# Algebra & Number Theory

Volume 12

2018

No. 9


## Microlocal lifts and quantum unique ergodicity on $GL_2(\mathbb{Q}_p)$


Paul D. Nelson



msp

ALGEBRA AND NUMBER THEORY 12:9 (2018)
dx.doi.org/10.2140/ant.2018.12.2033

# Microlocal lifts and quantum unique ergodicity on $GL_2(\mathbb{Q}_p)$

Paul D. Nelson

We prove that arithmetic quantum unique ergodicity holds on compact arithmetic quotients of $\mathrm{GL}_2(\mathbb{Q}_p)$ for automorphic forms belonging to the principal series. We interpret this conclusion in terms of the equidistribution of eigenfunctions on covers of a fixed regular graph or along nested sequences of regular graphs.

Our results are the first of their kind on any $p$-adic arithmetic quotient. They may be understood as analogues of Lindenstrauss's theorem on the equidistribution of Maass forms on a compact arithmetic surface. The new ingredients here include the introduction of a representation-theoretic notion of "$p$-adic microlocal lifts" with favorable properties, such as diagonal invariance of limit measures; the proof of positive entropy of limit measures in a $p$-adic aspect, following the method of Bourgain–Lindenstrauss; and some analysis of local Rankin–Selberg integrals involving the microlocal lifts introduced here as well as classical newvectors. An important input is a measure-classification result of Einsiedler–Lindenstrauss.



## 1. Introduction

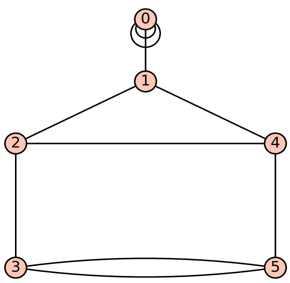
**1.1.** ***Overview.*** Let $p$ be a prime number. This article is concerned with the limiting behavior of eigenfunctions on compact arithmetic quotients of the group $G := \mathrm{GL}_2(\mathbb{Q}_p)$. A rich class of such quotients is parametrized by the definite quaternion algebras $B$ over $\mathbb{Q}$ that split at $p$. A maximal order $R$ in such an algebra and an embedding $B \hookrightarrow M_2(\mathbb{Q}_p)$ give rise to a discrete cocompact subgroup $\Gamma := R[1/p]^\times$ of $G$. Fix one such $\Gamma$. The corresponding arithmetic quotient $\boldsymbol{X} := \Gamma\backslash G$ is then compact; in interpreting this, it may help to note that the center of $\Gamma$ is the discrete cocompact subgroup $\mathbb{Z}[1/p]^\times$ of $\mathbb{Q}_p^\times$. In adelic terms, we may identify $\boldsymbol{X}$ with $B^\times\backslash B_{\mathbb{A}}^\times / B_\infty^\times \prod_{\ell\neq p} R_\ell^\times$ (see Section 2.1 for notation).

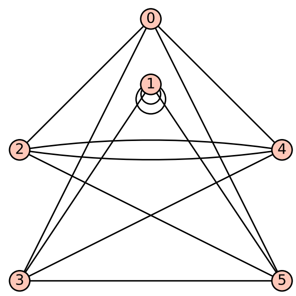
*MSC2010:* primary 58J51; secondary 22E50, 37A45.
*Keywords:* arithmetic quantum unique ergodicity, microlocal lifts, representation theory.

The space $\boldsymbol{X}$ is a $p$-adic analogue of the cotangent bundle of an arithmetic hyperbolic surface, such as the modular surface $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$. It comes with commuting families of Hecke correspondences $T_\ell$ indexed by the primes $\ell \neq p$ (see Section 3.1). To zeroth approximation, the space $\boldsymbol{X}$ is modeled by its minimal quotient $\boldsymbol{Y} := \boldsymbol{X}/K = \Gamma\backslash G/K$ by the maximal compact subgroup $K := \mathrm{GL}_2(\mathbb{Z}_p)$ of $G$. That quotient $\boldsymbol{Y}$ comes with an additional Hecke correspondence $T_p$. To simplify the exposition of Section 1.1, it will be convenient to assume that

$$(\text{the torsion subgroup of } \Gamma) = \{\pm 1\}. \tag{1}$$

Then $\boldsymbol{Y}$ may be safely regarded as an undirected $(p+1)$-regular finite multigraph (see [Vignéras 1980; Serre 2003; Lindenstrauss 2006b, §8]), whose adjacency matrix is $T_p$. The simplifying assumption (1) holds when the underlying quaternion algebra has discriminant (say) 73, in which case the graph $(\boldsymbol{Y}, T_p)$ may be depicted as follows when $p = 2, 3$:[1]




Such graphs and their eigenfunctions appear naturally in several contexts, and have been extensively studied since the pioneering work of Brandt [1943] and Eichler [1955]; they specialize to the $p$-isogeny graphs of elliptic curves in finite characteristic [Gross 1987, §2], provide an important tool for constructing spaces of modular forms [Pizer 1980], and their remarkable expansion properties have been studied and applied in computer science following [Lubotzky et al. 1988].

To study the space $\boldsymbol{X}$ at a finer resolution than that of its minimal quotient $\boldsymbol{Y}$, we introduce for each pair of integers $m, m'$ the notation $m..m' := \{m, m+1, \dots, m'\}$ and set

$$\boldsymbol{Y}_{m..m'} := \left\{ \begin{array}{l} \text{nonbacktracking paths } x = (x_m \to x_{m+1} \to \cdots \to x_{m'}) \\ \hfill \text{indexed by } m..m' \text{ on the graph } (\boldsymbol{Y}, T_p) \end{array} \right\}. \tag{2}$$

We will recall in Definition 10 the standard group-theoretic realization of $\boldsymbol{Y}_{m..m'}$ as a quotient of $\boldsymbol{X}$. We may and shall identify $\boldsymbol{Y}_{0..0}$ with $\boldsymbol{Y}$. For $m..m' \supseteq n..n'$, we define compatible surjections $\boldsymbol{Y}_{m..m'} \to \boldsymbol{Y}_{n..n'}$ by forgetting part of the path. For example, if $N \geqslant 0$, then the map $\boldsymbol{Y}_{-N..N} \to \boldsymbol{Y}_{0..0} = \boldsymbol{Y}$ sends a path $x$ as in (2) to its central vertex $x_0$. We define $L^2(\boldsymbol{Y}_{m..m'})$ with respect to the normalized counting measure, so that the maps $\boldsymbol{Y}_{m..m'} \to \boldsymbol{Y}_{n..n'}$ are measure-preserving.

We wish to study the asymptotic behavior of "eigenfunctions" in $L^2(\boldsymbol{Y}_{m..m'})$ as $|m - m'| \to \infty$. From the arithmetic perspective, there is a distinguished collection of such eigenfunctions, whose definition is

[1] The images were produced using the "Graph" and "BrandtModule" functions in Sage [2015].

analogous to that of the set of normalized classical holomorphic newforms of some given weight and level:

**Definition 1** (newvectors). Let $L^2_{\mathrm{new}}(\boldsymbol{Y}_{m..m'}) \subseteq L^2(\boldsymbol{Y}_{m..m'})$ denote the space of functions $\varphi : \boldsymbol{Y}_{m..m'} \to \mathbb{C}$ that are orthogonal to pullbacks from $\boldsymbol{Y}_{n..n'}$ whenever $n..n' \subsetneq m..m'$. Let $\mathcal{F}_{m..m'} \subseteq L^2_{\mathrm{new}}(\boldsymbol{Y}_{m..m'})$ be an orthonormal basis consisting of $\varphi$ for which:

- the pullback of $\varphi$ to $\boldsymbol{X} = \Gamma\backslash G$ generates an irreducible representation of $G = \mathrm{GL}_2(\mathbb{Q}_p)$ under the right translation action, and
- $\varphi$ is an eigenfunction of the Hecke operator $T_\ell$ (see Section 3.1) for all primes $\ell \neq p$.

It is known[2] then that $|\mathcal{F}_{m..m'}| \asymp |\boldsymbol{Y}_{m..m'}| \asymp p^{|m-m'|}$ for $|m-m'|$ sufficiently large. To simplify the exposition of Section 1.1, we focus on the symmetric intervals $-N..N$. Fix $n \in \mathbb{Z}_{\geqslant 0}$. Let $N \geqslant n$ be an integral parameter tending off to $\infty$. Denote by $\mathrm{pr} : \boldsymbol{Y}_{-N..N} \twoheadrightarrow \boldsymbol{Y}_{-n..n}$ the natural surjection. For $\varphi \in \mathcal{F}_{-N..N}$, we may define a probability measure $\mu_\varphi$ on $\boldsymbol{Y}_{-n..n}$ by setting

$$\mu_\varphi(E) := \frac{1}{|\boldsymbol{Y}_{-N..N}|} \sum_{x \in \boldsymbol{Y}_{-N..N} : \mathrm{pr}(x) \in E} |\varphi|^2(x).$$

For example, in the instructive special case $n = 0$, the measures $\mu_\varphi$ live on the base graph $\boldsymbol{Y}_{0..0} = \boldsymbol{Y}$ and assign to subsets $E \subseteq \boldsymbol{Y}$ the number

$$\mu_\varphi(E) = \frac{1}{|\boldsymbol{Y}_{-N..N}|} \sum_{x = (x_{-N} \to \cdots \to x_N) \in \boldsymbol{Y}_{-N..N} : x_0 \in E} |\varphi|^2(x),$$

which quantifies how much mass $\varphi : \boldsymbol{Y}_{-N..N} \to \mathbb{C}$ assigns to paths whose central vertex lies in $E$.

**Question 2.** Fix $n \in \mathbb{Z}_{\geqslant 0}$. Let $N \geqslant n$ traverse a sequence of positive integers tending to $\infty$. For each $N$, choose an element $\varphi_N \in \mathcal{F}_{-N..N}$. What are the possible limits of the sequence of measures $\mu_{\varphi_N}$ on the space $\boldsymbol{Y}_{-n..n}$?

The following conjecture has not appeared explicitly in the literature, but may be regarded nowadays as a standard analogue of the arithmetic quantum unique ergodicity conjecture of Rudnick–Sarnak [1994] (see [Sarnak 2011; Nelson et al. 2014]).

**Conjecture 3.** *In the context of Question 2, the uniform measure on $\boldsymbol{Y}_{-n..n}$ is the only possible weak limit. In other words, for any sequence $\varphi_N \in \mathcal{F}_{-N..N}$ and any $E \subseteq \boldsymbol{Y}_{-n..n}$,*

$$\lim_{N \to \infty} \mu_{\varphi_N}(E) = \frac{|E|}{|\boldsymbol{Y}_{-n..n}|}.$$

[2] One may verify this by applying the trace formula for $L^2(\Gamma\backslash G)$ to an element $f \in C_c^\infty(G)$, as in [Nelson 2017], that defines the orthogonal projection onto $L^2_{\mathrm{new}}(\boldsymbol{Y}_{m..m'})$, or alternatively by appealing to the Eichler/Jacquet–Langlands correspondence, which identifies $\mathcal{F}_{m..m'}$ with the set of normalized weight two newforms on $\Gamma_0(p^{|m-m'|} d_B)$, with $d_B$ the discriminant of $B$, and appealing to standard formulas for dimensions of spaces of newforms.

Conjecture 3 predicts that for any sequence $\varphi_N \in \mathcal{F}_{-N..N}$, the corresponding sequence of $L^2$-masses $\mu_{\varphi_N}$ equidistributes under pushforward to any fixed space $\boldsymbol{Y}_{-n..n}$. One can formulate this conclusion more concisely in terms of equidistribution on the compact space $\varprojlim \boldsymbol{Y}_{-n..n}$ of infinite bidirectional nonbacktracking paths, or equivalently, on the space $\boldsymbol{X} = \Gamma\backslash G$.

We note that the quantum unique ergodicity conjecture of Rudnick–Sarnak [1994] includes the case of *nonarithmetic* compact hyperbolic surfaces, while Conjecture 3, as formulated here, is specific to the arithmetic setting. We indicate in Remark 31 how one might formulate it more generally.

By explicating the triple product formula [Ichino and Ikeda 2010], one can show that Conjecture 3 follows from an open case of the subconvexity conjecture, which in turn follows from GRH; the latter can be shown to imply more precisely that

$$\mu_{\varphi_N}(E) = \frac{|E|}{|\boldsymbol{Y}_{-n..n}|} + O(p^{-(1+o(1))N/2}) \tag{3}$$

for fixed $n$. There are nowadays well-developed techniques (see for instance [Nelson 2016, §1.4]) to establish that:

- the prediction (3) holds for $\varphi_N$ outside a hypothetical exceptional subset of density $o(1)$,
- if (3) is true, it is essentially optimal, and
- Conjecture 3 holds for $\varphi_N$ outside a hypothetical exceptional subset of extremely small density $|\mathcal{F}_{-N..N}|^{-1/2+o(1)}$. (This may be understood as a very strong form of "quantum ergodicity," which would assert the analogous conclusion with density $o(1)$; compare with [Anantharaman and Le Masson 2015; Le Masson and Sahlsten 2017].)

The problem of eliminating such exceptions entirely (in the present setting and related ones) has proven subtle.

For context, we recall some instances in which the difficulty indicated above has been overcome; notation and terminology should be clear by analogy.

**Theorem 4** [Lindenstrauss 2006b]. *Let $\Gamma'\backslash\mathbb{H}$ be a compact hyperbolic surface attached to an order in a nonsplit indefinite quaternion algebra. Let $\varphi$ traverse a sequence of $L^2$-normalized Hecke–Laplace eigenfunctions on $\Gamma'\backslash\mathbb{H}$ with Laplace eigenvalue tending to $\infty$. Then the $L^2$-masses $\mu_\varphi$ equidistribute.*

**Theorem 5** (N, N–Pitale–Saha, Hu [Nelson 2011; Nelson et al. 2014; Hu 2018]). *Fix a natural number $q_0$. Let $q$ traverse a sequence of natural numbers tending to $\infty$. Let $\varphi$ be an $L^2$-normalized holomorphic Hecke newform on the standard congruence subgroup $\Gamma_0(q)$ of $\mathrm{SL}_2(\mathbb{Z})$. Then the pushforward to $\Gamma_0(q_0)\backslash\mathbb{H}$ of the $L^2$-mass of $\varphi$ equidistributes.*

We may of course specialize Theorem 5 to powers of a fixed prime:

**Theorem 6** (N, N–Pitale–Saha, Hu [Nelson 2011; Nelson et al. 2014; Hu 2018]). *Fix a prime $p$ and a nonnegative integer $n_0$. Let $n$ traverse a sequence of natural numbers tending to $\infty$. Let $\varphi$ be an $L^2$-normalized holomorphic Hecke newform on $\Gamma_0(p^n)$. Then the pushforward to $\Gamma_0(p^{n_0})\backslash\mathbb{H}$ of the $L^2$-mass of $\varphi$ equidistributes.*

Conjecture 3 is in the spirit of Theorem 6, save a crucial distinction to be discussed in due course (see Remark 19). Unfortunately, the method underlying the proof of Theorem 6, due to Holowinsky–Soundararajan [2010], is fundamentally inapplicable to Conjecture 3 due to its reliance on parabolic Fourier expansions, which are unavailable on the compact quotient $\boldsymbol{X}$. We will instead develop here a method more closely aligned with that underlying the proof of Theorem 4.

To describe our result, we must recall that the elements of $\mathcal{F}_{-N..N}$ may be partitioned according to the isomorphism class of the representation of $G = \mathrm{GL}_2(\mathbb{Q}_p)$ that they generate. Any such representation has unramified central character,[3] and for $N$ sufficiently large, is (isomorphic to) either:

- a (ramified) principal series representation (see Section 5.3), or
- a (supercuspidal) discrete series representation.

(See for instance [Schmidt 2002].) A (computable) positive proportion of elements of $\mathcal{F}_{-N..N}$ belongs to either category. The dichotomy here is analogous to that on $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ between Maass forms (principal series) and holomorphic forms (discrete series).

**Theorem 7** (main result). *The conclusion of Conjecture 3 holds if $\varphi_N$ belongs to the principal series.*

Theorem 7 represents the first genuine instance of arithmetic quantum unique ergodicity in the level aspect on a compact arithmetic quotient and also the first on any $p$-adic arithmetic quotient. It says that for a sequence $\varphi_N \in \mathcal{F}_{-N..N}$ belonging to the principal series, the corresponding $L^2$-masses equidistribute under pushforward to any fixed space $\boldsymbol{Y}_{-n..n}$.

**Remark 8.** Our result might be described concisely as *arithmetic quantum unique ergodicity on the path space over the fixed regular graph* $(\boldsymbol{Y}, T_p)$ and as contributing to the growing literature concerning quantum chaos on regular graphs (see [Brooks and Lindenstrauss 2010; 2013; Anantharaman and Le Masson 2015]). Alternatively, one could fix an auxiliary split prime $\ell \neq p$, regard $(\boldsymbol{Y}_{-N..N}, T_\ell)$ as traversing an inverse system of $(\ell+1)$-regular graphs, and interpret Theorem 17 as a form of arithmetic quantum unique ergodicity for such a sequence of graphs.

