Hyperbolic manifolds, discrete groups and ergodic theory

Course Notes
C. McMullen

February 18, 2017

Contents

1 Introduction .......................................................... 2
2 Ergodic theory .......................................................... 8
3 Geometry of the hyperbolic plane ..................................... 29
4 Arithmetic hyperbolic surfaces ....................................... 44
5 Dynamics on a hyperbolic surfaces ................................... 50
6 Dynamics on moduli space ............................................ 69
7 Orbit counting and equidistribution ................................... 73
8 Spectral theory .......................................................... 78
9 Mixing of unitary representations of $\text{SL}_n \mathbb{R}$ .................. 90
10 Amenability ............................................................ 95
11 The Laplacian .......................................................... 98
12 All unitary representations of $\text{PSL}_2(\mathbb{R})$ ...................... 107
13 Kazhdan’s property T .................................................. 109
14 Ergodic theory at infinity of hyperbolic manifolds ................. 116
15 Lattices: Dimension 1 .................................................. 119
16 Dimension 2 ............................................................ 121
17 Lattices, norms and totally real fields. ............................. 132
18 Dimension 3 ............................................................ 136
19 Dimension 4, 5, 6 ....................................................... 143
20 Higher rank dynamics on the circle ................................. 144
21 The discriminant–regulator paradox ................................. 147
A Appendix: The spectral theorem ..................................... 154
1 Introduction

These notes address topics in geometry and dynamics, and make contact with some related results in number theory and Lie groups as well.

The basic setting for dynamics is a bijective map $T : X \to X$. One wishes to study the behavior of orbits $x, T(x), T^2(x), \ldots$ from a topological or measurable perspective. In the former cases $T$ is required to be at least a homeomorphism, and in the latter case one usually requires that $T$ preserves a Borel probability measure $\mu$ on $X$. This means:

$$\mu(T^{-1}(E)) = \mu(E) \tag{1.1}$$

for any measurable set $E$, or equivalently

$$\int f \, d\mu = \int f \circ T \, d\mu$$

for any reasonable function $f$ on $X$.

More generally one can consider a flow $T : \mathbb{R} \times X \to X$ or the action of a topological group $T : G \times X \to X$. In topological dynamics, one requires that $T$ is a continuous map in the product topology; in measurable dynamics, that $T$ is a measurable map.

One can also consider the dynamics of an endomorphism $T$. Equation (1.1) defines the notion of measure preservation in this case as well.

**Ergodicity.** Given a measure-preserving map $T : X \to X$, a basic issue is whether or not this map is ‘irreducible’. More precisely, we say $T$ is ergodic if whenever $X$ is split into a disjoint union of measurable, $T^{-1}$-invariant sets,

$$X = A \sqcup B,$$

either $\mu(A) = 0$ or $\mu(B) = 0$.

**Examples of measure-preserving dynamical systems.**


2. The continued fraction map $x \mapsto \{1/x\}$ on $[0,1]$ preserves the Gauss measure $m = dx/(1 + x)$ (with total mass $\log 2$). This (together with ergodicity) allows us to describe the behavior of the continued fraction of a typical real number.
3. Stationary stochastic processes. A typical example is a sequence of independent, equally distributed random variables \(X_1, X_2, \ldots\). The model space is then \(X = \mathbb{R}^\mathbb{Z}_+\) with the product measure \(m = \prod m_i\), with all the factors \(m_i\) the same probability measure on \(\mathbb{R}\). The relevant dynamics then comes from the shift map,

\[
\sigma(X_1, X_2, \ldots) = (X_2, X_3, \ldots),
\]

which is measure–preserving.

In this setting the Kolmogorov 0/1 law states than any tail event \(E\) has probability zero or one. A tail event is one which is independent of \(X_i\) for each individual \(i\); for example, the event “\(X_i > 0\) for infinitely many \(i\)” is a tail event. The zero–one law is a manifestation of ergodicity of the system \((X, \sigma)\).

4. The Henon map. The Jacobian determinant \(J(x, y) = \det DT(x, y)\) of any polynomial automorphism \(T : \mathbb{C}^2 \to \mathbb{C}^2\), is constant. Indeed, \(J(x, y)\) is itself a polynomial, which will vanish at some point unless it is constant, and vanishing would contradict invertibility.

(The Jacobian conjecture asserts the converse: a polynomial map \(T : \mathbb{C}^n \to \mathbb{C}^n\) with constant Jacobian has a polynomial inverse. It is know for polynomials of degree 2 in \(n\) variables, and for polynomials of degree \(\leq 100\) in 2 variables.)

If \(T\) is defined over \(\mathbb{R}\), and \(\det DT = 1\), then \(T\) gives an area–preserving map of the plane.

In the Henon family \(H(x, y) = (x^2 + c – ay, x)\), measure is preserved when \(a = 1\).

5. Blaschke products on the circle, \(B(z) = z \prod (z-a_i)/(1-\overline{a_i}z)\), preserve \(m = |dz|/2\pi\). For the proof, observe that the measure of a set \(E \subset S^1\) is given by \(u_E(0)\), where \(u_E(z)\) is the harmonic extension of the indicator function of \(E\) to the disk. Since \(B(0) = 0\), we have

\[
m(B^{-1})(E) = u_{B^{-1}E}(0) = u_E \circ B(0) = u_E(0) = m(E).
\]

6. The full shift \(\Sigma_d = (\mathbb{Z}/d)\mathbb{Z}\).

7. The baker’s transformation. This is the discontinuous map given in base two by

\[
f(0.x_1x_2\ldots, 0.y_1y_2\ldots) = (0.x_2x_3\ldots, 0.x_1y_1y_2\ldots).
\]
This map acts on the unit square $X = [0, 1] \times [0, 1]$ by first cutting $X$ into two vertical rectangles, $A$ and $B$, then stacking them (with $B$ on top) and finally flattening them to obtain a square again. It is measurably conjugate to the full 2-shift.

8. Hamiltonian flows. Let $(M^{2n}, \omega)$ be a symplectic manifold. Then any smooth function $H : M^{2n} \to \mathbb{R}$ gives rise to a natural vector field $X$, characterized by $dH = i_X(\omega)$ or equivalently

$$\omega(X,Y) = YH$$

for every vector field $Y$. In the case of classical mechanics, $M^{2n}$ is the cotangent bundle to a manifold $N^n$ and $\omega = \sum dp_i \wedge dq_i$ is the canonical symplectic form in position–momentum coordinates. Then $H$ gives the energy of the system at each point in phase space, and $X = X_H$ gives its time evolution. The flow generated by $X$ preserves both $H$ (energy is conserved) and $\omega$. To see this, note that

$$\mathcal{L}_X(\omega) = di_\omega + i_X d\omega = d(dH) = 0 \quad \text{and} \quad XH = \omega(X,X) = 0.$$ 

In particular, the volume form $V = \omega^n$ on $M^{2n}$ is preserved by the flow.

When $M^{2n}$ is a Kähler manifold, we can write $X_H = J(\nabla H)$, where $J^2 = -I$. In particular, in the plane with $\omega = dx \wedge dy$, any function $H$ determines an area–preserving flow along the level sets of $H$, by rotating $\nabla H$ by 90 degrees so it becomes parallel to the level sets. The flow accelerates when the level lines get closer together.

9. Area–preserving maps on surfaces. The dynamics of a Hamilton flow on a surface are very tame, since each level curve of $H$ is invariant under the flow. (In particular, the flow is never ergodic with respect to area measure.) One says that such a system is completely integrable.

General area–preserving maps on surfaces exhibit remarkable universality, and many phenomena can be consistently observed (elliptic islands and a stochastic sea), yet almost none of these phenomena has been mathematically justified. (An exception is the existence of elliptic islands, which follows from KAM theory and generalizes to symplectic maps in higher dimensions.)

10. The geodesic flow on a Riemannian manifold preserves Liouville measure on the tangent bundle and a similar measure on the unit tangent
bundle. One can view the geodesic flow on \( N^n \) as an example of a Hamiltonian flow; given a Riemannian metric, \( H = \sum g_{ij} q_i q_j \) gives the kinetic energy of a particle and is preserved under its frictionless motion; and \( \omega^n \) gives the Liouville measure on the cotangent bundle of \( N \).

11. An intermediate regime is provided by the Hamilton flow attached to a closed 1-form \( \alpha \) on a surface, via \( i_X(\omega) = \alpha \). Closely related is the transversally measured foliation defined by \( \alpha \) itself. The dynamics of its first–return map give rise to interval–exchange transformations, since the transverse measure is preserved.

12. Billiards [KMS]. The motion of a billiard ball in a smoothly bounded plane region also preserves the natural measure on the unit tangent bundle to the region. One can imagine the double of the region as a closed surface, with curvature concentrated along the edge. In the case of a polygon, the curvature is concentrated at single points (the vertices). Many easily stated questions about billiards in polygons are open. For example, does every triangle have a periodic trajectory? Is the motion of billiards in almost every triangle ergodic?

Billiards in polygons can be related to the dynamics of a holomorphic 1-form on a Riemann surface.

Absence of invariant measures. It is worth noting that there are some situations where \( G \) acts on a compact metric space \( X \) and no invariant measure exists. This is the case, for example, when \( \text{SL}_2(\mathbb{Z}) \) acts on \( S^1 \) by Möbius transformations. In such a case, one can still ask if a given measure is ergodic in the sense above.

The space of invariant measures. On the other hand, once an invariant measure exists (on a compact metric space), an ergodic invariant measure exists. This results from the following three facts:

1. The space of invariant measures \( M \) forms a compact convex subset of \( C(X)^* \) in the weak* topology;

2. A measure in \( M \) is ergodic iff it is an extreme point; and

3. (Krein–Milman) a compact convex set is the closed convex hull of its extreme points.

**Theorem 1.1** Let \( T : X \to X \) be a homeomorphism of a nonempty, compact metric space. Then \( T \) admits an invariant measure.
Proof. Given $x \in X$, let $\mu_n$ be the average of $\delta$-masses at the points $x$, $T(x)$, $\ldots T^{n-1}(X)$. Then for any continuous function $f \in C(X)$, the values of $\mu_n(f)$ and $\mu_n(f \circ T)$ nearly agree. It follows that any weak* limit of $\mu_n$ is $T$-invariant. Such a limit exists by compactness of the space of probability measures on $X$ in the weak* topology.

Remark. The norm topology on $C(X)^*$ is rarely used. Note that $X$ embeds continuously into $C(X)^*$ in the weak* topology, by the map $x \mapsto \delta_x$. On the other hand, in the norm topology, the image of $X$ in $C(X)^*$ is a closed, discrete set.

A panorama of applications. We conclude this section by mentioning some geometric and arithmetic applications of ergodic theory.

1. Mostow rigidity. Thurston and Perelman proved that most compact topological 3-manifolds $M$ are hyperbolic; that is, $M$ is homeomorphic to a quotient $\mathbb{H}^3/\Gamma$ of hyperbolic space by a discrete group of isometries.

This ‘uniformization theorem’ is all the more remarkable because the hyperbolic structure, when it exists, is unique. More precisely, we have:

**Theorem 1.2 (Mostow)** Let $f : M^3 \to N^3$ be a homotopy equivalence between a pair of compact hyperbolic manifolds. Then $f$ can be deformed to an isometry.

The proof of this theorem is based on studying the ergodic theory $\Gamma$ acting on the sphere at infinity $S^2$ forming the boundary of hyperbolic space. The key point is that the action on $S^2 \times S^2$ is ergodic, or equivalently that the geodesic flow on $T_1M^3$ is ergodic.

As emphasized already by Klein in his Erlangen program, the study of hyperbolic 3-manifolds is closely related to the study of the Lie group $G = \text{SO}(3,1)$. Indeed, a hyperbolic 3-manifold is nothing more than a quotient

$$M = \Gamma \backslash G/K,$$

where $\Gamma \subset G$ is a discrete, torsion–free group and $K = \text{SO}(3)$ is a maximal compact subgroup. The study of the geodesic flow on the frame bundle of $M$ is equivalent to the study of the action of $A$ on $\Gamma \backslash G$, where $A$ is the diagonal subgroup of $\text{SO}(3,1)$. And the study of the action of $\Gamma$ on $S^2$ is equivalent to the study of its action on $G/AN$, where $N$ is a maximal unipotent subgroup of $G$.

From this perspective, ‘geometry’ is the study of $\Gamma \backslash G/K$ where $K$ is compact, and ‘dynamics’ is the study of $\Gamma$ acting on $G/H$ or of $H$ acting on...
Γ\G, where H is noncompact. In both cases we assume that Γ is discrete, and often that Γ\G has finite volume or is even compact.

2. The Oppenheim conjecture. A quadratic form in n variables over K is a homogeneous polynomial Q(x) of degree two in K[x₁,…,xₙ]. Usually we will assume that Q is nondegenerate, i.e. that the matrix Q(eᵢ,eⱼ) has nonzero determinant.

We say Q represents y if there exists an x ∈ Kⁿ, x ≠ 0, such that Q(x) = y.

Quadratic forms over R are determined by their signature (p,q), 0 ≤ p + q <= n. If Q represents zero over R, it is said to be indefinite. This is equivalent to requiring that its signature is not (n,0) or (0,n).

Meyer proved that an indefinite quadratic form over Z in 5 or more variables always represents zero. In other words, Q ∈ Z[x₁,…,x₅] represents zero iff Q represents zero over R.

In 1929, Oppenheim conjectured that an indefinite quadratic form over R in 5 or more variables ‘nearly’ represents zero over Z: for any ε > 0, there is an x ∈ Zⁿ, x ≠ 0, such that |Q(x)| < ε. A strong version of this result was finally proved by Margulis in 1987.

**Theorem 1.3** If Q(x₁,x₂,x₃) is a real quadratic form of signature (1,2), then either:

- Q is proportional to an integral form, and Q(Z³) is discrete, or
- Q(Z³) is dense in R.

The proof is an instance of a much stronger rigidity phenomenon for actions of unipotent groups, whose most general formulation is given by Ratner’s theorem. In the case at hand, one is reduced to analysis of the orbits of H = SO(2,1) on SL₃(Z)\SL₃(R). It is shown that such an orbit is either closed, or dense, giving the two possibilities above.

3. Littlewood’s conjecture. The original conjecture is that for any 2 real numbers a,b, we have

\[ \inf n\|na\| \cdot \|nb\| = 0, \]

where \( \|x\| \) is the fractional part of x. It is now known that the conjecture holds for all (a,b) outside a set of Hausdorff dimension zero (Einsiedler, Katok, Lindenstrauss).

The full conjecture would follow if one had a version of Ratner’s theorem for the semisimple group of diagonal matrices A ⊂ SL₃(R), acting again on SL₃(Z)\SL₃(R). More precisely, it would be sufficient to show that any bounded A orbit is closed (and hence homeomorphic to a torus \( \mathbb{R}^2/\mathbb{Z}^2 \)).
An equivalent conjecture is that whenever a lattice $\Lambda \subset \mathbb{R}^3$ has the property that $N(x) = |x_1x_2x_3|$ is bounded below (on its nonzero elements), the lattices comes from the ring of integers in a totally real cubic field.

If $(\alpha, \beta)$ satisfy Littlewood’s conjecture, then the lattice generated by $(1, \alpha, \beta)$, $(0, 1, 0)$ and $(0, 0, 1)$ has vectors of the form $(n, \|n\alpha\|, \|n\beta\|)$ with arbitrarily small norm, consistent with the conjecture just stated.

4. **Expanding graphs.** We mention in passing that the study of unitary representations of discrete groups like $\text{SL}_n(\mathbb{Z})$, which is a facet of ergodic theory since it includes the action of $\Gamma$ on $L^2(G/K)$, has led to the construction of explicit finite graphs with good expansion properties.

Indeed, $\text{SL}_3(\mathbb{Z})$ has Kazhdan’s property $T$, and thus once we fix a generating set $S$, the Cayley graphs for the quotients $\text{SL}_3(\mathbb{Z}/p)$ are a sequence of graphs of bounded degree with uniform expansion. This fact was noted by Margulis in 1975, when he was employed by the Institute of Problems of Information Transmission in the former Soviet Union. (Expanding graphs are useful, at least in principle, in networking problems.)

5. **Optimal billiards.** The problem here is to construct a polygon billiard tables $P \subset \mathbb{R}^2$ such that every billiard trajectory is either periodic, or uniformly distributed. (The simplest example of a table with optimal dynamics is a rectangle).

This problem is closely related to the study of the action of $H = \text{SL}_2(\mathbb{R})$, not on $\Gamma \backslash G$, but on
\[
\Omega M_g = \text{Mod}_g \backslash \Omega T_g,
\]
the moduli space of holomorphic 1-forms on Riemann surfaces of genus $g$. One expects a theorem like Ratner’s in this setting, and indeed such a result was proved very recently by Eskin and Mirzakhani.

6. **Irrational rotations.** We note that the ‘first example’ in ergodic theory, the irrational rotation of a circle, can be thought of as an instance of the Lie group discussion above: it concerns the action of $H$ on $\Gamma \backslash G$ where $G = \mathbb{R}$, $\Gamma = \mathbb{Z}$ and $H = \mathbb{Z} \theta$. We will discuss this example in detail below.

### 2 Ergodic theory

We now turn to a more systematic study of measurable dynamics. A useful reference for this section is [CFS].

**The basic setting.** The basic setting of ergodic theory is a measure-preserving transformation $T$ of a probability space $(X, \mathcal{B}, m)$. This means:
\[
m(T^{-1}A) = m(A) \forall A \in \mathcal{B}.
\]
Usually we assume $T$ is invertible; then we obtain an action of $\mathbb{Z}$ on $X$.

**Group actions.** More generally, the action of a topological group $G$ on a measure space $(X, B, m)$ is measurable if $T : G \times X \to X$ is measurable. This means $T^{-1}(E)$ is measurable for all measurable $E$. The Borel structure on $G$ is used to defined measurability in $G \times X$.

**How many measure spaces are there?** When $X$ is a compact metric space, the space of all measures on $X$ can be identified with the dual of $C(X)$. This is taken as the Riesz representation definition of a measure by Bourbaki (it used to be a theorem!). The most useful notation of convergence is *weak convergence*, meaning $\mu_n \to \mu$ if

$$\int f \mu_n \to \int f \mu$$

for all $f \in C(X)$. In this topology, the space of probability measures is convex and compact.

In a complete metric space, every open set $U$ arises as the preimage of an open set under a continuous function (e.g. $f(x) = d(x, \partial U)$). Thus the smallest $\sigma$-algebra that allows us to integrate all continuous function is the algebra of Borel sets. The elements of $C(X)^*$ can be regarded as measures with respect to this $\sigma$-algebra.

A *Borel map* between topological spaces is a map such that $f^{-1}(B)$ is Borel for all Borel sets $B$. (Note: a Lebesgue measurable function $f : [0, 1] \to [0, 1]$ is not always a Borel map, and it is not always integrable with respect to an arbitrary measure on $[0, 1]$.)

Borel maps are extremely flexible, and consequently we have the following:

**Proposition 2.1** Any Borel subset of a complete, separable metric space $X$ is Borel isomorphic to either $[0, 1]$ or to a countable set.

In particular, if $X$ itself is uncountable then $X$ is Borel isomorphic to $[0, 1]$. We can then transport any Borel measure $\mu$ on $X$ to $[0, 1]$. Taking the pushforward of this measure under the map $f(x) = \mu[0, x]$, leads to:

**Proposition 2.2** Any Borel probability measure on $[0, 1]$ without atoms is Borel isomorphic to Lebesgue measure.

For more on these foundational points, see [Mac], [Me, Prop 12.6], or [Roy].

**Poincaré recurrence:** Here is one of the most general results concern measure–preserving dynamics.
**Theorem 2.3** If \( m(A) > 0 \) then almost every \( x \in A \) is recurrent (we have \( T^i(x) \in A \) for infinitely many \( i > 0 \)).

**Proof.** Let \( N \) be the set of points in \( A \) whose forward orbits never return to \( A \). Then \( T^i(N) \cap N \subset T^i(N) \cap A = \emptyset \) for any \( i > 0 \). It follows that \( \langle T^i(N) \rangle \) are disjoint for all \( i \in \mathbb{Z} \); since \( \sum m(T^i(N)) \leq 1 \), we conclude \( m(N) = 0 \). Thus almost every \( x \in A \) returns at least once to \( A \). Replacing \( T \) with \( T^n \) for \( n \gg 0 \), we see almost \( x \) returns to \( A \) infinitely many times. \( \blacksquare \)

**The Rohlin-Halmos lemma.** Here is one more result that applies to general measure preserving maps.

**Theorem 2.4** Let \( T \) be a measure–preserving transformation whose periodic points have measure zero. Then for any \( n, \epsilon > 0 \) there exists a measurable set \( E \) such that \( E, T(E), \ldots, T^n(E) \) are disjoint and \( E \cup T(E) \cup \cdots \cup T^n(E) \) has measure at least \( 1 - \epsilon \).

See [CFS, §10.5]. It is interesting to work this out for particular cases, e.g. an irrational rotation of the circle.

**The category of measurable \( G \)-spaces.** Suppose \( G \) acts on \( X \) and \( Y \), preserving probability measures \( \mu_X \) and \( \mu_Y \). Then a *morphism* is a measurable map \( \phi : X \to Y \) such that \( \phi \) sends \( \mu_X \) to \( \mu_Y \), and \( \phi(gx) = g \cdot \phi(x) \) for (almost) every \( x \in X \).

One often says \( Y \) is a *factor* of \( X \) in this situation. Of course if \( \phi \) is invertible, then \( X \) and \( Y \) are isomorphic.

One of the fundamental questions in ergodic theory is: when are two dynamical systems measurably isomorphic? As we have seen, \( X \) and \( Y \) are often isomorphic as measure spaces, so all the action is in the behavior of \( T \).

**The spectral point of view: the unitary functor.** Here is a central theme in ergodic theory. Any measure–preserving map \( T \) induces a unitary operator \( U : H \to H \) on the Hilbert space \( H = L^2(X, m) \). Put differently, we have an arrow–reversing *functor*

\[
(X, m) \to L^2(X, m)
\]

from the category of measure spaces and quotient maps to Hilbert spaces and isometric inclusions. From this we obtain a function from the category of *measurable automorphisms* to the category of *unitary automorphisms*. On the level of objects, it sends

\[
(T : X \to X) \mapsto (U : L^2(X) \to L^2(X)),
\]
where $U$ is a unitary automorphism.

The same functor can be constructed for measurable $G$–actions, with the target the category of unitary representations of $G$.

Now the irreducible representations of a locally compact abelian group (such as $\mathbb{Z}$ or $\mathbb{R}$) are all one-dimensional. So an ergodic action is never irreducible from the point of view of representation theory, and finding its spectral decomposition gives additional structure and invariants.

**Spectral invariants.** The Hilbert space $L^2(X, m)$ comes equipped with a natural linear functional, namely

$$f \mapsto \int f \, d\mu = \langle f, 1 \rangle.$$  

It is frequently useful to eliminate the obvious invariant subspace of $L^2(X, m)$, namely the constant functions, by passing to the kernel:

$$L^2_0(X, m) = \{ f \in L^2(X, m) : \int f \, dm = 0 \}.$$  

Now assume $H$ is a separable Hilbert space. By the spectral theorem, there are two invariants which completely determine a unitary automorphism $U : H \to H$ up to isomorphism:

- Its spectral measure $\mu$, which is a finite Borel measure on $S^1$; and
- The multiplicity function $m : S^1 \to \{ 1, 2, \ldots, \infty \}$, which measures the size of each eigenspace.

From this data we can construct a Hilbert space bundle $\mathcal{H} \to S^1$ and its space of sections $L^2(S^1, \mu, \mathcal{H})$, with the norm:

$$\| f \|^2 = \int_{S^1} \| f(\lambda) \|^2 \, d\mu(\lambda).$$  

Then there is an isomorphism

$$\phi : H \cong L^2(S^1, \mu, \mathcal{H})$$  

such that if $\phi(x) = f$, then

$$\phi(Ux) = \lambda f(\lambda).$$  

Thus $(\mu, m)$ determines $U$ up to isomorphism. (See Appendix A for more details.)
Measure classes. Two measures are in the same measure class if they have the same sets of measure zero. It is not hard to show that \((\mu, m)\) and \((\mu', m')\) determine the same unitary operator iff \(\mu\) and \(\mu'\) are in the same measure class, and \(m = m'\) a.e. on \(S^1\) (with respect to that class). Although one speaks of ‘the spectral measure of \(U\),’ actually it is only the measure class that is determined by \(U\).

The finite–dimensional case. If \(H\) is finite–dimensional, then \(U\) has eigenvalues \(\lambda_1, \ldots, \lambda_n\) with multiplicities \(m_1, \ldots, m_n\). In this case \(\mu = \sum \delta_{\lambda_i}\) and \(m(\lambda_i) = m_i\). The values of \(m\) on the rest of the circle are irrelevant. Thus the spectral theorem diagonalizes \(U\).

Lebesgue spectrum. Let \(T: \mathbb{Z} \to \mathbb{Z}\) be the shift map, which preserves counting measure. Then \(T\) induces a unitary operator \(U: \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})\). Clearly this operator has no eigenvectors. Its spectrum is completely continuous. In fact, the map

\[
(a_i) \mapsto \sum a_iz^i
\]

gives an isomorphism between \(\ell^2(\mathbb{Z})\) and \(L^2(S^1)\) such that

\[
(Ua_i) = (a_{i-1}) \mapsto \sum a_{i-1}z^i = z \sum a_iz^i;
\]

that is, it sends \(U\) to the operator \(f(z) \mapsto af(z)\). Thus the spectral measure of \(U\) is Lebesgue measure on \(S^1\), and \(m(z) = 1\). In this case we say \(U\) has Lebesgue spectrum of multiplicity one.

The shift map on \(\ell^2(\mathbb{Z}) \otimes \mathbb{C}^n\) has Lebesgue spectrum of multiplicity \(n\).

Ergodicity. We say \(T\) is ergodic if whenever \(X\) is split into a disjoint union of measurable, \(T^{-1}\)-invariant sets,

\[
X = A \sqcup B,
\]

either \(mA = 0\) or \(m(X - A) = 0\). The following are equivalent:

1. \(T\) is ergodic.
2. If \(m(A) > 0\) then \(\bigcup_{i=0}^{\infty} T^i(A) = X\) is a set of full measure.
3. Any \(T\)-invariant measurable function is constant a.e.
4. The kernel of \((T - I)\) acting on \(L^2_0(X, m)\) is trivial.
5. The spectral measure of \(T\) acting on \(L^2_0(X, m)\) assigns no mass to \(\lambda = 1\).
Mixing. We say \( T \) is mixing if whenever \( A, B \subset X \) have positive measure, we have
\[
m(A \cap T^{-i}(B)) \to m(A)m(B)
\]
as \( i \to \infty \). (For an automorphism, we can use \( T^i \).) Clearly mixing implies ergodicity. It is equivalent to the following condition on functions: for any \( f, g \in L^2(X) \), we have
\[
\int f(x)g(T^i(x)) \, dx \to \int f \int g.
\]

Example. Shifts. Let \( I \) be a compact metric space (such as \([0,1]\)) with a probability measure \( \nu \). The 1-sided shift for \( I \) is given on the compact space
\[
X = \prod_1^\infty I,
\]
with the product measure \( \prod_1^\infty \nu \), by
\[
\sigma(x_1, x_2, \ldots) = (x_2, x_3, \ldots).
\]
It is easy to see that \( \sigma \) is a continuous, measure–preserving endomorphism of \( X \). (One can also form the 2-sided shift, which would be an automorphism.)

