M.E. & T.W.

Arithmetic Quantum Unique Ergodicity for $\Gamma\backslash\mathbb{H}$

February 4, 2010

Contents

1

Introduction to (A)QUE

Quantum unique ergodicity – QUE – is concerned with the distributional properties of high-frequency eigenfunctions of the Laplacian on a domain Ω, that is of solutions to the equation

$$\Delta\phi_j + \lambda_j \phi_j = 0$$

with the Dirichlet boundary conditions $\phi_j|_{\partial\Omega} = 0$ and normalization

$$\int_{\Omega} \phi_j^2 \,\mathrm{d}x \,\mathrm{d}y = 1,$$

where Δ is the appropriate Laplacian. There is a connection between the high-frequency states and the classical Hamiltonian dynamical system obtained by letting a billiard ball move inside Ω at unit speed, and bouncing off the boundary (if there is a boundary) of Ω with angle of incidence equal to the angle of reflection.

If the motion is integrable (for example, if Ω is a circle), then there are invariant sets with measure strictly between 0 and 1 (for example, if Ω is a circle, then the set of orbits tangent to a given concentric circle with radius in a given interval is an invariant set). At first the relationship between high-frequency eigenfunctions and distributional or ergodic properties may be surprising, but in the integrable case we can find eigenfunctions of the Laplacian that are highly localized on non-trivial invariant sets (see the survey [1, Figs. 1 & 2]). Even for the simplest of domains Ω a diversity of possibilities occurs. A special case to which ergodic methods may be applied comes from homogeneous dynamics – that is, to actions of subgroups of a Lie group G on quotients $\Gamma\backslash G$ of finite volume (see [12], [11] for examples and background). In this case quite strong conjectures were made by Rudnick and Sarnak [24].

1.1 (Arithmetic) Quantum Unique Ergodicity

In order to formulate the conjecture, we recall some notation (see [12, Chap. 9] for a detailed treatment).

- The upper half-plane model for the hyperbolic plane is
$$\mathbb{H} = \{z = x + \mathrm{i}y \in \mathbb{C} \mid y > 0\}.$$
- The group $\mathrm{SL}_2(\mathbb{R})$ acts transitively on $\mathbb{H}$ via Möbius transformations: the matrix $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ acts via
$$g : z \longmapsto g \cdot z = \frac{az + b}{cz + d}.$$
- Any subgroup $\Gamma \leqslant \mathrm{SL}_2(\mathbb{R})$ defines an associated quotient space $M = \Gamma\backslash\mathbb{H}$ under this action, and the quotient space inherits a measure vol_M from the volume measure $\mathrm{dvol}_M = \frac{1}{y^2}\,\mathrm{d}x\,\mathrm{d}y$ on $\mathbb{H}$, meaning that
$$\int_M f(x + \mathrm{i}y)\,\mathrm{dvol}_M = \int_F f(x + \mathrm{i}y)\frac{1}{y^2}\,\mathrm{d}x\,\mathrm{d}y,$$
where $F \subseteq M$ is a fundamental domain for the action of Γ on $\mathbb{H}$. This allows us to speak of function spaces like $L^2(M) = L^2_{\mathrm{vol}_M}(M)$.
- A sequence of measures $m_1, m_2, \ldots$ is said to converge weak* to a measure m, denoted by $m_i \underset{\text{weak*}}{\longrightarrow} m$ as $i \to \infty$, if
$$\int f\,\mathrm{d}m_i \longrightarrow \int f\,\mathrm{d}m$$
as $i \to \infty$ for any continuous function f with compact support.
- The *Laplacian* on M is the operator
$$\Delta = y^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right),$$
and a function $\phi \in C^\infty(M) \cap L^2(M)$ is an *eigenfunction* for Δ with *eigenvalue* λ if $\Delta\phi = \lambda\phi$. An eigenvalue is *normalized* if $\|\phi\|_2 = 1$.

Conjecture 1.1 (Quantum Unique Ergodicity; Rudnick–Sarnak). Let Γ be a discrete subgroup of $\mathrm{SL}_2(\mathbb{R})$ such that $M = \Gamma\backslash\mathbb{H}$ is compact. If $\{\phi_i \mid i \in \mathbb{N}\}$ are normalized eigenfunctions for Δ in $C^\infty(M)$ with corresponding eigenvalues $\{\lambda_i \mid i \in \mathbb{N}\}$ such that $\lambda_i \to \infty$ as $i \to \infty$, then
$$|\phi_i|^2\,\mathrm{dvol}_M \underset{\text{weak*}}{\longrightarrow} \mathrm{dvol}_M \tag{1.1}$$
as $i \to \infty$.

The motivation for this conjecture comes from physical considerations, but it has wide-ranging mathematical meaning. We address the motivation via a series of questions.

1.1.1 Why is Δ the differential operator studied?

(1) If $\nabla(f)$ denotes the total derivative of a function $f \in C_c^\infty(M)$ (that is, an infinitely differentiable function with compact support), then it can be shown that $\Delta(f)$ is equal to $-\nabla^*\overline{\nabla}(f)$, where ∇^* is the adjoint operator, and $\overline{\nabla}$ is the closed operator defined by ∇. This slightly mysterious observation suggests that Δ is a natural operator, and that its eigenvalues are negative.

(2) The operator Δ is the restriction of the *Casimir operator* Ω_c, which is a differential operator of degree two on $\mathrm{SL}_2(\mathbb{R})$ with unique invariance properties. In fact, Ω_c restricted to the space of functions on

$$\mathbb{H} = \mathrm{SL}_2(\mathbb{R})/\,\mathrm{SO}(2)$$

coincides with Δ (this will be discussed in more detail in Section 2.2). Here $\mathrm{SO}(2)$ denotes the special orthogonal group of matrices of the form $\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} = k_\theta$.

(3) In Schrödinger's quantum theory, the motion of a free (spinless, non-relativistic) quantum particle, moving in the absence of external forces on M, satisfies the equation

$$\mathrm{i}\frac{\partial\psi}{\partial t} = \Delta\psi.$$

This defines a *unitary evolution*, meaning that $\|\psi(\cdot,t)\|_{L^2(\mathrm{vol}_M)}$ is independent of t – so without loss of generality we may normalize and assume that $\|\psi(\cdot,t)\|_{L^2(\mathrm{vol}_M)} = 1$. The Born interpretation gives an empirical meaning to the "wave function" ψ by interpreting $|\psi|^2$ as the distribution of the position of the particle in the sense of probability. The eigenfunction equation $\Delta\psi = \lambda\psi$ corresponds to studying a particle with a given energy $-\lambda$. Thus the QUE conjecture concerns itself with the high-energy limit (also called the semi-classical limit). In fact the QUE conjecture implies a strengthening of the uncertainty principle: If ψ has a given large energy, then not only is the position of the particle uncertain, it is in fact almost equidistributed.

1.1.2 Why are there eigenfunctions?

(1) If M is compact, then the operator $(I - \Delta)^{-1}$ is a compact operator on $L^2(M)$. It follows that $L^2(M)$ is spanned by the eigenfunctions of Δ, and for every $\lambda \in \mathbb{R}$ the corresponding eigenspace

$$\{\phi \mid \Delta\phi = \lambda\phi\} \tag{1.2}$$

is finite-dimensional.

(2) If M is not compact, then $(I-\Delta)^{-1}$ is not a compact operator. In this case there may in general not be any eigenfunctions at all. However, if $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ (or a congruence subgroup, defined below), then one can again show that Δ has infinitely many eigenfunctions in $C^\infty(M) \cap L^2(M)$, and that the eigenspaces (1.2) are once again finite-dimensional. Rudnick and Sarnak also conjectured that in this case (1.1) should hold.

1.1.3 What other reasons are there to study eigenfunctions of Δ?

Apart from the quantum-mechanical interpretation in Section 1.1.1(3), the eigenfunctions of the Laplacian arise in many parts of mathematics.

(1) On compact quotients, they give the most canonical orthonormal basis of $L^2(M)$. This is part of the theory of harmonic analysis (the appropriate generalization of Fourier analysis) on M.
(2) The eigenfunctions, which are also called *Maass cusp forms*, are intimately related to L-functions in number theory.

1.1.4 The Result

Conjecture 1.1 is, in full generality, open. However, there are some important cases for which it is known. In order to describe these, we need to make a few more definitions. We call Γ a *congruence lattice over* $\mathbb{Q}$ if either

- Γ is a *congruence subgroup* of $\mathrm{SL}_2(\mathbb{Z})$, meaning that
$$\Gamma \supseteq \{\gamma \in \mathrm{SL}_2(\mathbb{Z}) \mid \gamma \equiv I_2 \bmod N\}$$
for some $N \geqslant 1$; or
- Γ is a lattice derived from a Eichler order in an $\mathbb{R}$-split quaternion division algebra over $\mathbb{Q}$.

The first type has the advantage of being quite concrete, and includes familiar examples like $\Gamma = \mathrm{SL}_2(\mathbb{Z})$; the second type has the advantage that in those cases the lattice is *uniform*, meaning that the quotient space $\Gamma\backslash\mathbb{H}$ is compact. In either case, it is possible to define a collection of additional operators $\{T_n\}$, called Hecke operators, which commute with Δ and with each other. These operators therefore act on the finite-dimensional eigenspaces (1.2), and are simultaneously diagonalizable. A *Hecke–Maass cusp form* is a joint eigenfunction $\phi \in C^\infty(M) \cap L^2(M)$ of Δ and of all the Hecke operators T_n for $n \geqslant 2$.

Lindenstrauss [18] and Soundarajan [28] together have shown the following, which we refer to as *arithmetic quantum unique ergodicity* (AQUE).

Theorem 1.2. *Let $M = \Gamma\backslash\mathbb{H}$, with Γ a congruence lattice over $\mathbb{Q}$. Then*

$$|\phi_i|^2 \,\mathrm{dvol}_M \underset{weak^*}{\longrightarrow} \mathrm{dvol}_M$$

as $i \to \infty$ for any sequence of Hecke–Maass cusp forms for which the Maass eigenvalues $\lambda_i \to -\infty$ as $i \to \infty$.

We briefly summarize some of the history leading up to this result. In 2001 Watson [30] showed this under the assumption of the Generalized Riemann Hypothesis (GRH), also obtaining under this assumption the optimal rate of convergence. In 2006 Lindenstrauss [18] obtained the result unconditionally, using ergodic methods, for lattices derived from Eichler orders and (almost) for congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$. For the latter case, Lindenstrauss showed that any weak*-limit is of the form $c\,\mathrm{dvol}_M$ for some $c \in [0,1]$ – in other words *escape of mass* to infinity was not ruled out. In 2009 Soundarajan [28] established, in a short paper of ten pages, that any weak*-limit is a probability measure – that is, escape of mass is not possible. Combined with [18], this proved Theorem 1.2.

1.2 Rigidity of Invariant Measures for the Geodesic Flow

Recall that the unit tangent bundle $\mathrm{T}^1\mathbb{H}$ of the hyperbolic plane is isomorphic to $\mathrm{PSL}_2(\mathbb{R}) = \mathrm{SL}_2(\mathbb{R})/\{\pm I_2\}$, by identifying the matrix $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with the point

$$\left(g \cdot \mathrm{i}, \frac{1}{(cz+d)^2}\mathrm{i}\right) \in \mathrm{T}^1\mathbb{H}.$$

Under this isomorphism, the geodesic flow (which, by definition, follows the geodesic determined by the arrow $(z, v) \in \mathrm{T}^1\mathbb{H}$ with unit speed, as illustrated in Figure 1.1) corresponds on $\mathrm{PSL}_2(\mathbb{R})$ to right-multiplication by

$$a_t = \begin{pmatrix} \mathrm{e}^{t/2} & \\ & \mathrm{e}^{-t/2} \end{pmatrix}$$

for $t \in \mathbb{R}$ (see [12, Chap. 9] for a detailed treatment).

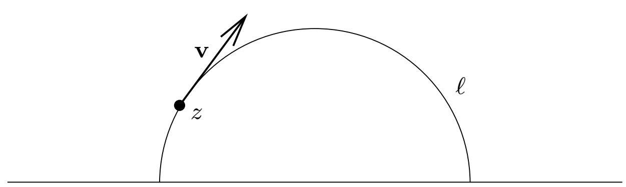

Fig. 1.1. The unique geodesic ℓ defined by a pair $(z, \mathbf{v})$.

This flow (that is, action of $\mathbb{R}$) naturally descends to the quotient

$$\mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R}),$$

or to any other quotient $\Gamma\backslash\mathrm{SL}_2(\mathbb{R})$. Recall that, by definition, Γ is a lattice if $X = \Gamma\backslash\mathrm{SL}_2(\mathbb{Z})$ supports an $\mathrm{SL}_2(\mathbb{R})$-invariant probability measure m_X, which we always assume. The measure m_X is also called the Haar measure of X and, if projected to $M = X/\mathrm{SO}(2)$, gives the normalized volume measure vol_M.

It is interesting to note that there are dense orbits of the geodesic flow on X. In fact, for almost every starting point $x \in X$ the orbit is *equidistributed* with respect to m_X, meaning that

$$\frac{1}{T}\int_0^T f(xa_t)\,\mathrm{d}t \longrightarrow \int_X f\,\mathrm{d}m_X$$

as $T \to \infty$ for any $f \in C_c(X)$. This is a consequence of the ergodicity of m_X with respect to the geodesic flow and Birkhoff's pointwise ergodic theorem. We mention these important but basic concepts from ergodic theory only in passing, as they will not be used in these lectures (see [12, Th. 2.30] for the pointwise ergodic theorem, [12, Sect. 4.4.2] for a discussion of generic points, [12, Sect. 9.5] for an account of Hopf's proof of ergodicity for the geodesic flow, and [12, Sect. 11.3] for an explanation of the 'Mautner phenomena' and ergodicity of the geodesic flow).

It is also interesting to note that there are many periodic orbits for the geodesic flow. For example, the matrix

$$\gamma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$$

is diagonalizable by some $k \in \mathrm{SO}(2)$ and has positive eigenvalues, so that

$$\mathrm{SL}_2(\mathbb{Z})ka_{t_0/2} = \mathrm{SL}_2(\mathbb{Z})\gamma k = \mathrm{SL}_2(\mathbb{Z})k$$

for some $t_0 > 0$, showing that $\mathrm{SL}_2(\mathbb{Z})k$ is periodic (as illustrated in Figure 1.2, where the usual fundamental domain for $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ is used; the details behind this form of illustration may be found in [12, Ch. 9]).

Clearly the periodic orbit is itself isomorphic to $\mathbb{R}/t_0\mathbb{Z}$, and the flow on the orbit corresponds under this isomorphism to translation on $\mathbb{R}/t_0\mathbb{Z}$. This gives rise to another type of invariant ergodic probability measure on $\Gamma\backslash\mathrm{SL}_2(\mathbb{R})$, namely the one-dimensional Lebesgue measure supported on a periodic orbit.

Taking convex combinations of m_X and one-dimensional Lebesgue measures on periodic orbits gives rise to many other invariant measures. However, these are not ergodic if they are proper convex combinations. One (of several) definitions of ergodicity for an invariant measure is extremality in the convex set of invariant probability measures. This implies, by Choquet's theorem, that any invariant probability measure on X is a convex combination* of invariant ergodic probability measures. Hence we would like to know if m_X as above and the periodic one-dimensional Lebesgue measures on periodic orbits are the only invariant ergodic probability measures for the geodesic flow. This

* This convex combination is really an integral over an entire probability space of ergodic measures; see Chapter 4, and [12, Ch. 6] for a detailed treatment.

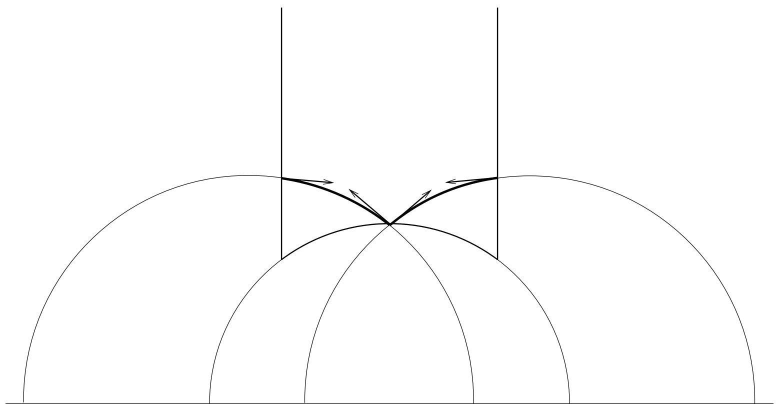

Fig. 1.2. A periodic orbit for the geodesic flow on $\mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R})$.

turns out to be very far from the truth; indeed for every $d \in [1, 3]$ there are many invariant ergodic probability measures for which the support* of the measure has Hausdorff dimension d.

We speak of *rigidity of invariant measures* for some group action if it is possible to give a complete classification of the invariant probability measures, and if the ergodic measures show a rich algebraic structure. This is, by the discussion, manifestly not the case for the geodesic flow. However, as the following theorem due to Lindenstrauss [18] shows, it is possible to give some (mild, and often checkable) additional conditions that characterize the Haar measure m_X. This theorem is related to earlier work of Rudolph [25], Host [14] and others on the unpublished conjecture of Furstenberg concerning measures on $\mathbb{R}/\mathbb{Z}$ invariant under $x \mapsto 2x \pmod 1$ and $x \mapsto 3x \pmod 1$, and of Katok and Spatzier [16] and of Einsiedler and Katok [8] on invariant measures for higher-rank diagonalizable flows in the direction of conjectures of Furstenberg, Katok and Spatzier, and Margulis. More surprising is that in [18] ideas from Ratner's work [23] on unipotent flows were also used, an unexpected connection because on the face of it the measures have very little structure with respect to the unipotent horocycle flow.

Theorem 1.3 (Lindenstrauss). *Let Γ be a congruence lattice over $\mathbb{Q}$, let $X = \Gamma\backslash\mathrm{SL}_2(\mathbb{R})$ and let μ be a probability measure satisfying the following properties:*

[I] *μ is* invariant *under the geodesic flow,*
[R]$_p$ *μ is* Hecke p-recurrent *for a prime p, and*
[E] *the* entropy *of every ergodic component of μ is positive for the geodesic flow.*

* The support of a measure μ is the smallest closed set A with $\mu(A) = 1$.

Then $\mu = m_X$ is the Haar measure on X.

The conditions are labeled [I] for invariance, $[\mathsf{R}]_p$ for recurrence and [E] for entropy. The method behind the theorem is more general, and has led to a number of further applications: Einsiedler, Katok and Lindenstrauss [9] applied this to obtain a partial result towards Littlewood's conjecture on simultaneous Diophantine approximation for pairs of real numbers, and Einsiedler, Lindenstrauss, Michel and Venkatesh [10] an application to the distribution of periodic orbits for the full diagonal flow on $\mathrm{SL}_3(\mathbb{Z})\backslash \mathrm{SL}_3(\mathbb{R})$.

For us, Theorem 1.3 will be used as a black box; we refer to the lecture notes of the Pisa Summer School in the Clay Mathematical Proceedings by Einsiedler and Lindenstrauss [6] for an introduction to the ideas and results needed in the proof. Instead we will focus on explaining the three assumptions in Theorem 1.3, and how they may be proved in order to deduce Theorem 1.2.

Finally, we note that conjecturally invariance [I] and recurrence $[\mathsf{R}]_p$ (for all primes p) should be sufficient to obtain the conclusion of Theorem 1.3. However, this is out of reach with current techniques in ergodic theory.

1.3 Introduction to Microlocal Lifts – Establishing Invariance

Šnirel′man [27], Colin de Verdière [5] and Zelditch [33] have shown in great generality (specifically, for any manifold on which the geodesic flow is ergodic with respect to the natural Liouvillian measure) that if one omits a subsequence of density zero from a complete sequence (ordered by eigenvalue) of eigenfunctions as in Conjecture 1.1, then for the remaining sequence one has

$$|\phi_i|^2 \,\mathrm{dvol}_M \underset{\text{weak*}}{\longrightarrow} \mathrm{dvol}_M.$$

An important component of their proof is the microlocal lift of any weak*-limit μ of the sequence $\big(|\phi_i|^2\,\mathrm{dvol}\big)$. The *microlocal lift* of such a limit is a measure $\widetilde{\mu}$ on the unit tangent bundle $\mathrm{T}^1 M$, invariant under the geodesic flow on $\mathrm{T}^1 M$, whose projection onto M is μ. The microlocal lift $\widetilde{\mu}$ is called a *quantum limit* of the sequence (ϕ_i). Rudnick and Sarnak also stated the following strengthened conjecture.

Conjecture 1.4 (QUE; Rudnick and Sarnak). Let $\Gamma \leqslant \mathrm{SL}_2(\mathbb{R})$ be a congruence lattice over $\mathbb{Q}$, or a discrete subgroup such that $M = \Gamma\backslash\mathbb{H}$ is compact. Then the Haar measure m_X for $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ is the only quantum limit.

What is actually shown in the proof of Theorem 1.2 is that any arithmetic quantum limit must be m_X. Here an *arithmetic quantum limit* is the microlocal lift $\widetilde{\mu}$ of a weak*-limit μ of $\big(|\phi_i|^2\,\mathrm{dvol}\big)$, where the ϕ_i are Hecke–Maass cusp forms.

1.4 Further Steps in the Proof

For the proof of Theorem 1.2, it is important that the construction of the microlocal lift has the property that the lift will have special properties arising from the assumption that the ϕ_i are eigenfunctions of the Hecke operators T_n, specifically properties $[\mathsf{R}]_p$ and $[\mathsf{E}]$. We will give such a construction of the micro-local lift in Chapter 2. This particular construction goes back to work of Wolpert [32], and could be described as a global construction of the micro-local lift. By this, we mean that we do not use charts of the manifold or other local tools, but instead define the lift on all of X at once.

It should now be clear how the proof of Theorem 1.2, assuming Theorem 1.3, works. After establishing property $[\mathsf{I}]$ in Chapter 2 we explain and prove the additional assumption $[\mathsf{R}]_p$ in Chapter 3 and property $[\mathsf{E}]$ in Chapter 4.

2

Establishing Invariance – The Micro-Local Lift

In this chapter we will assume for the most part that Γ is a lattice in $G = \mathrm{SL}_2(\mathbb{R})$, which by definition means that the homogeneous space $X = \Gamma\backslash G$ supports a probability measure m_X, called the Haar measure on X, which is invariant under the right action $x \mapsto xg^{-1}$ of $G \ni g$ on $X \ni x$. Recall that G acts on the upper half-plane $\mathbb{H}$ as in Section 1.1, and that if

$$K = \mathrm{SO}(2) = \left\{ k_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \mid \theta \in \mathbb{R} \right\} \tag{2.1}$$

denotes the orthogonal group, then

$$G/K \cong \mathrm{SL}_2(\mathbb{R})/\,\mathrm{SO}(2) \cong \mathbb{H},$$

where the coset $g\,\mathrm{SO}(2)$ corresponds to $g \cdot \mathrm{i} \in \mathbb{H}$. Moreover, if

$$U = \left\{ u_x = \begin{pmatrix} 1 & x \\ & 1 \end{pmatrix} \mid x \in \mathbb{R} \right\},$$

and

$$A = \left\{ a_y = \begin{pmatrix} y^{1/2} & \\ & y^{-1/2} \end{pmatrix} \mid y > 0 \right\}, \tag{2.2}$$

then we have the Iwasawa decomposition $G = UAK$ (which is the Gram–Schmidt orthonormalization process in disguise). This decomposition corresponds to the co-ordinates in $\mathbb{H}$ in the sense that $g = u_x a_y k_\theta$ maps i to $x + \mathrm{i}y$.

Next recall that the action of $g \in G$ on $\mathbb{H}$ is complex differentiable, and

$$\frac{d}{dz}\left(\frac{az+b}{cz+d}\right) = \frac{1}{(cz+d)^2}.$$

This gives rise to an action of G on the tangent bundle $\mathrm{T}\mathbb{H} = \mathbb{H} \times \mathbb{C}$ via

$$(\mathrm{D}\,g)(z,v) = (g \cdot z, \tfrac{1}{(cz+d)^2} v)$$

which preserves the hyperbolic Riemannian metric defined by

$$\langle (z,v_1),(z,v_2)\rangle_{\mathrm{T}_z\mathbb{H}} = \frac{1}{y^2}(v_1\cdot v_2),$$

where $z=y+\mathrm{i}y$ and $(v_1\cdot v_2)$ denotes the inner product of v_1 and v_2 viewed as elements of $\mathbb{C}\cong\mathbb{R}^2$. Here $\mathrm{T}_z\mathbb{H}=\{z\}\times\mathbb{C}$ is the tangent to $\mathbb{H}$ at z. The action $\mathrm{D}\,g$ for $g\in G$ restricted to the unit tangent bundle $\mathrm{T}^1\mathbb{H}=\{(z,v)\in \mathrm{T}\mathbb{H}\mid \|(z,v)\|_{\mathrm{T}_z\mathbb{H}}=1\}$ is almost simply transitive. In fact,

$$\mathrm{T}^1\mathbb{H}\cong\mathrm{PSL}_2(\mathbb{R})=\mathrm{SL}_2(\mathbb{R})/\{\pm I_2\}.$$

Finally, we let $M=\Gamma\backslash\mathbb{H}=\Gamma\backslash\mathrm{SL}_2(\mathbb{R})/\mathrm{SO}(2)$. Our goal in this chapter is to associate to every eigenfunction $\phi\in C^\infty(M)\cap L^2(M)$ of the hyperbolic Laplacian Δ a new (lifted) function $\widetilde{\phi}\in L^2(\Gamma\backslash G)=L^2(X)$ satisfying the following.

Theorem 2.1. *Let $\Gamma\leqslant\mathrm{SL}_2(\mathbb{R})$ be either a uniform lattice or a congruence lattice over $\mathbb{Q}$, and let $M=\Gamma\backslash\mathbb{H}$. Suppose that (ϕ_i) is an L^2-normalized* sequence of eigenfunctions of Δ in $C^\infty(M)\cap L^2(M)$, with the corresponding eigenvalues λ_i satisfying $|\lambda_i|\to\infty$ as $i\to\infty$, and assume that the weak*-limit μ of $|\phi_i|^2\,\mathrm{dvol}_M$ exists. If $\widetilde{\phi}_i$ denotes the sequence of lifted functions to be defined in Section* 2.5*, then (possibly after choosing a subsequence to achieve convergence) the weak*-limit $\widetilde{\mu}$ of $|\widetilde{\phi}_i|^2\,\mathrm{dvol}_M$ has the following properties:*

[L] *Projecting $\widetilde{\mu}$ on $X=\Gamma\backslash G$ to $M=\Gamma\backslash G/K$ gives μ.*
[I] *$\widetilde{\mu}$ is invariant under the right action of A.*

The measure $\widetilde{\mu}$ is called a microlocal lift *of μ, or a* quantum limit *of (ϕ_i).*

In fact, we will prove effective versions of [L] and [I], which will be statements about $|\phi|^2\,\mathrm{dvol}_M$ and $|\widetilde{\phi}|^2\,\mathrm{d}m_X$ involving negative powers of $|\lambda|$ in the error terms. In the proof we will make use of the following special properties of the eigenfunctions of Δ: The eigenfunctions are smooth, and moreover the eigenfunctions are bounded and all their partial derivatives are bounded.

2.1 The Lie Algebra and its Universal Enveloping Algebra

2.1.1 The Lie Algebra

Recall that

$$\mathfrak{sl}_2(\mathbb{R})=\{m\in\mathrm{Mat}_{22}(\mathbb{R})\mid \mathrm{tr}(m)=m_{11}+m_{22}=0\}$$

* This will always mean that $\|\phi\|_{L^2}=1$.

is the Lie algebra of $\mathrm{SL}_2(\mathbb{R})$. In fact, this follows quickly from

$$\det(\exp(m)) = \exp(\mathrm{tr}(m)).$$

Also recall that $g \in \mathrm{SL}_2(\mathbb{R})$ naturally acts on $m \in \mathfrak{sl}_2(\mathbb{R})$ by the adjoint representation (which is just a fancy name for conjugation),

$$\mathrm{Ad}_g(m) = gmg^{-1}.$$

This action satisfies

$$\exp(\mathrm{Ad}_g(m)) = g\exp(m)g^{-1}.$$

2.1.2 First Order Differential Operators

We now interpret $m \in \mathfrak{sl}_2(\mathbb{R})$ as a differential operator on $X = \Gamma\backslash\mathrm{SL}_2(\mathbb{R})$ (this would also make sense for any other smooth manifold carrying a smooth $\mathrm{SL}_2(\mathbb{R})$-action). Let $f \in C^\infty(X)$, then

$$(m * f)(x) = \left[\frac{\partial}{\partial t} f(x\exp(tm))\right]_{t=0}$$

defines a new smooth function $m * f \in C^\infty(X)$ (since $(x,t) \mapsto x\exp(tm)$ is smooth and f is smooth). It is easy to check, for example by applying the chain rule to the composition of $t \mapsto (t,t)$ with the smooth function

$$(t_1,t_2) \longmapsto f(x\exp(t_1\alpha_1 m_1)\exp(t_2\alpha_2 m_2))$$

on $\mathbb{R}^2$ at $t=0$ (and $(t_1,t_2)=(0,0)$ respectively) that the expression $m*f$ is linear in $m \in \mathfrak{sl}_2(\mathbb{R})$. Clearly, we should think of $(m*f)(x)$ as the partial derivative of f at x in the direction given by m.

Let us point out a few important special cases. The element

$$\mathcal{H} = \begin{pmatrix} \frac12 & \\ & -\frac12 \end{pmatrix} \in \mathfrak{sl}_2(\mathbb{R})$$

gives the differential operator in the direction of the geodesic flow corresponding to A. The element

$$\mathcal{W} = \begin{pmatrix} & 1 \\ -1 & \end{pmatrix} \in \mathfrak{sl}_2(\mathbb{R})$$

gives the differential operator in the direction of $K \cong \mathrm{SO}(2)$.

2.1.3 Second Order Differential Operators

If $m, w \in \mathfrak{sl}_2(\mathbb{R})$, then

$$C^\infty(\mathbb{R}) \ni f \longmapsto m*(w*f) = (m\circ w)*f \in C^\infty(\mathbb{R})$$

defines a second-order differential operator. As indicated, we will write $m \circ w$ for this operator, which is a formal product (and has, so far, nothing to do with matrix multiplication). By the linearity of $m * f$ in m discussed above, we see that $m \circ w$ is bilinear in m and w.

Recall that $\frac{\partial^2}{\partial t_1 \partial t_2} f = \frac{\partial^2}{\partial t_2 \partial t_1} f$ for any $f \in C^\infty(\mathbb{R}^2)$. This familiar relation does not generalize directly to our setting here. Instead we have the following important property.

Proposition 2.2. *For $m, w \in \mathfrak{sl}_2(\mathbb{R})$ we have*

$$m \circ w - w \circ m = [m, w]$$

where $[m, w] = mw - wm$ is the Lie bracket, defined by the difference of the matrix products. More concretely, this means that

$$m * (w * f) - w * (m * f) = ([m, w]) * f$$

for any $f \in C^\infty(X)$.