**Remark 9.** Assuming the multiplicity hypothesis that an element $\varphi \in \mathcal{F}_{-N..N}$ generating an irreducible principal series representation of $G$ is automatically an eigenfunction of the $T_\ell$ for $\ell \neq p$ (which is inspired by analogy from the conjectural simplicity of the spectrum of the Laplacian on $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$), Theorem 17 may be understood as telling us something new about individual finite graphs $(\boldsymbol{Y}, T_p)$, such as those pictured above, together with their realization as $\Gamma\backslash G/K$.

As indicated already, the proof of Theorem 7 is patterned on that of Theorem 4. An important ingredient in the proof of Theorem 4 is the existence of a measure $\mu$ on $\Gamma'\backslash \mathrm{SL}_2(\mathbb{R})$, called a *microlocal lift*, with the properties:

- $\mu$ lifts the measure $\lim_{j\to\infty} \mu_{\varphi_j}$ on $\Gamma'\backslash\mathbb{H}$.

[3] One may verify that "unramified central character" implies "trivial central character" in the present setup, but this special feature will not play an important role for us.

- $\mu$ is invariant under right translation by the diagonal subgroup of $\mathrm{SL}_2(\mathbb{R})$.
- $(\mu_{\varphi_j})_j \mapsto \mu$ is compatible with the Hecke operators (see [Silberman and Venkatesh 2007, Theorem 1.6] for details); this third property is that which is not obviously satisfied by the classical construction via charts and pseudodifferential calculus.

The known construction of $\mu$ with such properties, due to Zelditch and Wolpert (see [Zelditch 1987; Wolpert 2001; Lindenstrauss 2001]) and generalized by Silberman–Venkatesh [2007], relies heavily upon explicit calculation with raising and lowering operators in the Lie algebra of $\mathrm{SL}_2(\mathbb{R})$, which have no obvious $p$-adic analogue. One point of this paper is to introduce such an analogue and to investigate systematically its relationship to the classical theory of local newvectors. (The restriction to principal series in Theorem 7 then arises for the same reason that Lindenstrauss's argument does not apply to holomorphic forms of large weight: the absence of a "microlocal lift" invariant by a split torus.) The resulting construction may be of independent interest; for instance, it should have applications to the test vector problem (see Section 1.5 and Remark 50).

A curious subtlety of the argument, to be detailed further in Remark 26, is that the "lift" we construct is not a lift in the traditional sense (except against spherical observables, and even then only for $p \neq 2$). It instead satisfies a weaker "equidistribution implication" property which suffices for us. This subtlety is responsible for the most technical component of the argument (Section 6.3).

In the remainder of Section 1 we formulate our main result in a slightly more general setup (Section 1.2), introduce a key tool (Section 1.3), give an overview of the proof (Section 1.4), interpret our results in terms of $L$-functions (Section 1.5), and record some further remarks and open questions (Section 1.6).

**1.2. *Main results: general form.*** In this section we formulate a generalization of Theorem 4 in representation-theoretic language, which we adopt for the remainder of the paper.

**Definition 10.** Define the compact open subgroup

$$K_{m..m'} := \begin{bmatrix} \mathfrak{o} & \mathfrak{p}^{-m} \\ \mathfrak{p}^{m'} & \mathfrak{o} \end{bmatrix}^{\times}, \quad \mathfrak{o} := \mathbb{Z}_p, \mathfrak{p} := p\mathbb{Z}_p \tag{4}$$

of $G$. Each such subgroup is conjugate to $K_{0..n}$ for $n = m' - m \geqslant 0$, which is in turn analogous to the congruence subgroup $\Gamma_0(p^n)$ of $\mathrm{SL}_2(\mathbb{Z})$. Assuming (1), one has compatible bijections

$$\boldsymbol{X}/K_{m..m'} = \Gamma\backslash G/K_{m..m'} \xrightarrow{\cong} \boldsymbol{Y}_{m..m'},$$
$$\Gamma g K_{m..m'} \mapsto (x_m \to x_{m+1} \to \cdots \to x_{m'}) \text{ where } x_j := \Gamma g \begin{pmatrix} p^{-j} & \\ & 1 \end{pmatrix} K,$$

with $\boldsymbol{Y}_{m..m'}$ as defined in (2).

**Definition 11.** The space $\mathcal{A}(\boldsymbol{X})$ of *smooth* functions on $\boldsymbol{X}$ consists of all functions $\varphi : \boldsymbol{X} \to \mathbb{C}$ that are right-invariant under some open subgroup of $G$. An *eigenfunction* on $\boldsymbol{X}$ is an element $\varphi \in \mathcal{A}(\boldsymbol{X})$ that is a $T_\ell$-eigenfunction for each $\ell$ and that generates an irreducible representation of $G$ under the right translation action $g\varphi(x) := \rho_{\mathrm{reg}}(g)\varphi(x) := \varphi(xg)$. The *uniform measure* on $\boldsymbol{X}$, denoted simply $\int_{\boldsymbol{X}}$, is the probability

Haar coming from the $G$-action. An element $\varphi \in \mathcal{A}(\boldsymbol{X})$ is *$L^2$-normalized* if $\int_{\boldsymbol{X}} |\varphi|^2 = 1$. In that case, the *$L^2$-mass* of $\varphi$ is the probability measure $\mu_\varphi$ on $\boldsymbol{X}$ given by $\mu_\varphi(\Psi) := \int_{\boldsymbol{X}} \Psi|\varphi|^2$. Convergence of measures always refers to the weak sense, i.e., $\lim_{n\to\infty} \mu_n = \mu$ if for each fixed $\Psi \in \mathcal{A}(\boldsymbol{X})$, $\lim_{n\to\infty} \mu_n(\Psi) = \mu(\Psi)$. A sequence of measures *equidistributes* if it converges to the uniform measure.

**Definition 12.** We denote by $\mathcal{H} \subseteq \operatorname{End}(\mathcal{A}(\boldsymbol{X}))$ the ring generated by $\rho_{\text{reg}}(G)$ and the $T_\ell$, so that an eigenfunction in the sense of Definition 11 is an element of $\mathcal{A}(\boldsymbol{X})$ that generates an irreducible $\mathcal{H}$-submodule. We denote by $A(\boldsymbol{X})$ the set of irreducible $\mathcal{H}$-submodules of $\mathcal{A}(\boldsymbol{X})$, by $A_0(\boldsymbol{X}) \subseteq A(\boldsymbol{X})$ the subset consisting of those that are not one-dimensional, and by $\mathcal{A}_0(\boldsymbol{X}) \subseteq \mathcal{A}(\boldsymbol{X})$ the sum of the elements of $A_0(\boldsymbol{X})$, or equivalently, the orthogonal complement of the one-dimensional irreducible submodules.

A theorem of Eichler/Jacquet–Langlands implies that each $\pi \in A(\boldsymbol{X})$ occurs in $\mathcal{A}(\boldsymbol{X})$ with multiplicity one, so that $\mathcal{A}(\boldsymbol{X}) = \bigoplus_{\pi\in A(\boldsymbol{X})} \pi$ and $\mathcal{A}_0(\boldsymbol{X}) = \bigoplus_{\pi\in A_0(\boldsymbol{X})} \pi$. The one-dimensional elements of $A(\boldsymbol{X})$ are given by $\mathbb{C}(\chi \circ \det)$ for each character $\chi$ of the compact group $\mathbb{Q}_p^\times / \det(\Gamma)$, thus $A(\boldsymbol{X}) = \{\mathbb{C}(\chi \circ \det)\} \bigsqcup A_0(\boldsymbol{X})$.

**Definition 13.** Let $\chi_\pi : \mathbb{Q}_p^\times \to \mathbb{C}^\times$ denote the central character of $\pi$. For $\pi \in A_0(\boldsymbol{X})$, the *conductor* of $\pi$ has the form $C(\pi) = p^{c(\pi)}$, where $c(\pi)$ is the smallest nonnegative integer with the property that $\pi$ contains a nonzero vector $\varphi$ satisfying $g\varphi = \chi_\pi(d)g$ for all $g = \left(\begin{smallmatrix} * & * \\ * & d \end{smallmatrix}\right) \in K_{0..c(\pi)}$ [Casselman 1973a; Schmidt 2002].

**Definition 14.** Let $\pi \in A_0(\boldsymbol{X})$. For integers $m, m'$, a vector $\varphi \in \pi$ will be called a *newvector of support* $m..m'$ if $m' - m = c(\pi)$ and $g\varphi = \chi_\pi(d)\varphi$ for all $g = \left(\begin{smallmatrix} * & * \\ * & d \end{smallmatrix}\right) \in K_{m..m'}$. Local newvector theory [Casselman 1973a; Schmidt 2002] implies that the space of such vectors is one-dimensional, so if $\varphi$ is $L^2$-normalized, then the $L^2$-mass $\mu_\varphi$ depends only upon $\pi$ and $m..m'$, not $\varphi$. A vector $\varphi \in \pi$ will be called a *generalized newvector* if it is a newvector of support $m..m'$ for some $m, m'$. (We include the adjective "generalized" only to indicate explicitly that we are not necessarily referring to the traditional case $m..m' = 0..c(\pi)$, which will play no distinguished role here.)

**Remark 15.** The newvectors of support $m..m'$ that generate representations with unramified central character may be characterized more simply as those eigenfunctions $\varphi \in \mathcal{A}(\boldsymbol{X})$ (in the sense of Definition 11) which:

(1) are $K_{m..m'}$-invariant, or equivalently, descend to $\varphi : \boldsymbol{Y}_{m..m'} \to \mathbb{C}$, and

(2) are orthogonal to pullbacks from $\boldsymbol{Y}_{n..n'}$ whenever $n..n' \subsetneq m..m'$.

(The proof of this characterization is the same as the proof that local newvector theory [Casselman 1973a] recovers classical Atkin–Lehner theory [1970].) Under the torsion-freeness assumption (1), "orthogonal" can be taken to mean with respect to the normalized counting measure on $\boldsymbol{Y}_{m..m'}$; in general, one should take that induced by the uniform measure on $\boldsymbol{X}$. In this sense, Definition 14 is consistent with Definition 1.

**Definition 16.** We say that $\pi \in A_0(\boldsymbol{X})$ *belongs to the principal series* if the corresponding representation of $G$ does (see Section 5.3).

**Theorem 17** (equidistribution of newvectors II). *Let $\pi_j \in A_0(\mathbf{X})$ $(j = 1, 2, 3, \dots)$ be a sequence with $C(\bar{\pi}_j \times \pi_j) \to \infty$. Assume that $\pi_j$ belongs to the principal series. Let $\varphi_j \in \pi_j$ be an $L^2$-normalized generalized newvector. Then $\mu_{\varphi_j}$ equidistributes as $j \to \infty$.*

Theorem 17 specializes to Theorem 7 upon requiring the central character of $\pi_j$ to be unramified and restricting to newvectors of support $m..m' = -N..N$ for some $N$.

**Remark 18.** Unlike earlier works such as [Nelson 2011; Nelson et al. 2014; Hu 2018], we have allowed arbitrary central characters in Theorem 17. We note that the case of the argument in which the conductor of the central character is as large as possible relative to that of the representation is a bit more technically challenging than the others; see (26) and following.

**Remark 19.** Cases of Theorem 17 in which $m..m'$ is highly unbalanced, such as the most traditional case $m..m' = 0..n$ analogous to Theorem 6, are easier: they follow, sometimes with a power savings, from the triple product formula, the convexity bound for triple product $L$-functions, and nontrivial local estimates as in [Nelson et al. 2014; Hu 2018]. Cases in which $m..m'$ is balanced, such as the case $m..m' = -N..N$ illustrated in Section 1.1, do not follow from such local arguments and require the new ideas introduced here. This phenomenon is comparable to how the mass equidistribution on a hyperbolic surface $\Gamma' \backslash \mathbb{H}$ of a weight $k$ vector in a principal series $\pi \hookrightarrow L^2(\Gamma' \backslash \mathrm{SL}_2(\mathbb{R}))$ of parameter $t \to \infty$ follows from essentially local means for $t/k = o(1)$ but not for $k = 0$, or even for $k \ll t$; see [Zelditch 1992; Reznikov 2001] for some discussion along such lines. See also Remark 30 and footnote 12.

**1.3.** ***p-adic microlocal lifts.*** We turn to the key definitions that power the proof of the above results. We develop them slightly more precisely and algebraically than is strictly necessary for the consequences indicated above.

Let $k$ be a nonarchimedean local field with ring of integers $\mathfrak{o}$, maximal ideal $\mathfrak{p}$, normalized valuation $v : k \twoheadrightarrow \mathbb{Z} \cup \{+\infty\}$, and $q := \#\mathfrak{o}/\mathfrak{p}$. (The case $(k, \mathfrak{o}, \mathfrak{p}, q) = (\mathbb{Q}_p, \mathbb{Z}_p, p\mathbb{Z}_p, p)$ is relevant for the above application.)

To a generic irreducible representation $\pi$ of $\mathrm{GL}_n(k)$ one may attach a *conductor* $C(\pi) = q^{c(\pi)}$, with $c(\pi) \in \mathbb{Z}_{\geqslant 0}$; we recall this assignment in the most relevant case $n = 2$ in Section 5.3 and Section 5.5. One also defines $c(\omega)$ for each character $\omega$ of $\mathfrak{o}^\times$; it is the smallest integer $n$ for which $\omega$ has trivial restriction to $\mathfrak{o}^\times \cap 1 + \mathfrak{p}^n$.

For context, we record the local form of Definition 14:

**Definition 20** (newvectors). A vector $v$ in an irreducible generic representation $\pi$ of $\mathrm{GL}_2(k)$ is a *newvector of support $m..m'$* if $m' - m = c(\pi)$ and

$$\pi(g)v = \chi_\pi(d)v \text{ for all } g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}_2(\mathfrak{o}) \cap \begin{bmatrix} \mathfrak{o} & \mathfrak{p}^{-m} \\ \mathfrak{p}^{m'} & \mathfrak{o} \end{bmatrix}.$$

A *generalized newvector* is a newvector of some support.

Fix now for each nonnegative integer $N$ a partition $N = N_1 + N_2$ into nonnegative integers $N_1, N_2$ with the property that $N_1, N_2 \to \infty$ as $N \to \infty$. The precise choice is unimportant; one might take

$N_1 := \lfloor N/2 \rfloor$, $N_2 := \lceil N/2 \rceil$ for concreteness. Using this choice, we introduce the following class of vectors:

**Definition 21** (microlocal lifts). Let $\pi$ be a $\mathrm{GL}_2(k)$-module. A vector $v \in \pi$ shall be called a *microlocal lift* if:

- it is nonzero,
- it generates an irreducible admissible representation of $\mathrm{GL}_2(k)$, and
- there is a positive integer $N$ and characters $\omega_1, \omega_2$ of $\mathfrak{o}^\times$ so that $c(\omega_1/\omega_2) = N$ and

$$\pi(g)v = \omega_1(a)\omega_2(\det(g)/a)v \text{ for all } g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}_2(\mathfrak{o}) \cap \begin{bmatrix} \mathfrak{o} & \mathfrak{p}^{N_1} \\ \mathfrak{p}^{N_2} & \mathfrak{o} \end{bmatrix}.$$

In that case, we refer to $N$ as the *level* and $(\omega_1, \omega_2)$ as the *orientation* of $v$.

The observation that the special case $\omega_1 = 1$ of Definition 21 is similar to Definition 20 leads easily to the following characterization of microlocal lifts as twists of generalized newvectors from "extremal principal series" representations "$1 \boxplus \chi$" (see Section 6.1 for the proof):

**Lemma 22.** *An irreducible admissible representation $\pi$ of $\mathrm{GL}_2(k)$ contains a microlocal lift if and only if $\pi$ is an irreducible principal series representation $\pi \cong \chi_1 \boxplus \chi_2$ for which $N := c(\bar{\pi} \otimes \pi)/2 = c(\chi_1/\chi_2)$ is nonzero. In that case, the set of microlocal lifts is a disjoint union $\mathbb{C}^\times \varphi_+ \sqcup \mathbb{C}^\times \varphi_-$, where*

$$\mathbb{C}^\times \varphi_+ = \{\textit{microlocal lifts in } \pi \textit{ of level } N \textit{ and orientation } (\omega_1, \omega_2)\},$$
$$\mathbb{C}^\times \varphi_- = \{\textit{ microlocal lifts in } \pi \textit{ of level } N \textit{ and orientation } (\omega_2, \omega_1)\ \},$$

*with $\omega_i := \chi_i|_{\mathfrak{o}^\times}$. Explicitly, $\mathbb{C}^\times \varphi_+$ is the inverse image under the nonequivariant twisting isomorphism $\pi \to \pi \otimes \chi_1^{-1} \cong 1 \boxplus \chi_1^{-1}\chi_2$ of the set of nonzero newvectors of support $-N_1..N_2$. The set $\mathbb{C}^\times \varphi_-$ is described similarly, with the roles of $\omega_1$ and $\omega_2$ reversed.*

**Remark 23.** We briefly compare with the archimedean analogue inspiring Definition 21; a more complete exposition of this analogy seems beyond the scope of this article. Let $\pi$ be a principal series representation of $\mathrm{PGL}_2(\mathbb{R})$ of parameter $t \to \pm\infty$ with lowest weight vector $\varphi_0$ corresponding to a spherical Maass form of eigenvalue $\frac{1}{4} + t^2$ on some hyperbolic surface. The Zelditch–Wolpert construction[4] of a microlocal lift $\varphi_1$ of $\varphi_0$ is given up to normalizing factors in terms of standard raising/lowering operators $X^n$ for $n \in \mathbb{Z}$ (see [Wolpert 2001; Lindenstrauss 2001]) by $\varphi_1 := \sum_{n:|n|\leqslant t_1} X^n \varphi_0$, where $|t| = t_1 t_2$ with $t_1, t_2 \to \infty$ as $|t| \to \infty$. The choice $\varphi_2 := \sum_{n:|n|\leqslant t_1} (-1)^n X^n \varphi_0$ also works. The analogue of $(|t|, \varphi_1, \varphi_2, |.|^{it}, |.|^{-it})$ in the notation of Definition 21 and Lemma 22 is $(q^N, v_1, v_2, \chi_1, \chi_2)$ with $q := \#\mathfrak{o}/\mathfrak{p}$ and $v_1, v_2 \in \pi$ microlocal lifts of respective orientations $(\omega_1, \omega_2)$, $(\omega_2, \omega_1)$. The analogy may be obtained by comparing how $\mathrm{GL}_2(\mathfrak{o})$ acts on $v_1, v_2$ to how the Lie algebra of $\mathrm{PGL}_2(\mathbb{R})$ acts on $\varphi_1, \varphi_2$. The factorization $|t| = t_1 t_2$ is roughly analogous to the partition $N = N_1 + N_2$. It is also instructive to compare the formulas for $\varphi_1, \varphi_2$ in their induced models with those of Section 6.2.

