# QUANTUM UNIQUE ERGODICITY FOR LOCALLY SYMMETRIC SPACES II

LIOR SILBERMAN AND AKSHAY VENKATESH

Abstract. We prove the arithmetic quantum unique ergodicity (AQUE) conjecture for non-degenerate sequences of Hecke eigenfunctions on quotients $\Gamma\backslash G/K$, where $G \simeq \mathrm{PGL}_d(\mathbb{R})$, $K$ is a maximal compact subgroup of $G$ and $\Gamma < G$ is a lattice associated to a division algebra over $\mathbb{Q}$ of prime degree $d$.

The primary novelty of the present paper is a new method of proving positive entropy of quantum limits, which avoids sieves and yields better bounds than previous techniques. The result on AQUE is obtained by combining this with a measure-rigidity theorem due to Einsiedler-Katok, following a strategy first pioneered by Lindenstrauss.

## Contents



## 1. Introduction

1.1. **Result.** In this paper, we shall show (a slightly sharper version of) the following statement. For precise definitions we refer to §6 especially §6.2.

**Theorem 1.1.** *Let $\Gamma$ be a lattice in $\mathrm{PGL}_d(\mathbb{R})$, with $d$ prime, associated to a division algebra*[1] *, and $\psi_i$ a non-degenerate sequence of Hecke-Maass eigenfunctions on $Y := \Gamma\backslash\mathrm{PGL}_d(\mathbb{R})/\mathrm{PO}_d(\mathbb{R})$, normalized to have $L^2$-norm 1 w.r.t. the Riemannian volume $d$vol.*

*Then the measures $|\psi_i|^2 d\mathrm{vol}$ converge weakly to the Haar measure, i.e. for any $f \in C(Y)$,*

$$\lim_{i\to\infty} \int_Y |\psi_i|^2 f d\mathrm{vol} = \int_Y f d\mathrm{vol}. \tag{1.1}$$

[1] This means that $\Gamma$ is the image of $\mathcal{O}^\times$ in $\mathrm{PGL}_d(\mathbb{R})$, where $\mathcal{O}$ is an order in a $\mathbb{Q}$-division algebra so that $\mathcal{O}\otimes\mathbb{R} = M_d(\mathbb{R})$. We also impose a class number one condition, see §6.

In words, Theorem 1.1 asserts that the eigenfunctions $\psi_i$ are "quite spread out" – that they do not cluster too much on the manifold $Y$.

This theorem is a contribution to the study of the "Arithmetic Quantum Unique Ergodicity" problem. A detailed introduction to this problem may be found in our paper [11]. While it is hard to dispute that the spaces $Y$ are far too special to provide a reasonable model for the physical problem of "quantum chaos," both the statement (1.1) and the techniques we use to prove it seem to the authors to be of interest because of the scarcity of results concerning analysis of higher rank automorphic forms. In particular, we believe our techniques will find applications to analytic problems beside QUE.

Our strategy follows that of Lindenstrauss in his proof of the arithmetic QUE conjecture for quotients of the hyperbolic plane (the case $d = 2$ of the Theorem above) and has three conceptually distinct steps. Let $A \subset \mathrm{PGL}_d(\mathbb{R})$ be the subgroup of diagonal matrices.

(1) *Microlocal lift:* notation as above, any weak limit (as $i \to \infty$) of $|\psi_i|^2 \, d\mathrm{vol}$ may be lifted to an $A$-invariant measure $\sigma_\infty$ on $X := \Gamma\backslash\mathrm{PGL}_d(\mathbb{R})$.
(2) *Mass of small tubes:* If $\sigma_\infty$ is as in (1), then the $\sigma_\infty$-mass of an $\epsilon$-ball in $\Gamma\backslash\mathrm{PGL}_d(\mathbb{R})$ is $\ll \epsilon^{d-1+\delta}$, for some $\delta > 0$; note that the bound $\ll \epsilon^{d-1}$ is trivial from the $A$-invariance. [2]
(3) *Measure rigidity:* Any $A$-invariant measure satisfying the auxiliary condition prescribed by (2) must necessarily be a convex combination of algebraic measures. In our setting ($\Gamma$ associated to a division algebra of prime degree) this means it must be Haar measure.

* * *

In the context of Lindenstrauss' proof, the analogues of the steps 1, 2 and 3 are due, respectively, to S. Zelditch [13], Bourgain-Lindenstrauss [1] and Lindenstrauss [6]. It should be noted, however, that in the setting considered by Lindenstrauss, (3) is false as stated, and one of the most remarkable features of the paper [6] is the idea that "Hecke recurrence" could be used as an additional assumption to make such a statement valid. Further, the very existence of such a strategy was not at all clear until the appearance of [6].

We shall therefore concern ourselves with the higher rank case ($d > 2$), where step 1 – under a nondegeneracy condition – has been established by the authors in [11]. This novel part of the present paper is then the establishment of Step 2. Step 3 was established by Einsiedler-Katok-Lindenstrauss in [4].

*However*, we shall actually establish a stronger form of Step 2, which will allow us to use a weaker form of Step 3 due to Einsiedler-Katok. At the original time that this proof was completed, the work [4] was not available; moreover, it will be no effort to prove the stronger form of Step 2. In this stronger version, we shall replace small balls by tubes around an $M$-orbit, when $M \leq \mathrm{PGL}_d(\mathbb{R})$ is a Levi subgroup, and still obtain nontrivial bounds on the mass of such tubes.

1.2. **Bounding mass of tubes – vague discussion.** As was discussed in the previous Section, the main point of the present paper is to prove upper bounds for the mass of eigenfunctions in small tubes.

[2] This asserts, then, that $\sigma_\infty$ has some "thickness" transverse to the $A$-direction; for instance, it immediately implies that the dimension of the support of $\sigma_\infty$ is strictly larger than $d - 1$.

A *correspondence* on a manifold $X$, for our purposes, will be a subset $S \subset X \times X$ such that both projections are topological coverings. Such a correspondence induces an endomorphism of $L^2(X)$: pull back to $Y$ and push forward to $X$. We also think of a correspondence as a "multi-valued" or "set-valued" function $h_S$ from $X$ to $X$. In the latter view a correspondence induces a natural convolution action on functions on $X$, given by $(f * h_S)(x) = \sum_{y \in h_S(x)} f(y)$.

Two correspondences can be composed in a natural way and resulting algebra is, in general, non-commutative. However, the manifolds of interest to us ($X = \Gamma \backslash \mathrm{PGL}_d(\mathbb{R})$ with $\Gamma$ an arithmetic lattice) come equipped with a large algebra of *commuting* correspondences, the Hecke algebra $\mathcal{H}$. Moreover, these correspondences act on $L^2(X)$ by normal operators. We will then be interested in possible concentration of simultaneous eigenfunctions of the Hecke algebra.

As a concrete example, for $X = \mathrm{PGL}_d(\mathbb{Z}) \backslash \mathrm{PGL}_d(\mathbb{R})$ the Hecke correspondences are induced by *left* multiplication with $\mathrm{PGL}_d(\mathbb{Q})$: given $\gamma \in \mathrm{PGL}_d(\mathbb{Q})$ and a coset $x \in X$, we consider the set of products $\gamma g$ as $g$ varies over representatives in $\mathrm{PGL}_d(\mathbb{R})$ for $x$. It turns out that these products generate a finite set of cosets $h_\gamma(x) \subset X$. It is easy to check that the adjoint of $h_\gamma$ is $h_{\gamma^{-1}}$, but the commutativity of the Hecke algebra is more subtle. An important feature of the Hecke correspondences on $X$ is their equivariance w.r.t. the action of $G = \mathrm{PGL}_d(\mathbb{R})$ on $X$ on the right.

Returning to the general $X := \Gamma \backslash \mathrm{PGL}_d(\mathbb{R})$, let $\mathcal{T}(\epsilon)$ be a small subset of $G$, with its size in certain directions on the order of $\epsilon$ (for us $\mathcal{T}(\epsilon)$ will be a tube of width $\epsilon$ around a compact piece of a Levi subgroup of $G$). Our goal will be to prove a statement of the following type, for some fixed $\eta > 0$ depending only on $G$:

$$\text{Each } \mathcal{H}\text{-eigenfunction } \psi \in L^2(X) \text{ satisfies } \mu_\psi(x\mathcal{T}(\epsilon)) \ll \epsilon^\eta. \tag{1.2}$$

Here $\mu_\psi := |\psi|^2 \, d\mathrm{vol}$ is the product of the $\mathrm{PGL}_d(\mathbb{R})$-invariant measure and the function $|\psi|^2$, normalized to be a probability measure. (1.2) asserts that the eigenfunction $\psi$ cannot concentrate on a small tube. This is proven, in the cases of interest for this paper, in Theorem 5.4.

We will sketch here our approach to the proof. A basic form of the idea appeared in the paper [8] of Rudnick and Sarnak. It is essentially that, when $\psi$ is an eigenfunction of a correspondence $h \in \mathcal{H}$, if $\psi$ were large at some point $x$, it also tends to be quite large at points belonging to the orbit $h.x$. We can thereby "disperse" the local question of bounding the mass of a small tube to a global question about the size of $\psi$ throughout the manifold.

Say that the image of the point $x$ under $h \in \mathcal{H}$ is the collection of $N$ points $h.x = \{x_i\}$. Equivariance implies that the image of the tube $x\mathcal{T}(\epsilon)$ under $h$ is the collection of tubes $\{x_i\mathcal{T}(\epsilon)\}$. For any $t \in \mathcal{T}(\epsilon)$, we have, then

$$\lambda_h \psi(xt) = \sum_{i=1}^{N} \psi(x_i t),$$

where $\lambda_h$ is so that $h.\psi = \lambda_h \psi$.

Squaring, applying Cauchy-Schwarz and integrating over $t \in \mathcal{T}(\epsilon)$ gives:

$$\begin{aligned} \mu_\psi(x\mathcal{T}(\epsilon)) &\leq \frac{N}{|\lambda_h|^2} \sum_{i=1}^{N} \mu_\psi(x_i\mathcal{T}(\epsilon)) \\ &\leq \frac{N}{|\lambda_h|^2} \max_i \#\left\{j | x_i\mathcal{T}(\epsilon) \cap x_j\mathcal{T}(\epsilon) \neq \emptyset\right\}. \end{aligned} \tag{1.3}$$

If the tubes $x_i\mathcal{T}(\epsilon)$ are *disjoint*[3] and, furthermore, $|\lambda_h|$ is "large" (w.r.t. $N$), (1.3) yields a good upper bound for $\mu_\psi(x\mathcal{T}(\epsilon))$.

The issue of choosing $h$ so that $\lambda_h$ is "large" turns out to be relatively minor. The solution is given in Lemma 5.2 and proved in the Appendix. In essence, this amounts to "amplification on a higher rank group."

The more serious problem is that the tubes $x_i\mathcal{T}(\epsilon)$ might not be disjoint. It must be emphasized that this issue is not a technical artifact of the proof but related fundamentally to the analytic properties of eigenfunctions on such locally symmetric spaces; the peculiar behavior of sizes of eigenfunctions (cf. the Rudnick-Sarnak example of a sequence of eigenfunctions on a hyperbolic 3-manifold with large $L^\infty$-norms) is in fact connected to precisely phenomena of this nature.

There are two natural ways to continue:

(1) Choose a subset of translates $\{x_i\}$ for which we can prove an analogue of (1.3) and such that the tubes $x_i\mathcal{T}(\epsilon)$ are disjoint.
(2) Strengthen (1.3) and choose an appropriate $h$ for which the multiplicities of intersections are not too high *on average* over the translates.

