RECENT PROGRESS ON
THE QUANTUM UNIQUE ERGODICITY CONJECTURE

PETER SARNAK

Abstract. We report on some recent striking advances on the quantum unique ergodicity, or QUE conjecture, concerning the distribution of large frequency eigenfunctions of the Laplacian on a negatively curved manifold. The account falls naturally into two categories. The first concerns the general conjecture where the tools are more or less limited to microlocal analysis and the dynamics of the geodesic flow. The second is concerned with arithmetic such manifolds where tools from number theory and ergodic theory of flows on homogeneous spaces can be combined with the general methods to resolve the basic conjecture as well as its holomorphic analogue. Our main emphasis is on the second category, especially where QUE has been proven. This note is not meant to be a survey of these topics, and the discussion is not chronological. Our aim is to expose these recent developments after introducing the necessary backround which places them in their proper context.

2. The general QUE conjecture

In Figure 1 the domains $\Omega_E$, $\Omega_S$, and $\Omega_B$ are an ellipse, a stadium and a Barnett billiard table, respectively. Superimposed on these are the densities of a consecutive sequence of high frequency eigenfunctions (states, modes) of the Laplacian. That is, they are solutions to

$$\begin{align*}
\Delta \phi_j + \lambda_j \phi_j &= 0 \quad \text{in } \Omega, \\
\phi \big|_{\partial \Omega} &= 0 \quad \text{(Dirichlet boundary conditions),} \\
\int_{\Omega} |\phi_j|^2 \, dx \, dy &= 1.
\end{align*}
$$

Here $\Delta = \text{div} \text{grad} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$, $\lambda_1 < \lambda_2 \leq \lambda_3 \cdots$ are the eigenvalues, and the eigenfunctions are normalized to have unit $L^2$-norm. The sequences are of 12 consecutive modes around the 5600th eigenvalue. They are ordered from left to right and then down, and the grayscale represents the probability density $|\phi|^2$, with zero white and larger values darker.

The difference in the distribution of the densities is striking and its source is well understood, either through the wave equation $\frac{\partial^2 u}{\partial t^2} = \Delta u$ on $\Omega \times \mathbb{R}$ and geometric optics or the Schrödinger equation $i \frac{\partial u}{\partial t} = \Delta u$ on $\Omega \times \mathbb{R}$ and its semiclassical analysis. Both point to the intimate connection between the high frequency states and the
corresponding classical Hamiltonian dynamics, which in the domain case is that of a billiard in $\Omega$ moving at unit speed and bouncing at the boundary according to the law that the angle of incidence equals the angle of reflection. The difference in terms of the classical dynamics between $E$ and the other two is that for the ellipse this motion is integrable while for $S$ it is ergodic \cite{Bi}, and $B$, being a dispersing Sinai billiard, is ergodic and strongly chaotic \cite{Si}, \cite{C-M}. In more detail, in the case of an ellipse, a billiard trajectory that is once tangent to a confocal conic will remain so forever; see Figure 2 (this goes back at least to \cite{Bos} and is a nice exercise in projective geometry). This second dynamically invariant quantity renders the integrability of $E$. For $S$ and $B$ the ergodicity ensures that almost all of the trajectories are dense in the space of unit directions at points of the domain and it
is equivalent to these trajectories being equidistributed with respect to the Liouville measure \( \mu = \frac{dx dy d\theta}{(2\pi \text{Area}(\Omega))} \), where \( \theta \) is the angle of the direction.

There are many questions that are asked about such high frequency eigenmodes; we focus on the most basic one concerning their distribution. The density \( \nu_\phi := |\phi(x, y)|^2 dx dy \) is a probability measure on \( \Omega \) which quantum mechanically is interpreted as the probability distribution associated with being in the state \( \phi \). Do these measures equidistribute as \( \lambda \to \infty \) or can they localize? In the case of \( E \) or more generally the quantization of any completely integrable Hamiltonian system, these measures (or rather their microlocal lifts, see below) localize on invariant tori in a well understood manner (see \[La\] and \[CdV1\] for example). For \( E \) this is what one is seeing in Figure 1E, the high frequency eigenstates are very accurately localized onto invariant tori with action-angle parameters appropriately quantized and their projections onto \( \Omega \) fill out the shaded regions given by confocal conics. The remarkable correspondence between high frequency modes for an integrable system and its invariant tori, is the reason that Bohr was able to develop quantum theory for the hydrogen atom using only classical orbits. For helium he had little success as its classical mechanics is the nonintegrable and partially chaotic three body problem as opposed to the integrable two body problem for hydrogen. Our interest is in systems like \( S \) and \( B \) for which the familiar techniques from microlocal analysis (that is, geometric optics and semiclassical analysis, see for example \[Mel\]) say much less about individual high frequency states and a theoretical analysis is problematic.

There is a natural extension of the measures \( \nu_\phi \) to \( T_1(\Omega) \), the space of unit directions over \( \Omega \), which measures their distribution in this larger phase space. This extension is denoted by \( \mu_\phi \) and is called the microlocal lift of \( \nu_\phi \) and can be given explicitly as follows:

For a smooth function \( f(x, z) \) on \( T_1(\Omega) = \Omega \times S^1 \), set

\[
(1) \quad \mu_\phi(f) = \langle Op(f) \phi, \phi \rangle,
\]
where $\langle \cdot, \cdot \rangle$ is the $L^2$ scalar product on $\Omega$.