Proof. Let $f \in C^\infty(X)$. Then

$$w * f(x) = \left[\frac{\partial}{\partial t_1} f(x \exp(t_1 w)) \right]_{t_1 = 0},$$

and so

$$\begin{aligned} m * (w * f)(x) &= \left[\frac{\partial}{\partial t_2} (w \cdot f)(x \exp(t_2 m)) \right]_{t_2 = 0} \\ &= \left[\frac{\partial^2}{\partial t_1 \partial t_2} f(x \exp(t_2 m) \exp(t_1 w)) \right]_{t_1 = t_2 = 0}. \end{aligned}$$

Here the order of ∂t_1 and ∂t_2 does not matter, but the order of $\exp(t_2 m)$ and $\exp(t_1 w)$ does. However, we may write

$$\begin{aligned} \exp(t_2 m) \exp(t_1 w) &= \exp(t_1 w) \left[\exp(-t_1 w) \exp(t_2 m) \exp(t_1 w) \right] \\ &= \exp(t_1 w) \exp(\mathrm{Ad}_{\exp(-t_1 w)}\, t_2 m), \end{aligned}$$

which leads to

$$\begin{aligned} m * (w * f)(x) &= \left[\frac{\partial^2}{\partial t_1 \partial t_2} f(x \exp(t_1 w)) \exp(t_2\, \mathrm{Ad}_{\exp(-t_1 w)}\, m) \right]_{t_1 = t_2 = 0} \\ &= \left[\frac{\partial}{\partial t_1} \left(\mathrm{Ad}_{\exp(-t_1 w)}(m) \cdot f \right) (x \exp(t_1 w)) \right]_{t_1 = 0}. \end{aligned}$$

Applying the chain rule to the composition of the maps $t_1 \mapsto (t_1, t_1)$ and $(t, t') \mapsto \mathrm{Ad}_{\exp(-tw)} \cdot f(x \exp(t' w))$, we get

$$m * (w * f)(x) = \left[\frac{\partial}{\partial t}\left(\mathrm{Ad}_{\exp(-tw)}\, m\right) \cdot f(x)\right]_{t=0} + w * (m * f).$$

However, we note that $\mathrm{Ad}_{\exp(-tw)}\, m \in \mathfrak{sl}_2(\mathbb{R})$ and that $v \mapsto v * f(x)$ is linear in v. Hence it is sufficient to calculate

$$\begin{aligned}\left[\frac{\partial}{\partial t}\left(\mathrm{Ad}_{\exp(-tw)} m\right)\right]_{t=0} &= \left[\frac{\partial}{\partial t}\left(1 - tw + \frac{t^2}{2}w^2 + \cdots\right) m \left(1 + tw + \frac{t^2}{2}w^2 + \cdots\right)\right]_{t=0} \\ &= mw - wm, \end{aligned} \tag{2.3}$$

which gives $m \circ w = (mw - wm) + w \circ m$ as desired. □

For future reference, we note that (2.3) also shows that

$$\mathrm{ad}_w(m) = [w, m] = \left[\frac{\partial}{\partial t}\, \mathrm{Ad}_{\exp(tw)}\, m\right]_{t=0}. \tag{2.4}$$

2.1.4 The Universal Enveloping Algebra

The universal enveloping algebra $\mathscr{E}$ of $\mathfrak{sl}_2(\mathbb{R})$ is defined to be the formal associative infinite-dimensional algebra obtained by taking linear combinations of a unit 1, elements of $\mathfrak{sl}_2(\mathbb{R})$, and formal products $m_1 \circ m_2 \cdots m_d$ of all orders with $d \geqslant 2$ with the identification $m \circ w - w \circ m = [m, w] \in \mathfrak{sl}_2(\mathbb{R})$, together with all its formal consequences.

This construction should be contrasted with the more familiar construction of a Lie algebra from an associative algebra. Given any associative algebra A over the reals, defining the Lie bracket by $[u, v] = uv - vu$ gives A the structure of a Lie algebra.

The construction of the universal enveloping algebra does not reverse this process, but does associate to any Lie algebra $\mathfrak{l}$ over $\mathbb{R}$ a 'universal' (that is, most general) unital associative algebra $\mathscr{E}(\mathfrak{l})$ with the property that the Lie algebra constructed from $\mathscr{E}(\mathfrak{l})$ as above contains $\mathfrak{l}$. The real constraint in the construction of $\mathscr{E}(\mathfrak{l})$ from $\mathfrak{l}$ is to preserve the representation theory of $\mathfrak{l}$ in the sense that representations correspond one-to-one to modules over the associative algebra $\mathscr{E}(\mathfrak{l})$. In the case of a Lie algebra comprising elements that act as infinitesimal transformations, the universal enveloping algebra acts via differential operators. We refer to Knapp [17, Chap. III] for a more detailed discussion.

2.2 The Casimir Operator and the Laplacian

Recall that $\mathrm{SL}_2(\mathbb{R})$ acts on the Lie algebra $\mathfrak{sl}_2(\mathbb{R})$ by the adjoint representation (conjugation). This action extends to the universal enveloping algebra by defining

$$\mathrm{Ad}_g(m_1 \circ m_2) = (\mathrm{Ad}_g\, m_1) \circ (\mathrm{Ad}_g\, m_2), \tag{2.5}$$

and analogously on all higher order terms – this can be done coherently since $\mathrm{Ad}_g([u,v]) = [\mathrm{Ad}_g(u), \mathrm{Ad}_g(v)]$.

Recall (or this may be easily checked) that there is no non-zero element in $\mathfrak{sl}_2(\mathbb{R})$ which is fixed under Ad_g for all $g \in \mathrm{SL}_2(\mathbb{R})$; this makes the following result a little surprising.

Proposition 2.3. *Let*

$$\mathcal{H} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}, \mathcal{U}^+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \mathcal{U}^- = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

and

$$\mathcal{W} = \mathcal{U}^+ - \mathcal{U}^- = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Then the degree-2 element – called the Casimir element *(or* Casimir operator*) – defined by*

$$\begin{aligned} \Omega_c &= \mathcal{H}\circ\mathcal{H} + \tfrac{1}{2}\left(\mathcal{U}^+\circ\mathcal{U}^- + \mathcal{U}^-\circ\mathcal{U}^+\right) \\ &= \mathcal{H}\circ\mathcal{H} + \tfrac{1}{4}\left(\mathcal{U}^+ + \mathcal{U}^-\right)\circ\left(\mathcal{U}^+ + \mathcal{U}^-\right) - \tfrac{1}{4}\mathcal{W}\circ\mathcal{W}, \end{aligned}$$

is fixed under the action of $\mathrm{SL}_2(\mathbb{R})$ *(equivalently, under the derived action of* $\mathfrak{sl}_2(\mathbb{R})$*).*

Notice that the matrices $\mathcal{H}$ and $\mathcal{U}^{\pm}$ satisfy the fundamental identities

$$\left[\mathcal{H},\mathcal{U}^{\pm}\right] = \pm\mathcal{U}^{\pm}, \quad \left[\mathcal{U}^+,\mathcal{U}^-\right] = 2\mathcal{H}. \tag{2.6}$$

Proof of Proposition 2.3. Notice that the space $\mathcal{E}_{\leqslant 2}$ of elements with degree no more than 2 in the universal enveloping algebra is finite dimensional (indeed, it has dimension $1+3+6$), and that after choosing a basis the action is given by a smooth (in fact, polynomial) map

$$\phi : \mathrm{SL}_2(\mathbb{R}) \to \mathrm{GL}(\mathcal{E}_{\leqslant 2}).$$

Now $\mathrm{SL}_2(\mathbb{R})$ is connected, so Ω_c is fixed under the action of $\mathrm{SL}_2(\mathbb{R})$ if and only if

$$\frac{\partial}{\partial t}\mathrm{Ad}_{\exp(tw)}(\Omega_c) = 0 \tag{2.7}$$

for all $t \in \mathbb{R}$ and $w \in \mathfrak{sl}_2(\mathbb{R})$. However,

$$\mathrm{Ad}_{(t+t_0)w} = \mathrm{Ad}_{\exp(t_0 w)}\mathrm{Ad}_{\exp(tw)},$$

so it is sufficient to show (2.7) at $t = 0$.

Next notice that for a second-order term $m_1 \circ m_2$ we have to apply the product rule (which is a consequence of the appropriate use of the chain rule) in the calculation of $\frac{\partial}{\partial t}\mathrm{Ad}_{\exp(tw)}(m_1 \circ m_2)$ as the adjoint action is extended using (2.5). Finally, we have

$$\left[\frac{\partial}{\partial t}\operatorname{Ad}_{\exp(tw)}(m)\right]_{t=0}=[w,m]$$

by (2.4).

We now check that (2.7) holds at $t=0$ when w takes the values $\mathcal{H},\mathcal{U}^+$, and $\mathcal{U}^-$ which suffices to prove (2.7) at $t=0$ for all w, since those three elements span $\mathfrak{sl}_2(\mathbb{R})$.

Taking $w=\mathcal{H}$ gives

$$\begin{aligned}\left[\frac{\partial}{\partial t}\operatorname{Ad}_{\exp(t\mathcal{H})}(\Omega_c)\right]_{t=0} &= [\mathcal{H},\mathcal{H}]\circ\mathcal{H}+\mathcal{H}\circ[\mathcal{H},\mathcal{H}]\\ &\quad+\tfrac{1}{2}\Big([\mathcal{H},\mathcal{U}^+]\circ\mathcal{U}^- +\mathcal{U}^+\circ[\mathcal{H},\mathcal{U}^-]\\ &\qquad+[\mathcal{H},\mathcal{U}^-]\circ\mathcal{U}^+ +\mathcal{U}^-\circ[\mathcal{H},\mathcal{U}^+]\Big)\\ &= 0+0+\tfrac{1}{2}\big(\mathcal{U}^+\circ\mathcal{U}^- -\mathcal{U}^+\circ\mathcal{U}^- -\mathcal{U}^+ +\mathcal{U}^-\mathcal{U}^+\big)=0\end{aligned}$$

by (2.6). For $w=\mathcal{U}^+$ we also have

$$\begin{aligned}\left[\frac{\partial}{\partial t}\operatorname{Ad}_{\exp(t\mathcal{U}^+)}(\Omega_c)\right]_{t=0} &= [\mathcal{U}^+,\mathcal{H}]\circ\mathcal{H}+\mathcal{H}\circ[\mathcal{U}^+,\mathcal{H}]\\ &\quad+\tfrac{1}{2}\Big(0+\mathcal{U}^+\circ[\mathcal{U}^+,\mathcal{U}^-]\\ &\qquad+[\mathcal{U}^+,\mathcal{U}^-]\circ\mathcal{U}^+ +0\Big)\\ &= -\mathcal{U}^+\circ\mathcal{H}-\mathcal{H}\circ\mathcal{U}^+ +\tfrac{1}{2}\big(2\mathcal{U}^+\circ\mathcal{H}+\mathcal{H}\circ\mathcal{U}^+\big)=0,\end{aligned}$$

since $[\mathcal{U}^+,\mathcal{H}]=-[\mathcal{H},\mathcal{U}^+]=-\mathcal{U}^+$, by (2.6). The case $w=\mathcal{U}^-$ is similar.

Finally, the second formula for Ω_c is simply an algebraic reformulation using the definition of $\mathcal{W}$. □

We are interested in the Casimir element Ω_c because of the next result. For this, notice that any function f on $\Gamma\backslash\mathbb{H}$ can also be thought of* as a K-invariant function on $\Gamma\backslash G$.

Proposition 2.4. *For $f\in C^\infty(\Gamma\backslash\mathbb{H})\subseteq C^\infty(\Gamma\backslash\operatorname{SL}_2(\mathbb{R}))$, we have*

$$\Omega_c * f = y^2\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)f=\Delta f. \tag{2.8}$$

Sketch of Proof. We use the co-ordinate system

$$(x,y,\theta)\in\mathbb{R}\times(0,\infty)\times(\mathbb{R}/2\pi\mathbb{Z})$$

---

* Of course this also gives a lift, which we might denote $\widetilde{\phi}_{\mathrm{wrong}}$ of ϕ satisfying the desired lifting property [L](see p. 2.1). However, having made this choice it is not clear how to show the invariance property [I].

corresponding to the Iwasawa decomposition $u_x a_y k_\theta \in \mathrm{SL}_2(\mathbb{R})$, and use this to interpret the actions of the differential operators $\mathcal{H} * f$ and $(\mathcal{U}^+ + \mathcal{U}^-) * f$. As $\mathcal{W}$ is the derivative along K, we have $(\mathcal{W} \circ \mathcal{W}) * f = 0$ for $f \in C^\infty(\Gamma\backslash\mathbb{H})$.

We claim that

$$\mathcal{H}* = \left(-y \sin 2\theta \tfrac{\partial}{\partial x} + y \cos 2\theta \tfrac{\partial}{\partial y} + \tfrac{1}{2} \sin 2\theta \tfrac{\partial}{\partial \theta}\right), \tag{2.9}$$

$$\tfrac{1}{2}\left(\mathcal{U}^+ + \mathcal{U}^-\right) * = \left(y \cos 2\theta \tfrac{\partial}{\partial x} + y \sin 2\theta \tfrac{\partial}{\partial y} - \tfrac{1}{2} \cos 2\theta \tfrac{\partial}{\partial \theta}\right) \tag{2.10}$$

as operators on $C^\infty(\mathrm{SL}_2(\mathbb{R}))$. As every $f \in C^\infty(\mathrm{SL}_2(\mathbb{R}))$ is the restriction of a smooth function, also denoted f on $\mathrm{Mat}_{22}(\mathbb{R})$, we may use this ambient vector space to check the properties (2.9) and (2.10). Here notice that $\mathcal{H}*$ (and, similarly, $\frac{1}{2}(\mathcal{U}^+ + \mathcal{U}^-)*$) is defined by the right action, so that $\mathcal{H}*$ corresponds to taking the derivative in the direction of the vector

$$u_x a_y k_\theta \mathcal{H} \in \mathrm{Mat}_{22}(\mathbb{R}).$$

On the other hand, $\frac{\partial}{\partial x}$, $\frac{\partial}{\partial y}$, and $\frac{\partial}{\partial \theta}$ are the partial derivatives in the x, y, θ-coordinates of $g = u_x a_y k_\theta$, and so $\frac{\partial}{\partial x}$ corresponds to (the derivative in the direction of)

$$u_x \mathcal{U}^+ a_y k_\theta,$$

while $\frac{\partial}{\partial y}$ corresponds to

$$\frac{1}{y} u_x a_y \mathcal{H} k_\theta,$$

and $\frac{\partial}{\partial \theta}$ to

$$u_x a_y k_\theta \mathcal{W}.$$

Of these correspondences, perhaps the most mysterious is the one between $\frac{\partial}{\partial y}$ and $\frac{1}{y} u_x a_y \mathcal{H} k_\theta$, so we will explain this more carefully. Clearly, the derivative of $u_x a_y a_{\exp(t)} k_\theta$ with respect to t at $t = 0$ is $u_x a_y \mathcal{H} k_\theta$. However,

$$a_y a_{\exp(t)} = a_{y \exp(t)},$$

and so the derivative is y times the direction corresponding to $\frac{\partial}{\partial y}$. Using this, one can now calculate

$$\begin{aligned}
\mathcal{H}* &= -y \sin 2\theta \tfrac{\partial}{\partial x} + y \cos 2\theta \tfrac{\partial}{\partial y} + \tfrac{1}{2} \sin 2\theta \tfrac{\partial}{\partial \theta} \\
&= u_x \left(-y \sin 2\theta \mathcal{U}^+\right) a_y k_\theta + u_x a_y \left(\cos 2\theta \mathcal{H}\right) k_\theta + u_x a_y k_\theta \left(\tfrac{1}{2} \sin 2\theta \mathcal{W}\right) \\
&= u_x a_y \left(-\sin 2\theta \mathcal{U}^+ + \cos 2\theta \mathcal{H} + \tfrac{1}{2} \sin 2\theta \mathcal{W}\right) k_\theta \\
&= u_x a_y \begin{pmatrix} \frac{1}{2} \cos 2\theta & -\frac{1}{2} \sin 2\theta \\ -\frac{1}{2} \sin 2\theta & -\frac{1}{2} \cos 2\theta \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \\
&= \tfrac{1}{2} u_x a_y \begin{pmatrix} \cos 2\theta \cos\theta + \sin 2\theta \sin\theta & \cos 2\theta \sin\theta - \sin 2\theta \cos\theta \\ \cos 2\theta \sin\theta - \sin 2\theta \cos\theta & -\sin 2\theta \sin\theta - cos 2\theta \cos\theta \end{pmatrix},
\end{aligned}$$

which agrees with the calculation above. The proof for (2.10) is similar.

Assuming the relations (2.9) and (2.10), we find that $\mathcal{H} * \mathcal{H} *$ expands to

$$\Big(-y\sin 2\theta\tfrac{\partial}{\partial x}+y\cos 2\theta\tfrac{\partial}{\partial y}+\tfrac{1}{2}\sin 2\theta\tfrac{\partial}{\partial\theta}\Big)*\Big(-y\sin 2\theta\tfrac{\partial}{\partial x}+y\cos 2\theta\tfrac{\partial}{\partial y}+\big[\tfrac{1}{2}\sin 2\theta\tfrac{\partial}{\partial\theta}\big]\Big)$$

$$\begin{aligned}
&= \underline{y^2\sin^2 2\theta\tfrac{\partial^2}{\partial x^2}} - y^2\sin 2\theta\cos 2\theta\tfrac{\partial^2}{\partial x\partial y}\left[-\tfrac{y}{2}\sin^2 2\theta\tfrac{\partial^2}{\partial x\partial\theta}\right]\\
&\quad - \Big(-y^2\sin 2\theta\cos 2\theta\tfrac{\partial^2}{\partial x\partial y} - y\sin 2\theta\cos 2\theta\tfrac{\partial}{\partial x}\Big)\\
&\quad + \Big(\underline{y^2\cos^2 2\theta\tfrac{\partial^2}{\partial y^2}} + y\cos^2 2\theta\tfrac{\partial}{\partial y} + [\cdots]\Big)\\
&\quad + \tfrac{1}{2}\sin 2\theta\left(-2y\cos 2\theta\tfrac{\partial}{\partial x} - 2y\sin 2\theta\tfrac{\partial}{\partial\theta}\right) + [\cdots]\,,
\end{aligned}$$

where we have used the product rule several times, and indicate by $[\cdots]$ the terms involving $\frac{\partial}{\partial\theta}$, which are not important when this operator is applied to a function $f \in C^\infty(M)$. Similarly, one calculates

$$\Big(y\cos 2\theta\tfrac{\partial}{\partial x}+y\sin 2\theta\tfrac{\partial}{\partial y}-\tfrac{1}{2}\cos 2\theta\tfrac{\partial}{\partial\theta}\Big)*\Big(y\cos 2\theta\tfrac{\partial}{\partial x}+y\sin 2\theta\tfrac{\partial}{\partial y}-\big[\tfrac{1}{2}\cos 2\theta\tfrac{\partial}{\partial\theta}\big]\Big)$$

$$\begin{aligned}
&= \underline{y^2\cos^2 2\theta\tfrac{\partial^2}{\partial x^2}} + y^2\sin 2\theta\cos 2\theta\tfrac{\partial^2}{\partial x\partial y} - [\cdots]\\
&\quad + \Big(y^2\sin 2\theta\cos 2\theta\tfrac{\partial^2}{\partial x\partial y} + y\sin 2\theta\cos 2\theta\tfrac{\partial}{\partial x}\Big)\\
&\quad + \Big(\underline{y^2\sin^2 2\theta\tfrac{\partial^2}{\partial y^2}} + y^2\sin^2 2\theta\tfrac{\partial}{\partial y}\Big) + [\cdots]\\
&\quad - \tfrac{1}{2}\cos 2\theta\Big(-2y\sin 2\theta\tfrac{\partial}{\partial x} + 2y\cos 2\theta\tfrac{\partial}{\partial y}\Big) + [\cdots]\,.
\end{aligned}$$

Studying the above two expressions, we see that the underlined expressions together give Δ, while the remaining ten terms cancel out. □

2.3 *K*-finite Vectors, Raising and Lowering Operators

2.3.1 *K*-eigenfunctions and *K*-finite vectors

We may normalize the Haar (that is, Lebesgue) measure m_K on K to satisfy $m_K(K) = 1$. Then, using the usual Fourier expansion, the characters

$$e_n(k_\theta) = \mathrm{e}^{\mathrm{i}n\theta} \tag{2.11}$$

for $n \in \mathbb{Z}$ and $\theta \in [0, 2\pi)$ form a complete orthonormal basis of $L^2(K)$. Also, for $f \in C^\infty(K)$ there is a Fourier expansion

$$f = \sum_{n\in\mathbb{Z}} c_n e_n,$$

where $c_n = \langle f, e_n \rangle_{L^2(K)} \in \mathbb{C}$, converges absolutely since integration by parts shows that

$$|c_n| = |\langle c_n, e_n \rangle = |\left\langle f', \tfrac{1}{\mathrm{i}n} e_n \right\rangle| = \tfrac{1}{n^2} |\langle f'', e_n \rangle| \ll_f \tfrac{1}{n^2}.$$

Finally, note that

$$\begin{aligned} c_n e_n(k_\psi) &= \int f(k_\theta) \overline{e_n(k_\theta)} \, \mathrm{d}m_K e_n(k_\psi) \\ &= \int f(k_\theta) e_n(k_\psi k_\theta^{-1}) \, \mathrm{d}m_K \\ &= (f * e_n)(k_\psi) \end{aligned}$$

can also be obtained by convolution.

This generalizes to functions $f \in C^\infty(\mathrm{SL}_2(\mathbb{R}))$ as follows. We define

$$f_n(x) = f *_K e_n(x) = \int_K f(x k_\theta) e_n(-k_\theta) \, \mathrm{d}m_K(k_\theta),$$

which satisfies

$$\begin{aligned} f_n(x k_\psi) &= \int_K f(x k_\psi k_\theta) e_n(-k_\theta) \, \mathrm{d}m_K(k_\theta) \\ &= \int_K f(x k_{\theta'}) e_n(-k_{\theta'}) \, \mathrm{d}m_K(k_\theta) e_n(k_\psi) \\ &= f_n(x) e_n(k_\psi) \end{aligned}$$

by the substitution $k_{\theta'} = k_\psi k_\theta$. In other words, f_n is an eigenfunction for the right action of K on $C^\infty(\mathrm{SL}_2(\mathbb{R}))$ corresponding to the character e_n on K.

If we also assume that $f \in C_c^\infty(\mathrm{SL}_2(\mathbb{R}))$ then, again by integration by parts,

$$\begin{aligned} \|f_n\|_\infty &= \left\| \int_K f(x k_\theta) e_n(-k_\theta) \, \mathrm{d}m_K \right\|_\infty \\ &= \frac{1}{n^2} \left\| \int \left((\mathcal{W} \circ \mathcal{W}) * f \right) (x k_\theta) e_n(-k_\theta) \, \mathrm{d}m_K \right\|_\infty \ll_f \frac{1}{n^2}, \end{aligned}$$

which shows that $\sum_{n \in \mathbb{Z}} f_n$ converges uniformly to f. Similar results hold for $f \in L^2(X)$ and $f \in C_c^\infty(X)$ if $X = \Gamma \backslash \mathrm{SL}_2(\mathbb{R})$. For any of these spaces we define an associated subspace

$$\mathcal{A}_n = \{ f \mid f(x k_\theta) = e_n(k_\theta) f(x) \} \tag{2.12}$$

for $n \in \mathbb{Z}$; we will call the elements of $\mathcal{A}_n$ *K-eigenfunctions of weight n*.

In the case of the space $C^\infty(X)$, a simple calculation shows that $\mathcal{A}_\ell \mathcal{A}_n \subseteq \mathcal{A}_{\ell+n}$ for $\ell, n \in \mathbb{Z}$. A function that belongs to

$$\bigoplus_{n=-N}^{N} \mathcal{A}_n$$

for some N will be called K*-finite*.

Finally, note that for $L^2(X)$ the spaces of K-eigenfunctions of different weights are orthogonal, since for $f_\ell \in \mathcal{A}_\ell$, $f_n \in \mathcal{A}_n$ with $\ell \neq n$, there exists some θ with $e_{n-\ell}(k_\theta) \neq 1$, which forces

$$\langle f_\ell, f_n \rangle_{L^2(X)} = \langle f_\ell(xk_\theta), f_n(xk_\theta) \rangle_{L^2(X)} = e_\ell(k_\theta) e_{-n}(k_\theta) \langle f_\ell f_n \rangle$$

to be zero.

2.3.2 Raising and Lowering Operators

In Section 2.1.2 we showed that elements of $\mathfrak{sl}_2(\mathbb{R})$ are, in a natural sense, differential operators on $X = \Gamma \backslash \mathrm{SL}_2(\mathbb{R})$. This extends to

$$\mathfrak{sl}_2(\mathbb{C}) = \mathfrak{sl}_2(\mathbb{R}) + \mathrm{i}\mathfrak{sl}_2(\mathbb{R})$$

simply by defining the extension

$$(m_1 + \mathrm{i} m_2) * f = m_1 * f + \mathrm{i}(m_2 * f)$$

for $m_1, m_2 \in \mathfrak{sl}_2(\mathbb{R})$ and $f \in C^\infty(X)$ to be linear.

Almost all of the discussion in Section 2.1 also holds for the complex Lie algebra $\mathfrak{sl}_2(\mathbb{C})$. For example, the complex version of Proposition 2.2 follows quickly from its real counterpart as follows. If $m = m_1 + \mathrm{i} m_2$ and $w = w_1 + \mathrm{i} w_2$ with $m_1, m_2, w_1, w_2 \in \mathfrak{sl}_2(\mathbb{R})$, then

$$\begin{aligned}
(mw - wm) * f &= (m_1 + \mathrm{i} m_2) * (w_1 f + \mathrm{i} w_2 f) - (w_1 + \mathrm{i} w_2) * (m_1 f + \mathrm{i} m_2 f) \\
&= m_1 * (w_1 * f) - m_2 * (w_2 * f) + \mathrm{i}\,(m_1 * (w_2 * f) + m_2 * (w_1 * f)) \\
&\quad - w_1 * (m_1 * f) + w_2 * (m_2 * f) - \mathrm{i}\,(w_2 * (m_1 * f) + w_1 * (m_2 * f)) \\
&= ([m_1, w_1] - [m_2, w_2]) * f + \mathrm{i}\,([m_1, w_2] + [m_2, w_1]) * f
\end{aligned}$$

by collecting the appropriate terms and applying Proposition 2.2. On the other hand, using the complex matrix products we also see in the same manner that

$$[m, w] = mw - wm = ([m_1, w_1] - [m_2, w_2]) + \mathrm{i}\,([m_1, w_2] + [m_2, w_1]),$$

which implies the desired complex version.

We note, however, that $m * f$ for $m \in \mathfrak{sl}_2(\mathbb{C}) \smallsetminus \mathfrak{sl}_2(\mathbb{R})$ does not correspond to the partial derivative of f along the (non-existent) direction m. It is just, as defined, the complex linear combination of the partial derivatives along two (possibly different) directions.

Definition 2.5. *The* raising operator *is the element of* $\mathfrak{sl}_2(\mathbb{C})$ *given by*

$$\mathcal{E}^+ = \frac{1}{2}\begin{pmatrix} 1 & \mathrm{i} \\ \mathrm{i} & -1 \end{pmatrix} = \mathcal{H} + \frac{\mathrm{i}}{2}(\mathcal{U}^+ + \mathcal{U}^-), \tag{2.13}$$

and the lowering operator *is the element of* $\mathrm{SL}_2(\mathbb{C})$ *given by*

$$\mathcal{E}^- = \frac{1}{2}\begin{pmatrix} 1 & -\mathrm{i} \\ -\mathrm{i} & -1 \end{pmatrix} = \mathcal{H} - \frac{\mathrm{i}}{2}(\mathcal{U}^+ + \mathcal{U}^-). \tag{2.14}$$

The next proposition explains the terminology.

Proposition 2.6. *The operators $\mathcal{E}^+$ and $\mathcal{E}^-$ raise and lower in the sense that*

$$\begin{aligned} \mathcal{E}^+ &: \mathcal{A}_n \to \mathcal{A}_{n+2}, \\ \mathcal{E}^- &: \mathcal{A}_n \to \mathcal{A}_{n-2} \end{aligned}$$

for all $n \in \mathbb{Z}$.

Proof. We start by reformulating the definition of the subspace $\mathcal{A}_n$ in (2.12). We claim that

$$\mathcal{A}_n = \{f \mid \mathcal{W} * f = \mathrm{i}nf\}. \tag{2.15}$$

To see this, suppose first that $f(xk_\theta) = \mathrm{e}^{\mathrm{i}n\theta} f(x)$ as in (2.12). Then, taking the derivative along θ at $\theta = 0$ gives $\mathcal{W} * f = \mathrm{i}nf$, so the right-hand side of (2.15) contains $\mathcal{A}_n$. Conversely, assume that f is a smooth function with $\mathcal{W} * f = \mathrm{i}nf$. By the definition of $\mathcal{W}$, this means that

$$\left[\frac{\partial}{\partial\theta} f(xk_\theta)\right]_{\theta=0} = \mathrm{i}nf(x)$$

for all x, so

$$\frac{\partial}{\partial\theta}\left[\mathrm{e}^{-\mathrm{i}n\theta} f(xk_\theta)\right]_{\theta=\psi} = -\mathrm{i}nf(xk_\psi) + \mathrm{i}nf(xk_\psi) = 0,$$

in other words, $\mathrm{e}^{-\mathrm{i}n\theta} f(xk_\theta)$ is constant as required. Thus the right-hand side of (2.15) is contained in $\mathcal{A}_n$, proving the claim.

Notice that the matrix $\mathcal{W} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ can be diagonalized over $\mathbb{C}$, the column vectors of $m = \begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix}$ are the eigenvectors of $\mathcal{W}$, and that

$$\mathcal{W}m = \begin{pmatrix} \mathrm{i} & 0 \\ 0 & -\mathrm{i} \end{pmatrix} = 2\mathrm{i}m\mathcal{H}.$$

Finally, we calculate that

$$\mathcal{E}^+ m = \tfrac{1}{2}\begin{pmatrix} 1 & \mathrm{i} \\ \mathrm{i} & -1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix} = \tfrac{1}{2}\begin{pmatrix} 0 & 2 \\ 0 & 2\mathrm{i} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = m\mathcal{U}^+$$

and similarly $\mathcal{E}^- m = m\mathcal{U}^-$.

From this we see from (2.6) that

$$[\mathcal{W}, \mathcal{E}^\pm] = 2\mathrm{i}m[\mathcal{H}, \mathcal{U}^\pm]m^{-1} = \pm 2\mathrm{i}m(\mathcal{U}^\pm)m^{-1} = \pm 2\mathrm{i}\mathcal{E}^\pm,$$

which implies that

$$\begin{aligned}\mathcal{W} \cdot (\mathcal{E}^\pm \cdot f) &= \mathcal{E}^\pm \cdot (\mathcal{W} \cdot f) + [\mathcal{W}, \mathcal{E}^\pm] \cdot f \\ &= n\mathrm{i}\mathcal{E}^\pm \cdot f \pm 2\mathrm{i}\mathcal{E}^\pm \cdot f \\ &= (n \pm 2)\mathrm{i}\mathcal{E}^\pm \cdot f\end{aligned}$$

for $f \in \mathcal{A}_n$, as needed. □

We will also need formulas that relate $\mathcal{E}^\pm$ to the other differential operators that we introduced earlier. Note that complex conjugation extends to $\mathfrak{sl}_2(\mathbb{C}) = \mathfrak{sl}_2(\mathbb{R}) + \mathrm{i}\mathfrak{sl}_2(\mathbb{R})$ in the natural way.