In fact, both of these ideas can be succesfully implemented in our setting. The first technique was the one used in the work of Rudnick-Sarnak and Bourgain-Lindenstrauss. The quantitative version of Bourgain-Lindenstrauss required a sieving argument. Our original proof was based on a further refinement of this technique, which avoided sieves entirely by using some geometry of buildings. A presentation of that proof may be found in the PhD thesis of the first author, [10]. In this paper we shall present the second technique, which has not appeared in print.

As it turns out, the second technique – when carefully implemented – has very significant advantages over the first. It require much weaker multiplicity bounds. This enables one to avoid sieving (we also obtain considerably better quantitative bounds but this seems rather pointless at present). The key to this implementation is Lemma 3.4, which replaces the "max" of (1.3) by a suitable *average* intersection multiplicity.

1.3. **Spectrum of quotients. Significance of division algebras.** More generally, the second technique can be interpreted as an implementation of the following philosophy:

*The analytic behavior of Hecke eigenfunctions on $\Gamma\backslash G$ along orbits of a subgroup $H \subset G$ is controlled by the spectrum of quotients $L^2(G_p/H_p)$.*

Here $G_p$ is the $p$-adic group corresponding to $G$, and $H_p \subset G_p$ is a $p$-adic Lie subgroup "corresponding" to $H$ in a suitable sense. In the situation of this paper, $G_p = \mathrm{PGL}_d(\mathbb{Q}_p)$ for almost all $p$, $H$ will be a real Levi subgroup, and $H_p$ will be a

[3] It suffices for the number of tubes intersecting a given one to be uniformly bounded independenly of $\epsilon$.

torus. While hints of this may already be found in the work of Burger and Sarnak, we hope to systematically develop this point of view in [9].

In this context, the possibilities for the subgroup $H_p$ that can occur are closely related to the $\mathbb{Q}$-structure of the group underlying $G$. In general, the fewer $\mathbb{Q}$-subgroups $\mathbf{G}$ has, the fewer the possibilities for $H_p$. This underlies the reason that we have only treated quotients $\Gamma\backslash G$ arising from division algebras of prime rank in the present paper: the corresponding $\mathbb{Q}$-groups have very few subgroups. As one passes to general $\Gamma\backslash G$, the possibilities for $H_p$ become wilder, and eventually the methods of this paper do not seem to give much information.

1.4. **Organization of this paper.** In Section 2 we describe our setup in the general setting of algebraic groups. Further notation regarding our special case of division algebras is discussed in Section 4.1.

Section 3 contains the derivation of our main techical result, Lemma 3.4, giving a bound for the integral on a small set of the squared modulus of an eigenfunction of an equivariant correspondence. This is a version of (1.3) which can be used for non-disjoint translates. The bound we obtain depends on the average multiplicity of intersection among the translates of the tube as well as on covering properties of the tubes (easily understood in natural applications).

In Section 4 we define the kind of tubes we shall be interested in. Specializing to the case of division algebras we analyze the intersection pattern of the translates of a small tube by an element of the Hecke algebra. Using a diophantine argument we show that under suitable hypotheses the intersection pattern is controlled by a number field embedded in the division algebra. Here, the fact that we are dealing with (the units of) a division algebra and not a general algebraic group is important.

Section 5 then obtains the desired power-law decay of the mass of small tubes. Although the considerations of this section are fairly general, we remain in the context of division algebras for clarify of exposition.

Finally in Section 6 we give recall our previous result ("step 1" of the strategy) and prove our main Theorem.

1.5. **Acknowledgements.** This paper owes a tremendous debt both to Peter Sarnak and Elon Lindenstrauss. It was Sarnak's realization, developed throughout the 1990s, that the quantum unique ergodicity problem on arithmetic quotients was a question that had interesting structure and interesting links to the theory of $L$-functions; it was Lindenstrauss' paper [6] which introduced ergodic-theoretic methods in a decisive way. Peter and Elon have both given us many ideas and comments over the course of this work, and it is a pleasure to thank them.

The second author was supported by a Clay Research Fellowship, and NSF Grant DMS–0245606.

Both authors were partly supported by NSF Grant DMS–0111298; they would also like to thank the Institute for Advanced Study for providing superb working conditions.

## 2. Notation

We shall specify here the "general" notation to be used throughout the paper.

Later sections (after Section 3) will use, in addition to these notations, certain further setup about division algebras. This will be explained in §4.1.

Let $\mathbf{G}$ be a anisotropic[4] semisimple group over $\mathbb{Q}$. We choose an embedding $\rho : \mathbf{G} \to \mathrm{SL}_n$.

Let $G = \mathbf{G}(\mathbb{R})$, $G = NAK$ a Cartan decomposition, $K_{\mathrm{f}} \subset \mathbf{G}(\mathbb{A}_{\mathrm{f}})$ an open compact subgroup with the property that $\rho(K_{\mathrm{f}}) \subset \prod_p \mathrm{SL}_n(\mathbb{Z}_p)$. We set $X = \mathbf{G}(\mathbb{Q})\backslash\mathbf{G}(\mathbb{A})/K_{\mathrm{f}}, Y = X/K$. They are both compact, by virtue of the assumption that $\mathbf{G}$ is anisotropic. We let $d\mathrm{vol}$ be the natural probability measures on either $X$ or $Y$: in both cases, the projection of the $\mathbf{G}(\mathbb{A})$-invariant probability measure on $\mathbf{G}(\mathbb{Q})\backslash\mathbf{G}(\mathbb{A})$.

For any subset $S \subset G$ or $S \subset \mathbf{G}(\mathbb{A})$, we denote by $\overline{S}$ the image of $S$ in $X$ under the natural projections $G \to X$ and $G(\mathbb{A}) \to X$.

For $\gamma \in \mathbf{G}(\mathbb{Q})$, we set $\mathrm{denom}(\gamma)$ to be the least common multiple of the denominators of matrix entries of $\rho(\gamma)$. For $g_p \in \mathbf{G}(\mathbb{Q}_p)$, we define in a similar way $\mathrm{denom}_p(g_p) \in p^{\mathbb{N}}$. This extends to the group $\mathbf{G}(\mathbb{A}_{\mathrm{f}})$ in a natural way: $\mathrm{denom}(g_{\mathrm{f}}) = \prod_p \mathrm{denom}_p(g_p)$ for $g_{\mathrm{f}} \in \mathbf{G}(\mathbb{A}_{\mathrm{f}})$. The function denom on $\mathbf{G}(\mathbb{A}_{\mathrm{f}})$ is left- and right- invariant under $K_{\mathrm{f}}$, in view of our choices.

We normalize the Haar measures $dx$ on $X$, $dk$ on $K$ and $dy$ on $Y$ to have total mass 1 (here $dy$ is the pushforward of $dx$ under the the projection from $X$ to $Y$).

Fixing a model of $\mathbf{G}$ that is smooth and semisimple over $\mathbb{Z}[1/N]$, we say a prime $p$ is *good* if $p$ does not divide $N$, the projection $K_p$ of $K_{\mathrm{f}}$ to $\mathbf{G}(\mathbb{Q}_p)$ coincides with $\mathbf{G}(\mathbb{Z}_p)$, and $K_{\mathrm{f}}$ contains $K_p$ (using the natural embedding $\mathbf{G}(\mathbb{Q}_p) \hookrightarrow \mathbf{G}(\mathbb{A}_{\mathrm{f}})$). Then all but finitely many $p$ are *good*. If $p$ is not good, we say it is *bad*.

We denote by $\mathcal{H}$ the Hecke algebra of $\prod_{p\ good} K_p$ bi-invariant functions on $\prod_{p\ good} \mathbf{G}(\mathbb{Q}_p)$. It forms a commutative algebra under convolution. It acts in a natural way on functions on $X$. We can identify an element of $\mathcal{H}$ with a $K_{\mathrm{f}}$-invariant function on $\mathbf{G}(\mathbb{A}_{\mathrm{f}})/K_{\mathrm{f}}$ which is, moreover, supported on $K_{\mathrm{f}} \cdot \prod_{p\ good} \mathbf{G}(\mathbb{Q}_p)$. We shall abbreviate the latter condition to: *supported at good primes.*

If $\mathbf{H} \subset \mathbf{G}$ is a $\mathbb{Q}$-subgroup, we say a prime $p$ is $\mathbf{H}$-good if it is good for $\mathbf{G}$ and moreover $\mathbf{H}(\mathbb{Q}_p) \cap K_p$ is a maximal compact subgroup of $\mathbf{H}(\mathbb{Q}_p)$.

We fix compact subsets $\Omega_\infty \subset G, \Omega \subset \mathbf{G}(\mathbb{A})$.

**Definition 2.1.** We call $\psi \in L^2(X)$ a *Hecke eigenfunction* if it is a joint eigenfunction of the Hecke algebra $\mathcal{H}$. We set $\mu_\psi = c|\psi|^2 d\mathrm{vol}$, where the constant $c$ is chosen so that $\mu_\psi$ is a probability measure on $X$.

In §3 we will deal with an abstract $\mathbf{G}$ as above. In §4 — §6 we specialize to the case of $\mathbf{G}$ arising from a division algebra of prime rank, $D$. The extra notation necessary for this specialization will be set up in §4.1.

**Notational convention:** We shall allow the implicit constants in the notation $\ll$ and $O(\cdot)$ to depend on $\mathbf{G}$ and the data $\rho, K_{\mathrm{f}}, \Omega, \Omega_\infty$ and the choice of smooth model given above, *without explicit indication.* In other words, the notation $A \ll B$ means that there exists a constant $c$, which may depend on all the data specified in this section, so that $A \leq cB$. In Section 4.1 we will set up some further data concerning division algebras, and, as we specify there, we shall allow implicit constants after that point to depend on these extra data also.

## 3. Bounds on the mass of tubes

[4]This assumption is not really necessary; for our purposes, we only need to work in compact subsets of the quotient $\mathbf{G}(\mathbb{Q})\backslash\mathbf{G}(\mathbb{A})$. However, it will simplify matters at points, and since our eventual application is only to this case, we impose it from the start.

3.1. **A covering argument.** Let $B_0 \subset \Omega_\infty$ be an open set containing the identity. We set

$$B := B_0 \cdot B_0^{-1}, B_2 = B \cdot B, B_3 = B \cdot B \cdot B, \dots \tag{3.1}$$

In this section, we shall discuss estimating from above $\mu_\psi(\overline{xB})$ for $x \in G$. [5] We will repeatedly use the following fact. If $S \subset G$ is a compact subset and $x \in X$ is arbitrary, the cardinality of the fibers of the map $S \mapsto X$ defined by $s \mapsto x.s$ is bounded in a fashion that depends on $S$ but not on $x$. This is an immediate consequence of the compactness of $X$: indeed, there exists an open neighbourhood $U$ of the identity in $G$ so that $u \in U \mapsto xu \in X$ is injective for any $x \in X$. In particular, we may apply this remark when $S = B, B_2, B_3$. The size of the relevant fibers will therefore be bounded by a constant depending on $\Omega_\infty$.

**Lemma 3.1.** *There exists a covering of $X$ by translates $x_\alpha B$ so that any set $zB$, for $z \in X$, intersects at most $\frac{\mathrm{vol}B_3}{\mathrm{vol}B_0}$ of the $x_\alpha B$.*

*Proof.* Choose a maximal subset $\{x_\alpha\} \subset X$ with the property that $x_\alpha B_0$ are disjoint.