$$\text{(2)} \quad Op(f)\psi(x) = \int_{\mathbb{R}^2} e^{\langle x, \xi \rangle} \hat{\psi}(\xi) f(x, \frac{\xi}{|\xi|}) \, d\xi,$$

$$\text{(3)} \quad \hat{\psi}(\xi) = \int_{\mathbb{R}^2} \psi(x) e(-\langle x, \xi \rangle) \, dx$$

is the Fourier transform of $\psi$, and

$$e^{\langle x, \xi \rangle} = e^{2\pi i(x_1 \xi_1 + x_2 \xi_2)}.$$

$Op(f)$ is a zeroth order pseudo-differential operator with symbol $f$. Note that if $f(x, z)$ is a function of $x$ alone, then $Op(f) \psi(x) = f(x) \psi(x)$ and hence $\mu_\phi$ is indeed an extension of $\nu_\phi$. It is known ([Sh1], [Sh2], [Ze1])\(^1\) that $\mu_\phi$ is asymptotically positive; that is, if $f \geq 0$, then $\lim \mu_\phi(f) \geq 0$ as $\lambda \to \infty$. Hence any weak limit of the $\mu_\phi$’s is a probability measure. We call such a limit $\beta$, a quantum limit. The key insight due to Shnirelman ([Sh1]) is that any such quantum limit is invariant under the Hamilton billiard flow on $T_1(\Omega)$. To see this, let $\exp(i\sqrt{-\Delta} t)$ denote the wave operator (that is, the operator that evolves the solution to the wave equation), then

$$\text{(4)} \quad \langle \exp(-i\sqrt{-\Delta} t) \, Op(f) \exp(i\sqrt{-\Delta} t) \phi_j, \phi_j \rangle = \mu_{\phi_j}(f), \quad \text{for all } t.$$

On the other hand by the fundamental theorem of Egorov ([E] on propagation of singularities for the wave equation, $\exp(-i\sqrt{-\Delta} t) \, Op(f) \exp(i\sqrt{-\Delta} t)$ is a zeroth order pseudo-differential operator with leading symbol $Op(f_0 H_1)$, where $H_t$ is the Hamilton flow at time $t$. Thus as $\lambda \to \infty$, this together with (4) leads to the $H$-invariance of $\beta$.

The discussion above applies with almost no changes when $\Omega$ is replaced by a compact Riemannian manifold $(M, g)$, which for simplicity we assume has no boundary. The Laplacian is replaced by the Laplace-Beltrami operator $\triangle_g$ for the metric and the classical mechanics is that of motion by geodesics on $X = T_1(M)$, the space of unit tangent vectors over $M$. For an eigenfunction $\phi$ of $\triangle_g$ on $M$, we form as above the probability measure $\nu_\phi = |\phi(x)|^2 \, dv(x)$ on $M$ (here $dv$ is the volume element) and its microlocal lift $\mu_\phi$ to $T_1(M)$. There is a technical point here in that $M$ has no global coordinates and in defining pseudo-differential operators of, say, zeroth order, the lower order symbols depend on the coordinates used. This means that the microlocal lift (and more generally the process of quantization) is not unique, however since these differences are of lower order, in large $\lambda$-limit they disappear and thus lead to the same quantum limits $\beta$, that is, weak limits of the $\mu_\phi$’s. As before, any quantum limit is invariant under the geodesic flow.

We are interested in the case where the geodesic flow is ergodic, meaning that the only flow invariant subsets of $T_1(M)$ are either of zero or full $\mu$-measure, where $\mu$ is the Liouville measure (that is the natural measure on $T_1(M)$ coming from the metric $g$). In this case Birkhoff’s ergodic theorem implies that $\mu$-almost all geodesics are $\mu$-equidistributed in $T_1(M)$. There is a corresponding high frequency eigenfunction analogue of this ergodicity, called quantum ergodicity, which is formulated and proven in [Sh1], [Sh2], [Ze1], and [CdV2]. It asserts that if the geodesic flow is ergodic, then almost all (in the sense of density) of the eigenfunctions become equidistributed with respect to $\mu$. That is, if $\{\phi_j\}_{j=0}^{\infty}$ is an orthonormal basis of eigenfunctions with $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \lambda_3 \cdots$, then there is a subsequence $j_k$ of integers of full density, such that $\mu_{j_k} \to \mu$ as $k \to \infty$. One can see the domain

\(^1\)Actually, these apply to the case of no boundary; see [G-L] and [Z-Z] for this case.
version of this theorem \cite{G-L} and \cite{Z-Z} in action even in the small samples in Figures 1S and 1B.

The basic question is whether there can be other quantum limits, that is, subsequences on which the $\mu_\phi$‘s behave differently. If $M$ has (strictly) negative curvature, then the geodesic flow is very well understood thanks to works of Hopf, Morse, Sinai, Bowen, and others. It is ergodic and strongly chaotic in all senses. The periodic geodesics are isolated and are unstable, and there is no restriction on how they may distribute themselves as their period increases. In this context Colin de Verdiere \cite{CdV3} and also Zelditch asked an insightful question as to whether the most singular flow invariant measure, namely the arclength measure on a closed geodesic, can be a quantum limit. (These are called \textit{strong scars} in the terminology of \cite{Hel}.) In \cite{R-S} this question is answered in the negative for arithmetic surfaces (see §2) and based on that and a careful examination of conflicting interpretations of numerical experiments (\cite{Hel}, \cite{A-S}, \cite{Bog}, \cite{Be}) the following was put forth:

\textbf{QUE Conjecture} \cite{R-S} 2 If $M$ is a compact negatively curved manifold, then $\mu_\phi \to \mu$ as $\lambda \to \infty$. Put another way, $\mu$ is the only quantum limit. If true this says that even in the semiclassical limit the quantum mechanics of such strongly chaotic systems does not reflect the finer features of the classical mechanics. More recently, Barnett \cite{Ba} developed a numerical method which allows him to compute modes for $\Omega_B$ (which is the domain analogue of variable negative curvature) as large as the 700000th one, and his results confirm QUE for such general systems.

