathbb{C} \), we set

\[ \mu(T, \Psi) := \sum_i \langle v_i, \Psi v'_i \rangle. \]

We verify readily (as in [13, \S 26.3]) that \( |\mu(T, \Psi)| \leq \| T \|_1 \| \Psi \|_{L^{\infty}}, \) where \( \| \cdot \|_1 \) denotes the trace norm. We may thus extend the assignment \( T \mapsto \mu(T, \Psi) \) continuously to any trace class operator \( T \) on \( \pi \), and in particular, to the integral operators

\[ \pi(f) := \int_{g \in G} f(g) \pi(g) \, dg \]

attached to \( f \in C_c^\infty(G) \) and our choice of Haar measure \( dg \) on \( G \). Equivalently, we may express \( \mu(T, \Psi) \) as the absolutely convergent sum

\[ \mu(T, \Psi) = \sum_{v \in B(\pi)} \langle T v \cdot \Psi, v \rangle, \]

where \( B(\pi) \) is an orthonormal basis for \( \pi \) consisting of \( K \)-isotypic vectors.

Recall that we have fixed some nonzero mean zero even eigenfunctions \( \Psi_1 \in \sigma_1, \Psi_2 \in \sigma_2 \). We define hermitian forms \( \mathcal{V} \) and \( \mathcal{M} \) on \( C_c^\infty(G) \) as follows:

- \( \mathcal{V}(f) := \sum_{\tau \in \Lambda^0} t_\tau \mu(\pi(\tau), \Psi_1) \mu(\pi(\tau), \Psi_2). \) (By partial integration, such sums converge rapidly, as explained in [16, \S 4].)
- For a function \( f : G \to \mathbb{C} \), set

\[ \mathcal{S} f(g) := \frac{f(g) + f(g^{-1})}{2} \]

and, for \( g \in G \),

\[ g \cdot f(x) := f(g^{-1}xg) \]

- If \( \sigma_1 \neq \sigma_2 \), then \( \mathcal{M}(f) := 0. \) If \( \sigma_1 = \sigma_2 =: \sigma \), then

\[ \mathcal{M}(f) := c_B L^{(S)}(\sigma, \frac{1}{2}) I(f), \]

with \( c_B \) as in (1.3) and

\[ I(f) := \int_{g \in G} \langle g \cdot \mathcal{S} f, \mathcal{S} f \rangle_G(g\Psi_1, \Psi_2). \]

Here \( \langle \cdot, \cdot \rangle_G \) denote the inner product in \( L^2(G) \).

For any real vector space \( V \), we denote by \( \mathcal{S}(V) \) the space of Schwartz functions. Recall that \( M = M_2(\mathbb{R}) \) denotes the \( 2 \times 2 \) matrix algebra. For each \( \tau \in F^\times \) and \( f \in C_c^\infty(G) \), we define \( \nabla^\tau f \in \mathcal{S}(M) \) by the formula

\[ \nabla^\tau f(x) := 1_M(x) \begin{pmatrix} W(\tau \det(x)) \\ \frac{W(\tau \det(x))}{|\tau \det(x)|} \det(f(x)) \end{pmatrix}, \]
where \( pr : M^\times \to G \) denotes the natural projection and \( W \in C_c^\infty (\mathbb{R}^\times) \) is a nonzero test function that we fix once and for all.

The motivation for introducing the sums \( \mathcal{V} (f) \) is that for suitable \( f \), they will be seen to approximate the basic variance sums of interest. The "expected main terms" \( \mathcal{M} (f) \) arose in the prequel after some calculations involving theta functions and the Rallis inner product formula; we refer to §10 for some heuristic discussion about why one should expect \( \mathcal{V} (f) \approx \mathcal{M} (f) \) for nice enough \( f \). The operators \( \nabla^r \) should be understood as associating to a function on the multiplicative group \( G \) its "thickening" on the additive group \( M \). They are at the heart of the method developed in the prequel, where it was shown that \( \nabla^r f \) is the kernel of a theta function with \( L^2 \)-norm proportional to \( \mathcal{V} (f) \).

Let \( M^0 \ll M \) denote the trace zero subspace. We identify \( \mathbb{R} \) with the subspace of scalar matrices in \( M \). We then have an orthogonal decomposition \( M = \mathbb{R} \oplus M^0 \).

For \( y \in \mathbb{R}^\times \), we denote by \( D_y \) the operator on \( \mathcal{S} (M) \) given by normalized scaling of the \( M_0 \) component: for \( \Phi \in \mathcal{S} (M), t \in \mathbb{R}, u \in M^0, \)
\[
D_y \Phi (t + u) := |y|^{3/2} \Phi (t + yu).
\]
It extends to a unitary operator on \( L^2 (M) \).

We are now prepared to state a specialization of the main result of [16, Part 1].

**Theorem 2.** There is a finite subset \( X \) of \( F^\times \) and a collection \( (\mathcal{E}_{\tau_1, \tau_2})_{\tau_1, \tau_2 \in X} \) of sesquilinear forms on \( \mathcal{S} (M) \) so that for each \( f \in C_c^\infty (G), \)
\[
\mathcal{V} (f) = \mathcal{M} (f) + \sum_{\tau_1, \tau_2 \in X} \mathcal{E}_{\tau_1, \tau_2} (\nabla^{\tau_1} f, \nabla^{\tau_2} f). \tag{2.1}
\]

Moreover, there is a continuous seminorm \( C \) on the Schwartz space \( \mathcal{S} (M) \) so that for all \( y \in \mathbb{R}^\times \) and \( \phi_1, \phi_2 \in \mathcal{S} (M), \)
\[
|\mathcal{E}_{\tau_1, \tau_2} (D_y \phi_1, D_y \phi_2)| \leq \frac{\log (|y| + |y|^{-1})}{|y| + |y|^{-1}} C (\phi_1) C (\phi_2). \tag{2.2}
\]

**Proof.** This follows from [16, Thm 2], as specialized in [16, §4.4]. In the notation of that reference, we have \( \mathcal{V} (f) = V_f (\Phi_1, \Phi_2) \) and \( \mathcal{M} (f) = M_f (\Phi_1, \Phi_2) \). We take for \( \mathcal{E}_{\tau_1, \tau_2} \) the functional "\( \mathcal{E} \)" constructed in the final paragraph of [16, §9.3.4]. The estimate (2.2) is obtained by applying the "main estimate" of [16, Thm 2] with \( s \) any element of the metaplectic double cover of \( SL_2 (\mathbb{R}) \) lifting diag \((y, y^{-1})\), see also [16, §2.1.5].

Before explaining how we plan to apply this result, we set some asymptotic notation and terminology. We consider henceforth a sequence \( \{ h \} \) of positive reals \( h \) tending to zero, as in the statement of Theorem 1. By an "\( h \)-dependent element" of a set \( U \), we mean a map \( \{ h \} \to U \), which we understand colloquially as an element \( u \in U \) that depends (perhaps implicitly) upon the parameter \( h \). The word "fixed" will be taken to mean "independent of \( h \)." Our default convention is that quantities not labeled "fixed" may depend upon \( h \), but we will usually mention this dependence for the sake of clarity. Standard asymptotic notation is defined accordingly: \( A = O (B) \), \( A \ll B \) and \( B \gg A \) mean that \( |A| \leq c |B| \) for some fixed \( c \geq 0 \), while \( A \asymp B \) means that \( A \ll B \ll A \); the meaning of an infinite exponent as in \( A = O (h^\infty) \) is that the indicated estimate holds upon substituting for \( \infty \) any fixed positive quantity. The fixed quantities \( c \) may of course depend upon any previously mentioned fixed quantities. We always assume that \( h \) is small enough
with respect to any mentioned fixed quantities. (For instance, we may speak of an h-dependent element $\pi \in A_0$ satisfying $1/2 < -\hbar^2 \lambda_\pi < 1$; its microlocal lift $\mu_\pi$ is an h-dependent distribution on $X$ that satisfies $|\mu_\pi(\Psi)| = O(1)$ for fixed $\Psi$ as in (1.1); for fixed $\Psi \in C^\infty(X)$, we have $(\varphi_\pi, \Psi) = O(h^\infty)$.)

We now reduce the proof of Theorem 1 to some local problems. By an approximation argument, it suffices to show that there is a fixed $\delta > 0$ so that for each fixed nonnegative $k \in C^\infty_c(\mathbb{R}_{>0})$,

$$h \sum_{\pi \in A_0} \tau_\pi k(h^2 \lambda_\pi)^2 \mu_\pi(\Psi_1) \overline{\mu_\pi(\Psi_2)} = c_k L(S)(\alpha, \frac{1}{2}) \int_{s \in H} \langle s\Psi^w_1, \Psi^w_2 \rangle + O(h^\delta), \quad (2.3)$$

where $c_k := c_B \int_{t \in \mathbb{R}_{>0}} k(-t^2)^2 \frac{dt}{2\pi}$. Indeed, it is enough to show this for a class $\mathcal{K}$ of h-dependent nonnegative functions $k \in C^\infty_c(\mathbb{R}_{>0})$ with the following properties:

- (\(\mathcal{K}\) is “controlled”) Each $k \in \mathcal{K}$ is supported on a fixed compact subset of $\mathbb{R}_{>0}$ and bounded from above by a fixed quantity.
- (\(\mathcal{K}\) is “sufficiently rich”) For each fixed nonnegative $k_0 \in C_c(\mathbb{R}_{>0})$ we may find $k, k_+ \in \mathcal{K}$ so that $|k - k_0| \leq k_+$ and $\int k_+ \to 0$ as $h \to 0$.

We construct such a class $\mathcal{K}$ explicitly in §5. In §6 and beyond, we construct for each $k \in \mathcal{K}$ an h-dependent element $f \in C^\infty_c(G)$ and show that

$$\mathcal{V}(f) = h \sum_{\pi \in A_0} \tau_\pi k(h^2 \lambda_\pi)^2 \mu_\pi(\Psi_1) \overline{\mu_\pi(\Psi_2)} + O(h^\delta) \quad (2.4)$$

and

$$\mathcal{I}(f) = \int_{s \in H} \langle s\Psi^w_1, \Psi^w_2 \rangle \int_{t > 0} k(-t^2)^2 \frac{dt}{2\pi} + O(h^\delta) \quad (2.5)$$

and

$$\mathcal{E}_{\tau_1, \tau_2}(\nabla^{\tau_1} f, \nabla^{\tau_2} f) \ll h^{1-\delta'} \quad (2.6)$$

for fixed $\tau_1, \tau_2 \in F^\times$, where $\delta' = \delta'(\delta) > 0$ is a fixed quantity with $\delta' \to 0$ as $\delta \to 0$. The required estimate (2.3) then follows from the identity (2.1).

We briefly indicate the proofs of the above estimates. The idea behind the construction of $f$ (completed in §6) is to arrange that $\pi(f)$ is an approximate weighted projector onto a “space” spanned by unit vectors $v \in \pi$ for which $\langle v|\Psi, v \rangle \approx \mu_\pi(\Psi)$; this leads to (2.4). The orbit method and philosophy advocated in [13] and summarized in §3.7 is a suitable tool for constructing and studying such approximate projectors. (The analogous p-adic problem treated in [16] is simpler precisely because of the availability of exact projectors in the p-adic Hecke algebra.) For the proof (§7) of the main term estimate (2.5), we pull the inner product $\langle \cdot, \cdot \rangle_G$ back to the Lie algebra, apply Parseval, and disintegrate the resulting integral along the coadjoint orbits; the subgroup $H$ then arises naturally as the stabilizer of the “limiting microlocal support” of the $\mu_\pi$. The error estimate (2.6), proved in §8, is ultimately a consequence of (2.2) and the fact that the function $f$ that we construct concentrates just above the scale $1 + O(h) \subseteq G$ and barely oscillates below that scale.

3. Preliminaries

3.1. Lie algebra. Let $\mathfrak{g}$ denote the Lie algebra of $G$. We denote by $\mathfrak{g}_\mathbb{C}$ its complexification and by $\mathfrak{g}^*_\mathbb{C}$ the complex dual. We will often identify $\mathfrak{g}_\mathbb{C}$ with the space
of linear functions on \( g^*_C \). We work with the following basis elements for \( g_C \):

\[
X := \frac{1}{2i} \begin{pmatrix} 1 & i \\ i & -1 \end{pmatrix}, \quad Y := \frac{1}{2i} \begin{pmatrix} 1 & -i \\ -i & -1 \end{pmatrix}, \quad W := \frac{1}{2i} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.
\]

These satisfy \([X, Y] = -2W, [W, X] = X\) and \([W, Y] = -Y\). The map \( \theta \mapsto e^{i\theta W} \) defines an isomorphism from \( \mathbb{R}/2\pi\mathbb{Z} \) to \( K^1 \). The complex conjugation on \( g_C \) is given by \( -\overline{X} = Y \) and \( -\overline{W} = W \).

The center of the universal enveloping algebra of \( g_C \) is the one variable polynomial ring \( \mathbb{C}[\Omega] \), where

\[
\Omega := W^2 - \frac{XY + YX}{2} = W(W - 1) - XY = W(W + 1) - YX.
\]

The ring \( \text{Sym}(g_C)^G \) of \( G \)-invariant polynomials on \( g^*_C \) is generated by the polynomial \( \Lambda := W^2 - XY \). The Harish–Chandra isomorphism \( \mathbb{C}[\Omega] \xrightarrow{\cong} \mathbb{C}[\Lambda] \) is given in this case by \( \Omega \mapsto \Lambda - 1/4 \).

We identify \( g^*_C \) with \( g_C \) via the trace pairing \( (x, \xi) \mapsto \text{trace}(x\xi) \). We identify the real and imaginary duals \( g^r \) and \( i g^i \) of \( g \) with the subspaces of \( g^*_C \) taking real and imaginary values on \( g \), respectively. We abbreviate \( g^i := ig^r \); it identifies with the Pontryagin dual of \( g \) via the natural pairing \( g \times g^\vee \ni (x, \xi) \mapsto e^{i(x, \xi)} \in \mathbb{C}^1 \).