Given $z \in X$, we must have $zB_0 \cap x_\alpha B_0 \neq \emptyset$ for some $\alpha$. This means precisely that $X = \bigcup x_\alpha B$.

Next, fix $z \in X$. For any $\alpha$ so that $x_\alpha B \cap zB \neq \emptyset$, we may choose $\varpi_\alpha \in B \cdot B^{-1}$ so that $x_\alpha = z\varpi_\alpha$. Then the sets $z\varpi_\alpha B_0 \subset X$ are all disjoint.

A necessary condition for this is that the sets $\varpi_\alpha B_0$, considered as subsets of $G$, are disjoint. Since each $\varpi_\alpha B_0$ belongs to $B_3$, their number is bounded by $\frac{\mathrm{vol}B_3}{\mathrm{vol}B_0}$. □

**Lemma 3.2.** *Let $\nu$ be a probability measure on $X$ and $y_1, \dots, y_r \in X$.*

*Then*

$$\sum_{i=1}^{r} \nu(y_i B)^{1/2} \ll \frac{\mathrm{vol}B_3}{\mathrm{vol}B_0} \left(\#\left\{(i,j) : y_i B_2 \cap y_j B_2 \neq \emptyset\right\}\right)^{1/2}$$

*Proof.* Choose a collection $x_\alpha$ as in Lemma 3.1. Each set $y_i B$ is covered by at most $\frac{\mathrm{vol}B_3}{\mathrm{vol}B_0}$ sets $x_\alpha B$. Clearly $\nu(y_i B)^{1/2} \leq \sum_{\alpha : x_\alpha B \cap y_i B \neq \emptyset} \nu(x_\alpha B)^{1/2}$, and so

$$\begin{aligned} \left(\sum_i \nu(y_i B)^{1/2}\right)^2 &\leq \left(\sum_\alpha \nu(x_\alpha B)^{1/2} \sum_{i : y_i B \cap x_\alpha B \neq \emptyset} 1\right)^2 \\ &\leq \sum_\alpha \nu(x_\alpha B) \sum_\alpha \left(\sum_{i : y_i B \cap x_\alpha B \neq \emptyset} 1\right)^2 \end{aligned} \tag{3.2}$$

Now, $\sum_\alpha \nu(x_\alpha B) \leq \frac{\mathrm{vol}B_3}{\mathrm{vol}B_0}$ because each $z \in X$ belongs to at most $\frac{\mathrm{vol}B_3}{\mathrm{vol}B_0}$ of the $x_\alpha B$.

Moreover,

$$\sum_\alpha \left(\sum_{i : y_i B \cap x_\alpha B \neq \emptyset} 1\right)^2 = \#\{i, j, \alpha : y_i B \cap x_\alpha B \neq \emptyset, y_j B \cap x_\alpha B \neq \emptyset\}$$

[5] In fact, all our estimates really work for $\mu_\psi(xB)$ for $x \in X$, as the notation is meant to suggest. We prove our results only for $x \in G$ simply to keep notation to a minimum. If $G$ acts with a single orbit on $X$, as we assume in our final applications anyway, there is no difference at all.

For given $i, j$, the pertinent set of $\alpha$ is nonzero only if $y_i B_2 \cap y_j B_2 \neq \emptyset$. If it is nonempty, it has size at most $\mathrm{vol}(B_3)/\mathrm{vol}(B_0)$. □

3.2. **General bound.** Let $\psi$ be a Hecke eigenfunction on $X$, $\mu_\psi$ the associated probability measure.

Let $h \in \mathcal{H}$, which we can think of as a function $s \mapsto h_s$ on $\mathbf{G}(\mathbb{A}_{\mathrm{f}})/K_{\mathrm{f}}$, supported on good primes. Let $S \subset \mathbf{G}(\mathbb{A}_{\mathrm{f}})/K_{\mathrm{f}}$ be the support of $h$.

Because $\psi$ is a Hecke eigenfunction, there is $\Lambda_h \in \mathbb{C}$ so that:

$$\Lambda_h \psi(x) = \sum_{s \in S} h_s \psi(x.s) \quad (x \in G) \tag{3.3}$$

Note that $x.s \in \mathbf{G}(\mathbb{A})/K_{\mathrm{f}}$ and therefore $\psi(x.s)$ makes sense. The following Lemma bounds $\mu_\psi(\overline{xB})$ in terms of the average mass of certain Hecke translates of $\overline{xB_2}$.

**Lemma 3.3.** *Let $x \in G$. Then*

$$\mu_\psi(\overline{xB}) \ll \frac{\left(\sum_{s \in S} |h_s| \mu_\psi(\overline{xB}.s)^{1/2}\right)^2}{|\Lambda_h|^2}.$$

*Proof.* This follows by squaring out equation (3.3), integrating over $B$ and applying Cauchy-Schwarz. We use the fact that $b \in B \mapsto x.s.b \in X$ has fibers whose cardinality is bounded in terms of $\Omega_\infty$ (see discussion in §3.1). □

The next Lemma clarifies that the only necessary input to bound $\mu_\psi(\overline{xB})$ is an estimate for the average intersection multiplicity of Hecke translates of $\overline{xB}$.

**Lemma 3.4.** *Suppose that $h$ is supported on $S \subset \mathbf{G}(\mathbb{A}_{\mathrm{f}})/K_{\mathrm{f}}$ and $|h| \leq 1$.*

*Then for $x \in G$*

$$\mu_\psi(\overline{xB}) \ll \left(\frac{\mathrm{vol}B_3}{\mathrm{vol}B_0}\right)^2 |\Lambda_h|^{-2} \left(\#\{s, s' \in S : \overline{xB}_2.s \cap \overline{xB}_2.s' \neq \emptyset\}\right) \tag{3.4}$$

*Proof.* This follows from the prior Lemma and Lemma 3.2 (applied to the set $\{y_i\} = \{xs : s \in S\} \subset X$). □

Lemma 3.4 is our key technical Lemma. Setting $N = \#S$ as in the introduction we may rewrite the right-hand-side as:

$$\left(\frac{\mathrm{vol}B_3}{\mathrm{vol}B_0}\right)^2 \frac{N}{|\Lambda_h|^2} \cdot \frac{1}{N} \sum_{s \in S} \#\left\{s' \in S : \overline{xB}_2.s \cap \overline{xB}_2.s' \neq \emptyset\right\}.$$

The key feature of (3.4) then becomes apparent. The right-hand side now depends only on an "average" intersection number of Hecke translates, not on a "worst case" intersection number like $\sup_s \#\{s' : \overline{xB}_2.s \cap \overline{xB}_2.s' \neq \emptyset\}$ as in (1.3). Indeed, [1] relied on controlling this latter quantity. Both this and the case of necessity of dealing with split tori (see corrigendum to that paper) required the use of the sieve.

## 4. A Diophantine Lemma

In this section, we shall prove a Lemma that will be used to control the intersections for the present case of interest: when $\mathbf{G}$ arises from a division algebra of prime degree and $B$ is a tube around a piece of a Levi subgroup. Roughly, we show that all the $\gamma \in \mathbf{G}(\mathbb{Q})$ that lie "very close" to a Levi subgroup (where "very close"

really means "very close relative to the denominator of $\gamma$") must all lie on a single $\mathbb{Q}$-torus.

Using this we will show that the intersection pattern of Hecke translates of tubes of this type are controlled by the torus.

An argument of this type (in the case $\dim_{\mathbb{Q}} D = 4$) already appears in [1]; this argument does not suffice for the higher rank case, however, because [1] uses commutativity of the relevant Levi subgroups in an important way.[6] Although we shall use fairly *ad hoc* arguments for concreteness, there exist very soft and general arguments for this type of problem, which we hope to develop further in [9].

4.1. **Extra notations for Sections 4 – 6.** Let $D$ be a division algebra over $\mathbb{Q}$ of prime degree $d$, split over $\mathbb{R}$, and fix a lattice $D_{\mathbb{Z}} \subset D$ (i.e. a free $\mathbb{Z}$-submodule of maximal rank) and a Euclidean norm $\|\cdot\|$ on $D \otimes_{\mathbb{Q}} \mathbb{R}$. In other words, we have chosen a norm on $D \otimes \mathbb{Q}_v$ for all $v$, finite or infinite.

Let $\mathbf{G}$ the projectivized group of units (=invertible elements) of $D$, so $G = \mathbf{G}(\mathbb{R})$ is isomorphic to $\mathrm{PGL}_d(\mathbb{R})$. The Lie algebra of $G$ is identified with a quotient of $D \otimes \mathbb{R}$; as such, the norm on $D \otimes \mathbb{R}$ gives rise to a norm on the Lie algebra of $G$ and thus to a left-invariant Riemannian metric on $G$.

We fix extra data ($\rho, K_{\mathrm{f}}, \Omega, \Omega_\infty$ and a smooth model) for the group $\mathbf{G}$, as discussed in §2. In the rest of this paper, the implicit constants in the notations $\ll$ and $O(\cdot)$ will be allowed to depend on $D, D_{\mathbb{Z}}$, the norm $\|\cdot\|$ and this extra data, without explicitly indicating this.

It should be noted that we do not assume that $D_{\mathbb{Z}}.D_{\mathbb{Z}} \subset D_{\mathbb{Z}}$; on the other hand, clearly there is an integer $K = O(1)$ so that $D_{\mathbb{Z}}.D_{\mathbb{Z}} \subset K^{-1}D_{\mathbb{Z}}$.

4.2. **Diophantine lemma.** For $x \in D$, we set

$$\mathrm{denom}(x) := \inf\{m \in \mathbb{N} : \ mx \in D_{\mathbb{Z}}\}. \tag{4.1}$$

**Lemma 4.1.** *Let $S \subset D \otimes \mathbb{R}$ be a proper $\mathbb{R}$-subalgebra.*

*For $c > 0$ sufficiently large (in fact, depending only on $d$) and for $c' > 0$ sufficiently small (in fact, depending only on $D, D_{\mathbb{Z}}, \|\cdot\|$), the set of $x \in D$ satisfying*

$$\|x\| \le R, \inf_{s \in S} \|x - s\| \le \varepsilon, \mathrm{denom}(x) \le M \tag{4.2}$$

*is contained in a proper subalgebra $F \subset D$ as long as*

$$\varepsilon R^c M^c < c' \tag{4.3}$$

In other words: points of $D$ near a proper subalgebra of $D \otimes \mathbb{R}$ lie on a proper $\mathbb{Q}$-subalgebra of $D$. This proof will not use the fact that $D$ is a *division* algebra, nor the fact that it is of prime rank.

*Proof.* We use here $f_1(d), f_2(d), \dots$ to denote positive quantities that depend on the rank $d$ alone.

Let $s = \dim(S) + 1$. Then there is a polynomial function $G : D^s \to \mathbb{Q}$, with integral coefficients with respect to $D_{\mathbb{Z}}$, so that $G(\alpha_1, \dots, \alpha_s) = 0$ exactly when

[6] Here is a toy model of the type of reasoning we use in what follows. Let $\ell$ be a line segment in $\mathbb{R}^2$ of length 1. Suppose $P_i = (x_i, y_i)$, for $1 \le i \le 3$, are points in $\mathbb{R}^2$ with rational coordinates all of which lie within $\varepsilon$ of $\ell$, and let $M$ be an upper bound for the denominators of all $x_i, y_i$. Then, if $\varepsilon < \frac{1}{10}M^{-6}$, the $P_i$ are themselves co-linear. Indeed, the area of the triangle formed by $P_1, P_2, P_3$ is a rational number with denominator $\le 2M^6$. On the other hand the area of this triangle is $\le 2\varepsilon$, whence the conclusion.