Lemma 2.7. *We have* $\overline{\mathcal{E}^+} = \mathcal{E}^-$, $\mathcal{E}^+ + \mathcal{E}^- = 4\mathcal{H}$, *and*

$$\Omega_c = \mathcal{E}^- \circ \mathcal{E}^+ - \tfrac{1}{4}\mathcal{W} \circ \mathcal{W} - \tfrac{\mathrm{i}}{2}\mathcal{W} = \mathcal{E}^+ \circ \mathcal{E}^- - \tfrac{1}{4}\mathcal{W} \circ \mathcal{W} + \tfrac{\mathrm{i}}{2}\mathcal{W}.$$

Proof. The first two formulas are immediate from the definitions of $\mathcal{E}^+$ and $\mathcal{E}^-$ in (2.13) and (2.14). Moreover, the second formula for Ω_c follows from the first by complex conjugation. To derive the first formula for Ω_c, we calculate

$$\begin{aligned}\mathcal{E}^- \circ \mathcal{E}^+ - \tfrac{1}{4}\mathcal{W} \circ \mathcal{W} - \tfrac{1}{2}\mathrm{i}\mathcal{W} &= \left(\mathcal{H} - \tfrac{\mathrm{i}}{2}(\mathcal{U}^+ + \mathcal{U}^-)\right) \circ \left(\mathcal{H} + \tfrac{\mathrm{i}}{2}(\mathcal{U}^+ + \mathcal{U}^-)\right) \\ &\qquad - \tfrac{1}{4}\mathcal{W} \circ \mathcal{W} - \tfrac{1}{2}\mathrm{i}\mathcal{W} \\ &= \underbrace{\mathcal{H} \circ \mathcal{H} + \tfrac{1}{4}(\mathcal{U}^+ + \mathcal{U}^-) \circ (\mathcal{U}^+ + \mathcal{U}^-) - \tfrac{1}{4}\mathcal{W} \circ \mathcal{W}}_{\Omega_c} \\ &\qquad + \tfrac{\mathrm{i}}{2}\mathcal{H} \circ (\mathcal{U}^+ + \mathcal{U}^-) - \tfrac{\mathrm{i}}{2}(\mathcal{U}^+ + \mathcal{U}^-) \circ \mathcal{H} - \tfrac{\mathrm{i}}{2}\mathcal{W} \\ &= \Omega_c + \tfrac{\mathrm{i}}{2}[\mathcal{H}, \mathcal{U}^+ + \mathcal{U}^-] - \tfrac{\mathrm{i}}{2}\mathcal{W} \\ &= \Omega_c + \tfrac{\mathrm{i}}{2}(\mathcal{U}^+ - \mathcal{U}^-) - \tfrac{\mathrm{i}}{2}\mathcal{W} = \Omega_c.\end{aligned}$$

□

Lemma 2.8. *If $f \in C^\infty$ is K-finite, and $m \in \mathfrak{sl}_2(\mathbb{C})$, then $m * f$ is K-finite as well.*

Proof. This is clear for $m = \mathcal{W}$ and for $m = \mathcal{E}^\pm$, but $\mathcal{W}, \mathcal{E}^+$, and $\mathcal{E}^-$ span $\mathfrak{sl}_2(\mathbb{C})$. □

2.4 Differential Operators on $L^2(X)$ and their Adjoints

Recall that the action of $\mathrm{SL}_2(\mathbb{R})$ on $L^2(\Gamma\backslash\mathrm{SL}_2(\mathbb{R}))$ defined by

$$(g\cdot f)(x) = f(xg)$$

for $f \in L^2(\Gamma\backslash\mathrm{SL}_2(\mathbb{R}))$ and $g \in \mathrm{SL}_2(\mathbb{R})$ is unitary (since the right action of $\mathrm{SL}_2(\mathbb{R})$ on $\Gamma\backslash\mathrm{SL}_2(\mathbb{R})$ preserves the measure). On the other hand, elements of $\mathfrak{sl}_2(\mathbb{R})$ do not even define continuous operators on $L^2(\Gamma\backslash\mathrm{SL}_2(\mathbb{R}))$. Instead they define so-called *unbounded operators*, which are linear maps defined on a dense subset, called the *domain* of the operator. In fact, for $m \in \mathfrak{sl}_2(\mathbb{R})$ the operator $m*$ is well-defined on the subspace $C_c^\infty(\Gamma\backslash\mathrm{SL}_2(\mathbb{R}))$ (and also on some larger subspaces) which is dense in $L^2(\Gamma\backslash\mathrm{SL}_2(\mathbb{R}))$ with respect to $\|\cdot\|_2$.

The *adjoint of an unbounded operator* $\mathcal{O}$ defined on a dense subset $D \subseteq H$ of a Hilbert space H and mapping to a Hilbert space H' is the map that associates to $w \in H'$ the element $\mathcal{O}^*w \in H$ with the property

$$\langle \mathcal{O}v, w\rangle_{H'} = \langle v, \mathcal{O}^*w\rangle_H$$

for all $v \in D$. However, the element $\mathcal{O}^*w$ may not always exist; the domain $D^* \subseteq H'$ of $\mathcal{O}^*$ is defined to be the set of all $w \in H'$ for which $\mathcal{O}^*w$ exists. It is easy to check that for $w \in D^*$ the element $\mathcal{O}^*w$ is unique (that is, $\mathcal{O}^*$ is a map from D^* to H), that D^* is a linear subspace of H', and that $\mathcal{O}^*$ is linear on D^*.

We note that the definition of $\mathcal{O}^*$ is highly sensitive to both the operator $\mathcal{O}$ and its domain. For example, restricting $\mathcal{O}$ to a smaller (but still dense) subspace of D may change the meaning of $\mathcal{O}^*$.

2.4.1 Differential Operators on $L^2(\mathrm{SL}_2(\mathbb{R}))$

Proposition 2.9. *Let $m \in \mathfrak{sl}_2(\mathbb{C})$, and consider $m*$ as an unbounded operator on $L^2(\mathrm{SL}_2(\mathbb{R}))$ with domain $C_c^\infty(\mathrm{SL}_2(\mathbb{R}))$. Then*

$$(m*)^*f = -\overline{m} * f$$

for $f \in C_c^\infty(\mathrm{SL}_2(\mathbb{R}))$.

By definition, this means that for $f_1, f_2 \in C_c^\infty(\mathrm{SL}_2(\mathbb{R}))$ we have

$$\langle m * f_1, f_2\rangle = -\langle f_1, \overline{m} * f_2\rangle .$$

As we will see, this is simply a generalization of integration by parts.

Proof of Proposition 2.9. Recall that for $f_1, f_2 \in C_c^\infty(\mathbb{R})$ we have

$$0 = \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}x}(f_1\overline{f_2})\,\mathrm{d}x = \langle f_1', f_2\rangle_{L^2(\mathbb{R})} + \langle f_1, f_2'\rangle_{L^2(\mathbb{R})}\,, \tag{2.16}$$

where the first equality is an immediate consequence of the fact that the functions are both compactly supported. Similarly, for $f_1, f_2 \in C^\infty(K)$ we have

$$0 = \int_K \mathcal{W}(f_1\overline{f_2})\,\mathrm{d}m_K = \langle \mathcal{W}f_1, f_2\rangle_{L^2(K)} + \langle f_1, \mathcal{W}f_2\rangle_{L^2(K)}\,. \tag{2.17}$$

Now take $f_1, f_2 \in C_c^\infty(\mathrm{SL}_2(\mathbb{R}))$ and notice that

$$\begin{aligned}\left\langle \mathcal{U}^+ f_1, f_2\right\rangle_{L^2(\mathrm{SL}_2(\mathbb{R}))} &= \int_{\mathrm{SL}_2(\mathbb{R})} \left(\mathcal{U}^+ f_1\right)(x)\overline{f_2}(x)\,\mathrm{d}m_{\mathrm{SL}_2(\mathbb{R})}\\ &= \int_{KA}\int_U \left(\mathcal{U}^+ f_1\right)(k_\theta a_y u_x) f_2(k_\theta a_y u_x)\,\mathrm{d}x\,\mathrm{d}\mu(k_\theta a_y),\end{aligned} \tag{2.18}$$

since the Haar measure $m_{\mathrm{SL}_2(\mathbb{R})}$ can be written as a direct product $\mu\times m_U$ for some measure μ on KA. Now the inner integral (over U) in (2.18) is precisely of the form $\langle f_1', f_2\rangle_{L^2(\mathbb{R})}$ in (2.16). Hence we get

$$\begin{aligned}\left\langle \mathcal{U}^+ f_1, f_2\right\rangle &= -\int_{KA}\int_U f_1\left(\mathcal{U}^+ f_2\right)\,\mathrm{d}m_U\,\mathrm{d}\mu\\ &= -\left\langle f_1, \mathcal{U}^+ f_2\right\rangle.\end{aligned}$$

The proof for $\mathcal{U}^-$ is similar, and the proof for $\mathcal{W}$ uses (2.17) together with the fact that $m_{\mathrm{SL}_2(\mathbb{R})}$ can be written as the product $m_{UA}\times m_K$. Finally, notice that $\langle m\cdot f_1, f_2\rangle$ is linear in $m\in\mathfrak{sl}_2(\mathbb{C})$ and that $\langle f_1, m\cdot f_2\rangle$ is complex linear. This completes the proof. □

2.4.2 Differential Operators on $L^2(X)$ for Compact Quotients

Proposition 2.10. *Let $\Gamma < \mathrm{SL}_2(\mathbb{R})$ be a uniform lattice (so that the quotient space $X = \Gamma\backslash\mathrm{SL}_2(\mathbb{R})$ is compact), let $m\in\mathfrak{sl}_2(\mathbb{C})$ and consider $m*$ as an unbounded operator on $L^2(X)$ with domain $C^\infty(X)$. Then $(m*)^*f = -\overline{m}*f$ for $f\in C^\infty(\mathrm{SL}_2(\mathbb{R}))$.*

Proof. Recall from [12, Prop. 9.14] that for a discrete subgroup $\Gamma < \mathrm{SL}_2(\mathbb{R})$ and $x\in X = \Gamma\backslash\mathrm{SL}_2(\mathbb{R})$ there exists some $r = r_x > 0$ such that

$$B_r^{\mathrm{SL}_2(\mathbb{R})} \ni g \longmapsto xg \in B_r^X(x)$$

is injective. Moreover, for compact subsets of $\mathrm{SL}_2(\mathbb{R})$, such an injectivity radius $r>0$ can be chosen uniformly. Therefore, for a uniform lattice Γ we can write $X = O_1\cup\cdots\cup O_\ell$, where each $O_j = B_r^X(x_j)$ is an injective image of $B_r^{\mathrm{SL}_2(\mathbb{R})}$.

Let $\chi_1,\dots,\chi_\ell$ be a smooth partition of unity adapted to the decomposition

$$X = O_1 \cup \cdots \cup O_\ell,$$

which means that $\chi_j \in C^\infty(X)$, $\chi_j \geqslant 0$, $\mathrm{Supp}(\chi_j) \subseteq O_j$ for $j = 1, \ldots, \ell$, and $\sum_{j=1}^{\ell} \chi_j = 1$. Also choose functions $\psi_1, \ldots, \psi_\ell \in C^\infty(X)$ with the property that $\mathrm{Supp}(\psi_j) \subseteq O_j$ and with $\psi_j \equiv 1$ on a neighborhood of $\mathrm{Supp}(\chi_j)$. Then, for $f_1, f_2 \in C^\infty(X)$ and $m \in \mathfrak{sl}_2(\mathbb{C})$,

$$\begin{aligned}\langle m * f_1, f_2 \rangle_{L^2(X)} &= \sum_{j=1}^{\ell} \langle m * f_1, \chi_j f_2 \rangle_{L^2(X)} \\ &= \sum_{j=1}^{\ell} \langle m * (\psi_j f_1), \chi_j f_2 \rangle_{L^2(X)} .\end{aligned}$$

However, $\psi_j f_1, \chi_j f_2 \in C_c^\infty(O_j) \subseteq L^2(O_j)$, which we may identify with

$$C_c^\infty(B_r^{\mathrm{SL}_2(\mathbb{R})}) \subseteq L^2(\mathrm{SL}_2(\mathbb{R})).$$

Hence we may apply Proposition 2.9 to get

$$\begin{aligned}\langle m * f_1, f_2 \rangle_{L^2(X)} &= -\sum_{j=1}^{\ell} \langle \psi_j f_1, \overline{m} * (\chi_j f_2) \rangle_{L^2(X)} \\ &= -\sum_{j=1}^{\ell} \langle f_1, \overline{m} * (\chi_j f_2) \rangle_{L^2(X)} \\ &= -\langle f_1, \overline{m} * f_2 \rangle_{L^2(X)} .\end{aligned}$$

□

2.4.3 Differential Operators on $L^2(X)$ for Non-compact Quotients

Proposition 2.10 concerning a compact quotient easily generalizes to non-compact quotients if we restrict attention to $f_1, f_2 \in C_c^\infty(X)$. However, this will not be sufficient for our purposes as, for example, the eigenfunctions of Δ are not compactly supported. Rather (at least in the case where Γ is a congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$) they belong to

$$\mathcal{C} = \{ f \in C_{\mathrm{bd}}^\infty(X) \mid Df \in C_{\mathrm{bd}}^\infty(X) \text{ for all } D \in \mathcal{E}(\mathfrak{sl}_2(\mathbb{R})) \} \tag{2.19}$$

where $C_{\mathrm{bd}}^\infty(X) = C^\infty(X) \cap L^\infty(X)$.

Proposition 2.11. *Let $\Gamma \subseteq \mathrm{SL}_2(\mathbb{Z})$ be a congruence subgroup. For $f_1, f_2 \in \mathcal{C}$ and $m \in \mathfrak{sl}_2(\mathbb{C})$ we have*

$$\langle m * f_1, f_2 \rangle = -\langle f_1, \overline{m} * f_2 \rangle .$$

PROOF. Given some $\varepsilon > 0$ there exists some compact set $L \subseteq X$ such that $\|f_2 \mathbf{1}_{X \smallsetminus L}\|_2 < \varepsilon$ and $\|(m * f_2)\mathbf{1}_{X \smallsetminus L}\|_2 < \varepsilon$. We choose a smooth function $\phi \in C_c(X)$ with $0 \leqslant \phi \leqslant 1$, $\phi \equiv 1$ on a neighborhood of L, and such that $\|\overline{m} \cdot \phi\|_\infty \leqslant 1$ (this can be found concretely as a function $\phi \in C_c^\infty(\Gamma \backslash \mathbb{H})$). Furthermore, choose $\psi \in C_c(X)$ with $\psi \equiv 1$ on a neighborhood of $\operatorname{Supp} \phi$. With these functions we have, by applying the case of $C_c^\infty(X)$-functions, that

$$\begin{aligned} \langle m * f_1, f_2 \rangle &= \langle m * f_1, \phi f_2 \rangle + \mathrm{O}\left(\|m * f_1\|_2 \|(1-\phi) f_2\|_2\right) \\ &= \langle m * (\psi f_1), \phi f_2 \rangle + \mathrm{O}_{f_1}(\varepsilon) \\ &= -\langle \psi f_1, \overline{m} * (\phi f_2) \rangle + \mathrm{O}_{f_1}(\varepsilon) \\ &= -\langle f_1, (\overline{m} * \phi) f_2 + \phi(\overline{m} * f_2) \rangle + \mathrm{O}_{f_1}(\varepsilon) \\ &= -\langle f_1, \phi(\overline{m} * f_2) \rangle + \mathrm{O}\left(\|f_1\|_2 \|f_2 \mathbf{1}_{X \smallsetminus L}\|_2\right) + \mathrm{O}_{f_1}(\varepsilon) \\ &= -\langle f_1, \overline{m} * f_2 \rangle + \mathrm{O}\left(\|f_1\|_2 \|\overline{m} * f_2 \mathbf{1}_{X \smallsetminus L}\|_2\right) + \mathrm{O}_{f_1}(\varepsilon). \end{aligned}$$

As $\varepsilon > 0$ was arbitrary, the proposition follows. □

2.4.4 Some Formulas

In this section we assemble some further useful formulas relating the operators defined above. Let $\mathcal{C}$ be defined as in (2.19), which for X compact coincides with $C^\infty(X)$.

Corollary 2.12. *If $f \in \mathcal{C}$, then*

$$\begin{aligned} (\Omega_c *)^* f &= \Omega_c * f, \quad \textit{and} \\ (\mathcal{E}^\pm *)^* f &= -\mathcal{E}^\mp * f. \end{aligned}$$

Notice that in all of the classes of functions defined above (for example, the spaces $C_c^\infty(\mathrm{SL}_2(\mathbb{R}))$, $C^\infty(X)$ for compact X, $\mathcal{C}$ for non-compact X) the outcome of a differential operator or of its adjoint is in the same class, and so we can apply the operator many times. In this way we also arrive at formulas for the adjoint of elements of the universal enveloping algebra. For example, if f_1, f_2 are elements of the class considered, and $m, w \in \mathfrak{sl}_2(\mathbb{C})$, then

$$\begin{aligned} \langle m * (w * f_1), f_2 \rangle &= -\langle w * f_1, \overline{m} * f_2 \rangle \\ &= \langle f_1, \overline{w} * (\overline{m} * f_2) \rangle \end{aligned}$$

and so $((m * w)*)^* = (\overline{w} * \overline{m})*$.

PROOF OF COROLLARY 2.12. The first formula follows easily from the identity

$$\Omega_c = \mathcal{H} \circ \mathcal{H} + \tfrac{1}{2}(\mathcal{U}^+ \circ \mathcal{U}^- + \mathcal{U}^- \circ \mathcal{U}^+)$$

and the comment above. The second follows from the definitions in (2.13) and (2.14). □

Corollary 2.13. *If $f \in \mathcal{C} \cap \mathcal{A}_n$ satisfies $\Omega_c \cdot f = \lambda f$ with $\lambda = -(\frac{1}{4} + r^2)$ then*

$$\|\mathcal{E}^+ * f\|_2 = |\mathrm{i}r + \tfrac{1}{2} + \tfrac{1}{2}n| \|f\|_2, \text{ and}$$
$$\|\mathcal{E}^- * f\|_2 = |\mathrm{i}r + \tfrac{1}{2} - \tfrac{1}{2}n| \|f\|_2.$$

We will use these formulas repeatedly in our construction of the micro-local lift. Notice, for example, that $f = \phi \in \mathcal{C} \cap \mathcal{A}_0$ if $\phi \in C^\infty_{\mathrm{bd}}(M)$ is an eigenfunction of Δ.

Proof of Corollary 2.13. We have $(\mathcal{E}*)^* = -\mathcal{E}^- *$, and so

$$\begin{aligned} \|\mathcal{E}^+ * f\|_2^2 &= \left\langle \mathcal{E}^+ * f, \mathcal{E}^+ * f \right\rangle \\ &= -\left\langle (\mathcal{E}^- \circ \mathcal{E}^+) * f, f \right\rangle \\ &= -\left\langle (\Omega_c + \tfrac{1}{4}\mathcal{W} \circ \mathcal{W} + \tfrac{\mathrm{i}}{2}\mathcal{W}) * f, f \right\rangle \end{aligned}$$

by Lemma 2.7. Applying this to a function $f \in \mathcal{A}_n$ with eigenvalue λ gives

$$\|\mathcal{E}^+ \cdot f\|_2^2 = \left(-\lambda + \tfrac{1}{4}n^2 + \tfrac{1}{2}n\right) \|f\|_2^2.$$

As $\lambda = -\left(\frac{1}{4} + r^2\right)$ by definition of r, we have

$$\left(-\lambda + \tfrac{1}{4}n^2 + \tfrac{1}{2}n\right) = r^2 + \tfrac{1}{4} + \tfrac{1}{2}n + \tfrac{1}{4}n^2 = |\mathrm{i}r + \tfrac{1}{2} + \tfrac{1}{2}n|^2,$$

which gives the first formula of the corollary. The second follows by the same procedure, using the formula $(\mathcal{E}^- *)^* = -\mathcal{E}^+ *$ and the third formula of Lemma 2.7. □

2.5 The Micro-Local Lift

Recall that our goal is to define, for every eigenfunction ϕ of Δ on $M = X/K$, a new function $\widetilde{\phi}$ on X such that $\int f|\phi|^2 \,\mathrm{dvol}_M \approx \int f|\widetilde{\phi}|^2 \,\mathrm{d}m_X$ for every $f \in C_c^\infty(M)$, and $|\widetilde{\phi}|^2 \,\mathrm{d}m_X$ defines a measure that is (in a suitable sense which will be described in Corollary 2.20) almost invariant under the geodesic flow, which will then prove Theorem 2.1. After all the preparations above, we are ready to define $\widetilde{\phi}$.

Definition 2.14. *Inductively define functions by*

$$\phi_0(x) = \phi(xK) \in \mathcal{A}_0,$$

and

$$\phi_{2n+2} = \frac{1}{\mathrm{i}r + \frac{1}{2} + n} \mathcal{E}^+ * \phi_{2n} \in \mathcal{A}_{2n+2} \text{ for } n \geqslant 0, \tag{2.20}$$

$$\phi_{2n-2} = \frac{1}{\mathrm{i}r + \frac{1}{2} - n} \mathcal{E}^- * \phi_{2n} \in \mathcal{A}_{2n-2} \text{ for } n \leqslant 0. \tag{2.21}$$

Proposition 2.15. *The functions defined in Definition* 2.14 *have the following properties.*

(1) $\|\phi_{2n}\|_2 = 1$ *for all* $n \in \mathbb{Z}$.
(2) *The formulas* (2.20) *and* (2.21) *hold for all* $n \in \mathbb{Z}$*; equivalently,*

$$\mathcal{E}^- \mathcal{E}^+ \phi_{2n} = \left(\lambda - n^2 - n\right) \phi_{2n}$$

for all $n \in \mathbb{Z}$.

Proof. Property (1) follows from Corollary 2.13, since $\phi \in \mathcal{A}_{2n}$ by induction. For (2), recall that

$$\mathcal{E}^- \circ \mathcal{E}^+ = \Omega_c + \tfrac{1}{4}\mathcal{W} \circ \mathcal{W} + \tfrac{\mathrm{i}}{2}\mathcal{W},$$

which implies that

$$\left(\mathcal{E}^- \circ \mathcal{E}^+\right) * \phi_{2n} = \left(\lambda - n^2 - n\right) \phi_{2n}$$

for all $n \in \mathbb{Z}$, since $\Omega_c *$ commutes with the action of $\mathfrak{sl}_2(\mathbb{C})$, which is used to define ϕ_{2n} inductively starting from ϕ_0, and ϕ_0 has $\Omega_c * \phi_0 = \lambda\phi_0$. Now suppose that $n \geqslant 0$; then by definition and the identities above,

$$\begin{aligned}
\mathcal{E}^- * \phi_{2n+2} &= \frac{1}{\mathrm{i}r + \frac{1}{2} + n} \mathcal{E}^- \circ \mathcal{E}^+ * \phi_{2n} \\
&= \frac{-(\frac{1}{4} + r^2) - n^2 - n}{\mathrm{i}r + \frac{1}{2} + n} \phi_{2n} \\
&= \frac{-|\mathrm{i}r + \frac{1}{2} + n|^2}{\mathrm{i}r + \frac{1}{2} + n} \phi_{2n} \\
&= \left(\mathrm{i}r - \tfrac{1}{2} - n\right) \phi_{2n},
\end{aligned}$$

which coincides with (2.21) for $(n+1)$ instead of n. The proof of (2.20) for all $n \in \mathbb{Z}$ is similar. □

Definition 2.16. *Define, for* $N = N(\lambda)$ *to be chosen later,*

$$\widetilde{\phi} = \frac{1}{\sqrt{2N+1}} \sum_{n=-N}^{N} \phi_{2n},$$

Notice that $\|\widetilde{\phi}\|_2 = 1$ since each ϕ_{2n} has norm 1 by Proposition 2.15 and the distinct terms in the sum are mutually orthogonal since $\phi_{2n} \in \mathcal{A}_{2n}$.

2.5.1 Almost Lifts

The next result is the main step towards establishing property [L] in Theorem 2.1.

Theorem 2.17. *Let $\phi \in L^2(M)$ be an eigenfunction of Δ with corresponding eigenvalue $\lambda = -(\frac{1}{2} + r^2)$. If $f \in C_c^\infty(M)$, then*

$$\begin{aligned} \int f|\widetilde{\phi}|^2 \, \mathrm{d}m_X &= \langle f\phi, \phi \rangle_{L^2(M)} + \mathrm{O}\left(Nr^{-1}\right) \\ &= \int f|\phi|^2 \, \mathrm{dvol}_M + \mathrm{O}\left(Nr^{-1}\right). \end{aligned}$$

More generally, if f is a K-finite function in $C_c^\infty(M)$, then

$$\int f|\widetilde{\phi}|^2 \, \mathrm{d}m_X = \left\langle f \sum_{n=-N}^{N} \phi_{2n}, \phi \right\rangle_{L^2(X)} + \mathrm{O}_f\left(\max\{N^{-1}, Nr^{-1}\}\right)$$

Proof. Recall that $\mathcal{A}_\ell \mathcal{A}_n \subseteq \mathcal{A}_{\ell+n}$, which we will use below without explicit reference. Suppose that

$$f \in \sum_{\ell=-L}^{L} \mathcal{A}_{2\ell}$$

is K-finite (the case $L = 0$ is the first case of the theorem). Then

$$\left\langle f\widetilde{\phi}, \widetilde{\phi} \right\rangle = \frac{1}{2N+1} \sum_{m,n=-N}^{N} \langle f\phi_{2m}, \phi_{2n} \rangle,$$

by definition of $\widetilde{\phi}$.

The case $Nr^{-1} \geqslant 1$ is true, but not so interesting for our purposes:

$$\int f|\widetilde{\phi}|^2 \, \mathrm{d}m_X = \mathrm{O}\left(\|f\|_\infty\right) = \mathrm{O}_f(1),$$

and for $n \notin [-L, L]$ we see that

$$f\phi_{2n} \in \sum_{\ell=-L}^{L} \mathcal{A}_{2(\ell+n)}$$

is orthogonal to $\phi \in \mathcal{A}_0$, which also shows that

$$\left\langle f \sum_{n=-N}^{N} \phi_{2n}, \phi \right\rangle = \mathrm{O}\left(\|f\|_\infty L\right) = \mathrm{O}_f(1).$$

Suppose therefore that $Nr^{-1} \leqslant 1$. Then by Proposition 2.15(2) ((2.20) applied for ϕ_{2m-2}), we have

$$\begin{aligned}
\langle f\phi_{2m}, \phi_{2n}\rangle &= \frac{1}{\mathrm{i}r + \frac{1}{2} + m - 1} \left\langle f\mathcal{E}^{+} * \phi_{2m-2}, \phi_{2n}\right\rangle \\
&= \frac{1}{\mathrm{i}r - \frac{1}{2} + m} \left[\left\langle \mathcal{E}^{+} * (f\phi_{2m-2}), \phi_{2n}\right\rangle - \underbrace{\left\langle (\mathcal{E}^{+} * f)\phi_{2m+2}, \phi_{2n}\right\rangle}_{\mathrm{O}_f(1)} \right] \\
&= \frac{-1}{\mathrm{i}r - \frac{1}{2} + m} \left\langle f\phi_{2m-2}, \mathcal{E}^{-} * \phi_{2n}\right\rangle + \mathrm{O}_f(r^{-1}),
\end{aligned}$$

where we have also used the relation $(\mathcal{E}^{+}*)^{*} = -(\mathcal{E}^{-})$ from Corollary 2.12. Together with (2.21) for ϕ_{2n-2}, this gives

$$\begin{aligned}
\langle f\phi_{2m}, \phi_{2n}\rangle &= \frac{-(-\mathrm{i}r + \frac{1}{2} - n)}{\mathrm{i}r - \frac{1}{2} + m} \langle f\phi_{2m-2}, \phi_{2n-2}\rangle + \mathrm{O}_f(r^{-1}) \\
&= \frac{\mathrm{i}r - \frac{1}{2} + m + (n-m)}{\mathrm{i}r - \frac{1}{2} + m} \langle f\phi_{2m-2}, \phi_{2n-2}\rangle + \mathrm{O}_f(r^{-1}) \\
&= \langle f\phi_{2m-2}, \phi_{2n-2}\rangle + \mathrm{O}_f\left(\tfrac{|n-m|}{r}\right) + \mathrm{O}_f\left(r^{-1}\right). \qquad (2.22)
\end{aligned}$$

However, recall that $f \in \sum_{\ell=-L}^{L} \mathcal{A}_{2\ell}$, so that either $|n-m| \leqslant 2L = \mathrm{O}_f(1)$ or

$$\langle f\phi_{2m}, \phi_{2n}\rangle = 0 = \langle f\phi_{2m-2}, \phi_{2n-2}\rangle .$$

Therefore, we may write $O_f\left(\frac{|n-m|}{r}\right) = \mathrm{O}_f(r^{-1})$ in (2.22). Iterating this equation $|n| \leqslant N$ times gives

$$\langle f\phi_{2m}, \phi_{2n}\rangle = \left\langle f\phi_{2(m-n)}, \phi_0\right\rangle + \mathrm{O}_f(Nr^{-1}).$$

Summing over $m, n \in [-N, N]$ with $|n-m| \leqslant 2L$ (there are $\mathrm{O}_f(N)$ summands), and dividing by $2N+1$ gives

$$\begin{aligned}
\left\langle f\widetilde{\phi}, \widetilde{\phi}\right\rangle &= \frac{1}{2N+1} \sum_{m,n=-N}^{N} \left\langle f\phi_{2(m-n)}, \phi_0\right\rangle + \mathrm{O}_f(Nr^{-1}) \\
&= \sum_{\ell=-L}^{L} \frac{2N+1-|\ell|}{2N+1} \langle f\phi_{2\ell}, \phi_0\rangle + \mathrm{O}_f(Nr^{-1}), \qquad (2.23)
\end{aligned}$$

since $2N+1-|\ell|$ is the number of ways in which ℓ can be written as $m-n$ with $m, n \in [-N, N]$. Now $\frac{2N+1-|\ell|}{2N+1} = 1 + \mathrm{O}_f\left(N^{-1}\right)$ and $\langle f\phi_{2\ell}, \phi_0\rangle = \mathrm{O}_f(1)$, we finally get

$$\begin{aligned}
\left\langle f\widetilde{\phi}, \widetilde{\phi}\right\rangle &= \left\langle f \sum_{\ell=-L}^{L} \phi_{2\ell}, \phi_0 \right\rangle + \mathrm{O}_f(N^{-1}) + \mathrm{O}_f(Nr^{-1}) \\
&= \left\langle f \sum_{\ell=-N}^{N} \phi_{2\ell}, \phi_0 \right\rangle + \mathrm{O}_f(N^{-1}) + \mathrm{O}_f(Nr^{-1})
\end{aligned}$$

since $\langle f\phi_{2\ell}, \phi_0\rangle = 0$ if $|\ell| > L$, as stated in the theorem. If $f \in \mathcal{A}_0$, then the last step is not needed and (2.23) is already the claim of the theorem. □

This gives us [L] in Theorem 2.1.

Corollary 2.18. *Suppose that $N = N(\lambda)$ is a function of λ chosen so that $Nr^{-1} = \mathrm{O}(N|\lambda|^{-1/2}) \to 0$ as $|\lambda| \to \infty$. Assume also that (ϕ_i) is a sequence of Maass cusp forms, with corresponding eigenvalues $|\lambda_i| \to \infty$ as $i \to \infty$, and that $|\phi_i|^2 \,\mathrm{dvol}_M$ converges weak*. Then any weak*-limit of $|\widetilde{\phi}|^2 \,\mathrm{d}m_X$ projects to the weak*-limit of $|\phi_i|^2 \,\mathrm{dvol}_M$.*

Proof. This follows from Theorem 2.17, since a Borel measure is uniquely determined by how it integrates smooth functions in $C_c^\infty(M)$. □

2.5.2 Almost-Invariance

Clearly up to this point we could have chosen $N = 0$. However, for the proof of [I] in Theorem 2.1 we will need the second formula in Theorem 2.17 and the following result, both of which require $N \to \infty$ as $|\lambda| \to \infty$.

Theorem 2.19 (Zelditch). *If $f \in C_c^\infty(X)$ is a K-finite function, and N is sufficiently large (the lower bound depends on f), then*

$$\left\langle [(r\mathcal{H} + \mathcal{V}) * f] \sum_{n=-N}^{N} \phi_{2n}, \phi_0 \right\rangle = 0$$

for some fixed degree-two differential operator $\mathcal{V}$. In particular,

$$\left\langle (\mathcal{H} * f) \sum_{n=-N}^{N} \phi_{2n}, \phi_0 \right\rangle = \mathrm{O}_f(r^{-1}).$$

As we will see, the proof only uses the identities

$$\mathcal{E}^- \circ \mathcal{E}^+ * \phi_0 = \Omega_c * \phi_0 = \lambda\phi_0,$$

$$\Omega_c * \sum_{n=-N}^{N} \phi_{2n} = \lambda \sum_{n=-N}^{N} \phi_{2n},$$

and the product rule, which will produce extra terms.