We occasionally work with the coordinates and basis elements

\[
g \ni x := \begin{pmatrix} x_1/2 \\ x_2 \\ -x_1/2 \end{pmatrix} = \sum_{j=1,2,3} x_j e_j,
\]

\[
g^\vee \ni \xi = i \begin{pmatrix} \xi_1 \\ \xi_2 \\ -\xi_1 \end{pmatrix} = \sum_{j=1,2,3} \xi_j e_j^*,
\]

so that the natural pairing is given by \( (x, \xi) \mapsto e^{i\sum x_j \xi_j} \). We note that the invariant polynomial \( \Lambda \) is given in this optic by \( \Lambda(\xi) = \det(\xi/i) = -\xi_1^2 - \xi_2 \xi_3 \).

The following elements, defined for \( t \in \mathbb{R} \neq 0 \), will occur frequently in our analysis:

\[
\xi(t) := i \begin{pmatrix} t \\ -\bar{t} \end{pmatrix} \in g^\vee.
\]

We note that \( X(\xi(t)) = Y(\xi(t)) = t \), while \( W(\xi(t)) = 0 \), hence \( \Lambda(\xi(t)) = -t^2 \). We note also that the \( G \)-stabilizer of \( \xi(t) \) is the diagonal subgroup \( H \).

We fix norms \( |.| \) on all of the above spaces.

### 3.2. Coadjoint orbits.

A coadjoint orbit \( O \) is a \( G \)-orbit on \( g^\vee \); in particular, it is a smooth manifold. The origin \( \{0\} \) is a zero-dimensional coadjoint orbit. The other coadjoint orbits are two-dimensional and of the form

\[
O(\lambda) := \{ \xi \in g^\vee - \{0\} \text{ with } \Lambda(\xi) = \lambda \}
\]

for some \( \lambda \in \mathbb{R} \). If \( \lambda = 0 \), then \( O(\lambda) \) is the regular subset of the nilcone; if \( \lambda > 0 \), it is a two-sheeted hyperboloid; if \( \lambda < 0 \), it is a one-sheeted hyperboloid. The orbit of \( \xi(t) \) is \( O(-t^2) \).

We equip any two-dimensional coadjoint orbit \( O \) with its normalized symplectic measure \( C_e(g^\vee) \ni a \mapsto \int_Q a \), corresponding to the 2-form \( \sigma \) on \( O \) described as follows (see for instance [13, §6] or [7] for further details, and the calculations of §7 for some explicit formulas). For each \( \xi \in O \), the tangent space \( T_\xi O \) identifies with the space of vectors \( \{ x \cdot \xi : x \in g \} \), where \( x \cdot \xi \in g^\vee \) is defined by differentiating the action of \( G \) on \( g^\vee \). The component \( \sigma_\xi \) of \( \sigma \) at \( \xi \) is then given by \( \sigma_\xi(x \cdot \xi, y \cdot \xi) := \)
Lemma. Let estimate: the invariant polynomial $\Lambda$ is nonzero at $\omega$ such of the open neighborhood a coadjoint orbit, and we have $\int_{x \in hO} a(x) = h \int_{x \in O} a(hx)$. We record a simple estimate:

Proof. The main point is that each $\omega \in g^\wedge$ with $|\omega| > 1$, and let $0 < r \leq 1$. For any two-dimensional coadjoint orbit $O$, the symplectic volume of the subset $\{\xi \in O : |\xi - \omega| < r\}$ of $O$ is $O(r^2)$.

3.3. Representations. Let $\pi$ be an irreducible unitary representation of $G$. Then $\Omega$ acts on the smooth subspace of $\pi$ by some real scalar $\Omega_\pi$, and refer to it as the infinitesimal character of $\pi$. Up to isomorphism, we may classify $\pi$ as follows:

- The one-dimensional representations (the trivial representation $C$ and the sign representation $\mathbb{C}(\text{sgn} \circ \text{let})$, for which $\Omega_\pi = 0$ and $\lambda_\pi = 1/4$).
- The discrete series representations $\pi(k) (k \in \mathbb{Z}_{\geq 1})$, for which $\Omega_\pi = k(k-1)$ and $\lambda_\pi = (k - 1/2)^2$. (We note that $\pi(k)$ is often denoted $D_{2k}$.)
- The unitary principal series representations $\pi(t, \varepsilon)$, with
  - (i) $t \in \mathbb{R}$ and $\varepsilon \in \{\pm 1\}$ or
  - (ii) $t \in i(-1/2, 1/2) - \{0\}$ and $\varepsilon = 1$ (the “complementary series”),

obtained by normalized parabolic induction of the character $\text{diag}(y, 1) \mapsto \text{sgn}(y)^\varepsilon |y|^{it}$, for which $\Omega_\pi = -1/4 - t^2$ and $\lambda_\pi = -t^2$.

The only equivalences are that $\pi(t, \varepsilon) \cong \pi(-t, \varepsilon)$. The tempered irreducibles are the $\pi(k)$ and $\pi(t, \varepsilon)$ with $t \in \mathbb{R}$.

Suppose that $\pi$ is not one-dimensional. We may then realize it as follows. If $\pi = \pi(t, \varepsilon)$, set $Q := \mathbb{Z}$; if $\pi = \pi(k)$, set $Q := \{q \in \mathbb{Z} : |q| \geq k\}$ and $\varepsilon := 1$.

We regard $L^2(Q)$ as a Hilbert space with respect to the counting measure, with basis elements given by the $\delta$-masses $e_q$ at each $q \in Q$. It contains the dense subspace $C_c(Q)$ consisting of the finitely-supported elements. We verify readily that the following formulas define an infinitesimally unitary $(g, K)$-module structure on $C_c(Q)$, corresponding to a representative for the isomorphism class of $\pi$:

$$X e_q = (q(q + 1) - \Omega_\pi)^{1/2} e_{q+1},$$

$$Y e_{q+1} = (q(q + 1) - \Omega_\pi)^{1/2} e_q,$$

$$W e_q = q e_q, \quad e^{i\Omega W} e_q = e^{i\Omega} e_q, \quad \text{diag}(-1, 1) e_q = (-1)^\varepsilon e_{-q}.$$
• if \(dg\) is any Haar measure on \(G\), then there is a unique Haar measure \(dx\) on \(g\) so that for \(g = \exp(x)\) with \(x \in G\), we have \(dg = \jac(x)dx\). We say in this case that \(dg\) and \(dx\) are compatibly normalized.

**Lemma.** Let \(\pi\) be a tempered irreducible unitary representation of \(G\). Set \(O_\pi := O(\lambda_\pi)\). For \(x \in G\), we have the identity of generalized functions
\[
\chi_\pi(\exp(x)) = \jac(x)^{-1/2} \int_{\xi \in O_\pi} e^{i(x, \xi)}.
\]

See for instance [13, §6] and references. This says concretely that for each \(\phi \in C_c^\infty(G)\) and Haar measure \(dx\) on \(g\), the operator \(\int_{x \in G} \phi(x) \pi(\exp(x)) dx\) on \(\pi\) belongs to the trace class and has trace \(\int_{\xi \in O_\pi} (\int_{x \in G} \phi(x) \jac(x)^{-1/2} e^{i(x, \xi)} dx)\).

### 3.5. Construction of \(\mu_\pi\). Let \(\pi \in A_0\) with \(\lambda_\pi < 0\). Then \(\pi \cong \pi(t, \varepsilon)\) with \(t = \sqrt{-\lambda_\pi} > 0\). Recall that we have chosen a unit vector \(\varphi_\pi \in \pi\) invariant by \(K^1\). The microlocal lift \(\mu_\pi\) of \(\pi\) is defined on \(K\)-finite smooth functions \(\Phi : X \to \mathbb{C}\) as follows. Set \(\varphi_0 := \varphi_\pi\) and \(s := 1/2 + it\). Define \(\varphi_q\) for \(q \in \mathbb{Z}\) recursively by the formulas \(iX\varphi_q = (s + q)\varphi_{q+1}\) and \(iY\varphi_q = (s - q)\varphi_{q-1}\). Then
\[
\mu_\pi(\Phi) := \sum_{q \in \mathbb{Z}} \langle \varphi_0, \Phi, \varphi_q \rangle.
\]

### 3.6. Branching coefficients. Let \(\pi, \sigma \in A_0\).

**Lemma 1.** If \(\sigma\) is not even, then \(\langle \varphi_1, \Phi, \varphi_2 \rangle = 0\) for all \(\varphi_1, \varphi_2 \in \pi\) and \(\Phi \in \sigma\). In particular, \(\mu_\pi(\Phi) = 0\).

We give the proof below after some otherwise relevant preliminaries.

Assume temporarily that for each \(p \in S\), the involutory Hecke operator \(T_p\) acts trivially (i.e., with eigenvalue +1 rather than −1) on \(\sigma\). The triple product formula [5] then implies that for eigenfunctions \(\varphi_1, \varphi_2 \in \pi\) and \(\Phi \in \sigma\),
\[
|\langle \varphi_1, \Phi, \varphi_2 \rangle|^2 = \mathcal{L}(\pi, \sigma) \int_{g \in G} \langle g\varphi_1, \varphi_1 \rangle \overline{\langle g\varphi_2, \varphi_2 \rangle} \langle g\Phi, \Phi \rangle,
\]
where \(\mathcal{L}(\pi, \sigma)\) is nonnegative real given explicitly in terms of special values of \(L\)-functions; in particular,
\[
\mathcal{L}(\pi, \sigma) \propto \frac{L(\pi \otimes \pi \otimes \sigma, \frac{1}{2})}{L(\ad \pi, 1)^2 L(\ad \sigma, 1)},
\]
where \(L(\cdots)\) denotes the finite part of an \(L\)-function.

**Proof of lemma 1.** Since the distributions \(\mu_\pi\) are invariant by the involutory Hecke operators \(T_p\) (\(p \in S\)), the conclusion is clear if some such operator acts nontrivially on \(\sigma\), so suppose otherwise that each such operator acts trivially. The global root number of \(\sigma\) is then the same as the local root number at the distinguished real place \(q\), which, by hypothesis, is −1. Therefore \(L(\sigma, \frac{1}{2}) = 0\). Since \(L(\pi \otimes \pi \otimes \sigma, \frac{1}{2}) = L(\ad \pi \otimes \sigma, \frac{1}{2}) L(\sigma, \frac{1}{2})\), we have also \(\mathcal{L}(\pi, \sigma) = 0\). The conclusion follows now from (3.2). \(\square\)

**Lemma 2.** Let \(\sigma \in A_0\) be fixed and even. Let \(\pi\) be an \(h\)-dependent element of \(A_0\) with \(\lambda_\pi < 0\) and \(h^2 \lambda_\pi \approx 1\).
(i) Let Ψ ∈ σ be a fixed eigenfunction. Then

\[ |\mu_\pi(\Psi)|^2 \ll h \mathcal{L}(\pi, \sigma). \]  \hspace{1cm} (3.4)

(ii) There is a fixed eigenfunction Ψ ∈ σ so that

\[ |\mu_\pi(\Psi)|^2 \gg h \mathcal{L}(\pi, \sigma). \]  \hspace{1cm} (3.5)

(iii) We have

\[ h \mathcal{L}(\pi, \sigma) \ll 1. \]  \hspace{1cm} (3.6)

Proof. We may assume that Ψ is a $K_1^1$-eigenvector. There are two cases:

- σ is a principal series representation $\pi(t, \varepsilon)$. Our assumption that $\pi$ is even then implies that $\varepsilon = 1$.
- σ is a discrete series representation $\pi(k)$.

Explicit formulas for the matrix coefficient integral of (3.2) in terms of Γ-factors follow from work of Watson [24] and Ichino [5] in the first case and from work of Woodbury [19, Appendix] in the second case. Applying Stirling’s asymptotics to these formulas gives the upper bound (3.4). For the lower bound (3.5), we choose Ψ to be a $K_1^1$-eigenvector of smallest nonnegative weight and appeal again to the explicit formulas. The final estimate (3.6) follows from (3.5) and the trivial bound $\mu_\pi(\Psi) \ll 1$.

Remark. Since we require here estimates rather than explicit formulas, we sketch an alternative proof of lemma 2. Using (3.2), we may write $|\mu_\pi(\Psi)|^2 = \mathcal{L}(\pi, \sigma)|\mu_{\text{loc}}^\pi(\Psi)|^2$, say. One can show by arguments as in §7 and [14, §6.3] that the leading order asymptotics as $h \to 0$ of $|\mu_{\text{loc}}^\pi(\Psi)|^2$ are given by a constant multiple of $h \int_{s \in H} \langle s \Psi, \Psi \rangle$. As in [12, §3.3.1], we may write $\int_{s \in H} \langle s \Psi, \Psi \rangle \approx |\ell(\Psi)|^2$, where $\ell$ is described in the Kirillov model $\Psi \mapsto W_\Psi$ of $\sigma$ (with respect to some fixed nontrivial character) by the absolutely convergent integral $\ell(\Psi) = \int_{y \in \mathbb{R}^n} W_\Psi(y) \frac{dy}{h}$. Thus $|\mu_{\text{loc}}^\pi(\Psi)|^2 \ll h$; moreover, if $\ell(\Psi) \neq 0$, then we can replace “$\ll$” with “$\asymp$”.

3.7. Operator calculus. In this subsection we recall some properties of the operator calculus developed in [13]. We denote by $\pi$ an $h$-dependent unitary representation of $G$ and by $\pi_{\infty}$ its subspace of smooth vectors.

3.7.1. The basic operator assignment. We fix once and for all a cutoff $\chi \in C_c^\infty(G)$ with the following properties:

- The support of $\chi$ is sufficiently small.
- $\chi$ is $[0, 1]$-valued, $\chi(-x) = \chi(x)$, and $\chi = 1$ in a neighborhood of the origin.