The first recent development connected with QUE that we review concerns the stadium domain $S$. Unlike geodesic motion on a negatively curved $M$, billiards in $S$ have a family of periodic orbits of period twice the distance between the parallel sides and which corresponds to billiards bouncing back and forth between these sides. The numerical computations (see numbers 10 and 11 in Figure 1S) as well as a direct construction of approximate eigenfunctions called quasi-modes (\cite{H-O}, \cite{Zel2}, \cite{Do}) indicate that there is a subsequence of modes whose $\mu_\phi$‘s converge to the singular measure supported on the totality of such bouncing balls.\footnote{The terminology is meant to reflect the feature of a uniquely ergodic dynamical system, that being one for which there is only one invariant measure or every orbit is equidistributed.} A rigorous proof that such \textit{bouncing ball modes} exist in the limit remained elusive. The difficulty being that there may be many modes whose eigenvalues are close to that of the quasi-mode eigenvalue. In \cite{Ha} Hassell examines the variation of the eigenvalues of the stadium $S_t$, where $t$ is the length of the edge indicated, as $t$ varies. He shows that for generic $t$ there is sufficient separation of the eigenvalues from the quasi-eigenvalues to ensure that there are genuine modes which are not equidistributing with respect to $\mu$.

\textbf{Theorem 1 (Ha).} For almost all $t$ the stadium $S_t$ has a quantum limit $\beta_t$ which gives positive mass to the bouncing ball trajectories and, in particular, $S_t$ is not QUE.

Turning to the negatively curved $M$‘s of the QUE conjecture, we point out a potential spoiling role that high eigenvalue multiplicities may play. Let $V_\lambda$ be the space of eigenfunctions on $M$ with eigenvalue $\lambda$. If $m(\lambda) = \dim V_\lambda$ is very large, one can choose $\phi$ in $V_\lambda$ for which $\mu_\phi$ is badly distributed. Thus implicitly

\textbf{Theorem 1 (Ha).} For almost all $t$ the stadium $S_t$ has a quantum limit $\beta_t$ which gives positive mass to the bouncing ball trajectories and, in particular, $S_t$ is not QUE.

Turning to the negatively curved $M$‘s of the QUE conjecture, we point out a potential spoiling role that high eigenvalue multiplicities may play. Let $V_\lambda$ be the space of eigenfunctions on $M$ with eigenvalue $\lambda$. If $m(\lambda) = \dim V_\lambda$ is very large, one can choose $\phi$ in $V_\lambda$ for which $\mu_\phi$ is badly distributed. Thus implicitly
the QUE conjecture asserts that these multiplicities cannot be very large in terms of $\lambda$, and any proof of QUE would have to address this multiplicity issue, perhaps indirectly. The best known upper bound for $m(\lambda)$ is proven using the wave equation and geometric optics (see [Ber]) and asserts that $m(\lambda) \ll \lambda^{(n-1)/2} / \log \lambda$, where $n = \dim M$. We expect, say for $n = 2$, that $m(\lambda) = O_\epsilon(\lambda^\epsilon)$, for any $\epsilon > 0$.

An important step in understanding quantum limits in the negatively curved setting was taken recently by Anantharaman [An]. Her result concerns the entropy $h(\beta)$ [Si2] of a quantum limit $\beta$. The entropy of the dynamical system $(T_1(M), G_t, \beta)$, where $G_t$ is the geodesic flow, is a measure of the complexity of the $\beta$-flow, and it was introduced into the QUE problem by Lindenstrauss in the context of arithmetic QUE (see Section 2). For example, if $\beta$ is the arc-length measure on a closed geodesic, then $h(\beta) = 0$, while for the Liouville measure, $h(\mu) > 0$.

In a little more detail, the entropy measures the exponential growth rate of distinct orbits of length $T$ supported by $\beta$, as $T \to \infty$.

**Theorem 2 ([An]).** If $\beta$ is a quantum limit on $T_1(M)$, where $M$ is a compact negatively curved manifold, then $h(\beta) > 0$.

While the result was expected to be true, the fact that one could prove it in this generality came as a surprise. Her proof is based on a direct estimation of the entropy, by a delicate combination of information about $\beta$ that is gotten by pushing the known semiclassical asymptotics to their natural limit, coupled with information obtained from the global hyperbolic dynamics of the geodesics. This novel combination of these quite different inputs allows her to infer this highly nontrivial information about individual eigenfunctions. In particular, this theorem answers the Colin de Verdiere/Zelditch question emphatically: The arc-length measure on a closed geodesic on a negatively curved manifold is never a quantum limit. The proof that $h(\beta) > 0$ comes with an explicit lower bound, which has been sharpened in [A-N] and [Ri]. Specifically, in [Ri] Riviére shows that for $M$ a negatively curved surface, $h(\beta) \geq \frac{1}{2} \int_{T_1(M)} \lambda_+ d\beta$, where $\lambda_+$ is the positive Lyapunov exponent. In the setting of the quantum cat map that we mention at the very end the analogue of this lower bound is sharp. The positive entropy theorem allows for the multiplicity $m(\lambda)$ to be as large as the upper bound mentioned earlier since it applies equally well with $\beta$ replaced by weak limits of $\mu_{\psi}$, where $\psi$ is a function on $M$ whose spectral projection lies in an interval $[\lambda, \lambda + H]$, with $H$ as large as $c_M \lambda^{7/2} / \log \lambda$, $c_M$ a positive constant depending on $M$. For a more detailed account of these recent developments concerning quantum limits, see Zelditch [Zel4].