The shift space gives a good model for a sequence of independent trials of a random variable. Namely, suppose \((I, \nu) = ([0,1], dx)\) and \( f \in L^1(I) \). Then we can define a sequence of random variables \( Y_i : X \to \mathbb{R} \) by \( Y_i(x) = f(x_i) \). These are independent, and equally distributed. Then \((Y_1(x), Y_2(x), \ldots)\) is a model for repeated trials of the same experiment or measurement.

Theorem 2.5 The 1-sided shift is mixing, and hence ergodic.

Proof. The space \( C(X) \) is dense in \( L^2(X, \mu) \), and by the Stone–Weierstrass theorem, the algebra \( A \subset C(X) \) of functions that depend on only finitely many coordinates \((x_1, x_2, \ldots)\) is dense in \( C(X) \). But for any pair of functions \( f, g \in A \), there exists an \( N \) such that \( f \) and \( g \circ T^i \) depend on different coordinates for all \( i \geq N \). Thus
\[
\int f(x)g(T^i(x)) \, dx = \int f \int g
\]
by Fubini’s theorem, for \( i \geq N \). Mixing now follows from density of \( A \) in \( L^2(X) \). \( \blacksquare \)
Variants. The same argument shows the 2-sided shift is mixing. One can also consider, for any finite directed graph \( G \), the subset \( \Sigma \subset E(G)^{\mathbb{Z}} \) consisting of all bi-infinite paths in \( G \). This is shift-invariant and carries many natural invariant measures. If \( G \) is strongly connected, meaning there is a directed path from \( a \) to \( b \) for any \( a, b \in V(G) \), then \( (\Sigma, \sigma) \) has a unique measure of maximal entropy and this measure is mixing.

Even for the shift on two symbols, \( \{0,1\}^{\mathbb{Z}} \), there are a great many invariant measures, for example the atomic measure spread along a periodic cycle.

**Theorem 2.6** An automorphism \( T \) of a compact Abelian group \( G \) is ergodic with respect to Haar measure iff \( \hat{T} \) has no periodic points in \( \hat{G} \), other than the trivial character.

**Group automorphisms. Harmonic analysis on the torus.** Consider the action of a matrix \( T \in \text{SL}_d(\mathbb{Z}) \) on the torus \( X = (S^1)^d = \mathbb{R}^d/\mathbb{Z}^d \). Given \( (x_1, \ldots, x_d) \in \mathbb{R}^n \) the corresponding point on \( X \) can be described as \( z = (z_1, \ldots, z_d) \) where \( z_i = e^{2\pi i x_i} \).

The (Pontryagin) dual of the compact group \( X = (S^1)^d \) is the discrete group \( \hat{X} = \mathbb{Z}^d = \text{Hom}(X, S^1) \). The character given by \( n = (n_1, \ldots, n_d) \in \mathbb{Z}^d \) is given by
\[
\chi_n(x) = \exp(2\pi i \langle n, x \rangle) = z^n = z_1^{n_1} \cdots z_d^{n_d}.
\]
Since \( T \) is a group automorphism, \( \chi_n \circ T \) is also a character; indeed,
\[
\chi_n(T(x)) = \exp(2\pi i \langle n, T(x) \rangle) = \exp(2\pi i \langle T^*(n), t \rangle) = \chi_{T^*n}.
\]
Thus if \( f \in L^2(X) \) is given by the Fourier series
\[
f(x) = \sum a_n \chi_n(x),
\]
then we have
\[
f \circ T = \sum a_{(T^*)^{-1}n} \chi_n(x).
\]
In other words, the unitary operator \( U : L^2(X) \to L^2(X) \) associated to \( f \) is isomorphic to the operator on \( L^2(\mathbb{Z}^d) \) induced by a linear automorphism \( S = (T^*)^{-1} \) with the same behavior as \( T \).

In particular, if \( f \) is \( T \)-invariant, then \( a_n \) is an \( S \)-invariant function on \( \mathbb{Z}^d \). Now if \( f \) is non-constant, then \( a_n \neq 0 \) for some \( n \neq 0 \). Since \( \sum |a_n|^2 \) is finite, this implies \( S^p(n) = n \) for some \( p > 0 \). Conversely if \( S^p(n) = n \), then \( f(x) = \sum_{i=1}^p \chi_n(T^i x) \) is a nonconstant, \( T \)-invariant function. Since \( S \) is conjugate to \( T^{-1} \), this shows:
**Theorem 2.7** $T$ is ergodic if and only if $T$ has no periodic point on $\mathbb{Z}^d - \{0\}$.

**Corollary 2.8** $T$ is ergodic iff no eigenvalue of $T$ is a root of unity.

**Corollary 2.9** $T$ is ergodic iff $T$ is mixing. In fact $T|L^2_0(X)$ is isomorphic, as a unitary operator, to the shift map on $\bigoplus_1^\infty \ell^2(\mathbb{Z})$.

**Remark.** Of course the shift map’s operator is itself conjugate to multiplication by $z$ on $L^2(S^1)$. In particular, its spectral measure is Lebesgue measure of multiplicity one. This shows:

**Theorem 2.10** If $T$ is ergodic, then $T|L^2_0(X)$ has Lebesgue spectrum of infinite multiplicity.

**Proof.** If $T$ has no periodic points, then the same is true for $S$. Thus the nonzero orbits of $S$ consist of countably many copies of $\mathbb{Z}$, with $S$ acting by translation. 

**Examples.** In $\text{SL}_2(\mathbb{Z})$, $T$ is ergodic iff $T$ is hyperbolic. If $T$ is parabolic then it preserves a foliation by closed circles, and any function constant on these leaves is invariant. If $T$ is elliptic then $X/T$ is a nice space and hence there are abundant invariant functions.

**Corollary 2.11** All hyperbolic toral automorphisms are unitarily conjugate.

It is an interest challenge to show they are not all spatial conjugate. Here a useful invariant is the entropy, introduced by Kolmogorov for exactly this purpose.

**Space of invariant measures.** A hyperbolic toral automorphism has a dense set of periodic points, since it acts by permutation on $X[n]$, the points of order $n$. Thus it clearly admits many invariant measures besides Lebesgue measure. The Fourier transform of such a measure (other than Lebesgue) cannot tend to zero in $\hat{X} = \mathbb{Z}^d$.

**Group endomorphisms: the doubling map.** An even simpler example of mixing is provided by the measure–preserving endomorphism of $G = S^1 = \mathbb{R}/\mathbb{Z}$ given by $T(x) = 2x \mod 1$. Then $\overline{T}$ acts on $\tilde{X} = \mathbb{Z}$ by $n \mapsto 2n$. From this it is immediate that $T$ is mixing. Intuitively, if $f$ is a function of mean zero, then $f(T^n x) = f(2^n x)$ is a sound wave with very high frequency, that is hard to hear with a fixed listening device.
It is easy to see that the doubling map has a huge set of invariant measures. For example, the coin-flipping measures $\mu_p$, where heads come up with probability $p$ and determine the binary digits of $x \in [0,1]$, are all $T$-invariant, mutually singular, mixing and ergodic. (They are equivalent to a shift system). The many periodic points of $T$ are another source of invariant measures.

One might also consider the tripling map, $S(x) = 3x \mod 1$. It is a famous open problem, due to Furstenburg, to show that the only nonatomic measure invariant under both $S$ and $T$ is Lebesgue measure.

**Automorphisms of compact Abelian groups.** In general if $G$ is a compact Abelian group, then its continuous characters

$$\hat{G} = \text{Hom}(G, S^1)$$

form a discrete group, and have:

**Translations in a group. An irrational rotation of the circle.** Let $S^1 = \mathbb{R}/\mathbb{Z}$, let $\theta$ be an irrational number, and define a rotation $T : S^1 \to S^1$ by

$$T(x) = x + \theta \mod 1.$$ 

This is one of the most basic examples in dynamics.

Our first observation is that $T$ is minimal. That is, every orbit of $T$ is dense in $S^1$. This can easily be proved directly, or it can be related to the following important fact:

**Theorem 2.12** A closed subgroup of a Lie group is a Lie group.

Next we observe:

**Theorem 2.13** An irrational rotation of $S^1$ is ergodic.

**Proof.** Let $A \subset S^1$ be a $T$-invariant set of positive measure. Pick a point of Lebesgue density and move it around to a dense subset of $S^1$. From this one can deduce that $m(A)$ is as close to one as you like. ■

An easy exercise is to show that a rotation is not mixing, which implies:

**Theorem 2.14** An irrational rotation of $S^1$ is not isomorphic to a shift map.
(On the other hand, the map $f(z) = z^2$ on $S^1$ is measurably isomorphic to the 1-sided two shift.)

**Harmonic analysis on the circle.** Here is an alternate and useful proof of ergodicity, due to Weyl, that actually proves much more — it shows that $T$ is *uniquely ergodic*. We will study, for a function $f$ on $S^1$, the sums along orbits given by

$$ (S_nf)(x) = \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x). $$

**Theorem 2.15** For any $f \in C(S^1)$, $S_nf$ converges uniformly to the constant function $\int_{S^1} f(x) \, dx$ as $n \to \infty$.

**Proof.** Since $\|S_nf\| \leq \|f\|$ in the sup-norm on $C(X)$, it suffices to give the proof for a dense set in $C(S^1)$. Let us first see what happens for $f(z) = z^k$, $z = \exp(2\pi i x)$, $k \in \mathbb{Z}$. If $k = 0$ we have $S_nf = 1$ for all $n$ as we wish. Otherwise, we note that $f$ is an eigenfunction of $T$, namely

$$ f(T(z)) = \zeta^k f(z), $$

where $\zeta = \exp(2\pi i \theta)$. Thus by summing the geometric series, we find

$$ (S_nf)(z) = f(z) \cdot \frac{1}{n} \frac{1 - \zeta^{nk}}{1 - \zeta^k} \to 0 $$

as $n \to \infty$. Here we have used the fact that $\theta$ is irrational to insure that the denominator does not vanish. The convergence is uniform, so the same holds true for any finite sum,

$$ f(z) = \sum_{-N}^{N} a_k z^k. $$

By the Stone–Weierstrass theorem, such sums are dense in $C(S^1)$, and the theorem follows.  

**Unique ergodicity.** A homeomorphism of a topological space is *uniquely ergodic* if it has only one invariant probability measure.

**Corollary 2.16** An irrational rotation is uniquely ergodic.
Proof. Let $\mu$ be a $T$-invariant probability measure on $S^1$. Then for all $f \in C(X)$, we have
\[
\int f \, d\mu = \int S_n(f) \, d\mu \to \left( \int f \, dx \right) \, d\mu = \int f \, dx.
\]
Thus $\mu = dx$.

Corollary 2.17 For any interval $I = (a, b)$ on the circle, and any $x \in S^1$, the amount of time the orbit of $x$ spends in $I$ is a multiple of its length. That is,
\[
\lim_{N \to \infty} \frac{1}{N} \left| \{0 \leq i < N : T^i(x) \in I \} \right| \to \frac{|I|}{|S^1|}.
\]

Proof. Approximate $\chi_I$ from above and below by continuous functions, and apply the preceding results.

Distribution of decimal digits. (Arnold.) Write the powers of 2 in base 10:

$1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768, 65536 \ldots$

(The last two numbers are well-known to computer scientists of a certain generation, since they are one more than the largest signed and unsigned integers that an be represent in 16 bits.)

How often is the first digit equal to 1? Using the preceding Corollary, it is easy to see that the answer is about 30%. The same predominance of 1s occurs in many different types of real-world data, and is used in forensic finance to help detect fraud. (See: Benford’s Law.) It has also been observed that the beginning of a table of logarithms is usually much more worn than the end (Simon Newcomb).

Uniform distribution. Quite generally, a sequence $a_n \in S^1$ is said to be uniformly distributed if
\[
\frac{1}{N} \sum_{i=1}^{n} f(a_n) \to \int_{S^1} f(x) \, dx
\]
for any $f \in C(X)$. This means that $\delta$-masses spread out evenly along this sequence converge weakly to Lebesgue measure.
Weyl’s theorem shows that \((n\theta)\) is uniformly distributed whenever \(\theta\) is irrational. A great many other sequences are known to be uniformly distributed; for example, the sequence \((n^\alpha)\) for \(0 < \alpha < 1\). It is suspected that the sequence \((3/2)^n \mod 1\) and similar sequences are uniformly distributed, but even their density is unknown.

It is known that \(\theta^n\) is not equidistributed when \(\theta\) is a Salem number, i.e. an algebraic integer \(x\) all of whose conjugates satisfy \(|x'| \leq 1\).

**Translations in general.** Here is a general argument that also proves ergodicity of an irrational rotation, without using Fourier series.

**Theorem 2.18** Let \(T : G \to G\) be translation in a compact group. Then \(T\) is uniquely ergodic with respect to Haar measure iff \(T\) has a dense orbit.

**Proof.** Let \(T(x) = gx\), and let \(H\) be the closure of the group generated by \(g\). If \(\mu\) is invariant under \(\langle T^n \rangle\), then it is also invariant under \(H\). So if \(H = G\), \(\mu\) must be Haar measure (which is unique). Otherwise, \(G/H\) is a standard Borel space, so there are many different \(T\)-invariant functions on \(G\).

**Remark.** In particular, if \(G\) is a nonabelian group like \(\text{SO}(n), n > 2\), then \(T\) cannot be ergodic since it generates an Abelian subgroup. By similar considerations, one can show that a rotation of \(S^n, n \geq 2\), is never ergodic. (How large can an orbit closure be on \(S^3\)?)

**Minimality and unique ergodicity.** A topological dynamical system is *minimal* if every orbit is dense. (Compare *transitive*, meaning there exists a dense orbit.) The following result is clear for group translations, but holds more generally.

**Theorem 2.19** If \(T\) is uniquely ergodic, and its invariant measure has full support, then \(T\) is minimal.

**Proof.** Let \(f\) be supported in an arbitrary open set \(U \neq \emptyset\), with \(\int f \, d\mu = 1\). Then \(S_n(x, f) \to 1\) for any \(x \in X\). Therefore the orbit of \(X\) enters \(U\), so it is dense.

**Minimal but not ergodic.** Furstenberg constructed an analytic diffeomorphisms of the torus which is minimal but not uniquely ergodic; see [Me, §II.7]. There are also well-studied examples of interval exchange maps with the same property.
Cocycles and coboundaries. Now let $T : X \to X$ be a homeomorphism on a compact metric space. A function $f \in C(X)$ is a coboundary if $f = g - g \circ T$ for some $g \in C(X)$.

Clearly $S_n(g, x) \to 0$ for a coboundary, in fact $|S_n(g, x)| = O(1/n)$. (Note that a constant function cannot be coboundary.)

**Theorem 2.20** The following are equivalent.

1. $T$ is uniquely ergodic.
2. For all $f \in C(X)$, $(S_n f)(x)$ converges pointwise to a constant function.
3. For all $f \in C(X)$, $(S_n f)(x)$ converges uniformly to a constant function.
4. We have $C(X) = C \oplus \text{Im}(I - T)$. That is, every continuous function is the sum of a constant function and a uniform limit of coboundaries.

**Proof.** By the Hahn–Banach theorem, the space of invariant measures is given by

$$M(X)^T = \text{Im}(T - I)^\perp \cong (C(X)/\text{Im}(T - I))^*.$$ 

This shows the first and last conditions are equivalent. As for the others, just note that if $S_n f \to A(f)$ pointwise, then it also converges in $L^1$, and hence for any invariant measure we have

$$\int f \, d\mu = \int S_n f \, d\mu \to A(f).$$

Thus $\mu$ is unique.

**Aside:** Finite sums of coboundaries. Given a function $f$ on $S^1$ of mean zero, can we always write $f$ as a coboundary rather than as a limit of coboundaries? For simplicity we focus on the case where $f(z) = \sum a_n z^n$ is in $L^2(S^1)$. Then if $f = g - g \circ T$, where $g = \sum b_n z^n$, we find

$$b_n = \frac{a_n}{1 - \lambda^n}.$$ 

Since $\lambda^n$ is dense on the circle, the denominators accumulate at zero and in general we cannot have $\sum |b_n|^2 < \infty$. So the answer is no.

However every function $f \in L^2(S^1)$ with mean zero is the sum of 3 boundaries for the full rotation group.
Theorem 2.21 Any function \( f \in L^2(S^1) \) with \( \int f = 0 \) is the sum of 3 coboundaries for the rotation group: that is, \( f = \sum_1^3 g_i - g_i \circ R_i \), with \( g_i \in L^2(S^1) \) and \( R_i(z) = \lambda_i z \) a rotation.

Corollary 2.22 Any rotation-invariant linear functional on \( L^2(S^1) \) is a multiple of Lebesgue measure.

Proof of Theorem. Write \( f(z) = \sum a_n z^n \) with \( \sum |a_n|^2 < \infty \). We wish to find \( \lambda_i \in S^1 \) and \( g_i = \sum b^n_i z^n \in L^2(S^1) \), \( i = 1, 2, 3 \), such that

\[
a_n = \sum_1^3 b^n_i (1 - \lambda^n_i).
\]

Thinking of \( b_n \) and \( \lambda^n \) as vectors in \( \mathbb{C}^3 \), we can choose \( b_n \) to be a multiple of \( \lambda^n \), in which case \( |b_n|^2 = |a_n|^2 / |1 - \lambda^n|^2 \).

Now consider, on the torus \( (S^1)^3 \subset \mathbb{C}^3 \), the function \( F(\lambda) = |1 - \lambda|^2 \). Near its singularity \( \lambda = (1,1,1) \) this function behaves like \( 1/r^2 \), which is integrable on \( \mathbb{R}^3 \). (This is where the 3 in the statement of the Theorem comes from.) Also, since \( \lambda \mapsto \lambda^n \) is measure-preserving, we find that

\[
\int_{(S^1)^3} F(\lambda) d\lambda = \int_{(S^1)^3} F(\lambda^n) d\lambda
\]

and thus

\[
\int_{(S^1)^3} \sum_n |b_n|^2 = \int_{(S^1)^3} \sum_n \frac{|a_n|^2}{|1 - \lambda^n|^2} d\lambda = \sum |a_n|^2 \|F\|_1 < \infty.
\]

Thus almost every triple of rotations \( (\lambda_1, \lambda_2, \lambda_3) \) works.

Pointwise convergence in general? We now return to the setting of an irrational rotation \( T : S^1 \to S^1 \). We have seen that the frequency with which an orbit of \( T \) visits an interval \( U = (a, b) \subset S^1 \) is given by \( m(U)/m(S^1) \).

It is natural to expect, more generally, that the amount of time an orbit of \( T \) spends in an open set \( U \subset S^1 \) is proportional to the measure of \( U \). This cannot, however, be true for every point, since we could take \( U \) itself to be a neighborhood of a complete orbit, with small measure.

The ergodic theorem. The fact that orbits are uniformly distributed in this stronger sense is a consequence of a much more general theorem, that is central in the subject. It is a result in pure measure theory, called the Birkhoff–Khinchin ergodic theorem.
**Theorem 2.23** Let $T$ be a measure-preserving transformation of a probability space $(X, \mu)$. Then for any $f \in L^1(X)$, there exists a $T$-invariant function $\overline{f} \in L^1(X)$ such that 

$$(S_nf)(x) \to \overline{f}(x)$$

for almost every $x \in X$. Moreover, $\|S_nf - \overline{f}\|_1 \to 0$.

**The maximal theorem.** We will give two proofs of this important result. The first hinges on the following result.

**Theorem 2.24 (The maximal theorem)** Given $f \in L^1(X, \mu)$, let $A = \{x : \sup_n (S_nf)(x) > 0\}$. Then $\int_A f \, d\mu \geq 0$.

**Proof.** We will use the fact that the integral of a coboundary is zero. Let $F_n(x) = f(x) + f(Tx) + \cdots + f(T^{n-1}x)$ be the running sum of $f$, let $M_n(x) = \max\{F_1(x), \ldots, F_n(x)\}$, and let $M^*_n(x) = \max\{F_0(x), \ldots, F_n(x)\}$.

Note that $M^*_n(x) \geq 0$, that $M_1 \leq M_2 \leq \cdots$, and that $A$ is the increasing union of the sets $A_n = \{x : M_n(x) > 0\}$. Note also that we have $M_{n+1}(x) = M^*_n(Tx) + f(x)$, which nearly expresses $f$ as a coboundary. We now observe that

$$\int_{A_n} f = \int_{A_n} M_{n+1} - M^*_n \circ T$$

$$\geq \int_{A_n} M_n - M^*_n \circ T \quad \text{because } M_n \text{ increases with } n$$

$$= \int_{A_n} M^*_n - M^*_n \circ T \quad \text{since } M_n = M^*_n \text{ on } A_n$$

$$= \int_X M^*_n - \int_{A_n} M^*_n \circ T \quad \text{since } M^*_n(x) = 0 \text{ outside } A_n$$

$$\geq \int_X M^*_n - M^*_n \circ T \quad \text{since } M^*_n \geq 0$$

$$= 0 \quad \text{since } T \text{ is measure-preserving.}$$

Since $A = \bigcup A_n$ the Theorem follows. □
Proof of the ergodic theorem. Suppose to the contrary that \((S_n f)(x)\) does not converge a.e. Then for some \(a < b\), the set

\[
E = \{ x : \lim \inf (S_n f)(x) < a < b < \lim \sup (S_n f)(x) \} \tag{2.1}
\]

has positive measure. Note that \(E\) is also \(T\)-invariant. Since we are aiming at a contradiction, we may now assume \(X = E\). We may also rescale and translate \(f\) so that \(a = -1 < 1 = b\).

Clearly \(\sup (S_n f)(x) > 0\) for every \(x \in E\); in fact \(\sup (S_n f)(x) > 1\). By the maximal theorem,

\[
\int_E f \geq 0.
\]

By the same reasoning, \(\inf (S_n f)(x) < 0\) and hence \(\int_E f \leq 0\).

Thus \(\int_E f = 0\). But we may apply the same reasoning to \(f - c\) for any small value of \(c\), and hence \(\int_E f = c\). This is a contradiction, so \((S_n f)(x)\) converges pointwise a.e.

Let \(F(x) = \lim (S_n f)(x)\). We wish to show that \(F \in L^1(\Omega)\) and \(\| S_n f - F \|_1 \to 0\). Since any \(L^1\) function is the difference of two positive functions, it suffices to treat the case where \(f \geq 0\). Then by Fatou’s Lemma, we have

\[
\| F \|_1 = \int F = \int \lim S_n f \leq \lim \inf \int S_n f = \int f = \| f \|_1,
\]

so \(F\) is in \(L^1\). Moreover, if \(f\) is bounded then \(S_n f \to F\) in \(L^1(\Omega)\) by the dominated convergence theorem. For the general case, write \(f\) as the sum \(f_0 + r\) of two positive functions, with \(f_0\) bounded and \(\| r \|_1 < \epsilon\). We have \(F \in L^1(\Omega)\). Moreover, if \(f\) is bounded then \(S_n f \to F\) in \(L^1\) by the dominated convergence theorem. But we can write \(f = f_0 + r\) where \(f_0\) is bounded and \(\| r \|_1 < \epsilon\). Then \(S_n f_0 \to F_0\) in \(L^1(\Omega)\). \(R(x) = \lim (S_n r)(x)\) satisfies \(F_0 + R = F\), and \(\| R \|_1 \leq \| r \|_1 < \epsilon\). Thus:

\[
\limsup \| S_n f - F \|_1 \leq \limsup \| S_n f_0 - F_0 \|_1 + \limsup \| S_n r - R \|_1 \leq 2\epsilon.
\]

It follows that \(S_n f \to F\) in \(L^1(\Omega)\) and the proof is complete. 

\[\square\]
Corollary 2.25 If \( T \) is ergodic, then for every \( f \in L^1(X) \), we have \( (S_n f)(x) \to \int_X f \, d\mu \) for almost every \( x \in X \).

Uniform distribution, reprise. Here is a reformulation of the ergodic theorem in the case of a topological dynamical system.

Theorem 2.26 Let \( T : X \to X \) be a homeomorphism of a compact metric space, preserving the probability measure \( \mu \). Suppose \( T \) is ergodic. Then almost every orbit of \( T \) is uniformly distributed with respect to \( \mu \).

The example of hyperbolic dynamics on the torus shows that one cannot expect every orbit to be uniformly distributed.

The strong law of large numbers. To illustrate the power of this result, we will use it to prove an important theorem in probability theory. Let \( X_1, X_2, \ldots \) be independent, identically distributed random variables, with \( E(|X_1|) < \infty \).

Theorem 2.27 The averages \( (X_1 + \cdots + X_n)/n \) converge to \( E(X_1) \) almost surely.

Proof. Choose a function \( F : [0,1] \to \mathbb{R} \) with the same distribution as \( X_1 \), and apply the ergodic theorem to the shift space \( \prod_1^\infty [0,1] \), with \( f(x_1, x_2, \ldots) = F(x_1) \).

Example. For almost every \( x \in [0,1] \), exactly 10% of its decimal digits are equal to 7.

Flows. We remark that the ergodic theorem also holds for flows: given a measure–preserving action \( H_t \) of \( \mathbb{R} \), and \( f \in L^1(X, \mu) \), there exists a flow invariant function \( F \in L^1(X, \mu) \) such that

\[
(S_r f)(x) = \frac{1}{2r} \int_{-r}^r f(H_t x) \, dt \to F(x)
\]

a.e. and in \( L^1 \). This version will be very useful when we study actions of Lie groups and, in particular, their 1-parameter subgroups, as well as geodesic flows.

von Neumann’s ergodic theorem. Suppose \( T \) is not ergodic. Then \( S_n f \to F \), but what is \( F \)? A succinct answer, at least for \( f \in L^2(X) \), is provided by von Neumann’s ergodic theorem. The answer, in brief is that \( F \) is simply the projection of \( f \) to the space \( L^2(X)^T \) of \( T \)-invariant functions.
The statement in fact has nothing to do with ergodic theory; it is a property of unitary operators.

The proof of von Neumann’s theorem is easier than the proof of the Birkhoff–Khinchin ergodic theorem. In mid–1931, Koopman made the connection between measurably dynamics and unitary operators. By October of the same year, von Neumann had proved the ergodic theorem below and communicated it to Birkhoff. By December, Birkhoff had completed the proof of his pointwise result, and arranged for it to be published before von Neumann’s result. (Both appeared in the Proc. Nat. Acad. Sci.)

**Theorem 2.28 (von Neumann’s ergodic theorem)** Let $U : H \to H$ be a linear isometry of Hilbert space, and let $P$ be projection of $H$ onto the subspace of $U$-invariant vectors. Then for any $f \in H$,

$$\frac{1}{n} \sum_{0}^{n-1} U^i(f) \to P(f).$$

**Proof.** The result clearly holds for $U$-invariant $f$, and for coboundaries $f = g - Ug$. (The latter are perpendicular to the invariant $f$, and their averages converge to zero). It also holds for limits of coboundaries. So it suffices to show $H$ is a direct sum of the invariant $f$ and the closure of the coboundaries.

Now suppose $f$ is perpendicular to the coboundaries. Then we find $f$ is $U^*$-invariant, since $\langle f, g - Ug \rangle = 0 = \langle f - U^*f, g \rangle$ for all $g$. If $U$ is an invertible isometry, then $U = U^*$ and we are done.