For any $h$-dependent Schwartz function $a \in S(\mathfrak{g}^\vee)$, we may define the following objects (see [13, §2] for details):

- $a^\vee : \mathfrak{g} \to \mathbb{C}$, the inverse Fourier transform of $a$.
- $a_h : \mathfrak{g}^\vee \to \mathbb{C}$ the $h$-dependent function given by rescaling: $a_h(\xi) := a(h \xi)$.
- $a_h^\vee : \mathfrak{g} \to \mathbb{C}$, the inverse Fourier transform of the rescaling, thus $a_h^\vee(x) = h^{-\frac{3}{2}} a^\vee(x/h)$.
- $\chi a_h^\vee \in C_c^\infty(\mathfrak{g})$, the cutoff of $a_h^\vee$.
- The compactly-supported smooth distribution $\chi(x) a_h^\vee(x) dx$ on $\mathfrak{g}$, which is supported near the origin.
- The pushforward under the exponential map $x \mapsto g = \exp(x)$ of this distribution, which may be written $\widetilde{\text{Op}}_h(a)(g) dg$ for some $\widetilde{\text{Op}}_h(a) \in C_c^\infty(G)$ supported near the identity; explicitly, $\widetilde{\text{Op}}_h(a)(\exp(x)) = \text{jac}^{-1}(x) \chi(x) a_h^\vee(x)$. 
• An h-dependent integral operator $\text{Op}_h(a : \pi)$ on $\pi^\infty$, abbreviated $\text{Op}_h(a)$ when $\pi$ is clear by context, given by

$$\text{Op}_h(a : \pi) := \pi(\overline{\text{Op}_h(a)}) = \int_{x \in \mathcal{G}} \chi(x)(a^\gamma_h(x)\pi(\exp(x)))\,dx.$$ 

3.7.2. Adjoint. The operator $\text{Op}_h(a)$ extends to a bounded operator on $\pi$ with adjoint $\text{Op}_h(a^\dagger)$. In particular, if $a$ is real-valued, then $\text{Op}_h(a)$ is self-adjoint and $\text{Op}_h(a)^2$ is positive-definite.

3.7.3. Symbol classes. For $\xi$ belonging to any normed space (e.g., $\mathfrak{g}^\wedge$), we set $\langle \xi \rangle := (1 + |\xi|^2)^{1/2}$.

For fixed $0 < \delta < 1/2$ and $m \in \mathbb{Z}$, we write $S^m_\delta$ (denoted “$S^m|h^\delta|$” in [13, §4]) for the space of h-dependent functions $a : \mathfrak{g}^\wedge \to \mathbb{C}$ such that for each fixed multi-index $\alpha \in \mathbb{Z}^{\dim(\mathfrak{g})}$, the corresponding partial derivative $\partial^\alpha a$ enjoys for each $\xi \in \mathfrak{g}^\wedge$ the upper bound

$$\partial^\alpha a(\xi) \ll h^{-\delta|\alpha|/\langle \xi \rangle^m - |\alpha|}.$$ 

(The implied constant is thus allowed to depend upon $\alpha$, but not upon $h$ or $\xi$.) We extend the definition to $m = \infty$ or $m = \infty$ by taking unions or intersections. For instance, an h-independent Schwartz function defines an element of $S^{-\infty}_\delta$, while a polynomial of fixed degree $m \in \mathbb{Z}_{\geq 0}$ and coefficients $O(1)$ defines an element of $S^m_\delta$. Elements of $S^{-\infty}_\delta$ are in particular h-dependent Schwartz functions on $\mathfrak{g}^\wedge$, so the operators $\text{Op}_h(a) := \text{Op}_h(a : \pi)$ may be defined as above.

3.7.4. Smoothing operators. We denote by $\Psi^{-\infty} := \Psi^{-\infty}(\pi)$ the space of h-dependent operators $T$ on $\pi^\infty$ with the property that for any fixed collection $x_1, \ldots, x_m, y_1, \ldots, y_n \in \mathfrak{g}$, the operator norm of $\pi(x_1 \cdots x_m)T\pi(y_1 \cdots y_n)$ is $O(1)$. This is easily seen to be equivalent to the definition of $[13, \S3]$. It is verified in [13, §12.3] (see part (iii) of Theorem 9) that if $\pi$ is irreducible, then

$$T \in \Psi^{-\infty} \implies \text{the trace norm of } T \text{ is } O(1). \quad (3.7)$$

Given an h-dependent scalar $c$ and vector space $V$ consisting of h-dependent quantities, we denote by $cV$ the vector space of h-dependent quantities of the form $cv$, with $v \in V$. We write $h^\infty V$ for the intersection of $h^\delta V$ taken over all fixed $\eta \in \mathbb{R}$. In particular, we may define $h^\infty \Psi^{-\infty}$; we will regard it as the space of “negligible” operators on $\pi^\infty$.

3.7.5. Composition. For $\phi_1, \phi_2 \in C_c^{\infty}(\mathfrak{g})$ supported near the origin, let $\phi_1 \star \phi_2 \in C_c^{\infty}(\mathfrak{g})$ denote the function for which the distribution $(\phi_1 \star \phi_2)(x)\,dx$ on $\mathfrak{g}$ is the pullback of the convolution on $G$ of the images under pushforward of the distributions $\phi_1(x)\,dx$ and $\phi_2(x)\,dx$ on $\mathfrak{g}$. For $a, b \in S^{-\infty}_\delta$, it is verified in [13, §2.5, §4.6] that the (rescaled) star product $a \star_h b := (\chi a^\gamma_h \star \chi b^\gamma_h)^\wedge$ defines an element of $S^{-\infty}_\delta$ which enjoys the composition formula

$$\text{Op}_h(a) \text{Op}_h(b) \equiv \text{Op}_h(a \star_h b) \quad (mod \ h^\infty \Psi^{-\infty}). \quad (3.8)$$

The failure of (3.8) to be an equality is an artefact of the cutoff $\chi$. 

12 PAUL D. NELSON
3.7.6. Equivariance. It follows from \[13, \S 5.5\] that for \( g \in G \) belonging to a fixed compact subset, 
\[
\text{Op}_h(g \cdot a) \equiv \pi(g) \text{Op}_h(a) \pi(g)^{-1} \quad (\text{mod } h^\infty \Psi^{-\infty}),
\]
where \( g \cdot a(\xi) := a(g^{-1} \cdot \xi) \). The error comes from the failure of the cutoff \( \chi \) to be exactly \( G \)-invariant. It will be convenient to assume that \( \chi \) is exactly \( K \)-invariant (by averaging a given cutoff, for instance). We then have 
\[
\text{Op}_h(g \cdot a) = \pi(g) \text{Op}_h(a) \pi(g)^{-1} \quad \text{for all } g \in K.
\]

3.7.7. Star product extension and asymptotics. It is shown in \[13, \S 4.6\] that the star product extends to a compatible family of maps \( \star_h : S^m_\delta \times S^n_\delta \rightarrow S^{m+n}_\delta \) enjoying the asymptotic expansion: for fixed \( J \in \mathbb{Z}_{\geq 0} \), 
\[
a \star_h b \equiv \sum_{0 \leq j < J} h^j a \star^j b \quad (\text{mod } h^{(1-2\delta)J} S^{m+n-J}_\delta),
\]
with \( \star^j \) a fixed polynomial-coefficient differential operator, of order \( j \) in each variable, homogeneous of degree \( j \), satisfying the mapping property \( \star^j : S^m_\delta \times S^n_\delta \rightarrow h^{-2\delta j} S^{m+n-j}_\delta \) and given in the simplest case by \( a \star^0 b = ab \).

3.7.8. Extended operator assignment. It is shown in \[13, \S 5.6\] that \( \text{Op}_h \) extends to a compatible family of maps 
\[
\text{Op}_h : S^m_\delta \rightarrow \{ \text{h-dependent operators on } \pi^\infty \}
\]
for which the composition and equivariance properties (3.8), (3.9) and thus (3.10) remain valid.

3.7.9. Polynomial symbols. It is verified in \[13, \S 5.2\] that if \( p \in S^m_\delta \) is an \( h \)-dependent polynomial function (corresponding to some \( h \)-dependent element of \( \text{Sym}(g_C) \)), then 
\[
\text{Op}_h(p) = \pi(\text{sym}(p_h)),
\]
where \( \text{sym} \) denotes the symmetrization map from \( \text{Sym}(g_C) \) to the universal enveloping algebra of \( g_C \) and (as above) \( p_h(\xi) = p(h\xi) \).

3.7.10. Trace estimates. It is shown in \[13, \S 12.3\] that if \( \pi \) is irreducible and tempered (so that the coadjoint orbit \( \mathcal{O}_\pi \) as well as its rescaling \( h\mathcal{O}_\pi \) may be defined), then for \( a \in S^{\delta-2}_\delta \), the operator \( h \text{Op}_h(a) \) is trace-class, with trace asymptotics described for each fixed \( J \in \mathbb{Z}_{\geq 0} \) by 
\[
\text{trace}(h \text{Op}_h(a)) = \sum_{0 \leq j < J} h^j \int_{h\mathcal{O}_\pi} \mathcal{D}_j a + O(h^{(1-\delta)J}),
\]
where \( \mathcal{D}_j \) is a fixed constant coefficient differential operators of pure degree \( j \), with \( \mathcal{D}_0 a = a \). In particular, 
\[
\text{trace}(h \text{Op}_h(a)) \ll 1.
\]
3.7.11. Clean-up. It follows from [13, §10.3] that if \( \pi \) is irreducible (so that its infinitesimal character \( \lambda_\pi \in \mathbb{R} \) may be defined) and \( a \in S^\infty_\delta \) has the property that the image under the invariant polynomial \( \Lambda : \mathfrak{g}^\wedge \to \mathbb{R} \) of the support of \( a \) is separated by at least \( h^{1/2-\varepsilon} \) from \( h^2 \lambda_\pi \) for some fixed \( \varepsilon > 0 \), then

\[
\text{Op}_h(a) \in (\lambda_\pi)^{-\infty} h^\infty \Psi^{\infty},
\]

(3.15)

where as usual \( \langle \lambda_\pi \rangle := (1 + |\lambda_\pi|^2)^{1/2} \). In particular, the trace norm of \( \text{Op}_h(a) \) is \( \text{O}(\lambda_\pi)^{-\infty} h^\infty \).

(We note a potential point of notational confusion: the rescaled infinitesimal character that we denote here by \( h \lambda_\pi \in \mathbb{R}^\wedge \) is written “\( h \lambda_\pi \in [\mathfrak{g}^\wedge] \cong \mathbb{R}^\wedge \)” in [13]; see [13, §9] for details.)

4. Characterizing microlocal lifts via their symmetry

The methods of §2 apply most naturally to the distributions \( \Psi \mapsto \mu(\pi(f), \Psi) \) attached to integral operators \( \pi(f) \) with \( f \in C^\infty_c(G) \), but the construction of the microlocal lift \( \mu_\pi \) recorded in §3.5 is in terms of differential operators. We thus encounter the problem of constructing \( \mu_\pi \), or at least some asymptotically equivalent distributions, using integral operators.

We begin with some motivational remarks. Recall the asymptotic notation and terminology set in §2. Let \( \pi \) be an \( h \)-dependent element of \( A_0 \) with \( \lambda_\pi < 0 \) and \( h^2 \lambda_\pi \asymp 1 \). Set \( v := \sum_{q \in \mathbb{Z}, |q| \leq h^{-1/2}} \varphi_q \in \pi \), where \( c > 0 \) is chosen so that \( v \) is a unit vector, and \( T := v \otimes \pi \in \pi \otimes \pi \). It follows from calculations of Wolpert [26, §5] (see [9, §3] for a concise account) that for fixed eigenfunctions \( \Psi \),

\[
\mu(T, \Psi) = \langle v \Psi, v \rangle = \mu_\pi(\Psi) + \text{O}(h^{1/2}).
\]

(4.1)

Set \( t := \sqrt{-h^2 \Omega_\pi} = \sqrt{-h^2 \lambda_\pi} + \text{O}(h) \asymp 1 \). We verify readily that \( \pi(hX)v = tv + \text{O}(h^{1/2}) \), \( \pi(hY)v = tv + \text{O}(h^{1/2}) \), and \( \pi(hW)v = \text{O}(h^{1/2}) \); equivalently, for fixed \( Z \in \mathfrak{g} \), \( \pi(hZ)v = Z(\xi(t))v + \text{O}(h^{1/2}) \); in other words, \( v \) is an approximate eigenvector under the first-order differential operators on \( \pi^\infty \) defined by Lie algebra elements, with eigenvalue described by \( \xi(t) \in \mathfrak{g}^\wedge \). We will verify below that some variants of these observations concerning \( v \), phrased in terms of \( T \), give sufficient conditions for (more precise forms of) the estimate (4.1) to hold. Turning to details:

**Definition.** Let \( \pi \) be an \( h \)-dependent irreducible unitary representation of \( G \). Let \( T \) be an \( h \)-dependent positive-definite trace class operator on \( \pi \) such that \( \text{trace}(T) \ll 1 \). Let \( \omega \) be an \( h \)-dependent element of \( \mathfrak{g}^\wedge \) with \( |h \omega| \ll 1 \). Let \( 0 < \delta \leq 1/2 \) be fixed. We say that \( T \) is \( \delta \)-localized at \( \omega \) if for each fixed \( n \in \mathbb{Z}_{\geq 0} \) and each \( h \)-dependent polynomial function \( p : \mathfrak{g}^\wedge_0 \to \mathbb{C} \) of degree \( O(1) \) and coefficients \( O(1) \) that vanishes to order at least \( n \) at \( h \omega \), we have

\[
\text{trace}(\text{Op}_h(p)T) \ll h^{n\delta},
\]

(4.2)

where \( \text{Op}_h(p) := \text{Op}_h(p : \pi) \) is as given by (3.12).

One can verify that the \( T \) considered above is \( 1/2 \)-localized at \( \xi(t) \); we will not need this fact, so we omit the proof.

We may construct integral operators satisfying the above definition:

**Lemma 1** (Integral operators attached to localized symbols are localized). Let \( 0 < \delta < 1/2 \) be fixed. Let \( \omega \in \mathfrak{g}^\wedge \) with \( |h \omega| \asymp 1 \). Let \( a \in h^{-\delta} S^\infty_\delta \) be real-valued.
Let $\pi$ be an $h$-dependent tempered irreducible unitary representation of $G$. Set

$T := h \, \text{Op}_h(a : \pi)^2$.

Suppose that every element of $\text{supp}(a) \cap h \, \mathcal{O}_\pi$ is of the form $h \omega + O(h^3)$. Then $T$ is $\delta$-localized at $\omega$, and

$$\text{trace}(T) = \int_{h \, \mathcal{O}_\pi} a^2 + O(h^{1 - \delta}) = O(1). \tag{4.3}$$

Proof. We have $a^2(\xi) \ll h^{-25}$. By the lemma of §3.2 and the hypotheses concerning $|h \omega|$ and the support of $a$, the set $h \mathcal{O}_\pi \cap \text{supp}(a)$ has symplectic volume $O(h^{25})$. The required trace estimate (4.3) thus follows from §3.7.10. In particular, the operator $T$ is positive-definite with $\text{trace}(T) \ll 1$.