$\alpha_1, \dots, \alpha_s$ span a linear space of dimension $\leq s-1$. For example one may use the sum of the squares of the minors of a suitable matrix. The degree of $G$ is $f_1(d)$ and the size of its coefficients is $O(1)$.

Take $x_1, \dots, x_s$ belonging to the set defined by (4.2). There are $y_1, \dots, y_s \in S$ so that $\|x_i - y_i\| \leq \varepsilon$. Then $G(x_1, \dots, x_s) \ll R^{f_2(d)}\varepsilon$. On the other hand, if $G(x_1, \dots, x_s) \neq 0$ then, because $\mathrm{denom}(x_i) \leq M$, we must have $G(x_1, \dots, x_s) \gg M^{-f_3(d)}$. It follows that, if a condition of the type (4.3) holds for suitable $c, c'$ as stated, then $x_1, \dots, x_s$ span a $\mathbb{Q}$-linear space of dimension $s-1$.

Now let $X$ be the $\mathbb{Q}$-algebra spanned by those $x$ satisfying (4.2). It is clear that $X$ is, in fact, spanned by monomials in such $x$ of length at most $\dim_{\mathbb{Q}} D$. Each such monomial $y$ satisfies $\|y\| \ll R^{f_4(d)}, \inf_{s\in S}\|y - s\| \ll R^{f_5(d)}\varepsilon, \mathrm{denom}(y) \gg M^{f_6(d)}$. It follows that – increasing $c$ and decreasing $c'$ in (4.3) as necessary – it follows that the $\mathbb{Q}$-subalgebra generated by all solutions to (4.2) has dimension $\leq s-1$, in particular, is a proper subalgebra of $D$. □

We have defined two notions of denominator associated to this setup: for $\alpha \in D^\times$, we have just defined a denominator w.r.t. the lattice $D_{\mathbb{Z}}$ (see (4.1)). For $\gamma \in \mathbf{G}(\mathbb{Q}) = D^\times/\mathbb{Q}^\times$, we have the denominator $\mathrm{denom}(\gamma)$ defined in Section 2. We now clarify the relation between the two notions.

**Lemma 4.2.** *Let $\gamma \in \mathbf{G}(\mathbb{Q}) = D^\times/\mathbb{Q}^\times$ satisfy $\mathrm{denom}(\gamma) \leq M$ and belong to a compact subset $E \subset \mathbf{G}(\mathbb{R})$. Then there exists $\alpha \in D^\times$ lifting $\gamma$ so that:*

$$\mathrm{denom}(\alpha) \ll_E M^c, \|\alpha^{-1}\|, \|\alpha\| \ll_E 1$$

*where $c$ is a constant depending only on the isomorphism class of $\mathbf{G}$.*

*Proof.* In fact, let $\tilde{\mathbf{G}}$ be the algebraic group corresponding to $D^\times$, i.e. $\tilde{\mathbf{G}}(R) = (D \otimes_{\mathbb{Q}} R)^\times$ if $R$ is a ring containing $\mathbb{Q}$.

Then $\tilde{\mathbf{G}}$ and $\mathbf{G}$ are affine algebraic groups. Moreover, the map $\tilde{\mathbf{G}} \to \mathbf{G}$ *admits an algebraic section over a Zariski-open set $U \subset \mathbf{G}$*, as follows from Hilbert's theorem 90.[7]

One may, by translating $U$, find a finite collection of open sets $U_1, \dots, U_h$ which cover $\mathbf{G}$, and so that $\tilde{\mathbf{G}} \to \mathbf{G}$ admits a section $\theta_j : U_j \to \tilde{\mathbf{G}}$ over each $U_j$.

It follows from this that there exists $\alpha \in D^\times$ lifting $\gamma$ so that

$$\mathrm{denom}(\alpha), \mathrm{denom}(\alpha^{-1}) \ll M^c$$

where $c$ is a constant depending only on the choice of sets $U_j$ and the sections, i.e. only the isomorphism class of $\mathbf{G}$.

From this bound, it follows in particular that $M^{-c} \ll_E \|\alpha\| \ll_E M^c$. The lower bound is clear; for the upper bound, we use the fact that $\alpha$ projects to the compact subset $E \subset \mathbf{G}(\mathbb{R})$.

---

[7] Let $\mathbf{G}^{(1)}$ denote the group of elements of norm 1 in $D^\times$. It is a (geometrically) irreducible variety, because $\mathrm{SL}_n$ is an irreducible variety. The map $\mathbf{G}^{(1)} \to \tilde{\mathbf{G}}$ is a covering map (i.e. étale) and its kernel is the group of $d$th roots of unity. Let $E$ be the function field of $\mathbf{G}$, considered as $\mathbb{Q}$-variety. The generic point in $\eta \in \mathbf{G}(E)$ does not lift to a point of $\mathbf{G}^{(1)}(E)$, but it does at least to lift to a point of $\tilde{\eta} \in \mathbf{G}^{(1)}(\tilde{E})$ for some finite extension $\tilde{E}/E$, which we may assume to be Galois and to contain the $d$th roots of unity. Then $\sigma \mapsto \tilde{\eta}^\sigma/\tilde{\eta}$ defines a 1-cocycle of $\mathrm{Gal}(\tilde{E}/E)$ valued in the group of $d$th roots of unity. By Hilbert's theorem 90, there exists $\tilde{e} \in \tilde{E}$ so that this cocycle is $\sigma \mapsto \tilde{e}^\sigma/\tilde{e}$. Adjusting $\tilde{\eta}$ by $\tilde{e}$ gives a $\tilde{E}$-valued point of $\tilde{\mathbf{G}}$, which is invariant under $\mathrm{Gal}(\tilde{E}/E)$ and therefore is indeed an $E$-valued point of $\tilde{\mathbf{G}}$. This gives the desired section.

Let $p/q$ be a rational number satisfying $\|\alpha\| < p/q < 2\|\alpha\|$. We may choose $p, q$ so that $\max(p,q) \ll M^c$. Replacing $\alpha$ by $q\alpha/p$, we obtain a representative $\alpha$ for $\gamma$ that satisfies:

$$\mathrm{denom}(\alpha) \ll_E M^{2c}, \|\alpha\| \asymp_E 1$$

We increase $c$ as necessary.

Finally, the bound for $\|\alpha^{-1}\|$ follows from the bound for $\|\alpha\|$ together with the fact that $\alpha$ projects to the compact set $E \subset \mathbf{G}(\mathbb{R})$. □

4.3. **Tubes around Levi subgroups and the intersection pattern of their translates.** Let $A$ be the subgroup of diagonal matrices in $G$. Let $a \in A$ be nontrivial. We fix a compact neighbourhood $C$ of the identity in the centralizer $Z_G(a)$ of $a$ inside $G$, and let $B(C,\varepsilon)$ be a $\varepsilon$-neighbourhood of $C$ inside $G$. We shall bound the $\mu_\psi$-mass of sets of the form $\overline{xB(C,\varepsilon)} \subset X$.

Recall that we have fixed compact subsets $\Omega \subset \mathbf{G}(\mathbb{A})$ and $\Omega_\infty \subset G$.

We will assume that $B(C,\varepsilon) \subset \Omega_\infty$. In particular, $C \subset \Omega_\infty$. Consequently, in view of the convention discussed in §2, we shall not explicitly indicate that implicit constants in $\ll$ or $O(\dots)$ depend on $C$.

**Lemma 4.3.** *There is $c_2, c_4 > 0$, depending only on the isomorphism class of $\mathbf{G}$ and $c_1, c_3 = O_C(1)$, so that*

*For any $x = (x_\infty, x_\mathrm{f}) \in \Omega$ and any $0 < \varepsilon < 1/2$ there exists a subfield $F \subset D$ so that:*

(1) *If $s, s' \in \mathbf{G}(\mathbb{A}_\mathrm{f})$ both have denominator $\leq c_1\varepsilon^{-c_2}$ and are so that*

$$x\overline{B(C,\varepsilon)}s \cap \overline{xB(C,\varepsilon)}s' \neq \emptyset \text{ in } \mathbf{G}(\mathbb{Q})\backslash\mathbf{G}(\mathbb{A})/K_\mathrm{f}$$

*then there exists $\gamma \in F^\times/\mathbb{Q}^\times \subset \mathbf{G}(\mathbb{Q})$ with*

$$\gamma xB(C,\varepsilon)s \cap xB(C,\varepsilon)s' \neq \emptyset \text{ in } \mathbf{G}(\mathbb{A})/K_\mathrm{f}. \tag{4.4}$$

(2) *$F$ is generated by $\alpha \in D^\times$, so that $\alpha D_\mathbb{Z} + D_\mathbb{Z}\alpha \subset D_\mathbb{Z}$, and with $\|\alpha\| \leq c_3\varepsilon^{-c_4}$.*

*Proof.* By assumption there is – after replacing $s, s'$ by suitable elements of $sK_\mathrm{f}$ and $s'K_\mathrm{f}$ respectively – an element $\gamma \in \mathbf{G}(\mathbb{Q})$ so that $\gamma x_\mathrm{f} s = x_\mathrm{f} s'$ (equality in $\mathbf{G}(\mathbb{A}_\mathrm{f})$) and, moreover, $\gamma \in x_\infty B(C,\varepsilon)B(C,\varepsilon)^{-1}x_\infty^{-1}$ (equality in $\mathbf{G}(\mathbb{R})$).

The former inclusion implies, in particular, that $\mathrm{denom}(\gamma) \ll c_1'\varepsilon^{-c_2'}$, where $c_i'$ depends on $c_i$, and $c_i' \to 0$ as $c_i \to 0$ for $i = 1, 2$.

The latter inclusion shows that $\gamma$ lies in a fixed compact subset of $\mathbf{G}(\mathbb{R})$ depending only on $\Omega, C$.

The element $\gamma$ belongs to $D^\times/\mathbb{Q}^\times$. We may choose a representative $\alpha \in D^\times$ for $\gamma$ as in Lemma 4.2. In that case $\alpha$ lies in a fixed compact subset of $(D \otimes \mathbb{R})^\times$ – depending only on $\Omega, C$ – and $\mathrm{denom}(\alpha) \ll c_1''\varepsilon^{-c_2''}$, where (for $i = 1, 2$) $c_i''$ depends on $c_i$ and $c_i'' \to 0$ as $c_i' \to 0$.

Let $E$ be the subalgebra of $D \otimes \mathbb{R}$ that centralizes $x_\infty a x_\infty^{-1}$. The assertion that $\gamma \in x_\infty B(C,\varepsilon)B(C,\varepsilon)^{-1}x_\infty^{-1}$ shows that $\alpha$ is "close" to $E$; in fact, it is clear that

$$\inf_{e \in E} \|\alpha - e\| \ll \varepsilon$$

and moreover $\|\alpha\| \ll 1$ (because $\alpha$ lies in a fixed compact subset of $(D \otimes \mathbb{R})^\times$).

By Lemma 4.1 we see that, if $c_1'', c_2''$ are sufficiently small – this occurs, in particular, if $c_1, c_2$ are sufficiently small – then all such $\alpha$ necessarily belong to a proper $\mathbb{Q}$-subalgebra of $D$; because $D$ has prime degree, this must be a field $F$. Analyzing

this reasoning shows that $c_2$ may be taken to depend only on the isomorphism class of $\mathbf{G}$, whereas $c_1 = O(1)$. This proves (4.4).