3. Arithmetic QUE

The cases for which QUE has been established are arithmetic manifolds, and these are the subject of this section. We begin by restricting to dimension two and $M$ is of constant curvature say $K = -1$. The universal cover of $M$ is the hyperbolic plane $\mathbb{H}$, with its line element $ds = |dz|/y$ and its orientation preserving isometry group $G = PSL_2(\mathbb{R})$ acting by linear fractional transformations. Thus $M$ is realized as $\Gamma \backslash \mathbb{H}$ where $\Gamma$ is a discrete subgroup of $G$. We will also allow such quotients which are noncompact but of finite area. For these there is also continuous spectrum coming from the theory of Eisenstein series developed by Selberg and for which the QUE conjecture can be suitably formulated. $M$ is called arithmetic if
the group $\Gamma$ is constructed by arithmetic means (see [Ka] for a discussion). The basic and most important example is $\Gamma = SL_2(\mathbb{Z})$, that is, the group of $2 \times 2$ integer matrices of determinant 1. The quotient is the modular surface $Y$, it is not compact as shown in the familiar Figure 3.

An eigenfunction of the Laplacian on $Y$ is a function $\phi$ on $\mathbb{H}$ satisfying (there is also the continuous spectrum whose definition is similar):

\begin{align}
\phi(\gamma z) &= \phi(z), \quad \text{for } \gamma \in SL_2(\mathbb{Z}), \\
\Delta \phi &= y^2 \left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) \phi = -\lambda \phi, \\
\phi &\in L^2(Y).
\end{align}

Compact arithmetic surfaces are constructed similarly using integral matrices associated with quaternion division algebras [Ka]. The most important property that distinguishes arithmetic surfaces from the general constant curvature surface is that they carry a large family of algebraic correspondences which in turn give rise to the family of Hecke operators. These are linear operators on $L^2(M)$ which commute with each other and with $\Delta$. For example, if $Y$ is the modular surface, then for $n \geq 1$, the Hecke operator $T_n$ is defined by

\begin{equation}
T_n \psi(z) = \sum_{\substack{a \neq n \\ b \mod d}} \psi \left( \frac{az + b}{d} \right).
\end{equation}

One checks that if $\psi(\gamma z) = \psi(z)$ for $\gamma \in PSL_2(\mathbb{Z})$, then $T_n \psi$ is also $PSL_2(\mathbb{Z})$ invariant (see [Se1]).

The $T_n$’s are normal operators and hence this whole ring of Hecke operators together with $\Delta$ can be simultaneously diagonalized. If as is expected and is confirmed by numerical experiments, the Laplace spectrum of $Y$ is simple, then any eigenfunction $\phi$ of $\Delta$ is automatically an eigenfunction of the full Hecke ring. In any case in this arithmetic setting we always assume that $\phi$ is an eigenfunction of the Hecke ring. As discussed above one can always choose an orthonormal basis.

---

4 Actually, by “arithmetical” we mean that the group is furthermore a congruence group [P-S-R].
of such eigenfunctions. More generally, when we refer to arithmetic QUE we are assuming that our states are always Hecke eigenstates. This is the means by which we circumvent the issue of the unlikely, but potentially possible, high multiplicities \( m(\lambda) \). It is known that the multiplicities of the spaces of such joint Hecke eigenfunctions is small and, in particular for \( Y \), it is one. These eigenfunctions are called Maass forms and they are basic objects in modern number theory. As such, one might expect, and it is certainly the case, that this theory can be used to study the QUE question as well as many other interesting questions associated with the high frequency states for arithmetic manifolds; see \([Sa1] , [Sa2] , [Mark] , [Zel3]\). Figure 4 depicts the densities of a sequence of Maass forms on \( Y \) with shading and frequencies similar to those in Figure 1. The densities are less regular than those of the Barnett stadium.

The first results on QUE were obtained in \([L-S1]\) and \([Ja]\) where a suitably modified form of the equidistribution was established for the continuous spectrum for a noncompact arithmetic surface. According to our terminology, these are surfaces \( \Gamma \backslash \mathbb{H} \) where \( \Gamma \) is a congruence (see \([F-R]\) subgroup of \( SL_2(\mathbb{Z}) \) and in particular for \( Y \). A key point in the analysis is an explicit relation between \( \mu_\phi(f) \) (where \( \phi \) and \( f \) are Hecke eigenforms and \( f \) is possibly a holomorphic such form, see below) and special values of related Rankin-Selberg \( L \)-functions on their critical lines. One of the primary reasons for studying automorphic forms is that they give rise to families of \( L \)-functions generalizing Riemann’s zeta function and have properties similar to it. Via this relation, the QUE problem becomes one of estimating from above the corresponding special value. The convexity bound for such values is what one gets from a simple complex analytic interpolation, and it falls just short of what is needed. For example, for the Riemann zeta function

\[
\zeta(s) = \sum_{n=1}^{\infty} n^{-s} = \prod_p (1 - p^{-s})^{-1},
\]
the convexity bound on the critical line $s = \frac{1}{2} + it$, is (see [13])

$$\zeta\left(\frac{1}{2} + it\right) = O\left((|t| + 1)^{1/4}\right).$$  

(8)

The improvement or subconvex bound that is often needed is that there is $\delta > 0$ such that

$$\zeta\left(\frac{1}{2} + it\right) = O((|t| + 1)^{1/4 - \delta}).$$  

(9)