To verify the localization property, fix $n \in \mathbb{Z}_{>0}$ and let $p$, as above, be an $h$-dependent polynomial of degree $O(1)$ and coefficients $O(1)$ that vanishes to order $\geq n$ at $h \omega$. We must check then that $\text{trace}(\text{Op}_h(p)T) \ll h^{n \delta}$.

We pause to observe that for each $q \in S_0^\infty$ and each fixed $J \in \mathbb{Z}_{>0}$,

$$\text{trace}(\text{Op}_h(q)T) = \sum_{0 \leq j_1, j_2 < J} h^{j_1+j_2} \int_{h \, \mathcal{O}_\pi} q \ast^{j_1} (a \ast^{j_2} a) + O(h^{J'}), \tag{4.4}$$

where $J'$ is fixed and $J' \to \infty$ as $J \to \infty$. This estimate follows from the composition formula (3.8), the star product asymptotics (3.11) and the trace estimate (3.13), using (3.7) and (3.14) to clean up the remainders. Since $h \mathcal{O}_\pi \cap \text{supp}(a)$ has symplectic volume $O(h^{25})$, we have also for fixed $j_1, j_2 \geq 0$ that

$$\int_{h \, \mathcal{O}_\pi} q \ast^{j_1} (a \ast^{j_2} a) \ll h^{25} \|q \ast^{j_1} (a \ast^{j_2} a)\|_{L^\infty(h \, \mathcal{O}_\pi)}. \tag{4.5}$$

Returning to the proof of the lemma, choose a ball $B_1$ with origin $h \omega$ and radius $\asymp h^\delta$ so that $\text{supp}(a) \cap h \mathcal{O}_\pi \subseteq B_1$. Let $B_2$ denote the ball with the same origin as $B_1$ but twice the radius. Choose $\phi \in S_\delta^{-\infty}$ taking the value 1 on $B_1$ and the value 0 on the complement of $B_2$. We may then decompose $p = \phi p + (1 - \phi)p$. We apply the above estimates with $q = \phi p$ and $q = (1 - \phi)p$:

- Our assumptions on $p$ imply that $\phi p \in h^{n \delta} S_\delta^{-\infty}$. By (4.4) and the mapping properties of $\ast^j$, the symbol $h^{j_1+j_2} \phi p \ast^{j_1} (a \ast^{j_2} a)$ belongs to $h^{-25+n \delta} S_\delta^{-\infty}$ and thus has $L^\infty$-norm $O(h^{-28+n \delta})$. It follows that

  $$\text{trace}(\text{Op}_h(\phi p)T) = O(h^{n \delta} + h^{J'}).$$

- By construction, $h \mathcal{O}_\pi \cap \text{supp}(1 - \phi) \cap \text{supp}(a) = \emptyset$, so $(1 - \phi)p \ast^{j_1} (a \ast^{j_2} a)$ vanishes identically on $h \mathcal{O}_\pi$, and thus

  $$\text{trace}(\text{Op}_h((1 - \phi)p)T) = O(h^{J'}).$$

We conclude by combining these estimates and taking $J$ large enough. \hfill \Box

We verify next the promised relationship between the above definition and $\mu_\pi$.

**Lemma 2** (Some localized operators define microlocal lifts). Fix a mean zero even eigenfunction $\Psi \in \sigma \in \mathcal{A}_0$. Let $\pi$ be an $h$-dependent element of $\mathcal{A}_0$ such that $\lambda_\pi < 0$ and $h^2 \lambda_\pi \asymp 1$. Abbreviate $\mathcal{L} := \mathcal{L}(\pi, \sigma)$. Let $T$ be an $h$-dependent positive-definite trace class operator on $\pi$ with $\text{trace}(T) \ll 1$. Set $\omega := \xi(\sqrt{-1} \lambda_\pi) \in \pi^\vee$, so that $|h \omega| \asymp 1$. Fix $0 < \delta \leq \frac{1}{2}$.

Suppose that $T$ is $\delta$-localized at $\omega$. Then

$$\mu(T, \Psi) = \text{trace}(T) \mu_\pi(\Psi) + O(h^\delta \sqrt{h \mathcal{L}} + h^\infty). \tag{4.6}$$
Remark 1. The estimate (4.6) implies in particular that \( \mu(T, \Psi) = \text{trace}(T)\mu_\pi(\Psi) + O(h^\delta) \), but this weaker estimate is inadequate for our applications, in which we exploit crucially that \( L \) is “bounded on average” (see §6).

Remark 2. Although lemma 2 is formulated in terms of \( L \)-values, it does not fundamentally exploit the arithmeticity of \( X = \Gamma \backslash G \). What matters are the properties of \( L \) enunciated in §3.6, which make sense for general finite volume quotients (see [13, §1.4], [2]).

Remark 3. Lemma 2 may be used to give a proof of the asymptotic \( H \)-invariance of the measures \( \mu_\pi \), roughly in the spirit of [13, §26.5]; the relevant observations are that

- if \( T \) is \( \delta \)-localized at \( \omega \) and \( g \in G \) is fixed, then \( \pi(g)T\pi(g)^{-1} \) is \( \delta \)-localized at \( g \cdot \omega \), and
- \( H \) stabilizes the elements \( \xi(t) \).

The proof of lemma 2 is a bit tedious, but not difficult, and unrelated to the main novelties of this work. It is basically a quantification of the arguments used to prove (4.1). (Indeed, it is instructive to note that (4.6) recovers (4.1).) For these reasons we postpone the proof to Appendix A.

5. Constructing microlocal lifts via integral operators

Recall the coordinates \( \xi = \begin{pmatrix} \xi_1 & \xi_3 \\ \xi_2 & -\xi_1 \end{pmatrix} \) on \( g^\wedge \). We henceforth fix some \( 0 < \delta < 1/2 \). We denote by \( \tilde{K} \) the set of all real-valued \( a \in h^{-\delta}S_\delta^{-\infty} \) with the following properties:

- \( a(\xi) = 0 \) unless \( \xi_1 > 0 \) and \( \xi_1 \approx 1 \) and \( \xi_2, \xi_3 \ll h^\delta \); equivalently, \( a \) is supported in a \( O(h^\delta) \) neighborhood of some fixed compact subset of \( \{ \xi(t) : t > 0 \} \). In particular, \( a \) vanishes identically on \( O(\lambda) \) unless \( \lambda < 0 \) and \( |\lambda| \approx 1 \).
- We have
  \[
  a(-(w \cdot \xi)) = a(\xi) \tag{5.1}
  \]
  for all \( \xi \in g^\wedge \), where \( w = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in G \). Equivalently, \( a(\xi) \) is invariant under swapping the coordinates \( \xi_2 \) and \( \xi_3 \).

To each \( a \in \tilde{K} \) we attach an \( h \)-dependent element \( k \) of \( C_c^\infty(\mathbb{R},<0) \) by the formula

\[
  k(\lambda) := \int_{O(\lambda)} a^2.
\]

We note that the support condition for \( k \) follows from that for \( a \). We denote by \( K \) the set of \( h \)-dependent functions \( k \) arising in this way. We verify readily that \( K \) has the properties enunciated in §2.

We henceforth work with one such \( k \) together with a corresponding symbol \( a \in \tilde{K} \). We fix \( C \geq 10 \) sufficiently large that \( a(\xi) = 0 \) unless \( 1/3C < \xi_1 < C/3 \). Then \( k(\lambda) = 0 \) unless \( 1/2C < -\lambda < C/2 \).

We say that an \( h \)-dependent irreducible unitary representation \( \pi \) of \( G \) is good if \( \lambda_\pi < 0 \) and \( 1/C \leq -h^2\lambda_\pi \leq C \), and otherwise that \( \pi \) is bad. Informally, the bad \( \pi \) are those whose rescaled infinitesimal characters are sufficiently separated from the support of \( a \) that they play little role in our analysis.
Lemma. Let $\pi$ be an $h$-dependent irreducible unitary representation of $G$. Set
$$T := h \text{Op}_h(a)^2,$$
where $\text{Op}_h(a) := \text{Op}_h(a : \pi)$.

(i) $T$ defines a positive-definite trace-class operator on $\pi$. If $\pi$ is good, then
$$\text{trace}(T) = k(h^2 \lambda_\pi) + O(h^{1-\delta}).$$

If $\pi$ is bad, then
$$\text{trace}(T) = O(h^\infty \lambda_\pi^{-\infty}).$$

(ii) Fix a mean zero even eigenfunction $\Psi \in \sigma \in A_0$, and suppose that $\pi \in A_0$. If $\pi$ is good, then
$$\mu(T, \Psi) = k(h^2 \lambda_\pi) \mu_\pi(\Psi) + O(h^\delta \sqrt{hL} + h^\infty),$$
where $L := L(\pi, \sigma)$. If $\pi$ is bad, then
$$\mu(T, \Psi) = O(h^\infty \lambda_\pi^{-\infty}).$$

Proof. If $\pi$ is bad, then the estimate (5.3) follows from §3.7.11, while (5.5) follows from (5.3) and the inequality $|\mu(T, \Psi)| \leq \text{trace}(T) \|\Psi\|_{L^\infty}$. Suppose that $\pi$ is good. Then $\pi$ is tempered, $\lambda_\pi < 0$ and $h^2 \lambda_\pi \asymp 1$. Moreover, every element of $\text{supp}(a) \cap hO_\pi$ is of the form $h \omega + O(h^\delta)$ with $\omega = \xi(t)$, $t = \sqrt{-\Omega_\pi}$. The hypotheses of (4.3) are thus satisfied, while the conclusion gives (5.3). To deduce (5.4), we combine the lemmas of §4.

6. Quantum variance sums via integral operators

Recall that we have chosen $k \in K$ and a corresponding symbol $a \in \tilde{K}$. Then
$$f := h^{3/2} \tilde{\text{Op}}_h(a) * \tilde{\text{Op}}_h(a)$$
(here $*$ denotes convolution in $C^\infty_c(G)$ with respect to our chosen Haar measure) is an $h$-dependent positive-definite element of $C^\infty_c(G)$, supported in a fixed small neighborhood of the identity element. For each unitary representation $\pi$ of $G$, we have $\pi(f) = h^{1/2} T_\pi$ with $T_\pi := h \text{Op}_h(a : \pi)^2$ as in the lemma of §5. Recall from §1.3 that
$$V(f) = \sum_{\pi \in A_0} \iota_\pi \mu_\pi(\pi(f), \Psi_1) \mu_\pi(\pi(f), \Psi_2) = h \sum_{\pi \in A_0} \iota_\pi \mu(T_\pi, \Psi_1) \mu(T_\pi, \Psi_2).$$

The purpose of this section is to verify the claimed estimate (2.4) relating $V(f)$ to the quantum variance of microlocal lifts, which we copy here for convenience:
$$V(f) = h \sum_{\pi \in A_0} \iota_\pi k(h^2 \lambda_\pi)^2 \mu_\pi(\Psi_1) \mu_\pi(\Psi_2) + O(h^\delta).$$

(2.4)

We begin with an $a$ priori bound:

Lemma. Fix $C \geq 1$, and fix an even $\sigma \in A_0$. Then
$$h^2 \sum_{\pi \in A_0} \iota_\pi L(\pi, \sigma) \ll 1.$$
Proof. This can be deduced using an approximate functional equation and the
Kuznetsov formula as in the work of Luo–Sarnak–Zhao (who in fact obtain and
require asymptotic formulas with strong error terms rather than merely upper
bounds (6.2) of the expected order of magnitude). For completeness, we record
a self-contained proof of (6.2). We assume $k \in K$ chosen so that $k(\lambda) \geq 1$
whenever $C^{-1} \leq -\lambda \leq C$. Let $\pi$ be as in (6.2), so that $|k(h^2 \lambda_\pi)|^2 \geq 1$. Set $T_\pi := h \text{Op}_h(a : \pi)^2$. Fix an eigenfunction $\Psi \in \sigma$ for which (3.4) holds, so that $|k(h^2 \lambda_\pi)|^2 |\mu_\pi(\Psi)|^2 \gg h \mathcal{L}(\pi, \sigma)$. It follows then by (5.4) that $|\mu(T_\pi, \Psi)|^2 \gg |h \mathcal{L}(\pi, \sigma)|$. Thus the LHS of
(6.2) is bounded by a fixed multiple of $h \sum_{\pi \in A_0} t_\pi |\mu(T_\pi, \Psi)|^2$, which is just $\mathcal{V}(f)$ specialized to $\Psi_1 = \Psi_2 = \Psi$. The identity (2.1) and the estimates (2.5) and (2.6)
give an asymptotic formula for $\mathcal{V}(f)$ which implies in particular that $\mathcal{V}(f) \ll 1$.
This completes the proof. (We note that the proofs of the estimates (2.5) and (2.6), given below, do not depend upon the lemma that we proving, so our argument
is non-circular.) \qed

We now verify (2.4). The contribution from bad $\pi$ to (6.1) is adequately estimated using (5.5) and the very weak Weyl law $h^{100} \sum_{\pi \in A_0} \xi_\pi (\lambda_\pi)^{-10} \ll h^{10}$, say. If $\pi$ is good, then we see by (3.4) and (5.4) that
\[
\mu(T_\pi, \Psi_1) \mu(T_\pi, \Psi_2) = k(h t_\pi)^2 \mu_\pi(\Psi_1) \mu_\pi(\Psi_2) + O(h^{1+\delta} \sqrt{L_1 L_2} + h^\infty)
\]
where $L_j := \mathcal{L}(\pi, \sigma_j)$. To discard the error, we apply Cauchy–Schwarz followed by
the above lemma, which gives for $j = 1, 2$
\[
h \sum_{\pi \in A_0} (h^{1+\delta} L_j + h^\infty) \ll h^\delta.
\]
The proof of (2.4) is now complete. In the following sections we will verify (2.5) and (2.6), thereby completing the proof of Theorem 1.