Also, there exists $K \in \mathbb{N}$ so that $D_{\mathbb{Z}}.D_{\mathbb{Z}} \subset K^{-1}D_{\mathbb{Z}}$. Then $\alpha.D_{\mathbb{Z}} \subset K^{-1}.\mathrm{denom}(\alpha)^{-1}.D_{\mathbb{Z}}$ and similarly for $D_{\mathbb{Z}}.\alpha$.

Replacing $\alpha$ with $\alpha' = K.\mathrm{denom}(\alpha).\alpha$, we see that $\alpha' D_{\mathbb{Z}} + D_{\mathbb{Z}}\alpha' \subset D_{\mathbb{Z}}$ and $\|\alpha'\| \leq c_3 \varepsilon^{-c_4}$, where $c_3 = O_C(1)$ and $c_4$ depends only on the isomorphism class of $\mathbf{G}$. □

We next show that there are only a few bad primes.

**Lemma 4.4.** *Let $F \subset D$ be a subfield, and let $\mathbf{T}_F \subset \mathbf{G}$ be the torus defined by $F$, i.e. the centralizer of $F$ in $\mathbf{G}$. Suppose that $F$ is generated, over $\mathbb{Q}$, by an element $\alpha$ satisfying $\alpha D_{\mathbb{Z}} + D_{\mathbb{Z}}\alpha \subset D_{\mathbb{Z}}$. Then the number of primes which fail to be $\mathbf{T}_F$-good is at most $O(1 + \log \|\alpha\|)$.*

*Proof.* Let $K'_{\mathrm{f}}$ be the stabilizer[8] of $D_{\mathbb{Z}}$ inside $\mathbf{G}(\mathbb{A}_{\mathrm{f}})$.

Then $\mathbf{T}_F(\mathbb{Q}_p) \cap K'_{\mathrm{f}}$ is maximal compact inside $\mathbf{T}_F(\mathbb{Q}_p)$ as long as the maximal compact subring of $(F \otimes \mathbb{Q}_p)$ preserves $D_{\mathbb{Z}} \otimes \mathbb{Z}_p$ under both left and right multiplication. This will be so, in particular, at any prime where the maximal compact subring of $(F \otimes \mathbb{Q}_p)$ equals $\mathbb{Z}_p[\alpha]$. This will always be the case if $p$ does not divide the discriminant of the ring $\mathbb{Z}[\alpha]$.

From this, one deduces that there at most $O(1 + \log \|\alpha\|)$ primes for which $\mathbf{T}_F(\mathbb{Q}_p) \cap K'_{\mathrm{f}}$ fails to be maximal compact in $\mathbf{T}_F(\mathbb{Q}_p)$. But, for all but $O(1)$ primes, the intersection $\mathbf{G}(\mathbb{Q}_p) \cap K'_{\mathrm{f}}$ coincides with $\mathbf{G}(\mathbb{Q}_p) \cap K_{\mathrm{f}}$. So there are at most $O(1 + \log \|\alpha\|)$ primes for which $\mathbf{T}_F(\mathbb{Q}_p) \cap K_{\mathrm{f}}$ fails to be maximal compact in $\mathbf{T}_F(\mathbb{Q}_p)$. □

## 5. Bounds on the mass of tubes, II

Notations continue as in §4.1. In particular, $\mathbf{G}$ is the projective group of units of a division algebra $D$.

It should be noted that the considerations of the present section do not depend on this assumption in an essential way, i.e., they carry through for $\mathbf{G}$ semisimple. On the other hand, the results of this section are based on an assumption $\star$ (stated below) which we can in any case verify only in special cases, including the case of §4.1. Therefore we have restricted to this special case from the start.

We define "tubes" $B_0 := B(C, \varepsilon)$ as in Section 4.3. Set $B = B_0 B_0^{-1} \supset B(C, \varepsilon)$. There exists a compact subset $C' \subset Z_G(a)$ and a constant $M = O(1)$ such that $B$, $B_2$ $(= B_1.B_1)$ and $B_3$ are subsets of $B(C', M\varepsilon)$. Here notations are as (3.1). Also, $\frac{\mathrm{vol}B_3}{\mathrm{vol}B_0} \ll 1$.[9]

5.1. **Sets $S$ for which intersections are controlled by tori.** Let $Q$ be so that $Q/2$ is larger than any bad prime for $\mathbf{G}$, and let $\ell$ be fixed. (In practice, $Q \to \infty$ as $B$ grows small, whereas $\ell$ is fixed depending only on $\mathbf{G}$).

[8] Note that $\mathbf{G}(\mathbb{A}_{\mathrm{f}})$ acts naturally on lattices inside $D$, in a fashion derived from the conjugation action of $\mathbf{G}$ on $D$. Indeed, if $V$ is a $\mathbb{Q}$-vector space, the group $\mathrm{GL}(V \otimes_{\mathbb{Q}} \mathbb{A}_{\mathrm{f}})$ acts naturally on lattices inside $V$.

[9] In this and in the statement $M = O(1)$, the implicit constant certainly depends on $C$; however, $C$ was assumed to belong to $\Omega_\infty$, and we have permitted implicit constants to depend on $\Omega_\infty$ without explicit mention.

Set $S_p = \{g_p \in \mathbf{G}(\mathbb{Q}_p)/K_p : \mathrm{denom}_p(g_p) \leq p^\ell\}$. Initially we shall consider the set of translates given by $\bigcup_{p\in[Q/2,Q]} S_p$, where we identify $\mathbf{G}(\mathbb{Q}_p)/K_p$ with a subset of $\mathbf{G}(\mathbb{A}_\mathrm{f})/K_\mathrm{f}$ in the natural way.

Lemma 4.3 can now be rephrased[10] as establishing (for $Q^l \ll \varepsilon^{-c_2}$) the following condition. In words, it states that *intersections between Hecke translates of $B_2$ by $\cup S_p$ all arise from a $\mathbb{Q}$-torus:*

$\boxed{\star = \star_{B,Q,\ell}}$: For any $x \in \Omega_\infty$ there is a $\mathbb{Q}$-torus $\mathbf{T} \subset \mathbf{G}$ so that

$$\overline{xB}_2 s \cap \overline{xB}_2 s' \neq \emptyset \quad \text{in } X$$

with $s, s' \in \bigcup_{p\in[Q/2,Q]} S_p$ only if there is $\gamma \in \mathbf{T}(\mathbb{Q})$ so that for these $s, s'$,

$$\gamma x B_2 s \cap x B_2 s' \neq \emptyset \text{ in } \mathbf{G}(\mathbb{A})/K_\mathrm{f}.$$

Moreover, in this case Lemma 4.3,(2) and Lemma 4.4 assert that the number of primes in $[Q/2, Q]$ which are $\mathbf{T}$-bad is very small.

### 5.2. **Bounds on intersection multiplicities, assuming $\star$.**

**Lemma 5.1.** *Suppose that condition $\star_{B,Q,\ell}$ is satisfied. Take $x \in \Omega_\infty$ and let $\mathbf{T}$ be the torus specified by $\star_{B,Q,\ell}$. Assume that $S \subset \bigcup_p S_p \subset \mathbf{G}(\mathbb{A}_f)/K_\mathrm{f}$, with the union taken over $p \in [Q/2, Q]$ which are $\mathbf{T}$-good.*

*Notations being as in Lemma 3.4, we have*

$$\#\{s, s' \in S : \overline{xB}_2.s \cap \overline{xB}_2.s' \neq \emptyset\} \ll_\ell Q^2 + |S|$$

*Proof.* Consider any intersection $\overline{xB}_2.s \cap \overline{xB}_2.s' \neq \emptyset$ in $X$ when $s, s' \in S$. This means that there is $\gamma \in \mathbf{T}(\mathbb{Q})$ so that $\gamma x B_2 s \cap x B_2 s' \neq \emptyset$ in $\mathbf{G}(\mathbb{A})/K_\mathrm{f}$.

We may assume $s \in S_p, s' \in S_q$ when $p, q$ are $\mathbf{T}$-good primes in the range $[Q/2, Q]$, and distinguish two cases according to whether $p = q$ or not.

(1) $q = p$. In this case,

$$\gamma \in (xB_2B_2^{-1}x^{-1}).\mathbf{T}(\mathbb{Q}_p).K_\mathrm{f}.$$

For any fixed $p$, the number of $K_\mathrm{f}$-cosets contained in $\mathbf{T}(\mathbb{Q}_p).K_\mathrm{f}$ satisfying $\mathrm{denom}_p \leq \ell$ is $O_\ell(1)$, as is easily seen. In fact, because $p$ is $\mathbf{T}$-good, the quotient $\mathbf{T}(\mathbb{Q}_p)/\mathbf{T}(\mathbb{Q}_p) \cap K_\mathrm{f}$ is a free abelian group of rank $\leq \dim(\mathbf{T})$. Pick generators $t_1, \ldots, t_r$ for this quotient; they generate a discrete subgroup. We need to show that the number of $(e_1, \ldots, e_r) \in \mathbb{Z}^r$ so that $\mathrm{denom}_p(t_1^{e_1} \ldots t_r^{e_r}) \leq p^\ell$ is $O_\ell(1)$. To see this, pass to an extension of $\mathbb{Q}_p$ where $\rho(t_i)$ become diagonalizable.

Therefore, $\gamma$ is an element of $\mathbf{G}(\mathbb{Q})$ so that:

(a) considered as an element of $\mathbf{G}(\mathbb{R})$, $\gamma$ belongs to

$$xB_2^{-1}B_2x^{-1},$$

which in turn is contained in a compact set depending only on $\Omega_\infty$;

(b) considered as an element of $\mathbf{G}(\mathbb{A}_\mathrm{f})$, $\gamma$ belongs to $O_\ell(1)$ right $K_\mathrm{f}$-cosets.

The number of possibilities for $\gamma$ is therefore $O_\ell(1)$.

Since $(s, s')$ is determined by $(s, \gamma)$, it follows the number of possibilities for $(s, s')$ in the case "$p = q$" is at most $O_\ell(|S|)$.

[10]Note that, in what follows, we are applying Lemma 4.3 with $B(C, \varepsilon)$ replaced by the larger set $B_2 \subset B(C', M.\varepsilon)$. It is easy to see that the stated result holds even though this larger set may not be contained in $\Omega_\infty$.

(2) $p \neq q$.

In this case, $s$ belongs to $\mathbf{T}(\mathbb{Q}_p).K_p$ and $s'$ belongs to $\mathbf{T}(\mathbb{Q}_q).K_q$. By an argument already given, the number of $K_p$-cosets contained in $\mathbf{T}(\mathbb{Q}_p).K_p$ and satisfying $\mathrm{denom}_p \leq \ell$ is $O_\ell(1)$, and similarly with $q$ replacing $p$. It follows that the number of possibilities for $(s, s')$ is $O_\ell(1)$ for *given* $p, q$.

The total number of possibilities for $(s, s')$ in the case "$p \neq q$" is therefore $O_\ell(Q^2)$.

□

5.3. **"Amplification:" Correspondences $h_p$ for which the eigenvalues are large.** We are about to invoke Lemma 3.4, using Lemma 5.1 to bound the right-hand side of (3.4). The function $h$ used in Lemma 3.4 will be a sum of functions $h_p$ on $\mathbf{G}(\mathbb{Q}_p)/K_p$.