For reasons such as the one at hand, supplying such subconvex bounds for various automorphic $L$-functions has become a central problem in the theory of $L$-functions; see [13] and [14]. There is no doubt about the truth of subconvexity since it and optimally sharp bounds follow from the expected generalized Riemann Hypothesis for these $L$-functions. (For $\zeta$, one gets the bound $\zeta\left(\frac{1}{2} + it\right) = O_e((1 + |t|)^\epsilon)$, for any $\epsilon > 0$, known as the Lindeloff Hypothesis.) In the case of the continuous spectrum QUE, the required subconvex bounds were known and due to Weyl [We] for the Riemann zeta function and to [Mu] and [Go], for the degree two $L$-functions that present themselves here. In his thesis Watson (see [Wa]), building on the work [K-H], establishes a general explicit formula relating the integral of a product of three automorphic forms on an arithmetic surface to the special value of a degree 8 $L$-function. In the case of $Y$ and three (Hecke eigen) Maass cusp forms $\phi_1, \phi_2, \phi_3$ whose $L^2$-norms are as always normalized to be one, his formula reads

$$\begin{align*}
216 | \int_Y \phi_1(z)\phi_2(z)\phi_3(z) \frac{dxdy}{y^2} | & = \frac{\pi^4 \Lambda(\frac{1}{2}, 1, \phi_1, \phi_2, \phi_3) \Lambda(1, \text{sym}^2 \phi_1) \Lambda(1, \text{sym}^2 \phi_2) \Lambda(1, \text{sym}^2 \phi_3)}{\Lambda(1, \text{sym}^2 \phi_3) \Lambda(1, \text{sym}^2 \phi_2) \Lambda(1, \text{sym}^2 \phi_1)}.
\end{align*}$$

(10)

Here $\Lambda(s, \phi_1 \times \phi_2 \times \phi_3)$ is the completed degree 8 $L$-function associated with the tensor product $\phi_1 \otimes \phi_2 \otimes \phi_3$ and $\Lambda(s, \text{sym}^2 \phi_3)$ is a similar degree 3 $L$-function. Taking $\phi_1 = \phi_2 = \phi_3$ and $\phi_3$ fixed, the left-hand side is $\nu_{\phi_3}(\phi_3)^2$ and hence the QUE question, which asserts that this quantity goes to 0 (since $\int_Y \phi_3 \frac{dxdy}{y^2} = 0$), becomes one of estimating the value at $s = \frac{1}{2}$ of this family of $L$-functions as $\lambda \to \infty$. In this way Watson’s formula shows that the subconvexity feature is a general one, and that the full QUE for all arithmetic surfaces would follow from subconvexity for these degree 8 $L$-functions. Moreover, the Riemann Hypothesis for these $L$-functions yields the optimal rate of equidistribution of the measures $\mu_\phi$. With this there was no longer any doubt about the truth of QUE, at least in the arithmetic surface setting. Subconvexity for certain degree 4 $L$-functions was established in [13] and [14], from which QUE followed for dihedral forms. These forms are still special ones and are characterized by the degree 8 $L$-function factoring into ones of degree 4. While a lot of progress has been made on the subconvexity problem (see [14], [15]), what is needed for the general arithmetic QUE remains out of reach, and the solution of the problem took a quite different route.

The phase space $T_1(M) = T_1(\Gamma \backslash \mathbb{H})$ can be identified naturally with $\Gamma \backslash G$, and the geodesic flow $G_t$ with the diagonal $A$-flow given by $\Gamma g \mapsto \Gamma g \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}$, $a \in \mathbb{R}^*$. As we noted before, any quantum limit $\beta$ on $\Gamma \backslash G$ is $A$-invariant, but this alone is far from determining $\beta$. Motivated by ideas from measure rigidity for higher rank diagonal actions on homogeneous spaces, which we discuss briefly below in the context of products of $PSL_2(\mathbb{R})$’s, Lindenstrauss [11] established the following striking measure classification result. To describe it, we need to use further
terminology and notions such as $p$-adic numbers. We discuss the case of the modular surface $Y$, but it applies to any arithmetic surface. Fix a prime $p$ and let $X = \text{SL}_2(\mathbb{Z}[\frac{1}{p}]) \backslash \text{SL}_2(\mathbb{R}) \times \text{PGL}_2(\mathbb{Q}_p)/(1 \times \text{PGL}_2(\mathbb{Z}_p))$. Here $\mathbb{Q}_p$ denotes the $p$-adic numbers, $\mathbb{Z}_p$ the $p$-adic integers, and $\text{SL}_2(\mathbb{Z}[\frac{1}{p}])$ is embedded in the product $\text{SL}_2(\mathbb{R}) \times \text{PGL}_2(\mathbb{Q}_p)$ diagonally. $X$ carries the natural $A$ action by multiplication by $\left\{ \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}, 1 \right\}$ on the right and it is naturally foliated by leaves isomorphic to the $p+1$ regular tree $\tau_p = \text{PGL}_2(\mathbb{Q}_p)/\text{PGL}_2(\mathbb{Z}_p)$ (Se2). Let $\beta$ be a measure on $X$ which is $A$-invariant and for which all ergodic components of $\beta$ have positive entropy and which is $\tau_p$-recurrent (meaning that for $B \subset X$ with $\beta(B) > 0$ and for which almost all $x$ in $B$, there are infinitely many points $x'$ in the $\tau$-leaf of $x$ which are also in $B$). The conclusion is that $\beta$ is in fact $\text{SL}_2(\mathbb{R})$ invariant. To apply this to the QUE problem, Lindenstrauss exploits the fact that our eigenstates on $M$ are eigenfunctions of the Hecke operator $T_p$, which allows us to consider the $\mu_\phi$'s, in a standard way, as measures on $X$ and to consider the corresponding limits on this space. To verify the conditions in his measure classification, we use the full Hecke ring along the lines of [R-S] where these were used originally to show that the singular support of an arithmetic quantum limit cannot be a closed geodesic. In [B-L] this argument as well as the one in [Wo] is generalized vastly to show that the entropy of any ergodic component of $\beta$ is positive. The recurrence property of the $\tau_p$ foliation is easier to establish, and by putting all these threads together Lindenstrauss shows that any quantum limit is of the form $c \mu_\psi$, for some constant $0 \leq c \leq 1$. In the compact case $c$ must be 1 and we have:

**Theorem 3** ([Li1], Arithmetic QUE (compact case)). Let $M$ be a compact arithmetic surface. Then the only quantum limit is the Liouville measure $\mu$.