Proof. First notice that for $f_1, f_2 \in \mathcal{C}$, we have

$$\begin{aligned} 0 &= \langle f_1 f_2, \mathcal{W} * \phi_0\rangle = \langle \mathcal{W} * (f_1 f_2), \phi_0\rangle \\ &= \langle (\mathcal{W} * f_1) f_2, \phi_0\rangle + \langle f_1 \mathcal{W} * f_2, \phi_0\rangle, \end{aligned}$$

or equivalently

$$\langle (\mathcal{W} * f_1) f_2, \phi_0 \rangle = - \langle f_1 (\mathcal{W} * f_2), \phi_0 \rangle \,. \tag{2.24}$$

Also notice that we may write (2.20) as

$$\mathcal{E}^+ * \sum_{n=-N}^{N} \phi_{2n} = (\mathrm{i}r - \tfrac{1}{2}\mathcal{W} - \tfrac{1}{2}) * \sum_{n=-N+1}^{N+1} \phi_{2n+2}$$

and (2.21) as

$$\mathcal{E}^- * \sum_{n=-N}^{N} \phi_{2n} = (\mathrm{i}r + \tfrac{1}{2}\mathcal{W} - \tfrac{1}{2}) * \sum_{n=-N-1}^{N-1} \phi_{2n-2}.$$

We define the shorthand

$$\psi = \sum_{n=-N}^{N} \phi_{2n},$$

so that $\mathcal{E}^\pm * \psi$ and $(\mathrm{i}r \mp \frac{1}{2}\mathcal{W} - \frac{1}{2}) * \psi$ differ by the sum of two K eigenfunctions of weight about $\pm 2N$. In the formulas below we will look at inner products of the form $\langle F\mathcal{E}^\pm * \psi, \phi\rangle$ with $F \in \{f, \mathcal{E}^+ f, \mathcal{E}^- f\}$, which are all K-finite with weights in the range $\{-2L-2, \ldots, 2L+2\}$.

This difference between $\mathcal{E}^\pm * \psi$ and $(\mathrm{i}r \mp \frac{1}{2}\mathcal{W} - \frac{1}{2}) * \psi$ is not significant once N is large enough. In fact, if $N > L + 2$, then $F\phi_{\pm 2N}$, $F\phi_{\pm 2N \pm 2}$ is orthogonal to ϕ_0. Finally, recall that $\mathcal{E}^+ + \mathcal{E}^- = 4\mathcal{H}$ and $\Omega_c = \mathcal{E}^- \cdot \mathcal{E}^+ - \frac{1}{4}\mathcal{W} \cdot \mathcal{W} - \frac{\mathrm{i}}{2}\mathcal{W}$ by Lemma 2.7.

With these formulas and the product rule for differentiation, we get

$$\begin{aligned}
\lambda \langle f\psi, \phi_0 \rangle &= \left\langle f\psi, \mathcal{E}^- \circ \mathcal{E}^+ * \phi_0 \right\rangle \\
&= \left\langle \mathcal{E}^- \circ \mathcal{E}^+ * (f\psi), \phi_0 \right\rangle \\
&= \Big\langle (\mathcal{E}^- \circ \mathcal{E}^+ * f)\psi + (\mathcal{E}^+ * f)(\mathcal{E}^- * \psi) + (\mathcal{E}^- * f)(\mathcal{E}^+ * \psi) \\
&\qquad\qquad + f(\mathcal{E}^- \circ \mathcal{E}^+ * \psi), \phi_0 \Big\rangle \\
&= \Big\langle (\mathcal{E}^- \circ \mathcal{E}^+ * f)\psi + (\mathcal{E}^+ * f)(\mathrm{i}r + \tfrac{\mathrm{i}}{2}\mathcal{W} - \tfrac{1}{2}) * \psi \\
&\qquad + (\mathcal{E}^- * f)(\mathrm{i}r - \tfrac{\mathrm{i}}{2}\mathcal{W} - \tfrac{1}{2}) * \psi + f(\Omega_c + \tfrac{1}{2}\mathcal{W} \circ \mathcal{W} + \tfrac{\mathrm{i}}{2}\mathcal{W}) * \psi, \phi_0 \Big\rangle \,.
\end{aligned}$$

Notice that on the right-hand side, one term is equal to

$$\langle f(\Omega_c \cdot \psi), \phi_0 \rangle = \lambda \langle f\psi, \phi_0 \rangle \,,$$

which is the term we started with on the left-hand side. This is the first of two miracles: these are the only terms involving λ, and we may cancel them. Next we look at the terms that contain r (which is of size $|\lambda|^{1/2}$, and so are the main terms after the terms in λ have been canceled). The terms in r, and this is the second miracle, comprise

$$\left\langle (\mathcal{E}^+ * f)\mathrm{i}r\psi + (\mathcal{E}^- * f)\mathrm{i}r\psi, \phi_0 \right\rangle = 4\mathrm{i}r \left\langle (\mathcal{H} * f)\psi, \phi_0 \right\rangle .$$

For the remaining terms, we use (2.24) (once or twice as required) to arrive at

$$\begin{aligned}0 &= 4\mathrm{i}r\,\langle(\mathcal{H}{*}f)\psi,\phi_0\rangle + \big\langle(\mathcal{E}^-\circ\mathcal{E}^+ * f)\psi + (\mathcal{E}^+{*}f)(\tfrac{\mathrm{i}}{2}\mathcal{W}-\tfrac{1}{2})*\psi \\ &\qquad + (\mathcal{E}^-{*}f)(-\tfrac{\mathrm{i}}{2}\mathcal{W}-\tfrac{1}{2})*\psi + f(\tfrac{1}{2}\mathcal{W}\circ\mathcal{W}+\tfrac{\mathrm{i}}{2}\mathcal{W})*\psi,\phi_0\big\rangle \\ &= 4\mathrm{i}r\,\langle(\mathcal{H}{*}f)\psi,\phi_0\rangle + 4\mathrm{i}\,\langle(\mathcal{V}{*}f)\psi,\psi\rangle\end{aligned}$$

for some degree-two operator $\mathcal{V}$, which is independent of ϕ, λ, and r.

The second formula follows from the first by recalling that

$$\|\phi\|_2 = \|\sum_{n=-N}^{N}\phi_{2n}\|_2 = \sqrt{2N+1},$$

and the observation that $\|\mathcal{V}f\|_\infty = \mathrm{O}_f(1)$. □

Corollary 2.20. *Suppose that N is defined as a function of λ so as to ensure that $Nr^{-1}\to 0$ and $N^{-1}\to 0$ as $|\lambda|\to\infty$. Assume that (ϕ_i) is a sequence of Maass cusp forms with corresponding eigenvalues $|\lambda_i|\to\infty$ as $i\to\infty$, and that $|\phi_i|^2\,\mathrm{dvol}_M$ converges weak*. Then any weak*-limit of $|\widetilde{\phi_i}|^2\,\mathrm{d}m_X$ is invariant under the geodesic flow.*

We may define $N=\lceil r^{1/2}\rceil$, which is then of size $|\lambda|^{1/4}$. Then Corollary 2.20 applies, and gives [I] in Theorem 2.1. Moreover, Corollary 2.18 also applies, so that together Theorem 2.1 will follow. For the proof of Corollary 2.20 we need the following lemma.

Lemma 2.21. *Let $f\in C_c^\infty(\Gamma\backslash\mathrm{SL}_2(\mathbb{R}))$ and $w\in\mathfrak{sl}_2(\mathbb{C})$. Then the K-finite approximations*

$$f_{[-L,L]} = f *_K \sum_{\ell=-L}^{L} e_\ell$$

(where e_ℓ is the character defined in (2.11)) converge uniformly to f as $L\to\infty$. Moreover, $\mathcal{H}\cdot f_{[-L,L]}$ converges uniformly to $\mathcal{H}\cdot f$ as $L\to\infty$.

Proof. Just as in Section 2.4.2, we can consider the case of $f\in C_c^\infty(\mathrm{SL}_2(\mathbb{R}))$ only. Let

$$f_n(x) = f *_K e_n(x) = \int_K f(xk_\theta)e(k_\theta^{-1})\,\mathrm{d}m_K.$$

By the discussion in Section 2.3.1, $\|f_n\|\ll_f \frac{1}{n^2}$, and from this one can quickly show that $f_{[-L,L]}\to f$ uniformly as $L\to\infty$. For the second claim, we wish to estimate

$$\begin{aligned}[\mathcal{H}*f_n](x) &= \left[\frac{\partial}{\partial t}\int f_n(x\exp(t\mathcal{H})k_\theta)e_n(k_\theta^{-1})\,\mathrm{d}m_K\right]_{t=0} \\ &= \int\left[\frac{\partial}{\partial t}f_n(xk_\theta\exp(t\,\mathrm{Ad}_{k_\theta}^{-1}(\mathcal{H})))\right]_{t=0} e(k_\theta^{-1})\,\mathrm{d}m_K \\ &= \int\left(\mathrm{Ad}_{k_\theta}^{-1}(\mathcal{H})*f\right)(xk_\theta)e(k_\theta^{-1})\,\mathrm{d}m_K.\end{aligned}$$

Now notice that $\mathrm{Ad}_{k_\theta}^{-1}(\mathcal{H}) * f(x)$ is a smooth function of x and θ – in fact it can be written as a linear combination (with coefficients depending smoothly on θ), of $\mathcal{H} * f(x)$ and $\mathcal{U}^{\pm} * f(x)$. Therefore, we may use integration by parts with respect to θ twice to deduce that

$$\|\mathcal{H} * f_n\|_\infty = \mathrm{O}_f(1/n^2).$$

It follows that $\mathcal{H} * f_{[-L,L]}$ converges to some function $h \in C_c^\infty(\mathrm{SL}_2(\mathbb{R}))$. However, if $g \notin (\operatorname{Supp} f)K$ then

$$f_{[-L,L]}(ga_T) = \int_0^T w * f_{[-L,L]}(ga_t)\,\mathrm{d}t \to \int_0^T h(ga_t)\,\mathrm{d}t$$

as $L \to \infty$, which shows that

$$f(ga_T) = \int_0^T h(ga_t)\,\mathrm{d}t,$$

$\mathcal{H} * f = h$, and the lemma. □

Proof of Corollary 2.20. Assume that $|\widetilde{\phi}_i|\,\mathrm{d}m_X$ converges weak* to μ, and let $f \in C_c^\infty$ be a K-finite function. Then $\mathcal{H}*f$ and $\mathcal{V}*f$ are K-finite by Lemma 2.8. Given ϕ_i, let us write $\phi_{i,2n}$ for the functions in Definition 2.14, and let

$$\psi_i = \sum_{n=-N(\lambda_i)}^{N(\lambda_i)} \phi_i.$$

By Theorems 2.17 and (2.19) we have

$$\begin{aligned}\int_X \mathcal{H}*f|\widetilde{\phi}_i|^2\,\mathrm{d}m_X &= \langle \mathcal{H}*f\psi_i, \phi_0\rangle + \mathrm{O}_f(N^{-1} + Nr^{-1}) \\ &= \mathrm{O}_f(N^{-1} + Nr^{-1}),\end{aligned}$$

and so

$$\int_X \mathcal{H}*f\,\mathrm{d}\mu = 0.$$

If $f \in C_c^\infty$ (without the additional assumption that f is K-finite), then Lemma 2.21 implies that $\mathcal{H}*f$ is a uniform limit of a sequence $(\mathcal{H}*f_n)$ of K-finite functions in C_c^∞. It follows that

$$\int_X \mathcal{H}*f\,\mathrm{d}\mu = 0$$

for any $f \in C_c^\infty$. Now let $T \in \mathbb{R}$. Then we have

$$f(xa_T) - f(x) = \int_0^T (\mathcal{H} * f)(xa_t)\,\mathrm{d}t$$

(where a_t is defined as in (2.2)), and

$$\int_X f(xa_t)\,\mathrm{d}\mu - \int_X f(x)\,\mathrm{d}\mu = \int_0^T \int_X \mathcal{H} * (a_t f)(x)\,\mathrm{d}\mu(x)\,\mathrm{d}t = 0$$

since $(a_t \cdot f)(x) = f(xa_t)$ defines a function $a_t \cdot f \in C_c^\infty$. Thus

$$\int_X f(xa_T)\,\mathrm{d}\mu = \int f(x)\,\mathrm{d}\mu$$

for any $T \in \mathbb{R}$, so μ is invariant under the geodesic flow. $\square$

3

Hecke Operators and Recurrence

In this chapter we define the Hecke operators, define the recurrence assumption $[\mathsf{R}]_p$, and prove the recurrence assumption for any arithmetic quantum limit on $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$.

3.1 Hecke Operators

3.1.1 The Classical Definition of T_p

Definition 3.1. *Let $M = \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$, let p be a prime, and let f be a function on M. Then the action of the Hecke operator T_p on f is defined by*

$$\left(T_p(f)\right)(z) = \frac{1}{p+1}\left[f(pz) + \sum_{j=0}^{p-1} f\left(\frac{z+j}{p}\right)\right]. \tag{3.1}$$

Here we interpret f as a function on $\mathbb{H}$ satisfying the periodicity laws

$$f(\gamma \boldsymbol{\cdot} z) = f(z)$$

for all $\gamma \in \mathrm{SL}_2(\mathbb{Z})$. We will see later in Section 3.1.3 that $T_p(f)$ is again an $\mathrm{SL}_2(\mathbb{Z})$-periodic function, and therefore may be viewed as a well-defined function on M. The reader may check this now by considering the condition of periodicity with respect to the matrices $\begin{pmatrix} 1 & 1 \\ & 1 \end{pmatrix}$ and $\begin{pmatrix} & -1 \\ 1 & \end{pmatrix}$ that together generate $\mathrm{SL}_2(\mathbb{Z})$.

3.1.2 $\mathrm{SL}_2(Z)\backslash\mathrm{SL}_2(\mathbb{R})$ and $\mathrm{PGL}_2(\mathbb{Z})\backslash\mathrm{PGL}_2(\mathbb{R})$ are Isomorphic

Recall that

$$\begin{aligned}\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R}) &= \left\{\mathbb{Z}^2 g \mid g \in \mathrm{SL}_2(\mathbb{R})\right\} \\ &= \left\{\Lambda \subseteq \mathbb{R}^2 \mid \Lambda \text{ is a unimodular lattice in } \mathbb{R}^2\right\},\end{aligned}$$

where a *lattice* Λ in $\mathbb{R}^2$ is a $\mathbb{Z}$-module spanned by two linearly independent vectors, equivalently $\Lambda = \mathbb{Z}^2 g$ for some $g \in \mathrm{GL}_2(\mathbb{R})$, and a lattice $\Lambda = \mathbb{Z}^2 g$ is *unimodular* if its covolume $\mathrm{covol}(\Lambda) = |\det(\mathrm{g})|$ is equal to 1. Notice that swapping the rows of g has the effect of changing the sign of $\det(g)$, but does not change the lattice defined by g.

On the other hand,

$$\begin{aligned}\mathrm{PGL}_2(\mathbb{Z})\backslash \mathrm{PGL}_2(\mathbb{R}) &= \left\{\left[\mathbb{Z}^2 g\right] \mid g \in \mathrm{GL}_2(\mathbb{R})\right\} \\ &= \left\{[\Lambda] \mid \Lambda \text{ is a lattice in } \mathbb{R}^2\right\},\end{aligned}$$

where $[\Lambda]$ denotes the equivalence class of the lattice Λ under homothety. That is, $\Lambda' \in [\Lambda]$ if and only if $\Lambda' = a\Lambda$ for some $a \in \mathbb{R}_{>0}$.

The map sending a unimodular lattice Λ to its equivalence class $[\Lambda]$ defines a map from $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ to $\mathrm{PGL}_2(\mathbb{Z})\backslash \mathrm{PGL}_2(\mathbb{R})$. This map is injective, since the relation $\Lambda' = a\Lambda$ implies that $\mathrm{covol}(\Lambda') = \mathrm{a}^2\mathrm{covol}(\Lambda)$, and therefore $a = 1$ if both Λ' and Λ are unimodular (notice in this connection that $-I \in \mathrm{SL}_2(\mathbb{Z})$, a property that does not hold in $\mathrm{SL}_d(\mathbb{Z})$ for odd d). Moreover, the map $\Lambda \mapsto [\Lambda]$ is also surjective, since for any equivalence class $[\Lambda]$, there exists a unimodular representative $\Lambda' = \frac{1}{\mathrm{covol}(\Lambda)^{1/2}}\Lambda$.

3.1.3 The Second Definition of T_p

We now extend the definition of T_p given in Section 3.1.1 to the space of functions on $\mathrm{PGL}_2(\mathbb{Z})\backslash \mathrm{PGL}_2(\mathbb{R})$. For this, notice once again that any function f on $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ gives rise to a function (also denoted f) on $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R}) \cong \mathrm{PGL}_2(\mathbb{Z})\backslash \mathrm{PGL}_2(\mathbb{R})$.

Definition 3.2. *Let $X = \mathrm{PGL}_2(\mathbb{Z})\backslash \mathrm{PGL}_2(\mathbb{R})$, let p be a prime, and let f be a function on X. Then the Hecke operator T_p is defined by sending f to the normalized sum*

$$(T_p(f))\,([\Lambda]) = \frac{1}{p+1} \sum_{\substack{\Lambda' \subseteq \Lambda,\\ [\Lambda:\Lambda']=p}} f\left([\Lambda']\right) \tag{3.2}$$

of the values of f on all the sublattices Λ' of Λ with index p.

For $g \in \mathrm{PGL}_2(\mathbb{R})$ and a function f on X, we define $g\boldsymbol{\cdot}f$ to be the function $(g\boldsymbol{\cdot}f)\,([\Lambda]) = f\,([\Lambda]g)$.

Proposition 3.3. *If $\Lambda \subseteq \mathbb{R}^2$ is a lattice, then there are $p+1$ subgroups $\Lambda' \subseteq \Lambda$ with index p. The expression in* (3.2) *defines a function on the homothety classes $[\Lambda] \in \mathrm{PGL}_2(\mathbb{Z})\backslash \mathrm{PGL}_2(\mathbb{R})$ of lattices $\Lambda \subseteq \mathbb{R}^2$. If f is a function on $M = \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H} \cong X/\mathrm{SO}(2)$ then so is $T_p(f)$, and $T_p(f)$ agrees with the definition* (3.1)*. Finally, for any $g \in \mathrm{PGL}_2(\mathbb{R})$, we have $g\boldsymbol{\cdot}T_p(f) = T_p(g\boldsymbol{\cdot}f)$.*

Proof. If $\Lambda' \subseteq \Lambda$ has index p, then $\Lambda/\Lambda' \cong \mathbb{Z}/p\mathbb{Z}$, and so $p\Lambda \subseteq \Lambda'$. Thus index p subgroups $\lambda' \subseteq \Lambda$ are in one-to-one correspondence with lines (equivalently, index p subgroups) of the two-dimensional vector space $\Lambda/p\Lambda \cong (\mathbb{Z}/p\mathbb{Z})^2$ over the field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. After choosing a basis of Λ, there is a one-to-one correspondence between index p subgroups of Λ and points in the projective line $\mathbb{P}^1(\mathbb{F}_p)$ over the field $\mathbb{F}_p$. Note that $|\mathbb{P}^1(\mathbb{F}_p)| = \frac{p^2-1}{p-1} = p+1$.

It is clear that if we replace the lattice Λ by $a\Lambda$ for some $a \in \mathbb{R}_{>0}$, then we replace each index p subgroup $\Lambda' \subseteq \Lambda$ with the index p subgroup $a\Lambda' \subseteq a\Lambda$. This shows that $T_p(f)$ is indeed a function on X.

Now if $g \in \mathrm{GL}_2(\mathbb{R})$, then the index p subgroups of Λg are all of the form $\Lambda' g$ as Λ' varies over all index p subgroups of Λ. This shows that

$$[g \centerdot T_p(f)]\,([\Lambda]) = T_p(f)\,([\Lambda]g) = \frac{1}{p+1} \sum_{\Lambda' \subseteq \Lambda} f\,([\Lambda']g) = T_p(g \centerdot f)[\Lambda].$$

Now let f_M be a function on $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$. Then the associated function on X is given by

$$f_X([\Lambda]) = f_M(g \centerdot \mathrm{i})$$

where we assume that $\Lambda = \mathbb{Z}^2 g$ with $\det(g) > 0$. This is well-defined, since replacing Λ by $a\Lambda$ with $a \in \mathbb{R}_{>0}$ replaces g by ga, with $(ga)\centerdot\mathrm{i} = g\left((aI)\centerdot\mathrm{i}\right) = g\centerdot\left(\frac{a\mathrm{i}}{a}\right) = g\centerdot\mathrm{i}$. Moreover, choosing a different basis of Λ (with positive determinant) corresponds to replacing g with γg for some $\gamma \in \mathrm{SL}_2(\mathbb{Z})$. Neither of these changes affects the value of $f_M(g\centerdot\mathrm{i})$. Thus, $T_p(f_X)$ gives a function with

$$k \circ T_p(f_X) = T_p(k \circ f_X) = T_p(f_X)$$

for all $k \in \mathrm{SO}(2)$, which shows that $T_p(f_X)$ induces a function on

$$X/\mathrm{SO}(2) \cong \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H} = M.$$

We now calculate the function $T_p(f_X)$ as a function on M, and will arrive at the expression in Definition 3.1. Let $z \in \mathbb{H}$, so that

$$z = x + \mathrm{i}y = \begin{pmatrix} 1 & x \\ & 1 \end{pmatrix} \begin{pmatrix} y & \\ & 1 \end{pmatrix} \centerdot \mathrm{i}$$

corresponds to the lattice

$$\Lambda = \mathbb{Z}^2 \begin{pmatrix} 1 & x \\ & 1 \end{pmatrix} \begin{pmatrix} y & \\ & 1 \end{pmatrix},$$

which is spanned by the vectors

$$(1,0) \begin{pmatrix} 1 & x \\ & 1 \end{pmatrix} \begin{pmatrix} y & \\ & 1 \end{pmatrix} = (y, x)$$

and

$$(0,1)\begin{pmatrix}1 & x\\ & 1\end{pmatrix}\begin{pmatrix}y & \\ & 1\end{pmatrix}=(0,1).$$

It follows that the index p subgroups of Λ are given by

$$\langle p(y,x),(0,1)\rangle$$

and

$$\langle (y,x)+j(0,1),(0,p)\rangle$$

for $j=0,\dots,p-1$. The first is the lattice corresponding to pz. Each of the lattices of the second type is homothetic to

$$\langle \tfrac{1}{p}(y,x+j),(0,1)\rangle,$$

which corresponds to $\frac{z+j}{p}$ for $j=0,\dots,p-1$. Therefore, $T_p(f_X)$ as defined by Definition 3.2, when viewed as a function on M, gives $T_p(f_M)$ as defined by Definition 3.1 if f_X is the function on X defined by f_M on M. □

Corollary 3.4. *For any p the p-Hecke operator T_p commutes with any differential operator $m\in\mathfrak{sl}_2(\mathbb{C})$ (or even any element of the enveloping algebra of $\mathfrak{sl}_2(\mathbb{C})$).*

Proof. Since $T_p(g\bullet f)=g\bullet T_p(f)$ by Proposition 3.3, we have

$$m*(T_p(f))(x)=\left[\frac{\partial}{\partial t}T_p(f)(x\exp(tm))\right]_{t=0}=T_p(m*f)(x)$$

for any $f\in C^\infty(X)$. □

3.1.4 The p-adic Extension $X_{\infty,p}$ of $X_\infty=\mathrm{PGL}_2(\mathbb{Z})\backslash\mathrm{PGL}_2(\mathbb{R})$

Recall that $\mathbb{Q}_p$ is the completion of $\mathbb{Q}$ with respect to the non-Archimedean norm defined by $|0|_p=0$ and

$$\left|\frac{m}{n}p^k\right|_p=p^{-k}$$

for $m,n\in\mathbb{Z}\smallsetminus\{0\}$ with $p\nmid mn$ and $k\in\mathbb{Z}$. Then $\mathbb{Q}_p$ is a locally compact non-discrete field, and the closure $\mathbb{Z}_p$ of $\mathbb{Z}$ in $\mathbb{Q}_p$ is a maximal compact open subring of $\mathbb{Q}_p$, called the ring of p-adic integers. Finally, recall that $\mathbb{Z}[\frac{1}{p}]$ is a dense subset of $\mathbb{Q}_p$, but that under the diagonal embedding

$$\{(a,a)\mid a\in\mathbb{Z}[\tfrac{1}{p}]\}\subseteq\mathbb{R}\times\mathbb{Q}_p$$

is a discrete and co-compact subgroup (for background on properties of the p-adic numbers, see Weil [31]).

For a ring R with a unit 1, and group of units $R^\times$, we define

$$\mathrm{PGL}_2(R) = \left\{ \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mid a,b,c,d \in R, \det\gamma \in R^\times \right\} / \sim$$

where $\gamma \sim \gamma'$ if there is a scalar $r \in R^\times$ with $\gamma = r\gamma'$. If R is a topological ring, then there is an inherited topology on $\mathrm{GL}_2(R)$ and thence on $\mathrm{PGL}_2(R)$.

Proposition 3.5. *Embed* $\mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])$ *as a subset of* $\mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_p)$ *diagonally, by sending* γ *to* (γ,γ)*. Then* $\mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])$ *is a lattice in* $\mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_p)$*, and the double quotient*

$$\mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}]) \backslash \mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_p) / \mathrm{PGL}_2(\mathbb{Z}_p) \tag{3.3}$$

is naturally isomorphic to

$$\mathrm{PGL}_2(\mathbb{Z}) \backslash \mathrm{PGL}_2(\mathbb{R}). \tag{3.4}$$

Proof. We note that $\mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Z}_p)$ is an open subgroup of $\mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_p)$. If $(\gamma,\gamma) \in \mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Z}_p)$ for some $\gamma \in \mathrm{PGL}(\mathbb{Z}[\frac{1}{p}])$ then, after modifying γ by a scalar if necessary, we may assume that $\gamma \in \mathrm{GL}_2(\mathbb{Z})$. Therefore, $\gamma = I$ or (γ,γ) is not close to (I,I). That is, the diagonally embedded subgroup $\mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])$ is a discrete subgroup. Moreover, this also shows that the orbit of the identity coset $\mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])$ under the action of $\mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Z}_p)$ is isomorphic to $\mathrm{PGL}_2(\mathbb{Z}) \backslash \mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Z}_p)$ (where $\mathrm{PGL}_2(\mathbb{Z})$ is again embedded diagonally). We claim that this orbit $\mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])\mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Z}_p)$ of the open subgroup is actually all of the space in (3.3). This then implies that $\mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])$ is a lattice, and gives the isomorphism between (3.3) and (3.4).

The claim on the other hand is equivalent to

$$\mathrm{PGL}_2(\mathbb{Q}_p) = \mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])\,\mathrm{PGL}_2(\mathbb{Z}_p),$$

which may be seen as follows. Given an element

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PGL}_2(\mathbb{Q}_p),$$

we claim that we may multiply on the left by elements of $\mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])$ and on the right by elements of $\mathrm{PGL}_2(\mathbb{Z}_p)$ to obtain the identity. Notice that left multiplication corresponds to row operations and right multiplication to column operations. Applying a column operation, we can ensure that $|a|_p \geqslant |b|_p$ and so $\frac{b}{a} \in \mathbb{Z}_p$. Thus we can multiply on the right by $\begin{pmatrix} 1 & -\frac{b}{a} \\ & 1 \end{pmatrix}$ to obtain $\begin{pmatrix} a & 0 \\ c & d' \end{pmatrix}$. Multiplying on the left by $\begin{pmatrix} 1 & \\ \alpha & 1 \end{pmatrix}$ with some $\alpha \in \mathbb{Z}[\frac{1}{p}]$ very close to $-\frac{c}{a}$ we obtain $\begin{pmatrix} a & \\ a\alpha + c & d' \end{pmatrix}$, where we may assume that $|a\alpha + c|_p \leqslant |d'|_p$. This allows us to multiply (since $d' \neq 0$) on the right by $\begin{pmatrix} 1 & \\ -\frac{a\alpha+c}{d'} & 1 \end{pmatrix}$ to finally obtain $\begin{pmatrix} a & \\ & d' \end{pmatrix}$. Using diagonal matrices in $\mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])$ and $\mathrm{PGL}_2(\mathbb{Z}_p)$, the claim and the proposition follows. $\square$

We note that the orbit of a point in

$$\mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])\backslash\mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_p) \tag{3.5}$$

under $\mathrm{PGL}_2(\mathbb{Q}_p)$ is mapped under the isomorphism between (3.3) and (3.4) in Proposition 3.5 to a set isomorphic to

$$T = \mathrm{PGL}_2(\mathbb{Q}_p)/\,\mathrm{PGL}_2(\mathbb{Z}_p). \tag{3.6}$$

This holds since $\mathrm{PGL}(\mathbb{Z}[\frac{1}{p}])$ is embedded injectively into $\mathrm{PGL}_2(\mathbb{R})$, which implies that the stabilizer of any point in (3.5) under the action of $\mathrm{PGL}_2(\mathbb{Q}_p)$ is trivial.

The space T in (3.6) may be thought of as the p-adic analog of the hyperbolic plane – $\mathrm{PGL}_2(\mathbb{Z}_p)$ is a maximal compact subgroup of $\mathrm{PGL}_2(\mathbb{Q}_p)$, just as $\mathrm{SO}(2)$ is a maximal compact subgroup of $\mathrm{SL}_2(\mathbb{R})$. We may also think of T_p as being the p-adic Laplace operator. In fact T has the structure of a $(p+1)$-regular tree, and T_p is the Laplace operator on this tree in the sense of graph theory (see the discussion in Section 3.2.1).

3.1.5 Defining T_p as a Convolution Operator

Notice that $K_p = \mathrm{PGL}_2(\mathbb{Z}_p)$ is a compact and open subgroup of $\mathrm{PGL}_2(\mathbb{Q}_p)$, which shows that

$$K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p$$

must be a finite disjoint union of cosets $K_p g_p$ for some elements $g_p \in \mathrm{PGL}_2(\mathbb{Q}_p)$. The next lemma makes this explicit.

Lemma 3.6.

$$K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p = K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} \sqcup \bigsqcup_{j=0}^{p-1} K_p \begin{pmatrix} 1 & j \\ & p \end{pmatrix} = K_p \begin{pmatrix} 1 & \\ & p \end{pmatrix} K_p.$$

Proof. Note that $K_p g_1 = K_p g_2$ if and only if $g_1 g_2^{-1} \in K_p$. For $g_1 = \begin{pmatrix} 1 & j_1 \\ & p \end{pmatrix}$ and $g_2 = \begin{pmatrix} 1 & j_2 \\ & p \end{pmatrix}$ with $j_1 \neq j_2$, we have

$$g_1 g_2^{-1} = \begin{pmatrix} 1 & j_1 \\ & p \end{pmatrix} \begin{pmatrix} 1 & -\frac{j_2}{p} \\ & \frac{1}{p} \end{pmatrix} = \begin{pmatrix} 1 & (j_1 - j_2)p^{-1} \\ & 1 \end{pmatrix},$$

so that $K_p \begin{pmatrix} 1 & j_1 \\ & p \end{pmatrix} \neq K_p \begin{pmatrix} 1 & j_2 \\ & p \end{pmatrix}$ are disjoint. The disjointness of these to $K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix}$ is proved similarly. Also notice that $\begin{pmatrix} p & \\ & 1 \end{pmatrix}, \begin{pmatrix} 1 & j \\ & p \end{pmatrix} \in K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p$ for $j = 0, 1, \ldots, p-1$.

Finally let $g \in K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p$, which we may write as

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{Mat}_{22}(\mathbb{Z}_p)$$

and $|\det(g)|_p = \frac{1}{p}$. If $a \in \mathbb{Z}_p^\times$ (the case of $c \in \mathbb{Z}_p^\times$ can be reduced to this one by multiplying with $\begin{pmatrix} & 1 \\ 1 & \end{pmatrix} \in K_p$ on the left), then we may multiply with $\begin{pmatrix} \frac{1}{a} & \\ -\frac{c}{a} & 1 \end{pmatrix}$ (respectively, $\begin{pmatrix} 1 & \\ & \frac{1}{d} \end{pmatrix}$) on the left to obtain

$$K_p g = K_p \begin{pmatrix} 1 & b' \\ & d' \end{pmatrix} = K_p \begin{pmatrix} 1 & b'' \\ & p \end{pmatrix},$$

where $|d'|_p = p$, equivalently $d' = d_1 p$ with $d_1 \in \mathbb{Z}_p^\times$. Now there exists some j in $\{0, 1 \ldots, p-1\}$ with $b'' \equiv j \pmod{p}$, so that $\begin{pmatrix} 1 & \frac{-b''+j}{p} \\ & 1 \end{pmatrix} \in K_p$ which gives

$$K_p g = K_p \begin{pmatrix} 1 & j \\ & p \end{pmatrix}.$$

If, on the other hand $a, c \in p\mathbb{Z}_p$, then one of them, say a, must belong to $p\mathbb{Z}_p^\times$ since $\det(g) \in p\mathbb{Z}_p^\times$ by assumption. Multiplying on the left by $\begin{pmatrix} (\frac{a}{p})^{-1} & \\ -\frac{c}{a} & 1 \end{pmatrix}$ (respectively, by $\begin{pmatrix} 1 & -\frac{b_1}{d_1} \\ & d_1^{-1} \end{pmatrix}$) gives

$$K_p g = K_p \begin{pmatrix} p & b_1 \\ & d_1 \end{pmatrix} = K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix},$$

since we must have $d_1 \in \mathbb{Z}_p^\times$.