To select the functions $h_p$ we use the following Lemma. It is the adaptation of the following fact, used extensively in the amplification method and also in [1]: if $f$ is a modular form on the upper-half plane which is an eigenfunction of the $p$-Hecke operator $T_p$, then there exists $m \in \{p, p^2\}$ so that the eigenvalue of $T_m$ on $f$ is greater than $\frac{1}{2}m^{1/2}$ in absolute value. Here, normalizations are so that $T_p$ acts on the identity function with eigenvalue $p + 1$.

**Lemma 5.2.** *There exists $\ell, c, \alpha$ and a set of primes $\mathcal{P}$ of positive density*[11] *– all depending only on the isomorphism class of $\mathbf{G}$ – with the following property.*

*Let $\psi$ be a $\mathcal{H}$-eigenfunction. For each $p \in \mathcal{P}$ there exists a $K_p$-invariant function $h_p$ supported on $S_p = \{g_p : \mathrm{denom}(g_p) \leq p^\ell\}$ and*

(1) *$|h_p| \leq 1$ and $\alpha p^c \geq \#\,\mathrm{supp}(h_p) \geq \alpha^{-1}p$, where* supp *denotes support;*
(2) *$\Lambda_{h_p}$ is positive and $\Lambda_{h_p} \geq \alpha^{-1}\,\mathrm{supp}(h_p)^{1/2}$*

*Here $\Lambda_{h_p}$ is the scalar by which $h_p$ acts on the fixed Hecke eigenfunction $\psi$.*

We note it is almost certainly possible to arrange the conclusion for all sufficiently large $p$, but we have no need of this, and doing it only for a set of primes of positive density eases the notation in proof. The proof is not too difficult, but requires some extra notation from the theory of $p$-adic groups. We have therefore postponed it to the Appendix. The set $\mathcal{P}$ is taken to be the set of primes where a fixed maximal torus inside $\mathbf{G}$ splits.

5.4. **Conclusion.**

**Lemma 5.3.** *Let $\psi$ be a Hecke eigenfunction on $X$. Suppose condition $\star_{B,Q,\ell}$ is verified, where $\ell$ is as Lemma 5.2.*

*Then, for $x \in \Omega_\infty$,*

$$\mu_\psi(\overline{xB}) \ll Q^{1.01}g^{-2},$$

*where $g$ is the number of $\mathbf{T}$-good primes in $\mathcal{P} \cap [Q/2, Q]$, where $\mathcal{P}$ is as in Lemma 5.2 and $\mathbf{T}$ is as $\star_{B,Q,\ell}$.*

*Proof.* Let $x \in \Omega_\infty$ and let $\mathbf{T}$ be a torus as in $\star_{B,Q,\ell}$. For each $\mathbf{T}$-good prime $p \in [Q/2, Q] \cap \mathcal{P}$, let $h_p$ be as in the previous Lemma.

[11] i.e. for a set of primes $\mathcal{P}$ that satisfies $\#\mathcal{P} \cap [1, Q] \sim \frac{\rho Q}{\log Q}$ for some $\rho > 0$

Now the support of $h_p$ satisfies $p \ll \#\operatorname{supp}(h_p) \ll p^c$, where $c$ depends only on $\mathbf{G}, \ell$. Therefore, by a dyadic decomposition argument, there exists $\gg_{\mathbf{G},\ell} \frac{g}{\log Q}$ $\mathbf{T}$-good primes $p \in [Q/2, Q]$ so that $\#\operatorname{supp}(h_p)$ lies in a fixed dyadic interval $[A, 2A]$, with $A \gg Q$.

In what follows, all sums over $p$ mean sums over this set of primes. Set $h = \sum_p h_p$. Then, by what we have proven in Lemma 3.4 and Lemma 5.1,

$$\mu_\psi(\overline{xB}) \ll \left(\frac{\mathrm{vol}B_3}{\mathrm{vol}B_0}\right)^2 \frac{Q^2 + \sum_p \#\operatorname{supp}(h_p)}{\left(\sum_p \#\operatorname{supp}(h_p)^{1/2}\right)^2} \ll \left(\frac{\mathrm{vol}B_3}{\mathrm{vol}B_0}\right)^2 Q^{0.001}(Q/g^2 + g^{-1}).$$

Finally, as observed at the start of this section, $\frac{\mathrm{vol}B_3}{\mathrm{vol}B_0} \ll 1$. □

We now combine this Lemma with the results of Section 4, which give conditions under which $\star_{B,Q,\ell}$ is true.

**Theorem 5.4.** *Let $\mathbf{G}$ be the projectivized group of units of an division algebra $D$ of prime degree over $\mathbb{Q}$. Let $\psi$ be a Hecke eigenfunction on $X = \mathbf{G}(\mathbb{Q})\backslash\mathbf{G}(\mathbb{A})/K_{\mathrm{f}}$.*

*Let $\Omega_\infty \subset G$ be compact and let $B(C, \varepsilon) \subset \Omega_\infty$ be a tube as in Section 4.3.*

*Then there is $c > 0$, depending only on the isomorphism class of $\mathbf{G}$ so that, uniformly over $x \in \Omega_\infty$,*

$$\mu_\psi(xB(C, \varepsilon)) \ll \varepsilon^c.$$

*Proof.* Let $x \in \Omega_\infty$.

Recall, as remarked at the start of the present section, that $B_3 \subset B(C', M.\varepsilon)$, for a suitable compact set $C'$ and a suitable constant $M$.

Lemma 4.3 applied to $B(C', M.\varepsilon)$ [12] has shown that property $\star_{B,Q,\ell}$ holds so long as

$$Q^\ell \ll \epsilon^{-c_2}, \tag{5.1}$$

where $c_2$ depends only on the isomorphism class of $\mathbf{G}$.

If $Q$ satisfies this constraint, conclusion (2) of Lemma 4.3 in combination with Lemma 4.4 shows that at most $\ll \log \varepsilon$ primes $p$ are not $\mathbf{T}$-good, where $\mathbf{T}$ is the torus occurring in the definition of $\star_{B,Q,\ell}$. We thus have $g \gg Q/\log Q$ in the notation of Lemma 5.3. That Lemma then shows that $\mu_\psi(xB(C, \varepsilon)) \ll Q^{-0.98}$.

Choosing $Q$ as large as allowable under (5.1) yields the desired result. □

## 6. The AQUE problem and the application of the entropy bound.

We now return to the AQUE problem discussed in the Introduction, recall our previous work on this problem, and explain how our main theorem concerning AQUE is deduced.

### 6.1. Quantum unique ergodicity on locally symmetric spaces.

**Problem 6.1.** (QUE on locally symmetric spaces; Sarnak) Let $G$ be a connected semi-simple Lie group with finite center. Let $K$ be a maximal compact subgroup of $G$, $\Gamma < G$ a lattice, $X = \Gamma\backslash G$, $Y = \Gamma\backslash G/K$. Let $\{\psi_n\}_{n=1}^\infty \subset L^2(Y)$ be a sequence of normalized eigenfunctions of the ring of $G$-invariant differential operators on $G/K$, with the eigenvalues w.r.t. the Casimir operator tending to $\infty$ in absolute value. Is

[12] instead of $B(C, \epsilon)$; it is easy to see the proof works verbatim even though $B(C', M.\varepsilon)$ need not be contained in $\Omega_\infty$

it true that $\bar{\mu}_{\psi_n} := |\psi_n|^2 d\mathrm{vol}$ converge weak-* to the normalized projection of the Haar measure to $Y$?

In the paper [11] we have obtained Theorem 6.2, recalled below, constructing the microlocal lift in this setting. We needed to impose a non-degeneracy condition on the sequence of eigenfunctions (the assumption essentially amounts to asking that all eigenvalues tend to infinity, at the same rate for operators of the same order.) For the precise definition of *non-degenerate*, we refer to [11, Section 3.3].

With $K$ and $G$ as in Problem 6.1, let $A$ be as in the Iwasawa decomposition $G = NAK$, i.e. $A = \exp(\mathfrak{a})$ where $\mathfrak{a}$ is a maximal abelian subspace of $\mathfrak{p}$. For $G = \mathrm{PGL}_d(\mathbb{R})$ and $K = \mathrm{PO}_d(\mathbb{R})$, one may take $A$ to be the subgroup of diagonal matrices with positive entries. Let $\pi\colon X \to Y$ be the projection. We denote by $dx$ the $G$-invariant probability measures on $X$, and by $dy$ the projection of this measure to $Y$.

**Theorem 6.2.** *Let $\psi_n \subset L^2(Y)$ be a non-degenerate sequence of normalized eigenfunctions, whose eigenvalues approach $\infty$. Then, after replacing $\psi_n$ by an appropriate subsequence, there exist functions $\tilde{\psi}_n \in L^2(X)$ and distributions $\mu_n$ on $X$ such that:*

1. *(Lift) The projection of $\mu_n$ to $Y$ coincides with $\bar{\mu}_n$, i.e. $\pi_*\mu_n = \bar{\mu}_n$.*
2. *Let $\sigma_n$ be the measure $|\tilde{\psi}_n(x)|^2 dx$ on $X$. Then, for every $g \in C_c^\infty(X)$, we have $\lim_{n\to\infty}(\sigma_n(g) - \mu_n(g)) = 0$.*
3. *(Invariance) Every weak-* limit $\sigma_\infty$ of the measures $\sigma_n$ (necessarily a positive measure of mass $\leq 1$) is $A$-invariant.*
4. *(Equivariance). Let $E \subset \mathrm{End}_G(C^\infty(X))$ be a $\mathbb{C}$-subalgebra of bounded endomorphisms of $C^\infty(X)$, commuting with the $G$-action. Noting that each $e \in E$ induces an endomorphism of $C^\infty(Y)$, suppose that $\psi_n$ is an eigenfunction for $E$ (i.e. $E\psi_n \subset \mathbb{C}\psi_n$). Then we may choose $\tilde{\psi}_n$ so that $\tilde{\psi}_n$ is an eigenfunction for $E$ with the same eigenvalues as $\psi_n$, i.e. for all $e \in E$ there exists $\lambda_e \in \mathbb{C}$ such that $e\psi_n = \lambda_e\psi_n, e\tilde{\psi}_n = \lambda_e\tilde{\psi}_n$.*

We first remark that the distributions $\mu_n$ (resp. the measures $\sigma_n$) generalize the constructions of Zelditch (resp. Wolpert). Although, in view of (2), they carry roughly equivalent information, it is convenient to work with both simultaneously: the distributions $\mu_n$ are canonically defined and easier to manipulate algebraically, whereas the measures $\sigma_n$ are patently positive and are central to the arguments of the present paper.

The existence of the microlocal lift already places a restriction on the possible weak-* limits of the measures $\{\bar{\mu}_n\}$ on $Y$. For example, the $A$-invariance of $\mu_\infty$ shows that the support of any weak-* limit measure $\bar{\mu}_\infty$ must be a union of maximal flats. Following Lindenstrauss, we term the weak-* limits $\sigma_\infty$ of the lifts $\sigma_n$ *quantum limits*.

More importantly, Theorem 6.2 allows us to pose a new version of the problem:

**Problem 6.3.** (QUE on homogeneous spaces) In the setting of Problem 6.1, is the $G$-invariant measure on $X$ the unique non-degenerate quantum limit?

The main result of this paper is the resolution of Problem 6.3 for certain higher rank symmetric spaces, in the context of *arithmetic* quantum limits. We refer to [11, Section 1.4] for a further discussion of the significance of these spaces and how the introduction of arithmetic helps to eliminates degeneracy.