For a general isometry we still have $\langle f, g \rangle = \langle Uf, Ug \rangle$ for all $f, g \in H$, and hence $U^*U = I$. Now if $f$ is $U^*$ invariant, then

$$\|f - Uf\|^2 = \langle f, f \rangle + \langle Uf, Uf \rangle - 2\langle f, Uf \rangle.$$  

But $\langle Uf, Uf \rangle = \langle f, f \rangle$ since $U$ is an isometry, and $\langle f, Uf \rangle = \langle U^*f, f \rangle$ since $f$ is $U^*$ invariant. Thus $f$ is $U$-invariant as desired. 

This last part of the argument allows one to treat measurable endomorphisms, like $T(z) = z^2$ on $S^1$, which gives rise to an isometric map $U : L^2(S^1) \to L^2(S^1)$ which is not onto.

**The spectral perspective.** The spectral theorem for a unitary operator provides an equally transparent proof of the ergodic theorem above.

For convenience we assume $H$ is separable. Then by the spectral theorem, there is an isomorphism from $H$ to $\oplus L^2(S^1, \mu, m)$ where $\mu$ is a probability measure and $m$ is the multiplicity function. With this model for $H$,
the \( n \)th sum in the ergodic theorem is given by

\[
S_n(f) = \frac{1}{n} \left( 1 + z + \cdots + z^{n-1} \right) f(z) = \frac{1}{n} \frac{1 - z^n}{1 - z} f(z).
\]

Clearly

\[
\| (S_n f)(z) \| \leq \| f(z) \|
\]

and \((S_n f)(z) \to 0\) pointwise, for \( z \neq 1 \). By the dominated convergence theorem, this implies that \( S_n(f) \) converges to the unique function with \( F(1) = f(1) \) and \( F(z) = 0 \) for \( z \neq 1 \). This is exactly the projection of \( f \) onto the space of \( U \)-invariant functions.

**Forward and backward sums.** By von Neumann’s result we now can succinctly say what the ergodic averages \( S_n(f) \) converge to: for \( f \in L^2 \) it is just the projection of \( f \) onto the \( T \)-invariant functions.

In particular, we have:

**Corollary 2.29** For any measure-preserving map \( T \), the averages of \( f \) along the forward and backward orbits of \( T \) agree a.e.

**Another approach to the Birkhoff–Khinchin ergodic theorem.** (cf. Keane, [BKS, p.42], who attributes the idea to Kamae (1982)).

Let \( E \subset X \) be a measurable set, let \( S_n(x) = \sum_{i=0}^{n-1} \chi_E(T^i x) \) be the number of visits to \( E \) up to time \( n \), and let \( A_n(x) = S_n(x)/n \) be the average amount of time spent in \( E \).

**Theorem 2.30** \( A_n(x) \) converges almost everywhere.

**Proof.** Let \( \overline{A}(x) = \limsup A_n(x) \); our goal is to show that

\[
\int \overline{A}(x) \, d\mu(x) \leq \mu(E).
\]

The same argument will show the average of the liminf is at least \( \mu(E) \), so the limsup and liminf agree a.e. and we’ll be done.

Fixing \( \epsilon > 0 \) let \( \tau(x) \) be the least \( i > 0 \) such that \( A_i(x) > \overline{A}(x) - \epsilon \). In other words it is how long we have to wait to get the average close to its limsup. Obviously \( \tau(x) \) is finite.

Now suppose the wait is bounded, i.e. \( \tau(x) < N \) for all \( x \). Then we can compute \( A_n(x) \) for \( n \gg N \) by waiting (at most \( N \) steps) until the average exceeds \( \overline{A}(x) - \epsilon \), waiting that long again, and so on. Thus for
all $n$ large enough, we have $A_n(x) \geq \overline{A}(x) - 2\epsilon$ (more precisely, $A_n(x) \geq (1 - N/n)(\overline{A}(x) - \epsilon)$). Since $\int A_n(x) = \mu(E)$ for any $n$ we are done.

Now suppose the wait is not bounded. Still, we can choose $N$ large enough that the points where $\tau(x) > N$ have measure less than $\epsilon$. Adjoin these points to $E$ to obtain $E'$, and let $A'_n(x)$ denote the average number of visits to $E'$. This time we can again compute $A'_n(x)$ by waiting at most $N$ steps for the average to exceed $\overline{A}(x) - \epsilon$, since if $\tau(x) > N$ then $x$ is already in $E'$ (and one step suffices). Thus once again, $A'_n(x) \geq \overline{A}(x) - 2\epsilon$ for all $n$ sufficiently large. Since $\int A'_n(x) = \int E' \leq \mu(E) + \epsilon$, we have shown $\int \overline{A}(x) \geq \mu(E)$ in this case as well. \hfill \qed

Ergodicity of foliations on a torus. We now return to the study of a hyperbolic toral endomorphism $T : X \rightarrow X$, $X = \mathbb{R}^2/\mathbb{Z}^2$. For convenience we assume $T$ has eigenvalues $\lambda > 1 > \lambda^{-1} > 0$. The expanding and contracting directions of $T$ determine foliations $F^u$ and $F^s$ of the torus by parallel lines. Two points $x, y$ lie in the same leaf of the stable foliation $F^s$ if and only if $d(T^n x, T^n y) \rightarrow 0$ as $n \rightarrow +\infty$. Thus the leaves of $F^s$ are those contracted by $T$. The unstable foliation has a similar characterization.

Since their leaves are contracted and expanded by $T$, these foliations have no closed leaves. In other words their slopes on $\mathbb{R}^2/\mathbb{Z}^2$ are irrational.

We say a foliation $F$ of $(X, m)$ is ergodic if every saturated measurable set $A$ satisfies $m(A) = 0$ or $m(A) = 1$. (A set is saturated if it is a union of leaves of $F$.)

**Theorem 2.31** Any irrational foliation of the torus is ergodic.

**Proof.** Consider a circle $S^1$ transverse to one of the foliations, $F$. Then the first return map of a leaf to $S^1$ induces an irrational rotation $R$ on the circle. The ergodic average of $f$ along a leaf depends only on the leaf, so it descends to an $R$-invariant function on $S^1$. Since $R$ is ergodic, this function is constant, so $F$ is ergodic. \hfill \qed

Flows on a torus. It is similar easy to see that an irrational flow on a torus is ergodic, indeed uniquely ergodic. That is, if

$H_t(x) = x + te$

on $X = \mathbb{R}^2/\mathbb{Z}^2$, and the slope of $e \in \mathbb{R}^2$ is irrational, $e \neq 0$, then for any $f \in C(X)$ the average of $f(H_t(x))$ over $t \in [-r, r]$ satisfies

$(S_r f)(x) \rightarrow \int f$
as $r \to \infty$, uniformly on $X$.

**Hopf’s proof of the ergodicity of a toral automorphism.** Here is an elegant application of the ergodic theorem and foliations.

**Theorem 2.32** Let $T \in \text{SL}_2(\mathbb{Z})$ be a hyperbolic automorphism of a torus $X$. Then $T$ is ergodic.

**Proof.** Given $f \in C(X)$ and consider the limit $F_+(x)$ of the forward ergodic averages $S_n(f, x)$. If $x_1$ and $x_2$ are on the same leaf of the contracting foliation $\mathcal{F}^s$, then $d(T^n x_1, T^n x_2) \to 0$ as $n \to \infty$, so by continuity of $f$ we have $F_+(x_1) = F_+(x_2)$ (if both exist). By Fubini’s theorem, this means $F_+$ is constant along the leaves of $\mathcal{F}^s$, in the sense that $F_+$ is constant on almost every leaf.

Similarly, $F_-$ is constant along $\mathcal{F}^u$. But the forward and backward averages agree, so $F_+ = F_-$ is constant.

Since continuous functions are dense in $L^2$, and their ergodic averages are constants, we find that all $T$-invariant functions are constant, and hence $T$ is ergodic. 

**Aside: Fubini foiled for non–measurable sets.** Hopf’s proof relies on Fubini’s theorem to see that a set of positive measure saturated with respect to $\mathcal{F}^u$ and $\mathcal{F}^s$ is the whole space. However, using the continuum hypothesis it is easy to construct a set $A \subset [0, 1] \times [0, 1]$ such that $A$ meets horizontal line in sets of measure zero and vertical lines in sets of measure one. Namely, let $A = \{(x, y) : x < y\}$ with respect to a well-ordering coming from a bijection $I \doteq \Omega$ where $\Omega$ is the first uncountable ordinal.

**Mixing of a toral automorphism.** We can also use foliations to give another proof that $T$ is mixing.

The idea of the proof is that for any small box $B$, the image $T^n(B)$ is just a long rectangle along one of the leaves of the expanding (unstable) foliation. Since that foliation is ergodic, $B$ becomes equidistributed, so $m(A \cap T^n B) \to m A m B$.

To make this precise let $T \in \text{SL}_2(\mathbb{Z})$ be a hyperbolic automorphism of the torus $X = \mathbb{R}^2/\mathbb{Z}^2$. For simplicity assume that the leading eigenvalue of $T$ is $\lambda > 1$, with unit eigenvector $e$. Consider the translation flow

$$H_t(x) = x + te.$$ 

The key point is that $T$ and $H_t$ together generate a solvable group; we have

$$T H_t(x) = T(x + te) = T(x) + \lambda te = H_{\lambda t} T;$$
and that the flow $H_t$ is ergodic.

Now let $Tf(x) = f(Tx)$, and let $S_r f$ denote the average of $f(H_t x)$ over $t \in [-r, r]$. Then the relation just shown implies that

$$S_r T = TS_{\lambda r},$$

as operators on $L^2(X)$. Note also that $T^* = T^{-1}$ and $H_t^* = H_{-t}$ but $S_r^* = S_r$, since the interval $[-r, r]$ is symmetric.

**Theorem 2.33** $T$ is mixing.

**Proof.** Consider $f, g \in C(X)$ with $\int f = \int g = 0$. Since $f$ is continuous, for $r$ small, the norm of $f - S_r f$ is also small. Thus:

$$\langle f, T^n g \rangle \approx \langle S_r f, T^n g \rangle = \langle f, S_r T^n g \rangle = \langle f, T^n S_{\lambda r} g \rangle = \langle T^{-n} f, S_{\lambda r} g \rangle.$$

By ergodicity of the flow $H_t$, as $n \to \infty$ we have $S_{\lambda r} g \to 0$ in $L^2(X)$, while $\|T^{-n} f\| = \|f\|$. Thus $\langle f, T^n g \rangle \to 0$, which is mixing. ■

**Remark.** The unit speed flow $H_t$ is the analogue of the horocycle flow on a hyperbolic surface, to be studied in the next section.

### 3 Geometry of the hyperbolic plane

In this section we discuss the geometry of surfaces of constant negative curvature, in preparation for the study of their dynamics.

**Geometric surfaces.** The uniformization theorem states that every simply-connected Riemann surface $X$ is isomorphic to $\hat{\mathbb{C}}, \mathbb{C} \text{or} \mathbb{H}$, where $\mathbb{H} = \{z : \text{Im}(z) > 0\}$. The same theorem holds for (most) complex orbifolds of dimension one. (We must exclude those with no universal cover, which are genus 0 and signature $(n)$ or $(n, m)$ with $1 < n < m$.)

Each of these spaces carries a conformal metric of constant curvature, which we can normalize to be 1, 0 and $-1$. These metrics are given by:

$$\frac{2|dz|}{1 + |z|^2} \text{ on } \hat{\mathbb{C}}, \frac{|dz|}{1 - |z|^2} \text{ on } \mathbb{C}, \text{ and } \frac{2|dz|}{1 - |z|^2} \text{ on } \Delta \cong \mathbb{H}.$$

The last metric is canonical, but the other two are not, i.e. the conformal automorphism groups of $\hat{\mathbb{C}}$ and $\mathbb{C}$ are larger than their isometry groups. Nevertheless:
Any Riemann surface $X$ inherits a conformal metric of constant curvature from its universal cover.

We will now consider these covering spaces and their quotient surfaces in turn.

**Spherical geometry.** The Riemann sphere $\hat{\mathbb{C}}$ and the unit sphere $S^2 \subset \mathbb{R}^3$ can be naturally conformally identified by stereographic projection, in such way that $0$ and $\infty$ are the south and north poles of $S^2$, and $S^1 = \{z : |z| = 1\}$ is the equator. The group

$$\text{SL}_2(\mathbb{C}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : ad - bc = 1 \right\}$$

acts on $\hat{\mathbb{C}} \cong \mathbb{C}P^1$ by Möbius transformations. The action is not quite faithful, but it is often convenient to take the action as it comes rather than passing to $\text{PSL}_2(\mathbb{C})$, which acts faithfully.

If we think of $\hat{\mathbb{C}}$ as $\mathbb{P}\mathbb{C}^2$, then the basic Hermitian metric

$$\langle z, w \rangle = z_1 \overline{w}_1 + z_2 \overline{w}_2$$

gives rise to the spherical metric. Thus its isometry group corresponds to the subgroup preserving this form, viz.

$$\text{SU}(2) = \left\{ \begin{pmatrix} a & b \\ -\overline{b} & \overline{a} \end{pmatrix} : |a|^2 + |b|^2 = 1 \right\}.$$

Using $a$ and $b$ as coordinates, we see $\text{SU}(2) \cong S^3 \subset \mathbb{C}^2$; in particular, $\text{SU}(2)$ is simply–connected.

Alternatively, since $S^2$ is simply the unit sphere in $\mathbb{R}^3$ with the Euclidean metric, its group of orientation preserving isometries is given by

$$\text{SO}(3) = \{ A \in \text{SL}_3(\mathbb{R}) : A^tIA = I \},$$

where $I$ is the identity matrix.

**Geodesics.** The geodesics on the sphere are *great circles*. A typical geodesic, parameterized by arclength, is the equator:

$$\gamma(s) = (\sin s, \cos s, 0),$$

of total length $2\pi$. If we replace $S^2$ by $\mathbb{R}P^2$, then these circles give an honest instance of planar geometry: there is a unique line through any 2 points, two
lines meet in a single point, etc. An oriented geodesic circle $C_p$ is uniquely determined by its ‘center’ $p$, via

$$C_p = p^\perp \cap S^2.$$  

**Triangles.** In this geometry the interior angles of a triangle $T$ satisfy

$$\alpha + \beta + \gamma > \pi;$$

in fact, the angle defect satisfies

$$\alpha + \beta + \gamma - \pi = \text{area}(T).$$

(For example, the area of a quadrant — an equilateral right triangle – is $4\pi/8 = \pi/2 = 3(\pi/2) - \pi.$)

**Meaning of the inner product.** For any 2 points $p, q \in S^2$, it is clear that

$$\langle p, q \rangle = \cos d(p, q).$$

where distance is measured in the spherical metric. Similarly, the dihedral angle between two great circles satisfies

$$\langle p, q \rangle = \cos \theta(C_p, C_q)$$

For example, if $C_p$ and $C_q$ are the same circle with opposite orientations, then $q = -p$, $\theta(C_p, C_q) = \pi$ and $\langle p, q \rangle = -1$.

**Constructing a triangle with given angles.** It is easy to convince oneself, by a continuity argument, that for any angles $0 < \alpha, \beta, \gamma < \pi$ whose sum exceeds $\pi$, there exists a spherical triangle $T$ with these interior angles.

Here is a rigorous proof. We see the construct oriented great circles $C_a$, $C_b$, $C_c$ that define the sides of $T$ with a cyclic orientation. With respect to the basis $(a, b, c)$ for $\mathbb{R}^3$, the inner product would have to become

$$Q(\alpha, \beta, \gamma) = \begin{pmatrix} 1 & -\cos \alpha & -\cos \beta \\ -\cos \alpha & 1 & -\cos \gamma \\ -\cos \beta & -\cos \gamma & 1 \end{pmatrix}.$$  

(3.1)

We note that

$$\det Q(\alpha, \beta, \gamma) = 1 - \cos^2 \alpha - \cos^2 \beta - \cos^2 \gamma - 2\cos(\alpha)\cos(\beta)\cos(\gamma).$$

**Exercise.** We have $\det Q(\alpha, \beta, \gamma) > 0$ iff $\alpha + \beta + \gamma > \pi$. 
Thus in the case of excess angle, $Q$ has signature $(1,2)$ or $(3,0)$. But the first two vectors clearly span a subspace of signature $(2,0)$, so the signature is $(3,0)$. Thus there is a change of basis which sends $Q$ to the standard Euclidean metric, and sends $a,b,c$ to three points on $S^2$ defining the required circles.

**Exercise.** Give an explicit $(2,3,5)$ triangle on $S^2$, i.e. a triangle with inner angles $(\alpha,\beta,\gamma) = (\pi/2, \pi/3, \pi/5)$.

**Solution.** Let us take $a = (1,0,0)$ and $b = (0,1,0)$ to be the centers of the first two circles, so that $\theta(C_a,C_b) = \pi/2$. Then the final circle $C_c$ must satisfy $\langle c,a \rangle = -\cos \beta$ and $\langle c,b \rangle = -\cos \gamma$; thus

$$c = (-\cos \pi/3, -\cos \pi/5, \sqrt{1-\cos^2 \pi/3-\cos^2 \pi/5}).$$

To find the vertices, we observe that $C_a$ and $C_b$ meet along the intersection of $S^2$ with the line through the cross-product $a \times b$. Thus the vertices come from the 3 cross products $a \times b$, $b \times c$ and $c \times a$. Their projections to the $(x,y)$ plane are approximately $(0,0)$, $(0,0.5257)$ and $(0,0.3568)$.

**Orbifolds.** The nontrivial discrete subgroups of $SO(3)$ all have torsion, and hence the corresponding quotients $X = \Gamma \backslash S^2$ are all orbifolds. They are easily enumerated; aside from the cyclic and dihedral groups, which yield the $(n,n)$ and $(2,2,n)$ orbifolds, the 5 Platonic solids (3 up to duality) give the orbifolds $(2,3,3)$, $(2,3,4)$ and $(2,3,5)$. These are simply the compact orientable 2-dimensional orbifolds with

$$\chi(X) = 2 - 2g - n + \sum_{i=1}^{n} \frac{1}{p_i} > 0,$$

with the bad orbifolds $(p)$ and $(p,q)$, $1 < p < q$, excluded.

**Hyperbolic geometry.** We now turn to constant curvature $-1$, which will be our main concern. There are at least 5 useful models for the hyperbolic plane $\mathbb{H}$.

**I.** The first is the upper halfplane $\mathbb{H} \subset \mathbb{C}$, with the metric

$$\rho = \frac{|dz|}{y},$$

where $z = x + iy$. The group $SL_2(\mathbb{R})$ acts on $\mathbb{H}$ by Möbius transformations and gives all of its conformal symmetries. These symmetries are all isometries; thus any Riemann surface covered by $\mathbb{H}$ has a canonical hyperbolic
metric. The geodesics are circles perpendicular to the real axis. A typical geodesic parameterized by arclength is the imaginary axis:

\[ \gamma(s) = i \exp(s). \]

Another is the unit circle:

\[ \gamma(s) = \tanh s + i \sech s. \]

(Recall that \( \tanh^2(s) + \sech^2(s) = 1 \).) Note also that

\[
\| \gamma'(s) \|_\rho^2 = \frac{|\gamma'(s)|^2}{(\text{Im } \gamma(s))^2} = \cosh^2(s)(\sech^4(s) + \sech^2(s) \tanh^2(s)) = \sech^2(s) + \tanh(s) = 1.
\]

In particular, the ‘parallel’ at distance \( s \) from the imaginary axis is the line through the origin and \( \gamma(s) \), hence its angle \( \alpha \) with the real axis satisfies

\[ \cos \alpha = \tanh(s). \]

As another consequence, using the fact that \( \rho = |dz|/y \), we see that passing to the parallel at distance \( s \) from a geodesic expands distances by a factor of \( \cosh s \).

We remark that \( \mathbb{H} \) can be interpreted as the space of lattices \( L \subset \mathbb{C} \) with a chosen, oriented basis, modulo the action of \( \mathbb{C}^* \), via

\[ \tau \in \mathbb{H} \iff \mathbb{Z} \oplus \mathbb{Z}\tau \subset \mathbb{C}. \]

II. The second model is the unit disk:

\[ \mathbb{H} \cong \Delta = \{ z : |z| < 1 \} \subset \mathbb{C}. \]

This is also called the Poincaré model for hyperbolic space.

Here the symmetry group becomes

\[ \text{SU}(1, 1) = \left\{ \begin{pmatrix} a & b \\ \bar{b} & \bar{a} \end{pmatrix} : |a|^2 - |b|^2 = 1 \right\}. \]

Indeed, as a subset of \( \hat{\mathbb{C}} \cong \mathbb{P}\mathbb{C}^2 \), the unit disk is the space of positive lines for the Hermitian metric

\[ \langle z, w \rangle = z_1 \overline{w}_1 - z_2 \overline{w}_2 \]
on \( \mathbb{C}^2 \), where the slope \( z = z_2/z_1 \) of a complex line gives the complex coordinate on \( \mathbb{P}\mathbb{C}^2 \).

From the perspective of Hodge theory, this explains why \( \mathbb{H} \) is the space of marked complex tori. If \( \Sigma_1 \) is a smooth oriented surface of genus 1, then a complex structure is the same as a splitting

\[
H^1(\Sigma_1, \mathbb{C}) = H^{1,0} \oplus H^{0,1}.
\]

The intersection form gives the first space a Hermitian metric of signature \((1,1)\), and the line \(H^{1,0}\) must be positive since its classes are represented by holomorphic 1-forms.

We remark that the unit disk is also a model for \( \mathbb{CH}^1 \). It can be argued that the ‘correct’ invariant metric on \( \mathbb{CH}^n \cong B(0,1) \subset \mathbb{C}^n \) is given by

\[
\rho = \frac{|dz|}{1-|z|^2}.
\]

This metric has constant curvature \(-1\) on the totally real locus \( \mathbb{RH}^n \subset \mathbb{CH}^n \). It then has curvature \(-4\) on complex tangent planes; in particular, \( \mathbb{CH}^1 \) has curvature \(-4\).

We also note that any simply-connected region \( U \subset \mathbb{C} \), other than \( \mathbb{C} \) itself, carries a natural hyperbolic metric \( \rho_U \) provided by the Riemann mapping to \( U \). Using the Schwarz lemma and the Koebe \( 1/4 \)-theorem, one can show that \( \rho_U \) is comparable to the \( 1/d \)-metric; that is,

\[
\frac{1}{2d(z, \partial U)} \leq \rho_U(z) \leq \frac{2}{d(z, \partial U)},
\]

where \( d(\cdot, \cdot) \) is Euclidean distance. Knowing the hyperbolic metric up to a bounded factor is enough for many purposes, e.g. it allows one to define quasigeodesics, which are always a bounded distance from actual geodesics.

**III.** The third model is the *Minkowski model*. We give \( \mathbb{R}^3 \) the metric of signature \((2,1)\) with quadratic form

\[
(x, y, t)^2 = x^2 + y^2 - t^2.
\]

A model of \( \mathbb{H} \) is provided by the upper sheet of the ‘sphere of radius \( i \)’, i.e. the locus

\[
\mathcal{H} = \{ p : p \cdot p = -1 \text{ and } t > 0 \} = \{(x, y, z) : x^2 + y^2 = t^2 - 1, t > 0\}.
\]

Now the tangent space to \( \mathcal{H} \) at \( p \) is naturally identified with \( p^4 \), which is a positive-definite space; thus \( \mathcal{H} \) carries a natural metric, which is invariant under \( \text{SO}(2,1) \).
A typical geodesic in this model looks like a geodesic on the sphere, with hyperbolic functions replacing the spherical ones:
\[ \gamma(s) = (\sinh(s), 0, \cosh(s)) \]

is parameterized by arclength, since
\[ \|\gamma'(s)\|^2 = \cosh^2(s) - \sinh^2(s) = 1. \]

Every oriented geodesic corresponds to a unique point on the 1-sheeted hyperboloid

\[ \mathcal{G} = \{ p : p \cdot p = 1 \}, \]

by
\[ \gamma_p = p^\perp \cap \mathcal{H}. \]

**IV.** The fourth model is the **Klein model** \( \mathcal{B} \subset \mathbb{RP}^2 \). It is obtained by projectivizing the Minkowski model and observing that \( \mathcal{H} \) becomes the unit ball in the coordinates \([x, y, 1]\). In this model the group \( \text{SO}(2, 1) \subset \text{GL}_3(\mathbb{R}) \) acts by isometries. The geodesics are **straight lines** in \( \mathcal{B} \) and the space of **unoriented** geodesics can be identified with the Möbius band

\[ \mathcal{G}/(\pm 1) \cong \mathbb{RP}^2 - \mathcal{B}. \]

Geometrically, through any point \( p \) outside \( \mathcal{B} \) we have 2 lines tangent to \( \partial \mathcal{B} \), say \( L_1 \) and \( L_2 \), meeting it at \( q_1 \) and \( q_2 \); then the line \( q_1 q_2 \) is the geodesic \( \gamma_p \).

**V.** The fifth model is the **homogeneous space model**, \( \mathbb{H} = G/K \). Here \( G = \text{SL}_2(\mathbb{R}) \) and \( K = \text{SO}_2(\mathbb{R}) \). The identification with \( \mathbb{H} \) is transparent in the upper space picture, where \( \text{SO}_2(\mathbb{R}) \) is the stabilizer of \( z = i \).

One can think of \( \mathbb{H} \) as the space ellipses of area 1 centered at the origin in \( \mathbb{R}^2 \). Equivalently, \( \mathbb{H} \) is the space of inner products on \( \mathbb{R}^2 \) up to scale, or the space of complex structures on \( \mathbb{R}^2 \) agreeing with the standard orientation.

**Going between the models.** A direct relation between the Minkowski model and \( G \) can be obtained by considering the adjoint action of \( G \) on its Lie algebra \( \mathfrak{sl}_2\mathbb{R} \), in which the trace form \( \langle a, b \rangle = -\text{tr}(a \cdot b) \) of signature \((2, 1)\) is preserved.

To convert from Minkowski space to the Poincaré disk model \( \Delta \), think of \( \Delta \) as the southern hemisphere of \( S^2 \) (via stereographic projection). Then \( S^1_\infty \) is the equator. Under orthogonal projection to the equatorial plane, the geodesics in \( \Delta \) become the straight lines of the Klein model. In other words, the Klein model is just the southern hemisphere as seen from (an infinite distance) above.
With this perspective, it is clear that hyperbolic balls near $S^1_\infty$ in the Klein model are ellipses with major axes nearly parallel to $S^1_\infty$, becoming more eccentric as we near the circle.

**Complement: Euclidean space.** Next we explain how Euclidean space interpolates between spherical and hyperbolic geometry.

Imagine we are standing at the north pole $N$ of a very large sphere. Then spherical geometry is nearly the same as flat geometry. To keep the north pole in sight in $V = \mathbb{R}^3$, we can rescale so that $N = (0, 0, 1)$. Then our large sphere becomes the large, flat pancake determined by the equation

$$q_R(x, y, z) = \frac{x^2 + y^2}{R^2} + z^2 = 1,$$

with $R \gg 0$ the radius of the pancake. We would like to describe the limit $G$ of the group $SO(q_r)$ as $r \to \infty$. In the limit for the form $q(x, y, z) = z^2$ will be preserved. The ‘sphere’ $z^2 = 1$ has two components; let us stick to group elements which preserve each component, and hence preserve the $z$–coordinate.