We note for future reference that for $x \in G$
\[
f(\exp(x)) = h^{3/2} \text{jac}^{-1/2}(x) b_h^{\gamma}(x),
\]
where $b \in h^{-\delta} S^{-\infty}_\delta$ is characterized by the identity $b^\gamma = \text{jac}^{-1/2}(a \ast_h a)^\gamma$. By [13, §7.8], it admits an asymptotic expansion $b \sim \sum_{j \geq 0} b_j$ with $b_0 = a^\gamma$ and $b_j \in h^{-2\delta+1-2\delta} S^{-\infty}_\delta$; by this we mean that $b - \sum_{0 \leq j < J} b_j \in h^{-2\delta+1-2\delta} S^{-\infty}_\delta$ for each fixed $J \in \mathbb{Z}_{\geq 0}$.

7. Main term estimates

Before beginning the proof of (2.5), we establish a relevant integral formula. Recall that we have fixed a Haar measure $dg$ on $G$. Let $dx$ denote the compatibly
normalized Haar measure on $g$ (§3.4), and let $d\xi$ denote the corresponding dual
measure on $g^\wedge$, so that for instance $\phi(0) = \int_{x \in G} (\int_{\xi \in g^\wedge} \phi(\xi) e^{(x,\xi)} d\xi) dx$ for $\phi \in C_c^\infty(g^\wedge)$. Let $\Phi \in C_c(G)$ and $\phi_1, \phi_2 \in C_c(g^\wedge)$. The integral
\[
I := \int_{g \in G} f(g) \int_{\xi \in g^\wedge} \phi_1(g \cdot \xi) \phi_2(\xi) d\xi dg
\]
is then independent of these choices of measure. Our immediate aim is to express
$I$ in terms of the normalized symplectic measures on coadjoint orbits. We do this
under the assumption that the \( \phi_j \) are supported on the “negative cone” \( \{ \xi : \Lambda(\xi) < 0 \} = \cup_{t>0} \mathcal{O}(-t^2) \), which is the case relevant for our applications.

To state our result requires some notation. Let \( t > 0 \). Recall that \( \mathfrak{g}(t) \) (see (3.1)) has \( G \)-orbit \( \mathcal{O}(-t^2) \) and stabilizer \( H \) (the diagonal subgroup). For \( \xi, \eta \in \mathcal{O}(-t^2) \), set

\[
G_{\xi \to \eta} := \{ g \in G : g \cdot \eta = \xi \}.
\]

For any choice of elements \( x, y \in G \) with \( x \cdot \xi(t) = \xi \) and \( y \cdot \xi(t) = \eta \), we obtain a bijection \( H \cong G_{\xi \to \eta} \) given by \( s \mapsto xsy^{-1} \). We equip \( G_{\xi \to \eta} \) with the transport of the Haar measure on \( H \), as normalized in §1.2. The measure so-defined on \( G_{\xi \to \eta} \) is independent of the choice of \( x, y \).

**Lemma.** Let \( I \) be as defined above, with \( \phi_1, \phi_2 \) supported on \( \{ \xi : \Lambda(\xi) < 0 \} \). Then

\[
I = \int_{t>0} \int_{\xi, \eta \in \mathcal{O}(-t^2)} \phi_1(\xi) \phi_2(\eta) \left( \int_{G_{\xi \to \eta}} \Phi \right) \frac{dt}{2\pi}.
\]  

(7.1)

**Proof.** Although both sides of the identity are independent of all choices of Haar measure, it is convenient to make explicit choices for the proof. Recall the coordinates and notation of §(3.1). We assume that \( dx = \frac{1}{2} dx_1 dx_2 dx_3 \). This normalizes a Haar measure \( dg \) on \( G \), as well as the dual measure \( d\xi = \frac{1}{8\pi} d\xi_1 d\xi_2 d\xi_3 \) on \( \mathfrak{g}^\perp \).

We equip \( G/H \) with the quotient measure. We note in passing that \( H \) meets both connected components of \( G \), so the quotient \( G/H \) could be replaced by the quotient \( G^1/H^1 \) of connected groups in what follows.

Let \( t > 0 \). We first explicate \( \omega_{\xi(t)} \). Under the differentiated orbit map \( \mathfrak{g} \to T_{\xi(t)}(\mathcal{O}(-t^2)) \), we have \( e_2 \mapsto -2te_3^0 \) and \( e_3 \mapsto 2te_2^0 \). Thus \( \omega_{\xi(t)}(-2te_3^0 \wedge 2te_2^0) = \langle \xi(t), [e_2, e_3] \rangle / 2\pi i = 2t / 2\pi \), and so \( \omega_{\xi(t)} = \frac{1}{2\pi t} d\xi_2 \wedge d\xi_3 \).

Let \( \beta = \frac{1}{8\pi} d\xi_1 \wedge d\xi_2 \wedge d\xi_3 \) denote the differential form on \( \mathfrak{g}^\perp \) corresponding to \( d\xi \). Then \( \beta_{\xi(t)} = \frac{1}{2\pi} d\xi_1 \wedge \omega_{\xi(t)} \). This implies the integral formula

\[
\int_{\xi \in \mathfrak{g}^\perp} \phi(\xi) \, d\xi = \int_{t>0} \left( \int_{\mathcal{O}(-t^2)} \phi \right) \frac{t \, dt}{2\pi}
\]

(7.2)

for \( a \) satisfying the support condition enunciated above.

On the other hand, the Haar measure on \( H \) corresponds to the Haar measure on \( \mathfrak{h} = \text{Lie}(H) \) given by \( dx_1 \). The induced quotient measure on \( G/H \) corresponds to the differential form on \( \mathfrak{g}/\mathfrak{h} \) given at the origin by \( \frac{1}{2} dx_2 \wedge dx_3 \). Under the orbit map isomorphism \( \mathfrak{g}/\mathfrak{h} \cong T_{\xi(t)}(\mathcal{O}(-t^2)) \), we have \( \frac{1}{8\pi} dx_2 \wedge dx_3 \mapsto \pm \frac{1}{8\pi t} d\xi_2 \wedge d\xi_3 = \pm \frac{1}{t} \omega_{\xi(t)} \).

This implies the integral formula

\[
\int_{\mathcal{O}(-t^2)} \phi = t \int_{\mathfrak{g} \in \mathcal{G}/H} \phi(\xi(t)).
\]  

(7.3)

Combining the formulas established thus far, we obtain

\[
I = \int_{\mathfrak{g} \in \mathcal{G}} \Phi(g) \int_{t>0} \int_{\xi \in \mathcal{O}(-t^2)} \phi_1(\xi) \phi_2(\xi) \frac{t \, dt}{2\pi}
\]

\[
= \int_{\mathfrak{g} \in \mathcal{G}} \Phi(g) \int_{t>0} \int_{\mathfrak{g} \in \mathcal{G}/H} \phi_1(\xi(t)) \phi_2(y \cdot \xi(t)) \frac{t^2 \, dt}{2\pi}
\]

\[
= \int_{t>0} \int_{x,y \in \mathcal{G}/H} \phi_1(x \cdot \xi(t)) \phi_2(y \cdot \xi(t)) \int_{\mathfrak{g} \in \mathcal{G}/H} \Phi(xsy^{-1}) \frac{t^2 \, dt}{2\pi},
\]

which simplifies to the required formula. \( \square \)
We now verify (2.5), which we copy here for convenience:

\[ I(f) = \int_{x \in H} \langle s\Psi^1_\tau, \Psi^2_\tau \rangle \int_{t>0} k(-t^2)^2 \frac{dt}{2\pi} + O(h^\delta). \]  

(2.5)

Using (3.10) and (5.1), we compute that \( w \cdot f(g) = f(g^{-1}) \), hence \( \mathcal{G} f = \frac{t + w f}{2} \), and so

\[ I(f) = \int_{g \in G} \langle g \cdot f, f \rangle_G \Phi(g), \quad \Phi(g) := \langle g \Psi^1_\tau, \Psi^2_\tau \rangle. \]  

(7.4)

We compatibly normalize Haar measures on \( G \) and \( g^\wedge \) as above. By change of variables and Parseval, we then have

\[ \langle g \cdot f, f \rangle_G = \int_{x \in g} \text{jac}(x) f(g^{-1}) \exp(x) g f(\exp(x)) \, dx \]

\[ = h^3 \langle g \cdot b_1^\wedge, b_1^\wedge \rangle_g \]

\[ = \langle g \cdot b, b \rangle_{g^\wedge}. \]

As explained in [16, §2.1.6], all of these integrals converge absolutely. For instance, we have \( \Phi(g) \ll \| \text{Ad}(g) \|^{-\eta} \) for some fixed \( \eta > 0 \) by standard bounds for matrix coefficients, while \( \langle g \cdot b, b \rangle_{g^\wedge} \ll \| \text{Ad}(g) \|^{-1} \) by direct calculation. Since \( \int_{g \in G} \| \text{Ad}(g) \|^{-1-\eta} < \infty \), the required convergence follows. The same argument gives also that for any \( c_1, c_2 \in S_0^\wedge \),

\[ \int_{g \in G} \Phi(g) \langle g \cdot c_1, c_2 \rangle_{g^\wedge} \ll \| \text{Ad}(g) \|^{-1}. \]  

(7.5)

Using (7.5) as an \( a \text{ priori} \) estimate and inserting the asymptotic expansion \( b \sim \sum_{j \geq 0} b_j \) noted previously, we deduce that for each fixed \( N \geq 0 \) there is a fixed \( J \geq 0 \) so that

\[ I(f) = \sum_{0 \leq j_1 < j} I_{j_1, j_2} + O(h^N), \quad I_{j_1, j_2} := \int_{g \in G} \Phi(g) \langle g \cdot b_{j_1}, b_{j_2} \rangle_{g^\wedge}. \]

The above lemma gives, with \( \Phi'(\xi, \eta) := \int_{G_{\xi \leftarrow \eta}} \Phi', \) that

\[ I_{j_1, j_2} = \int_{t>0} \int_{[\xi, \eta] \in O(-t^2)} \Phi'(\xi, \eta) \langle \xi, b_{j_2}(\eta) \rangle_{b_{j_2}(\eta)} \frac{dt}{2\pi}. \]  

(7.6)

Using the orbit map at \( \xi(t) \in O(-t^2) \), we can view \( \Phi' \) as a function on \((G/H)^2\). Using the absolute convergence of the integral defining \( \Phi' \) and the smoothness of the vectors \( \Psi_1, \Psi_2 \), we may deduce that the function \( \Phi' \) is smooth (see [13, §18] for related arguments). In particular, \( \Phi' \) is Lipschitz near the origin, where it takes the value \( \Phi'(\xi(t), \xi(t)) = \int_{H} \Phi \). On the other hand, the factor \( b_{j_1}(\xi) b_{j_2}(\eta) \) vanishes unless \( \xi, \eta = \xi(t) + O(h^\delta) \); in that case, it is bounded in magnitude by \( O(h^{-\delta} + (1-2\delta)(j_1+j_2)) \), and we have \( \Phi'(\xi, \eta) = \int_{H} \Phi + O(h^\delta) \). The volume of the set of such pairs \( (\xi, \eta) \) is \( O(h^{4\delta}) \). It follows that

\[ I_{j_1, j_2} \ll h^{(1-2\delta)(j_1+j_2)}. \]

In particular, since \( \delta \) is sufficiently small, we have \( I_{j_1, j_2} \ll h^\delta \) if \((j_1, j_2) \neq (0, 0)\). On the other hand, since \( b_0 = a^2 \) and \( \int_{O(-t^2)} a^2 = k(-t^2) \), we have

\[ I_{0,0} = \left( \int_{H} \Phi \right) \int_{t>0} k(-t^2)^2 \frac{dt}{2\pi} + O(h^\delta). \]
This completes the proof of the required estimate for $\mathcal{I}(f)$.

Remark. An alternative proof may be obtained by first decomposing $\langle g \cdot f, f \rangle$ over the spectrum of $L^2(G)$ as the integral of the Hilbert–Schmidt inner products $\langle \pi(g)\pi(f)\pi(g)^{-1}, \pi(f) \rangle$. This decomposition is reflected above, in (7.6), at the level of coadjoint orbits.

8. Error estimates

We now prepare for the verification of (2.6), which requires some Lie-theoretic preliminaries. We begin with some trivial remarks concerning the complex plane which we hope convey a useful reference picture for the estimates to follow. There are two common choices of coordinates: the rectangular coordinates, described by real and imaginary part, and the polar coordinates, described by radius and angle. The former is adapted to addition, the latter to multiplication. If we restrict the radius to a fixed compact subset of the positive reals and the angle to a sufficiently small neighborhood of the origin, then the real part is likewise restricted to a fixed compact subset of the positive reals; moreover, the imaginary part and the angle are bounded from above and below by constant multiples of one another. These restrictions define a region in the complex plane. Given a scalar-valued function $\phi$ on that region and a small scaling parameter $h > 0$, there are two natural ways to rescale $\phi$ so that its support concentrates along the positive reals: in rectangular coordinates (by scaling the imaginary part) or in polar coordinates (by scaling the angle). The two classes of rescaled functions obtained in this way resemble one another. We aim now to record some analogues and elaborations of these observations with the complex numbers replaced by the $2 \times 2$ matrix algebra $M$.

Such considerations are natural because we are ultimately studying a problem in multiplicative harmonic analysis (the variance sums $\mathcal{V}(f)$ attached to test functions $f$ on the group $G = \text{PGL}_2(\mathbb{R})$) using additive harmonic analysis (via theta functions attached to Schwartz functions $\tau f$ on the matrix algebra $M = M_2(\mathbb{R})$).

We denote by $g = \mathfrak{sl}_2(\mathbb{R})$ the Lie algebra of $G$. We may identify $g$ with the subspace of traceless elements in $M$; then $M = \mathbb{R} \oplus g$, where $\mathbb{R}$ is the subspace of scalar matrices.