[4] We discuss here only the "positive measure" incarnation of that construction rather than the "distributional" one.

**Remark 24.** Le-Masson [2014] and Anantharaman–Le-Masson [2015] have introduced a notion of microlocal lifts on regular graphs and used that notion to prove some analogues of the quantum ergodicity theorem. Definition 21 serves different aims in that we do not explicitly vary the graph (except perhaps in the second sense indicated in Remark 8); it would be interesting to extend it further and compare the two notions on any domain of overlap.

For the remainder of Section 1.3, take $k = \mathbb{Q}_p$, so that $\mathrm{GL}_2(k) = G$. Definition 21 applies to $\pi \in A_0(\boldsymbol{X})$.

**Theorem 25** (basic properties of microlocal lifts). *Let $N$ traverse a sequence of positive integers tending to $\infty$, and let $\varphi \in \pi \in A_0(\boldsymbol{X})$ be an $L^2$-normalized microlocal lift of level $N$ on $\boldsymbol{X}$ with $L^2$-mass $\mu_\varphi$:*

- ***Diagonal invariance**: Any weak subsequential limit of the sequence of measures $\mu_\varphi$ is $a(\mathbb{Q}_p^\times)$-invariant.*
- ***Lifting property**: Suppose temporarily that $p \neq 2$, so that $\nu(2) = 0$. Let $\varphi' \in \pi$ be an $L^2$-normalized newvector of support $-N..N$, and let $\Psi \in \mathcal{A}(\boldsymbol{X})^K$ be independent of $N$ and right-invariant by $K := \mathrm{GL}_2(\mathbb{Z}_p)$. Then*
$$\lim_{N \to \infty} (\mu_\varphi(\Psi) - \mu_{\varphi'}(\Psi)) = 0.$$
- ***Equidistribution implication**: Suppose that $\mu_\varphi$ equidistributes as $N \to \infty$. Let $\varphi' \in \pi$ be an $L^2$-normalized generalized newvector. Then $\mu_{\varphi'}$ equidistributes as $N \to \infty$.*

Theorem 25 is established in Section 7 after developing the necessary local preliminaries in Section 5 and Section 6. The proof involves uniqueness of invariant trilinear forms[5] on $\mathrm{GL}_2$ and stationary phase analysis of local Rankin–Selberg integrals. Theorem 25 is essentially local, i.e., does not exploit the arithmeticity of $\Gamma \leqslant G$, and is stated here in a global setting only for convenience; see Theorem 49 for a local analogue.

**Remark 26.** The "lifting property" of Theorem 25 has been included only for the sake of illustration; it is not strictly necessary for the logical purposes of this paper. We have assumed $p \neq 2$ in its statement because the corresponding assertion is false when $p = 2$. For general $p$ and nonspherical observables $\Psi$, there does not appear to be any simple relationship between the quantities $\mu_\varphi(\Psi)$ and $\mu_{\varphi'}(\Psi)$ except that convergence to $\int_{\boldsymbol{X}} \Psi$ of the first implies that of the second (the "equidistribution implication"). The "lifting" relationship here is thus more subtle than that in [Lindenstrauss 2006b].

**1.4. *Equidistribution of microlocal lifts.*** Our core result (from which the others are ultimately derived) is the following:

**Theorem 27** (equidistribution of microlocal lifts). *Let $N$ traverse a sequence of positive integers tending to $\infty$. Let $\varphi \in \mathcal{A}(\boldsymbol{X})$ be an $L^2$-normalized microlocal lift of level $N$ on $\boldsymbol{X}$. Then $\mu_\varphi$ equidistributes.*

[5] It should be possible to avoid this comparatively deep fact in the proof of the first part of Theorem 25, but it is required by the application to subconvexity (Theorem 29), and the calculations required by that application already suffice here.

The proof depends upon an analogue of Lindenstrauss's celebrated result [2006b]. Here and throughout this article, "entropy" refers to the Kolmogorov–Sinai entropy of a measurable dynamical system (see, e.g., [Lindenstrauss 2006a, §8]).

**Theorem 28** (measure classification). *Let $\mu$ be a probability measure on $\boldsymbol{X}$, invariant by the center of $G$, with the properties*:

(1) *$\mu$ is $a(\mathbb{Q}_p^\times)$-invariant.*

(2) *$\mu$ is $T_\ell$-recurrent for some split prime $\ell \neq p$.*

(3) *The entropy of almost every ergodic component of $\mu$ is positive for the $a(\mathbb{Q}_p^\times)$-action.*

*Then $\mu$ is the uniform measure.*

We explain in Section 2 the specialization of Theorem 28 from a result of Einsiedler–Lindenstrauss [2008, Theorem 1.5]. To deduce Theorem 27, we apply Theorem 28 with $\mu$ any weak limit of the $L^2$-masses of a sequence of $L^2$-normalized microlocal lifts of level tending to $\infty$. Since $\boldsymbol{X}$ is compact, $\mu$ is a probability measure. The invariance hypothesis follows from the diagonal invariance of Theorem 25, while the $T_\ell$-recurrence and positive entropy hypotheses are verified below in Section 3 and Section 4. The proof of our main result Theorem 27 is then complete. Theorem 27 and the equidistribution implication of Theorem 25 imply Theorem 17.

**1.5. *Estimates for L-functions.*** For definitions of the $L$-functions and local distinguishedness, see [Piatetski-Shapiro and Rallis 1987; Ichino 2008]. We record the following because it provides an unambiguous benchmark of the strength of our results.

**Theorem 29** (weakly subconvex bound). *Fix $\sigma \in A_0(\boldsymbol{X})$. Let $\pi \in A_0(\boldsymbol{X})$ traverse a sequence with $C(\bar{\pi} \times \pi) \to \infty$. Assume that $\pi$ belongs to the principal series and that $\sigma \otimes \bar{\pi} \otimes \pi$ is locally distinguished. Then*

$$\frac{L(\sigma \times \bar{\pi} \times \pi, 1/2)}{L(\mathrm{ad}\,\pi, 1)^2} = o(C(\sigma \times \bar{\pi} \times \pi)^{1/4}). \tag{5}$$

The previously best known estimate for the LHS of (5) is the general weakly subconvex estimate of Soundararajan [2010], specializing here to $L \ll C^{1/4}/(\log C)^{1-\varepsilon}$ with $L := L\big(\sigma \times \bar{\pi} \times \pi, \frac{1}{2}\big)$, $C := C(\sigma \times \bar{\pi} \times \pi)$. The bound (5) improves upon that estimate in the unlikely (but difficult to exclude) case that $L(\mathrm{ad}\,\pi, 1)$ is exceptionally small, which turns out to be the most difficult one for equidistribution problems; see [Holowinsky and Soundararajan 2010] for further discussion.

Theorem 27 implies Theorem 29 after a local calculation with the triple product formula (see Section 7); in fact, the calculation shows that the two results are equivalent.

**Remark 30.** Theorem 29 implies Theorem 17, but the converse does not hold in general; a special case of the failure of that converse was noted and discussed at length in [Nelson et al. 2014, §1]. The present work may thus be understood as clarifying that discussion: the equivalence between subconvexity and equidistribution problems in the depth aspect is restored by working not with newvectors, but instead with the $p$-adic microlocal lifts introduced here.

**1.6.** ***Further remarks.***

**Remark 31.** Theorems 17 and 27 apply only to sequences of vectors $\varphi$ that generate irreducible $\mathcal{H}$-modules. One can ask whether the conclusion holds under the (hypothetically) weaker assumption that $\varphi$ generates an irreducible $G$-module. The problem formulated this way makes sense for any finite volume quotient $\Gamma' \backslash G$, not necessarily arithmetic; an affirmative answer would represent a $p$-adic analogue of the Rudnick–Sarnak quantum unique ergodicity conjecture [1994]. In that direction, we note that the method of Brooks–Lindenstrauss [2014] should apply in our setting, allowing one to relax the hypothesis of irreducibility under the full Hecke algebra to that under a single auxiliary Hecke operator $T_\ell$ for some fixed split prime $\ell \neq p$.

An affirmative answer to the question raised above would, by the (proof of the) equidistribution implication of Theorem 25, imply that the conclusion of Conjecture 3 remains valid on possibly nonarithmetic quotients $\Gamma' \backslash G$ under the hypothesis that $\varphi_N \in L^2_{\text{new}}(\boldsymbol{Y}_{-N..N})$ traverses a sequence of unit vectors that generate principal series representations of $\mathrm{GL}_2(\mathbb{Q}_p)$. (The analogous assertion for supercuspidal representations fails because such representations may be shown to occur with large multiplicity. A similar phenomenon is responsible for the subtlety in formulating holomorphic analogues of quantum unique ergodicity; see [Luo and Sarnak 2003; Holowinsky and Soundararajan 2010].)

**Remark 32.** Our results apply to principal series representations of conductor $p^N$ with $p$ fixed and $N \to \infty$. A natural question is whether one can establish analogous results for $N$ fixed, such as $N = 100$, and $p \to \infty$. We highlight here the weaker question of whether one can establish equidistribution (in a balanced case, cf. Remark 19) as $N \to \infty$ for $p$ satisfying $p \leqslant p_0(N)$ for some $p_0(N)$ tending *effectively* to $\infty$ as $N \to \infty$. Our results and a diagonalization argument imply an ineffective analogue.

**Remark 33.** The crucial local results of this article have been formulated and proved in generality, i.e., over any nonarchimedean local field. On the other hand, we have assumed in our global results that the subgroup $\Gamma$ of $G$ was constructed from a *maximal order* in a quaternion algebra *over* $\mathbb{Q}$. We expect that our results hold more generally:

(1) The statements and proofs of all our results except Theorem 29 extend straightforwardly to the case that $\Gamma$ arises from a fixed *Eichler order* in a quaternion algebra over $\mathbb{Q}$. To extend Theorem 29 in that direction would require some local triple product estimates at the "uninteresting" primes $\ell \neq p$ which we do not pursue here.

(2) Our results should extend to Eichler orders in totally definite quaternion algebras over *totally real number fields*, but some mild care is required in formulating such extensions when the class group has nontrivial 2-torsion: as observed in a related context in [Nelson 2012], there are sequences of dihedral forms that fail to satisfy the most naive formulation of quantum unique ergodicity.

(3) We expect our results extend to automorphic forms on definite quaternion algebras having fixed nontrivial infinity type; such an extension would require a more careful study of the measure classification input in Section 2.

(4) Over function fields, analogues of our results should follow more directly and in quantitatively stronger forms from Deligne's theorem and extensions of the triple product formula to the function field setting.

We leave such extensions to the interested reader.

***Organization of this paper.*** We verify the measure-classification (Theorem 28) and its hypotheses in Section 2, Section 3, and Section 4. We review the representation theory of $\mathrm{GL}_2(k)$ in Section 5. In Section 6 and Section 7, we prove our core results, notably Theorem 25, and their applications. Some additional results of independent interest are recorded along the way.

## 2. Measure classification

The purpose of this section is to deduce Theorem 28 from the following specialization to $\mathbb{Q}_p$ of a result of Einsiedler–Lindenstrauss [2008, Theorem 1.5]:

**Theorem 34.** *Let $G = G_1 \times G_2$, where $G_1$ is a semisimple linear algebraic group over $\mathbb{Q}_p$ with $\mathbb{Q}_p$-rank 1 and $G_2$ is a characteristic zero $S$-algebraic group. Let $\Gamma' \subset G$ be a discrete subgroup. Let $A_1$ be a $\mathbb{Q}_p$-split torus of $G_1$ and let $\chi$ be a nontrivial $\mathbb{Q}_p$-character of $A_1$ that can be extended to $C_{G_1}(A_1)$. Let $M_1 = \{h \in C_{G_1}(A_1) : \chi(h) = 1\}$. Let $\nu$ be an $A_1$-invariant, $G_2$-recurrent probability measure on $\Gamma' \backslash G$ such that*:

(1) *almost every $A_1$-ergodic component of $\nu$ has positive entropy with respect to some $a \in A_1$ with $|\chi(a)| \neq 1$, and*

(2) *for $\nu$-almost every $x \in \Gamma' \backslash G$, the group $\{h \in M_1 \times G_2 : xh = x\}$ is finite.*

*Then $\nu$ is a convex combination of homogeneous measures, each of which is supported on an orbit of a subgroup $H$ which contains a finite index subgroup of a semisimple algebraic subgroup of $G_1$ of $\mathbb{Q}_p$-rank one.*

To deduce Theorem 28 from Theorem 34 requires no new ideas, but we record a complete verification for completeness.

**2.1.** ***Consequences of strong approximation.*** Recall that $R$ is a maximal order in a definite quaternion algebra $B$. (For general background on quaternion algebras we mention [Vignéras 1980; Voight 2018; Nelson 2015, §2.2].)

For a prime $p$, we shall use the notations $B_p := B \otimes_{\mathbb{Q}} \mathbb{Q}_p$, $R_p := R \otimes_{\mathbb{Z}} \mathbb{Z}_p$. A superscripted (1) denotes "norm one elements," e.g., $B_p^{(1)} := \{b \in B_p^\times : \mathrm{nr}(b) = 1\}$. Denote by $\mathbb{A}_f$ the finite adele ring of $\mathbb{Q}$ and $\hat{B} := B \otimes_{\mathbb{Q}} \mathbb{A}_f$. (Thus $B_{\mathbb{A}} := B_\infty \times \hat{B}$ with $B_{\mathbb{A}} := B \otimes_{\mathbb{Q}} \mathbb{A}$, $B_\infty := B \otimes_{\mathbb{Q}} \mathbb{R}$, and $\mathbb{A}$ the adele ring of $\mathbb{Q}$.) Regard $B^\times$, $B_p^\times$, $R_p^\times$ as subsets of $\hat{B}^\times$ in the standard way.

**Lemma 35.** *Let $U$ be a subgroup of $\hat{B}^\times$ for which*:

(i) *There is a prime $p$ that splits $B$ for which $U$ contains an open subgroup of $\hat{B}^{(1)}$ containing $B_p^{(1)}$.*

(ii) *The image* $\mathrm{nr}(U)$ *of* $U$ *under the reduced norm* $\mathrm{nr} : \hat{B}^\times \to \mathbb{A}_f^\times$ *satisfies* $\mathbb{Q}_+^\times \mathrm{nr}(U) = \mathbb{A}_f^\times$.

*Then* $B^\times U = \hat{B}^\times$.