In the noncompact case there remained the possibility that some of the mass of $\mu_\phi$ escapes in the cusp, or that $c < 1$. However, recently Soundararajan [So1], using the multiplicativity of the Fourier coefficients in the expansion in the cusp of the forms $\phi$ (which is a consequence of being a Hecke eigenform) together with some clever arguments estimating the mass that $\nu_\phi$ can charge product sets in the cusp, shows that there can be no escape of mass, that is, $c = 1$ in Lindenstrauss’ theorem. Combining this with the previous results we have:

**Theorem 4** ([Ja], [Li1], [So1], Arithmetic QUE (noncompact)). Let $M$ be a noncompact arithmetic surface. Then QUE holds for both the continuous and discrete Hecke eigenfunctions on $M$.

Theorems 3 and 4 together establish QUE for arithmetic surfaces. The hyperbolic surfaces $M$ are also complex analytic Riemann surfaces, and as such the analogue of high frequency modes are holomorphic sections of high tensor powers of the canonical line bundle over $M$ or, what is the same thing, holomorphic forms of large weight. The question of the distribution of densities associated with sections arises more generally in complex and symplectic geometry, geometric quantization, as well as in quantizing symplectomorphisms on such manifolds (see [Dona] and [Zel5] for example). Our focus here is on the distribution question for the arithmetic sections, that is, for holomorphic Hecke eigenforms. (The algebraic correspondences in (6) give rise to similar Hecke operators which act on the space

---

5The reader may want to look first at the higher dimensional Hilbert modular case at the bottom of page 223 which is conceptually simpler.
RECENT PROGRESS ON QUE 221

of holomorphic forms of a given weight.) Such a form of weight \( k \) for \( \Gamma < PSL_2(\mathbb{R}) \) defines a probability density \( \nu_f \) on \( M = \Gamma \backslash \mathbb{H} \) using the Petersson inner product by setting

\[
\nu_f := |f(z)|^2 y^k \frac{dx dy}{y^2}.
\]

(11)

Here in the case that \( M \) is noncompact, we assume that \( f \) vanishes in the cusps, that is, \( f \) is a cusp form. By the Riemann-Roch theorem the dimension of the space of such forms grows linearly with \( k \). Thus in the arithmetic case we cannot drop the Hecke eigencondition if we are to expect the analogue of QUE, that is, the individual equidistribution of the \( \nu_f \)'s to hold (for example take a fixed form of weight \( k_0 \) and raise it to the \( k/k_0 \) power). Also there is no apparent \( A \)-invariant microlocal lift of the Hecke density \( \nu_f \) to \( \Gamma \backslash PSL_2(\mathbb{R}) \), and for this reason there is no known generalization of Lindenstrauss’ measure rigidity approach in this holomorphic setting. On the other hand as soon as the connection of such questions to special values of the \( L \)-function was made in the arithmetic setting, it was clear that the following QUE for holomorphic Hecke forms should be equally valid.

4. Arithmetic holomorphic QUE conjecture

Let \( M \) be an arithmetic surface. Then as \( k \to \infty \) with \( f \) a holomorphic Hecke eigenform of weight \( k \), \( \nu_f \to dA/Area(M) \), where \( dA \) is the hyperbolic area form.

Here too Watson’s explicit formula together with subconvexity for the corresponding degree 8 \( L \)-functions implies this holomorphic QUE. There is a very nice consequence in connection with the zeros of such \( f \)'s. In [N-V] and [S-Z] a general potential theoretic argument applied to \( \partial \bar{\partial} \log |f(z)| \) is used to show that if \( \nu_f \to dA/Area(M) \) and \( M \) is compact, then the zeros of \( f \) also become equidistributed in \( M \) with respect to \( dA/Area(M) \). This was extended in [Ru1] to the finite area noncompact cases. In particular, it follows that if the arithmetic holomorphic QUE conjecture is true, then the zeros of such holomorphic Hecke forms become equidistributed as \( k \to \infty \). In the next paragraphs we report on the recent works of Holowinsky and Soundararajan, [Ho], [So2] and [H-S], on this problem.

Soundararajan approaches the problem by seeking a more modest bound for the degree 8 \( L \)-functions in Watson’s formula. Instead of improving the convexity bound by a factor of \( k^{-\epsilon} \), he settles for \( (\log k)^{-\epsilon} \), which he calls weak subconvexity. The normalization factors in the denominator of (10) involve values of \( L \)-functions at \( s = 1 \) and which are potentially of size \( (\log k)^{-1} \) (G-H-L). Thus this weak subconvexity by itself does not do the full job; however, it is known [Lu] that with very few exceptions these special values at \( s = 1 \) are of more moderate size. Thus weak subconvexity allows Soundararajan to prove QUE for all but \( O(\epsilon k^\epsilon) \) of the \( \text{const} \times k \) Hecke eigenforms of weight \( k \) on \( M \). Moreover—and this is the striking feature—he establishes this weak subconvexity in the full generality of \( L \)-functions associated to automorphic cusp forms on the general linear group. The only assumption that he needs to make is that the corresponding forms satisfy the generalized Ramanujan Conjectures (see [Sa4]). Happily for holomorphic Hecke eigenforms on an arithmetic surface, the Ramanujan Conjectures are a celebrated theorem of Deligne [De]. To prove his general weak subconvexity, Soundararajan develops a far-reaching generalization of techniques in the theory of mean-values.

\[6\] Being such a general result, this will no doubt find many further applications.
Figure 5. Zeros of six holomorphic Hecke cusp forms of weight $k = 2000$ as computed by F. Stroemberg. A close inspection shows that each of these has a good number of its zeros on $\delta$, the union of the segments $\delta_1 = \{z : x = 0\}, \delta_2 = \{z : x = \frac{1}{2}\}, \delta_3 = \{z : |z| = 1\}$. These segments are distinguished in that the zeros of any such $f$ are symmetric under reflection in any one of these segments. In [G-S] it is shown that the number of zeros on $\delta$ of any such $f$ goes to infinity as $k$ goes to infinity.

of multiplicative functions, specifically showing that these do not vary much over certain ranges (see also [Hi] and [G-S]). This feature is the source of the sought-after cancellation in the sums over the multiplicative coefficients defining the $L$-functions in its Dirichlet series.