Let $\mathcal{R} \subseteq \mathbb{R}^+_+$ and $\mathcal{G} \subseteq g$ be precompact open subsets. We assume that $0 \in \mathcal{G}$ and that $\mathcal{G}$ is star-shaped: $h \mathcal{G} \subseteq \mathcal{G}$ for $h \in [0, 1]$. We assume that $\mathcal{G}$ is taken small enough in terms of $\mathcal{R}$; in particular, we assume that the map

$$\mathcal{R} \times \mathcal{G} \ni (r, x) \mapsto r^{1/2} \exp(x) \in \text{GL}^+_2(\mathbb{R})$$

is an analytic isomorphism onto its image, which we denote by $\mathcal{M}$. We write

$$(\rho, \theta) : \mathcal{M} \rightarrow \mathcal{R} \times \mathcal{G}$$

for the inverse isomorphism. Thus for $v \in \mathcal{M}$, we have $\rho(v) = \sqrt{\det(v)}$, while $\theta(v)$ is the logarithm of the image of $v$ in $G$. We informally regard $\rho(v)$ and $\theta(v)$ as the respective radial and angular parts of $v$.

Every element of $M_2(\mathbb{R})$ may be written uniquely in the form $t + u$, where $t \in \mathbb{R}$ and $u \in \mathfrak{g}$. Write $\gamma(t, u) = (\rho(t, u), \theta(t, u)) =: (r, x)$, say. Then

$$r = \sqrt{t^2 - u^2}, \quad x = \frac{1}{2} \log \frac{t + u}{t - u}.$$

We informally regard $(t, u)$ and $(r, x)$ as the respective rectangular and polar coordinates on $\mathcal{M}$. 
Since $G$ is small, we have $t \neq 0$. Thus $r, t$ are both constrained to lie in compact subsets of $\mathbb{R}^2$, and so
\[ |r| \asymp 1 \asymp |t|. \tag{8.1} \]
Moreover,
\[ |x| \asymp |u|. \tag{8.2} \]
Indeed, we may expand the analytic map
\[ \mathbb{R} \oplus g \ni (t, u) \mapsto \theta(t + u) \in G \tag{8.3} \]
as a Taylor series $\sum_{n \geq 1} c_n(u/t)^n$, and similarly for the inverse map. By (8.1), it follows that $x$ and $u$ tend to zero simultaneously and at the same rate. Since the magnitudes of both are bounded from above, we deduce (8.2).

We now fix a cutoff
\[ q \in C_c^\infty(M) \]
and define a family of maps of Schwartz spaces
\[ S(g) \rightarrow S(M) \]
\[ \phi \mapsto \Phi_h, \]
indexed by $h \in (0, 1)$, by the formula
\[ \Phi_h(t + u) := q(t + hu)\phi(h^{-1}\theta(t + hu)). \]
Informally, $\Phi_h$ is obtained from $\Phi_1$ by rescaling in two stages: first shrinking the angular support in polar coordinates, then stretching the imaginary support in rectangular coordinates. The reference picture discussed above hopefully renders the following observation unsurprising:

**Lemma.** The family of maps $\phi \mapsto \Phi_h$ is equicontinuous for the Schwartz topology.

**Proof.** The derivatives of the map $(t, u) \mapsto q(t + hu)$ are uniformly bounded.

Moreover, the derivatives of the map $(t, u) \mapsto h^{-1}\theta(t + hu)$ are bounded, uniformly for $t + hu \in \text{supp}(q)$. Indeed, we may write each such derivative as a convergent Taylor series; applying the triangle inequality gives a finite bound for the magnitude of this series, and the observation following (8.3) implies that this bound improves as $h$ decreases.

By the chain rule, we deduce that the derivatives of $\Phi_h$ at $t + u$ are dominated by derivatives of $\phi$ at those elements $h^{-1}\theta(t + hu)$ for which $t + hu \in \text{supp}(q)$. For such elements, the estimates (8.1) and (8.2) give $|t| \asymp 1$ and $|h^{-1}\theta(t + hu)| \asymp |u|$. Thus the rapid decay of the derivatives of $\Phi_h$ follows from that of $\phi$. $\square$

We now apply these considerations to establish (2.6), which we copy here for convenience:
\[ E_{\tau_1, \tau_2}(\bigwedge\tau_1 f, \bigwedge\tau_2 f) \ll h^{1-\delta}. \tag{2.6} \]
Let $\tau \in \{\tau_1, \tau_2\}$. We assume $R$ taken large enough to contain the support of $r \mapsto W(\tau r)$, and take $G$ small enough. We may assume that $\text{Op}$ was defined with respect to a cutoff supported in $G$. Set
\[ q^\tau(v) := \frac{W(\tau R(v)^2)}{|\tau R(v)^2|} \text{Jac}^{-1/2} \chi'(\theta(v)) \]
and $\phi := b^\tau$, with $b$ as in §6. We see then by unwinding the definitions that for $v \in M$,
\[ \bigwedge\tau f(v) = q^\tau(v) h^{-3/2} \phi(h^{-1}\theta(v)). \]
Let $\Phi_h^\tau$ be defined as in §8 using $q^\tau$ and $\phi$. Then $D^\tau f(t + h u) = h^{-3/2} \Phi_h^\tau(t + u)$, i.e.,
\[ D^\tau f = D_{1/h} \Phi_h^\tau, \]
so by Theorem 2,
\[ \mathcal{E}_{\tau_1, \tau_2}(D^\tau_{1/h} \Phi_h^\tau_1, D^\tau_{1/h} \Phi_h^\tau_2) \ll h^{1} \log(h^{-1}) C(\Phi_h^\tau_1) C(\Phi_h^\tau_2), \]
for some fixed continuous seminorm $C$. We appeal now to the (h-uniform) continuity of $\phi \mapsto \Phi_h^\tau$ noted in §8, together with the continuity of the map $a \mapsto b$ composed with the Fourier transform $b \mapsto b^\tau = \phi$, to write $C(\Phi_h^\tau_1) C(\Phi_h^\tau_2) = O(C(a)^2)$ for some fixed continuous seminorm $C$ on $\mathcal{S}(g^\wedge)$. The definition of $S_h^{-\infty}$ implies that $C(a) \ll h^{-N\delta}$ for some fixed $N \in \mathbb{Z}_{\geq 0}$, so we may conclude by taking $\delta' := (2N + 1)\delta$.

**Remark.** The proof of (2.6) recorded above is a bit different from that of the corresponding estimate in [16, §8], whose analogue here would be to exploit the smoothness of $\Psi$ and the diagonal $G$-invariance of $(f, \Psi) \mapsto \mu(\pi(f), \Psi)$ to “fatten up” the symbol $a$ under the adjoint action. The argument given here produces weaker estimates, but is a bit shorter and simpler.

The proof of our main result, Theorem 1, is now complete.

### 9. Removing the arithmetic weights

Here we fulfill the promise made in Remark 3 of §1.3 by explaining (a bit informally) how the modification factor (1.5) obtained by Sarnak–Zhao arises from the perspective of our method. Recall that $\iota_\pi = L(S)(\text{ad} \pi, 1)$. The idea is to write the desired unweighted variance sums
\[
\lim_{h \to 0} \sum_{0 < -h^2 \lambda_+ < 1} \mu_\pi(\Psi_1) \mu_\pi(\Psi_2)
\]
as the double limit
\[
\lim_{h \to 0} \lim_{S' \to S} V(S'), \quad V(S') := h \sum_{0 < -h^2 \lambda_+ < 1} L^{(S')}_{\pi}(\text{ad} \pi, 1) \mu_\pi(\Psi_1) \mu_\pi(\Psi_2), \quad (9.1)
\]
where $\lim_{S' \to S}$ denotes the limit taken over increasing finite subsets $S' \supseteq S$ of the set of finite primes of $F$, ordered by inclusion. We then try to swap the limits. The subtlety in making this precise is that the Euler product of $L(\text{ad} \pi, 1)$ fails to converge absolutely, but because 1 is at the edge of the critical strip, the failure is mild, so we at least expect the naive swapping of limits to produce the correct answer. The result [16, Thm 2] of the prequel applies to any $S' \supseteq S$: combining it with the estimates of this article shows that as far as main terms are concerned,
\[ V(S') \approx V(S) \prod_{p \in S' - S} c_\sigma(p), \]
where
\[ c_\sigma(p) := \frac{1}{\zeta_p(2) L_p(\sigma, \frac{1}{2})} \int_{g \in \text{PGL}_2(F_p)} \langle g \cdot f, f \rangle_{\text{PGL}_2(F_p)} \Phi(g), \quad (9.2) \]
where
- $\zeta_p(s) = (1 - |p|^{-s})^{-1}$ denotes the local zeta function for $F_p$,
- $L_p(\sigma, s) = (1 - \lambda_+(p) p^{-s} + p^{-2s})^{-1}$ denotes the local factor for $L(\sigma, s)$ at $p$,
- we fix an arbitrary Haar measure on $\text{PGL}_2(F_p)$ (the quantity $c_\sigma(p)$ will not ultimately depend upon this choice).
• $f$ is the normalized characteristic function $\text{vol}(J_p)^{-1}1_{J_p}$ of a maximal compact subgroup $J_p$ of $\text{PGL}_2(F_p)$,
• $g \cdot f(x) := f(g^{-1}xg)$ as usual, and
• $\Phi$ is the normalized bi-$J_p$-invariant matrix coefficient of the unramified representation of $\text{PGL}_2(F_p)$ corresponding to the action of $T_p$ on $\sigma$, so that for instance $\Phi(1) = 1$ and $\Phi(\text{diag}(\varpi, 1)) = \frac{\lambda_\sigma(p)}{|p|^{1/2} + |p|^{-1/2}}$, with $\varpi \in F_p$ a uniformizer.

The modification factor (1.5) is then explained by the following local calculation:

**Lemma.** $c_\sigma(p) = \frac{1}{\xi_p(2)}(1 - \frac{\lambda_\sigma(p)}{|p|^{1/2} + |p|^{-1/2}})$.

**Proof.** This follows by direct calculation with the Macdonald formula [3, Thm 4.6.6] and the Cartan decomposition; we leave it to the interested reader. □

10. Heuristics

In this section we record a heuristic derivation of the limiting variance (1.3) obtained in our main result (or more precisely, its unweighted variant discussed in §9). This serves both to check of the correctness of our results and to study the apparent deviation in the behavior of variance sums between arithmetic and non-arithmetic settings observed by Luo–Sarnak [11].

10.1. Overview. We revoke our general assumptions by taking for $\Gamma$ any discrete cocompact subgroup of $G$ (possibly non-arithmetic). The definitions of §1 adapt fairly painlessly to this setting, possibly after making some choices in the event of eigenvalue multiplicities. We fix $\Psi \in C^\infty(X)$ of mean zero (not necessarily an eigenfunction) and suppose given some unit vectors $v_\pi \in \pi$ for each $\pi$ in some varying family $F \subset A_0$ (e.g., we might take for $v_\pi$ the vectors defined at the beginning of §4.1, so that the microlocal lifts are given asymptotically by $\Psi \mapsto \langle v_\pi \Psi, v_\pi \rangle$). Our aim is to understand the asymptotics of the variance sums $\sum_{\pi \in F} |\langle v_\pi \Psi, v_\pi \rangle|^2$.

Translated into representation-theoretic language, the basic idea underlying the semiclassical predictions (see [4], [27, §15.6], [17, §4.1.3]) in the generic non-arithmetic setting is to postulate that

$$\langle v_\pi \Psi, v_\pi \rangle \approx \langle v_\pi' \Psi, v_\pi' \rangle$$

whenever $\pi'$ and $\pi$ are “close” (denoted $\pi' \approx \pi$) as quantified by their isomorphism classes under the group $G$. Then

$$\sum_{\pi \in F} |\langle v_\pi \Psi, v_\pi \rangle|^2 \approx \sum_{\pi \in F} E_{\pi' \approx \pi} |\langle v_\pi \Psi, v_\pi' \rangle|^2.$$  

The RHS of (10.2) may often be studied rigorously via semiclassical analysis, leading to predictions concerning the LHS.

This heuristic often requires some modification. One way that (10.1) can fail is when the representations $\pi \in F$ are self-dual, i.e., equal to their complex conjugates $\overline{\pi}$ (the representation-theoretic incarnation of “time-reversal symmetry”); in that case,

$$\langle v_\pi \Psi, v_\pi \rangle = \langle v_{\overline{\pi}} \Psi, v_{\overline{\pi}} \rangle \text{ with } \overline{v_\pi} \in \pi.$$  

Suppose for concreteness that $\overline{v_\pi} = wv_\pi$ for some involutory element $w \in G$. It follows then that the distributions $\Psi \mapsto \langle v_\pi \Psi, v_\pi \rangle$ are $w$-invariant. On the other hand, there is no obvious reason to suspect that the more general distributions...
\[ \Psi \mapsto \langle v_\pi \Psi, v_\pi \rangle \] are \( w \)-invariant when \( \pi' \neq \pi \), so (10.1) can fail, most obviously when \( w \Psi = -\Psi \). The simplest way to repair this failure is to restrict from the outset to observables \( \Psi \) for which \( w \Psi = \Psi \).

Another way that (10.1) can fail is if the space \( X \) admits a nontrivial correspondence \( T \). We may assume then that \( \pi \) and \( \pi' \) are \( T \)-eigenspaces with eigenvalues \( \lambda \) and \( \lambda' \). These eigenvalues may bias the asymptotics of \( \langle v_\pi \Psi, v_\pi \rangle \). The bias is most striking when \( T \) is an involution and \( T \Psi = \Psi \), in which case parity considerations imply that \( \langle v_\pi \Psi, v_\pi \rangle = 0 \) unless \( \lambda = \lambda' \). Thus (10.1) fails. We can repair it by strengthening the closeness condition \( \pi' \approx \pi \) to require also that \( \lambda' \approx \lambda \). The RHS (10.2) can now be estimated using semiclassical analysis on “\( G \times T \),” leading to a modification of the expected variance asymptotics. For instance, in the case of an involution, the modification is given by doubling; the factor \( 2^S \) in Theorem 1 may be explained in this way in terms of the involutory Hecke operators \( T_p \) (\( p \in S \)).

Such modified heuristics extend easily to finite commuting families of correspondence, but their further extension to arithmetic settings as in Theorem 1, with infinitely many commuting correspondences \( T_p \), requires some care. A naive approach is to run the heuristics first taking into account only those \( T_p \) for \( p \) belonging to some large finite set \( P \), and then to take the limit as \( P \) increases. We implement this naive approach in detail below. We will encounter main terms involving finite Euler products \( \prod_{p \in P} L_p(\sigma, \frac{1}{2}) \). Modulo the subtle business of identifying these with their formal limit, we will see that the resulting predictions are consistent with our rigorous results and also with the triple product formula and \( L \)-function analysis.