*Proof.* It is known (e.g., by Hasse–Minkowski) that $\mathrm{nr} : B^\times \to \mathbb{Q}_+^\times$ is surjective. Let $b \in \hat{B}^\times$ be given. By (ii), there exists $\gamma \in B^\times$ and $h \in U$ for which $\gamma b h \in \hat{B}^{(1)}$. Let $p$ be as in (i). The strong approximation theorem [Kneser 1966], applied to the simply connected semisimple algebraic group $B^{(1)}$ and its noncompact factor $B_p^{(1)}$, implies that $B^{(1)} B_p^{(1)}$ is dense in $\hat{B}^{(1)}$. By (i), we may write $\gamma b h = \delta h'$ for some $\delta \in B^{(1)}$ and $h' \in U$. Therefore $b = \gamma^{-1} \delta h' h^{-1}$ belongs to $B^\times U$, as required. ☐

Let $p$ be a split prime for $B$. For any prime $\ell$, one has $\mathrm{nr}(B_\ell^\times) = \mathbb{Q}_\ell^\times$; because $R$ is a maximal order (in particular, an Eichler order), one has moreover that $\mathrm{nr}(R_\ell^\times) = \mathbb{Z}_\ell^\times$. The hypotheses of Lemma 35 thus apply to $U = B_p^\times \prod_{\ell \neq p} R_\ell^\times$: (i) is clearly satisfied, while (ii) follows from the consequence $\mathbb{Q}_+^\times \mathbb{Q}_p^\times \prod_{\ell \neq p} \mathbb{Z}_\ell^\times = \mathbb{A}_f^\times$ of strong approximation for the ideles. For similar but simpler reasons, the hypotheses apply also to $U = B_p^\times B_\ell^\times \prod_{q \neq \ell, p} R_q^\times$. Thus

$$B^\times B_p^\times \prod_{\ell \neq p} R_\ell^\times = \hat{B}^\times = B^\times B_p^\times B_\ell^\times \prod_{q \neq \ell, p} R_q^\times.$$

We have $B^\times \cap \prod_{\ell \neq p} R_\ell^\times = R[1/p]^\times$ and $B^\times \cap \prod_{q \neq \ell, p} R_q^\times = R[1/p\ell]^\times$, whence the natural identifications

$$R[1/p]^\times \backslash B_p^\times / \mathbb{Q}_p^\times = B^\times \backslash \hat{B}^\times / \mathbb{Q}_p^\times \prod_{\ell \neq p} R_\ell^\times = R[1/p\ell]^\times \backslash B_p^\times B_\ell^\times / \mathbb{Q}_p^\times R_\ell^\times. \tag{6}$$

Since $\mathbb{Z}[1/p\ell]^\times \mathbb{Q}_p^\times \mathbb{Z}_\ell^\times = \mathbb{Q}_p^\times \mathbb{Q}_\ell^\times$, the RHS of (6) is unaffected by further reduction modulo $\mathbb{Q}_\ell^\times$, i.e.,

$$R[1/p]^\times \backslash B_p^\times / \mathbb{Q}_p^\times = R[1/p\ell]^\times \backslash B_p^\times B_\ell^\times / \mathbb{Q}_p^\times \mathbb{Q}_\ell^\times R_\ell^\times. \tag{7}$$

**2.2.** ***Deduction of Theorem 28.*** Let $p$ be a split prime for $B$. Identify $B_p^\times = \mathrm{GL}_2(\mathbb{Q}_p)$ and $\boldsymbol{X} = \Gamma \backslash \mathrm{GL}_2(\mathbb{Q}_p)$ as in Section 1. Let $\mu$ be a measure on $\boldsymbol{X}$ satisfying the hypotheses of Theorem 28. It is invariant under the diagonal torus of $\mathrm{GL}_2(\mathbb{Q}_p)$, which generates the latter modulo $\mathrm{SL}_2(\mathbb{Q}_p)$, so to prove that $\mu$ is the uniform measure, we need only verify that it is $\mathrm{SL}_2(\mathbb{Q}_p)$-invariant. To that end, we apply Theorem 34: Set $G_1 := \mathrm{PGL}_2(\mathbb{Q}_p) = B_p^\times / \mathbb{Q}_p^\times$, $G_2 := \mathrm{PGL}_2(\mathbb{Q}_\ell) = B_\ell^\times / \mathbb{Q}_\ell^\times$, $G := G_1 \times G_2$. Recall that $\Gamma = R[1/p]^\times$. Take for $\Gamma'$ the image of $R[1/p\ell]^\times$ in $G$. By strong approximation in the form (7), we may identify $\Gamma \backslash \mathrm{GL}_2(\mathbb{Q}_p) / \mathbb{Q}_p^\times$ with $\Gamma' \backslash G / \mathrm{PGL}_2(\mathbb{Z}_\ell)$ and $\mu$ with a right $\mathrm{PGL}_2(\mathbb{Z}_\ell)$-invariant measure $\nu$ on $\Gamma' \backslash G$. Our task is then to verify that $\nu$ is invariant by the image of $\mathrm{SL}_2(\mathbb{Q}_p)$. Take for $A_1$ the diagonal torus in $G_1$ and for $\chi : A_1 \to \mathbb{Q}_p^\times$ the map $\chi(\mathrm{diag}(y_1, y_2)) := y_1/y_2$. We have $C_{G_1}(A_1) = A_1$. The group $M_1$ is trivial, hence each $\{h \in M_1 \times G_2 : xh = x\}$ is trivial. The hypotheses of Theorem 28 are satisfied, so $\nu$ is invariant by some finite index subgroup $H_1$ of some semisimple algebraic subgroup of $G_1$ (of $\mathbb{Q}_p$-rank one) that contains $A_1$. The smallest such $H_1$ is the image of $\mathrm{SL}_2(\mathbb{Q}_p)$, so we conclude.

## 3. Recurrence

In this section we formulate and verify the $T_\ell$-recurrence hypothesis required by Theorem 28. The argument here is as in [Lindenstrauss 2006b, §8] except that we allow general central characters; for completeness, we record a proof of the key estimate in that case. The proof is simple; a key insight of Lindenstrauss [2006b] is that the condition enunciated here is useful for the present purposes.

### 3.1. *Hecke operators.*

**3.1.1.** *Summary of facts.* For a positive integer $n$ coprime to $p$, the Hecke operator $T_n \in \operatorname{End}(\mathcal{A}(\boldsymbol{X}))$ is defined by $T_n\varphi(x) := \sum_{\alpha \in M_n/M_1} \varphi(\alpha^{-1}x)$, where $M_n := R[1/p] \cap \operatorname{nr}^{-1}(n\mathbb{Z}[1/p]^\times)$, so that $M_1 = \Gamma$. These operators commute with one another and also with $\rho_{\text{reg}}(G)$. Given a scalar element $m$, let us introduce the general abbreviation $z(m)$ for the corresponding quaternion. For $m \in \mathbb{Q}^\times$, we abbreviate $z(m) := \rho_{\text{reg}}(z(m))$. If $\ell \mid \operatorname{disc} B$, then the operator $T_\ell$ is an involution modulo the action of the center, namely $T_\ell^2 = T_{\ell^2} = z(\ell^{-1})$; otherwise, $T_\ell$ is induced by a correspondence of degree $\ell + 1$. The adjoint of $T_n$ is $T_n^* = z(n)T_n$, and one has the composition formula

$$T_m T_n = \sum_{d \in \mathbb{Z}_{\geqslant 1} : d \mid \gcd(m,n),\ \gcd(d, \operatorname{disc} B) = 1} d \cdot z(d^{-1}) T_{mn/d^2}. \tag{8}$$

**3.1.2.** *Derivations.* Since we are unaware of a convenient reference for the facts recalled above, we briefly indicate how they fall out from the adelic picture and the structure of the local Hecke algebras. (The reader is strongly encouraged to skip this section, which we have included only for completeness.) With notation as in Section 2.1, let us abbreviate $H_\ell := B_\ell^\times$, $J_\ell := R_\ell^\times$ and $H := \prod_{\ell \neq p} H_\ell$ and $J := \prod_{\ell \neq p} J_\ell$, so that $J$ is a compact open subgroup of $H$ and $G \times H = \hat{B}^\times$. By strong approximation as in Section 2.1, the map $G \ni x \mapsto (x, 1) \in G \times H$ induces a bijection $\boldsymbol{X} = \Gamma \backslash G \xrightarrow{\sim} B^\times \backslash (G \times H/J)$. In this way, we may identify each $\varphi \in \mathcal{A}(\boldsymbol{X})$ with a right-$J$-invariant function $\Phi : B^\times \backslash (G \times H) \to \mathbb{C}$, called the *lift* of $\varphi$. Equip $H$ with the Haar measure assigning volume one to $J$. Then the algebra $\mathcal{H} := C_c^\infty(J \backslash H/J)$, under convolution, acts on $\mathcal{A}(\boldsymbol{X})$ by translating the corresponding lifts. The algebra $\mathcal{H}$ decomposes as a restricted tensor product of local Hecke algebras $\mathcal{H}_\ell = C_c^\infty(J_\ell \backslash H_\ell / J_\ell)$, where again we normalize so that $J_\ell$ has volume one. These local Hecke algebras may be described as follows:

- Suppose $\ell \mid \operatorname{disc}(B)$, i.e., that $\ell$ does not split $B$, so that $B_\ell$ is a quaternion division algebra. Then $J_\ell$ is the kernel of the map $H_\ell \to \mathbb{Z}$ sending an element to the valuation of its reduced norm. This induces an isomorphism from $\mathcal{H}_\ell$ to the group algebra $\mathbb{C}[\mathbb{Z}]$. In other words, $\mathcal{H}_\ell$ has a basis given by the characteristic functions $T_{\ell^n}$ of those $x \in H_\ell$ with reduced norm of valuation $n$, and we have $T_{\ell^m} T_{\ell^n} = T_{\ell^{m+n}}$. We note that $T_{\ell^{2n}}$ is the characteristic function of $J\ell z(\ell^n) J_\ell$, where as usual $z(y) \in H_\ell$ denotes the scalar element corresponding to $y \in \mathbb{Q}_\ell^\times$.
- Suppose $\ell \nmid \operatorname{disc}(B)$, i.e., that $\ell$ splits $B$. Then $H_\ell \cong \operatorname{GL}_2(\mathbb{Q}_\ell)$ and $J_\ell \cong \operatorname{GL}_2(\mathbb{Z}_\ell)$. Let $T_{\ell^n} \in \mathcal{H}_\ell$ denote the characteristic function of $H_\ell^{(\ell^n)}$, where $H_\ell^{(\ell^n)}$ denotes the set of all $k \in R_\ell$ with reduced norm of

valuation $n$. Then $T_{\ell^m} T_{\ell^n} = \sum_{j=0}^{|m-n|} \ell^j z(\ell^j) T_{\ell^{m+n-2j}}$, with $z(y)$ as before. The Hecke algebra $\mathcal{H}_j$ is generated by the $T_{\ell^n}$ together with the characteristic functions of $J_\ell z(y) J_\ell$ taken over $y \in \mathbb{Q}_\ell^\times / \mathbb{Z}_\ell^\times$.

In summary, the algebra $\mathcal{H}$ is generated by:

- For each $m \in \prod_{\ell \neq p} \mathbb{Q}_\ell^\times / \mathbb{Z}_\ell^\times$, the characteristic function of $J z(m^{-1}) J = J z(m^{-1}) = z(m^{-1}) J$.
- For each $n \in \prod_{\ell \neq p} (\mathbb{Q}_\ell^\times \cap \mathbb{Z}_\ell) / \mathbb{Z}_\ell^\times$, the characteristic function of the double-$J$-coset

$$H^{(n)} := \left\{ k \in \prod_{\ell \neq p} R_\ell : \mathrm{nr}(\ell) \in n \prod_{\ell \neq p} \mathbb{Z}_\ell^\times \right\}.$$

Let us denote the operators on $\mathcal{A}(\boldsymbol{X})$ obtained in the first case by $\widetilde{z(m)}$ and in the second by $\widetilde{T}_n$. Since $\mathbb{Q}^\times \prod_\ell \mathbb{Z}_\ell^\times = \prod_\ell \mathbb{Q}_\ell^\times$, we may assume in the first case that $m$ is represented by an element of $\mathbb{Q}^\times$ coprime to $p$; we then verify readily, using the identity $\Phi(x, z(m^{-1})) = \Phi(z(m)x, 1)$, that $\widetilde{z(m)} = z(m)$ as defined above. In the second case, we note first that we may assume that $n$ is a positive integer coprime to $p$. Using strong approximation as in Section 2.1, we see then that the natural map $M_n / M_1 \to H^{(n)}/J$ is bijective. Decomposing $H^{(n)}$ into right $J$-cosets, it follows readily that $\widetilde{T}_n = T_n$. Thus the operators $T_n$ and $z(m)$ generate the same subalgebra of $\mathrm{End}(\mathcal{A}(\boldsymbol{X}))$ as $\mathcal{H}$ does. The relations stated in Section 3.1.1 follow from the corresponding local relations given above.

**3.2.** ***Spherical averaging operators.*** Let $n$ be a positive integer coprime to $p$. The operator $T_n$ on $\mathcal{A}(\boldsymbol{X})$ is induced by the *correspondence* on $\boldsymbol{X}$, denoted also by $T_n$, given for $x \in \boldsymbol{X}$ by the multiset (i.e., formal sum) $T_n(x) := \sum_{s \in M_n/\Gamma} s^{-1} x$. Thus $T_n \varphi(x) = \sum_{y \in T_n(x)} \varphi(y)$. Denote by $M_n^{\mathrm{prim}}$ the set of all *primitive* elements of $M_n$, i.e., those that are not divisible inside $R[1/p]$ by any divisor $d > 1$ of $n$. Then $M_n^{\mathrm{prim}}$ is right-invariant by $\Gamma$, and one has $M_n = \bigsqcup_{d^2 \mid n} z(d) M_{n/d^2}^{\mathrm{prim}}$. Denote by $S_n$ the "Hecke sphere" correspondence $S_n(x) := \sum_{s \in M_n^{\mathrm{prim}}/\Gamma} s^{-1} x$; it likewise induces an operator $S_n$ on $\mathcal{A}(\boldsymbol{X})$ given by $S_n \varphi(x) := \sum_{s \in M_n^{\mathrm{prim}}/\Gamma} \varphi(s^{-1} x) = \sum_{y \in S_n(x)} \varphi(y)$, and one has

$$T_n(x) = \sum_{d^2 \mid n} z(d^{-1}) S_{n/d^2}(x). \tag{9}$$

**3.3.** ***Recurrence.*** Let $\ell \neq p$ be a *split prime*, that is to say, a prime that splits the quaternion algebra underlying the construction of $\Gamma$, so that the Hecke operator $T_\ell$ has degree $\ell + 1$.

**Definition 36.** Let $Z$ denote the center of $G$. A finite $Z$-invariant measure $\mu$ on $\boldsymbol{X}$ is called *$T_\ell$-recurrent* if for each Borel subset $E \subseteq \boldsymbol{X}$ and $\mu$-almost every $x \in E$, there exist infinitely many positive integers $n$ for which $S_{\ell^n}(x) \cap E \neq \varnothing$.

**Theorem 37** (Hecke recurrence). *Let $\mu$ be any subsequential limit of a sequence of $L^2$-masses $\mu_\varphi$ of $L^2$-normalized automorphic forms $\varphi \in \pi \in A_0(\boldsymbol{X})$. Then $\mu$ is $T_\ell$-recurrent.*[6]

The proof of Theorem 37 reduces via measure-theoretic considerations as in [Lindenstrauss 2006b; Brooks and Lindenstrauss 2014] to that of the following:

[6]It suffices to assume only that $\varphi$ is a $T_\ell$-eigenfunction.

**Lemma 38.** *There exists $c_0 > 0$ and $n_0 \geqslant 1$ so that for each split prime $\ell$ and $\varphi \in \pi \in A_0(\boldsymbol{X})$ and $x \in \boldsymbol{X}$, one has $\sum_{k \leqslant n} \sum_{y \in S_{\ell^k}(x)} |\varphi(y)|^2 \geqslant c_0 n |\varphi(x)|^2$ for all natural numbers $n \geqslant n_0$.*

*Proof.* By a theorem of Eichler, Shimura and Igusa, $\pi$ is tempered,[7] hence there exist $\alpha, \beta \in \mathbb{C}^{(1)}$ (the Satake parameters) so that $\lambda_\pi(\ell) = \alpha + \beta$; one then has more generally for $n \in \mathbb{Z}_{\geqslant 1}$ that

$$\lambda_\pi(\ell^n) = \frac{\alpha^{n+1} - \beta^{n+1}}{\alpha - \beta}. \tag{10}$$

By (9), one has $T_{\ell^n} = \sum_{k \leqslant n : k \equiv n(2)} z(\ell^{(k-n)/2}) S_{\ell^k}$. Conversely, $S_{\ell^k} = T_{\ell^k} - 1_{k \geqslant 2} z(\ell^{-1}) T_{\ell^{k-2}}$. Since $\pi$ has a unitary central character, there is $\theta \in \mathbb{C}^{(1)}$ so that $z(\ell^{-1})\varphi = \theta\varphi$ for all $\varphi \in \pi$. Thus, denoting by $\ell^{k/2}\sigma_k \in \mathbb{C}$ the scalar by which $S_{\ell^k}$ acts on $\pi$, one obtains $\sigma_k = \lambda(\ell^k) - 1_{k \geqslant 2} \theta \ell^{-1} \lambda(\ell^{k-2})$, which expands for $k \geqslant 2$ to

$$\sigma_k = \frac{\gamma_1 \alpha^k - \gamma_2 \beta^k}{\alpha - \beta}, \tag{11}$$

with $\gamma_1 := \alpha - \theta \ell^{-1} \alpha^{-1}$, $\gamma_2 := \beta - \theta \ell^{-1} \beta^{-1}$. Note that $|\gamma_1|, |\gamma_2| \geqslant \frac{1}{2}$.