Now in the dual space $V^*$ we have a dual quadratic form

$$q^*_R(x, y, z) = R^2(x^2 + y^2) + z^2.$$

Rescaling, we see see for any $A \in G = \lim_{R \to \infty} SO(q_R)$, $A$ preserves $q(x, y, z) = z^2$ and $A^t$ preserves $q^*(x, y, z) = x^2 + y^2$. The set of all such matrices, with our convention that $z$ itself is fixed in $V$, is given by

$$G = \left\{ \begin{pmatrix} a & b & u \\ c & d & v \\ 0 & 0 & 1 \end{pmatrix} : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SO(2) \right\} \subset GL(V).$$

On the plane $z = 1$, any $g$ as above acts by $(x, y, 1) \mapsto (x', y', 1)$, where

$$(x', y') = g(x, y) = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} u \\ v \end{pmatrix}.$$ 

Such $g$ are exactly the orientation–preserving isometries of the Euclidean plane.

Alternatively, and more directly, the limiting $g$ preserve (i) the limiting form $q = z^2$, and (ii) its rescaled limit $q' = x^2 + y^2$ on the plane annihilated by $q$.

**Spaces of horocycles and geodesics; the circle at infinity; $T^1\mathbb{H}$.** Next we briefly example the other homogeneous spaces for $G = \text{PSL}_2(\mathbb{R})$. 

36
Theorem 3.1  Up to coverings, $G$ has 5 homogeneous spaces, namely $T_1\mathbb{H}$, $\mathbb{H}$, $G = (S^1_\infty \times S^1_\infty - \text{diag})$, $L = \mathbb{R}^2 - \{0\}$ and $S^1_\infty$

Proof. Every connected subgroup of $G$ is conjugate to $\{1\}$, $K$, $A$, $N$, or $AN$. □

The circle at infinity $S^1_\infty$. The space $S^1_\infty = (AN)\backslash \mathbb{H}$ is the space of asymptotic equivalence classes of geodesics on $\mathbb{H}$.

In the Minkowski model, $S^1_\infty$ is the space of lines in the light cone.

The space of geodesics $\mathcal{G}$. The space $\mathcal{G}$ of all oriented geodesics in $\mathbb{H}$ can be described as

$$\mathcal{G} = (S^1_\infty \times S^1_\infty - \text{diag}) = \tilde{\mathbb{RP}}^2 - \Delta = A\backslash G.$$ 

The Möbius band $\mathbb{RP}^2 - \Delta$ is a natural model for the space of unoriented geodesics in the Klein model. Namely two points on $S^1_\infty$ determine a pair of tangent lines meeting in $\mathbb{RP}^2$.

Identifying $S^1_\infty$ with $\mathbb{R} \cup \infty$, the invariant measure on $\mathcal{G}$ is given by

$$\frac{dx \, dy}{|x - y|^2}.$$ 

In the Minkowski model, $\mathcal{G}$ can be described as the one-sheeted hyperboloid $x^2 + y^2 = t^2 + 1$ in the Minkowski model, consisting of the vectors with $\ell(v) = 1$. Under projectivization it gives the Möbius band. The geodesic corresponding to $v$ is the intersection of the plane $v^\perp$ with $\mathbb{H}$.

Upon projectivization to $\mathbb{RP}^2$, the space of geodesics $\mathcal{G}$ becomes the double cover of the Möbius band $\mathcal{G}' = \mathbb{RP}^2 - \mathbb{H}$. The points $p \in \mathcal{G}'$ correspond to unoriented geodesics $\gamma = [a, b] \subset \mathbb{H}$, such that the lines $ap$ and $bp$ are tangent to $S^1_\infty$.

The Minkowski form induces a Lorentz metric on $\mathcal{G}$, and the null geodesics through $p$ are the extensions of the lines $ap$ and $bp$. A pair of points $p, q \in \mathcal{G}$ are causally related (i.e. connected by a timelike path) iff the corresponding geodesics are disjoint.

Horocycles. A horocycle $H \subset \mathbb{H}$ is, in the Poincaré model, a circle tangent to $S^1_\infty = \partial \mathbb{H}$.

Geometrically, a horocycle through $p$ can be defined in any space by taking the limit of a sequence of spheres $S(c_n, r_n)$, where $c_n \to \infty$ and $r_n = d(p, c_n)$. In spherical geometry, there is no way for the centers to escape; and in Euclidean geometry, the horocycles are just straight lines.
But in hyperbolic geometry, the horocycles have geodesic curvature $+1$. They interpolate between spheres and parallels of geodesics.

**The space of horocycles $\mathcal{L}$.** The horocycles correspond bijectively to points in one component of the light cone: 

$$\mathcal{L} = \{ v \in \mathbb{R}^{2,1} : \langle v, v \rangle = 0, v \neq 0 \}.$$  

To each such $v$ we associate the locus 

$$H_v = \{ w \in \mathcal{H} : \langle w, v \rangle = -1 \}.$$  

Recall that for a center $c \in \mathcal{H}$ we have $\langle c, c \rangle = -1$ and the sphere of radius $r$ corresponds to the locus 

$$S(c, r) = \{ v \in \mathcal{H} : \langle v, c \rangle = -\cosh(r) \}.$$  

Now if $c_n \to \infty$ its limiting direction lies in the light cone, and if $r_n = \langle v_0, c_n \rangle$, $v_0 \in \mathcal{H}$, then 

$$\frac{c_n}{\cosh r_n} \to p \in \mathcal{L}, \quad \text{and} \quad S(c_n, r_n) \to H_p.$$  

The space $\mathcal{L} = G/N$ of all horocycles can also be described as $\mathcal{L} = \mathbb{R}^2 - \{0\}$ with the invariant measure $dx \, dy$; or as the bundle of nonzero vectors over $S^1_\infty$. The horocycle $H_v$ corresponding to a vector $v$ is the set of points in $\mathbb{H}$ from which $v$ has visual measure 1. The direction of $v$ gives the orientation of $H_v$.

The restriction of the Minkowski form gives a degenerate metric on the space of horocycles $\mathcal{L} = \mathbb{R}^2 - \{0\}$. Given $v \in \mathbb{R}^2$, the length of a vector $w$ at $v$ is $|v \wedge w|$. Since the area form on $\mathbb{R}^2$ is preserved, so is this metric.

Geometrically, if we have a family of horocycles $H_t$ with centers $c_t \in S^1_\infty$, then the rate of motion $dH_t/dt$ is equal to the visual length of the vector $dc_t/dt$ as seen from any point on $H_t$. (Note that the visual length is the same from all points.)

**Invariant measures on $\mathcal{H}$, $\mathcal{L}$ and $\mathcal{G}$.**

**Theorem 3.2** The natural $SO(2,1)$-invariant measure on the light cone is given by $dx \, dy/|t|$.

**Proof.** Let $q(x, y, t) = x^2 + y^2 - t^2$. Then $q$ is $SO(2,1)$ invariant, the light-cone $\mathcal{L}$ is a level set of $q$, and $dq \neq 0$ along $\mathcal{L}$. Therefore an invariant volume form $\omega$ on $\mathcal{L}$ is uniquely determined by the requirement that 

$$\omega \wedge dq = dV = dx \, dy \, dt$$
(since the latter is also SO(2,1)-invariant). But clearly
\[
\frac{dx
dy}{t} \wedge dq = \frac{dx
dy}{t} \wedge (2x
dx + 2y
dy + 2t
dt) = 2
dx
dy
dt,
\]
so \(\omega\) is proportional to \(dx
dy/t\).

This vector field is clearly SO(2,1)-invariant, and by duality it determines the invariant measure on \(\mathcal{H}\). Since the vector field blows up like \(1/t\) near the origin, so does the invariant volume form on \(\mathcal{H}\).

Yet another point of view: for any level set \(L = \{v : q(v) = r\}\), consider the band \(L_\epsilon = \{v : |q(v) - r| < \epsilon\}\). Then \(dV\mid_{L_\epsilon}\) is SO(2,1) invariant, and as \(\epsilon \to 0\), it converges (after rescaling) to \(dV/dq\). This yields:

**Theorem 3.3** The SO(2,1)-invariant measures on \(\mathcal{H}\), \(\mathcal{L}\) and \(\mathcal{G}\) are given by
\[
\omega = \frac{dV}{dq} = \frac{dx
dy}{\sqrt{x^2 + y^2 - r}},
\]
where \(r = -1, 0, 1\) respectively.

**Corollary 3.4** The invariant measure of the Euclidean ball \(B(0, R)\) intersected with \(\mathcal{H}\), \(\mathcal{L}\) or \(\mathcal{G}\) is asymptotic to \(CR\) for \(R \gg 0\).

**Proof.** The measure of a ball \(B(0, R)\) in the plane with respect to the measure \(dx
dy/\sqrt{x^2 + y^2}\) grows like \(R\).

**Remark.** By the same reasoning, the invariant measure on \(\mathcal{H}\) for SO\((n,1)\) is \(dx_1
dx_2\ldots
dx_n/t\), and the measure of \(B(0, R) \cap \mathcal{H}\), \(B(0, R) \cap \mathbb{H}^n\) and \(B(0, r) \cap \mathcal{G}\) all grow like \(R^{n-1}\). (For \(n = 1\) they grow like log \(R\).)

These estimates are useful for Diophantine counting problems related to quadratic forms.

**Hyperbolic trigonometry.** We now return to hyperbolic geometry and describe how to do actual computations.

If \(p, q \in \mathcal{H}\), so \(p^2 = q^2 = -1\), then we have
\[
\langle p, q \rangle = -\cosh(d(p, q)).
\]

To check this, just note that
\[
\langle \gamma(0), \gamma(s) \rangle = -\cosh(s).
\]
On the other hand, if \( p, q \in G \), then the angle \( \theta \) between the oriented lines they determine satisfies
\[
\cos(\theta(\gamma_p, \gamma_q)) = \langle p, q \rangle.
\]
To see this, consider the case where \( p, q \in G \) lie on the circle \( t = 0 \), so \( x^2 + y^2 = 1 \); then \( \langle p, q \rangle \) is just the usual inner product in the \((x, y)\) plane.

What if the lines don’t meet? Then we have
\[
\langle p, q \rangle = \pm \cosh d(\gamma_p, \gamma_q),
\]
where the sign is determined by the orientations. Finally for \( p \in H \) and \( q \in G \), we have
\[
\langle p, q \rangle = \pm \sinh d(p, \gamma_q).
\]
This can be checked by letting \( p = (\sinh s, 0, \cosh s) \) and \( q = (1, 0, 0) \).

**Constructing a triangle with given angles.** Consider again the matrix \( Q(\alpha, \beta, \gamma) \) defined in equation (3.1), but now with \( 0 \leq \alpha + \beta + \gamma < \pi \). Then \( \det Q < 0 \), and again there is an evident plane of signature \((2, 0)\), so we can conclude that \( Q \) has signature \((2, 1)\). Changing coordinates to obtain the standard quadratic form of this signature, the original basis elements become points of \( G \) defining the 3 sides of a triangle with inner angles \( \alpha, \beta \) and \( \gamma \).

**The \((2, 3, 7)\) triangle.** One of the most significant triangles in hyperbolic geometry is the one with inner angles \( \pi/2, \pi/3, \) and \( \pi/7 \). Reflections in the sides of this triangle give the discrete group \( \Gamma \subset \text{Isom}(\mathbb{H}) \) of minimal covolume.

Let us construct this triangle concretely in the Poincaré model, \( \Delta \). We can assume two sides are given by the lines \( x = 0 \) and \( y = 0 \). We just need to compute the center and radius of the circle \( C \) defining the third side.

Now the first 2 sides correspond, in the Minkowski model, to the planes normal to \( c_1 = (1, 0, 0) \) and \( c_2 = (0, 1, 0) \). The third circle should be normal to \( c_3 = (x, y, t) \), where \( c_3 \) satisfies
\[
\langle c_3, c_i \rangle_{i=1}^3 = (\cos \pi/3, \cos \pi/7, 1).
\]
This gives immediately
\[
c_3 = (\cos \pi/3, \cos \pi/7, t)
\]
where
\[
1 + t^2 = \cos^2 \pi/3 + \cos^2 \pi/7.
\]
Projectivizing, we find $C$ should be centered at $z = x + iy$ where

$$(x, y) = \left( \frac{\cos \pi/3}{t}, \frac{\cos \pi/7}{t} \right).$$

The point $z$ lies outside the unit disk, and the radius $r$ of the corresponding circle can be represented by a segment along a line tangent to $\partial \Delta$, so we find

$$1 + r^2 = |z|^2,$$

and hence $r = 1/t$. In the case at hand, we find

$$c = (2.01219, 3.62585), r = 4.02438.$$

In general, any point $p = (x, y, t)$ with $p^2 = 1$ corresponds to a circle with center $c = (x/t, y/t)$ and radius $r = 1/t$.

**Tessellation.** By similar calculations, it is easy to find circles whose reflections generate tessellations of $\mathbb{H}$. The examples coming from a triangle with $0^\circ$ angles, a square with $45^\circ$ angles, and a pentagon with $90^\circ$ angles are shown in Figure 1.

![Figure 1. Groups and Circles.](image)

**Constructing a pair of pants.** By the same token, we can construct a pair of pants with cuffs (and waist) of arbitrary length. That is, we can find 3 geodesics bounding a region with specified positive distances $L_1, L_2, L_3$ between consecutive geodesics. This amounts to showing that $\det Q(L_1, L_2, L_3) < 0$, where

$$Q(L_1, L_2, L_3) = \begin{pmatrix} 1 & -\cosh L_1 & -\cosh L_2 \\ -\cosh L_1 & 1 & -\cosh L_3 \\ -\cosh L_2 & -\cosh L_3 & 1 \end{pmatrix}.$$
These are the basic building blocks of compact hyperbolic surfaces, and the origin of Fenchel–Nielsen coordinates on Teichmüller space.

**Examples of hyperbolic surfaces.** It is easy to give concrete, finite-volume examples of hyperbolic surfaces using arithmetic. Start with the group $\Gamma(1) = \text{SL}_2(\mathbb{Z})$; its quotient $X(1) = \mathbb{H}/\Gamma(1)$ is the $(2,3,\infty)$ orbifold. By the ‘highest point’ algorithm, a fundamental domain for $\text{SL}_2(\mathbb{Z})$ is given by the region

$$F = \{ z : |\text{Re}(z)| \leq 1/2 \text{ and } |z| \geq 1 \}.$$ 

Clearly

$$F \subset F' = \{ z : |\text{Re}(z)| \leq 1/2 \text{ and } |\text{Im}(z)| > \sqrt{3}/2 \},$$

and it is easy to see that the integral of $|dz|^2/y^2$ over $F'$ is finite, so $\mathbb{H}/\text{SL}_2(\mathbb{Z})$ has finite volume. The set $F'$ is a substitute for a strict fundamental domain, called a *Siegel set*. It has the property that every orbit meets $F'$ at least once and at most $N$ times, for some $N$. For such a set, finiteness of the area of $F'$ is equivalent to finiteness of the volume of the quotient space.

It is a general fact that an arithmetic quotient of a semisimple group, $G(\mathbb{R})/G(\mathbb{Z})$, has finite volume. This was proved by Borel and Harish–Chandra, using Siegel sets. Borel once said he spent a good part of his career figuring out how to never compute a fundamental domain.

For more examples one can consider the covers $X(n) = \mathbb{H}/\Gamma(n)$ of $X(1)$, defined using the congruence subgroup

$$\Gamma(n) = \{ A \in \Gamma(1) : A \equiv \text{id} \mod n \}.$$ 

In general the symmetry group of $X(n)/X(1)$ is given by

$$G(n) = \text{PSL}_2(\mathbb{Z}/n),$$

which has order 6 when $n = 2$ (since $I = -I$ in this case). When $n = p > 2$ is prime, we have

$$|G(p)| = \frac{(p^2 - 1)(p^2 - p)}{2(p - 1)} = \frac{p(p^2 - 1)}{2},$$

and the number of cusps is $(p^2 - 1)/2$. For example:

1. $X(2)$ is the triply-punctured sphere, with symmetry group $S_3$;
2. $X(3)$ is a tetrahedron with its 4 face centers removed, and symmetry group $A_4$;
3. $X(4)$ is a cube with its 6 face centers removed, and symmetry group $S_4$; and

4. $X(5)$ is a dodecahedron with its 12 face centers removed, and symmetry group $A_5$. But we can continue past the Platonic solids; e.g.

5. $X(7)$ is the Klein surface of genus 3, built out of 24 heptagons the face centers removed, and a symmetry group of order 168.

**Constructions of compact surfaces.**

1. *Polygons.* A surface of genus two can be built from a regular octagon with internal angles of 45°, and similar for a surface of genus $g > 2$.

2. *Pairs of pants.* Surfaces can also be built by gluing together pairs of pants. All finite volume surfaces can be obtained in this way.

3. *Orbifolds.* One can construct finite volume orbifolds and then take finite covers to get compact surfaces. The smallest volume hyperbolic orbifold is the $(2, 3, 7)$ orbifold, with Euler characteristic $1 - (1/2) - (1/3) - (1/7) = 1/42$. As a complex space, it is obtained by compactifying $X(7)$. It is covered by the Klein quartic (a canonical curve in $\mathbb{P}^2$), which has genus 2, Euler characteristic 4 and hence deck group of order 168 as mentioned above. Since $X/\text{Aut}(X)$ is a hyperbolic orbifold for any compact $X$, the symmetry group of a hyperbolic Riemann surface satisfies

$$|\text{Aut}(X)| \leq 42(2g - 2) = 84(g - 1).$$

Equality is only obtained when $X$ covers the $(2, 3, 7)$ orbifold. For example, equality is not obtained for genus 2; in this case, the largest possible automorphism group is of order 48 (and it is related to $X(4)$).

**Finite volume surfaces; structure of the thin part.** So we can have an image in mind for noncompact, finite–volume surfaces, we mention the basic structure theorem regarding the thin part of a hyperbolic surface, $X_\epsilon \subset X$. This is the set where the injectivity radius of $X$ is less than $\epsilon$. If $X = \mathbb{H}/\Gamma$, then

$$\widetilde{X}_\epsilon = \{ x \in \mathbb{H} : \inf_{\gamma \neq \text{id}} d(x, \gamma x) \leq 2\epsilon \}.$$

We then have:

**Proposition 3.5** There exists an $\epsilon > 0$ such that for all hyperbolic surfaces, every component $U$ of $X_\epsilon$ is either a collar neighborhood of a short geodesic, or a horoball neighborhood of a cusp.
Caveats. In particular, we have $\pi_1(U) \cong \mathbb{Z}$, and $\text{area}(U) > C(\epsilon > 0)$. Thus if $X$ has finite volume, its thin part has only finitely many components.

Sketch of the proof. Let $X = \Gamma \backslash \mathbb{H}$. Suppose $d(x, gx) < \epsilon$ and $d(x, hx) < \epsilon$, with neither $g$ nor $h$ equal to the identity. Then both $g$ and $h$ are close to the identity in $G$; any significant twist would make them elliptic.

Suppose that the discrete group $(g, h)$ is not cyclic. Then the centralizers $Z(g)$ and $Z(h)$ meet only at the identity. Consequently, the commutator $g_1 = [h, g] = hgh^{-1}g^{-1}$ is even closer to the identity; on the order of $\epsilon^2$. It can belong to the centralizer of $g$ or $h$, but not both; taking the commutator repeatedly, we obtain a sequence of elements $g_n \in \Gamma$, $g_n \to \text{id}$, $g_n \neq \text{id}$, contradicting discreteness.

Thus for each point $x \in \tilde{X}$, there a unique maximal cyclic subgroup $\langle g \rangle \subset \Gamma$, generated by a group element that translates $x$ a small distance. The components associated to different $g$ are disjoint, and hence the quotient of each such component by its cyclic group injects into $X$.

Example. What is the fractional part of $\pi$? An extreme example, for the case of a cusp, is provided by the triply-punctured sphere $X = \mathbb{H}/\Gamma(2)$. The balls $B(p/q)$ resting on $z = p/q$ with height $1/q^2$, together with $B(1/0) = \{z : \text{Im}(z) \geq 1\}$, form a symmetric packing by maximal horoballs. They descend to horoballs $H_1, H_2, H_3$ on $X$, with 2 complementary triangular regions, $T_1$ and $T_2$.

The stabilizer of $B(1/0)$ is generated by $z \mapsto z + 2$, so the length of $\partial H_i = 2$. Thus the area of $H_i$ is 2 for each $i$. By Gauss–Bonnet, the area of $X$ is $2\pi = 6.28318\ldots$. Hence we have

$$\text{area}(T_i) = (2\pi - 6)/2 = \{\pi\} = 0.14159265\ldots$$

This example is extreme, in the following sense:

*For any hyperbolic surface $X$, there exist disjoint horoball neighborhoods $B_i$ of the cusps of $X$ such that the length of $\partial B_i$ is 2 for all $i$.*

4 Arithmetic hyperbolic surfaces

In this section we discuss connections between hyperbolic surfaces and the arithmetic of quadratic forms. Algebraic groups provide a source of concrete examples of lattice like $\text{SL}_2(\mathbb{Z})$, and also cocompact lattices as we will
present below. The cocompactness comes from Mahler’s general compactness criterion for families of lattices.

**Mahler’s compactness criterion.** Let \( \mathcal{L}_n = \text{SL}_n(\mathbb{R})/\text{SL}_n(\mathbb{Z}) \). We can think of \( \mathcal{L}_n \) as the space of (unmarked) unimodular lattices \( \Lambda \subset \mathbb{R}^n \), i.e. discrete subgroups isomorphic to \( \mathbb{Z}^n \) of covolume one. Alternatively, \( \mathcal{L}_n \) is the moduli space of baseframed, \( n \)-dimensional flat tori \( T \), normalized to have volume 1. The correspondence is by \( \Lambda \mapsto \mathbb{R}^n/\Lambda \).

Let \( L(\Lambda) = \inf\{|v| : v \in \Lambda, v \neq 0\} \). The injectivity radius of \( T = \mathbb{R}^n/\Lambda \) is \( L(\Lambda)/2 \).

**Theorem 4.1 (Mahler)** For any \( r > 0 \), the set of \( \Lambda \in \mathcal{L}_n \) such that \( L(\Lambda) \geq r \) is compact.

**Proof.** Choose a basis \((v_1, \ldots, v_n)\) for \( \Lambda \) with sup \( |v_i| \) minimized. Then \( |v_i| \geq r \) and \( |v_i - v_j| \geq r \). On the other hand, since the volume of \( T \) is one, a lower bound on its injectivity radius gives an upper bound on its diameter. Thus \( |v_i| \leq D(r) \). Thus the basis \((v_i)\) ranges in a compact set, and any limit of such \((v_i)\)’s is still linearly independent, so we obtain a compact subset of \( M_n \).

See §16 and its sequels for more on lattices in \( \mathbb{R}^n \).

**Arithmetic examples of compact hyperbolic spaces: dimension one.** Consider the form

\[
q(x, t) = x^2 - Dt^2,
\]

with \( D \) square free, e.g. \( D = 2 \). Equivalently, assume \( D \) is chosen so \( q \) does not represent zero.

**Theorem 4.2** The group \( \Gamma = \text{SO}(q, \mathbb{Z}) \subset \text{SO}(q, \mathbb{R}) \) is cocompact. In fact \( \Gamma \backslash \mathbb{H} \) is a circle.

**Idea of the proof.** Consider \( \text{SO}(q, \mathbb{R})/\text{SO}(q, \mathbb{R}) \) inside \( \mathcal{L}_2 = \text{SL}_2(\mathbb{R})/\text{SL}_2(\mathbb{Z}) \) and apply Mähler’s compactness criterion. The case of surfaces will be treated below.

**Orbit counting in dimension one.**

**Theorem 4.3** For any \( n \neq 0 \),

\[
N(R) = |\{(x, t) \in \mathbb{Z}^2 : q(x, t) = n\}| \sim C \log R
\]

for some \( C > 0 \).
Proof. The group $\Gamma = \text{SO}(q, \mathbb{R})$ acts discretely, by orientation-preserving isometries on $\mathbb{H}^1$, so its action is cyclic and $X = \mathbb{H}^1/\Gamma$ is a compact manifold (a circle). The integral solutions to $q(x, t) = 1$ descend to a discrete, hence finite subset of $X$. Thus the number of solutions $N(R)$ meeting the Euclidean ball $B(R)$ of radius $R$ grows like $\text{vol}(B(R) \cap \mathbb{H}^1)$ for the invariant volume form. But the invariant volume, projected down to the $x$-line, is proportional to $dx/x$, so we get $N(R) \sim C \log R$.

Fundamental units. The quadratic form $q(x, t)$ can be thought of as the norm in the field $K = \mathbb{Q}(\sqrt{D})$, since $(x + \sqrt{D}y)(x - \sqrt{D}y) = x^2 - Dy^2$. The units $U$ (elements of norm 1) in $K$ form a group under multiplication, and in fact we have $U \cong \text{SO}(q, \mathbb{Z})$. The solutions $S$ to $q(x, t) = n$ are just the integers of norm $n$, which fall into finitely many orbits under the action of the units.

The case $q(x, t) = 1$ is special — in this case $U = S$ (up to $\pm 1$) and there is only one orbit.

For example, $3 - 2 \cdot 2^2 = 1$ gives a unit $\omega = 3 + 2\sqrt{2}$, so $\omega^2, \omega^3, \ldots$ give solutions

$$17^2 - 2 \cdot 12^2 = 289 - 2 \cdot 144 = 1,$$

$$99^2 - 2 \cdot 70^2 = 9801 - 2 \cdot 4900 = 1,$$

etcetera.

Arithmetic examples of compact hyperbolic surfaces. Consider a quadratic form

$$q(x, y, t) = x^2 + y^2 - Dt^2$$

on $\mathbb{R}^3$ with $D > 0$ an integer. Suppose $q$ does not represent zero over $\mathbb{Z}$; that is, the light cone for $q$ is disjoint from $\mathbb{Z}^3$. This is the case if $D \equiv 3 \mod 4$. Then we have:

Theorem 4.4 The quotient space $\text{SO}(q, \mathbb{R})/\text{SO}(q, \mathbb{Z})$ is compact.

Since $\text{SO}(q, \mathbb{R})$ is conjugate in $\text{SL}_3(\mathbb{R})$ to $\text{SO}(2, 1)$, we obtain infinitely many examples of cocompact hyperbolic orbifolds. (Note that $\text{SO}(q, \mathbb{Z})$ is clearly discrete.) Suitable finite covers of these give compact hyperbolic surfaces.

Proof. Consider the image of $X = \text{SO}(q, \mathbb{R})/\text{SO}(q, \mathbb{Z})$ in the moduli space of flat 3-tori, $L_3 = \text{SL}_3(\mathbb{R})/\text{SL}_3(\mathbb{Z})$. Since $\text{SO}(q, \mathbb{Z}) = \text{SO}(q, \mathbb{R}) \cap \text{SL}_3(\mathbb{Z})$, the map $X \to L_3$ is an inclusion. We will show use M"ahler’s compactness criterion to show that $X$ is a compact subset of $L_3$. 

46
$X \subset M$ is the set of lattices of the form $SO(q, \mathbb{R}) \cdot \mathbb{Z}^3$. The quadratic form $q$ assumes integral values on any lattice $\Lambda \in X$, since $q(Ax) = q(x)$ for any $A \in SO(q, \mathbb{R})$. Thus $|q(x)| \geq 1$ for any $x \neq 0$ in any $\Lambda \in X$. Therefore the length of the shortest vector in $\Lambda$ is bounded below, so $X$ lies in a compact subset of $M$.

Geometrically, the hyperboloids $q(x) = 1$ and $q(x) = -1$ separate $\Lambda - \{0\}$ from the origin.

It remains to check that $X$ is closed; equivalently, that $SO(q, \mathbb{R}) \cdot SL_3(\mathbb{Z})$ is closed in $SL_3(\mathbb{R})$. This results from the fact that the set of integral quadratic forms is discrete.