Finally, notice that K_p contains $\begin{pmatrix} & 1 \\ 1 & \end{pmatrix}$, which shows that

$$K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p = K_p \begin{pmatrix} 1 & \\ & p \end{pmatrix} K_p.$$

□

We normalize the Haar measure $m_{\mathrm{PGL}_2(\mathbb{Q}_p)}$ so that $m_{\mathrm{PGL}_2(\mathbb{Q}_p)}(K_p) = 1$, which then by Lemma 3.6 implies that

$$m_{\mathrm{PGL}_2(\mathbb{Q}_p)}\left(K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p\right) = p + 1.$$

Recall that if $\chi \in L^1(H)$ and the group H acts on X, then for functions $f \in L^2(X)$, we may define the convolution

$$f * \chi(x) = \int_H f(xh^{-1})\chi(h)\,\mathrm{d}m_H(h).$$

Proposition 3.7. *For a function f on $\mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])\backslash \mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_p)$, we may obtain the p-Hecke operator by the convolution*

$$T_p(f) = \frac{1}{p+1} f * \mathbf{1}_{K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p}. \tag{3.7}$$

For a function f on $\mathrm{PGL}_2(\mathbb{Z})\backslash \mathrm{PGL}_2(\mathbb{R})$, this agrees with T_p as defined in Definition 3.2.

Proof. Notice that for $k \in K_p$ we have

$$\begin{aligned} f * \mathbf{1}_{K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p}(xk) &= \int_{\mathrm{PGL}_2(\mathbb{Q}_p)} f(xkh^{-1})\mathbf{1}_{K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p}(h)\,\mathrm{d}m(h) \\ &= f * \mathbf{1}_{K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p}(x), \end{aligned}$$

which shows that $f * \mathbf{1}_{K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p}$ is invariant under K_p and so can be considered as a function on $\mathrm{PGL}_2(\mathbb{Z})\backslash \mathrm{PGL}_2(\mathbb{R})$ by Proposition 3.5.

Suppose now that $f \in L^2(\mathrm{PGL}_2(\mathbb{Z})\backslash \mathrm{PGL}_2(\mathbb{R}))$. We may consider f as a function on $\mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])\backslash \mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_p)$ with the property that

$$f(xk) = f(x)$$

for all $k \in K_p$. Then

$$\frac{1}{p+1} f * \mathbf{1}_{K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p}(x) = \frac{1}{p+1}\left[f(x(I, \begin{pmatrix} p & \\ & 1 \end{pmatrix}^{-1})) + \sum_{j=0}^{p-1} f(x(I, \begin{pmatrix} 1 & j \\ & p \end{pmatrix}^{-1})) \right],$$

which we now evaluate further. We suppose that $x = \mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])(g, I)$ for $g \in \mathrm{PGL}_2(\mathbb{R})$, which is the general case by the first paragraph of the proof, and corresponds to the homothety class $[\mathbb{Z}^2 g]$. Then

$$\begin{aligned} x\left(I, \begin{pmatrix} 1 & j \\ & p \end{pmatrix}\right) &= \mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])\left(g, \begin{pmatrix} 1 & j \\ & p \end{pmatrix}^{-1}\right) \\ &= \mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])\left(\begin{pmatrix} 1 & j \\ & p \end{pmatrix} g, I\right) \end{aligned}$$

corresponds to the homothety class of the index p subgroup $\mathbb{Z}^2 \begin{pmatrix} 1 & j \\ & p \end{pmatrix} g \subseteq \mathbb{Z}^2 g$ for $j = 0, \ldots, p-1$. The case of $\begin{pmatrix} p & \\ & 1 \end{pmatrix}$ is similar, so that indeed $T_p(f)$, as defined in Definition 3.2, is equal to the right-hand side of (3.7). $\square$

Corollary 3.8. *T_p is a self-adjoint operator on $L^2(\mathrm{PGL}_2(\mathbb{Z})\backslash \mathrm{PGL}_2(\mathbb{R}))$ (and on $L^2(\mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])\backslash \mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_p))$).*

Proof. Recall that the adjoint of $(h\bullet)$ for $h \in \mathrm{PGL}_2(\mathbb{Q}_p)$ is $(h^{-1}\bullet)$, since the action of $\mathrm{PGL}_2(\mathbb{Q}_p)$ is unitary. This implies that

$$T_p^*(f) = \frac{1}{p+1} f * \mathbf{1}_{\left(K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p\right)^{-1}}.$$

However, as $\begin{pmatrix} p & \\ & 1 \end{pmatrix}\begin{pmatrix} 1 & \\ & p \end{pmatrix} = \begin{pmatrix} p & \\ & p \end{pmatrix} = I$ in $\mathrm{PGL}_2(\mathbb{Q}_p)$ we have

$$\begin{aligned} \left(K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p\right)^{-1} &= K_p \begin{pmatrix} p^{-1} & \\ & 1 \end{pmatrix} K_p \\ &= K_p \begin{pmatrix} 1 & \\ & p \end{pmatrix} K_p \\ &= K_p \begin{pmatrix} p & \\ & 1 \end{pmatrix} K_p \end{aligned}$$

by Lemma 3.6. $\square$

Corollary 3.9. *For any two primes $p_1 \neq p_2$ the two associated Hecke operators T_{p_1} and T_{p_2} on $\mathrm{PGL}_2(\mathbb{Z})\backslash \mathrm{PGL}_2(\mathbb{R})$ and on*

$$\mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p_1 p_2}])\backslash \mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_{p_1}) \times \mathrm{PGL}_2(\mathbb{Q}_{p_2})$$

commute.

Proof. It may be shown (as above) that

$$\mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p_1 p_2}])\backslash \mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_{p_1}) \times \mathrm{PGL}_2(\mathbb{Q}_{p_2})$$

is a common extension of

$$\mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p_1}])\backslash \mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_{p_1})$$

and

$$\mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p_2}])\backslash \mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_{p_2}),$$

and that T_{p_1} and T_{p_2} can again be defined by the convolution operators (3.7) associated to the group actions of $\mathrm{PGL}_2(\mathbb{Q}_{p_1})$, respectively $\mathrm{PGL}_2(\mathbb{Q}_{p_2})$. However, as these actions commute we obtain the corollary. □

3.2 Trees, Recurrence Relations, and Recurrence

In this section we will prove Hecke recurrence, which we now define.

Definition 3.10. *Let H be a σ-compact locally compact group, acting continuously on a σ-compact locally compact space Y. Let ν be a Borel probability measure on Y. Then the H-action is called* recurrent with respect to ν *if, for any Borel set $B \subseteq Y$ with $\nu(B) > 0$, there is a sequence (h_n) in H with* $h_n \to \infty$ such that $h_n \bullet y \in B$ for all $n \geqslant 1$.*

Notice that invariance of ν under the H-action implies recurrence automatically, but that recurrence is a much weaker requirement than invariance.

Since the p-adic group $\mathrm{PGL}_2(\mathbb{Q}_p)$ does not act on $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$, we have to modify the definition slightly to adapt to our situation.

Definition 3.11. *Let p be a prime, and let μ be a finite measure on the space*

$$X = \mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R}) \cong \mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])\backslash \mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_p)/\,\mathrm{PGL}_2(\mathbb{Z}_p).$$

Then μ is Hecke-p-recurrent if for any Borel set $B \subseteq X$ and μ-almost every $x = \Gamma(g_\infty, I)K_p \in B$ there exists a sequence (h_n) with $h_n \in \mathrm{PGL}_2(\mathbb{Q}_p)$, going to infinity, for which $\Gamma(g_\infty, h_n)K_p \in B$ for all $n \geqslant 1$.

In this chapter, we will prove the following theorem.

Theorem 3.12. *Let $X = \mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$, let p be a prime, and let (ϕ_i) be a sequence in $L^2_{m_X}$ of eigenfunctions of T_p with $\|\phi_i\|_2 = 1$. Then any weak*-limit μ of $|\phi_i|^2\,\mathrm{d}m_X$ is Hecke-p-recurrent.*

* This means that for any compact set $A \subseteq H$ there is some $N = N(A)$ such that $n \geqslant N \implies h_n \notin A$.

3.2.1 The Tree $\mathbf{PGL_2}(\mathbb{Q}_{\boldsymbol{p}})/\mathbf{PGL_2}(\mathbb{Z}_{\boldsymbol{p}})$ and its Laplacian

In this section we show that $\mathrm{PGL}_2(\mathbb{Q}_p)/\mathrm{PGL}_2(\mathbb{Z}_p)$ is a $(p+1)$-regular tree, and define the graph-theoretic Laplacian as the averaging operator over the nearest neighbors.

Lemma 3.13. *For any prime p, the quotient $\mathrm{PGL}_2(\mathbb{Q}_p)/\mathrm{PGL}_2(\mathbb{Z}_p)$ is naturally identified with the space of homothety equivalence classes $[\Lambda]$ of two-dimensional free $\mathbb{Z}_p$-submodules of the form $\Lambda = g\mathbb{Z}_p^2$ of the column space $\mathbb{Q}_p^2$ for $g \in \mathrm{GL}_2(\mathbb{Q}_p)$.*

Proof. Note that $\mathrm{GL}_2(\mathbb{Q}_p)$ acts transitively on the set of two-dimensional $\mathbb{Z}_p$-submodules, so that $\mathrm{PGL}_2(\mathbb{Q}_p)$ acts transitively on homothety equivalence classes. Moreover, $\mathrm{PGL}_2(\mathbb{Z}_p)$ is the stabilizer subgroup of $[\mathbb{Z}_p^2]$, proving the lemma. □

We now let $\mathcal{N} \subseteq \mathrm{Mat}_{22}(\mathbb{Z}_p)$ be a set of $(p+1)$ matrices with determinants in $p\mathbb{Z}_p^\times$ such that any one-dimensional subspace over $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ of $\mathbb{F}_p^2 \cong \mathbb{Z}_p^2/p\mathbb{Z}_p^2$ is of the form $\mathrm{Im}\,(h \pmod{p\mathbb{Z}_p}) = h\mathbb{Z}_p^2/p\mathbb{Z}_p^2$ for some (and hence for precisely one) $h \in \mathcal{N}$. For example, we could take

$$\mathcal{N} = \left\{ \begin{pmatrix} 1 & \\ & p \end{pmatrix}, \begin{pmatrix} p & j \\ & 1 \end{pmatrix} \mid 0 \leqslant j \leqslant p-1 \right\}.$$

Lemma 3.14. *If $g \in \mathrm{Mat}_{22}(\mathbb{Z}_p) \smallsetminus p\,\mathrm{Mat}_{22}(\mathbb{Z}_p)$, then there is a unique sequence $h_1, \dots, h_\ell \in \mathcal{N}$ with $g \in h_1 \cdots h_\ell \,\mathrm{GL}_2(\mathbb{Z}_p)$. Moreover, ℓ has the property that $\det(g) \in p^\ell \mathbb{Z}_p^\times$, and for any $k \in \{1, \dots, \ell-1\}$ we have*

$$\ker(h_k(\mathrm{mod p\mathbb{Z}_p})) \neq \mathrm{Im}(h_{k+1}(\mathrm{mod p\mathbb{Z}_p})).$$

Proof. As $g \notin p\,\mathrm{Mat}_{22}(\mathbb{Z}_p)$, we have either $\det(g) \in \mathbb{Z}_p^\times$ and $g \in \mathrm{GL}_2(\mathbb{Z}_p)$ or

$$(g\mathbb{Z}_p^2 + p\mathbb{Z}_p^2)/p\mathbb{Z}_p^2 \subseteq \mathbb{Z}_p^2/p\mathbb{Z}_p^2 \cong \mathbb{F}_p^2$$

is a line in the two-dimensional plane over $\mathbb{F}_p$. Hence

$$g\mathbb{Z}_p^2 + p\mathbb{Z}_p^2 = h_1\mathbb{Z}_p^2$$

for precisely one matrix $h_1 \in \mathcal{N}$. Equivalently,

$$g = h_1 + pg'$$

for some $g' \in \mathrm{Mat}_{22}(\mathbb{Z}_p)$. Applying h_1^{-1} we get

$$g_1 = h_1^{-1}g = I + ph_1^{-1}g' \in \mathrm{Mat}_{22}(\mathbb{Z}_p)$$

since $\det(h_1) \in p\mathbb{Z}_p^\times$ by assumption, and so $ph_1^{-1} \in \mathrm{Mat}_{22}(\mathbb{Z}_p)$. Now either $g_1 \in \mathrm{GL}_2(\mathbb{Z}_p)$, or we may apply the argument above again. Iterating, we find $g \in h_1 \cdots h_\ell \,\mathrm{GL}_2(\mathbb{Z}_p)$ as required.

To see that the sequence constructed is unique, assume that we also have $g \in h'_1 \cdots h'_\ell \operatorname{GL}_2(\mathbb{Z}_p)$ with $h'_1, \dots, h'_\ell \in \mathcal{N}$. This implies that

$$\det(g) \in \left(p^{\ell'}\mathbb{Z}_p^\times\right) \cap \left(p^\ell \mathbb{Z}_p^\times\right),$$

and so $\ell' = \ell$. Moreover,

$$g\mathbb{Z}_p^2 + p\mathbb{Z}_p^2 \subseteq h'_1\mathbb{Z}_p^2 + p\mathbb{Z}_p^2$$

must equal $h_1\mathbb{Z}_p^2 + p\mathbb{Z}_p^2$, and so $h'_1 = h_1$ by our assumption on $\mathcal{N}$. Now apply h_1^{-1}, and repeat the argument.

For the final claim, notice that if the kernel of the matrix h_k modulo $\mathbb{Z}_p$, considered as a representation of a linear map from the column space $\mathbb{F}_p^2$ to itself, coincides with the image of the matrix h_{k+1} modulo $p\mathbb{Z}_p^2$, then $h_k h_{k+1} \in p\operatorname{Mat}_{22}(\mathbb{Z}_p)$. Hence $g \in p\operatorname{Mat}_{22}(\mathbb{Z}_p)$, contradicting the assumption on g. □

Proposition 3.15. *For any prime p the quotient $\operatorname{PGL}_2(\mathbb{Q}_p)/\operatorname{PGL}_2(\mathbb{Z}_p)$ is a $(p+1)$-regular tree, where two equivalence classes $[\Lambda_1], [\Lambda_2]$ of two two-dimensional $\mathbb{Z}_p$-submodules $\Lambda_1, \Lambda_2 \subseteq \mathbb{Q}_p^2$ are adjacent if the representatives Λ_1, Λ_2 can be chosen so that $\Lambda_1 \subseteq \Lambda_2$ and $[\Lambda_2 : \Lambda_1] = p$. Equivalently, the neighbors of $\operatorname{PGL}_2(\mathbb{Z}_p)$ are $h\operatorname{PGL}_2(\mathbb{Z}_p)$ for $h \in \mathcal{N}$, and the neighbors of $g\operatorname{PGL}_2(\mathbb{Z}_p)$ are $gh\operatorname{PGL}_2(\mathbb{Z}_p)$ for $h \in \mathcal{N}$. If*

$$g \in h_1 \cdots h_\ell \operatorname{GL}_2(\mathbb{Z}_p) \smallsetminus p\operatorname{Mat}_{22}(\mathbb{Z}_p),$$

then the neighbors are also given by

$$h_1, \dots, h_{\ell-1}\operatorname{PGL}_2(\mathbb{Z}_p)$$

and

$$h_1, \dots, h_\ell h_{\ell+1}\operatorname{PGL}_2(\mathbb{Z}_p)$$

with $h_{\ell+1} \in \mathcal{N}$ and $\operatorname{Im}(h_{\ell+1} \pmod{p\mathbb{Z}_p}) \neq \ker(h_\ell \pmod{p\mathbb{Z}_p})$.

Notice that Proposition 3.15 and Lemma 3.14 together imply that

$$\begin{aligned}\{w \sim_\ell \operatorname{PGL}_2(\mathbb{Z}_p)\} &= \{h_1 \cdots h_\ell \operatorname{PGL}_2(\mathbb{Z}_p) \mid h_1, \dots, h_\ell \in \mathcal{N}, p \nmid h_1 \cdots h_\ell\} \\ &= \{g_p \operatorname{PGL}_2(\mathbb{Z}_p) \mid g_p \in \operatorname{Mat}_{22}(\mathbb{Z}_p), p \nmid g_p, \det g_p \in p^\ell \mathbb{Z}_p^\times\}.\end{aligned}$$

Proof of Proposition 3.15. We have to show that the three definitions of adjacency agree, and that they define a $(p+1)$-regular tree. For this, notice first that the first definition in the proposition is indeed symmetric: If $\Lambda_1 \subseteq \Lambda_2$ and $[\Lambda_2 : \Lambda_1] = p$ then $p\Lambda_2 \subseteq \Lambda_1$, and $[\Lambda_1 : p\Lambda_2] = p$ also. Next suppose that $\Lambda = \mathbb{Z}_p^2$, so that by choice of $\mathcal{N}$ we have $[h\mathbb{Z}_p^2]$ for $h \in \mathcal{N}$ are all neighbors of $[\mathbb{Z}_p^2]$. Applying a linear map $g \in \operatorname{GL}_2(\mathbb{Z}_p)$ to $\mathbb{Z}_p^2$ and its index p subgroups $\{h\mathbb{Z}_p^2 \mid h \in \mathcal{N}\}$, we see that the neighbors of $[g\mathbb{Z}_p^2]$ are $[gh\mathbb{Z}_p^2]$ for $h \in \mathcal{N}$. This shows the equivalence of the first and second definitions.

Suppose now that $g = h_1 \cdots h_\ell$, and let $h' \in \mathcal{N}$ be the matrix for which $\operatorname{Im}(h' \pmod{p\mathbb{Z}_p}) = \ker(h_\ell \pmod{p\mathbb{Z}_p})$. Then $h_\ell h' \in p\operatorname{Mat}_{22}(\mathbb{Z}_p)$ and $\det(h_\ell h') \in p^2\mathbb{Z}_p^\times$ so that $\frac{1}{p}h_\ell h' \in \mathrm{GL}_2(\mathbb{Z}_p)$. Therefore,

$$h_1 \cdots h_{\ell-1} h_\ell h' \operatorname{PGL}_2(\mathbb{Z}_p) = h_1 \cdots h_{\ell-1} \operatorname{PGL}_2(\mathbb{Z}_p)$$

which shows that the first two and the last definition of adjacency agree. Also notice that for $h \in \mathcal{N} \smallsetminus \{h'\}$, we still have

$$h_1 \cdots h_\ell h \in \operatorname{Mat}_{22}(\mathbb{Z}_p) \smallsetminus p\operatorname{Mat}_{22}(\mathbb{Z}_p).$$

It remains to show that $\operatorname{PGL}_2(\mathbb{Q}_p)/\operatorname{PGL}(\mathbb{Z}_p)$ is a $(p+1)$-regular tree. First, it is clear that every vertex $g\operatorname{PGL}_2(\mathbb{Q}_p)$ has $(p+1)$ neighboring vertices. Next notice that the neighbors of the neighbors of the base vertex $[\mathbb{Z}_p^2]$ give, by the definition and Lemma 3.14, apart from the base vertex, $p(p+1)$ new distinct vertices, which are the element of the sphere of radius two in the graph. Similarly, the set of neighbors of the elements of the sphere of radius $\ell \geqslant 1$ consists of the elements of the sphere of radius $\ell - 1$ and $p^\ell(p+1)$ new distinct elements of the sphere of radius $\ell+1$. Finally, the graph $\operatorname{PGL}_2(\mathbb{Q}_p)/\operatorname{PGL}_2(\mathbb{Z}_p)$ is also connected by Lemma 3.14. $\square$

To simplify the notation we will write $v, w, \ldots$ for the vertices of the tree $\operatorname{PGL}_2(\mathbb{Q}_p)/\operatorname{PGL}_2(\mathbb{Z}_p)$, write $v \sim w$ if v and w are neighbors, write $v \sim_\ell w$ if v and w have distance ℓ in the tree, and write $v \sim_{\leqslant \ell} w$ if v and w have distance no more than ℓ in the tree structure.

Definition 3.16. *Let p be a prime, and let f be a function on the quotient space $\operatorname{PGL}_2(\mathbb{Q}_p)/\operatorname{PGL}_2(\mathbb{Z}_p)$. Then the tree-Laplacian or Hecke operator T_p is defined by the normalized sum*

$$T_p(f)(v) = \frac{1}{p+1} \sum_{w \sim v} f(w)$$

of the values of f on the neighbors w of v.

3.2.2 The Embedded Trees

We already showed (see the discussion at the end of Section 3.1.4) that for any point

$$\operatorname{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])(g, I)\operatorname{PGL}_2(\mathbb{Z}_p)$$

in

$$X_\infty = \operatorname{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])\backslash \operatorname{PGL}_2(\mathbb{R}) \times \operatorname{PGL}_2(\mathbb{Q}_p)/\operatorname{PGL}_2(\mathbb{Z}_p),$$

the image in X_∞ of the orbit

$$\operatorname{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])\{g\} \times \operatorname{PGL}_2(\mathbb{Q}_p) \subseteq X_{\infty,p}$$

(obtained by taking the right quotient by $K_p = \mathrm{PGL}_2(\mathbb{Z}_p)$) is isomorphic to the $(p+1)$-regular tree

$$\mathrm{PGL}_2(\mathbb{Q}_p)/\,\mathrm{PGL}_2(\mathbb{Z}_p)$$

discussed in Section 3.2.1. We will call this image the *embedded Hecke tree* through $\mathrm{PGL}_2(\mathbb{Z})g$.

Proposition 3.17. *The Hecke operator T_p as defined in Definition* 3.2 *agrees with the tree-Laplacian T_p as defined in Definition* 3.16 *when applied to each embedded Hecke tree. Write $S_{p^k}(f)(v) = \sum_{w\sim_k v} f(w)$ for the summing operator over the distance k neighbors $w \in T$ of $v \in T$. Then S_{p^k} applied to each embedded Hecke tree gives the operator*

$$S_{p^k}(f)\,(\mathrm{PGL}_2(\mathbb{Z})g) = \sum_{\substack{\eta\in\mathrm{GL}_2(\mathbb{Z})\backslash\,\mathrm{Mat}_{22}(\mathbb{Z}),\\ p\nmid\eta,\det(\eta)=p^k}} f(\mathrm{PGL}_2(\mathbb{Z})\eta g), \tag{3.8}$$

where the notation $\mathrm{GL}_2(\mathbb{Z})\backslash\,\mathrm{Mat}_{22}(\mathbb{Z})$ means that we take, from each $\mathrm{GL}_2(\mathbb{Z})$ orbit (under left multiplication) in $\mathrm{Mat}_{22}(\mathbb{Z})$ satisfying the additional properties, one representative in the summation.

Proof. Note that the second part of the proposition implies the first: If $k = 1$ then $T_p = \frac{1}{p+1}S_p$ and $\eta \in \mathrm{Mat}_{22}(\mathbb{Z})$ with $\det\eta = p$ defines an index p subgroup $\Lambda_\eta = \mathbb{Z}^2\eta g$ of $\Lambda = \mathbb{Z}^2 g$ as in Definition 3.2. For this, also notice that $\Lambda_\eta = \Lambda_{\eta'}$ if and only if $\mathbb{Z}^2\eta = \mathbb{Z}^2\eta'$ and if and only if $\mathrm{GL}_2(\mathbb{Z})\eta = \mathrm{GL}_2(\mathbb{Z})\eta'$.

Now let $k \geqslant 1$ be arbitrary. By Proposition 3.15, $w \in T$ is of distance k to $v = \mathrm{PGL}_2(\mathbb{Z}_p) \in T = \mathrm{PGL}_2(\mathbb{Q}_p)/\,\mathrm{PGL}_2(\mathbb{Z}_p)$ if $w = g_p\,\mathrm{PGL}_2(\mathbb{Z}_p)$ for some $g_p \in \mathrm{Mat}_{22}(\mathbb{Z})$ with $p \nmid g_p$ and $\det(g_p) = p^k$. Also by Proposition 3.15, we can in fact choose g_p to be a product of k elements of $\mathcal{N}$ (where $\mathcal{N}$ is chosen as in the discussion before Lemma 3.14), so that $g_p \in \mathrm{Mat}_{22}(\mathbb{Z})$.

We define $\eta = p^k g_p^{-1} \in \mathrm{Mat}_{22}(\mathbb{Z})$ with

$$\det\eta = p^{2k}\det(g_p)^{-1} = p^k.$$

We claim that $p \nmid \eta$. To see this, notice that $p|\eta$ implies that $\eta_1 = p^{-1}\eta \in \mathrm{Mat}_{22}(\mathbb{Z})$, $\det(\eta_1) = p^{k-2}$, and

$$g_p = p^k(p^k g_p^{-1})^{-1} = p^{k-1}\eta_1^{-1} = p(p^{k-2}\eta_1^{-1}) \in p\,\mathrm{Mat}_{22}(\mathbb{Z}),$$

which is a contradiction of the choice of g_p. Clearly

$$g_{p,1}\,\mathrm{PGL}_2(\mathbb{Z}_p) = g_{p,2}\,\mathrm{PGL}_2(\mathbb{Z}_p)$$

if and only if the associated $\eta_i = p^k g_{p,i}^{-1}$ satisfy

$$\mathrm{GL}_2(\mathbb{Z})\eta_1 = \mathrm{GL}_2(\mathbb{Z})\eta_2.$$

Since

$$\begin{aligned}\mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])(g, g_p) &= \mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])(\eta^{-1}, \eta^{-1})(g, g)\\ &\in \mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])(\eta^{-1} g) \times \mathrm{PGL}_2(\mathbb{Z}_p),\end{aligned}$$

this correspondence between g_p and η proves the proposition. □

3.2.3 Chebyshev Polynomials of the Second Kind and Recurrence Relations for T_p

Definition 3.18. *The* Chebyshev polynomials of the second kind *are the polynomials $U_n \in \mathbb{Z}[x]$ defined recursively by*

$$\begin{aligned}U_0(x) &= 1,\\ U_1(x) &= 2x, \text{ and}\\ U_{n+1}(x) &= 2xU_n(x) - U_{n-1}(x)\end{aligned}$$

for $n \geqslant 1$.

Lemma 3.19. *If $x = \cos\theta$, then*

$$U_n(x) = \frac{\sin[(n+1)\theta]}{\sin\theta}, \tag{3.9}$$

and if $x = \cosh\theta$, then

$$U_n(x) = \frac{\sinh[(n+1)\theta]}{\sinh\theta}.$$

The lemma is easily checked using the standard addition formulas for $\sin\theta$ and $\sinh\theta$ (which are identical).

Proof of Lemma 3.19. First notice that if $x = \cos\theta$, then

$$\frac{\sin 2\theta}{\sin\theta} = \frac{2\sin\theta\cos\theta}{\sin\theta} = 2x,$$

which proves the case $n = 1$. Now assume that the lemma holds for $n - 1$ and for n, for some $n \geqslant 2$. Then

$$\begin{aligned}\frac{\sin[(n+2)\theta]}{\sin\theta} &= \frac{\sin[(n+1)\theta+\theta]}{\sin\theta} + \frac{\sin[(n+1)\theta-\theta]}{\sin\theta} - U_{n-1}(x)\\ &= \frac{\sin[(n+1)\theta]\cos\theta}{\sin\theta} + \frac{\cos[(n+1)\theta]\sin\theta}{\sin\theta}\\ &\qquad + \frac{\sin[(n+1)\theta]\cos(-\theta)}{\sin\theta} + \frac{\cos[(n+1)\theta]\sin(-\theta)}{\sin\theta} - U_{n-1}(x)\\ &= 2xU_n(x) - U_{n-1}(x) = U_{n+1}(x),\end{aligned}$$

giving the case $n+1$. As the addition formula for $\sinh\theta$ is the same as the addition formula for $\sin\theta$, the proof of the second case is identical. □

The relationship between regular trees and Chebyshev polynomials – and hence the reason for our interest in these polynomials – is revealed in the next lemma.

Lemma 3.20. *Let $\mathbf{U}_n$ be the operator that maps any function f on a $(p+1)$-regular tree to the function $\mathbf{U}_n(f)$ defined by*

$$\mathbf{U}_n(f)(v) = \frac{1}{p^{n/2}} \sum_{\substack{k \leqslant n,\\ k \equiv n (\mathrm{mod} 2)}} \sum_{w \sim_k v} f(w),$$

where as before $w \sim_k v$ means that w and v have distance k in the tree T. Then the sequence of operators $(\mathbf{U}_n)$ satisfies

$$\begin{aligned} \mathbf{U}_0 &= I, \\ \mathbf{U}_1 &= \frac{p+1}{\sqrt{p}} T_p = \frac{1}{\sqrt{p}} S_p, \text{ and} \\ \mathbf{U}_{n+1} &= \frac{1}{\sqrt{p}} S_p \circ \mathbf{U}_n - \mathbf{U}_{n-1} \end{aligned}$$

for $n \geqslant 1$, where

$$S_p(f)(v) = \sum_{w \sim v} f(w)$$

is the summing operator over the neighbors.

Proof. The cases $n=0$ and $n=1$ hold trivially by definition. For $n \geqslant 2$, we need to calculate the product

$$\begin{aligned} \frac{1}{\sqrt{p}} S_p \left(\mathbf{U}_n(f)\right)(v) &= \frac{1}{\sqrt{p}} \sum_{v' \sim v} \mathbf{U}_n(f)(v') \\ &= \frac{1}{p^{(n+1)/2}} \left(\sum_{v' \sim v} \sum_{\substack{k \leqslant n,\\ k \equiv n (\mathrm{mod} 2)}} \sum_{w \sim_k v'} f(w) \right). \end{aligned}$$

Here there are two possibilities for the distance from w to v. The distance could be $k+1$, in which case v' is the unique element with distance 1 to v in the direction of the path to w. The distance could be $k-1$, in which case v' could be any of the p neighbors of v away from the direction of the path to w (see Figure 3.1 for the case $p=2$). There is one exceptional case in this description: If $k=1$ and $w=v$ then all $(p+1)$ choices of neighbors of v give rise to $w=v$.

We reorder the summation, and then split the inner summation into two sums depending on the two cases, keeping track of the multiplicities coming from the choices for v' in the second case, giving

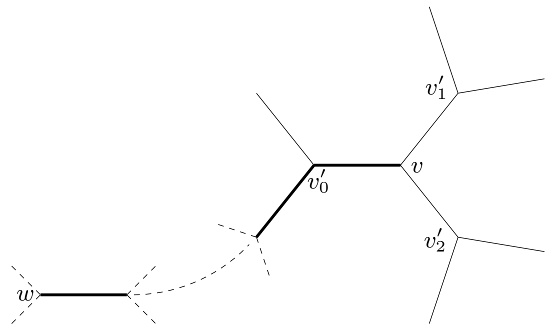

Fig. 3.1. In the $(p+1)$-regular tree there is always a unique geodesic path from v to w. Here v_0' is the unique neighbor of v on this path, and v_1', v_2' are further from w than v is.

$$\frac{1}{p^{(n+1)/2}} \sum_{\substack{k\leqslant n,\\ k\equiv n(\mathrm{mod}2)}} \left(\sum_{w\sim_{k+1}v} f(w) + p \sum_{w\sim_{k+1}v} f(w) + \delta_{k-1} f(v) \right).$$

In the case $k=0$ we adopt the convention that the sum over $w \sim_{-1} v$ sums over an empty set and so can be ignored. The extra term involving

$$\delta_{k1} = \begin{cases} 0 & k \neq 1, \\ 1 & k = 1 \end{cases}$$

corrects the multiplicity as discussed above. Shifting the summation over k to a summation over $\ell = k+1$ (respectively $\ell = k-1$), we get

$$\frac{1}{p^{(n+1)/2}} \sum_{\substack{\ell\leqslant n+1,\\ \ell\equiv n+1(\mathrm{mod}2)}} \sum_{w\sim_\ell v} f(w) + \frac{1}{p^{(n-1)/2}} \sum_{\substack{\ell\leqslant n-1,\\ \ell\equiv n-1(\mathrm{mod}2)}} \sum_{w\sim_\ell v} f(w)$$

which is equal to

$$\mathbf{U}_{n+1}(f)(v) + \mathbf{U}_{n-1}(f)(v),$$

proving the lemma. □

Lemma 3.21. *For* $\mathbf{U}_n' = (-1)^n \mathbf{U}_n$ *the following recurrence formulas hold:*

$$\begin{aligned} \mathbf{U}_0' &= I, \\ \mathbf{U}_1' &= -\frac{1}{\sqrt{p}} S_p, \ \textit{and} \\ \mathbf{U}_{n+1}' &= \left(-\frac{1}{\sqrt{p}} S_p\right) \mathbf{U}_n' - \mathbf{U}_{n-1}' \end{aligned}$$

for $n \geqslant 1$.