We turn to the main argument. For $m, k \in \mathbb{Z}_{\geqslant 0}$, Cauchy–Schwarz gives

$$\ell^m |\lambda_\pi(\ell^m)\varphi(x)|^2 = |T_{\ell^m}\varphi(x)|^2 \leqslant (1 + \ell^{-1})\ell^m \sum_{y \in T_{\ell^m}(x)} |\varphi(y)|^2,$$
$$\ell^k |\sigma_k \varphi(x)|^2 = |S_{\ell^k}\varphi(x)|^2 \leqslant (1 + \ell^{-1})\ell^k \sum_{y \in S_{\ell^k}(x)} |\varphi(y)|^2,$$

whence by (9) that $\sum_{k \leqslant n} \sum_{y \in S_{\ell^k}(x)} |\varphi(y)|^2 \gg |\varphi(x)|^2 c_{\pi,\ell}(n)$ with

$$c_{\pi,\ell}(n) := \sum_{k \leqslant n} |\sigma_k|^2 + \max_{m \leqslant n} |\lambda_\pi(\ell^m)|^2. \tag{12}$$

Our task thereby reduces to verifying that $c_{\pi,\ell}(n) \gg n$, uniformly in $\pi$ and (unimportantly) $\ell$. Suppose this estimate fails. Then there is a sequence of integers $j \to \infty$ and tuples $(\pi, n, \ell) = (\pi_j, n_j, \ell_j)$ as above, depending upon $j$, so that $n \to \infty$ as $j \to \infty$ and $c_{\pi,\ell}(n) = o(n)$. Here asymptotic notation refers to the $j \to \infty$ limit, and for quantities $A, B = A_j, B_j$ depending (implicitly) upon $j$, we write $A \ll B$ for $\limsup_{j \to \infty} |A_j / B_j| < \infty$ and $A \lll B$ or $A = o(B)$ for $\limsup_{j \to \infty} |A_j / B_j| = 0$; the notations $A \gg B$ and $A \ggg B$ are defined symmetrically. We shall derive from this supposition a contradiction. By passing to subsequences, we may consider separately cases in which the Satake parameters $\alpha, \beta$ of $\pi$, as defined above, satisfy:

(i) $|\alpha - \beta| \ggg 1/n$, or

(ii) $|\alpha - \beta| \ll 1/n$.[8]

[7] As in the references, the nontempered case may be treated more simply.

[8] The standard argument considers cases for which $|\alpha - \beta| \gg 1/n$ and $|\alpha - \beta| \lll 1/n$. We have found the present division slightly more efficient.

In case (i), we have $|1-\alpha\bar{\beta}|^{-1} \lll n$, and so upon expanding the square and summing the geometric series,

$$c_{\pi,\ell}(n) \geqslant \sum_{k\leqslant n} |\sigma_k|^2 = \frac{|\gamma_1|^2 n + |\gamma_2|^2 n + o(n)}{|\alpha-\beta|^2} \geqslant \frac{n/3}{|\alpha-\beta|^2} \gg n.$$

In case (ii), one has $|\alpha-\beta|^{-1}/10 \gg n$, so the largest positive integer $m \leqslant n$ for which $m|\alpha-\beta| < \frac{1}{10}$ satisfies $m \gg n$, and (10) gives $c_{\pi,\ell}(n) \geqslant |\lambda_\pi(\ell^m)|^2 \gg m^2 \gg n^2 \geqslant n$. In either case, we derive the required contradiction. □

## 4. Positive entropy

In this section we verify the entropy hypothesis required by Theorem 28. The basic ideas here are due to Bourgain–Lindenstrauss [2003] following earlier work of Rudnick–Sarnak [1994] and Lindenstrauss [2001] and followed by later developments of Silberman–Venkatesh [≥ 2018] and Brooks–Lindenstrauss [2014]. Those works dealt with archimedean aspects; the present $p$-adic adaptation is obtained by replacing the role played by the discreteness of $\mathbb{Z}$ in $\mathbb{R}$ with that of $\mathbb{Z}[1/p]$ in $\mathbb{R}\times\mathbb{Q}_p$. We also give a new formulation of the basic line of attack (Lemma 41) emphasizing convolution over covering arguments (compare with [Silberman and Venkatesh ≥ 2018, Lemma 3.4]), which may be of use in other contexts.

Call $\varepsilon > 0$ *admissible* if it belongs to the image of $|.| : \mathbb{Q}_p^\times \to \mathbb{R}_+^\times$. For a compact open subgroup $C$ of $\mathbb{Q}_p^\times$ and admissible $\varepsilon > 0$ set

$$B(U,\varepsilon) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in G : a, d \in U, |b|, |c| \leqslant \varepsilon \right\}.$$

We refer to [Lindenstrauss 2006a, §8] for definitions and basic facts concerning (Kolmogorov–Sinai) entropy. As in [Lindenstrauss 2006a, §8; Bourgain and Lindenstrauss 2003; Silberman and Venkatesh ≥ 2018, Theorem 6.4], the following criterion suffices:

**Theorem 39** (positive entropy on almost every ergodic component). *For each compact subset $\Omega$ of $G$, there exists $U$ as above and $C, c > 0$ so that for all admissible $\varepsilon \in (0,1)$, all $L^2$-normalized $\varphi \in \pi \in A_0(\boldsymbol{X})$, and all $x \in \Omega$, one has $\mu_\varphi(xB(U,\varepsilon)) \leqslant C\varepsilon^c$.*

Let us henceforth fix $\Omega$ as in Theorem 39. We then take for $U$ any open subgroup of $\mathfrak{o}^\times$ with the property that for small enough $\varepsilon$, one has

$$xB(U,\varepsilon)x^{-1} \subseteq K \text{ for all } x \in \Omega, \tag{13}$$

$$gB(U,\varepsilon)g^{-1} \cap \Gamma = \{1\} \text{ for all } g \in G. \tag{14}$$

(Let us recall why it is possible to do this. Since $K$ is open, we may find for each $x \in \Omega$ a pair $(U,\varepsilon)$ so that (13) holds. Since $\Omega$ is compact, we may find one pair that works for every $x$. Similarly, since $\Gamma$ is discrete in $G$, we may find for each $g \in G$ a pair $(U,\varepsilon)$ so that (14) holds. The validity of (14) depends only upon the class of $g$ in the quotient $\Gamma\backslash G$, which is compact, so we may again find one pair that works every $g$.)

We now state two independent lemmas, prove Theorem 39 assuming them, and then prove the lemmas.

**Lemma 40** (bounds for Hecke returns). *For all small enough admissible $\varepsilon \in (0, 1)$, all $n \in \mathbb{Z}_{\geqslant 1}$ coprime to $p$ and satisfying $n < \sqrt{1/2}\varepsilon^{-1}$, all $m \in \mathbb{Q}^\times$ with numerator and denominator coprime to $p$, and all $x \in \Omega$, the set $S := M_n \cap z(m)xB(U, \varepsilon)x^{-1}$ has cardinality $\#S \leqslant 6\prod_{p^k \| n}(k+1)$. In particular, $\#S \leqslant 2^{13}$ if $n$ has at most 10 prime divisors counted with multiplicity.*

**Lemma 41** (geometric amplification). *Let $(c_\ell)_{\ell \in \mathbb{Z}_{\geqslant 1}}$ be a finitely supported sequence of scalars. Set $T := \sum_\ell c_\ell T_\ell/\sqrt{\ell}$ and $T^a := \sum_\ell |c_\ell| T_\ell^*/\sqrt{\ell}$. Let $\varphi \in \mathcal{A}(\boldsymbol{X})$, $\psi, \nu \in C_c^\infty(G)$. Define $\Psi \in \mathcal{A}(\boldsymbol{X})$ by $\Psi(g) := \sum_{\gamma \in \Gamma} |\psi|(\gamma g)$ and $\psi * \nu \in C_c^\infty(G)$ by $\psi * \nu(x) := \int_{y \in G} \psi(xy)\nu(y)$. Then*

$$\|T\varphi(\psi * \nu)\|_{L^2(G)} \leqslant \|\varphi\|_{L^2(\boldsymbol{X})} \|T^a \Psi\|_{L^2(\boldsymbol{X})} \|\nu\|_{L^2(G)}.$$

*Proof of Theorem 39.* We have $T\varphi = \lambda\varphi$ with $\lambda := \sum c_\ell \lambda_\pi(\ell)$, where $T_\ell \varphi = \sqrt{\ell}\lambda_\pi(\ell)\varphi$. Abbreviate $J := B(U, \varepsilon)$; it is a group. Let $x \in \Omega$. Take $\psi := 1_{xB(U,\varepsilon)} \geqslant 0$ and $\nu := e_J := \mathrm{vol}(J)^{-1} 1_J$. Then $1_{xB(U,\varepsilon)} = |\psi * \nu|^2$. By (14), we have $\mu_{T\varphi}(|\psi * \nu|^2) = \|T\varphi(\psi * \nu)\|_{L^2(G)}^2$, and so by Lemma 41, $\mu_\varphi(xB(U, \varepsilon)) \leqslant |\lambda|^{-1} \|T^a \Psi\|_{L^2(\boldsymbol{X})} \|\nu\|_{L^2(G)}$. The square $\|T^a\Psi\|_{L^2(\boldsymbol{X})}^2$ is a linear combination of terms $\langle T_\ell^*\Psi, T_{\ell'}^*\Psi\rangle = \langle T_{\ell'} T_\ell^* \Psi, \Psi\rangle$ to which we apply the Hecke multiplicativity (8) and the unfolding: for $m, n \in \mathbb{Z}_{\geqslant 1}$,

$$\langle z(m) T_n^* \Psi, \Psi\rangle \|\nu\|_{L^2(G)}^2 = \int_{g \in G} \sum_{s \in M_n} \psi(z(m)sg)\psi(g)\, \mathrm{vol}(J)^{-1} = \#M_n \cap z(m^{-1})xJx^{-1}. \tag{15}$$

By Lemma 40, we thereby obtain

$$\mu_\varphi(xB(U, \varepsilon))^2 \leqslant 2^{13} |\lambda|^{-2} \sum_{\ell, \ell'} |c_\ell c_{\ell'}| \sum_{d \mid (\ell, \ell')} d/\sqrt{\ell\ell'}$$

provided that $c_\ell$ is supported on integers $\ell \leqslant 2^{-1/4}\varepsilon^{-1/2}$ having at most 5 prime factors counted with multiplicity. A standard choice of $c_\ell$ completes the proof. For completeness, we record a variant of the choice from [Venkatesh 2010, §4.1]: Set $L := (1/\varepsilon)^{0.1}$. Denote by $\mathcal{L}$ the set consisting of all $\ell = q$ or $\ell = q^2$ taken over primes $q \in [L, 2L]$; each such $q$ splits $B$ provided $\varepsilon$ is small enough. Set $c_\ell := 0$ unless $\ell \in \mathcal{L}$, in which case $c_\ell := L^{-1}\log(L)\,\mathrm{sgn}(\lambda_\pi(\ell))^{-1}$. We have $\sum_\ell |c_\ell| \asymp 1$ and $|c_\ell| \leqslant L^{-1}\log(L)$, while Iwaniec's trick $|\lambda_\pi(q)|^2 + |\lambda_\pi(q^2)| \geqslant 1$, a consequence of (8), implies $\lambda \asymp 1$. With trivial estimation we obtain $\mu_\varphi(xB(U, \varepsilon)) \ll L^{-1/2}(\log L)^{O(1)} \ll \varepsilon^{0.01}$, as required. □

*Proof of Lemma 40.* Observe first, thanks to (13) and $n\mathbb{Z}[1/p]^\times \cap (\mathbb{Q}_+ \times \mathbb{Z}_p) = \{n\}$ and $z(m) \in K$, that $S \subseteq M_n \cap K = R(n) := \{\alpha \in R : \mathrm{nr}(\alpha) = n\}$. Given $s, t \in S$, their commutator $u := sts^{-1}t^{-1}$ thus satisfies $\mathrm{nr}(u) = 1$ and $n^2 u = sts^\iota t^\iota \in R$, hence $\mathrm{tr}(u) \in n^{-2}\mathbb{Z}$. Since $S$ is conjugate to a subset of the preimage in $M_2(\mathfrak{o})$ of the upper-triangular Borel in $M_2(\mathfrak{o}/\mathfrak{q})$ with $\mathfrak{q} := \{x \in \mathfrak{o} : |x| \leqslant \varepsilon^2\}$, and the commutator of that preimage is contained in the preimage of the unipotent, one has $|\mathrm{tr}(u) - 2|_p \leqslant \varepsilon^2$. Since $B$ is definite, $|\mathrm{tr}(u)|_\infty \leqslant 2|\mathrm{nr}(u)|_\infty^{1/2} = 2$. The integer $a := n^2\,\mathrm{tr}(u) - 2n^2$ thus satisfies $|a|_\infty |a|_p \leqslant 2n^2\varepsilon^2 < 1$ and so must be zero, i.e., $\mathrm{tr}(u) = 2$; since $B$ is nonsplit, $u = 1$. In summary, any two elements of $S$ commute.

Since $B$ is nonsplit and definite, $S$ is contained in the set $\mathcal{O}(n)$ of norm $n$ elements in some imaginary quadratic order $\mathcal{O} \subset R$. Thus $\#S \leqslant \#\mathcal{O}(n) \leqslant \#\mathcal{O}^\times \cdot \#\{I \subseteq \mathcal{O} : \mathrm{nr}(I) = n\} \leqslant 6 \prod_{p^k || n}(k+1)$. □

*Proof of Lemma 41.* Write $M := R[1/p]$. We may express the operator $T$ by the formula $T\varphi(x) = \sum_{s \in M/\Gamma} h_s \varphi(s^{-1}x)$ for some finitely supported coefficients $h_s$; then $T^a\Psi(x) = \sum_{s \in M/\Gamma} |h_s| \Psi(sx)$. Abbreviate $I := \|T\varphi(\psi * \nu)\|_{L^2(G)}$. By the triangle inequality and a change of variables $x \mapsto sx$, we have

$$I \leqslant \sum_{s \in M/\Gamma} |h_s| \left( \int_{x \in G} |\varphi|^2(x) |\psi * \nu(sx)|^2 \right)^{1/2}.$$

By a change of variables, $\psi * \nu(sx) = \int_{y \in G} \psi(sy) \nu_y^*(x)$ with $\nu_y^*(x) := \nu(x^{-1}y)$. By the triangle inequality, $I \leqslant \int_{y \in G} \sum_{s \in M/\Gamma} |h_s| |\psi(sy)| \|\varphi \nu_y^*\|_{L^2(G)}$. We unfold $\int_{y \in G} \sum_{s \in M/\Gamma} = \int_{y \in X} \sum_{s \in \Gamma \backslash M} \sum_{\gamma \in \Gamma}$, giving $I \leqslant \int_{y \in X} T^a \Psi(y) \|\varphi \nu_y^*\|_{L^2(G)}$. We conclude via Cauchy–Schwartz and the identity $\int_{y \in X} \|\varphi \nu_y^*\|^2_{L^2(G)} = \|\nu\|^2_{L^2(G)} \|\varphi\|^2_{L^2(X)}$. □

## 5. Representation-theoretic preliminaries

**5.1.** ***Generalities.*** Let $k$ be a nonarchimedean local field with maximal order $\mathfrak{o}$, maximal ideal $\mathfrak{p}$, normalized valuation $\nu : k \to \mathbb{Z} \cup \{+\infty\}$, and $q := \#\mathfrak{o}/\mathfrak{p}$. Fix Haar measures $dx, d^\times y$ on $k, k^\times$ assigning volume one to maximal compact subgroups. Fix a nontrivial unramified additive character $\psi : k \to \mathbb{C}^{(1)}$. Set $G := \mathrm{GL}_2(k)$.

**5.2.** ***Some notation and terminology.*** For $x \in k$ and $y_1, y_2 \in k^\times$, set

$$n(x) := \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}, \qquad n'(x) := \begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix},$$
$$\mathrm{diag}(y_1, y_2) := \begin{pmatrix} y_1 & 0 \\ 0 & y_2 \end{pmatrix}, \qquad w := \begin{pmatrix} & -1 \\ 1 & \end{pmatrix},$$

and $a(y) := \mathrm{diag}(y, 1)$, $z(y) := \mathrm{diag}(y, y)$. Say that a vector $v$ in some $\mathrm{GL}_2(k)$-module $\pi$ is *supported on* $m..m'$, for integers $m, m'$ with $m \leqslant m'$, if $v$ is invariant by $n(\mathfrak{p}^{-m})$ and $n'(\mathfrak{p}^{m'})$, and that $v$ has *orientation* $(\omega_1, \omega_2)$, for characters $\omega_1, \omega_2$ of $\mathfrak{o}^\times$, if $\pi(\mathrm{diag}(y_1, y_2))v = \omega_1(y_1)\omega_2(y_2)v$ for all $y_1, y_2 \in \mathfrak{o}^\times$.

**5.3.** ***Principal series representations.*** For characters $\chi_1, \chi_2 : k^\times \to \mathbb{C}^\times$, denote by $\pi = \chi_1 \boxplus \chi_2$ the *principal series representation* of $G$ realized in its *induced model* as a space of smooth functions $v : G \to \mathbb{C}$ satisfying $v(n(x)\,\mathrm{diag}(y_1, y_2)g) = |y_1/y_2|^{1/2} \chi_1(y_1)\chi_2(y_2)v(g)$ for all $x \in k$ and $y_1, y_2 \in k^\times$ and $g \in G$. A sufficient condition for $\pi$ to be irreducible is that $c(\chi_1/\chi_2) \neq 0$ (see, e.g., [Schmidt 2002]). If $\chi_1, \chi_2$ are unitary, then $\pi$ is unitary; an invariant norm is given by $\|v\|^2 := \int_{x \in k} |v(n'(x))|^2 \, dx$ (see, e.g., [Knapp 1986, (7.1)]). The log-conductor is $c(\pi) = c(\chi_1) + c(\chi_2)$ and the central character is $\chi_\pi = \chi_1\chi_2$ (see, e.g., [Schmidt 2002]).