The starting point for Holowinsky is yet a third approach to QUE which depends critically on $M$ being noncompact, and it was used in [L-S1] and [L-S2] to investigate similar problems. Using the cusps of $M$, one develops the holomorphic form in a Fourier expansion and the QUE problem can be reduced to estimating shifted convolution (which are quadratic) sums involving these coefficients. One expects, and the quantitative forms of QUE demand, that there is a lot of cancellation due to the signs of the coefficients of such (varying) forms. Holowinsky’s novel idea is to forgo this cancellation and to exploit the fact that the mean-values of the absolute values of these coefficients is of size $(\log k)^{-\delta}$. The source of this phenomenon is that if $f$ is not dihedral (and we can assume this since QUE is known for these), then the distribution of the coefficients at primes are expected to follow a Sato-Tate distribution, and enough towards the latter is known by the work of Shahidi [K-S] to exploit this feature (see [E-M-S] for the case of a fixed $f$). Using a sieving argument (see also [Na] for a general such inequality in the context of multiplicative functions) Holowinsky is able to give a bound for the shifted convolution sums which

\footnote{A proof of this was recently posted; see [B-G-H-T].}
improves the trivial bound by a factor of a small power of \( \log k \). Here too Deligne’s theorem is used as a critical ingredient. To apply this bound to QUE, normalization factors involving the values at \( s = 1 \) of associated \( L \)-functions intervene again. In this way Holowinsky is also able to establish QUE for all but \( O_{s}(k') \) of the forms of weight \( k \), for a noncompact arithmetic surface \( M \).

The miracle—and it is not uncommon for such “luck” to be at the heart of a breakthrough—is that there are no common exceptions to Holowinsky and Soundararajan’s treatments. Soundararajan’s is unconditional as long as the value at \( s = 1 \) of the related symmetric square \( L \)-function is not very small (essentially as small as \( (\log k)^{-1} \), but if this is so, then one can show that most of the Fourier coefficients of \( f \) are even smaller at primes, and with this Holowinsky’s treatment becomes unconditional. So combining their results, one has

**Theorem 5 ([HS], Holomorphic QUE (noncompact surfaces)).** Let \( M \) be a noncompact arithmetic surface and \( f \) a holomorphic Hecke cusp form of weight \( k \). Then as \( k \to \infty \),

\[
\begin{align*}
\text{(i)} & \quad \nu_{f} \to dA/(\text{Area}M). \\
\text{(ii)} & \quad \text{The zeros of } f \text{ become equidistributed in } M \text{ with respect to } dA/(\text{Area}M).
\end{align*}
\]

Part (ii) is an elegant and fundamental fact about zeros of holomorphic Hecke eigenforms which at the present is proved using the full force of this QUE theory. On the face of it (ii) is far weaker than (i), so perhaps there is a more elementary proof of (ii)? As far as arithmetic QUE for surfaces goes, the only case that remains open is for holomorphic forms on a compact arithmetic surface. Soundararajan’s arguments apply in this case but not Holowinsky’s.

To end, we discuss briefly the higher dimensional cases of arithmetic QUE. The Hilbert modular varieties are the closest to the arithmetic surface case. Let \( n \geq 2 \), and let \( K \) be a totally real (that is, all Galois embeddings \( \sigma_1, \sigma_2, \ldots, \sigma_n \) of \( K \) are real) number field of degree \( n \). Let \( O_K \) be its ring of integers and \( \Gamma = SL_2(O_K) \) the corresponding group of \( 2 \times 2 \) matrices. Via the embedding of \( \Gamma(\gamma) \to (\sigma_1(\gamma),\sigma_2(\gamma),\ldots,\sigma_n(\gamma)) \) into \( G = SL_2(\mathbb{R}) \times \cdots \times SL_2(\mathbb{R}) \), we get a discrete subgroup of \( G \). The quotient \( M = \Gamma\backslash \mathbb{H} \times \mathbb{H} \times \cdots \times \mathbb{H} \) is a finite volume locally symmetric space, it is also a complex manifold and is known as a Hilbert modular variety.

The QUE problem for high frequency eigenstates can be formulated in this context. In the philosophy of diagonalizing geometrically defined commuting operators, we consider functions \( \phi(z_1, z_2, \ldots, z_n) \) which are \( \Gamma \)-invariant, are eigenfunctions of the full ring of differential operators on \( \mathbb{H} \times \mathbb{H} \times \cdots \times \mathbb{H} \) which commute with the action of \( \Gamma \). That is, \( \phi \) is a simultaneous eigenfunction of \( \Delta_{z_1}, \Delta_{z_2}, \ldots, \Delta_{z_n} \) (and not just of \( \Delta = \Delta_{z_1} + \cdots + \Delta_{z_n} \)). The probability density \( \nu_{\phi} = |\phi(z)|^2dv(z) \) has a natural microlocal lift \( \mu_{\phi} \), to \( \Gamma\backslash G \) see [LJ2]. Note that since \( n \geq 2 \), this is not the unit tangent bundle of \( M \) and correspondingly the geodesic flow on \( T_1(M) \) is not ergodic. The new feature is that, being eigenfunctions of all the \( \Delta_{z_j} \)’s, the quantum limits of the \( \mu_{\phi} \)’s on \( \Gamma\backslash G \) are invariant under the full multiparameter diagonal subgroup