This is easily proved using Lemma 3.20 by keeping track of signs. The next proposition is the main step towards the goal of establishing Hecke p-recurrence.

Proposition 3.22. *There exists an absolute constant $c_0 > 0$ with the following property. Suppose that T is a $(p+1)$-regular tree, and let f be a non-zero eigenfunction of S_p (or, equivalently, of T_p) with real eigenvalue $\lambda \in \mathbb{R}$. Then*

$$\sum_{w \sim_{\leqslant n} v} |f(w)|^2 \geqslant c_0 n |f(v)|^2$$

for all $n \geqslant 0$.

We emphasize that the constant c_0 is independent of n, λ, and f (and also of p, which is less important for our purposes). We define the operator

$$S_{p^k}(f)(v) = \sum_{w \sim_k v} f(w),$$

which can easily be defined as a linear combination of $\mathbf{U}_k$ and $\mathbf{U}_{k-1}$, which in turn are polynomials in S_p by Lemma 3.20. This shows that S_p, $\mathbf{U}_k$, and S_{p^k} all commute.

Proof of Proposition 3.22 in the 'non-tempered' case $|\lambda| > 2\sqrt{p}$. We may assume that $\lambda > 0$, for otherwise we could simply work with the operators $-S_p$ and $\mathbf{U}'_k$ as in Lemma 3.21, instead of S_p and $\mathbf{U}_k$. We set $\cosh\theta = \frac{\lambda}{2\sqrt{p}}$, so that

$$\mathbf{U}_0(f) = f = U_0(\cosh\theta) f,$$
$$\mathbf{U}_1(f) = \frac{\lambda}{\sqrt{p}} f = U_1(\cosh\theta) f,$$

and, by Lemma 3.20, Definition 3.18 and induction,

$$\begin{aligned}
\mathbf{U}_{n+1}(f) &= \frac{1}{\sqrt{p}} S_p \mathbf{U}_n(f) - \mathbf{U}_{n-1}(f) \\
&= \frac{1}{\sqrt{p}} S_p(f) U_n(\cosh\theta) - U_{n-1}(\cosh\theta) f \\
&= \left[2\cosh\theta U_n(\cosh\theta) - U_{n-1}(\cosh\theta)\right] f \\
&= U_{n+1}(\cosh\theta) f.
\end{aligned}$$

By Lemma 3.19, this shows that for $n = 2m$,

$$\sum_{k=0}^{m} \sum_{w \sim_{2k} v} f(w) = p^m \frac{\sinh(2m+1)\theta}{\sinh\theta} f(v).$$

However, we have

$$\frac{\sinh(2m+1)\theta}{\sinh\theta} \geqslant (2m+1) \tag{3.10}$$

by convexity of the function $\theta \mapsto \sinh\theta$. Together with the Cauchy–Schwartz inequality, applied on the set $\{w \mid w \sim_{2k} v, k = 0, \dots, m\}$ of cardinality

$$1 + p(p+1) + \cdots + p^{2m-1}(p+1)$$

we get the inequality

$$\begin{aligned} p^m(2m+1)|f(v)| &= \left|\sum_{k=0}^{m} \sum_{w\sim_{2k}v} f(w)\right| \\ &\leqslant \left(\sum_{k=0}^{m} \sum_{w\sim_{2k}v} |f(w)|^2\right)^{1/2} \left(1+p(p+1)+\cdots+p^{2m-1}(p+1)\right)^{1/2}. \end{aligned}$$

Using the geometric series we may bound the last square root by

$$2p^{1/2}(p+1)^{1/2}\left(1+p^2+\cdots+p^{2m-2}\right)^{1/2} \leqslant 4p\left(\frac{p^{2m}-1}{p^2-1}\right)^{1/2} \leqslant 8p^m,$$

where we have used (without much concern for optimality) the factor 2 repeatedly to ignore the leading 1 in the original expression and the ratio between p and $p+1$. Squaring, we deduce that

$$\sum_{k=0}^{m} \sum_{w\sim_{2k}v} |f(w)|^2 \geqslant \tfrac{1}{4^2}(2m+1)^2|f(v)|^2,$$

which implies the proposition in this case. □

The proof of Proposition 3.22 above for the ‘non-tempered’ case in which $|\lambda| > 2\sqrt{p}$ was quite direct, since $\theta \mapsto \frac{\sinh(2m+1)\theta}{\sinh\theta}$ is monotone in m, and is easily bounded. In the second case, where we will use the formula 3.9, more care is needed.

Proof of Proposition 3.22 in the ‘tempered’ case $|\lambda| \leqslant 2\sqrt{p}$. Once more we may assume that $\lambda > 0$, and set $\cos\theta = \frac{\lambda}{2\sqrt{p}}$ for some $\theta \in [0, \frac{\pi}{2}]$, so that

$$\begin{aligned} \mathbf{U}_n(f) &= U_n(\cos\theta) f \\ &= \frac{\sin[(n+1)\theta]}{\sin\theta} f, \end{aligned}$$

just as in the beginning of the proof of the non-tempered case. Let us write $\lambda_{p^{2k}}$ for the eigenvalue of f under $S_{p^{2k}}$, so that

$$\sum_{k=0}^{m} \lambda_{p^{2k}} = p^m \frac{\sin[(2m+1)\theta]}{\sin\theta}. \tag{3.11}$$

Note that if θ and $(2m+1)\theta$ are sufficiently small, say $m\theta \leqslant c_1$, then $\sin\theta$ (respectively $\sin(2m+1)\theta$) is approximately θ (respectively $(2m+1)\theta$). In particular, we have in this case

$$\frac{\sin[(2m+1)\theta]}{\sin\theta} \geqslant m,$$

which is a weaker version of (but still similar to) the estimate in (3.10). Arguing just as in the non-tempered case, this implies that there exists some absolute constant $c_2 > 0$ with

$$\sum_{k=0}^{m} \sum_{w \sim_{2k} v} |f(w)|^2 \geqslant c_2 m |f(v)|^2, \tag{3.12}$$

whenever $m = 1$ (which holds trivially if $c_2 \leqslant 1$) or $m\theta \leqslant c_1$.

To handle the further cases, we again apply Cauchy–Schwartz, applied this time on the set $\{w \mid w \sim_{2k} v\}$, to obtain

$$\begin{aligned} \sum_{w \sim_{2k} v} |f(w)|^2 &\geqslant \frac{1}{(p+1)p^{2k-1}} \Bigg| \sum_{w \sim_{2k} v} f(w) \Bigg|^2 & (3.13) \\ &= \frac{1}{(p+1)p^{2k-1}} |\lambda_{p^{2k}}| |f(v)|^2 \\ &\geqslant \tfrac{1}{2} \left[\tfrac{\sin[(2k+1)\theta]}{\sin\theta} - \tfrac{\sin[(2k-1)\theta]}{p\sin\theta} \right]^2 |f(v)|^2, & (3.14) \end{aligned}$$

where we used the inequality $(p+1) \leqslant 2p$ and (3.11) for $m = k$ and $m = k-1$ to express $\lambda_{p^{2k}}$ as a difference. Since $p \geqslant 2$, we have

$$\left| \frac{\sin[(2k+1)\theta]}{\sin\theta} - \frac{\sin[(2k-1)\theta]}{p\sin\theta} \right| \geqslant c_3 \frac{1}{|\sin\theta|} \geqslant c_3 \tag{3.15}$$

for some absolute constant $c_3 > 0$, whenever

$$(2k+1)\theta \pmod{\pi} \in I = [\tfrac{\pi}{4} - \tfrac{\pi}{100}, \tfrac{3\pi}{4} + \tfrac{\pi}{100}]$$

as in this case $|\sin[(2k+1)\theta]| \geqslant \frac{1}{2} + c_3$.

We now show how the estimates (3.12) and (3.14), (3.15) together prove the proposition. We claim that there exists some absolute constants c_4, c_5 such that if $m\theta \geqslant c_4$ then

$$\sum_{k=1}^{m} \mathbf{1}_I \left((2k+1)\theta \pmod{\pi}\right) > c_5 m. \tag{3.16}$$

This then implies that we can use the estimate (3.15) for at least $c_5 m$ of the various $k \in [0, m]$, giving

$$\sum_{k=0}^{m} \sum_{w \sim_{2^k} v} |f(w)|^2 \geqslant \tfrac{1}{2} c_3 c_5 m |f(v)|^2 \tag{3.17}$$

whenever $m\theta \geqslant c_4$.

Together with (3.12) we get the proposition: If $m = 1$ or $m\theta < c_1$, then (3.12) applies. If $m\theta > c_4$ then (3.17) applies. Thus

$$\sum_{k=0}^{m} \sum_{w \sim_{2^k} v} |f(w)|^2 \geqslant \underbrace{\min\{c_2 \tfrac{c_1}{2c_4}, \tfrac{1}{2} c_3 c_5\}}_{c_0} m |f(v)|^2 \tag{3.18}$$

holds for any m. In fact (3.18) is weaker than (3.17) if $m\theta \geqslant c_4$, and if $m\theta \leqslant c_4$ then $m' = \max\{1, \lfloor \frac{c_1}{c_4} m \rfloor\}$ satisfies (3.12) and $m' \geqslant \frac{c_1}{2c_4} m$, which implies (3.18).

Thus it remains to prove the claim in (3.16). To simplify the notation, we divide by π and consider the interval

$$I = [\tfrac{1}{4} - \tfrac{1}{100}, \tfrac{3}{4} + \tfrac{1}{100}] \subseteq [0, 1],$$

whose length strictly exceeds $\frac{1}{2}$. Moreover, we will write $A \ll B$ if there exists an absolute constant $c > 0$ with $A \leqslant cB$.

The proof of the claim consists of several cases. We start by assuming $\theta \in [0, \frac{1}{4}]$, in which case $2\theta \leqslant \frac{1}{2}$. For any $\alpha \in [0, \frac{1}{4}]$, there exists some $k \in \mathbb{Z}$ with $\alpha + 2k\theta \in [\frac{1}{4}, \frac{3}{4}] \subseteq I$. This k satisfies $2k\theta \leqslant \frac{3}{4}$, so $k \leqslant \frac{3}{8\theta}$. Indeed, more is true: If θ is sufficiently small, then

$$\left|\{k \in [0, \tfrac{3}{8\theta}] \cap \mathbb{Z} \mid \alpha + 2k\theta \in I\}\right| \geqslant \tfrac{3}{16\theta},$$

since the length of I exceeds $\frac{1}{2}$. In either case this shows that

$$\left|\{k \in [0, \tfrac{3}{8\theta}] \cap \mathbb{Z} \mid \alpha + 2k\theta \in I\}\right| \gg \tfrac{1}{\theta}. \tag{3.19}$$

Now suppose that $m\theta \geqslant c_4 = 1$. Then we may split the sequence

$$\theta, 3\theta, \dots, (1 + 2m)\theta$$

into consecutive subintervals

$$\begin{aligned}
&\alpha_0 = \theta, 3\theta, \dots, (1 + 2\ell_0)\theta \in [0, 1), \\
&\alpha_1 = (1 + 2\ell_0 + 2)\theta, \alpha_1 + 2\theta, \dots, \alpha_1 + 2\ell_2\theta \in [1, 2),
\end{aligned}$$

and so on, where the lengths ℓ_i satisfy $\theta \ll \ell_i \ll \frac{1}{\theta}$ (and the last interval may be shorter). Applying (3.19) to each of the these intervals (except the last one) and summing, we obtain

$$|\{k \in \mathbb{Z} \cap [0, m] \mid 1 + 2k\theta \pmod 1 \in I\}| \gg m \tag{3.20}$$

as required. By the symmetry of the claim, this also deals with the case $\theta \in [\frac{3}{4}, 1]$.

Now suppose $\theta \in [\frac{1}{4}, \frac{3}{4}]$, and write $\theta = \frac{1}{2} + \kappa$ for some κ with $|\kappa| \leqslant \frac{1}{4}$. Then we are interested in the sequence

$$\theta, \theta + 2\kappa, \theta + 4\kappa, \ldots, \theta + 2m\kappa \pmod 1.$$

As before, the angle of rotation 2α is less than $\frac{1}{2}$, and so we must have visits to the interval I modulo 1. There is one difference to the previous case: If κ is small then, since θ lies in the interior of I, we have

$$\theta, \theta + 2\alpha, \ldots, \theta + k\alpha \in I$$

for some $k \gg \frac{1}{\alpha}$. In other words, in this case the sequence starts with a long segment in I, followed by a comparable length segment outside I, and so on. This shows (3.20) for all m. If κ is bigger, then the visits to I are frequent enough to also give (3.20). □

Proposition 3.22 has an interesting corollary. Even though we will not need this directly, our arguments below are closely related. Recall that the operator S_p is self-adjoint, since it can be expressed as a polynomial in T_p with real coefficients.

Corollary 3.23. *For $p \geqslant 2$, the operator S_p restricted to $L^2(T)$ for the $(p+1)$-regular tree T has no eigenfunctions.*

Corollary 3.24. *Let μ be a weak*-limit of $|\phi_i|^2 \,\mathrm{d}m_X$ as in Theorem 3.12, and let $f \in C_c(X)$ be non-negative. Then*

$$\int \sum_{k=0}^{n} S_{p^k}(f) \geqslant c_0 n \int f \,\mathrm{d}\mu.$$

The same holds for any $f \geqslant 0$.

Here S_n is the operator defined on the $(p+1)$-regular tree T, extended to

$$\mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])\backslash \mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_p) / \mathrm{PGL}_2(\mathbb{Z}_p),$$

where, by the discussion on page 42 at the end of Section 3.1.4, every set of the form

$$\mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])(g_\infty, g_p)\mathrm{PGL}(\mathbb{Q}_p) / \mathrm{PGL}_2(\mathbb{Z}_p)$$

is identified with $\mathrm{PGL}_2(\mathbb{Q}_p) / \mathrm{PGL}_2(\mathbb{Z}_p)$.

Proof of Corollary 3.24. Let $\phi \in L^2_{m_X}$, then

$$\begin{aligned} \int S_{p^k}(f)|\phi|^2 \,\mathrm{d}m_X &= \left\langle \sum_{k=0}^{n} S_{p^k}(f), |\phi|^2 \right\rangle_{L^2_{m_X}} \\ &= \left\langle f, \sum_{k=0}^{n} S_{p^k}(|\phi|^2) \right\rangle_{L^2_{m_X}}, \end{aligned}$$

since S_{p^k} can be expressed as a polynomial in T_p with real coefficients, and T_p is self-adjoint. By Proposition 3.22,

$$\sum_{k=0}^{n} S_{p^k}(|\phi|^2) \geqslant c_0 n|\phi|^2,$$

which shows that

$$\int \sum_{k=0}^{n} S_{p^k}(f)|\phi|^2 \, \mathrm{d}m_X \geqslant c_0 n \int f|\phi|^2 \, \mathrm{d}m_X.$$

Applying this to $\phi = \phi_i$ and taking the weak*-limit, we get the statement in the corollary for any positive $f \in C_c(X)$.

Now let $K \subseteq X$ be compact, and choose some positive sequence (f_n) in $C_c(X)$ with $f_n \searrow \mathbf{1}_K$. Then by dominated convergence the estimate for each f_n implies the same property for $f = \mathbf{1}_K$. Similarly, if $F = \bigcup_n K_n$ is a countable union of compact sets, the monotone convergence theorem implies the estimate for $f = \mathbf{1}_F$. Finally, if $B \subseteq X$ is any Borel set, then there exists some F of this form with $F \subseteq B$ and $\mu(B \smallsetminus F) = 0$. This shows that

$$\int \sum_{k=0}^{n} S_{p^k}(\mathbf{1}_B) \, \mathrm{d}\mu \geqslant \int \sum_{k=0}^{n} S_{p^k}(\mathbf{1}_F) \, \mathrm{d}\mu \geqslant c_0 n \int \mathbf{1}_F \, \mathrm{d}\mu = c_0 n \int \mathbf{1}_B \, \mathrm{d}\mu.$$

Approximating a measurable non-negative function from below by simple functions, the corollary follows from monotone convergence. □

Proof of Theorem 3.12. Let $B \subseteq X$ be Borel measurable, with $\mu(B) > 0$. Suppose that μ is not Hecke-p-recurrent, and that for some positive measure subset B' of B we have that

$$x = \Gamma(g_\infty, I)K_p \in B', \Gamma(g_\infty, h)K_p \in B$$

implies that h belongs to some compact subset $M \subseteq \mathrm{PGL}_2(\mathbb{Q}_p)$ (which may depend on x).

As $\mathrm{PGL}_2(\mathbb{Q}_p)$ is σ-compact, this shows that there exists a Borel subset $B \subseteq X$ with $\mu(B) > 0$, and a compact subset $M \subseteq \mathrm{PGL}_2(\mathbb{Q}_p)$ with the property that

$$x = \Gamma(g_\infty, I)K_p \in B, \Gamma(g_\infty, h)K_p \in B$$

implies $h \in M$.

We may replace B with another subset of B to simplify the argument. Let $x_0 = \Gamma(g_0, I)K_p \in \operatorname{Supp} \mu|_B$. By the discussion at the end of Section 3.1.4 on page 42, there is a bijective correspondence between

$$\Gamma\{g_\infty\} \times \mathrm{PGL}_2(\mathbb{Q}_p) / \mathrm{PGL}_2(\mathbb{Z}_p)$$

and

$$\mathrm{PGL}_2(\mathbb{Q}_p)/\,\mathrm{PGL}_2(\mathbb{Z}_p).$$

Therefore, there exists some $\delta > 0$ such that

$$h \in M \implies h \in \mathrm{PGL}_2(\mathbb{Z}_p),$$

or

$$\mathsf{d}_X\left(\Gamma(g_0, I)K_p, \Gamma(g_0, h)K_p\right) > 2\delta.$$

Now let $B' = B \cap B_\delta(x)$, and suppose that

$$x = \Gamma(g_\infty, I)K_p \in B', \Gamma(g_\infty, h)K_p \in B'$$

where we may assume that $\mathsf{d}_{\mathrm{SL}_2(\mathbb{R})}(g_0, g_\infty) < \delta$. Then $h \in M$ by the choice of B, and

$$\mathsf{d}_X\left(\Gamma(g_0, h)K_p, \Gamma(g_\infty, h)K_p\right) < \delta$$

since the action of $\mathrm{SL}_2(\mathbb{R}) \subseteq \mathrm{PGL}_2(\mathbb{R})$ (which is used to measure distances in X) commutes with the action of $\mathrm{PGL}_2(\mathbb{Q}_p)$ on

$$\Gamma\backslash\mathrm{PGL}_2(\mathbb{R}) \times \mathrm{PGL}_2(\mathbb{Q}_p).$$

Therefore, we must have $h \in K_p = \mathrm{PGL}_2(\mathbb{Z}_p)$. To summarize: We may assume that $B \subseteq X$ has positive measure, and

$$\Gamma(g_\infty, I)K_p, \Gamma(g_\infty, h)K_p \in B$$

implies $h \in K_p$. In other words, for $x \in B$ the only point of the tree in X associated to x that belongs to B is x.

We now apply Corollary 3.24 to $f = \mathbf{1}_B$ and see that

$$\int \sum_{k=0}^{n} S_{p^k}(\mathbf{1}_B)\,\mathrm{d}\mu \geqslant c_0 n \mu(B)$$

for all $n \geqslant 1$. We claim that $\sum_{k=0}^{n} S_{p^k}(\mathbf{1}_B) \leqslant 1$, which then gives the contradiction as $1 \geqslant c_0 n \mu(B)$ cannot hold for all $n \geqslant 1$.

If $\mathbf{U}_n(\mathbf{1}_B)(y) \geqslant 2$ for some n and $y \in X$, then the tree through y contains at least two points x_1, x_2 within distance no more than n that belong to B. However, this is impossible as in this case x_2 belongs to the tree through x_1 (with distance no more than $2n$) which contradicts the construction of B. □

By Proposition 3.17, the operator $\mathbf{U}_n$, as defined in Lemma 3.20, when applied to an embedded Hecke tree gives rise to the operator

$$\mathbf{U}_n(f)(\mathrm{PGL}_2(\mathbb{Z})g) = \frac{1}{p^{n/2}} \sum_{\substack{\eta \in \mathrm{GL}_2(\mathbb{Z})\backslash \mathrm{Mat}_{22}(\mathbb{Z}), \\ \det \eta = p^n}} .$$

Notice here that the absence of the condition $p \nmid \eta$ appearing in (3.8) gives rise to the various distances $k \leqslant n$ with $k \equiv n$ modulo 2. This operation is (with this normalization) usually called the p^n-Hecke operator.

3.3 Ruling out Scarring on Periodic Orbits

In this section we will show that arithmetic quantum limits on $\mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R})$ assign zero mass to compact geodesic orbits. The possibility of positive mass on periodic orbits for quantum limits is usually referred to as *scarring**. As we will see, Hecke-p-recurrence for all (or almost all) primes is enough to prove the following theorem, due to Lindenstrauss [18] (extending earlier work of Rudnick and Sarnak [24]). Recall that

$$A = \left\{ \begin{pmatrix} \mathrm{e}^t & \\ & \mathrm{e}^{-t} \end{pmatrix} \mid t \in \mathbb{R} \right\} \subseteq \mathrm{SL}_2(\mathbb{R})$$

denotes the diagonal subgroup.

Theorem 3.25. *Let $X = \mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R})$, and suppose that $\phi_i \in L^2(X)$ are Hecke eigenfunctions with $\|\phi_i\|_2 = 1$. Let μ be a weak*-limit of $|\phi_i|^2\,\mathrm{d}m_X$. Then $\mu(xA) = 0$ for any periodic (that is, compact) A-orbit xA for $x \in X$.*

The proof comprises three steps. First, we show that a periodic orbit xA corresponds to a real quadratic number field $\mathbb{F}$. Second, we discuss $\mathbb{F} \otimes_{\mathbb{Q}} \mathbb{Q}_p$, and choose p to be an 'inert' prime. Finally, we show that $\mu(xA) > 0$ and Hecke-p-recurrence are mutually exclusive.

Proposition 3.26. *If $x = \mathrm{SL}_2(\mathbb{Z})g \in X = \mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R})$ has periodic A-orbit, then $xa = x$ for some $a \in A \smallsetminus \{I\}$, and so*

$$ga = \gamma g \tag{3.21}$$

for some $\gamma \in \mathrm{SL}_2(\mathbb{Z})$. The algebra $\mathbb{F} = \mathbb{Q}(\gamma) \subseteq \mathrm{Mat}_{22}(\mathbb{Q})$ is a real quadratic field extension of $\mathbb{Q}$, and $\gamma \in \mathscr{O}_{\mathbb{F}}^{\times}$ is an algebraic unit.

Proof. The existence of a and γ is clear, as $A \cong \mathbb{R}$, and the only non-trivial compact quotients of $\mathbb{R}$ are of the form $\mathbb{R}/t\mathbb{R}$ for some $t \neq 0$. Suppose therefore that $a = \begin{pmatrix} \mathrm{e}^t & \\ & \mathrm{e}^{-t} \end{pmatrix}$ for some $t \neq 0$. Then (3.21) shows that the eigenvalues of γ must be $\mathrm{e}^t, \mathrm{e}^{-t} \neq \pm 1$. It follows that the characteristic polynomial of γ (with leading and trailing coefficients equal to ± 1) must be irreducible. This implies that $\mathbb{F}$ is a field, and that γ is an algebraic integer and unit in $\mathbb{F}$, and that the Galois embeddings (which may be obtained by conjugating $\mathbb{F}$ by g, and reading off one of the diagonal entries) are both real. □

We now recall some basic algebraic number theory for quadratic number fields $\mathbb{F}|\mathbb{Q}$. If $\mathscr{O}_{\mathbb{F}}$ denotes the *ring of algebraic integers*, then $\mathscr{O}_{\mathbb{F}}$ is a *Dedekind domain*, and every ideal $\mathscr{J} \subseteq \mathscr{O}_{\mathbb{F}}$ has a unique prime factorization

* For an introduction to the physical meaning behind (the absence of) this phenomenon, see the brief survey by Anantharaman and Nonnenmacher [1].

$$\mathscr{J} = \prod_{\mathscr{P}} \mathscr{P}^{v_{\mathscr{P}}(\mathscr{J})},$$

where the product runs over all prime ideals $\mathscr{P} \subseteq \mathscr{O}_{\mathbb{F}}$. Applying this to the ideal $\mathscr{J} = p\mathscr{O}_{\mathbb{F}}$ for a fixed prime $p \in \mathbb{Z}$, there are three possibilities:

- $p\mathscr{O}_{\mathbb{F}} = \mathscr{P}^2$ for some prime ideal $\mathscr{P} \subseteq \mathscr{O}_{\mathbb{F}}$ – in this case p is said to be *ramified* for $\mathbb{F}|\mathbb{Q}$;
- $p\mathscr{O}_{\mathbb{F}} = \mathscr{P}$ is a prime ideal in $\mathscr{O}_{\mathbb{F}}$ – in this case p is said to be *inert* for $\mathbb{F}|\mathbb{Q}$;
- $p\mathscr{O}_{\mathbb{F}} = \mathscr{P}_1\mathscr{P}_2$ is a product of two distinct prime ideals $\mathscr{P}_1, \mathscr{P}_2 \subseteq \mathscr{O}_{\mathbb{F}}$ – in this case p is said to be *split* for $\mathbb{F}|\mathbb{Q}$.

We also recall the following standard fact about quadratic number fields.

Proposition 3.27. *Let $\mathbb{F}|\mathbb{Q}$ be a quadratic number field. For any prime p the following properties are equivalent:*

(1) *p is ramified or inert for $\mathbb{F}|\mathbb{Q}$;*
(2) *$\mathbb{F} \otimes_{\mathbb{Q}} \mathbb{Q}_p|\mathbb{Q}_p$ is a field extension;*
(3) *the group of units of the ring $\mathscr{O}_{\mathbb{F}}[\frac{1}{p}]$ contains*
   - *$s = 1$ if $\mathbb{F}$ is complex, and*
   - *$s = 2$ if $\mathbb{F}$ is real*

   multiplicatively independent elements.*

Our final ingredient from algebraic number theory is the following result, which is an easy consequence of Dirichlet's theorem on primes in arithmetic progression.

Proposition 3.28. *Let $\mathbb{F}|\mathbb{Q}$ be a real quadratic number field. Then there are infinitely many inert primes $p \in \mathbb{Z}$ for $\mathbb{F}|\mathbb{Q}$.*

Proof. Let $\lambda = \sqrt{d} \in \mathscr{O}_{\mathbb{F}}$ with $d \in \mathbb{N}$ square-free be chosen so that $\mathbb{F} = \mathbb{Q}[\lambda]$, and let N_0 have the property that for any prime $p \geqslant N_0$ the local rings $\mathscr{O}_{\mathbb{F}}^{(p)}$ and $(\mathbb{Z}[\lambda])^{(p)}$ coincide. Then a prime $p \geqslant N_0$ is inert if $p\mathscr{O}_{\mathbb{F}}^{(p)}$ is prime in $\mathscr{O}_{\mathbb{F}}^{(p)}$, or equivalently if the characteristic polynomial of λ is irreducible modulo p. As $\lambda = \sqrt{d}$, this is the case if and only if d is a non-square modulo p, that is if $\left(\frac{d}{p}\right) = -1$ (where $\left(\frac{d}{p}\right)$ denotes the Legendre symbol as usual). If $d = q_1 \cdots q_\ell$ is the prime factorization of d, then

$$\begin{aligned}\left(\frac{d}{p}\right) &= \left(\frac{q_1}{p}\right) \cdots \left(\frac{q_\ell}{p}\right) \\ &= (-1)^{(p-1)(q_1-1)/4} \left(\frac{p}{q_1}\right) \cdots (-1)^{(p-1)(q_\ell-1)/4} \left(\frac{p}{q_\ell}\right)\end{aligned}$$

* That is, $\mathscr{O}_{\mathbb{F}}(\frac{1}{p})^{\times} \cong F \times \mathbb{Z}^s$ where F is a finite group of roots of unity in $\mathbb{F}$. See [31] for the equivalence of (1) and (2); the equivalence of (1) and (3) is a consequence of Dirichlet's S-unit theorem.

by quadratic reciprocity (with the understanding that if some prime $q_i = 2$, then the corresponding factor $(-1)^{(p^2-1)/8}$ has to be included). In either case, this implies that $\left(\frac{d}{p}\right) = -1$ is equivalent to a finite list of congruence conditions concerning p modulo $8d$. Choose $a \in \mathbb{N}$ coprime to $8d$ such that $p \equiv a \pmod{d}$ implies $\left(\frac{d}{p}\right) = -1$. By Dirichlet's theorem, there are infinitely many primes p congruent to a modulo d, and so $\left(\frac{d}{p}\right) = -1$ for those infinitely many primes as required. □

In fact, we will only need one inert prime.

Proof of Theorem 3.25. Let $xA \subseteq X$ be a compact A-orbit. Assume, for the purposes of a contradiction, that $\mu(xA) > 0$, and let $\mathbb{F}$ be the real quadratic number field associated to xA in the sense of Proposition 3.26. Let $p \in \mathbb{Z}$ be an inert prime for $\mathbb{F}|\mathbb{Q}$. By Theorem 3.12, the limit μ is Hecke-p-recurrent. Therefore, on replacing $x = \mathrm{SL}_2(\mathbb{Z})g$ by some $x' \in xA$ if necessary, there exists a sequence (h_n) in $\mathrm{PGL}_2(\mathbb{Q}_p)$ with

$$h_n K_p \neq h_m K_p$$

for $n \neq m$ such that

$$\mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])(g, h_n)K_p \in \mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])(g, I)AK_p \cong xA.$$

We wish to remove the p-adic extension from this statement. For any h_n, there is some $\gamma_n \in \mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])$ with $\gamma_n h_n \in K_p$, or equivalently $\gamma_n^{-1}K_p = h_n K_p$. With this, we get

$$K_p\gamma_n \neq K_p\gamma_m \tag{3.22}$$

for $n \neq m$, and that

$$\begin{aligned}\mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])(g, h_n)K_p &= \mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])(\gamma_n, \gamma_n)(g, h_n)K_p \\ &= \mathrm{PGL}_2(\mathbb{Z}[\tfrac{1}{p}])(\gamma_n g, I)K_p\end{aligned}$$

corresponds in $\mathrm{PGL}_2(\mathbb{Z})\backslash\mathrm{PGL}_2(\mathbb{R})$ to

$$\mathrm{PGL}_2(\mathbb{Z})\gamma_n g = \mathrm{PGL}_2(\mathbb{Z})ga_n \in \mathrm{PGL}_2(\mathbb{Z})gA.$$

Modifying γ_n on the left by some element of $\mathrm{PGL}_2(\mathbb{Z})$ if necessary (which does not change any of its other properties) we may assume that

$$\gamma_n g = ga_n \in \mathrm{PGL}_2(\mathbb{R}).$$

This and (3.21) together shows that γ_n is diagonalized by the same matrix g as γ, and so commutes with γ. We now make the implicit scalars in the above equations in $\mathrm{PGL}_2(\mathbb{R})$ explicit, so as to obtain equations in $\mathrm{Mat}_{22}(\mathbb{R})$. As the equivalence class γ_n belongs to $\mathrm{PGL}_2(\mathbb{Z}[\frac{1}{p}])$, we may represent it by a matrix (which we again denote by γ_n) in $\mathrm{Mat}_{22}(\mathbb{Z})$ with determinant in $\pm p^{\mathbb{N}}$. As $\mathbb{F}$ is a two-dimensional vector space over $\mathbb{Q}$, and the centralizer of γ is also two-dimensional (after tensoring with $\mathbb{R}$, this follows from (3.21)), we have $\gamma_n \in \mathbb{F}$. However, as $\mathbb{Z}[\gamma_n]$ is a finitely-generated over $\mathbb{Z}$, $\gamma_n \in \mathscr{O}_{\mathbb{F}}$ must be an integer. Finally, the determinant $\det(\gamma_n)$ equals the norm $N_{\mathbb{F}|\mathbb{Q}}(\gamma_n) = \pm p^{\ell}$, which shows that $\gamma_n \in (\mathscr{O}_{\mathbb{F}}[\frac{1}{p}])^{\times}$.