The following "line model" parametrization of $\pi$ shall be convenient: for suitable $f \in C^\infty(k)$, define $v_f \in \pi$ by

$$v_f(g) := f(c/d)|\det(g)/d^2|^{1/2}\chi_1(\det(g)/d)\chi_2(d), \quad g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \tag{16}$$

If $\chi_1, \chi_2$ are unitary, then $\|v_f\|^2 = \int_k |f|^2$.

**5.4.** ***Generic representations.*** Recall that an irreducible representation $\sigma$ of $G$ is *generic* if it is isomorphic to an irreducible subspace $\mathcal{W}(\sigma, \psi)$ of the space of smooth functions $W : G \to \mathbb{C}$ satisfying $W(n(x)g) = \psi(x)W(g)$ for all $x, g \in k, G$; in that case, $\mathcal{W}(\sigma, \psi)$ is called the *Whittaker model* of $\sigma$. It is known that every nongeneric irreducible representation of $G$ is one-dimensional (see, e.g., [Schmidt 2002]).

For each $W \in \mathcal{W}(\sigma, \psi)$, denote also by $W$ the function $W : k^\times \to \mathbb{C}$ defined by $W(y) := W(a(y))$. The space $\mathcal{K}(\sigma, \psi)$ of functions $W : k^\times \to \mathbb{C}$ arising in this way from some $W \in \mathcal{W}(\sigma, \psi)$ is called the *Kirillov model* of $\sigma$. It is known that the natural map $\mathcal{W}(\sigma, \psi) \to \mathcal{K}(\sigma, \psi)$ is an isomorphism and that $\mathcal{K}(\sigma, \psi) \supseteq C_c^\infty(k^\times)$ (see, e.g., [Schmidt 2002]).

An irreducible principal series representation $\pi = \chi_1 \boxplus \chi_2$ is generic (see, e.g., [Schmidt 2002]); the standard intertwining map from $\pi$ to its $\psi$-Whittaker model $\mathcal{W}(\pi, \psi)$, denoted $\pi \ni v \mapsto W_v : \mathrm{GL}_2(k) \to \mathbb{C}$, is given by $W_v(g) := \int_{x \in k} v(wn(x)g)\psi(-x)\,dx$. In general, this integral fails to converge absolutely and must instead be interpreted via analytic continuation, regularization, or as a limit of integrals taken over the compact subgroups $\mathfrak{p}^{-n}$ of $k$ as $n \to \infty$ (see, e.g., [Bump 1997, p. 485]); for the sake of presentation, we ignore such technicalities in what follows.

**5.5.** ***Newvector theory.*** Recall Definition 20. Recall also from Section 1.3 that we have fixed decompositions $N = N_1 + N_2$ of every nonnegative integer $N$, with $N_1, N_2 \to \infty$ as $N \to \infty$.

**Theorem 42** (basic newvector theory). *Let $\pi$ be a generic irreducible representation of* $\mathrm{GL}_2(k)$ *and let $m \leqslant m'$ be integers. Then the space of vectors in $\pi$ supported on $m..m'$ and with orientation* $(1, \chi_\pi|_{\mathfrak{o}^\times})$ *has dimension* $\max(0, 1 + |m - m'| - c(\pi))$.

*In particular, let $\pi$ be any irreducible representation of* $\mathrm{GL}_2(k)$ *with ramified central character $\chi_\pi$. Denote by $V$ the space of vectors in $\pi$ supported on $-N_1..N_2$ with orientation* $(1, \chi_\pi|_{\mathfrak{o}^\times})$. *Then $V = 0$ unless $\pi$ is generic, in which case* $\dim V = \max(0, 1 + N - c(\pi))$.

*Proof.* For the first assertion, see [Casselman 1973a]. The generic case of the second assertion follows from the first assertion, so suppose $\pi$ is one-dimensional. Write $\pi = \chi \circ \det$ for some $\chi : k^\times \to \mathbb{C}^\times$. Since $\chi_\pi$ is ramified, the characters $(1, \chi_\pi|_{\mathfrak{o}^\times})$ and $(\chi|_{\mathfrak{o}^\times}, \chi|_{\mathfrak{o}^\times})$ of $\mathfrak{o}^\times \times \mathfrak{o}^\times$ are distinct, and so $V = 0$. □

**Lemma 43.** *Let $\pi$ be an irreducible generic representation of* $\mathrm{GL}_2(k)$ *with ramified central character $\chi_\pi$. Then $c(\chi_\pi) \leqslant c(\pi)$ with equality precisely when $\pi$ is isomorphic to an irreducible principal series representation $\chi_1 \boxplus \chi_2$ for which at least one of the inducing characters $\chi_1, \chi_2$ is unramified.*

*Proof.* This is well-known; see [Templier 2014, Lemma 3.1; Casselman 1973b, Proof of Proposition 2]. □

**Lemma 44.** *Let $\pi = \chi_1 \boxplus \chi_2$ be an irreducible principal series representation of $G$. Let $v \in \pi$ be a newvector of some support $m..m'$:*

(1) *If $\chi_1$ is ramified and $\chi_2$ is ramified, then $v = v_f$ as in* (16) *for $f$ a character multiple of the characteristic function of an $\mathfrak{o}^\times$-coset, thus $f = c\chi 1_{\varpi^n \mathfrak{o}^\times}$ for some $c \in \mathbb{C}$, $\chi : k^\times \to \mathbb{C}^\times$ and $n \in \mathbb{Z}$.*

(2) *If $\chi_1$ is unramified and $\chi_2$ is ramified, then $v = v_f$ for $f = c 1_{\mathfrak{a}}$ for some scalar $c$ and fractional $\mathfrak{o}$-ideal $\mathfrak{a} \subset k$.*

*Proof.* Both assertions are well-known in the special case $m = 0$ (see [Schmidt 2002]) and follow inductively in general using that $a(\varpi)$ bijectively maps newvectors of support $m..m'$ to those of support $m-1..m'-1$. □

**5.6.** ***Local Rankin–Selberg integrals.*** Let $\pi$ be an irreducible unitary principal series representation of $G := \mathrm{GL}_2(k)$ and $\sigma$ an irreducible generic unitary representation of $\mathrm{PGL}_2(k)$. We have the following special case of a theorem of D. Prasad:

**Theorem 45** [Prasad 1990]. *The space $\mathrm{Hom}_G(\sigma \otimes \bar{\pi} \otimes \pi, \mathbb{C})$, consisting of trilinear functionals $\ell : \sigma \otimes \bar{\pi} \otimes \pi \to \mathbb{C}$ satisfying the diagonal invariance $\ell(\sigma(g)v_1, \bar{\pi}(g)v_2, \pi(g)v_3) = \ell(v_1, v_2, v_3)$ for all $g \in G$ and all vectors, is one-dimensional.*

We may fix a nonzero element $\ell_{\mathrm{RS}} \in \mathrm{Hom}_G(\sigma \otimes \bar{\pi} \otimes \pi, \mathbb{C})$ as follows: Denote by $Z$ the center of $G$ and $U := \{n(x) : x \in k\}$. Equip the right $G$-space $ZU\backslash G$ with the Haar measure for which

$$\int_{g \in ZU\backslash G} \phi(g) = \int_{y \in k^\times} \int_{x \in k} \phi(a(y)n'(x)) \frac{d^\times y}{|y|} \, dx \tag{17}$$

for $\phi \in C_c(ZU\backslash G)$ (see [Michel and Venkatesh 2010, §3.1.5]). Realize $\pi$ in its induced model. For $W_1 \in \mathcal{W}(\sigma, \psi)$, $W_2 \in \mathcal{W}(\pi, \psi)$ and $v_3 \in \pi$, set $\ell_{\mathrm{RS}}(W_1, \overline{W}_2, v_3) := \int_{ZU\backslash G} W_1 \overline{W}_2 v_3$ (see [Michel and Venkatesh 2010, §3.4.1]). The definition applies in particular when $W_2$ is the image $W_v$ of some $v \in \pi$ under the intertwiner from Section 5.4.

The trick encapsulated by the following lemma (a careful application of "nonarchimedean integration by parts") shall be exploited repeatedly in Section 6.3:

**Lemma 46** (application of diagonal invariance). *Let $f \in C_c^\infty(k)$. Let $U_1$ be an open subgroup of $\mathfrak{o}^\times$ for which $\bar{f} \otimes f$ is $U_1$-*invariant *in the sense that $\bar{f}(ux) f(uy) = \bar{f}(x) f(y)$ for all $u, x, y \in U_1, k, k$. Let $W_1 \in \mathcal{W}(\sigma, \psi)$. Then*

$$\ell_{\mathrm{RS}}(W_1, \overline{W}_{v_f}, v_f) = \int_{x \in k,\, y \in k^\times,\, t \in k} f(x) \bar{f}\Big(x + \frac{y}{t}\Big) F(x, y, t; W_1, U_1) \frac{dt}{|t|} \, dx \, d^\times y,$$

*where $F(x, y, t; W_1, U_1) := \mathbb{E}_{u \in U_1} W_1(a(y)n'(x/u)) \chi_1 \chi_2^{-1}(ut) \psi(ut)$ with $\mathbb{E}_{u \in U_1}$ denoting an integral with respect to the probability Haar.*

*Proof.* Set $g := a(y)n'(x) = \left(\begin{smallmatrix} y & \\ x & 1 \end{smallmatrix}\right)$. For $t \in k$ one has $wn(t)g = \left(\begin{smallmatrix} -x & -1 \\ y+tx & t \end{smallmatrix}\right)$, hence

$$\begin{aligned} v_f(g) &= f(x)|y|^{1/2}\chi_1(y), \\ \bar{v}_f(wn(t)g) &= \bar{f}((y+tx)/t)|y/t^2|^{1/2}\chi_1^{-1}(y/t)\chi_2^{-1}(t), \\ v_f(g)\overline{W}_{v_f}(g) &= \int_{t\in k} v_f(g)\overline{v_f(wn(t)g)\psi(-t)}\,dt \\ &= |y| f(x) \int_{t\in k} \bar{f}\Big(x + \frac{y}{t}\Big)\chi_1\chi_2^{-1}(t)\psi(t)\,\frac{dt}{|t|}. \end{aligned}$$

Integrating against $W_1(a(y)n'(x))|y|^{-1}\,dx\,d^\times y$ gives that $\ell_{\mathrm{RS}}(W_1, \overline{W}_{v_f}, v_f)$ equals

$$\int_{x\in k,\ y\in k^\times,\ t\in k} f(x)\bar{f}\Big(x+\frac{y}{t}\Big)W_1(a(y)n'(x))\chi_1\chi_2^{-1}(t)\psi(t)\,\frac{dt}{|t|}\,dx\,d^\times y.$$

To obtain the claimed formula, we apply for $u \in U_1$ the substitutions $t \mapsto ut$, $x \mapsto x/u$, invoke the assumed $U_1$-invariance of $\bar{f} \otimes f$, and average over $u$. □

**5.7. *Gauss sums.*** We shall repeatedly use the following without explicit mention:

**Lemma 47.** *Let $U_1 \leqslant \mathfrak{o}^\times$ be an open subgroup and $\omega$ a character of $\mathfrak{o}^\times$. For $t \in k^\times$, set $H(t) := H(t, \omega, U_1) := \mathbb{E}_{u\in U_1}\omega(ut)\psi(ut)$, where $\mathbb{E}$ denotes integration with respect to the probability Haar.*

(1) *For fixed $U_1$, one has $H(t) = 0$ unless $-\nu(t) = c(\omega) + O(1)$, in which case $H(t) \ll C(\omega)^{-1/2}$, with implied constants depending at most upon $U_1$.*

(2) *Suppose $U_1 = \mathfrak{o}^\times$ and $c(\omega) > 0$. Then $H(t) = 0$ unless $-\nu(t) = c(\omega)$, in which case $H(t)$ is independent of $t$ and has magnitude $|H(t)| = cC(\omega)^{-1/2}$ for some $c > 0$ depending only upon $k$.*

*Proof.* For $U_1 = \mathfrak{o}^\times$, these are standard assertions concerning Gauss sums. The standard proof adapts to the general case (compare with [Michel and Venkatesh 2010, 3.1.14]). □

## 6. Local study of nonarchimedean microlocal lifts

Recall Definition 21 and the statement of Lemma 22. Retain the notation of Section 5.

**6.1. *Proof of Lemma 22: determination of microlocal lifts.*** For any character $\chi : k^\times \to \mathbb{C}^\times$, the nonequivariant twisting isomorphism $\pi \to \pi' := \pi \otimes \chi$ induces nonequivariant linear isomorphisms

$$\begin{aligned} V :=& \{\text{microlocal lifts in } \pi \text{ of orientation } (\omega_1, \omega_2)\} \\ \cong& \{\text{microlocal lifts in } \pi' \text{ of orientation } (\omega_1', \omega_2')\}, \end{aligned} \tag{18}$$

with $\omega_i' := \omega_i \cdot \chi|_{\mathfrak{o}^\times}$. We thereby reduce to verifying the conclusion in the special case $\omega_1 = 1$. Suppose $V \neq 0$. Write $\omega := \omega_2$. By the convention $N \geqslant 1$ of Definition 21, $\omega$ is ramified. The central character $\chi_\pi$ of $\pi$ restricts to $\omega$, hence is ramified; by Theorem 42, $\dim V = \max(0, 1 + c(\pi) - c(\chi_\pi))$, and so $V \neq 0$ only if $c(\pi) \geqslant c(\chi_\pi)$. By Lemma 43, the latter happens only if $c(\pi) = c(\chi_\pi)$ and $\pi$ has the indicated form, in which case $\dim V = 1$. The explicit description of $V$ now follows in general from (18).

**6.2. *Explicit formulas.*** Let $\pi := \chi_1 \boxplus \chi_2$ and $\omega_i := \chi_i|_{\mathfrak{o}^\times}$ with $N := c(\omega_1/\omega_2) \geqslant 1$.

**Lemma 48.** *Define $f_1, f_2 \in C^\infty(k)$ (as if in the "line model" of Section 5.3) by*

$$f_1(x) := 1_{\mathfrak{p}^{N_2}}(x), \quad f_2(x) := 1_{\mathfrak{p}^{N_1}}(1/x)|1/x|\chi_1^{-1}\chi_2(x)$$

*and $v_1, v_2 \in \pi$ in the induced model on $g = \left(\begin{smallmatrix} * & * \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}_2(k)$ by*

$$v_1(g) := v_{f_1}(g) = 1_{\mathfrak{p}^{N_2}}(c/d)\left|\frac{\det g}{d^2}\right|^{1/2} \chi_1(\det(g)/d)\chi_2(d), \tag{19}$$

$$v_2(g) := v_{f_2}(g) = 1_{\mathfrak{p}^{N_1}}(d/c)\left|\frac{\det g}{c^2}\right|^{1/2} \chi_1(\det(g)/c)\chi_2(c), \tag{20}$$

*and $W_1, W_2 \in \pi$ in the Kirillov model $\mathcal{K}(\pi, \psi)$ by*[9]

$$W_1(y) := 1_{\mathfrak{p}^{-N_1}}(y)|y|^{1/2}\chi_1(y), \quad W_2(y) := 1_{\mathfrak{p}^{-N_1}}(y)|y|^{1/2}\chi_2(y). \tag{21}$$

*Then $v_1, W_1$ and $v_2, W_2$ are microlocal lifts of orientations $(\omega_1, \omega_2)$ and $(\omega_2, \omega_1)$, respectively.*

*Proof.* The formulas for $W_1, v_1$ in the case $\chi_1 = 1$ and those for $W_2, v_2$ in the case $\chi_2 = 1$ follow from known formulas for standard newvectors [Schmidt 2002]; the general case follows from the twisting isomorphisms (18). □

**6.3. *Stationary phase analysis of local Rankin–Selberg integrals.*** In this section we apply stationary phase analysis to evaluate and estimate some local Rankin–Selberg integrals involving microlocal lifts and newvectors. We use these in Section 7 to prove Theorem 25 and Theorem 29. Retain the notation of Section 5.1. Let $\chi_1, \chi_2$ be unitary characters of $k^\times$ for which $N := c(\chi_1/\chi_2)$ is positive. Let $\pi = \chi_1 \boxplus \chi_2$ be the corresponding generic irreducible unitary principal series representation of $\mathrm{GL}_2(k)$, realized in its induced model and equipped with the norm given in Section 5.3. Equip the complex-conjugate representation $\bar{\pi}$ with the compatible unitary structure. Define the intertwiner $\pi \ni v \mapsto W_v \in \mathcal{W}(\pi, \psi)$ as in Section 5.4. Let $\sigma$ be a generic irreducible unitary representation of $\mathrm{PGL}_2(k)$, realized in its $\psi$-Whittaker model $\sigma = \mathcal{W}(\sigma, \psi)$.