6.2. **Results: Arithmetic QUE for division algebra quotients.** For brevity, we state the result in the language of automorphic forms; in particular, $\mathbb{A}$ is the ring of adèles of $\mathbb{Q}$.

Let $D/\mathbb{Q}$ be a division algebra of prime degree $d$, and let $\mathbf{G}$ be the associated projective general linear group, i.e. the quotient of the group of units in $D$ by its center. Assume that $\mathbf{G}$ is split at $\infty$, ie that $G = \mathbf{G}(\mathbb{R}) \simeq \mathrm{PGL}_d(\mathbb{R})$. Let $K_{\mathrm{f}}$ be an open compact subgroup of $\mathbf{G}(\mathbb{A}_{\mathrm{f}})$ such that $X = \mathbf{G}(\mathbb{Q})\backslash\mathbf{G}(\mathbb{A})/K_{\mathrm{f}}$ consists of a single $G$-orbit (this condition is mainly cosmetic: see Remark 6.3 in §6.3). Then there exists a discrete subgroup $\Gamma < G$ such that $X = \Gamma\backslash G$. Let $\mathcal{H}$ be the Hecke algebra, as defined in Section 2. It acts on $L^2(X)$. Set $Y = \Gamma\backslash G/K$ the associated locally symmetric space, where $K$ is the standard maximal compact subgroup inside $G$. $A$ will denote the maximal split torus of diagonal matrices in $G$.

The Theorem 5.4 implies:

**Theorem 6.4.** *Let $\tilde{\psi}_n \in L^2(X)$ be a sequence of $\mathcal{H}$-eigenfunctions on $X$ such that the associated probability measures $\sigma_n := |\tilde{\psi}_n(x)|^2 dx$ on $X$ converge weak-* to an $A$-invariant probability measure $\sigma_\infty$. Then every $a \in A\backslash\{1\}$ acts on every $A$-ergodic component of $\sigma_\infty$ with positive entropy.*

*Proof.* This is essentially a rephrasing of Theorem 5.4, where the uniformity of the estimate means it carries over to weak-* limits.

For a proof that this bound implies that $a$ acts with positive entropy see [5, Sec. 8]. While written for the case of quaternion algebras ($d = 2$), that discussion readily generalizes to our situation by modifying its "Step 2" to account for the action of $a$ on the Lie algebra. □

Using results on measure-rigidity due to Einsiedler and Katok, this has the following implication for the QUE problem:

**Theorem 6.5.** *Notations as above, let $\{\psi_n\}_{n=1}^{\infty} \subset L^2(Y)$ be a non-degenerate sequence of eigenfunctions for the ring of $G$-invariant differential operators on $G/K$ (cf.* [11, Sec. 3.3]*) which are also eigenforms of the Hecke algebra $\mathcal{H}$ (cf. Section 2). Such $\psi_n$ are also called Hecke-Maass forms.*

*Then the associated probability measures $\bar{\mu}_n$ converge weak-* to the normalized Haar measure on $Y$, and their lifts $\mu_n$ (see Theorem 6.2) converge weak-* to the normalized Haar measure $dx$ on $X = \Gamma\backslash\mathrm{PGL}_d(\mathbb{R})$.*

*Proof.* The case $d = 2$ is Lindenstrauss's theorem, and we will thus assume $d \geq 3$.

Passing to a subsequence, let $\psi_n \in L^2(Y)$ be a non-degenerate sequence of Hecke-Maass forms on $Y$ such that $\bar{\mu}_n \to \bar{\mu}_\infty$ weakly. Passing to a subsequence, let $\tilde{\psi}_n$ and $\sigma_n$ be as in Theorem 6.2 such that $\sigma_n \to \sigma_\infty$ weakly and $\sigma_\infty$ lifts $\bar{\mu}_\infty$. Then $\sigma_\infty$ is a non-degenerate arithmetic quantum limit on $X$. By Theorem 6.4, $\sigma_\infty$ is an $A$-invariant probability measure on $X$ such that every $a \in A \setminus \{1\}$ acts on almost every $A$-ergodic component of $\sigma_\infty$ with positive entropy. Remark also that the measure $\sigma_\infty$ is invariant under the finite group $Z_K(A)$, the centralizer of $A$ in $K$, by construction (see [9, Remark 1.7, (5)]).

Then [3, Thm. 4.1(iv)] shows that $\sigma_\infty$ has a unique ergodic component, $\mu_{\mathrm{Haar}}$. □

6.3. **Remarks on generalizations.**

(1) The assumption imposed that $G$ act with a single orbit in $\mathbf{G}(\mathbb{Q})\backslash\mathbf{G}(\mathbb{A}_{\mathrm{f}})/K_{\mathrm{f}}$ is, as we remarked, cosmetic. In general, if we remove this assumption, one

would still know – making analogous definitions – that quantum limits remain $G$-invariant. However, this would not quite be a complete answer since the space of $G$-invariant measures on $\mathbf{G}(\mathbb{Q})\backslash\mathbf{G}(\mathbb{A})/K_{\mathrm{f}}$ is now finite dimensional, and we would not know the relative measures of the different components.

(2) In all likelihood it is possible to obtain a version of Theorem 6.2 even relaxing the non-degeneracy assumption. We hope to discuss this elsewhere.

(3) We expect the techniques developed for the proof of Theorem 6.5 will generalize at least to some other locally symmetric spaces, the case of $D$ being the simplest; but there are considerable obstacles to obtaining a theorem for *any* arithmetic locally symmetric space at present. A discussion of some of these difficulties (in a more general context than the QUE problem) will appear in [9].

(4) It is also possible to prove results for the case where $\mathbf{G}$ is split, i.e. isomorphic to $\mathrm{PGL}_d$ over $\mathbb{Q}$. The proof is essentially the same except that in order to obtain the control of intersections by a number field we only use the diophantine Lemma when the points are close to an $\mathbb{R}$-split torus in $G$, i.e. when $a \in A$ is a regular element. Thus only these elements are known to act with positive entropy in the limit. The measure rigidity results of [4] can then be used. Since in that case the quotient is not compact this does not address the escape-of-mass question. Somewhat surprisingly, however, the normalization of the measure is already controlled by the degenerate Eisenstein series. Hence a sub-convexity result for the Rankin-Selberg $L$-function would control the escape, just as in the case of $\mathrm{GL}_2$.

## Appendix A. Proof of Lemma 5.2: how to construct a higher rank amplifier

To readers familiar with the usage of "amplification" in analytic number theory (as represented, for instance, in the work of Duke, Friedlander and Iwaniec): the Lemma 5.2 in effect represents a way to construct an amplifier in higher rank.

We shall use certain standard properties of split groups over $p$-adic fields. Standard references are [12] and [2]. In the proof that follows, the notations $\ll_{\mathbf{G}}$ and $\gg_{\mathbf{G}}$ will be used to denote implicit constants depending only on the isomorphism class of $\mathbf{G}$ (not, e.g. on the extra data fixed in Section 2).

Fix a number field $K$ over which $\mathbf{G}$ splits; let $\mathbf{T}_0$ be a maximal split torus in $\mathbf{G} \times_{\mathbb{Q}} K$; let $X_* = \mathrm{Hom}(\mathbb{G}_m, \mathbf{T}_0)$ and $X^* = \mathrm{Hom}(\mathbf{T}_0, \mathbf{G}_m)$. Let $\Phi \subset X^*$ be the set of roots for the action of $\mathbf{T}_0$ on the Lie algebra of $\mathbf{G}$. Let $\Phi_\rho \subset X^*$ be the set of roots for the action of $\mathbf{T}_0$ on $K^n$ via the representation $\rho$ (see §2). We fix a $W$-invariant inner product $\|\cdot\|$ on $X_*$.

We assume that $p$ is good, as we may, because the statement concerns only "sufficiently large" $p$. There is a set $\mathcal{P}$ of such $p$, of positive density, so that $K$ embeds into $\mathbb{Q}_p$. For each $p \in \mathcal{P}$, fix an embedding $K \hookrightarrow \mathbb{Q}_p$. We deal only with such $p \in \mathcal{P}$ in the sequel.

Fixing $R \geq 1$, let $x_1, \ldots, x_r$ be a set of representatives for all nonzero $W$-orbits contained in $\{x \in X_* : \|x\| \leq R\}$. Also, set $R' = \{\sup |\alpha(x_i)| : 1 \leq i \leq r, \alpha \in \Phi_\rho\}$.

We obtain from the torus $\mathbf{T}_0$ (and the choice of embedding $K \hookrightarrow \mathbb{Q}_p$) a maximal split torus $\mathbf{A}_p \subset \mathbf{G} \times_{\mathbb{Q}} \mathbb{Q}_p$. Put $A_p = \mathbf{A}_p(\mathbb{Q}_p)$. Fixing a positive system of roots for $A_p$, let $N_p$ be the subgroup corresponding to all the positive roots. For sufficiently

large $p$, we have the Iwasawa decomposition $\mathbf{G}(\mathbb{Q}_p) = N_p.A_p.K_p$. Let $\delta : A_p \to \mathbb{R}^\times$ be the character corresponding to the half-sum of positive roots, composed with $\|\cdot\|_p$ on $\mathbb{Q}_p$. Let $\mathfrak{a} := A_p/(A_p \cap K_p)$, a free abelian group of rank equal to the rank of $\mathbf{G}(\mathbb{Q}_p)$. The Weyl group $W$ of $A_p$ acts on $\mathfrak{a}$.

Then we may identify $\mathfrak{a}$ with $X_*$, uniquely up to the action of $W$ on either. Indeed, $\mathfrak{a}$ is identified with $\mathrm{Hom}(\mathbf{G}_m, \mathbf{A}_p)$: for, given $a \in \mathfrak{a}$, there exists a unique homomorphism $\theta : \mathbb{G}_m \to \mathbf{A}_p$ so that $\theta(p)$ and $a$ lie in the same $A_p \cap K_p$ coset. Moreover, $X_*$ is evidently identified (by extension of scalars from $K$ to $\mathbb{Q}_p$) with $\mathrm{Hom}(\mathbf{G}_m, \mathbf{A}_p)$, whence the result.

This yields the desired identification of $\mathfrak{a}$ with $X_*$. Transporting the inner product on $X_*$ gives a $W$-invariant inner product on $\mathfrak{a}$, also denoted $\|\cdot\|$.

Also, if $a(x) \in \mathfrak{a}$ corresponds to $x \in X_*$, it is clear that we have:

$$g \in K_p a(x) K_p \Rightarrow \mathrm{denom}_p(g) \leq \sup_{\alpha \in \Phi_\rho} p^{|\alpha(x)|}, \quad \text{large enough } p \tag{A.1}$$

Let $\mathfrak{a}^* = \mathrm{Hom}(\mathfrak{a}, \mathbb{C}^\times)$ and $\mathfrak{a}^*_{\mathrm{temp}}$ the subset of *unitary* characters. Then $\mathfrak{a}^*_{\mathrm{temp}}$ is a (compact) torus.

To any character $\nu : \mathfrak{a} \to \mathbb{C}^\times$, we may associate a spherical representaion[13] $\pi(\nu)$ of $\mathbf{G}(\mathbb{Q}_p)$ by extending $\nu\delta$ to $N_p A_p$ trivally on $N_p$, and inducing to $\mathbf{G}(\mathbb{Q}_p)$, and taking the unique spherical subquotient.