In more detail, suppose $A_i B_i \to C \in SL_3(\mathbb{R})$, where $A_i \in SO(q, \mathbb{R})$ and $B_i \in SL_3(\mathbb{Z})$. Then the quadratic forms $q_i(x) = q(A_i B_i x) = q(B_i x)$ assume integral values on $\mathbb{Z}^3$ and converge to $q_\infty(x) = q(Cx)$. Since a quadratic form is determined by its values on a finite set of points, we have $q_i = q_\infty$ for all $i \gg 0$. But then $q(B_i x) = q(B_j x)$ for $i, j \gg 0$, so $q(B_i B_j^{-1} x) = q(B_j B_i^{-1} x) = q(x)$. In other words, $B_i B_j^{-1} \in SO(q, \mathbb{R})$. Thus we can absorb the difference between $B_i$ and $B_j$ into $A_i$, and assume $B_i$ is constant for all $i \gg 0$. But $A_i B_i \to C$, so $A_i$ converges in $SO(q, \mathbb{R})$ and thus $SO(q, \mathbb{R}) \cdot SL_3(\mathbb{Z})$ is closed.

**Theorem 4.5** The group $SO(q, \mathbb{Z})$ contains a torsion-free subgroup of finite index.

**Proof.** It suffices to prove the same for $SL_n(\mathbb{Z})$. But if $A \in SL_n(\mathbb{Z})$ has finite order, then its eigenvalues are roots of unity of degree at most $n$ over $\mathbb{Q}$. There are only finitely many such roots of unity (since they satisfy a monic equation of degree $n$ with integral coefficients that are symmetric functions of roots of unity, hence bounded). Thus we can find a finite set of irreducible polynomials $p_i(X)$ such that for any $A$ of finite order (greater than one), $\det p_i(A) = 0$ for some $i$. We can also assume $\det p_i(I) \neq 0$.

Let $p$ be a prime which is large enough that $\det p_i(I) \neq 0 \mod p$ for all $i$. Let $\Gamma_n(p) \subset SL_n(\mathbb{Z})$ be the finite index subgroup of matrices congruent to the identity mod $p$. Then $\Gamma_n(p)$ is torsion free. ■
**Parabolics.** For general $D$ (with $q(x)$ possibly representing zero) the construction gives a lattice (but possibly noncompact).

**Example.** The case $D = 1$ gives $\text{PSL}_2(\mathbb{Z})$. Indeed, the adjoint action of $\text{SL}_2(\mathbb{R})$ on its Lie algebra $\text{sl}_2(\mathbb{R})$ preserves the form $q(A, B) = \text{tr}(AB)$. (This trace form comes from the identification of $\mathfrak{s}(V)$ with $\text{Hom}(V, V)$ associated to the tautological representation of $\text{SL}(V)$ acting on $V$.) With respect to the basis $k = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, $a = \left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, and $b = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, the form becomes

$$q(k, a, b) = 2(a^2 + b^2 - k^2).$$

Note that the compact direction $k$ is distinguished as being time-like, while the geodesic directions $a$ and $b$ are space-like. We can think of the $\text{sl}_2(\mathbb{R})$ with the form $q$ as a picture of an infinitesimal neighborhood of the identity in $\text{SL}_2(\mathbb{R})$, with the light-cone of parabolic elements bounding two cones of elliptic elements (distinguished by direction of rotation) and a connected region of hyperbolic elements.

**Finiteness of orbits.** Here is a concrete case of a general finiteness theorem for arithmetic groups.

**Theorem 4.6** For general $D$, the integral solutions to $q(x, y, t) = -u$ fall into a finite orbits under the action of $\text{SO}(q, \mathbb{Z})$.

**Proof.** The integral points just described lie on a copy of $\mathbb{H}$ for $q$. When the quotient is compact, this Theorem is clear, since the integral points are a discrete subset of $\mathbb{H}$.

When the quotient is noncompact but finite volume, the key point is that each cusp corresponds to an integral point $v$ on the light-cone $q(v, v) = 0$. Since $q(v, w)$ assumes integral values at integral points, the horoball neighborhood $|q(v, w)| < 1$ of the cusp is devoid of integral points. So again the integral points are discrete and confined to a compact region in the quotient, and hence finite in number. 

**Beyond $\mathbb{Q}$.** The surfaces just constructed are *arithmetic*; they are also related to quaternion algebras. The construction can be generalized to yield compact arithmetic hyperbolic $n$-manifolds for every $n \geq 2$. For $n = 3$ we can still use an integral indefinite form $q$ that does not represent zero. For $n \geq 4$ it is necessary to pass to number fields, since any positive integer is a sum of 4 squares. Cf. [BP, §E.3].

**Restriction of scalars.** Finally, we describe how compact hyperbolic surfaces can be constructed using quadratic forms over more general number fields than $\mathbb{Q}$. 

48
For example, the subgroup of $\Gamma$ of $SL_3(\mathbb{Z}[\sqrt{2}])$ that preserves the quadratic form

$$q(x, y, t) = x^2 + y^2 - \sqrt{2}t^2$$

embeds in $SO(2, 1) \times SO(3)$, using the two embeddings of $\mathbb{Q}(\sqrt{2})$ into $\mathbb{R}$. The projection to the first factor gives a discrete group $\Gamma \subset SO(2, 1)$, which has no unipotent elements because the second factor is compact.

Note that $q$ form does not represent zero over $\mathbb{Z}[\sqrt{2}]$; if it did, so would its Galois conjugate $q'$, but $q'$ is positive definite.

We claim $H/\Gamma$ is compact. The proof will combine Mahler’s compactness criterion with the method of restriction of scalars.

We have a quadratic space $(L, q) = (\mathbb{Z}[\sqrt{2}]^3, q)$ defined over $K = \mathbb{Q}(\sqrt{2})$. The two real places of $K$ give a map $v \mapsto (v, v')$ sending $L \to \mathbb{R}^3 \oplus \mathbb{R}^3$ whose image is a discrete lattice. Thus $L \in X_6$. On this lattice we have a ‘quadratic form’

$$Q : L \to \mathbb{Z}[\sqrt{2}] \cong \mathbb{Z}^2$$

given by $Q(v, w) = (q(v), q'(w))$, which extends to a form $Q : \mathbb{R}^6 \to \mathbb{R}^2$.

Letting $SO(Q)$ denote the symmetries of this form, we have

$$SO(Q) \cong SO(q) \times SO(q') \cong SO(2, 1) \oplus SO(3).$$

(Note that the two summands of $\mathbb{R}^6 = \mathbb{R}^3 \oplus \mathbb{R}^3$ can be distinguished as the radicals of $q$ and $q'$ respectively, so they must be preserved by any element of $SO(Q)$.)

Finally we have $\Gamma = SO(Q) \cap SL_6(\mathbb{Z})$, and thus

$$SO(2, 1)/\Gamma \cong SO(Q) \cdot L \subset X_6.$$ 

Since $Q$ does not represent zero, the orbit $SO(Q) \cdot L$ is bounded in $X_6$.

To complete the proof we need to show this orbit is closed. So suppose $L_n \to L'$. Then for $L_n$ large enough there is an obvious isomorphism $\iota_n : L_n \to L'$ close to the identity. Since the quadratic form takes integral values, this map is an isometry for $n$ large enough. Moreover $L_n$ is $T$-invariant so the same is true for $L'$ and the isometry commutes with the action of $T$. Thus $L'$ is in the $SO(Q)$ orbit of $L$.

**General arithmetic groups.** More generally and more functorially, one can consider an order $A$ in a totally real field $K \subset \mathbb{R}$ with $n$ real places. Then using these places to give maps $A \to \mathbb{R}$, we obtain a lattice

$$A \subset A \otimes_A \mathbb{R}^n \cong \mathbb{R}^n.$$
Now given a quadratic form \( q : L \to L \), where \( L \) is a rank \( d \) \( A \)-module, we similarly have a lattice

\[ L \subset L \otimes_A \mathbb{R}^n \cong \mathbb{R}^{nd}, \]

and a quadratic form

\[ Q : \mathbb{R}^{nd} \to \mathbb{R}^n \]

obtained by tensoring \( q \) with \( \mathbb{R}^n \) over \( A \). We then have

\[ \text{SO}(Q, \mathbb{R}) \cong \text{SO}(L \otimes_A \mathbb{R}^n) \cong \prod_{i=1}^d \text{SO}(p_i, q_i), \]

where \((p_i, q_i)\) is the signature of \( q \) at the \( i \)th place of \( K \). Using \( L \) to give \( \text{SO}(Q) \) an integral structure, we then have

\[ \Gamma = \text{SO}(q, A) \cong \text{SO}(Q, \mathbb{Z}) \subset \text{SO}(Q, \mathbb{R}). \]

It is then a general factor that \( \Gamma \) is a lattice in \( \text{SO}(Q, \mathbb{R}) \); so if \( q_i = 0 \) for all \( i > 1 \), by projection we get a lattice in \( \text{SO}(p_1, q_1) \).

**Parabolics and representations of zero.** In the construction above, the fact that \( q \) did not represent zero played an important role in the cocompactness of \( \Gamma \). Conversely, if \( q \) does represent zero, then we can find parabolic elements in \( \Gamma \); cf. Artin’s *Geometric Algebra*.

## 5 Dynamics on a hyperbolic surfaces

In this section we study the ergodic theory of the geodesic and horocycle flows on a hyperbolic surface. In particular we show these flows are mixing, and that the horocycle flow on a compact surface is minimal and uniquely ergodic. The latter statement is the simplest instance of Ratner’s theorem.

We begin by describing the geodesic, horocyclic and elliptic flows on the unit tangent bundle of a hyperbolic surface.

Let \( G = \text{Isom}^+(\mathbb{H}) \cong \text{PSL}_2(\mathbb{R}) \), and let \( K \subset G \) be the stabilizer of \( z = i \) in the upper halfplane model. Then we can identify \( \mathbb{H} \) with the homogeneous space \( G/K \), with \( G \) acting isometrically on the left. Then \( G \) itself can be identified with the unit tangent bundle \( T_1 \mathbb{H} \), if we take a vertical unit vector at \( z = i \) as the basepoint in \( T_1 \mathbb{H} \).

A *hyperbolic surface* \( Y \) is a complete 2-manifold of constant curvature \(-1\). Equivalently, \( Y = \Gamma \backslash \mathbb{H} \) where \( \Gamma \subset G \) is a discrete, torsion-free group.
Let $X = T_1^1 Y = T_1^1 (\Gamma \setminus \mathbb{H}) = \Gamma \setminus G$ be the unit tangent bundle to a hyperbolic surface $Y$.

The *geodesic flow* $g^t : X \to X$ is defined by moving distance $t$ along the geodesic through $x$. The *horocycle flow* $h^s : X \to X$ is defined by moving distance $s$ along the horocycle perpendicular to $x$ with $x$ pointing inward. (This means the horocycle rests on the positive endpoint of the geodesic through $x$.) Finally the *elliptic flow* $e^r : X \to X$ is defined by rotating $x$ through angle $r$ in its tangent space.

All three flows preserve the Liouville measure on $X$, which is the product of area measure in the base and angular measure in the fiber. Cf. [Ghys] for a brief introduction.

**Theorem 5.1** The geodesic, horocyclic and elliptic flows on a hyperbolic surface $Y = \mathbb{H}/\Gamma$ correspond to the 1-parameter subgroups

$$A = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix},$$

$$N = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}, \text{ and}$$

$$K = \begin{pmatrix} \cos r/2 & \sin r/2 \\ -\sin r/2 & \cos r/2 \end{pmatrix}$$

in $G = \text{PSL}_2 \mathbb{R}$, acting on the right on $X = T_1^1 Y = \Gamma \setminus G$.

**Proof.** Note that $K$ is the stabilizer of $z = i$, consistent with our identification of $\mathbb{H}$ with $G/K$. By their definitions, the flows $g_t$, $h_s$ and $e_r$ on $T_1^1 \mathbb{H}$ commute with the action of isometries of $\mathbb{H}$. Similarly, the right and left actions of $G$ on $G = T_1^1 \mathbb{H}$ also commute. Thus it suffices to check that the action of each 1-parameter subgroups above is correct at the basepoint $z = i$, corresponding to the identity element in the identification $G = T_1^1 \mathbb{H}$.

But the right and left actions agree at the identity element, so it suffices to verify the Theorem for the *isometric* (or left) action of each subgroup on $v$, where $v$ is the basepoint in $T_1^1 \mathbb{H}$, a vertical vector at $z = i$. This is clear:

(a) We have $g_t(z) = e^t z$ for $g_t \in A$, giving translation along the vertical geodesic.

(b) We have $h_s(z) = z + s$ for $h_s \in N$, giving translation along the horocycle $\text{Im} z = 1$.

(c) Finally $e_r \in K$ stabilizes $z$ and translates $v$ through angle $r$. $\blacksquare$
Conjugacy classes in $G$. Every element $g$ of $G - \{\text{id}\}$ is conjugate into exactly one of $K$, $A$ or $N$. These cases are characterized by $|\text{tr}(g)| < 2$, $> 2$ and $= 2$ respectively. The can be classified by looking at the fixed-points of $g$ in $\mathbb{H} \cup S_{1\infty}$.

Translation length. The classification of conjugacy classes is conveniently carried out by studying the translation length $T(g) = \inf_{x \in \mathbb{H}} d(x, gx)$.

The right hand side is a convex function of $x$. There are 3 cases:

- $T(g) = 0$, realized. Then $g$ has a fixed point, so it is elliptic.
- $T(g) = 0$, not realized. Consider any limit point $p \in S_{1\infty}$ of a sequence of points $x_n$ with $d(x_n, gx_n) \to 0$. Then it is easy to see $g(p) = p$ and $g'(p) = 1$ so $g$ is parabolic.
- $T(g) > 0$. In this case the geodesic from $x$ to $gx$ must be preserved at any point where $T(g)$ is realized, so $g$ is hyperbolic.

Renormalization. The horocycle flow can be obtained from either the geodesic flow or the elliptic flow by renormalization. That is, $h_k$ is a limit of rotations around centers $z_n \to \infty$ in $\mathbb{H}$; and also a limit of translations along geodesic $\gamma_n \to \infty$. In both cases one must slow the approximating flows down so they maintain unit speed at the basepoint.

Alternatively, consider the unit speed flow $F_k$ along curves of constant geodesic curvature $k$. Then $F_0$ is the geodesic flow; $F_{1-\epsilon}$ is the flow along distant parallels of geodesics; $F_1$ is the horocycle flow; $F_{1+\epsilon}$ is the flow along circles with distance centers; and the elliptic flow arises in the limit as $k \to \infty$. (Time rescaling is necessary for this last to exist.)

The polar decomposition $G = KAK$. Given two vectors $x, y \in T_1\mathbb{H}$, rotate the first so it points along the geodesic to the second; then apply the geodesic flow, then rotate once more. This decomposition is not quite unique (since $K = KK$).

The Iwasawa decomposition $G = KAN$. Draw the horocycle $H$ normal to $x$, and then find the geodesic normal to $H$ passing through $y$; finally rotate, upon arrival, so the image vector lines up with $y$. This decomposition is unique.

The Bruhat decomposition $G = N^tAN$. Two vectors determine 2 oriented horocycles, with a geodesic joining the nearest points on both. Note that we may have to flow backwards!
Closed geodesics and horocycles. Let $X$ be a compact hyperbolic surface. Every nontrivial homotopy class contains a unique closed geodesic; there are no closed horocycles.

If $X$ has finite volume, then the homotopy classes around cusps are represented by families of parallel horocycles, and by no geodesics.

Basic relations between the geodesic, horocyclic and elliptic flows. First, note that $h_s(x)$ and $g_t(x)$ converge to the same point on $S^1_{\infty}$. However the vector $x$ is rotated by almost $\pi$ when it is moved along the horocycle. Imagine a vector $x$ in $\mathbb{H}$, resting on the horizontal horocycle $\{z : \text{Im } z = 1\}$. Drawing the geodesic between $x$ and $h^s(x)$, we find

$$h^s(x) = e^{r+\pi}g^r(x),$$

where $t(s) \to \infty$ and $r(s) \to 0$ as $s \to \infty$.

Second, note that the inward parallel to a horocycle at distance $t$ is contracted in length by the factor $\exp(t)$. This implies

$$g_t h^{\exp(t)s}(x) = h^s g_t(x).$$

The second equation has an exact analogue for a toral automorphism, while the first does not. The second equation reflects the Anosov nature of the geodesic flow.

Theorem 5.2 The geodesic and horocycle flows on a hyperbolic surface of finite area are ergodic and mixing.

The proof will be in steps: a) $g^t$ is ergodic; b) $h^s$ is ergodic; c) $g^t$ is mixing; d) $h^s$ is mixing.

Ergodicity of the geodesic flow. Following Hopf’s argument for the ergodicity of a toral endomorphism, we first consider a compactly supported continuous function $f$ on $T_1Y$. By density in $L^2$, it suffices to show the ergodic average

$$F(x) = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(g^t x) dt$$

is constant a.e. for all such $f$.

Now we can study this average as a function of $x \in T_1(\mathbb{H})$ by lifting to the universal cover. Clearly $F(x)$ only depends on the geodesic $\gamma$ through $x$; this geodesic can be labeled $\gamma(a,b)$ for $a,b \in S^1_{\infty}$, so $F$ becomes a function $F(a,b)$.

Any two geodesics with the same forward endpoint are asymptotic, so $F(a,b)$ is independent of $a$ (when it exists). On the other hand, the ergodic
averages as \( t \to -\infty \) agree with the positive time ones a.e., so \( F(a,b) \) is also independent of \( b \). Therefore \( F \) is constant a.e., and thus the geodesic flow is ergodic.

Put differently, the space of oriented geodesics \( \mathcal{G} = (S^1_\infty \times S^1_\infty) - \text{diag} \) has an obvious pair of foliations coming from the product structure. The Hopf argument shows the ergodic averages are constant a.e. along almost every leaf, so \( F \) is constant.

One can also describe this proof in terms of three 1-dimensional foliations on \( T_1(\mathbb{H}) \), namely \( \mathcal{F}^u, \mathcal{F}^s \) and \( \mathcal{F}^c \) (the unstable and stable horocycle foliations, and the central or geodesic foliation).

**Ergodicity of the horocycle flow.** Now let \( f \in L^2(X) \) be invariant under the horocycle flow and of mean zero; we will show \( f = 0 \). Let \( G^t, H^s \) and \( E^r \) denote the operators on \( L^2(X) \) corresponding to the various flows. By equation (5.1), we have

\[
H^s = E^r G^t E^{\pi+r}
\]

where \( r \to 0 \) as \( t \to \infty \). Since \( H^s f = f \), we have for any \( T > 0 \),

\[
f = \frac{1}{T} \int_0^T E^r G^t E^{\pi+r} f \, dt.
\]

Thus for any \( g \in L^2(X) \) we have

\[
\langle g, f \rangle = \lim_{T \to \infty} \langle E^{-r} g, \int_0^T G^t E^{\pi+r} f \, dt \rangle.
\]

Since \( E^{\pi+r} f \to E^\pi f \) as \( t \to \infty \), the right term in the inner product converges in \( L^2(X) \) to zero, by ergodicity of the geodesic flow and von Neumann’s ergodic theorem. The left term converges to \( g \). Therefore \( \langle g, f \rangle = 0 \) and so \( f = 0 \).

**Mixing of the geodesic flow.** We have seen:

\[
h^s g^t = g^t h^{\exp(t)s}.
\]

The proof of mixing of the geodesic flow now follows exactly as in the case of a toral automorphism, using this equation. That is, for \( f_0, f_1 \in C_0(X) \), we have for small \( s \),

\[
\langle f_0, g^t f_1 \rangle \approx \langle h^{-s} f_0, g^t f_1 \rangle = \langle f_0, h^s g^t f_1 \rangle = \langle f_0, g^t h^{\exp(t)s} f_1 \rangle = \langle g^{-t} f_0, h^{\exp(t)s} f_1 \rangle.
\]
Fix a small $S$, and average both sides over $s \in [0, S]$. On the right-hand side we obtain, for large $t$,

$$F_t = \frac{1}{S} \int_0^S h^{\exp(t)s} f_1 \, ds \approx \int_X f_1 = \langle f_1, 1 \rangle$$

by ergodicity of $h^s$. In particular, the difference is small in $L^2$. Therefore

$$\langle f_0, g^t f_1 \rangle \approx \langle g^{-t} f_0, F_t \rangle \approx \langle g^{-t} f_0, 1 \rangle \langle f_1, 1 \rangle = \langle f_0, 1 \rangle \langle f_1, 1 \rangle$$

for large $t$, which is mixing of $g^t$.

Here is the intuitive picture of mixing of $g^t$. Consider a small box $B$ in the unit tangent bundle; $B$ is a packet of vectors all pointing in approximately the same direction, say with spread $\theta$. Then $g^t(B)$ is concentrated along an arc of length $t\theta$ along a circle of radius $t$. This circular arc approximates a horocycle; since the horocycle flow is ergodic, we obtain mixing.

**Mixing of the horocycle flow.** We use the relation $h^s = e^{t+\pi} g^t e^r$ again. From this it follows that $\int \alpha(h^s x) \beta(x) = \int \alpha(e^{t+\pi} g_t x) \beta(e^t x)$, and the latter is approximately $\int \alpha(e^s g_t x) \beta(x)$ for $s$ large. By mixing of the geodesic flow this in turn is near to $\int \alpha \beta$, so the horocycle flow is mixing.

**Mixing of the action of $G$ on $\Gamma \backslash G$.**

**Theorem 5.3** If $g_n \to \infty$, then $\langle g_n \alpha, \beta \rangle \to \langle \alpha, 1 \rangle \langle \beta, 1 \rangle$.

**Proof.** Use the KAK decomposition of $G$. If $g_n = k_n a_n k'_n \to \infty$, we can pass to a subsequence such that $k_n \to k$ and $k'_n \to k'$. Then

$$\langle k_n a_n k'_n \alpha, \beta \rangle = \langle a_n k'_n \alpha, k_n^{-1} \beta \rangle \approx \langle a_n k' \alpha, k^{-1} \beta \rangle \to \langle \alpha, 1 \rangle \langle \beta, 1 \rangle.$$

---

**Topology of the $\Gamma$-action on $S^1_\infty$.** There is a great deal of information packaged in the action of $\Gamma \cong \pi_1(X)$ on the circle at infinity. To unpack it, we will look at the action on not just the circle, but also various powers of the circle as well. These actions become less chaotic as the power increases.

**Theorem 5.4** For any finite volume surface $Y = \mathbb{H}/\Gamma$,

(a) Every $\Gamma$ orbit on $S^1_\infty$ is dense;
(b) fixed-points are dense in $S^1_\infty$;
(c) $\Gamma$ acts ergodically with respect to the Lebesgue measure class on $S^1_\infty$. 

---

55
Proof. More generally consider any closed $\Gamma$-invariant set $E \subset S^1_\infty$. Then the closed convex hull of $E$ in $\mathbb{H}$ must be all of $\mathbb{H}$, since it descends to give a convex surface of $Y$ carrying the fundamental group. Therefore $E = S^1_\infty$. This prove (a) and (b).

For (c), given an invariant set $A \subset S^1_\infty$ of positive measure, we can extend $\chi_A$ to a bounded, $\Gamma$--invariant harmonic function $u(z)$ on $\mathbb{H}$, which descends then to $Y$. When $Y$ is compact we can apply the maximum principle to conclude that $u$ is constant and hence $A$ has full measure. For the case of finite volume we first apply removable singularities to fill in $u$ over the punctures of $Y$. \hfill \blacksquare

Action on $S^1_\infty \times S^1_\infty$. We remark that there is no $\sigma$-finite $\Gamma$--invariant measure in the Lebesgue measure class on $S^1_\infty$. However, once we go the two points on the circle, there is a natural $G$--invariant measure on

$$S^1_\infty \times S^1_\infty \cong \mathbb{R} \times \mathbb{R} \cong \mathcal{G} = G/A,$$

given by $dx \, dy/(x - y)^2$. Note that we must consider pairs of distinct points to get a correspondence to oriented geodesics.

Theorem 5.5 Let $X = \mathbb{H}/\Gamma$ be a finite area surface.

(a) The action of $\Gamma$ on $\mathcal{G}$ is transitive (there exists a dense orbit).
(b) A point $x \in \mathcal{G}$ has a discrete $\Gamma$-orbit iff the geodesic $\gamma_x \subset X$ is closed or joins a pair of cusps.
(c) Closed geodesics are dense in $T_1X$.
(d) The action of $\Gamma$ is ergodic for the (infinite) invariant measure on $\mathcal{G}$.

Proof. (a) and (d) follow from ergodicity of the geodesic flow on $\Gamma \backslash G$.

For (b), we note that a discrete $\Gamma$ orbit corresponds to a locally finite configuration of geodesics in $\mathbb{H}$, and hence a properly immersed geodesic in $X$, which must either be closed or divergent in both directions.

For (c), we use the a dense geodesic $\gamma : \mathbb{R} \rightarrow T_1X$ as provided by (a). Choose $s_n \rightarrow -\infty$, $t_n \rightarrow +\infty$ such that $d(\gamma(s_n), \gamma(t_n)) \rightarrow 0$ in $T_1X$. Then for $n$ large we can close $\gamma$ by perturbing it slightly so $\gamma(s_n) = \gamma(t_n)$, and then taking the geodesic representative $\delta_n$ of this loop. The closed geodesic is very close to $\gamma(t)$ for $t$ small compared to $s_n$ and $t_n$, so closed geodesics are dense.
Note: the same argument shows geodesics joining cusps are dense when $X$ is noncompact.

\textbf{Aside: The space of $Z = \Gamma \backslash \mathcal{G}$ of geodesics on $X$ as a non-commutative space.} Note that $Z$ is a smooth manifold, with natural measure theory and so on, but not Hausdorff. Every continuous function on $Z$ is constant (because there is a dense point); every measurable function on $Z$ is constant (by ergodicity of the geodesic flow).

Traditional spaces such as manifolds can be reconstructed from commutative algebras such as $C(Y)$, $L^\infty(Y)$, etc. For $Z$ these commutative algebras are trivial. To reconstruct $Z$, we need to associate to it \textit{non-commutative} algebras.

To make an interesting algebra on $Z$, we consider sections of bundles over $Z$. The most basic bundle is the Hilbert space bundle $H \to Z$ whose fiber over $p = [\Gamma x]$ is $H_p = \ell^2(\Gamma x)$. We can then build the bundle $\mathcal{B}(H)$ of bounded operators on the fiber; this is a bundle of algebras.

Note that every element $f$ of $L^\infty(\mathcal{G})$ gives a section of this bundle, by consider the multiplication operator $f(\gamma x)$ on $L^2(\Gamma x)$. This shows the bundle $H \to Z$ is \textit{nontrivial}. Indeed, by ergodicity, any section of the \textit{trivial} Hilbert space bundle over $Z$ is constant almost everywhere.

Additional (unitary) sections of the bundle of algebras $\mathcal{B}(H)$ are obtained by considering the action of $\Gamma$ on the fibers. Together these produce a non-commutative algebra $\mathcal{A}$ that in some sense records the space $Z$ (Connes).

\textbf{Action on $S^1_\infty \times S^1_\infty \times S^1_\infty$.} Let $T$ denote the set of \textit{ordered} 3-tuples of distinct points $(a, b, c)$ on the circle at infinity. This space has two components, depending on the orientation of the triangle with vertices $(a, b, c)$. Let $T_0$ denote one of these components. Then $G$ acts simply transitively on $T_0$. In other words, we can identify $T_0$ with $T_1 \mathbb{H}$.