**Theorem 49.** *Let $v \in \pi$ be a microlocal lift of orientation $(\chi_1|_{\mathfrak{o}^\times}, \chi_2|_{\mathfrak{o}^\times})$, let $v' \in \pi$ be a generalized newvector, and let $W_1 \in \sigma$.*

(I) *If $N$ is large enough in terms of $W_1$, then*

$$\ell_{\mathrm{RS}}(W_1, \overline{W}_v, v) = cq^{-N/2}\|v\|^2 \int_{y \in k^\times} W_1(y)\, d^\times y,$$

*where*[10] *$c := q^{N/2} \int_{t \in k^\times} \chi_1\chi_2^{-1}(t)\psi(t)\, dt/|t| \asymp 1$ is a complex scalar which is independent of $W_1$ and whose magnitude depends only upon $k$.*

[9] Recall that $\psi$ is assumed unramified.

[10] The integral defining $c$ should be interpreted in the usual way as (for instance) a limit of integrals over increasing finite unions of $\mathfrak{o}^\times$-cosets.

(II) *One has* $\ell_{\mathrm{RS}}(W_1 \otimes \overline{W}_{v'} \otimes v') \ll q^{-N/2}\|v'\|^2$ *with the implied constant depending at most upon* $W_1$.

(III) *Suppose that* $\nu(2) = 0$, $\chi_\pi$ *is unramified*, $\|v'\| = \|v\|$, *the support of* $v'$ *is* $-N..N$, $\sigma$ *is unramified, and* $W_1 \in \sigma$ *is spherical, so that* $N = c(\chi_1) = c(\chi_2)$ *and* $c(\pi) = 2N$. *Then* $\ell_{\mathrm{RS}}(W_1, \overline{W}_v, v) = \ell_{\mathrm{RS}}(W_1, \overline{W}_{v'}, v')$.

The most difficult assertion is (II), which is used only to deduce the equidistribution of newvectors (Theorem 17). Assertion (III) serves only the purpose of illustration (see the discussion after Theorem 25). The other main results of this article (Theorems 29, 27) require only (I), whose proof is very short.

*Proof of* (I). Without loss of generality, let $v = v_f$ with $f(x) := 1_{\mathfrak{p}^{N_2}}(x)$. Because $N_2$ is large in enough in terms of $W_1$, we have whenever $f(x) \neq 0$ that $W_1(a(y)n'(x/u)) = W_1(y)$ for all $u \in \mathfrak{o}^\times$. Lemma 46 gives after the simplifications $f(x)\bar{f}(x + y/t) = 1_{\mathfrak{p}^{N_2}}(x)1_{\mathfrak{p}^{N_2}}(y/t)$ and $1_{\mathfrak{p}^{N_2}}(x)F(x, y, t; W_1, \mathfrak{o}^\times) = 1_{\mathfrak{p}^{N_2}}(x)W_1(y)H(t)$ with $H(t) := \mathbb{E}_{u\in\mathfrak{o}^\times}\chi_1\chi_2^{-1}(ut)\psi(ut)$ that

$$\ell_{\mathrm{RS}}(W_1, \overline{W}_v, v) = \int_{y\in k^\times} W_1(y) \int_{x\in k} 1_{\mathfrak{p}^{N_2}}(x) \int_{t\in k} 1_{\mathfrak{p}^{N_2}}(y/t)H(t)\,\frac{dt}{|t|}\,dx\,d^\times y.$$

We have $W_1(y)H(t) = 0$ unless $|t| \asymp q^N$ and $|y| \ll 1$; because $N_1$ is large enough in terms of $W_1$, the factor $1_{\mathfrak{p}^{N_2}}(y/t) = 1$ is thus redundant. Since $\int_{x\in k} 1_{\mathfrak{p}^{N_2}}(x)\,dx = \int_k |f|^2 = \|v\|^2$, we obtain the required identity. □

*Proof of* (II). Suppose first that $\chi_1$ and $\chi_2$ are both ramified. In that case, Lemma 44 says that $v' = v_f$ with $f$ a character multiple of the characteristic function of some $\mathfrak{o}^\times$ coset. In particular,

$$f \text{ is supported on a coset of } \mathfrak{o}^\times\text{, and } \bar{f} \otimes f \text{ is } \mathfrak{o}^\times\text{-invariant.} \tag{22}$$

From the mod-center identity $a(y)n'(x) \equiv n(y/x)a(y/x^2)wn(1/x)$, we have

$$W_1(a(y)n'(x)) = \psi(y/x)W_1(a(y/x^2)wn(1/x)). \tag{23}$$

From (23) and standard bounds on Whittaker functions, we have[11]

$$\sup_{x\in k} \int_{y\in k^\times} |W_1(a(y)n'(x))|\,d^\times y \ll 1. \tag{24}$$

By (23), there exists a fixed open subgroup $U_1 \leqslant \mathfrak{o}^\times$ for which

$$W_1(a(y)n'(x/u)) = W_1(a(y)n'(x)) \times \begin{cases} 1 & \text{for } |x| \leqslant 1, \\ \psi((u-1)y/x) & \text{for } |x| \geqslant 1. \end{cases} \tag{25}$$

Without loss of generality, suppose $\int_k |f|^2 = 1$. We apply Lemma 46, split the integral according as $|x| \leqslant 1$ or not, and appeal to (24) and (25); our task thereby reduces to showing with

$$H_1(t) := \mathbb{E}_{u\in U_1}\chi_1\chi_2^{-1}(ut)\psi(ut),$$
$$H_2(t, y/x) := \psi(-y/x)\mathbb{E}_{u\in U_1}\chi_1\chi_2^{-1}(ut)\psi(u(t + y/x))$$

[11] See [Michel and Venkatesh 2010, 3.2.3], and recall that $\sigma$ is assumed generic and unitary.

that the quantities

$$I_1 := \sup_{y \in k^\times} \int_{t \in k^\times} \int_{x \in k : |x| \leqslant 1} |f(x) \bar{f}(x + y/t) H_1(t)| \, d^\times t \, dx \, d^\times y,$$
$$I_2 := \sup_{y \in k^\times} \int_{t \in k^\times} \int_{x \in k : |x| > 1} |f(x) \bar{f}(x + y/t) H_2(t, y/x)| \, d^\times t \, dx \, d^\times y$$

are $O(q^{-N/2})$. We have $H_1(t) = 0$ unless $|t| \asymp q^N$, in which case $H_1(t) \ll q^{-N/2}$; the set of such $t$ has $d^\times t$-volume $O(1)$, so an adequate estimate for $I_1$ follows from Cauchy–Schwartz applied to the $x$-integral. Similarly, $H_2(t, y/x) = 0$ unless $|t + y/x| \asymp q^N$, in which case $H_2(t, y/x) \ll q^{-N/2}$; the support condition on $f$ shows that $f(x) \bar{f}(x + y/t) = 0$ unless $|t + y/x| = |t|$, we may conclude once again by Cauchy–Schwartz.[12]

We turn to the case that one of $\chi_1, \chi_2$ is unramified. By the assumption $c(\chi_1/\chi_2) \neq 0$, the other one is ramified. By symmetry, we may suppose that $\chi_1$ is unramified and $\chi_2$ is ramified. By Lemma 44, we may suppose without loss of generality that $v' = v_f$ for $f = 1_{\mathfrak{a}}$ with $\mathfrak{a} \subset k$ a fractional $\mathfrak{o}$-ideal. Then $\bar{f} \otimes f$ is $\mathfrak{o}^\times$-invariant. We split the integral over $x \in k$ as above, and the same argument works for the range $|x| \leqslant 1$. The remaining range contributes

$$I_3 := \int_{x \in k : |x| > 1} \int_{y \in k^\times} \int_{t \in k^\times} 1_{\mathfrak{a}}(x) 1_{\mathfrak{a}}(x + y/t) H_3(t, y/x; x) \, dx \, d^\times y \, d^\times t, \tag{26}$$

where

$$H_3(t, y/x; x) := W_1(a(y/x^2) w n(1/x)) \mathbb{E}_{u \in U_1} \chi_1 \chi_2^{-1}(ut) \psi(u(t + y/x)).$$

A bit more care is required than in the above argument, which gives now an upper bound of $+\infty$; the problem is that the nonvanishing of $H_3(t, y/x; x)$ no longer restricts $t$ to a volume $O(1)$ subset of $k^\times$. We do better here by exploiting additional cancellation coming from the $y$-integral: Let $C_1, C_2$ be positive scalars, depending only upon $W_1, U_1$, so that

$$H_3(t, y/x; x) \neq 0 \Longrightarrow C_1 q^N < |y/x + t| < C_2 q^N. \tag{27}$$

If $|y/x| \geqslant C_2 q^N$, then $H_3(t, y/x; x) \neq 0$ only if $|t| = |y/x|$. If $|y/x| \leqslant C_1 q^N$, then $H_3(t, y/x; x) \neq 0$ only if $C_1 q^N < |t| < C_2 q^N$. Arguing as above, we reduce to considering the range $C_1 q^N < |y/x| < C_2 q^N$, in which $H_3(t, y/x; x) \neq 0$ only if $|t| < C_2 q^N$. The range $C_1 q^N \leqslant |t| < C_2 q^N$ may be treated as before, so we reduce to showing that

$$I_4 := \int_{\substack{x, y, t \in k, k^\times, k^\times : |x| > 1, \\ C_1 q^N < |y/x| < C_2 q^N, \ |t| < C_1 q^N}} 1_{\mathfrak{a}}(x) 1_{\mathfrak{a}}(x + y/t) H_3(t, y/x; x) \, dx \, d^\times y \, d^\times t = 0. \tag{28}$$

[12] The estimate just derived is essentially sharp when $f$ is supported in a fixed open subset of $k^\times$, but can be substantially sharpened when $f$ is “unbalanced” in the sense that its support tends sufficiently rapidly with $N$ either to zero or infinity. The possibility of such sharpening is the simplest case of the “weak subconvexity” phenomenon identified in [Nelson et al. 2014].

Note that the conditions defining the integrand imply that $|xt/y| < 1$. There is an open subgroup $U_2$ of $\mathfrak{o}^\times$, depending only upon $W_1$, $U_1$, so that

$$z \in U_2, |xt/y| < 1 \Longrightarrow \frac{t+zy/x}{t+y/x} \in U_1, \tag{29}$$

$$|x| > 1, z \in U_2 \Longrightarrow W_1(a(zy/x^2)wn(1/x)) = W_1(a(y/x^2)wn(1/x)), \tag{30}$$

$$|xt/y| < 1, z \in U_2 \Longrightarrow 1_{\mathfrak{a}}(x+zy/t) = 1_{\mathfrak{a}}(x+y/t). \tag{31}$$

For $N$ large enough in terms of $W_1$, $U_1$ and hence $U_2$, we have

$$|xt/y| < 1 \Longrightarrow \mathbb{E}_{z \in U_2} \chi_1^{-1}\chi_2(t+zy/x) = 0. \tag{32}$$

In $I_4$, we substitute $y \mapsto yz$ with $z \in U_2$ and average over $z$; by (31) and (30), our task reduces to establishing for $|xt/y| < 1$ that

$$\mathbb{E}_{z \in U_2} \mathbb{E}_{u \in U_1} \chi_1 \chi_2^{-1}(ut) \psi(u(t+zy/x)) = 0,$$

which follows from (32) after the change of variables $u \mapsto u(t+y/x)/(t+zy/x)$ suggested by (29). $\square$

*Proof of* (III). By (I), our task reduces to showing that

$$\ell_{\mathrm{RS}}(W_1, \overline{W}_{v'}, v') = cq^{-N/2} \|v'\|^2 \int_{y \in k^\times} W_1(y)\, d^\times y$$

with the same scalar $c$ as in (I). Suppose without loss of generality that $v' = v_f$ with $f := \chi_2 1_{\mathfrak{o}^\times}$. Note that $\bar{f} \otimes f$ is $\mathfrak{o}^\times$-invariant. If $f(x) \neq 0$, then $W_1(a(y)n'(x/u)) = W_1(y)$ for all $u \in \mathfrak{o}^\times$. Lemma 46 gives after the simplification $f(x)F(x, y, t; W_1, \mathfrak{o}^\times) = f(x)W_1(y)H(t)$ with $H$ as in the proof of (I) that

$$\ell_{RS}(W_1, \overline{W}_{v'}, v') = \int_{y \in k^\times} \int_{x \in k} \int_{t \in k} W_1(y) f(x) \bar{f}(x+y/t) H(t) \frac{dt}{|t|}\, dx\, d^\times y.$$

Because $\nu(2) = 0$, we have $c(\chi_2) = c(\chi_1\chi_2^{-1}) = N$. Thus if $W_1(y)f(x)H(t) \neq 0$, then $y, x, t \in \mathfrak{o}, \mathfrak{o}^\times, \varpi^{-N}\mathfrak{o}^\times$ and so $f(x)\bar{f}(x+y/t) = 1$. From $\int_{x \in k} 1_{\mathfrak{o}^\times}(x)\, dx = \int_k |f|^2 = \|v'\|^2$, we conclude. $\square$

**Remark 50.** [Michel and Venkatesh 2010, 3.4.2; 2010, (3.25)] and Theorem 49(I) imply the following: Let $v_2, v_3 \in \pi$ be microlocal lifts of the same orientation and $v_1 \in \sigma$, realized in its Kirillov model $\mathcal{K}(\sigma, \psi)$. The formula $\|v_1\|^2 := \int_{y \in k^\times} |v_1(y)|^2\, d^\times y$ is known to define an invariant norm on $\sigma$. Suppose that $N$ is large enough in terms of $v_1$. Then

$$\int_{g \in Z \backslash G} \prod_{i=1,2,3} \langle \pi_i(g) v_i, v_i \rangle = cq^{-N} \|v_2\|^2 \|v_3\|^2 \int_{y \in k^\times} \langle a(y) v_1, v_1 \rangle\, d^\times y$$

for some positive scalar $c \asymp 1$ depending only upon $k$. This identity solves the problem of producing a subconvexity-critical test vector for the local triple product period in the QUE case when the varying representation is principal series. It would be interesting to verify whether the supercuspidal case follows similarly using a modification of Definition 21 involving characters on an $\varepsilon$-neighborhood in $\mathrm{GL}_2(\mathfrak{o})$ of the points of a suitable nonsplit torus, where $\varepsilon \asymp C(\bar{\pi} \otimes \pi)^{-1/4}$.[13]

## 7. Completion of the proof

In this section, $\varphi \in \pi \in A_0(\boldsymbol{X})$ traverses a sequence of $L^2$-normalized microlocal lifts on $\boldsymbol{X}$ of level $N \to \infty$. Thus $\varphi$ and $\pi$, like most objects to be considered in this section, depend upon $N$, but we omit this dependence from our notation. We use the abbreviations *fixed* to mean "independent of $N$" and *eventually* to mean "for large enough $N$." Asymptotic notation such as $o(1)$ refers to the $N \to \infty$ limit. Our aim is to verify the conclusions of Theorem 25 and Theorem 29.

As $G$-modules, $\pi \cong \chi_1 \boxplus \chi_2$ for some unitary characters $\chi_1, \chi_2$ of $\mathbb{Q}_p^\times$ for which $c(\chi_1/\chi_2) = N$.

Recall our simplifying assumption that $R$ is a maximal order. This implies that for any irreducible $\mathcal{H}$-submodule $\pi'$ of $\mathcal{A}(\boldsymbol{X})$, the vector space underlying $\pi'$ is an irreducible admissible $G$-module. In other words, the local components at all places $v \neq p$ are one-dimensional.

The function $\varphi$ has unitary central character, so the measure $\mu_\varphi$ is invariant by the center. Let $\ell$ be a prime dividing the discriminant of $B$. Recalling from Section 3.1.1 that $T_\ell$ is an involution modulo the center, we see that it acts on $\pi$ by some scalar of magnitude one. Thus $\mu_\varphi$ is $T_\ell$-invariant. The natural space of observables against which it suffices to test $\mu_\varphi$ is thus

$$\mathcal{A}^+(\boldsymbol{X}) := \left\{ \Psi \in \mathcal{A}(\boldsymbol{X}) : \begin{array}{r} T_\ell \Psi = \Psi \text{ for } \ell \mid \operatorname{disc}(B), \\ z\Psi = \Psi \text{ for } z \in Z := \text{center of } G, \end{array} \right\}.$$

That space decomposes further as $\mathcal{A}^+(\boldsymbol{X}) = (\oplus_\chi \mathbb{C}(\chi \circ \det)) \oplus \mathcal{A}_0^+(\boldsymbol{X})$ where

- $\chi$ traverses the set of quadratic characters of the compact group $\mathbb{Q}_p^\times/\mathbb{Z}[1/p]^\times$ satisfying $\chi(\ell) = 1$ for $\ell \mid \operatorname{disc}(B)$, and
- $\mathcal{A}_0^+(\boldsymbol{X}) := \mathcal{A}^+(\boldsymbol{X}) \cap \mathcal{A}_0(\boldsymbol{X})$, which decomposes further as a countable direct sum $\mathcal{A}_0^+(\boldsymbol{X}) = \oplus_{\sigma \in A_0^+(\boldsymbol{X})} \sigma$ where we substitute $A$ for $\mathcal{A}$ to denote "irreducible submodules of."

Let $\sigma \in A^+(\boldsymbol{X})$ be fixed. It is either one-dimensional and of the form $\mathbb{C}(\chi \circ \det)$ for some $\chi$ as above, or belongs to $A_0^+(\boldsymbol{X})$ and is generic as a $G$-module. Denote by $\ell_{\mathrm{Aut}} : \sigma \otimes \bar{\pi} \otimes \pi \to \mathbb{C}$ the $G$-invariant functional defined by integration over $\boldsymbol{X}$.