The prime p was chosen inert for $\mathbb{F}|\mathbb{Q}$, which by Proposition 3.27 means that $(\mathscr{O}_{\mathbb{F}}[\frac{1}{p}])^{\times}$ contains $\mathscr{O}_{\mathbb{F}}^{\times} p^{\mathbb{Z}}$ as a finite-index subgroup. Also notice that

$$\mathscr{O} = \mathbb{F} \cap \mathrm{Mat}_{22}(\mathbb{Z})$$

is an order in $\mathbb{F}$, so that $\mathscr{O}^{\times} \subseteq \mathscr{O}_{\mathbb{F}}$ is also of finite index. Together, we deduce that there exists a finite list

$$\eta_1, \dots, \eta_\ell \in (\mathscr{O}_{\mathbb{F}}[\tfrac{1}{p}])^{\times}$$

with the property that

$$\bigcup_{i=1}^{\ell} \mathscr{O}^{\times} p^{\mathbb{Z}} \eta_i = (\mathscr{O}_{\mathbb{F}}[\tfrac{1}{p}])^{\times}.$$

As $\gamma_n \in (\mathscr{O}_{\mathbb{F}}[\frac{1}{p}])^{\times}$ for all $n \geqslant 1$, there exist m, n with $m \neq n$ for which

$$\mathscr{O}^{\times} p^{\mathbb{Z}} \gamma_n = \mathscr{O}^{\times} p^{\mathbb{Z}} \eta_i = \mathscr{O}^{\times} p^{\mathbb{Z}} \gamma_m.$$

However, since $\mathscr{O}^{\times} \subseteq \mathrm{GL}_2(\mathbb{Z})$, this implies that

$$\mathrm{PGL}_2(\mathbb{Z}_p)\gamma_n = \mathrm{PGL}_2(\mathbb{Z}_p)\gamma_m.$$

This contradiction to (3.22) implies that $\mu(xA) = 0$, completing the proof of the theorem. $\square$

4

Establishing Positive Entropy

In this chapter we discuss measure-theoretic entropy and show how it can be established for Hecke–Maass cusp forms by using the Hecke operators. For this we first give the general definitions in Section 4.1, and then consider a commutative case of how positive entropy can be established in Section 4.2. This commutative case is a toy case in preparation for the later picture. After some more preparations in Section 4.3, we prove positive entropy of almost all ergodic components in Section 4.4.

4.1 Measure-Theoretic Entropy

We start with the basic definitions of measure-theoretic entropy; for full details see [11] or [29].

4.1.1 Definitions

Let $(X, \mathscr{B}, \mu)$ be a probability space, and let $T : X \to X$ be a measure-preserving map. Then the (static) entropy of a countable measurable partition $\xi = \{A_1, A_2, \dots\}$ of X is

$$H_\mu(\xi) = -\sum_{A_i \in \xi} \mu(A_i) \log \mu(A_i).$$

The (dynamical) entropy of ξ with respect to T is defined by

$$h_\mu(T, \xi) = \lim_{n\to\infty} \frac{1}{n} H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right)$$

where $\bigvee_{i=0}^{n-1} T^{-i}\xi$ denotes the common refinement of the partitions

$$\xi, T^{-1}\xi, \dots, T^{-(n-1)}\xi.$$

It is not difficult to show that the limit above exists by subadditivity along the following lines. It is not difficult to show that $H_\mu(\xi \vee \eta) \leqslant H_\mu(\xi) + H_\mu(\eta)$ for countable partitions ξ and η, and $H_\mu(T^{-1}\xi) = H_\mu(\xi)$, so

$$H_\mu\left(\bigvee_{i=0}^{m+n-1} T^{-i}\xi\right) \leqslant H_\mu\left(\bigvee_{i=0}^{m-1} T^{-i}\xi\right) + H_\mu\left(\bigvee_{i=m}^{m+n-1} T^{-i}\xi\right)$$
$$= H_\mu\left(\bigvee_{i=0}^{m-1} T^{-i}\xi\right) + H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\xi\right).$$

Finally the entropy of T is defined to be

$$h_\mu(T) = \sup_{\xi} h_\mu(T, \xi),$$

where the supremum is taken over all countable partitions ξ of X with finite entropy (equivalently, over all finite partitions since a countable partition can be approximated by a finite partition and the static entropy is a continuous function of the partition in an appropriate metric).

4.1.2 (Non)-Ergodicity

Recall that a measure-preserving transformation T on $(X, \mathscr{B}, \mu)$ (equivalently, the T-invariant probability measure μ on $(X, \mathscr{B})$) is said to be ergodic if any (equivalently, all) of the following equivalent conditions are satisfied:

- $T^{-1}B = B$ for some $B \in \mathscr{B}$ implies that $\mu(B) \in \{0, 1\}$ (there are no non-trivial T-invariant sets);
- $\mu(T^{-1}B \triangle B) = 0$ for $B \in \mathscr{B}$ implies that $\mu(B) \in \{0, 1\}$ (there are no non-trivial almost T-invariant sets);
- $f \circ T = f$ for $f \in L^2_\mu$ implies that f is equal to a constant almost everywhere.

We note that many natural invariant measures (for example, Haar measures) are often ergodic unless there are obvious obstructions to ergodicity. However, it is easy to give examples of non-ergodic measures. If μ_1 and μ_2 are two singular T-invariant probability measures, then the convex combination

$$\nu = s\mu_1 + (1 - s)\mu_2 \tag{4.1}$$

is another invariant probability measure, which is not ergodic for $s \in (0, 1)$.

In fact, it is possible to present any invariant probability measure ν as a (generalized) convex combination of T-invariant ergodic probability measures μ_θ. However, in general this requires an integral representation* over

* The convex combination (4.1) is a special case, with $\Xi = \{\mu_1, \mu_2\}$, $\rho(\{\mu_1\}) = s$, and $\rho(\{\mu_2\}) = 1 - s$.

a probability space $(\Xi, \mathscr{B}_\Xi, \rho)$. This is a consequence of Choquet's theorem, and allows us to write

$$\nu = \int_\Xi \nu_\theta \,\mathrm{d}\rho(\theta), \tag{4.2}$$

which is called the ergodic decomposition of ν (see [12, Sect. 4.2]). It is interesting to note that it is possible to take $\Xi = X$, $\rho = \nu$, and to define for almost every $x \in X$ "its" ergodic component $\nu_x^{\mathscr{E}}$, which as the name suggests is a T-invariant and ergodic probability measure on X, so that

$$\nu = \int \nu_x^{\mathscr{E}} \,\mathrm{d}\nu(x) \tag{4.3}$$

(here $\mathscr{E}$ denotes the sub-σ-algebra of T-invariant sets in $\mathscr{B}$, and $\nu_x^{\mathscr{E}}$ is the conditional measure of ν at x with respect to $\mathscr{E}$). Moreover, $\nu_{Tx}^{\mathscr{E}} = \nu_x^{\mathscr{E}}$ (this is a consequence of the construction, which is described in [12, Sect. 6.1]).

4.1.3 Entropy and the Ergodic Decomposition

If $\nu = s\mu_1 + (1-s)\mu_2$ for two invariant probability measures μ_1 and μ_2 that are singular, then one can show (as a consequence of the convexity of the map $x \mapsto -\log x$) that

$$sH_{\mu_1}(\xi) + (1-s)H_{\mu_2}(\xi) \leqslant H_\nu(\xi) \leqslant sH_{\mu_1}(\xi) + (1-s)H_{\mu_2}(\xi) + \log 2.$$

For the dynamical entropy, this immediately implies that

$$h_\nu(T,\xi) = sh_{\mu_1}(T,\xi) + (1-s)h_{\mu_2}(T,\xi),$$

and hence

$$h_\nu(T) = sh_{\mu_1}(T) + (1-s)h_{\mu_2}(T).$$

This generalizes (but not immediately) to the ergodic decomposition in (4.2), giving

$$h_\nu(T,\xi) = \int_\Xi h_{\mu_\theta}(T,\xi) \,\mathrm{d}\rho(\theta)$$

and

$$h_\nu(T) = \int_\Xi h_{\mu_\theta}(T) \,\mathrm{d}\rho(\theta).$$

We refer to [11, Sect. 4.4] for a proof of this.

4.1.4 Positive Entropy of Almost All Ergodic Components

It should now be clear what the assumption [E] in Theorem 1.3 means: Let

$$T(x) = x \begin{pmatrix} \mathrm{e}^{1/2} & \\ & \mathrm{e}^{-1/2} \end{pmatrix}$$

be the time-one map of the geodesic flow on the quotient $X = \Gamma \backslash \mathrm{SL}_2(\mathbb{R})$ by a lattice Γ. Then a T-invariant measure ν on X has *positive entropy for almost all ergodic components* if ν is written as in (4.2) and $h_{\nu_\theta}(T) > 0$ for ρ-almost every $\theta \in \Xi$ (equivalently, if $h_{\nu_x^{\mathscr{E}}}(T) > 0$ for ν-almost every x).

This condition may look like an extremely difficult thing to check – in particular, if ν is not known concretely, then neither can the decomposition space $(\Xi, \mathscr{B}_\Xi, \rho)$ be known in an explicit way. However, the following observation makes condition [E] a little easier to check.

Proposition 4.1. *Let T be a measure-preserving transformation of a probability space $(X, \mathscr{B}, \nu)$, and suppose that the ergodic decomposition of ν is given by equation* (4.2). *If there exists a partition ξ of X with $H_\mu(\xi) < \infty$ such that for any T-invariant set B of positive measure one has*

$$h_{(\frac{1}{\nu(B)})\nu|_B}(T, \xi) > 0,$$

then $h_{\mu_\theta}(T) > 0$ for ρ-almost every $\theta \in \Xi$.

We will use Proposition 4.1 as the criterion for positivity of (almost) all ergodic components.

Proof of Proposition 4.1. Suppose that the conclusion does not hold, and use the ergodic decomposition from (4.3), so that the set

$$B = \{x \in X \mid h_{\nu_x^{\mathscr{E}}}(T) = 0\}$$

is measurable with positive measure. Since $\nu_{Tx}^{\mathscr{E}} = \nu_x^{\mathscr{E}}$, the set B is also T-invariant. Thus the ergodic decomposition of the measure $\frac{1}{\nu(B)}\nu|_B$ (that is, of ν restricted to B and then normalized to be a probability measure again) is given by

$$\frac{1}{\nu(B)}\nu|_B = \frac{1}{\nu(B)} \int_B \nu_x^{\mathscr{E}} \,\mathrm{d}\nu(x),$$

so that

$$h_{(\frac{1}{\nu(B)})\nu|_B}(T, \xi) = \frac{1}{\nu(B)} \int_B h_{\nu_x^{\mathscr{E}}}(T, \xi) \,\mathrm{d}\nu(x) = 0$$

by the choice of B, which contradicts the assumption of the proposition. □

4.1.5 Entropy as a Decay Rate

We defined the (static) entropy $H_\nu(\xi)$ of a partition ξ as the weighted average of $-\log \nu(A_j)$ over all the elements $A_j \in \xi$ of the partition. Therefore, if we wish to show that $h_\nu(T, \xi) \geqslant h$ for some $h > 0$, then it would be sufficient to find a constant $c > 0$ such that

$$H_\nu\left(\xi \vee \cdots \vee T^{-(n-1)}\xi\right) \geqslant nh - c$$

for all $n \geqslant 1$, which in turn would follow from

$$\nu(A) \leqslant \mathrm{e}^{-nh+c} \tag{4.4}$$

for all $A \in \xi \vee \cdots \vee T^{-(n-1)}\xi$. This leads to the following formulation of positive entropy for almost all ergodic components.

Corollary 4.2. *Let T be a measure-preserving transformation on the probability space $(X, \mathscr{B}, \nu)$, and let ξ be a countable partition of X with $H_\nu(\xi) < \infty$. If there exist constants $h > 0$ and $c > 0$ such that*

$$\nu(A) \leqslant \mathrm{e}^{-nh+c}$$

for all $A \in \xi \vee \cdots \vee T^{-(n-1)}\xi$ and all $n \geqslant 0$, then almost every ergodic component of ν has positive entropy. Moreover, $h_{\nu_\theta}(T, \xi) \geqslant h$ for ρ-almost every $\theta \in \Xi$ in the ergodic decomposition (4.2).

Proof. We will only prove the positivity, and leave it to the reader to modify Proposition 4.1 in order to obtain the final claim of the corollary (which, strictly speaking, is not needed later). Let $B \subseteq X$ be a measurable T-invariant set with positive measure. Then our assumption implies that

$$\frac{1}{\nu(B)}\nu|_B(A) \leqslant \frac{1}{\nu(B)}\mathrm{e}^{-nh+c}$$

for all $A \in \xi \vee \cdots \vee T^{-(n-1)}\xi$ and $n \geqslant 1$. As discussed above, this shows that

$$h_{(\frac{1}{\nu(B)})\nu|_B}(T, \xi) \geqslant h,$$

which by Proposition 4.1 gives positive entropy for almost all ergodic components as required. □

The assumed bound on $\nu(A)$ above may be called *effective positive entropy*, and is a much stronger assumption than simple positivity. In fact, the effective version of positivity of entropy has been shown for the arithmetic quantum limits by Bourgain and Lindenstrauss [2]. In fact, they show the assumption of Corollary 4.2 with $h = \frac{1}{9}$, when T is the time-one map of the geodesic flow on $\Gamma\backslash \mathrm{SL}_2(\mathbb{R})$, where Γ is a congruence lattice over $\mathbb{Q}$. Notice that $h_{m_X}(T) = 1$.

We will simplify their argument, reaching a weaker conclusion. This will nonetheless give positivity of entropy for almost all ergodic components by using the fundamental Shannon–McMillan–Breiman theorem (see [11] for a detailed discussion and references; the result was proved in increasingly general settings by Shannon [26], McMillan [20], Carleson [4], and Breiman [3]). The Shannon–McMillan–Breiman theorem may be viewed as the pointwise ergodic theorem of entropy theory. For any countable partition η of a probability space $(X, \mathscr{B}, \mu)$, we denote by $[x]_\eta$ the element of η containing x (this will also be called the atom of η containing x).

Theorem 4.3. *Let T be a measure-preserving transformation on $(X, \mathscr{B}, \mu)$, and let ξ be a countable partition of X with $H_\mu(\xi) < \infty$. Then*

$$-\frac{1}{n} \log \mu \left([x]_{\bigvee_{i=0}^{n-1} T^{-i}\xi} \right) \longrightarrow h_{\mu_x^{\mathscr{E}}}(T, \xi) \tag{4.5}$$

as $n \to \infty$ for μ-almost every x.

This result makes it easier to show positive entropy for almost all ergodic components, as one does not have to show (4.4) for all $n \geqslant 0$. Instead it is sufficient to show (4.4) for infinitely many n – since the convergence in (4.5) is guaranteed, then a similar bound must also hold for all other n.

4.2 An Abelian Case

We will explain the method of Bourgain and Lindenstrauss [2] for establishing positive entropy in the following much simpler (in comparison, a toy) situation. This simple exposition of the idea is taken from work of Einsiedler and Fish [7].

Let $S \subseteq \mathbb{N} \smallsetminus \{0\}$ be a multiplicative semigroup. If $S \subseteq a^{\mathbb{N}}$ for some $a > 1$, then S is called *lacunary* and in this case there is a multitude of S-invariant probability measures (under the natural action of multiplication by elements of S on $\mathbb{T} = \mathbb{R}/\mathbb{Z}$) and S-invariant closed infinite subsets of $X = \mathbb{R}/\mathbb{Z}$ (a set $A \subseteq \mathbb{T}$ is S-invariant if $sA \subseteq A$ for all $s \in S$). On the other hand, if $S \not\subseteq a^{\mathbb{N}}$ for any $a > 1$, then S is called *non-lacunary* and in this case Furstenberg [13] showed that there are only very few different S-invariant closed subsets.

The simplest non-lacunary semigroup in $\mathbb{N}$ is $S = 2^{\mathbb{N}}3^{\mathbb{N}}$, and the next result raises the problem of classifying the S-invariant probability measures.

Theorem 4.4 (Furstenberg). *If S is non-lacunary, then an S-invariant closed subset $A \subseteq \mathbb{T}$ is either finite (consisting of rational points) or equal to $\mathbb{T}$.*

Furstenberg also asked the question of whether the ergodic theoretic analog of Theorem 4.4 holds as well: that is, must a probability measure that is invariant and ergodic* for a non-lacunary semigroup be of finite support or be Lebesgue measure? It is straightforward† to see that for $S = \mathbb{N}_{>0}$ there are only two possible S-invariant and ergodic probability measures, namely the Dirac measure δ_0 and the Lebesgue measure $m_{\mathbb{T}}$. However, for general non-lacunary semigroups the conjecture is still open. There have been some partial results towards the conjecture, the strongest of which is due to Rudolph [25] and Johnson [15].

* A measure μ on $\mathbb{T}$ is ergodic for the action of S if $\mu(B \triangle s^{-1}B) = 0$ for all $s \in S$ implies that $\mu(B) \in \{0, 1\}$. If $S = a^{\mathbb{N}}$ is generated by a single element a, then a measure is ergodic for S if and only if it is ergodic for multiplication by a.

† For example, by seeing that the invariance forces the Fourier transform of the invariant measure to coincide with the Fourier transform of Lebesgue measure if it is not the point measure at 0.

Theorem 4.5. *Let S be a non-lacunary semigroup in $\mathbb{N}_{>0}$, and let μ be an S-invariant ergodic probability measure on $\mathbb{T}$. If $h_\mu(s) > 0$ for some $s \in S$, then $\mu = m_{\mathbb{T}}$ is Lebesgue measure.*

One may ask whether it is possible to give stronger conditions on S which would allow a complete classification of S-invariant ergodic probability measures without the entropy hypothesis (as done above in the trivial case $S = \mathbb{N}_{>0}$). This can indeed be done for the following class of semigroups.

Definition 4.6. *A semigroup $S \subseteq \mathbb{N}_{>0}$ has* polynomial density with exponent $\alpha > 0$ *if*

$$|S \cap [1, M]| \geqslant M^{\alpha} \tag{4.6}$$

for all sufficiently large M.

Theorem 4.7. *Let S be a semigroup of polynomial density. Then any S-invariant and ergodic probability measure on $\mathbb{T}$ is either supported on a finite set of rational points, or is the Lebesgue measure.*

We will prove Theorem 4.7 by establishing, under this stronger assumption on S, the entropy hypothesis in Theorem 4.5. This is similar to the proof of Theorem 1.2, where information about the Hecke operators will be used to prove the positive entropy hypothesis in Theorem 1.3.

Proof of Theorem 4.7. Fix some $s_0 \in S \setminus \{1\}$, write $T(x) = s_0 x$, and let μ be an S-invariant ergodic probability measure. Then we will show below that either μ is supported on a finite set, or $h_\mu(s_0) \geqslant \delta = \frac{\alpha \log s_0}{5}$. This will imply the theorem by the work of Rudolph and Johnson. Let

$$\xi = \{[0, \tfrac{1}{s_0}), [\tfrac{1}{s_0}), \dots, [\tfrac{s_0 - 1}{s_0}, 1)\}$$

be the partition corresponding to fixing the first digit in the s_0-ary expansion of real numbers $x \in [0, 1) \cong \mathbb{T}$. Notice that $\bigvee_{i=0}^{n-1} T^{-i}\xi$ is the partition corresponding to the first n digits in the s_0-ary expansion, and so comprises intervals of length $\frac{1}{s_0^n}$.

If $\mu(\mathbb{Q} \cap [0, 1)) > 0$ then by ergodicity we must have $\mu(\mathbb{Q} \cap [0, 1)) = 1$, and so μ must be supported on a finite set since each point in $\mathbb{Q} \cap [0, 1)$ has a finite orbit under S. Suppose therefore that μ is an S-invariant ergodic probability measure on $\mathbb{T}$ with $\mu(\mathbb{Q} \cap [0, 1)) = 0$.

Now assume that $x \in \mathbb{T}$ satisfies (4.5), with limit

$$h_{\mu_x^{\mathscr{E}}}(T, \xi) < \delta.$$

Then for large enough $n \geqslant n_0$ we would have

$$-\frac{1}{n} \log \mu \left([x]_{\bigvee_{i=0}^{n-1} T^{-i}\xi} \right) < \delta,$$

and so

$$\mu\left(B(x, s_0^{-n})\right) \geqslant \mathrm{e}^{-n\delta},$$

where

$$B(x, s_0^{-n}) = (x - s_0^{-n}, x + s_0^{-n}) \supseteq [x]_{\bigvee_{i=0}^{n-1} T^{-i}\xi}.$$

We define $M = M(n)$ by

$$M = 2\mathrm{e}^{n\delta/\alpha}, \tag{4.7}$$

and we may assume that (4.6) holds whenever $n \geqslant n_0$. Recall that $s \in S$ preserves the measure μ, so $\mu(\{x \mid sx \in B\}) = \mu(B)$ for any Borel set $B \subseteq \mathbb{T}$. This clearly implies that $\mu(sB) \geqslant \mu(B)$, and applying this to the interval $B = B(x, s_0^{-n})$ gives

$$\mu\left(sB(x, s_0^{-n})\right) \geqslant \mathrm{e}^{-n\delta},$$

and so

$$\sum_{s \in S \cap [1,M]} \mu\left(sB(x, s_0^{-n})\right) \geqslant 2M^{\alpha}\mathrm{e}^{-n\delta} > 1.$$

Therefore, there must be distinct elements $s, s' \in S \cap [1, M]$ with

$$sB(x, s_0^{-n}) \cap s'B(x, s_0^{-n}) \neq \varnothing.$$

Of course, this overlapping must be understood in $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, so if we identify $x \in \mathbb{T}$ with the corresponding element $x \in [0, 1) \subseteq \mathbb{R}$, then we have

$$s(x + v) = s'(x + v') + k,$$

where $|v|, |v'| < s_0^{-n}$, and $k \in \mathbb{Z}$. Thus

$$x = \frac{s'v' - sv}{s - s'} + \frac{k}{s - s'},$$

so by (4.7) and the definition of $\delta = \frac{\alpha \log s_0}{5}$,

$$\left|\frac{s'v' - sv}{s - s'}\right| \leqslant Ms_0^{-n} = M\mathrm{e}^{-n \log s_0} \ll MM^{-\frac{\alpha}{\delta}\log s_0} = M^{1-5} = M^{-4}$$

and

$$|s - s'| < M.$$

This already should be surprising – the real number x (about which we only assumed (4.5) with $h_{\mu_x^{\mathscr{E}}}(T, \xi) < \delta$) has a rational approximation of the shape

$$\left|x - \frac{p_1}{q_1}\right| \ll M^{-4}$$

with denominator $q_1 < M$.

We now apply the argument above to $2n$ in place of n; the quantity $M = M(n)$ is then squared, and we find a rational approximation

$$\left| x - \frac{p_2}{q_2} \right| \ll M^{-8}$$

with $q_2 < M^2$. Together these two approximations give

$$\left| \frac{p_1}{q_1} - \frac{p_2}{q_2} \right| \ll M^{-4} + M^{-8} \ll M^{-4}. \tag{4.8}$$

On the other hand, the left-hand side of (4.8) is either zero, or has

$$\left| \frac{p_1}{q_1} - \frac{p_2}{q_2} \right| \geqslant \frac{1}{q_1 q_2} \geqslant M^{-3}.$$

Assuming n_0 is large enough (which we may), the latter gives a contradiction, so we must have $\frac{p_1}{q_1} = \frac{p_2}{q_2}$, and therefore

$$\left| x - \frac{p_1}{q_1} \right| < M^{-8}.$$

Repeating this observation for $2n$, for $4n$, for $8n$ and so on, the same argument gives better and better approximation of x by $\frac{p_1}{q_1}$, and so we must have

$$x = \frac{p_1}{q_1} \in \mathbb{Q} \cap [0, 1).$$

We have shown that $h_{\mu_x^{\mathcal{E}}}(T, \xi) < \delta$ implies that $x \in \mathbb{Q} \cap [0, 1)$. Since

$$\mu(\mathbb{Q} \cap [0, 1)) = 0$$

we deduce that almost every ergodic component of μ has $h_{\mu_x^{\mathcal{E}}}(T, \xi) \geqslant \delta > 0$ as required. □

4.3 The Partition ξ

As discussed in Section 4.1 in general, and in Section 4.2 for a particular abelian example, we would like to find a partition ξ (either finite, or at least with finite entropy), for which the measure of elements of the common refinement

$$\bigvee_{i=-n}^{n} T^{-i} \xi$$

can be estimated. This then leads to positive entropy of almost all ergodic components for the time-one map T of the geodesic flow.

4.3.1 The Compact Case

If $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ is compact, then it is quite straightforward to construct a finite partition ξ which captures all the entropy in the sense that

$$h_\mu(T) = h_\mu(T, \xi)$$

for any T-invariant probability measure μ on X. We will not need this equality, only properties of the partition ξ.

Recall that if $a = \begin{pmatrix} \mathrm{e}^{1/2} & \\ & \mathrm{e}^{-1/2} \end{pmatrix}$, then

$$T(x) = xa$$

for $x \in X$. Also recall (for example, from [12]) that, since X is compact, there exists a uniform injectivity radius $r > 0$ for which

$$B_r^{\mathrm{SL}_2(\mathbb{R})}(I) \ni h \longmapsto xh \in B_r^X(x)$$

is an isometry (and, in particular, is a bijection) for any $x \in X$. We may assume that $r > 0$ is small enough to ensure that $B_r^{\mathrm{SL}_2(\mathbb{R})}(I)$ is the injective image of some small open neighborhood of $0 \in \mathfrak{sl}_2(\mathbb{R})$ under the exponential map. Let $\delta \in (0, r)$ be such that

$$aB_\delta^{\mathrm{SL}_2(\mathbb{R})}(I)a^{-1} \subseteq B_r^{\mathrm{SL}_2(\mathbb{R})}(I) \tag{4.9}$$

and

$$a^{-1}B_\delta^{\mathrm{SL}_2(\mathbb{R})}(I)a \subseteq B_r^{\mathrm{SL}_2(\mathbb{R})}(I). \tag{4.10}$$

Lemma 4.8. *Any partition ξ of X consisting of sets with diameter less than δ has the property that any element*

$$A \in \bigvee_{i=-n}^{n} T^{-i}\xi \tag{4.11}$$

is contained in a Bowen n-ball,

$$A \subseteq x\left(\bigcap_{i=-n}^{n} a^i B_\delta^{\mathrm{SL}_2(\mathbb{R})} a^{-i}\right)$$

for any $x \in A$.

In some sense the particular shape of the sets in the refinement in (4.11) is likely to be quite complicated, as it depends on a non-canonical choice of the partition ξ. The *Bowen n-ball*

$$x\left(\bigcap_{i=-n}^{n} a^i B_\delta^{\mathrm{SL}_2(\mathbb{R})} a^{-i}\right)$$

at x, on the other hand, by definition consists of all displacements xh of x by elements h small enough to have

$$\mathsf{d}(ha^i, Ia^i) = \mathsf{d}(a^{-i}ha^i, I) < \delta$$

for $i = -n, \dots, n$, and is therefore a relatively concretely described set. Notice that if $y = xh$ belongs to the Bowen n-ball, then

$$\mathsf{d}(ya^i, xa^i) < \delta$$

for $i = -n, \dots, n$. This does not imply that $y \in A$, but will nonetheless prove useful.

Proof of Lemma 4.8. Let ξ, A be as in the lemma, and let $x, y = xh \in A$. Here we may assume that $h \in B_\delta^{\mathrm{SL}_2(\mathbb{R})}(I)$ as x and y belong to the same element of the partition ξ. Applying the map T, we get

$$T(x), T(y) = xha = T(x)(a^{-1}ha).$$

Here $a^{-1}ha \in B_r^{\mathrm{SL}_2(\mathbb{R})}(I)$ by the choice of $\delta > 0$. Notice that $T(x), T(y)$ belong to the same partition element of ξ, and the diameter of this set is less than $\delta \leqslant r$. It follows that $T(y) = T(x)h_1$ for some $h_1 \in B_\delta^{\mathrm{SL}_2(\mathbb{R})}$. By choice of r, this shows that $h_1 = a^{-1}ha$.

Iterating this argument shows that

$$h \in \bigcap_{i=-n}^{n} a^i B_\delta^{\mathrm{SL}_2(\mathbb{R})} a^{-i},$$

as claimed in the lemma. □

4.3.2 The Non-compact Case

If $X = \Gamma \backslash \mathrm{SL}_2(\mathbb{R})$ is non-compact but of finite volume, then there exists a compact set $\Omega \subseteq X$ such that every A-orbit intersects Ω non-trivially. This may be seen geometrically for $\mathrm{SL}_2(\mathbb{Z}) \backslash \mathrm{SL}_2(\mathbb{R})$ since A-orbits can be represented in $\mathrm{SL}_2(\mathbb{Z}) \backslash \mathrm{T}^1\mathbb{H}$ by vertical lines or half-circles intersecting the real axis normally, and both types of orbit intersect the line $\Im(z) = 1$, which is mapped to a compact set Ω in $\mathrm{SL}_2(\mathbb{Z}) \backslash \mathrm{SL}_2(\mathbb{R})$. More generally, this follows from the structure of the cusps of finite volume quotients $\Gamma \backslash \mathrm{SL}_2(\mathbb{R})$ (see, for example, [12, Lem. 11.29]). Enlarging Ω if necessary to a bigger compact subset, we may also assume that the time-one map T of the geodesic flow has the property that for any $x \in X$ there exists at least one $n \in \mathbb{Z}$ with $T^n x \in \Omega$. Let r be the injectivity radius on Ω.

Proposition 4.9. *Let $X = \Gamma \backslash \mathrm{SL}_2(\mathbb{R})$ be a non-compact, finite volume, quotient. Then there exists a countable partition ξ of X containing $X \smallsetminus \Omega$ as one of its element, such that $x \in \Omega$, and*

$$x \in A \in \bigvee_{i=-n}^{n} T^{-i}\xi$$

implies that A is contained in the Bowen n-ball

$$A \subseteq x \bigcap_{i=-n}^{n} a^{-i} B_\delta^{\mathrm{SL}_2(\mathbb{R})} a^i,$$

where we choose δ small enough to satisfy (4.9) *and* (4.10)*. Furthermore, ξ has finite entropy for any T-invariant probability measure on X.*

We refer to [6] for a proof, and note that this construction is very similar to a particular portion of the proof [19] by Margulis and Tomanov of Ratner's measure classification theorem (see [21] and [22]).

4.4 Proving Positive Entropy

In Section 4.2 we assumed invariance under a large multiplicative subset of $\mathbb{N}$, but then only used the invariance in a very weak way to prove that images of certain intervals are big in measure, forcing there to be overlaps. Our discussion of the Hecke-operators in Chapter 3 does not give any invariance properties under $\mathrm{PGL}_2(\mathbb{Q}_p)$, but we will see that it can give the result that certain images of Bowen n-balls are big in measure, which will again force overlaps.

4.4.1 Mass of Distance $\leqslant 2$ Neighbors

We start with the following corollary to Lemma 3.20.