The identification can be made geometric by taking the geodesic $\gamma$ from $a$ to $c$, and dropping a perpendicular from $b$ to obtain a point $p \in \gamma$. Then the point $p$ together with the direction of $\gamma$ gives an element of $T_1 \mathbb{H}$.

Note that $T_0$ retracts onto the space of triples of the form $(a, \omega a, \omega^2 a)$ where $\omega^3 = 1$, i.e. it retracts onto the space of equilateral triangles, so evidently

$$\pi_1(G) = \pi_1(T_0) = \pi_1(T_1 \mathbb{H}) = \mathbb{Z}.$$ 

Now $T_0$ comes equipped with a 1-dimensional foliation obtained by varying $b$ while holding $a$ and $c$ fixed. This foliation covers the foliation of $T_1 X$ by geodesics. We have shown:
Theorem 5.6 The space of positively oriented triples of distinct points on $S^1_\infty$, modulo the action of $\pi_1(X)$, is homeomorphic to $T_1X$ with its foliation by the orbits of the geodesic flow.

What is the circle at infinity? Because of all these results, one is led to ask the question: does the action of $\pi_1(X)$ on the circle depend on the hyperbolic metric on $X$? We will see that, at least topological, it does not. Thus one would like to be able to construct the ‘circle at infinity’ canonically, in terms of topology rather than geometry. This can be done using the notion of quasi–geodesics, which tend to be stable under variation of the metric.

Quasi-geodesics. A path $\gamma(s)$ in $\mathbb{H}$, parameterized by arclength, is a quasigeodesic if $d(\gamma(s), \gamma(t)) > \epsilon |s-t|$ for all $s$ and $t$.

Theorem 5.7 Any quasigeodesic is a bounded distance from a unique geodesic.

Proof. Let $\delta_n$ be the hyperbolic geodesic joining $\gamma(-n)$ to $\gamma(n)$, and consider the $r$-neighborhood $N$ of $\gamma_n$ for $r \gg 0$. Then projection from $\partial N$ to $\delta_n$ shrinks distance by a factor of about $e^{-r}$.

Suppose $\gamma([a, b])$ with $[a, b] \subset [-n, n]$ is a maximal segment outside of $N$, with endpoints in $\partial N$. Then we can join $\gamma(a)$ to $\gamma(b)$ by running distance $r$ to $\delta_n$, and then running along the projection of $\gamma([a, b])$. The total length so obtained is $r + e^{-r} |a - b| > \epsilon |a - b|$, so we see $a$ and $b$ must be close (if we choose $r$ so $e^{-r} \ll \epsilon$).

Therefore excursions outside of $N$ have bounded length, so for another $R > r$ we have $\gamma[-n, n]$ contained in an $R$-neighborhood of $\delta_n$.

Taking a limit of the geodesics $\delta_n$ as $n \to \infty$ we obtain the theorem. ■

Coarser versions. More generally, a map $f : X \to Y$ between a pair of metric spaces is a quasi-isometry if there exists a $K > 1$ and $R > 0$ such that

$$Kd(x, x') \geq d(fx, fx') \geq d(x, x')/K - R$$

for all $x, x' \in X$. The factor of $R$ means that distances are preserved, up to a factor of $K$, whenever $x, x' \in X$ are far enough apart. The preceding argument also shows:

Theorem 5.8 The image of any quasi-isometry $f : \mathbb{Z} \to \mathbb{H}$ lies within a bounded distance of a complete geodesic.
Smoother versions. One can also control the behavior of a path by its curvature. Any path in $\mathbb{H}$ which is nearly straight is a quasi-geodesic.

**Theorem 5.9** Any curve $\gamma \subset \mathbb{H}$ with geodesic curvature bounded by $k < 1$ is a quasi-geodesic.

**Proof.** Consider the line $L_s$ orthogonal to $\gamma(s)$ at $p$. For $k < 1$ these lines are disjoint (the extreme case is a horocycle where $k = 1$). Also $(d/ds)d(L_0, L_s) > c(k)s$, where $c(k) \to 1$ as $k \to 0$. Thus $d(L_s, L_t) > c(k)|s - t|$ and this provides a lower bound for $d(\gamma(s), \gamma(t))$. □

**Theorem 5.10** A polygonal path $\gamma$ of segments of length at least $L$ and bends at most $\theta < \pi$ is a quasi-geodesic when $L$ is long enough compared to $\theta$. Also as $(L, \theta) \to (\infty, 0)$ the distance from $\gamma$ to its straightening tends to zero.

**Proof.** As before $L_s$ advances at a definite rate, indeed at a linear rate, except near the bends of $\gamma$. The size of the neighborhood of the bend that must be excluded tends to zero as $\theta \to 0$. Thus the geodesic representative of $\gamma$ is very close to $\gamma$ between the bends when $L$ is long. □

The circle at infinity. The next result shows that, as a topological space, the circle $S^1_\infty$ is canonically attached to any compact surface. There are many ways to view this result. To begin with, we will show:

**Theorem 5.11** Let $h : X \to Y$ be a diffeomorphism between closed hyperbolic surfaces. Then $\tilde{h} : \mathbb{H} \to \mathbb{H}$ extends to a homeomorphism $S^1_\infty \to S^1_\infty$ conjugating $\Gamma_X$ to $\Gamma_Y$.

**Proof.** The map $h$ is bilipschitz, so $\tilde{h}$ maps geodesics to quasigeodesics. Straightening these, we obtain a map $S^1_\infty \to S^1_\infty$. To check continuity, note that geodesics near infinity straighten to geodesics near infinity. □

Topological construction of $S^1_\infty$. The preceding result can be viewed more functorially as follows. First, associated to a closed surface $S$ is a the fundamental group $\pi_1(S)$. Choosing a finite generating set, we obtain a metric on $\pi_1(S)$ that is well–defined up to isometry. We can then functorially construct

$$\partial \pi_1(S) = \{\text{quasi-geodesics } \mathbb{N} \to \pi_1(S)\} / \sim$$
as a topological space. Here we regard two rays in $\pi_1(S)$ as equivalent if they are a bounded distance apart.

Next, we observe that the universal cover of a pointed, connected, compact manifold $(M, x)$ is quasi-isometric to $\pi_1(M, x)$. In particular, if $X = \mathbb{H}/\Gamma$ and we have an isomorphism $\pi_1(S) \cong \pi_1(X)$, then we get a natural homeomorphism

$$\partial \pi_1(S) \to S^1_\infty = \partial \mathbb{H}$$

by straightening quasigeodesics. This shows that $\partial \pi_1(S)$ is a circle and that this circle is naturally identified with $\partial \mathbb{H}$ once a marking of $X$ is chosen.

Finally if we have two marked hyperbolic surfaces, then both their circles at infinity are marked by $\partial \pi_1(S)$ and hence we obtain a homeomorphism on the circle at infinity.

**Flexibility and rigidity of foliations.** We are now ready to study the geodesic and horocycle foliations of $T_1X$, and their dependence on the hyperbolic structure on $X$. Indeed, Theorem 5.11 immediately yields:

**Theorem 5.12** If $X$ and $Y$ are homeomorphic closed surfaces, then the foliations of $T_1X$ and $T_1Y$ by orbits of the geodesic flow are topologically equivalent.

**Proof.** The actions of $\pi_1(X) \cong \pi_1(Y)$ on $S^1_\infty$ are topologically the same, and $T_1X$ together with its geodesic foliation can be reconstructed from the diagonal action on $(S^1)^3$.

Note however that we have:

**Theorem 5.13** Let $X$ and $Y$ be hyperbolic surfaces. Suppose the geodesic flows on $T_1X$ and $T_1Y$ are topologically conjugate. Then $X$ is isometric to $Y$.

**Proof.** The surfaces $X$ and $Y$ then have the same marked length spectrum, and this is enough to get an isometry.

**A glimpse of Mostow rigidity.** Here is the first inkling of rigidity for hyperbolic manifolds. The same argument will yield a much stronger conclusion in higher dimensions.

**Theorem 5.14** Let $X_i = \mathbb{H}/\Gamma_i$, $i = 1, 2$, be a pair of compact hyperbolic surfaces. Let $f : S^1_\infty \to S^1_\infty$ be a homeomorphism that sends the action of $\Gamma_1$ to the action of $\Gamma_2$.

Suppose $0 < f'(x) < \infty$ exists at one point $x \in S^1_\infty$. Then $f$ is a Möbius transformation, and $X_1$ and $X_2$ are isometric.
Proof. By assumption we have an isomorphism \( \gamma \mapsto \gamma' \in \Gamma' \) such that \( f(\gamma x) = \gamma' f(x) \) for all \( x \in S^1_{\infty} \) and \( \gamma \in \Gamma \). Choose coordinates on the upper halfplane such that \( S^1_{\infty} = \hat{\mathbb{R}} \) and \( f(0) = 0 \) and \( f'(0) = 1 \). Let \( A_n(x) = nx \). Then we have
\[
A_n f(A_n^{-1} x) \rightarrow x
\]
uniformly on \( S^1_{\infty} \), by the definition of differentiability. Since \( \Gamma G \) is compact, we can find \( \gamma_n \in \Gamma \) such that \( A_n \gamma_n' \) is bounded in \( G \). We then have
\[
A_n \gamma_n' f(\gamma_n^{-1} A_n^{-1} x) = A_n f(A_n^{-1} x) \rightarrow x.
\]
Pass to a subsequence such that \( A_n \gamma_n' \rightarrow A' \in G \). Since \( f \) is a homeomorphism, we then also have \( A_n \gamma_n \rightarrow A \in G \), and hence
\[
A' f(A^{-1} x) = x.
\]
This shows that \( f(x) \) is a Möbius transformation.

Corollary 5.15 The map \( f \) is either a Möbius transformation or it is completely singular (it sends a set of full measure to a set of measure zero).

Proof. A monotone function like \( f \) is differentiable a.e. We have just shown that \( f'(x) = 0 \) or \( \infty \) whenever it exists, so \( f \) is singular. (In general \( m(E) = \int_E f'(x) \, dx \) gives the absolutely continuous part of \( f^*(dx) \).)

Hölder continuity and quasi-symmetry. Here is a direct proof that if \( f \) is Lipschitz, then it is a Möbius transformation. Namely, for \( \lambda, \delta < 1 \), the geometric sequences \( \lambda^n \) and \( \delta^n \) are related by a Lipschitz map iff \( \lambda = \delta \). This implies that the lengths of corresponding geodesics agree.

Similarly, if we know that \( f, f^{-1} \in C^\alpha \), meaning
\[
|f(x) - f(y)| \leq M |x - y|^\alpha,
\]
then the lengths of geodesics change at most by a multiplicative factor controlled by \( \alpha \).

In fact, the mapping \( f \) is always quasi-symmetric, which implies Hölder continuity.

Deeper properties of the horocycle flow. We now turn to some properties of the horocycle flow that go beyond ergodicity and mixing. One should observe the sharp contrast with the geodesic flow in each case.
Minimality. We begin with the following important result, proved in 1936.
Every orbit of $AN$ is dense since $\Gamma$ has dense orbits on $S^1_\infty = AN \setminus G$.
We'll now see the sharper statement that the action of $N$ has only dense orbits. This means that for any $x, y \in S^1_\infty$ and $t \in \mathbb{R}$, there is a $\gamma \in \Gamma$ such that $\gamma(x) \approx y$ and $\gamma'(x) \approx t$.

**Theorem 5.16 (Hedlund)** The horocycle flow on a closed surface $X$ is minimal (every orbit is dense).

**Proof.** Let $X$ be a closed surface. Then by Zorn's lemma there exists in $T_1X$ a nonempty minimal closed set $F$ invariant under the horocycle flow. The orbit of every point in $F$ is dense in $F$. We will show $F = T_1X$.

Since $F$ is $N$-invariant, it will suffice to show $F$ is $A$-invariant (by the preceding result). Note that for any $a \in A$, $a(F) = N$-invariant (since $A$ normalizes $N$), and so by minimality if $aF \cap F \neq \emptyset$ then $aF = F$. So it suffices to show there are $a_n \to \text{id}$ in $A$ such that $a_n F$ meets $F$.

Now $X$ has no closed horocycles, so $F$ contains (uncountably) many horocycles, all dense in $F$. Consider two very nearly parallel horocycles, $H_1$ and $H_2$. In the upper halfplane model for $\mathbb{H}$ we can normalize so $H_1 = \{\text{Im } z = 1\}$ and $H_2$ is a circle of radius $R \gg 0$ resting on $\mathbb{R}$ at $z = 0$. Then in the region $\{\text{Im } z \leq 2\}$ these two horocycles are nearly parallel (in the Euclidean sense). In particular, there is a unit hyperbolic geodesic $\gamma$, normal to $H_1$ (a vertical segment), such that $\gamma$ is also almost normal to $H_2$.

By compactness of $X$, we can find two horocycles in $F$ that are joined by a common perpendicular. (This means in the universal cover $H_1$ and $H_2$ rest on the same point on $S^1_\infty$). Thus $a_1 F = F$ for some $a_1 \in A$. translating distance one. By the same argument one obtains $a_n F = F$ with $a_n \to \text{id}$. Thus $F$ is $AN$-invariant and hence $F = T_1X$.

**Remark: the horocycle flow and type $III_1$ factors.** Ergodicity of the horocycle has the following consequence in terms of the dynamics of $\Gamma$ on $S^1$: for almost every $x \in S^1$, the set

$$\{(\gamma(x), \gamma'(x)) : \gamma \in \Gamma\}$$

is dense in $S^1 \times \mathbb{R}_+$. Indeed, for a horocycle $H_x$ resting on $x \in S^1$, the image $\gamma(H_x)$ is a horocycle resting on $\gamma(x)$ with height determined by $\gamma'(x)$, so the density of this set follows from density of $H_x$ in $T_1(\mathbb{H}/\Gamma)$.

This is related to the concept of a type $III_1$ factor in von Neumann algebras: there is no $\Gamma$–invariant measure on the circle, and there is no
measure such that the Radon-Nikodym derivatives lie in a proper subgroup of $\mathbb{R}_+$.

**Unipotent renormalization.** Here is a broad perspective on the key step in Hedlund’s argument proving minimality of the horocycle flow.

One can regard a Lie group $G$ as the substrate of a geometry, following Klein’s Erlangen program. The associate geometric space is then $G/K$ for a suitable compact subgroup $K$ of $G$.

Here $K$ is the stabilizer of a point in $X$. Other subgroups $A \subset G$ label other types of geometric structures, namely those stabilized by $A$. For example, when $G = \text{PSL}_2(\mathbb{R})$ we can also consider points on $S^1$, horocycles, and geodesics as other geometric structures.

The space of structures of a given type can be identified with $G/A$, from which it acquires a topology.

One might ask, what are the possible configurations of two geometric objects of the same type? The answer is given by the quotient of $(G/A) \times (G/A)$ by the action of $G$, or equivalently the double coset space $A \backslash G/A$. For example, $K \backslash G/K$ gives the space of configurations of pairs of points in $X$. For $X = \mathbb{H}^3$, there is a unique invariant, namely the distance, and then $K \backslash G/K$ is $[0, \infty)$.

The space $A \backslash G/A$ can be visualized geometrically as the space of orbits of $A$ acting on $G/A$. Topologically, the double coset space may be poorly separated. A nice example is provided by $A = U$ in $G = \text{PSL}_2(\mathbb{R})$. Then $G/U \cong \mathbb{R}^2 - \{0\}$, and the orbits of $U$ correspond to the horizontal lines $L_c$ defined by $y = c$, $c \neq 0$, and the points on the $x$-axis. Notice that as $c \to 0$, the geometric limit of $L_c$ includes all the points on the $x$-axis, and hence there are infinitely many possibilities for the ‘limit’ of $[L_c] \in U \backslash G/U$ as $c \to 0$.

Geometrically, the points of $G/U$ correspond to horocycles in $\mathbb{H}^2$. For two typical horocycles $H_1$ and $H_2$, we have a unique geodesic joining their ‘centers’ (points in $H_i \cap S^1$), which we can use to define a ‘distance’ between them which serves as a complete invariant of the configuration $H_1 \cup H_2$. These pairs of horocycles correspond to the lines $L_c$. If, however, the horocycles have the same center, then there is a different invariant — the ratio of their radii — which describes their configuration. This invariant corresponds to the points of the $x$-axis.

The second ‘special’ invariant arises when the stabilizer of the configuration $H_1 \cup H_2$ is unusually large. Typically the stabilizer is 0-dimensional; but when $H_1$ and $H_2$ have the same center, the configuration has a 1-dimensional stabilizer.
In terms of groups (“homogeneous dynamics”), the phenomenon that the invariants of a pair of horocycles can ‘blow up’ in the limit can be described as follows.

**Theorem 5.17** Suppose \(g_n \to \text{id} \) in \(G - AU\). Then given any \(a \in A\), there exist \(u_n, u'_n \in U\) such that
\[
u_n g_n u'_n \to a.\]

An equivalent geometric statement is the following.

**Theorem 5.18** Let \(H_n \to H\) be a convergent sequence of horocycles in \(\mathbb{H}^2\), such that for all \(n\), \(H_n\) is not tangent to \(H\). Then for any \(r \in \mathbb{R}\) there exist \(u_n \in G_H\) such that
\[
u_n H_n \to H(r),\]
where \(H(r)\) is the unique horocycle parallel to \(H\) at signed distance \(r\).

**Proof.** We may assume \(H\) is the line \(y = 1\) in the upper halfplane model for \(\mathbb{H}^2\). Then \(H_n\) is a circle of radius \(r_n\) resting on the real axis, with \(r_n \to \infty\). For all \(n \gg 0\), \(H_n\) crosses the line \(y = e^r\) at a point of the form \(z_n = x_n + ie^r\). Let \(u_n(z) = z - x_n\). Then \(u_n H_n\) passes through the point \(ie^r\), and since its radius is tending to infinity, it converges to the line \(y = e^r\) as \(n \to \infty\). 

**Rigidity of the horocycle foliation.** Unlike the geodesic foliation, the topology of the horocycle foliation changes whenever we vary the metric on \(X\).

**Theorem 5.19 (Marcus, [Mrc])** Let \(X\) and \(Y\) be closed hyperbolic surfaces. If the horocycle foliations of \(T_1 X\) and \(T_1 Y\) are topologically conjugate, then \(X\) and \(Y\) are isometric.

**Sketch of the proof.** Let \(f : T_1 X \to T_1 Y\) be a homeomorphism respecting the horocycle foliations, and let \(F : T_1 \mathbb{H} \to T_1 \mathbb{H}\) be its lift to the unit tangent bundle of the universal covers of \(X\) and \(Y\). (This lift exists because \(f\) preserves the center of \(\pi_1(T_1 X)\) which corresponds to the fibers.) Then \(F\) is a quasi-isometry, it conjugates \(\Gamma_X\) to \(\Gamma_Y\), and it sends horocycles to horocycles.

Now two horocycles rest on the same point on \(S^1_\infty\) if and only if they are a bounded distance apart. Since \(F\) is a quasi-isometry, it respects this relationship between horocycles.
Moreover $F$ descends to a map between the spaces of horocycles, which is naturally $\mathbb{R}^2$ with the groups $\Gamma_X$ and $\Gamma_Y$ acting linearly. Thus we have a homeomorphism

$$f : \mathbb{R}^2 \to \mathbb{R}^2$$

such that (i) $f$ sends rays to rays and (ii) $f(\gamma v) = \gamma' f(v)$ for all $\gamma \in \Gamma_X$.

Because of (i), we have a function $T(t, v)$ on $\mathbb{R}_+ \times \mathbb{R}^2$ characterized by

$$f(tv) = T(t, v)f(v).$$

Since $f$ is continuous, so is $T(t, v)$. Moreover, since the actions of $\Gamma_X$ and $\Gamma_Y$ on $\mathbb{R}^2$ are linear, for each fixed $t$ we have $T(t, \gamma v) = T(t, v)$. Since $\Gamma$ has a dense orbit on $\mathbb{R}^2$ (e.g. by ergodicity), this implies $T(t, v)$ depends only on $t$. Then from the definition we have

$$T(st) = T(s)T(t),$$

and thus $T(t) = t^\alpha$ for some $\alpha \in \mathbb{R}_+$.

Now consider the expanding eigenvector $v$ for $\gamma \in \Gamma_X$, satisfying $\gamma v = \lambda v$ with $\lambda > 1$. It is related to the hyperbolic length of the corresponding geodesic on $X$ by $2 \log \lambda = L(\gamma)$. On the other hand, $v' = f(v)$ must be the expanding eigenvector for $\gamma'$, with eigenvalue

$$\lambda' = T(\lambda) = \lambda^\alpha.$$ 

This shows lengths on $X$ and $Y$ are related by

$$L(\gamma') = \alpha L(\gamma);$$

i.e. corresponding geodesics on $X$ and $Y$ have the proportional lengths.

But it is known that the number of geodesics with length $\ell(\gamma) \leq L$ on a compact hyperbolic surface grows like $e^{L/L}$. Therefore $\lambda = \pm 1$ and we have shown that corresponding geodesics have the same length. The length function on $\pi_1(X)$ determines the metric on $X$ uniquely, so we are done.

**Topological invariants of the horocycle foliation.** Here is a useful exercise: just given the topological action of $\Gamma$ on $\mathbb{R}^2$, and $g, h \in \Gamma$, determine if $L(g) > L(h)$. For the solution, let $\alpha$ and $\beta$ be the eigenvalues of $g$ and $h$, and observe that the eigenvectors of the conjugates of $g$ are dense in $\mathbb{R}^2$. Thus the expanding eigenvector $b$ for $g_2$ is a limit of expanding eigenvectors $a_n$ for $g_1$. We can then compare the limit of the segments $[a_n, \alpha a_n]$ to $[b, \beta b]$ and see which one contains the other.

**Unique ergodicity of the horocycle flow.** We now come to a strengthening of Hedlund’s topological result.
Theorem 5.20 (Furstenberg [Fur2]) The horocycle flow on a closed hyperbolic surface $X$ is uniquely ergodic.

We already know the horocycle flow is ergodic; now we want to show every orbit is distributed the same way that almost every orbit is.

Irrational rotations. To convey the spirit of the argument, we first reprove unique ergodicity of an irrational rotation $T : S^1 \to S^1$. Consider an interval $I = [a, b] \subset S^1$. Given $y \in S^1$, we need to show

$$A_n(y, I) = \frac{|\{i : T^i(y) \in I, 1 \leq i \leq n\}|}{n} \to m(I).$$

Fix any $\epsilon > 0$. Consider intervals $I' \supset I \supset I''$, slightly larger and smaller by $\epsilon$. By ergodicity, $A_n(x, I) \to m(I)$ for almost every $x$, and similarly for $I'$ and $I''$. Thus we can find a set of measure at least $1 - \epsilon$ and an $N$ such that $|A_n(x) - m(I)| < \epsilon$ for all $x \in E$ and $n \geq N$, and similarly for $I'$ and $I''$.

Since $m(E) > 1 - \epsilon$, there is a point $x \in E$ such that $|x - y| < \epsilon$. Then the orbits of these points are close: $|T^i x - T^i y| < \epsilon$ for all $i$. Therefore when $n > N$ we have

$$A_n(y, I) < A_n(x, I') < m(I') + \epsilon < m(I) + 2\epsilon,$$

and a reverse bound also holds by considering $I''$.

Therefore $A_n(y) \to m(I)$ and $T$ is uniquely ergodic.

Proof of unique ergodicity of the horocycle flow. (Adapted from Ratner and [Ghys, §3.3].) Through any $x$ in $T_1 X$, consider the flow box $Q(x, s_-, s_+, t)$ obtained by first applying the negative, then the positive horocycle flows for time $[-s_-, s_-]$ and $[-s_+, s_+]$ respectively, and then applying the geodesic flow for time $[-t, t]$. Since $X$ is compact, when the parameters $s_-, s_+, t$ are small enough we have $Q(x, s_-, s_+, t)$ embedded in $T_1 X$ for any $x$.

Fix one such flow box $Q_0$. Let $H(x, t) = h_{x}^{[0,t]}(x)$ denote horocycle flow line of length $t$ starting at $x$. Then by ergodicity of the horocycle flow with respect to Lebesgue measure $m(\cdot)$, we have

$$\lim_{t} \frac{\text{length}(H(x, t) \cap Q_0)}{t} = m(Q_0)$$

for almost every $x$. Moreover the limit is almost uniform: for any $\epsilon$ there exists a $T$ and an $X_0 \subset T_1 X$, $m(X_0) > 1 - \epsilon$, such that

$$\left| m(Q_0) - \frac{\text{length}(H(x, T) \cap Q_0)}{T} \right| < \epsilon$$

66
for all $x \in X_0$. We can also ensure that the same statement holds for slightly smaller and larger boxes, $Q'_0 \supset Q_0 \supset Q''_0$.

Now consider any point $y \in T_1X$. We will show
\[
\frac{1}{T}m\{t \in [0, T] : h^t_+(y) \in Q_0\} \approx m(Q_0)
\]
as well, the approximation becoming better as $T \to \infty$.

To this end, transform the whole picture by applying the geodesic flow $g^{\log T}$, which compresses the positive horocycle flow by the factor $T$. Under the geodesic flow we have
\[
g^uQ(x, s_-, s_+, t) = Q(g^u x, e^u s_-, e^{-u} s_+, t),
\]
and
\[
g^uH(x, t) = H(g^u x, e^{-u} t).
\]

Thus
\[
Q_1 = g^{\log T}Q_0 = Q(x', Ts_-, s_+/T, t),
\]
and $H(y, T)$ is transformed to a unit segment $H(y', 1)$. Finally $X_0$ is transformed to a set $X_1$, still of measure $1 - \epsilon$, such that the length of the part of $H(x, 1)$ inside $Q_1$ is almost exactly $m(Q_1) = m(Q_0)$.

Since $X_1$ almost has full measure, there is an $x \in X_1$ very close to $y'$. Now $H(x, 1)$ consists of very many short segments (of length about $s_+/T$) inside the highly flattened box $Q_1$. Moving $x$ slightly to $y'$, $H(x, 1)$ moves slightly to $H(y', 1)$. The edge effects can be ignored by the fact that $H(x, 1)$ also works for slightly larger and smaller boxes. More precisely, the length of $H(y, 1) \cap Q_1$ is bounded above by the length of $H(x, 1 + \epsilon) \cap Q'_1 \supset Q_1$, and there is a similar lower bound. Thus the lengths of $H(x, 1) \cap Q_1$ and $H(y', 1) \cap Q_1$ is almost the same, so the horocycle orbit through $y$ is equidistributed.

**Note:** In this last step we have used the fact that the injectivity radius of $X$ is bounded below to study the local picture of $Q_1$ and $H(x, 1)$.

**Slow divergence.** Here is an explanation of Ratner’s proof without applying $g^{\log T}$.

Consider an arbitrary point $y \in T_1Y$ and a set of good points $E \subset T_1Y$, of measure $1 - \epsilon$. We want to show the $h_+$-orbit, $H(y, T)$, cuts $Q_0$ in about the same length as $H(x, T)$.

Now any set $F$ with $m(F) > \epsilon$ meets $E$. The key idea is not to take $F$ to be a ball around $y$. Instead, $F$ is taken to be a long, narrow region around
the segment $H(y, S)$, where $S = \delta T$ and $\delta$ is small. The point is that we are willing to sacrifice $H(x, T)$ following $H(y, T)$ for time $2S$, as long as we can get the two segments to be close for the rest of the time.