10.2. General predictions. Turning to details, choose a Haar measure on \( G \) and denote by \( G^\wedge \) the tempered dual, equipped with Plancherel measure. Equip \( X = \Gamma \backslash G \) with the quotient Haar. Suppose given a nice subset \( \tilde{F} \) of \( G^\wedge \) and a nice function \( f : G \rightarrow \mathbb{C} \) such that

- for \( \pi \in \tilde{F} \), the operator \( \pi(f) \) is the orthogonal projection onto the line \( \mathbb{C} v_\pi \) spanned by some unit vector \( v_\pi \in \pi \), and
- for \( \pi \notin \tilde{F} \), we have \( \pi(f) = 0 \).

(In practice, such assumptions are satisfied exactly only for \( p \)-adic groups \( G \); for real groups, one should instead smoothly weight the family \( \tilde{F} \) and work with families of vectors in each \( \pi \in \tilde{F} \), as illustrated in the bulk of this paper. We omit such technicalities from this heuristic discussion to keep the exposition clean.) We then have the spectral decomposition \( f(g) = \int_{\pi \in \tilde{F}} \langle v_\pi, g \cdot v_\pi \rangle \) and the formula \( \int_{\pi \in \tilde{F}} |\langle g v_\pi, v_\pi \rangle|^2 = (g \cdot f, f) \), with \( g \cdot f(x) = f(g^{-1} x g) \) as before and the latter inner product taken in \( L^2(G) \). We take for \( \mathcal{F} \subset A_0 \) the set of all \( \pi \) whose isomorphism class belongs to \( \tilde{F} \). The pretrace formula reads \( \sum_{\pi \in \mathcal{F}} v_\pi(x) v_\pi(y) = \sum_{\gamma \in \Gamma} f(x^{-1} \gamma y) = \sum_{\gamma \in \Gamma} \int_{\pi \in \tilde{F}} \langle x v_\pi, \gamma y v_\pi \rangle \). Dividing this by the Weyl law \( |\mathcal{F}| \approx \text{vol}(X) \text{vol}(\tilde{F}) \) gives

\[
E_{\pi \in \mathcal{F}} v_\pi(x) v_\pi(y) \approx \frac{1}{\text{vol}(X)} \sum_{\gamma \in \Gamma} E_{\pi \in \tilde{F}} \langle x v_\pi, \gamma y v_\pi \rangle, \tag{10.3}
\]

where \( E \) denotes the average (taken with respect to the counting measure on \( \mathcal{F} \) and the Plancherel measure on \( \tilde{F} \)).
Suppose temporarily that \( \tilde{F} \) is sufficiently concentrated near some given \( \pi \in A_0 \) that
\[
\langle gv_{\pi'}, v_{\pi'} \rangle \approx \langle gv_{\pi}, v_{\pi} \rangle
\]
for all \( \pi' \in \tilde{F} \). Then (10.3) simplifies to
\[
E_{\pi' \in F} v_{\pi'}(x) v_{\pi'}(y) \approx \frac{1}{\text{vol}(X)} \sum_{\gamma \in \Gamma} \langle x v_{\pi}, \gamma y v_{\pi} \rangle. 
\] (10.4)

Assume that quantum ergodicity holds in the strong form
\[
\langle g(v_{\pi}\Psi), v_{\pi}\Psi \rangle \approx \frac{1}{\text{vol}(X)} \langle gv_{\pi}, v_{\pi} \rangle \langle g\Psi, \Psi \rangle,
\] (10.5)
at least on average over \( \pi \). From (10.4), (10.5) and “unfolding,” we obtain
\[
E_{\pi' \in F} |\langle v_{\pi}\Psi, v_{\pi'} \rangle|^2 \approx \frac{1}{\text{vol}(X)^2} \int_{g \in G} |\langle gv_{\pi}, v_{\pi} \rangle|^2 \langle g\Psi, \Psi \rangle. 
\] (10.6)

We now relax our assumption that \( \tilde{F} \) be concentrated and consider fairly general families. By the Weyl law, we expect
\[
\sum_{\pi \in \tilde{F}} |\langle gv_{\pi}, v_{\pi} \rangle|^2 \approx \text{vol}(X) \int_{\pi \in \tilde{F}} |\langle gv_{\pi}, v_{\pi} \rangle|^2 = \text{vol}(X) \langle g \cdot f, f \rangle. 
\] (10.7)

Suppose that the heuristic (10.1) holds. We may then apply (10.6) to the family \( \{\pi' : \pi' \approx \pi\} \), substitute the result into (10.2), and appeal to (10.7), giving the prediction
\[
\sum_{\pi \in \tilde{F}} |\langle v_{\pi}\Psi, v_{\pi} \rangle|^2 \approx \frac{1}{\text{vol}(X)} \int_{g \in G} \langle g \cdot f, f \rangle \langle g\Psi, \Psi \rangle 
\] (10.8)
subject to the modifications indicated above in the case of the “time-reversal symmetry” \( \pi = \overline{\pi} \) or the presence of nontrivial correspondences on \( X \).

We note that this argument applies to fairly general quotients of the form \( X = \Gamma \backslash G \). This generality will be exploited below.

10.3. Applications. We now specialize (10.8) in three ways.

10.3.1. Generic non-arithmetic lattices. First, we take \( G := \text{PSL}_2(\mathbb{R}), \Gamma \leq G \) a “generic” (i.e., trivial commensurator) non-arithmetic cocompact lattice, \( v_\pi \) as in §4.1, and \( \tilde{F} = \{ \pi : 0 < -h^2 \lambda_\pi < 1 \} \) (the relevant definitions apply equally well to \( \text{PSL}_2(\mathbb{R}) \) as to \( \text{PGL}_2(\mathbb{R}) \)). We may take \( f \) essentially (i.e., up to the constant factor \( h^{1/2} \)) as in §6, with \( k \) approximating the characteristic function of the interval \((-1, 0)\). We assume that \( \pi = \overline{\pi} \) and that \( \Psi \) is invariant by \( w = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in G \). By the analogue of (2.5) for \( \text{PSL}_2(\mathbb{R}) \), the prediction (10.8) simplifies to
\[
\frac{1}{2\pi \text{vol}(X)} \int_{u \in \mathbb{R}} \langle e^{u/2} \Psi, \Psi \rangle du,
\] (10.9)
which may be seen to agree with the prediction of [4].
10.3.2. Arithmetic lattices. Second, we address the setting of Theorem 1. We focus for simplicity on the diagonal case \( \Psi \in \sigma \in A_0 \), and assume that \( \Psi = \Psi^{\text{sym}} \) as in (1.4). Our task is accomplished most directly by applying (10.2) to the adelic quotient \( \Gamma \setminus G \), (notation as in §1.2), which has the effect of incorporating the nontrivial correspondences on the real quotient. We take for \( f \) a tensor product \( \otimes_p f_p \) over the places \( p \) of \( F \), with the local factor \( f_q \) at the distinguished real place \( q \) is as in the previous paragraph and the remaining \( f_p \) the normalized characteristic functions of compact open subgroups \( J_p \) as in §1.2. The (absolutely divergent) integral on the RHS of (10.8) factors (formally) as a product \( \prod_p I_p \) of local integrals over all places \( p \) of \( F \); at places other than \( q \), the component of \( \langle q \Psi, \Psi \rangle \) is the normalized bi-\( J_p \)-invariant matrix coefficient of \( \sigma \) at \( p \), as in §9, while at \( q \) we take the usual matrix coefficient. The local integrals \( I_p \) have already been computed, either in this work or its prequel. The local integral \( I_q \) is given by (7.4). For archimedean places \( p \) other than the distinguished real place \( q \), we have \( I_p = 1 \) by [16, (2.16)]. For finite primes \( p \in S \), we have \( I_p = 1 \) by [16, (2.15)]. For finite primes \( p \notin S \), we have \( I_p = L_p(\sigma, \frac{1}{2})(1 - \frac{\lambda_\sigma(p)}{|p|^{s/2} + |p|^{1/2}}) \) by §9. Modulo identifying \( \prod_{p \in S} L_p(\sigma, \frac{1}{2}) \) with \( L(S)(\sigma, \frac{1}{2}) \), we derive from (10.8) the prediction
\[
\sum_{0 < h \leq \lambda_q < 1} |\mu_\sigma(\Psi)|^2 \approx \left( \frac{c'_\sigma}{2\pi \text{vol}(X)} \right) \int_{u \in \mathbb{R}} \langle \left( e^{u/2} \ e^{-u/2} \right) \Psi, \Psi \rangle du,
\]
\[
c'_\sigma := 2^{#S+1} L(S)(\sigma, \frac{1}{2}) \prod_{p \notin S} (1 - \frac{\lambda_\sigma(p)}{|p|^{s/2} + |p|^{1/2}}).
\]
We verify readily that this prediction agrees with the unweighted variant of Theorem 1 discussed in §9.

10.3.3. Comparison with L-function heuristics. Thirdly, we verify that the predictions of §10.2, and hence likewise our main results, are (unsurprisingly) consistent with the triple product formula and standard heuristics for averages of families of \( L \)-functions. We include this discussion not only as a further check of our calculations, but also because we feel that it offers an interesting semiclassical perspective on the product formula itself.

We continue to take \( \Gamma \setminus G = G(F) \setminus G(A) \). Equip \( G(A) \) with Tamagawa measure, so that \( \text{vol}(\Gamma \setminus G) = 2 \), and factor the measure on \( G(A) \) over the places \( v \) of \( F \) in such a way that for \( p \notin S \), the local measure at \( G_p = G(F_p) \) assigns volume one to a maximal compact subgroup. The main result of [5] then says that for \( v_\pi \) and \( \Psi \) unramified outside \( S \),
\[
|\langle v_\pi \Psi, v_\pi \rangle|^2 = \frac{1}{8} \ell(S)(\pi, \sigma) w(\pi),
\]
where
\[
\ell(S)(\pi, \sigma) := \zeta_F^{(S)}(2) \frac{L(S)(\pi \otimes \pi \otimes \sigma, \frac{1}{2})}{L(S)(\text{ad} \pi, 1)^2 L(S)(\text{ad} \sigma, 1)}
\]
and \( w(\pi) := \prod_{v \in S} \int_{F_v} \int_{\prod_{v \in S} G_v} |\langle v_\pi, v_\pi \rangle|^2 |\langle g \Psi, \Psi \rangle| \). We assume (for simplicity, and without loss of generality) that our family \( \mathcal{F} \) has been chosen sufficiently concentrated that the weight \( \pi \mapsto w(\pi) \) is essentially constant over its support. By
comparing with (10.1) and (10.6), we see that our predictions translate to
\[
E_{\pi \in F} \ell^{(S)}(\pi, \sigma) \approx \frac{8}{\text{vol}(\Gamma \setminus G)^2} \prod_{p \notin S} I_p,
\]
(10.12)
where \(I_p\) is as above. (As before, the product diverges and is to be understood formally; in particular, it hides the factor \(L^{(S)}(\sigma, \frac{1}{2})\).) We may spectrally expand \(I_p\) as the integral over unramified \(\pi_p \in G_p^\wedge\) of the integral \(\int_{g \in G_p} \Xi_{\pi_p}(g)^2 \Xi_{\pi_p}(g)\) of normalized unramified matrix coefficients. Ichino–Ikeda [6, Theorem 1.2] have shown that the latter integral evaluates to the local Euler factor \(L(\pi, \sigma)\), so that in fact
\[
I_p = \int_{\text{unramified } \pi_p \in G_p^\wedge} \ell_p(\pi_p, \sigma_p).
\]
(10.13)
We may factor \(L(\pi \otimes \sigma \otimes \frac{1}{2}) = L(\text{ad } \pi \otimes \sigma, \frac{1}{2}) L(\sigma, \frac{1}{2})\). The family \(\pi \mapsto L(\text{ad } \pi \otimes \sigma, \frac{1}{2})\) is self-dual with positive root numbers (assuming \(\sigma\) even) and orthogonal symmetry type, so random matrix heuristics (see for instance [13, §1.2]) predict that
\[
E_{\pi \in F} \ell^{(S)}(\pi, \sigma) \approx 2 \prod_{p \notin S} I_p,
\]
with \(I_p\) as given by (10.13). Since \(8/\text{vol}(\Gamma \setminus G)^2 = 2\), those heuristics are consistent with ours.

Appendix A. Calculations with raising and lowering operators

Here we record the proof of lemma 2 of §4. We recall that our task is to verify, under certain assumptions, the estimate
\[
\mu(T, \Psi) = \text{trace}(T) \mu_\pi(\Psi) + O(h^d \sqrt{\text{vol } \mathcal{L}^\infty}).
\]
(2.6)
We have \(\pi \cong \pi(t, \varepsilon)\) with \(t > 0\). We realize \(\pi(t, \varepsilon)\) as \(L^2(\mathbb{Z})\) as in §3. There is then a unique equivariant (isometric) isomorphism \(j_\pi : L^2(\mathbb{Z}) \rightarrow \pi\) that maps the basis element \(\varphi_\pi\) to \(\varphi_\pi\). Thus \(\varphi_q\), as in the construction of \(\mu_\pi\), is equal to \(b(q) j_\pi(\varphi_q)\), where \(b(q)\) is defined recursively by
\[
b(q + 1) = b(q) \frac{i}{s + q} \sqrt{q(q + 1) - \Omega_\pi}.
\]
Since \(t \in \mathbb{R}\), we have \(|s + q|^2 = q(q + 1) - \Omega_\pi\), and so \(|b(q)| = 1\) for all \(q\). Moreover, since \(t > 0\), we have for fixed \(q\) that \(b(q) = 1 + O(1/t)\). Thus the vectors \(\varphi_q\) are asymptotically quite close to the \(j_\pi(\varphi_q)\).