Corollary 4.10. *If S_p (and S_{p^2}) denote the operators that sum functions on a $(p+1)$-regular tree T over the neighbors (and over the distance-two neighbors, respectively), then for any eigenfunction of S_p on T we have either*

$$|\lambda_p| \gg \sqrt{p}$$

or

$$|\lambda_{p^2}| \gg p,$$

where λ_{p^i} is the eigenvalue for S_{p^i} for $i = 1, 2$.

Proof. Notice that

$$\begin{aligned} S_p^2(f)(v) &= \sum_{w \sim v} S_p(f)(w) \\ &= \sum_{w \sim v} \sum_{w' \sim w} f(w') \\ &= \sum_{w' \sim_2 v} f(w') + (p+1) f(v), \end{aligned}$$

so $S_p^2 = S_{p^2} + (p+1)I$ and therefore

$$\lambda_p^2 = \lambda_{p^2} + p + 1,$$

giving the corollary. $\square$

Applying the Cauchy–Schwartz inequality (as in the proof of Theorem 3.12) gives another corollary.

Corollary 4.11. *If ϕ is an eigenfunction of S_p on the $(p+1)$-regular tree with real eigenvalue, then*

$$\sum_{w\sim v} |\phi(w)|^2 + \sum_{w\sim_2 v} |\phi(w)|^2 \gg |\phi(v)|^2$$

for any vertex v in T.

PROOF. Suppose that $|\lambda_p| \gg \sqrt{p}$. Then by Cauchy–Schwartz,

$$\sqrt{p}|\phi(v)| \ll |\lambda_p \phi(v)| = |\sum_{w\sim v} \phi(w)| \leqslant \left(\sum_{w\sim v} |\phi(w)|^2\right)^{1/2} (p+1)^{1/2},$$

which gives the corollary in this case. If $|\lambda_{p^2}| \gg p$ the same argument applies. $\square$

For arithmetic quantum limits this gives the following.

Corollary 4.12. *Let $X = \mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$, and suppose that $\phi_i \in L^2_{m_X}$ satisfies $\|\phi_i\|_2 = 1$ for all $i \geqslant 1$, and μ is a weak*-limit of the measures defined by $|\phi_i^2| \,\mathrm{d}m_X$. Then for any non-negative measurable function f we have*

$$\int \big(S_p(f) + S_{p^2}(f)\big) \,\mathrm{d}\mu \gg \int f \,\mathrm{d}\mu$$

*for any prime p, and so**

$$\int \Big(\sum_{p\leqslant Q} S_p(f) + S_{p^2}(f) \Big) \,\mathrm{d}\mu \gg Q^{1/2} \int f \,\mathrm{d}\mu,$$

where the sum is taken over all primes $p \leqslant Q$, for any Q.

PROOF. The proof is similar to the proof of Theorem 3.12. If ϕ is a bounded eigenfunction in $L^2(X)$ for S_p, then

$$\begin{aligned}\int \big(S_p(f) + S_{p^2}(f)\big) |\phi|^2 \,\mathrm{d}m_X &= \big\langle S_p(f) + S_{p^2}(f), |\phi|^2 \big\rangle \\ &= \big\langle f, S_p(|\phi|^2) + S_{p^2}(|\phi|^2) \big\rangle \\ &\gg \big\langle f, |\phi|^2 \big\rangle\end{aligned}$$

* The exponent $\frac{1}{2}$ of Q on the right-hand side could also be replaced by $1 - \mathrm{o}(1)$.

by Corollary 4.11. If $f \in C_c(X)$, then this implies that

$$\int \left(S_p(f) + S_{p^2}(f)\right) \mathrm{d}\mu \gg \int f \, \mathrm{d}\mu.$$

Using dominated and monotone convergence, this then extends to any measurable non-negative f as in the proof of Theorem 3.12.

The final claim follows from the argument above and the prime number theorem, which implies in particular that there are $\gg Q^{1/2}$ primes $p \leqslant Q$. □

If $f = \mathbf{1}_B$ is the characteristic function of a set $B \subseteq X$, then the result above implies that either the measure of B is small (in comparison to Q), or that

$$\sum_{p \leqslant Q} S_p(f) + S_{p^2}(f) > 1$$

at some points of X, which roughly speaking implies overlaps of the Hecke images of B.

4.4.2 The Volume of Bowen n-balls

Theorem 4.13. *If μ is a micro-local lift of an arithmetic quantum limit on $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$, then the entropy of the time-one map T of the geodesic flow is positive on almost every ergodic component.*

By our discussion of entropy of ergodic components in Proposition 4.1 and the choice of the partition in Proposition 4.9, it follows that in order to prove Theorem 4.13 it is sufficient to show that for almost every $x \in \Omega$ there are infinitely many $n \geqslant 1$ for which

$$\mu\left(x \bigcap_{i=-n}^{n} a^i B_\delta^{\mathrm{SL}_2(\mathbb{R})} a^{-i}\right) \leqslant \mathrm{e}^{-nh} \tag{4.12}$$

for some positive constant h. Indeed, if we have shown this, then for almost every $x \in X$ there is some $T^n(x) \in \Omega$ for which $h_{\mu_x^{\mathscr{E}}}(T,\xi) = h_{\mu_{T^n x}^{\mathscr{E}}}(T,\xi) \geqslant h > 0$ by Theorem 4.3.

Proof of Theorem 4.13. We fix some $x = \mathrm{PGL}_2(\mathbb{Z})g \in \Omega$ and set

$$B_n = x \bigcap_{i=-n}^{n} a^i B_\delta^{\mathrm{SL}_2(\mathbb{R})} a^{-i}.$$

Fix some $h < \frac{1}{16}$, and assume that

$$\mu(B_n) \geqslant \mathrm{e}^{-hn} \tag{4.13}$$

for all $n \geqslant n_0$. Our goal is to show that this forces x to belong to some concrete null set. Let $c > 0$ be the implicit constant appearing in Corollary 4.12, so that

$$\int \Big(\sum_{p \leqslant Q} S_p(f) + S_{p^2}(f) \Big) \, \mathrm{d}\mu \geqslant cQ^{1/2} \int f \, \mathrm{d}\mu, \tag{4.14}$$

for all $f \geqslant 0$. We define $Q(n) = c^{-2}\mathrm{e}^{2hn}$. Combining (4.14) with $f = \mathbf{1}_{B_n}$ with (4.13) gives

$$\int \Big(\sum_{p \leqslant Q(n)} S_p(\mathbf{1}_{B_n}) + S_{p^2}(\mathbf{1}_{B_n}) \Big) \, \mathrm{d}\mu \geqslant 1,$$

so that

$$\sum_{p \leqslant Q} \big(S_p(\mathbf{1}_{B_n}) + S_{p^2}(\mathbf{1}_{B_n}) \big) (y_n) > 1 \tag{4.15}$$

for some $y_n \in X$ and all $n \geqslant n_0$. We wish to rephrase this to give a more intrinsic property of the center x of the Bowen ball B_n.

Clearly (4.15) means that there are primes $p_1, p_2 \leqslant Q(n)$ (possibly the same prime) and $q_i \in \{p_i, p_i^2\}$ (again, possibly the same) with

$$S_{q_1}(\mathbf{1}_{B_n})(y_n) > 1$$

if $q_1 = q_2$, or

$$S_{q_1}(\mathbf{1}_{B_n})(y_n) + S_{q_2}(\mathbf{1}_{B_n})(y_n) > 1$$

if $q_1 \neq q_2$. We claim that, in either case, there exists $x_i = xb_i$ (which depends on $n \geqslant n_0$) with

$$b_i \in \bigcap_{i=-n}^{n} a^i B_\delta^{\mathrm{SL}_2(\mathbb{R})} a^{-i},$$

some $\gamma_i \in \mathrm{Mat}_{22}(\mathbb{Z})$ (not divisible by p_i) with $\det \gamma_i = q_i$,

$$\mathrm{PGL}_2(\mathbb{Z})\gamma_1 \neq \mathrm{PGL}(\mathbb{Z})\gamma_2, \tag{4.16}$$

finally $y_n = \mathrm{PGL}_2(\mathbb{Z})\gamma_1 g b_1 = \mathrm{PGL}_2(\mathbb{Z})\gamma_2 g b_2$ and

$$\gamma_1 g b_1 = \gamma_2 g b_2 \tag{4.17}$$

in $\mathrm{PGL}_2(\mathbb{R})$. Indeed, by Section 3.2.2, the fact that $S_{q_i}(\mathbf{1}_{B_n})(y_n) \geqslant 1$ means that $x_i = xb_i$ and $\gamma_i \in \mathrm{Mat}_{22}(\mathbb{Z})$ not divisible by p_i exists as claimed. In the case of $q_1 = q_2$, (4.16) is also guaranteed by the discussion in Section 3.2.2. Assume therefore that $q_1 \neq q_2$ but $\mathrm{PGL}_2(\mathbb{Z})\gamma_1 = \mathrm{PGL}_2(\mathbb{Z})\gamma_2$, or equivalently $\gamma_1 = c\eta\gamma_2$ for some $c \in \mathbb{Q}^\times$ and $\eta \in \mathrm{PGL}_2(\mathbb{Z})$. Taking determinants, this implies that

$$q_1 = \det \gamma_1 = c^2 \det \gamma_2 = c^2 q_2$$

and so $c = \frac{p_1}{p_2}$ and $q_i = p_2^2$ for $i = 1, 2$. As we assume for the moment that $q_1 \neq q_2$, $\gamma_1 = c\eta\gamma_2$ and $\gamma_2 \in \mathrm{Mat}_{22}(\mathbb{Z})$ implies that γ_i is divisible by p_i for $i = 1, 2$, which is a contradiction to the construction of γ_i. This proves the claim.

The claim may be further reformulated as follows: For every $n \geqslant n_0$ there exists some $\eta_n \in \mathrm{Mat}_{22}(\mathbb{Z}) \smallsetminus \mathbb{Q}\,\mathrm{GL}_2(\mathbb{Z})$ with $\det(\eta_n) \leqslant Q(n)^4$ such that

$$\frac{1}{\det(\eta_n)^{1/2}} g^{-1}\eta_n g \in \bigcap_{i=-n}^{n} a^i B_{2\delta}^{\mathrm{SL}_2(\mathbb{R})} a^{-1}. \tag{4.18}$$

Indeed, for any $n \geqslant n_0$ we can define

$$\eta_n = q_2 \gamma_2^{-1} \gamma_1$$

with determinant $q_1 q_2 \leqslant Q(n)^4$, and then (4.17) implies that

$$g^{-1}\eta_n g = b_2 b_1$$

in $\mathrm{PGL}_2(\mathbb{R})$, which shows (4.18) in $\mathrm{SL}_2(\mathbb{R})$. This is the desired property of $x = \mathrm{PGL}_2(\mathbb{Z})g$.

We now proceed to analyze the set on the right-hand side of (4.18). Assuming (as we may) that δ is sufficiently small, $B_{2\delta}^{\mathrm{SL}_2(\mathbb{R})}$ is in the injective image of a small neighborhood of $0 \in \mathfrak{sl}_2(\mathbb{R})$ under the exponential map. Note that $a\exp(v)a^{-1} = \exp(\mathrm{Ad}_a(v))$ and that Ad_a has the eigenvectors $\mathcal{H}, \mathcal{U}^-$, and $\mathcal{U}^+$ with eigenvalues 1, e^{-1} and e respectively (see Proposition 2.3). From this one sees quickly (much as in the argument in Section 4.3.1) that

$$\bigcap_{i=-n}^{n} a^i B_{2\delta}^{\mathrm{SL}_2(\mathbb{R})} a^{-1} \subseteq \exp\left([-\kappa,\kappa]\mathcal{H} + [-\kappa\mathrm{e}^{-n}, \kappa\mathrm{e}^{-n}]\mathcal{U}^- + [-\kappa\mathrm{e}^{-n}, \kappa\mathrm{e}^{-n}]\mathcal{U}^+\right) \tag{4.19}$$

for some $\kappa \ll \delta$. In other words, we should think of this set as an e^{-n}-small neighborhood of a bounded segment of the diagonal subgroup. Notice that any element of the set in (4.19) can also be written as $\exp(t\mathcal{H})\exp(v)$ with $|t| \ll \delta$ and $\|v\| \ll \delta\mathrm{e}^{-n}$.

Now recall the usual definition

$$[g,h] = g^{-1}h^{-1}gh$$

of the commutator of group elements g and h. Applying the commutator operation to η_n, η_{n+1} as in (4.18) for some $n \geqslant n_0$, we should get a very small element. Indeed

$$\begin{aligned} g^{-1}[\eta_n,\eta_{n+1}]g &= [\exp(t_n\mathcal{H})\exp(v_n), \exp(t_{n+1}\mathcal{H})\exp(v_{n+1})] \\ &= \exp(-v_n)\exp(-t_n\mathcal{H})\exp(-v_{n+1})\exp\left((-t_{n+1}+t_n)\mathcal{H}\right) \\ &\qquad \exp(v_n)\exp(t_{n+1})\exp(v_{n+1}) \\ &= \exp(-v_n)\left(\exp(-t_n\mathcal{H})\exp(-v_{n+1})\exp(t_n\mathcal{H})\right) \\ &\qquad \left(\exp(-t_{n+1}\mathcal{H})\exp(v_n)\exp(t_{n+1}\mathcal{H})\right)\exp(v_{n+1}) \end{aligned}$$

is the product of four terms each of distance $\ll \mathrm{e}^{-n}$ from the identity. As $\Omega \subseteq X$ is some fixed compact set, we may assume that $g \in \mathrm{SL}_2(\mathbb{R})$ also belongs to some fixed compact set, which shows that $[\eta_n, \eta_{n+1}]$ is of distance $\ll \mathrm{e}^{-n}$ from the identity.

On the other hand,

$$[\eta_n, \eta_{n+1}] \in \mathrm{Mat}_{22}(\mathbb{Q}),$$

has numerator $\leqslant Q(n)^8$, and has determinant 1. This implies that either $[\eta_n, \eta_{n+1}]$ is the identity, or the distance to the identity is $\gg \frac{1}{Q(n)^8}$. In the latter case,

$$\mathrm{e}^{-16hn} \ll \frac{1}{Q(n)^8} \ll \mathrm{e}^{-n}.$$

However, as $h < \frac{1}{16}$ by our choice of h, this inequality only holds for finitely many values of n, say for $n < n_0$. It follows that η_n and η_{n+1} commute for $n \geqslant n_0$.

Notice that (up to conjugation) $\mathrm{SL}_2(\mathbb{R})$ contains three types of elements close to the identity:

- elements of A (defined in (2.2) with y small);
- elements of $\mathrm{SO}(2)(\mathbb{R})$ (as defined in (2.1), with θ small); or
- elements of the unipotent subgroup $U = \left\{ u_s = \begin{pmatrix} 1 & s \\ & 1 \end{pmatrix} \mid s \in \mathbb{R} \right\}$ (with $|s|$ small).

In each case, the commutator of any of its non-central elements coincides with the subgroup. Also notice that the η_n are non-central since they are close to the identity but non-trivial. Thus the centralizer of η_n is the equal to the centralizer of η_{n+1}, respectively η_{n+2} and so on, and they all belong to the same centralizer subgroup.

We now consider the three cases of $A, \mathrm{SO}(2)$, and U, where the latter two are easier to rule out using purely geometric arguments. The union of the set of $\mathrm{SL}_2(\mathbb{R})$-conjugates of $\mathrm{SO}(2)(\mathbb{R})$ forms a compact set, *uniformly transverse* to A. One way to make this precise is to describe the set of Lie algebras that are conjugated to $\mathrm{Lie}\, U = \mathbb{R}\mathcal{U}^+$ or to $\mathrm{Lie}\, \mathrm{SO}(2)(\mathbb{R}) = \mathbb{R}\mathcal{W}$, as illustrated in Figure 4.1.

Assuming this description, we can rule out the cases where η_n is conjugated to an element of $\mathrm{SO}(2)(\mathbb{R})$ or of U quickly: In this case (4.18), (4.19) and Figure 4.1 together show that $\frac{1}{\det(\eta_n)^{1/2}} g^{-1} \eta_n g$ has distance $\ll \mathrm{e}^{-n}$ from the identity. However, $\det(\eta_n)^{1/2} \leqslant Q(n)$ and $\eta_n \in \mathrm{Mat}_{22}(\mathbb{Z})$. As η_n is not of the form $\begin{pmatrix} a & \\ & a \end{pmatrix}$, dividing by $\det(\eta_n)^{1/2}$ shows that the distance from $\frac{1}{\det(\eta_n)^{1/2}} g^{-1} \eta_n g$ to the identity is $\gg \frac{1}{Q(n)}$. These inequalities together are similar to, but much stronger than, what we knew before about $[\eta_n, \eta_{n+1}]$. Therefore, this gives a contradiction to the assumption that η_n is conjugated to an element of $\mathrm{SO}(2)(\mathbb{R}) \cup U$.

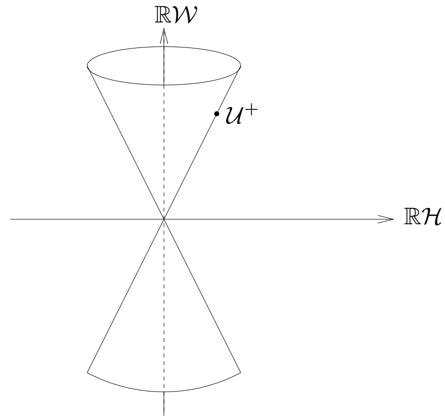

Fig. 4.1. The closed filled cone consists of all elements of $\mathfrak{sl}_2(\mathbb{R})$ that are conjugated to an element of $\mathbb{R}\mathcal{W} \cup \mathbb{R}\mathcal{U}^+$.

Instead of justifying the geometrical picture in Figure 4.1, we now give an alternative more formal argument leading to (almost) the same estimates, and the same conclusion. Assume once again that

$$\frac{1}{\det(\eta_n)^{1/2}} g^{-1}\eta_n g = \exp(t_n\mathcal{H})\exp(v_n)$$

with $\|v_n\| \ll \mathrm{e}^{-n}$ as in the discussion after (4.19) is conjugated to an element of $\mathrm{SO}(2)(\mathbb{R}) \cup U$. This implies that its trace must belong to $[-2, 2]$. However,

$$\exp(v_n) = I + w$$

for some $w = w(n) \in \mathrm{Mat}_{22}(\mathbb{R})$, and so

$$\mathrm{tr}\left[\begin{pmatrix} \mathrm{e}^{t_n/2} & \\ & \mathrm{e}^{-t_n/2}\end{pmatrix}\begin{pmatrix} 1+w_{11} & w_{12} \\ w_{21} & 1+w_{22}\end{pmatrix}\right] = \left(\mathrm{e}^{t_n/2} + \mathrm{e}^{-t_n/2}\right) + \mathrm{O}(\mathrm{e}^{-n}).$$

Thus

$$\mathrm{e}^{t_n/2} + \mathrm{e}^{-t_n/2} = 2\cosh(t_n/2) \leqslant 2 + \mathrm{O}(\mathrm{e}^{-n}),$$

which implies that $|t_n| \ll \mathrm{e}^{-n/2}$. Therefore, the distance of $\frac{1}{\det(\eta_n)^{1/2}} g^{-1}\eta_n g$ to the identity is $\ll \mathrm{e}^{-n/2}$. As before, this leads to a contradiction and leaves only one case open.

Assume now that $\eta_n, \eta_{n+1}, \dots$ are all conjugated to A. We claim that in fact

$$\frac{1}{\det(\eta_n)^{1/2}} g^{-1}\eta_n g \in A$$

for all $n \geqslant n_0$. Fix some $m \geqslant n_0$. Then, by (4.19),

$$\frac{1}{\det(\eta_n)^{1/2}} g^{-1}\eta_n g = \exp(t_n\mathcal{H} + w_n),$$

where $|t_n| \ll \delta$ and $\|w_n\| \ll \mathrm{e}^{-n}$. We must have $|t_n| \geqslant \mathrm{e}^{-n/2}$, for otherwise we have the same estimates on t_n and w_n as in the case of $\mathrm{SO}(2)(\mathbb{R}) \cup U$. We now normalize the direction in the Lie algebra to have norm one and get

$$\frac{1}{\det(\eta_n)^{1/2}} g^{-1}\eta_n g = \exp(s_n\widetilde{w}_n)$$

with $|s_n| \gg \mathrm{e}^{-n/2}$ and $\|\widetilde{w}_n\| = 1$. Moreover, this implies that

$$\widetilde{w}_n = \frac{1}{s_n}(t_n\mathcal{H} + w_n)$$

satisfies $\|\widetilde{w_n} - \frac{t_n}{s_n}\mathcal{H}\| \ll \mathrm{e}^{-n/2}$. Also recall that the centralizer of any non-central element is one-dimensional. This implies that $\widetilde{w}_{n+1} = \pm\widetilde{w}_n \in \mathbb{R}\mathcal{H}$, and so proves the claim.

To summarize, we have shown that if $x = \Gamma g \in \Omega$ satisfies (4.13) for all $n \geqslant n_0$ and some fixed $h < \frac{1}{16}$, then there exists some element

$$\eta \in \mathrm{Mat}_{22}(\mathbb{Z}) \smallsetminus \mathbb{Q}\,\mathrm{GL}_2(\mathbb{Z})$$

with $g^{-1}\eta g \in A$. We consider now two cases, depending on whether or not η is diagonalizable over $\mathbb{Q}$.

Assume first that $g_{\mathbb{Q}} \in \mathrm{GL}_2(\mathbb{Q})$ also satisfies

$$g_{\mathbb{Q}}^{-1}\eta g_{\mathbb{Q}} \in A,$$

so that $gg_{\mathbb{Q}}^{-1}$ conjugates this element of A to A, and so normalizes A. This implies that $g \in \frac{1}{\det(g_{\mathbb{Q}})} g_{\mathbb{Q}} A$ or $g \in \frac{1}{\det(g_{\mathbb{Q}})} g_{\mathbb{Q}} \begin{pmatrix} & -1 \\ 1 & \end{pmatrix} A$, and so

$$x \in \mathrm{SL}_2(\mathbb{Z}) \tfrac{1}{\det(g_{\mathbb{Q}})} g_{\mathbb{Q}} A \cup \mathrm{SL}_2(\mathbb{Z}) \tfrac{1}{\det(g_{\mathbb{Q}})} g_{\mathbb{Q}} \begin{pmatrix} & -1 \\ +1 & \end{pmatrix} A$$

belongs to a fixed set associated to η. The orbit $\mathrm{SL}_2(\mathbb{Z}) \frac{1}{\det(g_{\mathbb{Q}})} A$, and similarly also $\mathrm{SL}_2(\mathbb{Z}) \frac{1}{\det(g_{\mathbb{Q}})} \begin{pmatrix} & -1 \\ +1 & \end{pmatrix} A$, is divergent, and so is a null set for any A-invariant probability measure μ on $X = \mathrm{SL}_2(\mathbb{Z}) \backslash \mathrm{SL}_2(\mathbb{R})$ by Poincaré recurrence (see [12, Th. 2.11]). To see that this orbit is divergent, notice that the endpoints $g_{\mathbb{Q}}(0)$ and $g_{\mathbb{Q}}(\infty)$ of the geodesic line in $\mathbb{H}$ associated to the orbit $\mathrm{SL}_2(\mathbb{Z}) g_{\mathbb{Q}} A$ belong to $\mathbb{Q}$.

Assume now that $\eta \in \mathrm{GL}_2(\mathbb{Q})$ is not diagonalizable over $\mathbb{Q}$ but is over $\mathbb{R}$. Given η, we may choose some $g_{\mathbb{R}} \in \mathrm{SL}_2(\mathbb{R})$ with

$$g_{\mathbb{R}}^{-1}\eta g_{\mathbb{R}} \in A.$$

As above, this implies that $g \in g_{\mathbb{R}} A$ or $g \in g_{\mathbb{R}} \begin{pmatrix} & -1 \\ +1 & \end{pmatrix} A$. In this case the field $K = \mathbb{Q}[\eta]$ is a real quadratic field over $\mathbb{Q}$ and the two orbits

$$\mathrm{SL}_2(\mathbb{Z}) g_{\mathbb{R}} A$$

and

$$\mathrm{SL}_2(\mathbb{Z}) g_{\mathbb{R}} \begin{pmatrix} & -1 \\ +1 & \end{pmatrix} A$$

are periodic A-orbits (since the order $K \cap \mathrm{Mat}_{22}(\mathbb{Z})$ in K contains a non-trivial unit). By Theorem 3.25, any arithmetic quantum limit gives zero mass to such an orbit.

We summarize again: To any $\mathbb{R}$-diagonalizable $\eta \in \mathrm{GL}_2(\mathbb{Q}) \smallsetminus \mathbb{Q}\,\mathrm{GL}_2(\mathbb{Z})$ we may associate a μ-null set N_η so that $N = \bigcup_\eta N_\eta$ has the property that any $x \in \Omega$ which satisfies (4.13) for all $n \geqslant n_0$ belongs to N. Therefore, for μ-almost every $x \in \Omega$ we have (4.12) for infinitely many n. As discussed there, this implies that almost every ergodic component has positive entropy. $\square$

References

1. N. Anantharaman and S. Nonnenmacher, 'Chaotic vibrations and strong scars', *Newsletter of the Europ. Math. Soc.* (2009), 19–24. http://www.ems-ph.org/journals/newsletter/pdf/2009-12-74.pdf.
2. J. Bourgain and E. Lindenstrauss, 'Entropy of quantum limits', *Comm. Math. Phys.* **233** (2003), no. 1, 153–171. http://dx.doi.org/10.1007/s00220-002-0770-8.
3. L. Breiman, 'The individual ergodic theorem of information theory', *Ann. Math. Statist.* **28** (1957), 809–811. http://projecteuclid.org/euclid.aoms/1177706899.
4. L. Carleson, 'Two remarks on the basic theorems of information theory', *Math. Scand.* **6** (1958), 175–180. http://www.mscand.dk/article.php?id=1512.
5. Y. Colin de Verdière, 'Ergodicité et fonctions propres du laplacien', *Comm. Math. Phys.* **102** (1985), no. 3, 497–502. http://projecteuclid.org/euclid.cmp/1104114465.
6. M. Einsiedler and E. Lindenstrauss, *Diagonal actions on locally homogeneous spaces*, in *Homogeneous flows, moduli spaces and arithmetic*, in *Clay Math. Proc.* (Amer. Math. Soc., Providence, RI, 2010).
7. M. Einsiedler and A. Fish, 'Rigidity of measures invariant under the action of a multiplicative semigroup of polynomial growth on $\mathbb{T}$', *Ergodic Theory Dynam. Systems* **30** (2010), 151–157. http://dx.doi.org/10.1017/S014338570800103X.
8. M. Einsiedler and A. Katok, 'Rigidity of measures—the high entropy case and non-commuting foliations', *Israel J. Math.* **148** (2005), 169–238. http://dx.doi.org/10.1007/BF02775436.
9. M. Einsiedler, A. Katok, and E. Lindenstrauss, 'Invariant measures and the set of exceptions to Littlewood's conjecture', *Ann. of Math. (2)* **164** (2006), no. 2, 513–560. http://dx.doi.org/10.4007/annals.2006.164.513.
10. M. Einsiedler, E. Lindenstrauss, P. Michel, and A. Venkatesh, *Distribution of periodic orbits on homogeneous spaces*. http://arxiv.org/abs/math.DS/0607815.
11. M. Einsiedler, E. Lindenstrauss, and T. Ward, *Entropy in ergodic theory and homogeneous dynamics*. http://www.mth.uea.ac.uk/entropy.
12. M. Einsiedler and T. Ward, *Ergodic theory with a view toward number theory*, in *Graduate Texts in Mathematics* (Springer-Verlag, New York, 2010). http://www.mth.uea.ac.uk/ergodic.

13. H. Furstenberg, 'Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation', *Math. Systems Theory* **1** (1967), 1–49. `http://dx.doi.org/10.1007/BF01692494`.
14. B. Host, 'Nombres normaux, entropie, translations', *Israel J. Math.* **91** (1995), no. 1-3, 419–428. `http://dx.doi.org/10.1007/BF02761660`.
15. A. S. A. Johnson, 'Measures on the circle invariant under multiplication by a nonlacunary subsemigroup of the integers', *Israel J. Math.* **77** (1992), no. 1-2, 211–240. `http://dx.doi.org/10.1007/BF02808018`.
16. A. Katok and R. J. Spatzier, 'Invariant measures for higher-rank hyperbolic abelian actions', *Ergodic Theory Dynam. Systems* **16** (1996), no. 4, 751–778. `http://dx.doi.org/10.1017/S0143385700009081`.
17. A. W. Knapp, *Lie groups beyond an introduction*, in *Progress in Mathematics* **140** (Birkhäuser Boston Inc., Boston, MA, second ed., 2002). `http://www.springer.com/birkhauser/mathematics/book/978-0-8176-4259-4`.
18. E. Lindenstrauss, 'Invariant measures and arithmetic quantum unique ergodicity', *Ann. of Math. (2)* **163** (2006), no. 1, 165–219. `http://dx.doi.org/10.4007/annals.2006.163.165`.
19. G. A. Margulis and G. M. Tomanov, 'Invariant measures for actions of unipotent groups over local fields on homogeneous spaces', *Invent. Math.* **116** (1994), no. 1-3, 347–392.
20. B. McMillan, 'The basic theorems of information theory', *Ann. Math. Statistics* **24** (1953), 196–219. `http://projecteuclid.org/euclid.aoms/1177729028`.
21. M. Ratner, 'On Raghunathan's measure conjecture', *Ann. of Math. (2)* **134** (1991), no. 3, 545–607. `http://www.jstor.org/stable/2944357`.
22. M. Ratner, 'Raghunathan's conjectures for Cartesian products of real and p-adic Lie groups', *Duke Math. J.* **77** (1995), no. 2, 275–382. `http://dx.doi.org/10.1215/S0012-7094-95-07710-2`.
23. M. Ratner, 'Horocycle flows, joinings and rigidity of products', *Ann. of Math. (2)* **118** (1983), no. 2, 277–313. `http://www.jstor.org/pss/2007030`.
24. Z. Rudnick and P. Sarnak, 'The behaviour of eigenstates of arithmetic hyperbolic manifolds', *Comm. Math. Phys.* **161** (1994), no. 1, 195–213. `http://projecteuclid.org/euclid.cmp/1104269797`.
25. D. J. Rudolph, '$\times 2$ and $\times 3$ invariant measures and entropy', *Ergodic Theory Dynam. Systems* **10** (1990), no. 2, 395–406. `http://dx.doi.org/10.1017/S0143385700005629`.
26. C. E. Shannon, 'A mathematical theory of communication', *Bell System Tech. J.* **27** (1948), 379–423, 623–656. `http://doi.acm.org/10.1145/584091.584093`.
27. A. I. Šnirel′man, 'Ergodic properties of eigenfunctions', *Uspehi Mat. Nauk* **29** (1974), no. 6(180), 181–182. `http://mi.mathnet.ru/eng/umn4463`.
28. K. Soundarajan, 'Quantum unique ergodicity for $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$', *Ann. of Math. (2)*. To appear.
29. P. Walters, *An introduction to ergodic theory*, in *Graduate Texts in Mathematics* **79** (Springer-Verlag, New York, 1982). `http://www.springer.com/math/analysis/book/978-0-387-95152-2`.
30. T. Watson, 'Rankin triple products and quantum chaos', *Ann. of Math. (2)*. `http://arxiv.org/abs/0810.0425`. to appear.
31. A. Weil, *Basic number theory*, in *Die Grundlehren der mathematischen Wissenschaften, Band 144* (Springer-Verlag New York, Inc., New York, 1967). `http://www.springer.com/math/numbers/book/978-3-540-58655-5`.

32. S. A. Wolpert, ‘The modulus of continuity for $\Gamma_0(m)\backslash\mathbb{H}$ semi-classical limits’, *Comm. Math. Phys.* **216** (2001), no. 2, 313–323. `http://dx.doi.org/10.1007/PL00005549`.
33. S. Zelditch, ‘Uniform distribution of eigenfunctions on compact hyperbolic surfaces’, *Duke Math. J.* **55** (1987), no. 4, 919–941. `http://dx.doi.org/10.1215/S0012-7094-87-05546-3`.

Index