More precisely, $F$ is taken to have dimension $\delta$ in the $g^t$ direction, and $\delta/T$ in the $h_-$-direction. The latter two conditions insure that $H(x, T)$ follows $H(y, T)$ to within $O(\delta)$. Now the total volume of $F$ is about $S \times \delta \times \delta/T = \delta^3$, which does not depend on $T$. Thus when $\epsilon$ is small enough, $F$ meets $E$ and we are done.

To make this rigorous, we have to show $F$ maps injectively into the unit tangent bundle. For this it is useful to flow by $g^{\log T}$; then the dimensions of $F$ are independent of $T$.

**The case of finite area.** We just state the results:

**Theorem 5.21 (Dani [Dani])** Any ergodic measure on $T_1 Y$, invariant under the horocycle flow, is either Liouville measure or concentrated on a single closed horocycle.

**Theorem 5.22 (Dani and Smillie [DS])** Any horocycle is either closed, or equidistributed in $T_1 Y$.

See [Dani] and [DS] for details.

**Example.** Consider $\Gamma = \text{SL}_2 \mathbb{Z}$ acting on $\mathbb{R}^2 - \{0\} = G/N$. Then the orbit of a vector $(x, y) \in \mathbb{Z}^2$ is discrete; more generally, the orbit is discrete if $x/y \in \mathbb{Q}$. Otherwise, the orbit is dense.

**Proof.** Let $\mu$ be an invariant measure and let $y$ be a point whose orbit is distributed according to $\mu$. Then the preceding proof works to show $\mu = m$, so long as $y' = g^{\log T}(y)$ is recurrent as $T \to \infty$ (this recurrence allows us to work in a fixed compact subset of $X$, where the injectivity radius is bounded below). But if the horocycle through $y$ is not recurrent, it is closed. Thus the only other ergodic invariant measures are uniform length measure supported on closed orbits.

**Theorem 5.23 (Ratner)** Let $\Gamma \subset G$ be a lattice in a connected Lie group $G$, and let $U \subset G$ be a subgroup generated by Ad-unipotent elements. Then

(a) any $U$-invariant ergodic measure is the image of Haar measure on a subgroup $H$ such that $H \cap \Gamma$ is a lattice in $H$; and

(b) the closure of any $U$-orbit is the orbit of a closed subgroup $H$ of $G$, again such that $H \cap \Gamma$ is a lattice in $H$.  

68
Moreover, if \( U \) is a 1-parameter group, then any orbit \( Ux \) is uniformly distributed with respect to the natural Haar measure on \( Ux \) given by (a) and (b).

See [Ghys], [Rn] and references therein.

<table>
<thead>
<tr>
<th></th>
<th>Dynamics on compact hyperbolic surfaces</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geodesic flow</td>
<td>Horocycle flow</td>
</tr>
<tr>
<td>Ergodic</td>
<td>Ergodic</td>
</tr>
<tr>
<td>Mixing</td>
<td>Mixing</td>
</tr>
<tr>
<td>Countably many closed orbits</td>
<td>No closed orbits</td>
</tr>
<tr>
<td>Not minimal</td>
<td>Minimal (all orbits dense)</td>
</tr>
<tr>
<td>Positive entropy</td>
<td>Zero entropy</td>
</tr>
<tr>
<td>Topology of orbits depends only on genus of ( Y )</td>
<td>Topology of orbits determines geometry of ( Y )</td>
</tr>
<tr>
<td>Many ergodic measures</td>
<td>Uniquely ergodic</td>
</tr>
</tbody>
</table>

6 Dynamics on moduli space

We briefly recount the related theory of the \( \text{SL}_2(\mathbb{R}) \) action of the bundle \( Q\mathcal{M}_g \) of quadratic differentials \((X,q)\) with \( X \in \mathcal{M}_g \) and \( q \in Q(X) \). Here \( \dim Q(X) = 3g - 3 \) by Riemann–Roch. The fact that \( 3g - 3 \) is also the maximum number of disjoint simple geodesics on a Riemann surface of genus \( g \) already hints at the close connection we will see between quadratic differentials and simple closed curves.

Any such differential can be presented in the form \((X,q) = (P,dz^2)/\sim\) with \( P \) a polygon in \( \mathbb{C} \), and the action is given by

\[
A \cdot (X,q) = (A(P),dz^2).
\]

A beautiful feature of this action is that we have a natural map

\[
Q\mathcal{M}_g \to \mathcal{ML}_g \times \mathcal{ML}_g
\]

sending \((X,q)\) to the pair of foliations \((\alpha,\beta) = (\mathcal{F}(q),\mathcal{F}(-q))\). We have \( i(\alpha,\beta) = \int_X |q| \). In these coordinates the geodesic flow acts by sending
\((\alpha, \beta)\) to \((t\alpha, (1/t)\beta)\), \(t \in \mathbb{R}_+\). Thus geodesics in unit sphere bundle of \(Q\mathcal{M}_g\) can be identified with points in

\[ \mathbb{P}\mathcal{ML}_g \times \mathbb{P}\mathcal{ML}_g. \]

Using the Hopf argument (with some added finesse, because not all geodesic rays with the same endpoint are asymptotic), we obtain:

**Theorem 6.1 (Masur,Veech)** The Teichmüller geodesic flow on \(Q\mathcal{M}_g\) is ergodic.

Here we have not explained the invariant measure; it comes from periods and is invariant under \(\text{SL}_2(\mathbb{R})\) as well. Due to the basic theory of unitary representation of \(\text{SL}_2(\mathbb{R})\) we have already discussed, an immediate corollary is:

**Theorem 6.2** The geodesic flow on \(Q\mathcal{M}_g\) is mixing. In fact the full \(\text{SL}_2(\mathbb{R})\) action is mixing.

**Curve systems.** The connection to elements of \(\mathcal{ML}_g\) can be made simple and explicit in the case of differentials such that both \(q\) and \(-q\) are Strebel, that is they decompose the surface into cylinders.

Start with \(\alpha = \sum a_i A_i\) and \(\beta = \sum b_j B_j\), systems of disjoint simple curves on \(\Sigma_g\) with positive real weights. Assume these curves bind the surface, in the sense that the complementary regions are disks. At each point of intersection between \(A_i\) and \(B_j\), introduce a rectangle with dimensions \(a_i \times b_j\). These glue together to give a Riemann surface \(X\) and a quadratic differential \(q\). Each complementary \(4 + 2n\)-gon gives rise to a zero of multiplicity \(n\) for \(q\).

**Affine dynamics.** Using this construction one can find quadratic differentials whose stabilizers \(\text{SL}(X,q)\) are large. To this end, first suppose \(q\) decomposes into cylinders \(C_i\) with heights and circumferences \(h_i\) and \(c_i\). Their moduli are given by \(m_i = h_i/c_i\). If all these moduli are equal to a single constant \(m > 0\), then:

\[
\begin{pmatrix}
1 & 1/m \\
0 & 1
\end{pmatrix} \in \text{SL}(X,q).
\]

More generally, if the moduli are given by \((m_i) = m(n_i)\), with \(n_i\) relatively prime integers, then the same is true. The resulting affine transformation is a product of Dehn twists, \(\prod_1^{n_i} \tau_i^{m_i}\). Higher cylinders are twisted more.
Now let us return to a curve system \((A_i, B_j)\) and try to arrange that the moduli of the vertical and horizontal cylinders are all the same. For this it is useful to put the curves together into a single system \(C_i\), let the weights (heights) be \(h_i\), and let \(J_{ij} = i(C_i, C_j)\). Then for any choice of heights we have

\[
c_i = \sum_j J_{ij} h_i.
\]

So if we would like \(1/m_i = c_i/h_i = \mu > 0\) to be a positive constant, we need to solve the eigenvalue equation

\[
\mu h_i = c_i = \sum_j J_{ij} h_i.
\]

By the Perron–Frobenius theorem, there is a unique solution up to scale. With these weights we obtain a quadratic differential such that

\[
\Gamma = \langle (\begin{smallmatrix} 1 & \mu \\ 0 & 1 \end{smallmatrix}) \rangle \subset \text{SL}(X, q).
\]

These elements are represented by right Dehn multitwists \(\tau_A\) and \(\tau_B\). Note that

\[
\text{tr}(\tau_A \tau_B^{-1}) = 2 + \mu^2 > 2,
\]

so this product gives a pseudo–Anosov automorphism of \(X\). Similarly

\[
\text{tr}(\tau_A \tau_B) = 2 - \mu^2
\]

will give a pseudo–Anosov map if \(\mu^2 > 2\).

**General pseudo–Anosov maps.** Remarkably, any irreducible mapping–class \(f\) of infinite order can be represented by an essentially unique pseudo–Anosov map. That is, these elements of \(\pi_1(M_g)\) are the projects of closed \(A\)–orbits. By the Hopf argument applied to the stable and unstable manifolds of \(f\), we obtain:

**Theorem 6.3** Any pseudo–Anosov map is ergodic.

In fact these maps are mixing and their foliations are uniquely ergodic, just as for the torus case. More generally, if the Teichmüller ray determined by \(F(q)\) is recurrent in \(M_g\), then the folation \(F(q)\) is uniquely ergodic (Masur).

**The golden mean and the regular pentagon.** The simplest curve system is the \(A_4\) diagram in genus two. In general a curve system gives rise
to a rectangular intersection matrix $K$, indexed by the $A_i$’s and the $B_j$’s. We then have $\mu^2 = \rho(K K' )$. This makes computations easy. For $A_4$ we have $K = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$, and thus

$$\mu^2 = \rho(K K') = \gamma^2 = \frac{3 + \sqrt{5}}{2}.$$ 

Thus $\mu$ is the golden mean. Since $\mu < 2$ the group $\Gamma$ is a lattice, in fact it is the $(2, 5, \infty)$ triangle group. To see we have an element of order 5, note that

$$|\text{tr}(\tau_A \tau_B)| = \mu^2 - 2 = \gamma^{-1},$$

and $\gamma^{-1} = 2 \cos(2\pi/5) = \zeta_5 + \zeta_5^{-1}$.

The golden table. The differential $q$ can be realized using the symmetric $L$–shaped polygon with long sides of length $\gamma$ and short sides of length 1. This corresponds to the curve $y^2 = p(x)$ where the roots of $p(x)$ are at the 5 points

$$\zeta_5^k - \zeta_5^{-k} = \sin(2\pi k/5), \; k = 1, 2, \ldots, 5.$$ 

The regular pentagon. We can deduce by symmetry that the order 5 point in the SL$_2(\mathbb{R})$ orbit of $(X,q)$ corresponds to the algebraic curve $y^2 = x^5 - 1$, with $q$ the differential $dx^2/y^2$. Its unique zero comes from the point at infinity.

Using the fact that SL$(X,q)$ is a lattice, and the unique ergodicity criterion of Masur, one can then deduce:

**Theorem 6.4** Billiards in a regular pentagon have optimal dynamics. That is, every trajectory is either periodic or uniformly distributed.

Weighted curve systems. We can also attach positive integer weights to each of the curve $A_i$ and $B_j$, and then $\tau_A$ and $\tau_B$ become products of multitwists. If the weights are given by diagonal matrices $W_A$ and $W_B$, and the incidence matrix is $K$, then

$$\mu^2 = \rho(W_A K W_B K').$$

For example, with the weights $(3, 1, 1, 1)$ on the $A_4$ diagram we obtain

$$\mu^2 = \rho\left( \begin{pmatrix} 1 & 1 \\ 3 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \right) = \rho\left( \begin{pmatrix} 2 & 1 \\ 3 & 3 \end{pmatrix} \right) = \frac{5 + \sqrt{13}}{2} = 4.30277\ldots$$
Thus $4 - \mu^2$ is just less than $-2$, and in fact $\tau_A \tau_B$ gives the shortest geodesic on $\mathcal{M}_2$. (For $g \geq 10$ the length of the shortest geodesic on $\mathcal{M}_g$ is unknown.) We remark that

$$\mu^2 - 4 = \frac{1 + \sqrt{13}}{2} = \lambda + \lambda^{-1},$$

where $\lambda$ is the smallest Salem number of degree 4 (satisfying $x^4 - x^3 - x^2 - x + 1 = 0$).

7 Orbit counting and equidistribution

Reference for this section: [EsM].

Equidistribution of spheres.

**Theorem 7.1** Let $p \in X$ be a point on a hyperbolic surface of finite area. Then the spheres about $p$ are equidistributed on $X$. That is, for any compactly supported continuous function $f$ on $X$,

$$\frac{1}{\text{length}(S(p, R))} \int_{S(p, R)} f(s) \, ds \to \frac{1}{\text{area}(X)} \int_X f(x) \, dx$$

as $R \to \infty$.

**Proof.** Consider over $p$ a small symmetric ball $A$ in $T_1 Y$. Then the geodesic flow transports $A$ to a set $g_R(A)$, symmetric about $p$ and concentrated near $S(p, R)$. Pulling $f$ back to $T_1 X$ we have $\langle g_R \chi_A, f \rangle \to \langle \chi_A, 1 \rangle \langle f, 1 \rangle$ by mixing of the geodesic flow. But $(1/mA) \langle g_R \chi_A, f \rangle$ is almost the same as the average of $f$ over $S(p, R)$, by continuity.

**Orbit counting.**

**Theorem 7.2** For any lattice $\Gamma$ in the isometry group of the hyperbolic plane, and any $p, q \in \mathbb{H}^2$,

$$|B(p, R) \cap \Gamma q| \sim \frac{\text{area } B(p, R)}{\text{area } (\mathbb{H}^2/\Gamma)}.$$

**Proof.** Replace $q$ by a bump function $f$ of total mass 1 concentrated near $q$, and let $F = \sum_{\Gamma} f \circ \gamma$. Then the orbit count is approximately $\int_{B(p, R)} F(x) \, dx$. But area measure on $B(p, R)$ is a continuous linear combination of the probability measures on the circles $S(p, r)$, $0 < r < R$. For $r$
large, the average of $F(x)$ over $S(p, r)$ is close to the average of $F(x)$ over $X = \mathbb{H}/\Gamma$, by equidistribution. The latter is $1/\text{area}(X)$. Since the total measure of $B(p, R)$ is $\text{area}(B(p, R))$, we find the orbit count is asymptotic to $\text{area}(B(p, R))/\text{area}(X)$.

The error term in the hyperbolic orbit counting problem is studied in [PR].

For a Euclidean lattice $L \subset \mathbb{R}^n$, it is easy to see that

$$|L \cap B(0, R)| = \frac{\text{vol} B(0, R)}{\text{vol}(\mathbb{R}^n/L)} + O(R^{n-1}).$$

For $L = \mathbb{Z}^2$ the classical circle problem is to estimate the error term, which is usually written in the form

$$P(x) = \sum_{n \leq x} r(n) - \pi x$$

where $r(n)$ is the number of integer solutions to $a^2 + b^2 = n$. The estimate above gives $P(x) = O(x^{1/2})$. A typical modern bound is $P(x) = O(x^{7/22 + \epsilon})$ (Iwaniec and Mozzochi, 1988). Numerical evidence supports $P(x) = O(x^{1/4} \log x)$.

**Counting lifts of closed geodesics.** Let $\gamma \subset Y$ be a closed geodesic on a surface of finite area. Let $\tilde{\gamma} \subset \mathbb{H}$ be its lift to the universal cover; it is a locally finite configuration of geodesics. Fixing a point $p \in \mathbb{H}$, let $N(R)$ denote the number of distinct geodesics in $\tilde{\gamma}$ meeting $B(p, R)$.

In terms of the Minkowski model, $\gamma$ gives a point $x$ in the 1-sheeted hyperboloid $\mathcal{G}$ with a discrete orbit. The counting problem translates into knowing the number of points of $\Gamma x$ meeting a compact region corresponding to $B(p, R)$. But $\Gamma \backslash \mathcal{G}$ is not even Hausdorff! Although the orbit $\Gamma x$ is discrete, typical orbits in $\mathcal{G}$ are dense (by ergodicity of the geodesic flow). So how to solve the counting problem?

**Theorem 7.3** We have

$$N(R) \sim \frac{\text{area} B(\gamma, R)}{\text{area} Y}.$$

Here $\text{area} B(\gamma, R)$ is the immersed area of an $R$-neighborhood of $\gamma$ on $Y$.

One can show that

$$\text{area}(B(\gamma, R)) \sim \text{length}(\gamma) \text{area}(B(x, R)/\pi).$$
Proof. Using mixing of the geodesic flow, one can show the parallels $L_t$ of $\gamma$ at distance $t$ are equidistributed on $Y$. (The proof is similar to that for equidistribution of spheres.) Now consider the covering space $Z \to Y$ corresponding to $\pi_1(\gamma) \subset \pi_1(Y)$. Let $E \subset Z$ be the projection of $\Gamma_p$ to $Z$. Then it is not hard to see that $N(R)$ is the same as $|E \cap B(\gamma', R)|$, where $\gamma'$ is the canonical lift of $\gamma$ from $Y$ to $Z$. Projecting Dirichlet regions based at $\Gamma_p$ to $Z$ gives the heuristic for the count; it is justified by equidistribution of parallels to $\gamma$.

Counting integral points on level sets of quadratic forms. Fix $A \neq 0$ and $D > 0$. Consider the Diophantine problem of counting the number of integral solutions $N(R)$ to the equation

$$x^2 + y^2 = A + Dt^2$$

with $x^2 + y^2 \leq R^2$. For $A < 0$ this corresponds to counting orbits; for $A > 0$, to counting geodesics.

To fix the ideas, suppose $D \equiv 3 \text{ mod } 4$, and $A < 0$. Then the quotient $\text{SO}(q, \mathbb{R})/\text{SO}(q, \mathbb{Z})$ is compact, so only finitely many points on $Y = \mathbb{H}/\text{SO}(q, \mathbb{Z})$ correspond to integral points on the hyperboloid $x^2 + y^2 - Dt^2 = A$. For a given orbit $O$, the preceding discussion shows

$$N(R, O) \sim C \text{ vol}(B(0, R) \cap \mathbb{H}) \sim CR$$

(by our earlier estimate of the volume form on $\mathbb{H}$ and $\mathcal{H}$); and since there are only finitely many orbits we have

$$N(R) \sim C(A, D)R.$$  

For the case $A > 0$ the geodesic counting argument leads to the same conclusion.

In fact the same estimate holds for any $A \neq 0$, since even in the finite volume case, there are only finitely many orbits of integral points.

Counting integral points in $\mathbb{R}^{2+1}$. Consider the simplest case, namely the problem of counting solutions to

$$x^2 + y^2 + 1 = t^2,$$

associated to the level set $q = -1$ for the form $q(x, y, t) = x^2 + y^2 - t^2$.

**Theorem 7.4** Let $N(R)$ be the number of $(x, y)$ with $x^2 + y^2 \leq R^2$ that occur in integral solutions to $x^2 + y^2 + 1 = t^2$. Then $N(R) \sim R$ as $R \to \infty$. 

75
Proof. Let $\Gamma = \text{SO}(q, \mathbb{Z})$. We first show that all the integral solutions belong to a single $\Gamma$-orbit, namely the orbit of $(0, 0, 1)$.

Indeed, $\Gamma$ is essentially the group $\text{SL}_2(\mathbb{Z})$. Passing to $\mathbb{RP}^2$, consider the ideal quadrilateral $F \subset \mathbb{H} \cong \Delta$ which has vertices $(\pm 1, 0), (0, \pm 1)$. Then $F$ is a fundamental domain for a subgroup of $\Gamma$ (in fact for $\Gamma(2) \subset \text{SL}_2(\mathbb{Z})$.) The geodesic from $(1, 0)$ to $(0, 1)$ bounding $F$ is just the line $x + y = 1$, or $x + y = t$ in homogeneous coordinates. By moving solutions into $F$, we see the orbit of every integral point has a representative with $|x \pm y| \leq |t|$. But squaring this equation and using the assumption that $x^2 + y^2 + 1 = t^2$, we get $|2xy| \leq 1$. The only integral point with this property is $(x, y) = (0, 0)$.

By tiling $B(R) \cap \mathbb{H}^2$ with copies of $F$, we conclude that $N(R)$ grows like $\text{vol } B(R)/\text{vol}(F)$. Now projecting to the $(x, y)$-plane, $F$ becomes the region $|2xy| \leq 1$, $B(R)$ becomes the region $x^2 + y^2 \leq R^2$, and the hyperbolic area element becomes $dA = dx dy/\sqrt{x^2 + y^2 + 1}$. We then calculate that $\text{area}(F) = 2\pi$. Since $dA \sim drd\theta$, we get $\text{area } B(R) \sim 2\pi R$, and thus $N(R) \sim R$. 

Example: for radius $R = 100$, there are 101 solutions, generated from $(x, y)$ in the set

$\{(0, 0), (2, 2), (8, 4), (12, 12), (18, 6), (30, 18), (32, 8), (38, 34), (46, 22), (50, 10), (68, 44), (70, 70), (72, 12), (76, 28), (98, 14)\}$

by changing signs and swapping the coordinates. A few other values: $N(50) = 41; N(200) = 197; N(1000) = 993$; see Figure 2.

Quaternion algebras. A quaternion algebra $D$ over a field $K$ is a central simple associative algebra of rank 4. Alternatively, passing to the algebraic closure we have $D \otimes \overline{K} \cong M_2(\overline{k})$, a matrix algebra.

Every element $x \in D - k$ generates a quadratic field extension $K = k(x)$ of $k$, since this is true in the matrix algebra. Thus we obtain a trace, norm and conjugation involution on $D$.

Assume char $k \neq 2$. Then we can normalize $x \in D$ by subtracting $\text{tr}(x)/2$ to arrange that $\text{tr}(x) = 0$.

Let $i \in D - K$ be an element of trace zero. Then $i^2 = a \in k^*$. It is known that there exists an element $j \in k$ such that $ij = -ji$, and we can also arrange that $\text{tr}(j) = 0$. Write $j^2 = b \in k^*$. Then if we let $k = ij$ (not to be confused with the base field), we have $k^2 = -ab$.

In short, a quaternion algebra is specified by a pair of elements $(a, b) \in k^*$. Examples: The Hamilton quaternions correspond to $(a, b) = (-1, -1)$. The algebra $M_2(k)$ comes from $(a, b) = (1, 1)$. For the matrix algebra we
can take

\[ i = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad j = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \text{and} \ k = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \]

A quaternion algebra is *split* if \( D \cong M_2(k) \); otherwise it is *ramified*.

Let \( D_0 \subset D \) be the trace zero subset. Then \( D_0 = ki \oplus kj \oplus k(ij) \), with the norm given by

\[ N(xi + yj + tk) = q(x, y, t) = ax^2 + by^2 - abt^2. \]

Thus a quaternion algebra \( D \) determines a ternary quadratic form \( q \), whose zero locus gives a conic \( C \) in \( \mathbb{P}^2 \). The algebra \( D \) is split iff \( q \) represents zero iff \( C(k) \) has a point.

For example, a quaternion algebra over \( \mathbb{R} \) is unramified iff the form \( q(x, y, t) \) is *indefinite*.

To discuss 'integral points', we need to take a maximal order \( \mathcal{O} \subset D \), that is, a maximal subring that is finitely-generated as a \( \mathbb{Z} \)-module. For concreteness let us suppose \( k = \mathbb{Q} \). Then the group of units \( U \subset \mathcal{O} \) consists of elements with norm \( N(u) = \pm 1 \). We can also discuss the units \( U_1 \) with norm 1, and the quotient \( U' = U_1/(\pm 1) \).
For example, if $\mathcal{O} = M_2(\mathbb{Z})$, then $U = GL_2(\mathbb{Z})$, $U_1 = SL_2(\mathbb{R})$ and $U' = PSL_2(\mathbb{R})$.

If we suppose further that $D \otimes \mathbb{R}$ is split, i.e. that the form $q$ is indefinite, then we obtain an embedding $U_1 \subset SL_2(\mathbb{R})$ realizing $U_1$ is an arithmetic group. This groups is commensurable to $SO(q, \mathbb{Z})$.

At the same time, $U_1$ acts on the set $X_n$ of elements of norm $n$ and trace 0 in $\mathcal{O}$, i.e. on the solutions to the Diophantine equation $q(x, y, t) = n$. The space of orbits $X_n/U_1$ is finite.

Example: the integral solutions $X$ to $x^2 + y^2 + 1 = t^2$ form the level set $X_{-1}$ for the algebra $D \cong M_2(\mathbb{Z})$ with structure constants $(a, b) = (1, 1)$. Indeed, we can think of $X \subset M_2(\mathbb{Z})$ as the set of matrices with trace zero and determinant one, with $U_1 \cong SL_2(\mathbb{Z})$ acting by conjugation. Any matrix $A \in X$ is conjugate to $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, explaining again why $X$ consists of a single orbit.

By localizing, we can talk about the set of places of $k$ such that $D_v$ is ramified. Class field theory shows over a number field $k$, a quaternion algebra is ramified at a finite set of places $R$, that $|R|$ is even, that any even set arises, and that $R$ determines $D$ up to isomorphism.

More abstractly, we can regard the data $(a, b)$ as a pair of classes in $k^*/(k^*)^2 \cong H^1(k, \mathbb{Z}/2)$ in Galois cohomology, and the isomorphism class of $D$ is the cup product $a \wedge b$ in the Brauer group $H^2(k, \mathbb{Z}/2)$. For the rationals, we have

$$H^2(\mathbb{Q}, \mathbb{Z}/2) = \{ x \in \oplus_v (\mathbb{Z}/2)_v : \sum x_v = 0 \}.$$ 

8 Spectral theory

References for this section: [CFS], [Me].

Unitary representations. Let $G$ be a topological group, and let $H$ be a complex Hilbert space. Usually $H$ will be separable, and possibly finite-dimensional. We have $\langle \alpha, \beta \rangle = \overline{\langle \beta, \alpha \rangle}$.

A representation of $G$ on $H$ is a continuous linear action $G \times H \to H$. A representation is unitary if $\langle g\alpha, g\beta \rangle = \langle \alpha, \beta \rangle$ for all $\alpha, \beta \in H$.

Example. The action of $\mathbb{R}$ on $\ell^2(\mathbb{Z})$ by $g^t(a_n) = e^{int}a_n$ is unitary. Note that in the operator norm,

$$\lim_{t \to 0} \|g^t - g^0\| \neq 0.$$ 

This shows a representation is generally not given by a continuous map $G \to \mathcal{B}(H)$ in the operator norm topology.

78
Theorem 8.1  A unitary action of $G$ is continuous iff the matrix coefficient $\langle g\alpha, \beta \rangle$ is continuous on $G$ for any $\alpha, \beta \in \mathbb{H}$.

Proof. Suppose $g_n \to e$ in $G$, and $\alpha_n \to \alpha$; we wish to show $g_n\alpha_n \to \alpha$. Since $\|g_n\| = 1$, it suffices to show $g_n\alpha \to \alpha$, and we may assume $\|\alpha\| = 1$. But then $\langle g_n\alpha, \alpha \rangle \to 1$, and since $\|g_n\alpha\| = 1$ this implies $g_n\alpha \to \alpha$. \qed

Classification of unitary representations of locally compact abelian groups $G$ such as $\mathbb{Z}$ and $\mathbb{R}$. (Cf. [Nai], [RN], [Mac].) First, any irreducible (complex) representation is 1-dimensional, and given by a character $\chi: G \to \mathbb{C}^* = GL_1(\mathbb{C})$. Second, any representation can be decomposed into a direct integral over $\hat{G}$ of irreducibles. Note that $\hat{\mathbb{Z}} = S^1$, $\hat{\mathbb{R}} = \mathbb{R}$.

The direct integral depends on the following data:

(a) a measure class $[\mu]$ on $\hat{G}$, and
(b) a measurable function $m: \hat{G} \to \{0, 1, 2, \ldots, \infty\}$.