By a limiting argument, it will suffice to consider the case that \(T\) is a finite rank operator \(T = \sum_i j_\pi(v_i) \otimes j_\pi(v_i)\) attached to some finite orthogonal subset \(\{v_i\}\) of \(L^2(\mathbb{Z})\). For \(q, \xi \in \mathbb{Z}\), we set \(T(q, \xi) := \sum_i v_i(q) v_i(q + \xi)\), so that \(T = \sum_{q, \xi} T(q, \xi) j_\pi(\varphi_q) \otimes j_\pi(\varphi_{q+\xi})\), and \(\Psi(q, \xi) := \langle j_\pi(\varphi_q) \Psi, j_\pi(\varphi_{q+\xi}) \rangle\), so that
\[
\mu_\Psi(\Psi) = \sum_{\xi} b(\xi) \Psi(0, \xi)
\]
and
\[
\mu(T, \Psi) = \sum_{q, \xi} T(q, \xi) \Psi(q, \xi)
\]
and

\[ \text{trace}(T) = \sum_{i,q} |v_i(q)|^2 = \sum_q T(q, 0). \]

By Cauchy–Schwarz and the assumed trace estimate for \( T \),

\[ \sum_q |T(q, \xi)| \leq \sum_q |v_i(q) v_{i+1}(q + \xi)| \leq \sum_{q,i} |v_i(q)|^2 = \text{trace}(T) \ll 1. \quad (A.1) \]

Set \( \tau := \sqrt{-\Omega} \), so that \( \omega = \xi(\tau) \). We temporarily abbreviate \( X, Y, W := \pi(X), \pi(Y), \pi(W) \). We will make use of the following consequences of our assumption that \( T \) is \( \delta \)-localized at \( \omega \):

\[ \text{trace}((hW^n T) \ll h^{n\delta} \text{ for each fixed } n \in \mathbb{Z}_{\geq 0}, \quad (A.2) \]

\[ \text{trace}((hX - h \tau T) \ll h^{\delta}. \quad (A.3) \]

Indeed, we have \( X(h \omega) = Y(h \omega) = h \tau \) and \( W(h \omega) = 0 \), so the polynomial \( p = W^n \) vanishes to order \( n \) at \( h \omega \) and satisfies \( \text{sym}(p_h) = (hW)^n \), while the polynomial \( p = X - h \tau \) vanishes to order 1 at \( h \omega \) and satisfies \( \text{sym}(p_h) = hX - h \tau \).

By \((A.2)\) with \( n = 2 \), we have

\[ \sum_{q,i} |h q|^2 |v_i(q)|^2 = \text{trace}((hW^2 T) \ll h^{2\delta}. \]

Using Cauchy–Schwarz as above, it follows that for fixed \( \xi \),

\[ \sum_q |T(q, \xi)| \cdot |h q| \ll h^{\delta}. \quad (A.4) \]

We now fix \( 0 < \delta' < \delta \), and argue using \((A.2)\) for arbitrary fixed \( n \) that

\[ \sum_{q:|h q| \geq h^{\delta'}} |T(q, \xi)| \ll h^{\infty}. \quad (A.5) \]

We now investigate the consequences of \((A.3)\). We have

\[ hX T = \sum_{q, \xi} T(q, \xi) h \sqrt{\tau^2 + q(q + 1)} e_{q+1} \otimes \tau q \xi, \]

thus

\[ \text{trace}(hX T) = \sum_q h \sqrt{\tau^2 + q(q + 1)} T(q, 1). \quad (A.6) \]

Our assumptions on \( \pi \) imply that \( h \tau \ll 1 \). We estimate the latter sum in the range \( |h q| \geq h^{\delta'} \) using \((A.5)\) and in the range \( |h q| < h^{\delta'} \) using the Taylor expansion

\[ h \sqrt{\tau^2 + q(q + 1)} = h \tau + O(|h q|). \]

The contribution to \((A.6)\) of the remainder in this expansion is treated using \((A.4)\), and then we extend the sum to all \( q \) again using \((A.5)\). We obtain in this way that

\[ \text{trace}(hX T) = h \tau \sum_q T(q, 1) + O(h^{\delta}). \quad (A.7) \]

Since \( T(q, 1) = \sum_i v_i(q) v_i(q + 1) \), the estimate \((A.3)\) thus translates to

\[ \sum_{q,i} |v_i(q)|^2 = \sum_{q,i} v_i(q) v_i(q + 1) + O(h^{\delta}). \quad (A.8) \]
We deduce that
\[ \sum_{q,i} |v_i(q) - v_i(q + 1)|^2 \ll h^\delta \] (A.9)
by expanding the square and applying (A.8) twice. By iterating (A.9), we deduce
that for each fixed \( \xi \),
\[ \sum_{q,i} |v_i(q) - v_i(q + \xi)|^2 \ll h^\delta. \] (A.10)
By Cauchy–Schwarz, it follows finally that for each fixed \( \xi \),
\[ \sum_q T(q, \xi) = \text{trace}(T) + O(h^\delta). \] (A.11)
Recall that \( \Psi(q, \xi) = 0 \) unless \( |\xi| \leq C \) for some fixed \( C \). We have the trivial
bound
\[ |\Psi(q, \xi)| \leq \|\Psi\|_{L^\infty(X)} \ll 1 \] (A.12)
for all \( q, \xi \). Suppose now that \( |h q| \leq h^\delta \). We claim then that
\[ \Psi(q, \xi) \ll \sqrt{h L} + h^\infty \] (A.13)
and
\[ \Psi(q, \xi) = \Psi(0, \xi) + O(h \sqrt{h L}) \] (A.14)
We will prove these when \( q \geq 0 \); an analogous argument applies to negative \( q \).
For \( j \in \mathbb{Z} \geq 0 \) and \( q, \xi \in \mathbb{Z} \), let \( \Psi^j(q, \xi) \) be defined like \( \Psi(q, \xi) \), but with \( \Psi \) replaced
with \( X^j \Psi \). We will work in what follows with fixed values of \( j \), so that
\( \|X^j \Psi\| \ll 1 \).
By (3.6), we have for each fixed \( j \)
\[ \Psi^j(0, \xi) \ll \sqrt{h L}. \] (A.15)
We have also for fixed \( j \) the trivial bound
\[ \Psi^j(q, \xi) \ll 1, \] (A.16)
as in (A.12).
We now argue recursively using the following instance of “partial integration:”
the integral over \( X \) of \( \int_X (e^{q}(e_{q+\xi})X^j \Psi) \) vanishes. Expanding this out, we
obtain with \( f(q) := h \sqrt{\tau^2 + q(q + 1)} \) that
\[ f(q + \xi) \Psi^j(q, \xi) = f(q) \Psi^j(q + 1, \xi) + h \Psi^{j+1}(q, \xi + 1). \]
For \( q \) in the indicated range and \( \xi \ll 1 \), we have \( f(q) \asymp 1 \) and \( f(q + \xi) = f(q) + O(h) \).
Hence for such \( q \) and \( \xi \),
\[ \Psi^j(q + 1, \xi) - \Psi^j(q, \xi) \ll h(\Psi^j(q, \xi)) + \Psi^j(q + 1, \xi) + \Psi^{j+1}(q, \xi + 1) \]. (A.17)
Fix \( J \in \mathbb{Z}_{\geq 0} \) and then \( C \in \mathbb{R}_{\geq 1} \) sufficiently large, and set
\[ \beta_j(q) := C(1 + C h)^{-q} \sup_{\xi} |\Psi^j(q, \xi)|. \]
We consider \( q \geq 0 \) with \( |h q| \leq h^\delta \). Having chosen \( C \) large enough, the estimate
(A.17) implies
\[ \beta_j(q + 1) \leq \beta_j(q) + C h \beta_{j+1}(q) \quad (0 \leq j < J). \] (A.18)
Similarly, by (A.16),
\[ \beta_J(q) \leq 1. \] (A.19)
Thus the sequence of \((J + 1)\)-dimensional row vectors 
\[
\beta(q) := (\beta_0(q), \beta_1(q), \ldots, \beta_{J-1}(q), 1)
\]
satisfy 
\[
\beta(q) \leq \beta(0) M^q, \quad M := \begin{pmatrix}
1 & C h & 1 & \cdots & \cdots \\
C h & 1 & \cdots & \cdots \\
\vdots & \vdots & \ddots & \cdots & \vdots \\
\cdots & \cdots & \cdots & 1 & C h \\
C h & 1 & \cdots & \cdots & 1
\end{pmatrix}.
\]
We also have, by (A.15) and the estimate \((1 + C h)^q = 1 + O(h^\delta)\), the initial bound 
\[
\beta_j(0) \ll \sqrt{h L} \quad (0 \leq j < J) \quad (A.20)
\]
These lead to 
\[
\beta_j(q) \ll \sqrt{h L} + h^{(J-j)\delta} \quad (A.21)
\]
Taking \(j = 0\) and recalling that \(J\) was arbitrary, we obtain (A.13), and also its analogue for \(\Psi^1\); inserting the latter into the \(j = 0\) case of (A.17) then gives (A.14).

We now combine the above estimates to conclude. Expanding the definitions, we have 
\[
\mu(T, \Psi) = \text{trace}(T) \mu_\pi(\Psi) + S_1 + S_2 + S_3
\]
where 
\[
S_1 := \sum_q \sum_{\xi} (T(q, \xi) - \text{trace}(T)) \Psi(0, \xi), \quad (A.22)
\]
\[
S_2 := \sum_q T(q, \xi)(\Psi(q, \xi) - \Psi(0, \xi)), \quad (A.23)
\]
\[
S_3 := \text{trace}(T) \sum_{\xi} (1 - b(\xi)) \Psi(0, \xi). \quad (A.24)
\]
Using (A.13) and (A.11), we may see that \(S_1 \ll h^\delta \sqrt{h L}\). To bound \(S_2\), we estimate the contribution from \(|h q| \geq h^\delta\) via (A.5) and (A.13). We then estimate the remaining contribution via (A.14). We obtain 
\[
S_2 \ll \sqrt{h L} \sum_{\xi: |\xi| \leq C} \sum_q |T(q, \xi)| \cdot |h q| + h^\infty \ll h^\delta \sqrt{h L} + h^\infty. \quad (A.25)
\]
For \(S_3\), we use that \(\Psi(0, \xi) \ll \sqrt{h L}\) and that \(\Psi(0, \xi) \neq 0\) only if \(|\xi| = O(1)\), in which case \(b(\xi) = 1 + O(h)\); thus \(S_3 \ll h \sqrt{h L}\). This completes the proof of (4.6).

Acknowledgements. We gratefully acknowledge the support of NSF grant OISE-1064866 and SNF grant SNF-137488 during the work leading to this paper.

References

[1] Nalini Anantharaman and Steve Zelditch. Patterson-Sullivan distributions and quantum ergodicity. *Ann. Henri Poincaré*, 8(2):361–426, 2007.
[2] Joseph Bernstein and Andre Reznikov. Subconvexity bounds for triple \(L\)-functions and representation theory. *Ann. of Math. (2)*, 172(3):1679–1718, 2010.
[3] Daniel Bump. *Automorphic Forms and Representations*, volume 55 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1997.
[4] Mario Feingold and Asher Peres. Distribution of matrix elements of chaotic systems. *Phys. Rev. A* (3), 34(1):591–595, 1986.
[5] Atsushi Ichino. Trilinear forms and the central values of triple product \(L\)-functions. *Duke Math. J.*, 145(2):281–307, 2008.

[6] 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.

[7] A. A. Kirillov. Merits and demerits of the orbit method. *Bull. Amer. Math. Soc. (N.S.)*, 36(4):433–488, 1999.

[8] 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.

[9] Elon Lindenstrauss. On quantum unique ergodicity for \(\Gamma \backslash \mathbb{H} \times \mathbb{H}\). *Internat. Math. Res. Notices*, (17):913–933, 2001.

[10] Elon Lindenstrauss. Invariant measures and arithmetic quantum unique ergodicity. *Ann. of Math. (2)*, 163(1):165–219, 2006.

[11] Wenzhi Luo and Peter Sarnak. Quantum variance for Hecke eigenforms. *Ann. Sci. École Norm. Sup. (4)*, 37(5):769–799, 2004.

[12] Philippe Michel and Akshay Venkatesh. The subconvexity problem for \(GL_2\). *Publ. Math. Inst. Hautes Études Sci.*, (111):171–271, 2010.

[13] P. D. Nelson and A. Venkatesh. The orbit method and analysis of automorphic forms. *ArXiv e-prints*, May 2018.

[14] Paul D. Nelson. Microlocal lifts and quantum unique ergodicity on \(GL_2(\mathbb{Q}_p)\). *Algebra & Number Theory*, 12, 01 2016.

[15] Paul D. Nelson. Quantum variance on quaternion algebras, I. preprint, 2016.

[16] Paul D. Nelson. Quantum variance on quaternion algebras, II. preprint, 2017.

[17] Stéphane Nonnenmacher. Anatomy of quantum chaotic eigenstates. In *Chaos*, volume 66 of *Prog. Math. Phys.*, pages 193–238. Birkhäuser/Springer, Basel, 2013.

[18] Zeév Rudnick and Peter Sarnak. The behaviour of eigenstates of arithmetic hyperbolic manifolds. *Comm. Math. Phys.*, 161(1):195–213, 1994.

[19] Paul D. Nelson. Quantum variance on quaternion algebras, I. preprint, 2016.

[20] Paul D. Nelson. Quantum variance on quaternion algebras, II. preprint, 2017.

[21] Marie-France Vignéras. *Arithmétique des algèbres de quaternions*, volume 800 of *Lecture Notes in Mathematics*. Springer, Berlin, 1980.

[22] Scott A. Wolpert. Semiclassical limits for the hyperbolic plane. *Duke Math. J.*, 108(3):449–509, 2001.

[23] Steven Zelditch. Recent developments in mathematical Quantum Chaos. *ArXiv e-prints*, November 2009.

[24] Steven Zelditch. Pseudodifferential analysis on hyperbolic surfaces. *J. Funct. Anal.*, 68(1):72–105, 1986.

[25] Steven Zelditch. Uniform distribution of eigenfunctions on compact hyperbolic surfaces. *Duke Math. J.*, 55(4):919–941, 1987.

[26] Steven Zelditch. The averaging method and ergodic theory for pseudo-differential operators on compact hyperbolic surfaces. *J. Funct. Anal.*, 82(1):38–68, 1989.

[27] Peng Zhao. Quantum variance of Maass-Hecke cusp forms. *Comm. Math. Phys.*, 297(2):475–514, 2010.

ETH ZURICH, DEPARTMENT OF MATHEMATICS, RÄMISTRASSE 101, CH-8092, ZÜRICH, SWITZERLAND

E-mail address: paul.nelson@math.ethz.ch