**Lemma 51.** *Suppose $\sigma$ is one-dimensional and $\ell_{\mathrm{Aut}} \neq 0$. Then $\sigma$ is trivial eventually.*

*Proof.* Write $\sigma = \mathbb{C}(\chi \circ \det)$ for some quadratic character $\chi$. By Schur's lemma, $\pi \cong \chi_1 \boxplus \chi_2$ is isomorphic as a $G$-module to $\pi \otimes \chi \circ \det \cong \chi_1\chi \boxplus \chi_2\chi$, which happens (see, e.g., [Schmidt 2002]) only if either $\chi_1 = \chi_1\chi$, in which case $\chi$ is trivial, or $\chi_1 = \chi_2\chi$, in which case $c(\chi) = c(\chi_1/\chi_2) = N \to \infty$, which does not happen because $\chi$ is quadratic.[14] $\square$

[13] Added later: the recent work [Nelson et al. $\geq$ 2018] contains results in this direction.

[14] We use here that the local field $\mathbb{Q}_p$ is not a function field of characteristic 2.

We now prove Theorem 25. It suffices to verify that the various assertions hold for fixed $\Psi \in \sigma \in A^+(\boldsymbol{X})$. They are tautological if $\sigma$ is trivial, so by Lemma 51, we reduce to the case that $\sigma \in A_0^+(\boldsymbol{X})$ is generic. Fix an unramified nontrivial character $\psi : \mathbb{Q}_p \to \mathbb{C}^{(1)}$ and $G$-equivariant isometric isomorphisms $\sigma \cong \mathcal{W}(\sigma, \psi)$, $\pi \cong \chi_1 \boxplus \chi_2$. Denote by $\ell_{\mathrm{RS}} : \sigma \otimes \bar{\pi} \otimes \pi \to \mathbb{C}$ the trilinear form defined in Section 5.6. By Theorem 45 and the nonvanishing of $\ell_{\mathrm{RS}}$, there exists a complex scalar $\mathcal{L}^{1/2} \in \mathbb{C}$ so that

$$\ell_{\mathrm{Aut}} = \mathcal{L}^{1/2} \ell_{\mathrm{RS}}. \tag{33}$$

Theorem 49(I) implies that $\ell_{\mathrm{RS}}(\sigma(a(y))\Psi, \bar{\varphi}, \varphi) = \ell_{\mathrm{RS}}(\Psi, \bar{\varphi}, \varphi)$ holds eventually for fixed $y \in k^\times$; the required diagonal invariance then follows from (33). If $p \neq 2$ and $\varphi'$ is an $L^2$-normalized newvector of support $-N..N$ and $\Psi \in \sigma^K$ is spherical, then Theorem 49(III) gives $\ell_{\mathrm{RS}}(\Psi, \bar{\varphi}, \varphi) = \ell_{\mathrm{RS}}(\Psi, \bar{\varphi}', \varphi')$ eventually; the lifting property then follows from (33). For the equidistribution application, we reduce by Lemma 51 and (33) and Theorem 49(II) to showing that $\mathcal{L}^{1/2} = o(p^{N/2})$ holds under the hypothesis that for each fixed $\Psi_0 \in \sigma$, one has $\ell_{\mathrm{Aut}}(\Psi_0, \bar{\varphi}, \varphi) = o(1)$. Let $\Psi_0 \in \sigma \cong \mathcal{W}(\sigma, \psi)$ be given in the Kirillov model by the characteristic function of the unit group. By Theorem 49(I), $\ell_{\mathrm{RS}}(\Psi_0, \bar{\varphi}, \varphi) \asymp p^{-N/2}$ eventually, so our hypothesis and (33) give the required estimate for $\mathcal{L}^{1/2}$.

We turn to the proof of Theorem 29. Our assumptions on $\pi$ and $\sigma$ imply that $\sigma \in A_0^+(\boldsymbol{X})$ and that the adelizations of $\sigma, \bar{\pi}$ and $\pi$ at each $v \in S_B := \{\infty\} \cup \{\ell : \ell \mid \mathrm{disc}(B)\}$ are one-dimensional and have trivial tensor product, hence that the product of their normalized matrix coefficients is one; by Ichino's formula [Ichino and Ikeda 2010] and [Michel and Venkatesh 2010, 3.4.2], it follows that $L \asymp |\mathcal{L}^{1/2}|^2$, where $L$ denotes the LHS of (5) and $\mathcal{L}^{1/2}$ is as above (compare with Remark 50). By Theorem 27 and the argument of the previous paragraph, $\mathcal{L}^{1/2} = o(p^{N/2})$. Our goal is to show that $L = o(C^{1/4})$, where $C := C(\sigma \times \bar{\pi} \times \pi)$ is the global conductor; the contribution to $C$ from $v \in S_B$ is bounded, while the contribution from $p$ is

$$C(\sigma_p \otimes \chi_1^{-1}\chi_2) C(\sigma_p \otimes \chi_2^{-1}\chi_1) C(\sigma_p)^2 \asymp C(\chi_1^{-1}\chi_2)^4 = p^{4N}.$$

Thus $C \asymp p^{4N}$. The known estimate $\mathcal{L}^{1/2} = o(p^{N/2})$ thus translates to the goal $L = o(C^{1/4})$, as required.

## Acknowledgements

This paper owes an evident debt of ideas and inspiration to E. Lindenstrauss's work [2006b]; we thank him also for helpful feedback and interest. We thank M. Einsiedler for helpful discussions on measure classification and feedback on an earlier draft, P. Sarnak and A. Venkatesh for several helpful discussions informing our general understanding of microlocal lifts and microlocal analysis, A. Saha for helpful references concerning conductors, S. Jana for feedback on entropy bounds, Y. Hu for helpful clarifying questions, and E. Kowalski, Ph. Michel and D. Ramakrishnan for encouragement. We gratefully acknowledge the support of NSF grant OISE-1064866 and SNF grant SNF-137488 during the work leading to this paper. Finally, we thank the anonymous referee for many helpful corrections and suggestions concerning this work.

## References

[Anantharaman and Le Masson 2015] N. Anantharaman and E. Le Masson, "Quantum ergodicity on large regular graphs", *Duke Math. J.* **164**:4 (2015), 723–765. MR Zbl

[Atkin and Lehner 1970] A. O. L. Atkin and J. Lehner, "Hecke operators on $\Gamma_0(m)$", *Math. Ann.* **185** (1970), 134–160. MR Zbl

[Bourgain and Lindenstrauss 2003] J. Bourgain and E. Lindenstrauss, "Entropy of quantum limits", *Comm. Math. Phys.* **233**:1 (2003), 153–171. MR Zbl

[Brandt 1943] H. Brandt, "Zur Zahlentheorie der Quaternionen", *Jber. Deutsch. Math. Verein.* **53** (1943), 23–57. MR

[Brooks and Lindenstrauss 2010] S. Brooks and E. Lindenstrauss, "Graph eigenfunctions and quantum unique ergodicity", *C. R. Math. Acad. Sci. Paris* **348**:15-16 (2010), 829–834. MR Zbl

[Brooks and Lindenstrauss 2013] S. Brooks and E. Lindenstrauss, "Non-localization of eigenfunctions on large regular graphs", *Israel J. Math.* **193**:1 (2013), 1–14. MR Zbl

[Brooks and Lindenstrauss 2014] S. Brooks and E. Lindenstrauss, "Joint quasimodes, positive entropy, and quantum unique ergodicity", *Invent. Math.* **198**:1 (2014), 219–259. MR Zbl

[Bump 1997] D. Bump, *Automorphic forms and representations*, Cambridge Studies in Advanced Mathematics **55**, Cambridge University Press, 1997. MR Zbl

[Casselman 1973a] W. Casselman, "On some results of Atkin and Lehner", *Math. Ann.* **201** (1973), 301–314. MR Zbl

[Casselman 1973b] W. Casselman, "The restriction of a representation of $\mathrm{GL}_2(k)$ to $\mathrm{GL}_2(\mathfrak{o})$", *Math. Ann.* **206** (1973), 311–318. MR Zbl

[Eichler 1955] M. Eichler, "Zur Zahlentheorie der Quaternionen-Algebren", *J. Reine Angew. Math.* **195** (1955), 127–151. MR

[Einsiedler and Lindenstrauss 2008] M. Einsiedler and E. Lindenstrauss, "On measures invariant under diagonalizable actions: the rank-one case and the general low-entropy method", *J. Mod. Dyn.* **2**:1 (2008), 83–128. MR Zbl

[Gross 1987] B. H. Gross, "Heights and the special values of $L$-series", pp. 115–187 in *Number theory* (Montreal, Quebec, 1985), edited by H. Kisilevsky and J. Labute, CMS Conf. Proc. **7**, Amer. Math. Soc., Providence, RI, 1987. MR Zbl

[Holowinsky and Soundararajan 2010] R. Holowinsky and K. Soundararajan, "Mass equidistribution for Hecke eigenforms", *Ann. of Math.* (*2*) **172**:2 (2010), 1517–1528. MR Zbl

[Hu 2018] Y. Hu, "Triple product formula and mass equidistribution on modular curves of level $N$", *Int. Math. Res. Not.* **2018**:9 (2018), 2899–2943. MR

[Ichino 2008] A. Ichino, "Trilinear forms and the central values of triple product $L$-functions", *Duke Math. J.* **145**:2 (2008), 281–307. MR Zbl

[Ichino and Ikeda 2010] A. Ichino and T. Ikeda, "On the periods of automorphic forms on special orthogonal groups and the Gross–Prasad conjecture", *Geom. Funct. Anal.* **19**:5 (2010), 1378–1425. MR Zbl

[Knapp 1986] A. W. Knapp, *Representation theory of semisimple groups*, Princeton Mathematical Series **36**, Princeton University Press, 1986. MR Zbl

[Kneser 1966] M. Kneser, "Strong approximation", pp. 187–196 in *Algebraic Groups and Discontinuous Subgroups* (Boulder, Colorado, 1965), edited by A. Borel and G. D. Mostow, Proc. Sympos. Pure Math. **9**, Amer. Math. Soc., Providence, RI, 1966. MR Zbl

[Le Masson 2014] E. Le Masson, "Pseudo-differential calculus on homogeneous trees", *Ann. Henri Poincaré* **15**:9 (2014), 1697–1732. MR Zbl

[Le Masson and Sahlsten 2017] E. Le Masson and T. Sahlsten, "Quantum ergodicity and Benjamini–Schramm convergence of hyperbolic surfaces", *Duke Math. J.* **166**:18 (2017), 3425–3460. MR Zbl

[Lindenstrauss 2001] E. Lindenstrauss, "On quantum unique ergodicity for $\Gamma\backslash\mathbb{H}\times\mathbb{H}$", *Internat. Math. Res. Notices* 17 (2001), 913–933. MR Zbl

[Lindenstrauss 2006a] E. Lindenstrauss, "Adelic dynamics and arithmetic quantum unique ergodicity", pp. 111–139 in *Current developments in mathematics, 2004*, edited by D. Jerison et al., Int. Press, Somerville, MA, 2006. MR Zbl

[Lindenstrauss 2006b] E. Lindenstrauss, "Invariant measures and arithmetic quantum unique ergodicity", *Ann. of Math.* (*2*) **163**:1 (2006), 165–219. MR Zbl

[Lubotzky et al. 1988] A. Lubotzky, R. Phillips, and P. Sarnak, "Ramanujan graphs", *Combinatorica* **8**:3 (1988), 261–277. MR Zbl

[Luo and Sarnak 2003] W. Luo and P. Sarnak, "Mass equidistribution for Hecke eigenforms", *Comm. Pure Appl. Math.* **56**:7 (2003), 874–891. MR Zbl

[Michel and Venkatesh 2010] P. Michel and A. Venkatesh, "The subconvexity problem for $\mathrm{GL}_2$", *Publ. Math. Inst. Hautes Études Sci.* 111 (2010), 171–271. MR Zbl

[Nelson 2011] P. D. Nelson, "Equidistribution of cusp forms in the level aspect", *Duke Math. J.* **160**:3 (2011), 467–501. MR Zbl

[Nelson 2012] P. D. Nelson, "Mass equidistribution of Hilbert modular eigenforms", *Ramanujan J.* **27**:2 (2012), 235–284. MR Zbl

[Nelson 2015] P. D. Nelson, "Evaluating modular forms on Shimura curves", *Math. Comp.* **84**:295 (2015), 2471–2503. MR Zbl

[Nelson 2016] P. D. Nelson, "Quantum variance on quaternion algebras, I", preprint, 2016. arXiv

[Nelson 2017] P. D. Nelson, "Analytic isolation of newforms of given level", *Arch. Math.* (*Basel*) **108**:6 (2017), 555–568. MR Zbl

[Nelson et al. 2014] P. D. Nelson, A. Pitale, and A. Saha, "Bounds for Rankin–Selberg integrals and quantum unique ergodicity for powerful levels", *J. Amer. Math. Soc.* **27**:1 (2014), 147–191. MR Zbl

[Nelson et al. $\geq$ 2018] P. D. Nelson, A. Saha, and Y. Hu, "Some analytic aspects of automorphic forms on GL(2) of minimal type", To appear in *Comm. Math. Helv.*

[Piatetski-Shapiro and Rallis 1987] I. Piatetski-Shapiro and S. Rallis, "Rankin triple $L$ functions", *Compositio Math.* **64**:1 (1987), 31–115. MR Zbl

[Pizer 1980] A. Pizer, "An algorithm for computing modular forms on $\Gamma_0(N)$", *J. Algebra* **64**:2 (1980), 340–390. MR Zbl

[Prasad 1990] D. Prasad, "Trilinear forms for representations of GL(2) and local $\epsilon$-factors", *Compositio Math.* **75**:1 (1990), 1–46. MR Zbl

[Reznikov 2001] A. Reznikov, "Laplace–Beltrami operator on a Riemann surface and equidistribution of measures", *Comm. Math. Phys.* **222**:2 (2001), 249–267. MR Zbl

[Rudnick and Sarnak 1994] Z. Rudnick and P. Sarnak, "The behaviour of eigenstates of arithmetic hyperbolic manifolds", *Comm. Math. Phys.* **161**:1 (1994), 195–213. MR Zbl

[Sage 2015] W. A. Stein et al., "Sage Mathematics Software", 2015, available at http://www.sagemath.org. Version 6.7.

[Sarnak 2011] P. Sarnak, "Recent progress on the quantum unique ergodicity conjecture", *Bull. Amer. Math. Soc.* (*N.S.*) **48**:2 (2011), 211–228. MR Zbl

[Schmidt 2002] R. Schmidt, "Some remarks on local newforms for GL(2)", *J. Ramanujan Math. Soc.* **17**:2 (2002), 115–147. MR Zbl

[Serre 2003] J.-P. Serre, *Trees*, Springer, Berlin, 2003. MR Zbl

[Silberman and Venkatesh 2007] L. Silberman and A. Venkatesh, "On quantum unique ergodicity for locally symmetric spaces", *Geom. Funct. Anal.* **17**:3 (2007), 960–998. MR Zbl

[Silberman and Venkatesh $\geq$ 2018] L. Silberman and A. Venkatesh, "Quantum unique ergodicity for locally symmetric spaces II", available at http://www.math.ubc.ca/~lior/work/AQUE-nov6.pdf. Zbl

[Soundararajan 2010] K. Soundararajan, "Weak subconvexity for central values of $L$-functions", *Ann. of Math.* (2) **172**:2 (2010), 1469–1498. MR Zbl

[Templier 2014] N. Templier, "Large values of modular forms", *Camb. J. Math.* **2**:1 (2014), 91–116. MR Zbl

[Venkatesh 2010] A. Venkatesh, "Sparse equidistribution problems, period bounds and subconvexity", *Ann. of Math.* (2) **172**:2 (2010), 989–1094. MR Zbl

[Vignéras 1980] M.-F. Vignéras, *Arithmétique des algèbres de quaternions*, Lecture Notes in Mathematics **800**, Springer, Berlin, 1980. MR Zbl

[Voight 2018] J. Voight, "Quaternion algebras", Dartmouth College, 2018, available at https://math.dartmouth.edu/~jvoight/quat.html. Zbl

[Wolpert 2001] S. A. Wolpert, "The modulus of continuity for $\Gamma_0(m)\backslash\mathbb{H}$ semi-classical limits", *Comm. Math. Phys.* **216**:2 (2001), 313–323. MR Zbl

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

[Zelditch 1992] S. Zelditch, "On a "quantum chaos" theorem of R. Schrader and M. Taylor", *J. Funct. Anal.* **109**:1 (1992), 1–21. MR Zbl

Communicated by Philippe Michel
Received 2017-01-26 Revised 2018-04-09 Accepted 2018-07-15

paul.nelson@math.ethz.ch *Departement Mathematik, ETH, Zurich, Switzerland*

# Algebra & Number Theory

msp.org/ant



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

The subscription price for 2018 is US $340/year for the electronic version, and $535/year (+$55, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

Algebra & Number Theory (ISSN 1944-7833 electronic, 1937-0652 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

ANT peer review and production are managed by EditFLOW® from MSP.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

# Algebra & Number Theory

Volume 12 No. 9 2018





1937-0652(2018)12:9;1-8