For any $K_p$-bi-invariant function $k$ on $\mathbf{G}(\mathbb{Q}_p)$, let $\hat{k}(\nu)$ be the scalar by which $k$ acts on the spherical vector in $\pi(\nu)$. There is a unique $K_p$-bi-invariant function $\Xi_\nu$ on $\mathbf{G}(\mathbb{Q}_p)$ (the spherical function with parameter $\nu$) so that:

$$\hat{k}(\nu) = \int_{g \in \mathbf{G}(\mathbb{Q}_p)} k(g) \Xi_\nu(g) dg, \tag{A.2}$$

where the integral is taken w.r.t. the Haar measure that assigns mass 1 to $K_p$.

The map $k \mapsto \hat{k}(\nu)$ is an isomorphism between the space of compactly supported, $K_p$-bi-invariant functions on $\mathbf{G}(\mathbb{Q}_p)$, and the space of $W$-invariant "trigonometric polynomials" on $\mathfrak{a}$; here "trigonometric polynomial" means "finite linear combination of characters."

The inverse of the spherical transform is given by the following explicit formula:

$$k(g) = \int_{\nu \in \mathfrak{a}^*_{\mathrm{temp}}} \hat{k}(\nu) \Xi_\nu(g) d\mu(\nu) \tag{A.3}$$

where $\mu$ is the Plancherel measure on $\mathfrak{a}^*_{\mathrm{temp}}$. In our normalization, it is a probability measure.

Moreover, as $p \to \infty$, the measure $\mu = \mu_p$ converges weakly to a probability measure $\mu_\infty$ on $\mathfrak{a}^*_{\mathrm{temp}}$, which is a finite linear combination of characters.

Moreover, the $L^2$-norm is given by

$$\int_{\mathbf{G}(\mathbb{Q}_p)} |k(g)|^2 dg = \int_{\nu \in \mathfrak{a}^*_{\mathrm{temp}}} |\hat{k}(\nu)|^2 d\mu(\nu) \tag{A.4}$$

Now let $\nu_0 \in \mathfrak{a}^*$. Let $a_1 = a(x_1), \ldots, a_r = a(x_r) \in \mathfrak{a}$ correspond to $x_1, \ldots, x_r \in X_*$. For complex numbers $\mathbf{c} = c_j \in \mathbb{C}$ $(1 \leq j \leq r)$ consider the function $k_{\mathbf{c}}$ with

[13] Recall a spherical (irreducible) representation of $\mathbf{G}(\mathbb{Q}_p)$ is one that possesses a one dimensional space of $K_p$-invariants

spherical transform

$$\hat{k_{\mathbf{c}}}(\nu) = \sum_{j=1}^{r} c_j \sum_{w \in W} w\nu(a_j) \tag{A.5}$$

Note that if $\alpha_1, \ldots, \alpha_m$ are any nonzero complex numbers, then by a simple compactness argument.

$$\max_{j \neq 0, |j| \leq m} \left| \alpha_1^j + \cdots + \alpha_m^j \right| \geq c(m) > 0 \tag{A.6}$$

Take $R$ so large so that the set $\{a_i\}_{1 \leq i \leq r}$ contains a subset $S$ of the form $\{ja : j \neq 0, |j| \leq |W|\}$ for some $a \in \mathfrak{a}$ nonzero; moreover, if $R$ is sufficiently large, this may be done in such a way that the characters $\nu \to \nu(ja)$ of $\mathfrak{a}^*_{\text{temp}}$ integrate to zero against the measure $\mu_\infty$. This choice of $R$ may be done in a way that depends only on $\mathbf{G}$.

For this choice of $R$ and this choice of $S$, it follows from (A.6) that for any $\nu \in \mathfrak{a}^*$,

$$\sum_{j \in S} \left| \sum_{w \in W} w\nu(a_j) \right|^2 \gg_{\mathbf{G}} 1$$

Define $\mathbf{c}$ by $\mathbf{c}_j = \begin{cases} \overline{\sum_{w \in W} w\nu_0(a_j)}, j \in S \\ 0, j \notin S \end{cases}$. Set $\|\mathbf{c}\|_2^2 := \sum_{j=1}^r |c_j|^2$. Then $\|\mathbf{c}\|_2 \gg_{\mathbf{G}} 1$ and

$$\sup_j |\mathbf{c}_j| \ll \|\mathbf{c}\|_2, \sup_{\nu \in \mathfrak{a}^*_{\text{temp}}} |\hat{k}_{\mathbf{c}}(\nu)| \ll \|\mathbf{c}\|_2 \tag{A.7}$$

By (A.3), (A.7) and our comments about $\mu_\infty$, we have, for any $\varepsilon > 0$, the bound $|k_{\mathbf{c}}(1)| \leq \varepsilon \|\mathbf{c}\|_2$ for $p$ sufficiently large (in terms of $\varepsilon, \mathbf{G}$).

Moreover, by definition, $\hat{k}_{\mathbf{c}}(\nu_0) = \|\mathbf{c}\|_2^2$. By (A.4) and (A.7), there exists a constant $a > 0$ so that $\|k_{\mathbf{c}}\|_{L^2}^2 \leq a\|\mathbf{c}\|_2^2$ for large enough $p$.

We set $k = k_{\mathbf{c}} - k_{\mathbf{c}}(1)1_{K_p}$. Then $k(1) = 0$ and $|\hat{k}(\nu_0)| \gg_{\mathbf{G}} \|k\|_{L^2}$ for $p$ sufficiently large. By the Paley-Wiener theorem for $p$-adic groups, the function $k$ is supported in the set

$$K_p \cdot \{a \in \mathfrak{a}_{\text{temp}} : \|a\| \leq R\} \cdot K_p.$$

(We are unable to locate a suitable reference, but this statement is elementary and may be established directly using the proof of the bijectivity of the Satake transform [7, Theorem 3.3.6]. We indicate how to extract it from that proof. Let $\mathfrak{a}^+$ be the closed positive Weyl chamber within $\mathfrak{a}$; it is denoted '$T^{++}$" in the notations of [7].

Suppose $\alpha \in \mathfrak{a}^+$. It follows from the considerations of [7] – see comments after (3.3.10) – that if $\beta \in \mathfrak{a}^+$ is so that $\nu \mapsto \nu(\beta)$ occurs with a nonzero coefficient in $K_p\hat{\alpha}K_p$, then, necessarily, $\alpha - \beta$ belongs to the dual cone to $\mathfrak{a}^+$. From this it is easy to see that $\|\alpha\| > \|\beta\|$, unless $\alpha$ and $\beta$ coincide. It is also proven in (3.3.8') that the character $\nu \mapsto \nu(\alpha)$ indeed occurs in $K_p\hat{\alpha}K_p$.

Now, choose $\alpha \in \mathfrak{a}^+$ so that $k(K_p\alpha K_p) \neq 0$ and $\|\alpha\|$ is maximal subject to that restriction. We claim that $\nu \mapsto \nu(\alpha)$ necessarily occurs in $\hat{k}$. For, in view of the remarks above, if this were not the case there must exist $K_p\beta K_p$ in the support of $k$, with $\beta \in \mathfrak{a}^+$ and $\|\alpha\| < \|\beta\|$. This is a contradiction.)

This means that we may therefore write $k$ as a linear combination of the characteristic functions of double cosets $K_p a_i K_p$, with $1 \leq i \leq r$:

$$k = \sum_{r=1}^{R} e_i 1_{K_p a_i K_p}, \quad e_i \in \mathbb{C}$$

Then we may rephrase the fact that $|\hat{k}(\nu_0)| \gg_{\mathbf{G}} \|k\|_{L^2}$ as:

$$\left| \sum_{i=1}^{r} e_i \widehat{1_{K_p a_i K_p}}(\nu_0) \right|^2 \gg_{\mathbf{G}} \sum_i |e_i|^2 \int_{K_p a_i K_p} dg$$

It follows that there exists at least one $1 \leq i \leq r$ so that

$$|\widehat{1_{K_p a_i K_p}}(\nu_0)|^2 \gg_{\mathbf{G}} \int_{K_p a_i K_p} dg. \tag{A.8}$$

Also, from (A.1), we have $\mathrm{denom}_p(K_p a_i K_p) \leq p^{R'}$. It is easy to verify that (note $a_i$ does not belong to $K_p$!)

$$p^{R''} \gg_{\mathbf{G}} \int_{K_p a_i K_p} dg \gg_{\mathbf{G}} p, \tag{A.9}$$

where $R''$ depends only the isomorphism class of $\mathbf{G}$. The quantities of (A.9) could be explicitly computed, in fact, using [2, Section 3.5].

The desired result (Lemma 5.2) follows from (A.8) and (A.9), multiplying $1_{K_p a_i K_p}$ by a suitable complex number of absolute value 1. □

## References


[1] Jean Bourgain and Elon Lindenstrauss, *Entropy of quantum limits*, Comm. Math. Phys. **233** (2003), no. 1, 153–171.

[2] P. Cartier, *Representations of p-adic groups: a survey*, Automorphic forms, representations and $L$-functions (Oregon State Univ., Corvallis, Ore., 1977), Part 1, Proc. Sympos. Pure Math., XXXIII, American Mathematical Society, Providence, R.I., 1979, pp. 111–155. MR **81e:**22029

[3] Manfred Einsiedler and Anatole Katok, *Invariant measures on $G/\Gamma$ for split simple Lie groups $G$*, Comm. Pure Appl. Math. **56** (2003), no. 8, 1184–1221, Dedicated to the memory of Jürgen K. Moser. MR **2004e:**37042

[4] Manfred Einsiedler, Anatole Katok, and Elon Lindenstrauss, *Invariant measures and the set of exceptions to littlewood's conjecture*, `http://www.math.princeton.edu/~elonl/Publications/EKLslnr.pdf`.

[5] Elon Lindenstrauss, *Adelic dynamics and arithmetic quantum unique ergodicity*, to be published in the Proceedings of the 2005 Current Developments in Mathematics Conference.

[6] ———, *Invariant measures and arithmetic quantum unique ergodicity*, preprint (2003), (54 pages).

[7] I. G. Macdonald, *Spherical functions on a p-adic Chevalley group*, Bull. Amer. Math. Soc. **74** (1968), 520–525. MR MR0222089 (36 #5141)

[8] Zeév Rudnick and Peter Sarnak, *The behaviour of eigenstates of arithmetic hyperbolic manifolds*, Comm. Math. Phys. **161** (1994), no. 1, 195–213. MR **95m:**11052

[9] P. Sarnak and A. Venkatesh, *The size of an eigenfunction on an arithmetic manifold*, in preparation.

[10] Lior Silberman, *Arithemtic quantum chaos on locally symmetric spaces*, Ph.D. thesis, Princeton University, 2005, available at `http://www.math.harvard.edu/~lior/work/`.

[11] Lior Silberman and Akshay Venkatesh, *On quantum unique ergodicity for locally symmetric spaces I*, preprint available at `http://arxiv.org/abs/math/407413`.

[12] J. Tits, *Reductive groups over local fields*, Automorphic forms, representations and $L$-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part 1, Proc. Sympos. Pure Math., XXXIII, Amer. Math. Soc., Providence, R.I., 1979, pp. 29–69. MR **MR546588 (80h:**20064)

[13] Steven Zelditch, *Pseudodifferential analysis on hyperbolic surfaces*, J. Funct. Anal. **68** (1986), no. 1, 72–105. MR **87j:**58092

Lior Silberman, Department of Mathematics, Harvard University, One Oxford Street, Cambridge, MA 02138-2901, USA.

*E-mail address*: lior@math.harvard.edu

Akshay Venkatesh, Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA.

*E-mail address*: venkatesh@cims.nyu.edu