\[
A = \left\\{ \left( \begin{array}{cc} a_1 & 0 \\ 0 & a_1^{-1} \end{array} \right), \left( \begin{array}{cc} a_2 & 0 \\ 0 & a_2^{-1} \end{array} \right), \ldots, \left( \begin{array}{cc} a_n & 0 \\ 0 & a_n^{-1} \end{array} \right) : a_j \neq 0 \right\} \right\}
\]
of $G$. For $n \geq 2$ this is exactly the setting of the measure rigidity Conjecture of Margulis [Ma] (which in turn has its roots in the “$\times 2, \times 3$” conjecture of Furstenberg [Ru]) which asserts that any such measure on a homogeneous space $\Gamma \backslash G$ which is invariant under a rank 2 or higher diagonal group should be rigid. While no cases of this conjecture are proven, versions in which one assumes some form of positive entropy are known ([E-K-L], [Li1]) (and again these have their roots in similar such theorems for “$\times 2, \times 3$” [Ru]). The positive entropy is established by using the full Hecke algebra, as in [B-L], and together with an extension of the arguments in [So1], QUE for these Hecke-Maass forms on $M$ is established.

The holomorphic QUE for these Hilbert modular varieties asserts that if $f(z) = f(z_1, z_2, \ldots, z_n)$ is a holomorphic Hecke cusp form of weight $k = (k_1, k_2, \ldots, k_n)$, then the probability densities $\nu_f = |f(z)|^2 y_1^{k_1} \ldots y_n^{k_n}$ on $M$ become equidistributed with respect to $dv(z)/\text{Vol}(M)$, as $|k| = |(k_1, \ldots, k_n)|$ goes to infinity. A version of Watson’s formula valid over any number field has been derived in [Ich], and as before from it and the corresponding subconvexity for the corresponding $L$-functions, all aspects of arithmetic QUE for Hilbert modular varieties would follow. As Marshall shows in his thesis [Mars], this holomorphic setting is another one in which the Holowinsky-Soundararajan miracle occurs. Firstly, the Ramanujan Conjectures are known for these forms (see [Bl]), and the main issue is to develop the analogue of Holowinsky’s method in this higher dimensional setting. The potential theoretic arguments which allow one to pass from the equidistribution of the densities $\nu_f$ to those of the zero divisors $Z_f$ of $f$, extend to higher dimensional complex manifolds. This is formulated and proven in the context of holomorphic sections of high tensor powers of a positive Hermitian line bundle on a complex manifold in [S-Z] (Marshall shows that these also extend to the noncompact Hilbert modular setting).

**Theorem 6** ([Mars], Holomorphic QUE (Hilbert modular forms)). Let $M$ be a Hilbert modular variety and $f$ a holomorphic Hecke cusp form of weight $k = (k_1, \ldots, k_n)$. Then as $|k| \to \infty$ with $\min_j k_j \geq |k|^\delta$ for some fixed $\delta > 0$,

1. $\nu_f \to dv/\text{Vol}(M)$,
2. The zero divisors $Z_f$ become equidistributed in $M$, either as the $(1,1)$ currents $\partial \bar{\partial} \log |f|$ supported on $Z_f$ or as subsets (with possible positive multiplicity) of $M$ with their induced Riemannian $2n - 2$ volumes.

Given how little is known about the zero divisors of such Hecke eigenforms in several variables, part (ii) is a pleasing application of the arithmetic QUE theory.

Unlike the periods to special values of $L$-functions relation in (10), which appears to be rather special, the measure rigidity approach to arithmetic QUE extends to quite general compact arithmetic manifolds, as has been shown by Silberman and Venkatesh [S-V1], [S-V2]. The microlocal extension of the $\nu_\phi$’s goes naturally from the locally symmetric space $M = \Gamma \backslash G/K$ to $\Gamma \backslash G$, rather than to the unit tangent bundle of $M$. Its construction requires some elaborate representation theory and, as in the Hilbert modular case, the quantum limits are invariant under a full diagonal subgroup $A$, whose rank is that of $G/K$. Their proof of positive entropy for the corresponding quantum limits of Hecke eigenforms clarifies the role played by homogeneous subvarieties in this connection.

---

Footnote: In fact this is the setting in which the connection between QUE and measure rigidity was first noted [Li2].
Finally, we point out that in the symplectic setting the analogue of QUE for strongly chaotic transformations may fail. The mathematical model is that of quantizing a symplectic transformation of a compact symplectic manifold. The simplest (and very degenerate) case is that of a linear area preserving transformation of the torus $\mathbb{R}^2/\mathbb{Z}^2; (x, y)^t \to A(x, y)^t$ with $A \in SL_2(\mathbb{Z})$. If $|\text{trace}(A)| > 2$, then the dynamics of iterating $A$ is ergodic and strongly chaotic, and in the literature it goes by the name of cat map. The eigenstates of the corresponding quantization can be studied in depth (see [Ru2]) and in particular the analogues of Hecke operators and QUE for their eigenstates is known [K-R]. In this setting, when the eigenvalues of the quantization are maximally degenerate, the analogue of QUE can fail ([F-N-D]). Some take this as a warning about the original QUE conjecture. Note that even though QUE fails here, the positive entropy analogue of Anantharaman is still true and is sharp ([Br]).

ACKNOWLEDGMENTS

The author would like to thank Alex Barnett, Holger Then, and F. Stroemberg for supplying him with Figures 1, 4, and 5, respectively. The corresponding eigenstates were computed using their respective algorithms. The author also thanks S. Nonnenmacher and S. Zelditch for their comments on an earlier version of the paper.

ABOUT THE AUTHOR

The author is a professor of mathematics at Princeton University and at The Institute for Advanced Study. His interests include number theory and mathematical physics.

REFERENCES


9This silly but catchy name comes from the picture in the book Ergodic problems of classical mechanics by Arnold and Avez in which a picture of a cat and its deformation under a couple of iterates of $A$ is depicted.