From this data we form a measurable bundle $\mathcal{H}$ of Hilbert spaces $H_\chi$ over $\hat{G}$ such that $\dim H_\chi = m(\chi)$ a.e. (this bundle is trivial on the level sets of $m$).

Next consider $L^2(\hat{G}, [\mu])$, the Hilbert space of square-integrable $(1/2)$-densities for $[\mu]$. This space is functorially attached to the measure class rather than a measure $\mu$ within the class. It is canonically isomorphic to $L^2(\hat{G}, \mu)$, and for any $\nu = f\mu$ in the same measure class, we have a canonical isometry $L^2(\hat{G}, \mu) \to L^2(\hat{G}, \nu)$ given by $a(x) \mapsto f(x)^{-1/2}a(x)$. Morally, these half-densities are ‘differentials’ of the form $a(x)d\mu^{1/2}$, so
\[
a(x)d\mu^{1/2} = a(x) \left(\frac{dv}{d\mu}\right)^{-1/2}dv^{1/2}.
\]

Finally, let $H = L^2(\hat{G}, [\mu], \mathcal{H})$ be the Hilbert space of $L^2$ sections of $\mathcal{H}$. The inner product is given by
\[
\langle f, g \rangle = \int_{\hat{G}} \langle f(\chi), g(\chi) \rangle d\mu(\chi),
\]
where the pointwise inner products are taken in $H_\chi$. Then $G$ acts unitarily on $H$ by $(g \cdot f)(\chi) = \chi(g)f(\chi)$. This is the direct integral representation
\[
\pi = \int m(\chi) \cdot \chi d\mu(\chi).
\]
Theorem 8.2 Every unitary representation $\pi$ of $G$ on a separable Hilbert space is isomorphic to a direct integral of 1-dimensional representations. The measure class $\mu$ and multiplicity function $m(\chi)$ on $\hat{G}$ are uniquely determined by $\pi$.

Corollary 8.3 Two unitary operators are equivalent iff their spectral measure classes on $S^1$ are the same and their multiplicity functions agree a.e.

Simplest example: Let $U$ act on $H$ with $\dim H < \infty$. Then $U$ gives a representation of $G = \mathbb{Z}$, $\hat{G} = S^1$, we have the eigenspace decomposition $H = \bigoplus H_i$ with distinct eigenvalues $\lambda_i$; we can set $\mu = \sum \delta_{\lambda_i}$; and $m(\lambda) = \dim H_i$ if $\lambda = \lambda_i$, $m(\lambda) = 0$ otherwise.

Direct integral decomposition and Fourier analysis.

a) Let $\mathbb{R}$ act via rotation on $H = L^2(S^1)$; then $H = \oplus H_n = \int_{\mathbb{R}} H_t d\mu(t)$ where $x$ acts by $e^{ix \cdot t}$ on $H_t$; here $\mu$ is supported on the integers.

b) Let $\mathbb{R}^d$ act on $H = L^2(\mathbb{R}^d)$ by translation. Then $H = \int_{\mathbb{R}^d} H_t dt$, where $dt$ denotes Lebesgue measure. The decomposition of $f \in H$ according to $H_t$ is just the Fourier transform $\hat{f}(t)$.

Proof of the spectral decomposition for a unitary operator. First suppose $\xi$ is a cyclic vector, that is $\langle U^n \xi : n \in \mathbb{Z} \rangle$ is dense in the Hilbert space $H$. Consider the linear functional defined on trigonometric polynomials $f(z) = \sum_{-N}^N a_n z^n$ by

$$\mu(f) = \langle f(U) \xi, \xi \rangle.$$

It is known that if $f(z) \geq 0$ for $z \in S^1$ then $f(z) = g(z)\overline{g(z)}$ for some trigonometric polynomial $g(z)$. Also $g(U) = g(U)^*$. Thus $f \geq 0$ implies

$$\mu(f) = \langle g(U)^* g(U) \xi, \xi \rangle = \langle g(U) \xi, g(U) \xi \rangle \geq 0.$$

By this positivity, $|\mu(f)| = O(\|f\|_\infty)$, so $\mu$ defines a measure on $C(S^1)$.

Now map the dense subset of $L^2(S^1, \mu)$ spanned by trigonometric polynomials into $H$ by

$$\Phi : \sum_{-N}^N a_n z^n \mapsto \sum_{-N}^N a_n U^n \xi.$$

In other words, let the constant function 1 correspond to $\xi$ and arrange that multiplication by $z$ goes over to the action of $U$. Using the fact that $\langle U^j \xi, U^j \xi \rangle = \langle U^{j-j} \xi, \xi \rangle$ one can easily check that $\Phi$ is an isometry. Since $U^n \xi$ is dense, we obtain an isomorphism between $L^2(S^1, \mu)$ and $H$, expressing $U$ as a direct integral of irreducible representations of $\mathbb{Z}$.
Finally, if $H$ has no cyclic vector we can still write $H = \oplus H_i$, a countable direct sum of $U$-invariant subspaces, with $\xi_i$ a cyclic vector for $H_i$ (here we use the fact that $H$ is separable). Then we obtain an isomorphism $H = \oplus H_i$. Let $\mu = \sum \mu_i$, write $\mu_i = f_i \mu$, let $E_i$ be the indicator function of the set where $f_i > 0$ and let $m(x) = \sum \chi_{E_i}(x)$. Then $H$ is isomorphic to the direct integral over $(S^1, \mu)$ with multiplicity $m$.

**Spectral classification of irrational rotations.** Let $T$ be an irrational rotations of $S^1$ by $\lambda$. Then $L^2(S^1)$ decomposes as a sum of eigenspaces $\langle z^n \rangle$ with eigenvalues $\lambda^n$. Thus the spectral measure class is $\mu = \sum \delta_{\lambda^n}$, and the multiplicity function $d(\chi) = 1$.

Since the support of $\mu$ is a cyclic subgroup of $S^1$, we see $\lambda^{\pm 1}$ is uniquely determined by the unitary operator associated to $T$. In particular, a pair of irrational rotations are spectrally isomorphic $\iff \lambda_1 = \lambda_2$ or $\lambda_1^{-1} = \lambda_2$, in which case they are spatially isomorphic (use complex conjugation if necessary).

**Ergodicity is detected by the spectrum.** Indeed, $T$ is ergodic iff $m(1) = 1$. (Of course $T$ ergodic does not imply $U$ irreducible; indeed all irreducible representations of abelian groups are 1-dimensional.)

**Lebesgue spectrum.** We say $U : H \to H$ has *Lebesgue spectrum* of multiplicity $n$ if it is conjugate to multiplication by $z$ on $\bigoplus_1^n L^2(S^1, d\theta)$; equivalently, if $H$ decomposes as a direct integral over $S^1$ with Lebesgue measure and multiplicity $n$. (Infinite multiplicity is permitted).

Examples:

(a) Consider the shift $T : \mathbb{Z} \to \mathbb{Z}$ preserving counting measure. Then the corresponding $U$ on $\ell^2(\mathbb{Z})$ has Lebesgue spectrum. Indeed, the Fourier transform gives an isomorphism $\ell^2(\mathbb{Z}) \cong L^2(S^1, d\theta)$ sending $U$ to multiplication by $z$.

(b) The unitary shift on $\bigoplus_{\mathbb{Z}} H$ has Lebesgue spectrum of multiplicity $n = \dim H$.

An automorphism $T$ of a measure space is a *Lebesgue automorphism* if the associated unitary operator has Lebesgue spectrum.

It is remarkable that there is no known Lebesgue automorphism of finite multiplicity [Me, p.146].

**Theorem 8.4** Any ergodic toral automorphism $T$ has Lebesgue spectrum of infinite multiplicity.

Proof: Let $T$ be an ergodic automorphism of a torus $(S^1)^n$. Then $U$ is given (after Fourier transform) by the action of $T$ on the Hilbert space $H = \ell^2(\mathbb{Z}^d)$, using the characters as a basis for $L^2((S^1)^d)$. Given $\chi \neq 0$ in $\mathbb{Z}^d$,
the orbit \( \langle T^i(\chi) \rangle \) is countable and gives rise to a subspace of \( H \) isomorphic to \( \ell^2(\mathbb{Z}) \) on which \( U \) acts by the shift. Since \( \mathbb{Z}^d - \{0\} \) decomposes into a countable union of such orbits, we have the theorem.

**Corollary 8.5** Any two ergodic toral automorphisms are spectrally equivalent.

Note that the equivalence to \( \oplus_1^\infty \ell^2(\mathbb{Z}) \) is easy to construct concretely on the level of characters.

**Spectrum and mixing.** Any transformation with Lebesgue spectrum is mixing, since

\[ \langle U^n f, g \rangle = \int_{S^1} z^n \langle f(z), g(z) \rangle |dz| \rightarrow 0 \]

(indeed by the Riemann-Lebesgue lemma, the Fourier coefficients of any \( L^1 \) function tend to zero). Mixing can be similarly characterized in terms of a representative spectral measure: it is necessary and sufficient that \( \int z^n f \, d\mu \rightarrow 0 \) for any \( f \in L^2(S^1, \mu) \). (In particular mixing depends only on the measure class, not on the multiplicity).

**Spectral theory of the baker’s transformation \( T \):** it is also Lebesgue of infinite multiplicity. This can be seen at the level of characters, or using the fact that \( T \) is a \( K \)-automorphism (this means there is a Borel subalgebra \( A \) such that \( T(A) \) refines \( A \), \( \bigvee T^i(A) \) is the full algebra and \( \bigwedge T^{-i}(A) \) is the trivial algebra).

Hence all shifts and all toral automorphisms are spectrally isomorphic.

**Positive definite sequences and functions.** We now turn to a more systematic discussion of unitary representations of abelian groups; cf. [Ka].

A finite matrix \( a_{ij} \) is positive definite if

\[ \sum a_{ij} b_i b_j \geq 0 \]

for all \( b_i \).

A continuous complex function \( f(x) \) on an abelian group \( G \) is positive definite if the matrix \( a_{ij} = f(x_i - x_j) \) is positive definite for all finite sets \( x_i \in G \). The main cases of interest for us will be the cases \( G = \mathbb{Z} \) and \( G = \mathbb{R} \).

By considering a 2-point set \( (x_1, x_2) \) we find \( f(-x) = \overline{f(x)} \) and \( |f(x)| \leq f(0) \).

Examples.

Any unitary character \( f : G \rightarrow S^1 \) is positive definite. The positive definite functions form a convex cone. (We will see all positive definite functions are positive combinations of characters.)
If $U$ is a unitary operator, then
\[ a_n = \langle U^n \xi, \xi \rangle \]
is a positive definite sequence. Proof:
\[ \sum a_{i-j}b_i\overline{b_j} = \| \sum b_i U^i \xi \| \geq 0. \]
Similarly if $U^t$ is a unitary representation of $\mathbb{R}$, then $f(t) = \langle U^t \xi, \xi \rangle$ is positive definite.

It turns out every positive definite function $f(g)$ on $G$ is a linear combination of characters. What would that mean? It means there is a positive measure $\mu$ on $\hat{G}$ such that
\[ f(g) = \int \chi(g) d\mu(\chi). \]
But this says exactly that $f$ is the Fourier transform of a measure!

We will prove this result for the cases $G = \mathbb{Z}$ and $G = \mathbb{R}$.

**Theorem 8.6 (Herglotz)** A sequence $a_n$ is the Fourier series of a positive measure on $S^1$ if and only if it is positive definite.

**Proof.** Let $\mu = \sum a_n z^n$ formally represent the desired measure. Then if $f(z) = \sum b_n z^n$ has a finite Fourier series, we can formally integrate:
\[ \frac{1}{2\pi} \int f(z) d\mu(z) = \sum a_n b_{-n}. \]
It will suffice to show $\int f \, d\mu \geq 0$ whenever $f \geq 0$, since we have $\int 1 \, d\mu = 2\pi a_0$, and thus we will have $|\mu(f)| = O(\|f\|_{\infty})$ and $\mu$ will extend to a bounded linear functional on $C(S^1)$.

Now the point of positive definiteness is that
\[ \frac{1}{2\pi} \int |f(z)|^2 \, d\mu(z) = \sum_{j-i=n} a_i b_i \overline{b_j} + \sum_{i,j} a_{i-j} b_i b_j \geq 0. \]
(We could now finish as before, using the fact that every positive $f$ is of the form $|g(z)|^2$. We will take an alternate route).

Next note that there are trigonometric polynomials $k^N(z)$ such that:

- $\delta^N(z) = |k^N(z)|^2 \geq 0$, $\int_{S^1} \delta^N(z) |dz| = 1$, and $f \ast \delta^N \to f$ uniformly for any trigonometric polynomial $f$. 

83
In other words, $\delta^N$ form an approximate identity in the algebra $L^1(S^1)$ with respect to convolution, and $\delta^N$ tends to the distributional delta-function (which satisfies $f * \delta = \delta$).

To construct $\delta^N(z) = \sum d^n z^n$, note that $\hat{f} * \delta^N = \hat{f} \delta^N$, so we want $\delta^N \to 1$ as $N \to \infty$. The first thing to try is $k^N(z) = \sum_{-N}^N z^n$. Then $k^N * f \to f$ but unfortunately $k^N$ is not positive. To remedy this, set $\delta^N(z) = \frac{\lvert k^N(z) \rvert}{2N + 1}$. Then

$$d_n^N = \frac{\max(0, 2N + 1 - |n|)}{2N + 1},$$

and the desired properties are easily verified. (Note:

$$\delta^N(z) = \frac{1}{2N + 1} \left( \frac{\sin \frac{N+1}{2} \theta}{\sin \theta/2} \right)^2$$

is known as the Fejér kernel.)

Now since $\delta^N = |k_N|^2$ by positivity of $a_n$ we have $\int \delta^N(x + y) \, d\mu(x) \geq 0$ for any $y$. Thus

$$0 \leq \lim_N \int (f * \delta^N) \, d\mu = \int f \, d\mu$$

as desired. Here the limit is evaluated using the fact that $\hat{f} * \delta^N$ is a trigonometric polynomial of the same degree as $f$, whose coefficients converge to those of $f$.

Bochner’s theorem and unitary representations of $\mathbb{R}$.

**Theorem 8.7 (Bochner)** The positive definite functions $f : \mathbb{R} \to \mathbb{C}$ are exactly the Fourier transforms of finite measures on $\mathbb{R}$.

This means $f(t) = \int_{\mathbb{R}} e^{itz} \, d\mu(x)$.

**Corollary 8.8** Let $U^t$ be a unitary action of $\mathbb{R}$ on $H$ with cyclic vector $\xi$. Then there is:

- a probability measure $\mu$ on $\mathbb{R}$, and
- an isomorphism $H \to L^2(\mathbb{R}, \mu)$, such that
- $\xi$ corresponds to the constant function 1, and
- $U^t$ is sent to the action of multiplication by $e^{i xt}$. 

84
Proof. First let \( f(t) = \langle U^t \xi, \xi \rangle \). Then
\[
\sum f(t_i - t_j) a_i a_j = \| \sum a_i U^{t_i} \xi \|^2 \geq 0
\]
so \( f \) is positive definite.

Let \( A \subset L^2(\mathbb{R}, \mu) \) be the algebra spanned by the characters, i.e. those functions of the form \( g(x) = \sum_1^N a_i e^{ixt_i} \). (Note that the product of two characters is another character.) Map \( A \) to \( H \) by sending \( g \) to \( \sum_1^N a_i U^{t_i} \xi \). It is easily verified that this map is an isometry (by the definition of \( \mu \)), and that \( A \) is dense in \( L^2(\mathbb{R}, \mu) \) (by the Stone-Weierstrass theorem). The completion of this map gives the required isomorphism. ■

**Proof of Bochner’s theorem.** The argument parallels the proof for \( \mathbb{Z} \). Let \( x \) and \( t \) denote coordinates on \( \mathbb{R} \) and \( \hat{\mathbb{R}} \). On the real line, the Fourier transform and its inverse are given by:
\[
\hat{f}(t) = \int f(x) e^{-ixt} \, dx
\]
and
\[
f(x) = \frac{1}{2\pi} \int \hat{f}(t) e^{ixt} \, dt.
\]

Consider first the space \( A \) of functions \( f(x) \in L^1(\mathbb{R}) \) such that \( \hat{f}(t) \) is compactly supported. These are the analogues of trigonometric polynomials; they are analytic functions of \( x \).

Let \( \phi(t) \) be a positive definite function. Define a linear functional on \( A \) formally by
\[
\int f(x) \, d\mu = \frac{1}{2\pi} \int \hat{f}(-t) \phi(t) \, dt.
\]

We claim \( f \geq 0 \) implies \( \int f \, d\mu \geq 0 \). To see this we use approximate identities again. That is, define on the level of Fourier transforms,
\[
\hat{k}^T = 1 \text{ for } |t| \leq T,
\]
and
\[
\hat{\delta}^T = \frac{\hat{k}^T \ast \hat{k}^T}{2T} = 1 - \frac{|t|}{2T} \text{ for } |t| \leq 2T.
\]
Then
\[
k^T(x) = \frac{1}{2\pi} \frac{\sin(Tx)}{x/2}
\]

85
and
\[ \delta^T(x) = \frac{|k^T(x)|^2}{2T} \]
satisfies \( \int \delta^T(x) \, dx = 1 \).

By positive definiteness of \( \phi(t) \), for \( f \in A \) we have
\[ \int |f(x)|^2 \, d\mu \geq 0; \]
thus \( \int \delta^T(x + y) \, d\mu(x) \geq 0 \) for any \( y \). As before, for \( f \geq 0 \) in \( A \) we then have
\[ 0 \leq \int (\delta^T * f)(x) \, d\mu \to \int f(x) \, d\mu \]
and we have verified positivity.

Now enlarge \( A \) to the class \( S \) of \( f(x) \) such that \( \hat{f} \) decays rapidly at infinity. Then \( \hat{f} \in L^1 \), so the definition of \( \int f \, d\mu \) extends to \( S \), and the above argument generalizes to show positivity on \( S \).

The important difference between \( S \) and \( A \) is that \( S \) includes all \( C^\infty \) functions of compact support. We now show \( \int f \, d\mu = O(\|f\|_\infty) \) for such functions.

Let \( f \in S \) have compact support and satisfy \( 0 \leq f \leq 1 - \epsilon \). Observe that for \( T \) small, \( \delta^T(x) \) is almost constant near the origin, and
\[ \int \delta^T(x) \, d\mu = \frac{1}{2\pi} \int_{-2T}^{2T} \left( 1 - \frac{|t|}{2T} \right) \phi(t) \, dt \sim \frac{2T}{2\pi} \phi(0) \]
as \( T \to 0 \). Letting \( f_T(x) = (\pi/T) \delta^T(x) \) we have \( f_T(0) = 1 \) and \( \int f_T(x) \, d\mu \to \phi(0) \). If \( T \) is sufficiently small, then \( f \leq f_T \) and thus \( \int f \, d\mu \leq \phi(0) \). It follows that for any \( f \in S \) with compact support we have \( |\int f \, d\mu| \leq \phi(0)\|f\|_\infty \), and from this we find \( \mu \) is a measure of finite total mass. \hfill \qed

More generally, given a set \( X \), a function \( f : X \times X \to \mathbb{C} \) is positive definite (or of positive type) if
\[ \sum \lambda_i \overline{\lambda_j} f(x_i, x_j) \geq 0 \]
for any finite sequence \( x_i \in X \) and \( \lambda_i \in \mathbb{C} \). Cf. [HV, §5].

**Theorem 8.9** The kernel \( f \) is positive definite if and only if there exists a Hilbert space \( H \) and a map \( F : X \to H \) such that the linear span of \( F(X) \) is dense in \( H \) and
\[ f(x, y) = \langle F(x), F(y) \rangle. \]
The pair \((F, H)\) is unique up to isometry.
Proof. Consider the vector space $V = \mathbb{C}^X$ (each element of $X$ gives a basis vector); use $f$ to define an inner product on this space, take the completion, and mod out by vectors of length zero to obtain $H$. □

Entropy and information. We have seen that many measurable dynamical systems such as shifts and toral automorphisms are spectrally equivalent (since they are Lebesgue). Are they actually spatially conjugate? Kolmogorov and Shannon invented an invariant to distinguish them.

Suppose you learn that $x \in X$ belongs to $A \subset X$; how much information $I(A)$ have you gained about $x$? If we postulate that:

- $I(A) = f(m(A))$: the amount of information depends only on the measure of $A$; and
- $f(xy) = f(x) + f(y)$: so the information provided by independent events (in the sense of probability theory) is additive;

then we find that $f(x) = C \log x$; and to maintain positivity we set $I(A) = \log(1/m(A))$.

The entropy of a partition $\mathcal{A}$ of a measure space $X$ is the expected amount of information gained when you pick a point at random, and then learn which element of $\mathcal{A}$ contains it. It is thus given by

$$h(\mathcal{A}) = \int_X \sum_{A} -\chi_A \log m(A) = \sum_{A} m(A) \log(1/m(A)).$$

Entropy of a dynamical system. Given a partition $\mathcal{A}$ and dynamics $T : X \to X$, consider the information gained by measuring the position not only of $x$ but of $Tx$, $T^2x$, $\ldots$, $T^{n-1}x$ relative to $\mathcal{A}$. This is the same as measuring $x$ relative to the partition

$$\mathcal{A}(T,n) = \mathcal{A} \lor T^{-1}(\mathcal{A}) \lor \ldots \lor T^{-n+1}(\mathcal{A}).$$

We can measure the rate of growth of information along an orbit by:

$$h(T,\mathcal{A}) = \limsup_n \frac{h(\mathcal{A}(T,n))}{n}$$

and finally define the entropy of $T$ by

$$h(T) = \sup_{\mathcal{A}} h(T,\mathcal{A})$$

where the supremum is over all finite partitions.
Basic property of entropy. The most basic property of entropy is
\[ h(A \vee B) \leq h(A) + h(B). \]
In other words, the information from two measurements is maximized when the measurements are independent.

**Proof.** Let \( \phi(x) = x \log x \); then \( \phi \) is convex; the inequality will come from the fact that \( \phi(\sum a_i b_i) \leq \sum a_i \phi(b_i) \) when \( \sum a_i = 1, a_i \geq 0. \) Indeed, we have
\[ h(A \vee B) = -\sum_{i,j} m(A_i \cap B_j) \log m(A_i \cap B_j) \]
\[ = -\sum_i m(A_i) \sum_j \frac{m(A_i \cap B_j)}{m(A_i)} \left( \log m(A_i \cap B_j) + \log m(A_i) \right) \]
\[ = h(A) - \sum_{i,j} m(A_i) \phi \left( \frac{m(A_i \cap B_j)}{m(A_i)} \right) \]
\[ \leq h(A) - \sum_j \phi(m(B_j)) \]
\[ = h(A) + h(B). \]

From this inequality it follows that the lim sup in the definition of \( h(T, \mathcal{A}) \) is actually a limit.

**Entropy calculations.** Suppose the smallest \( T \)-invariant \( \sigma \)-algebra generated by \( \mathcal{A} \) is equal to all the measurable sets, up to sets of measure zero. Then we say \( \mathcal{A} \) is a generating partition.

**Theorem 8.10 (Kolmogorov-Sinai)** If \( \mathcal{A} \) is a generating partition, then \( h(T, \mathcal{A}) = h(T) \).

**Idea of the proof.** Given an arbitrary partition \( \mathcal{B} \), suppose we can find an \( i \) such that \( \mathcal{A}(T, i) > \mathcal{B} \) (the first partition is finer). Then \( \mathcal{A}(T, n+i) > \mathcal{B}(T, n) \) and thus
\[ h(T, \mathcal{A}) = \lim_{n} \frac{h(\mathcal{A}(T, i+n))}{n} \geq \lim_{n} \frac{h(\mathcal{B}(T, n))}{n} = h(\mathcal{B}). \]
Since \( h(T) = \sup h(T, \mathcal{B}) \) we would be done. The idea is that since \( \mathcal{A} \) is a generating partition, \( \mathcal{A}(T, i) \) does almost refine any given finite partition \( \mathcal{B} \).
Note: we always have $h(T) \leq h(A) \leq |A|$. Thus entropy constrains the size of a generating partition.

Examples.
Let $(T, X) = (S^1, z \mapsto \lambda z)$. Then $h(T) = 0$.

**Proof.** Take a pair of intervals as the partition $A$. Then $A(T, n)$ consists of at most $2n$ intervals, since the endpoints of the original intervals give at most $2n$ possible endpoints. The entropy of such a partition is maximized when the intervals have equal length; thus

$$\frac{h(A(T, n))}{n} \leq \frac{2n \cdot (2n)^{-1} \log 2n}{n} \to 0.$$  

**Theorem 8.11** Let $(T, X) = (\sigma, \Sigma_d)$. Then $h(T) = \log d$.

**Proof.** Let $A$ be the partition into $d$ blocks according to the first symbol. Then $A$ is a generating partition, $A(T, n)$ consists of $d^n$ blocks each of measure $d^{-n}$, and thus

$$\frac{h(A(T, n))}{n} = \frac{d^n \cdot d^{-n} \log d^n}{n} = \log d.$$

**Bernoulli shifts.** The measure space $B(p_1, \ldots, p_n)$ is $\Sigma_n$ with the product measure assigning probability $p_i$ to the $i$th symbol ($\sum p_i = 1$). The Bernoulli shift is the dynamical system in which the shift $\sigma$ acts on $B(p_1, \ldots, p_n)$. Its entropy is given by:

$$h(\sigma) = \sum -p_i \log p_i.$$  

**Theorem 8.12** (Ornstein) *Entropy is a complete invariant for Bernoulli shifts; two are measurably conjugate iff their entropies agree.*

**Toral automorphisms.** Let $T \in \text{SL}_n \mathbb{Z}$ represent a toral automorphism with eigenvalues $\lambda_i$. Then it can be shown that

$$h(T) = \sum_{|\lambda_i| > 1} \log |\lambda_i|.$$  

Conjecture: there is a constant $h_0$ such that $h(T) > h_0 > 0$ for any positive entropy toral automorphism of any dimension.

This is directly related to Lehmer’s conjecture, that for an algebraic integer $\lambda$, the product of the conjugates of $\lambda$ of norm greater than one is bounded away from one.
Theorem 8.13 If $h(T) > 0$ then the spectral decomposition of $U_T$ includes Lebesgue spectrum of infinite multiplicity.

See [Par, p.71].

9 Mixing of unitary representations of $\text{SL}_n \mathbb{R}$.

References for this section: [Zim].

Mixing and ergodicity of unitary representations. Let $G$ be a Lie group and let $L \subset G$ be a closed subgroup. Then $L$ acts unitarily on $H = L^2(\Gamma \backslash G)$ for any lattice $\Gamma$. Ergodicity and mixing of this action can be translated into facts about this unitary representation, and then generalized to all unitary representations. In this dictionary we get the following translations:

**Ergodicity.** Any $L$-invariant vector is $G$-invariant. (Since in the lattice case, only constants are $G$-invariant.)

**Mixing.** Suppose $H$ has no $G$-invariant vectors. (This is like restricting to $L^2_0(\Gamma \backslash G)$. Then for any $f_1, f_2 \in H$, we have $\langle gf_1, f_2 \rangle \to 0$ as $g \to \infty$ in $L$ (leaves compact subsets).

We will reprove these more general forms of mixing and ergodicity for $G = \text{SL}_2 \mathbb{R}$ and $\text{SL}_n \mathbb{R}$.

The $N$-trick for $\text{SL}_2 \mathbb{R}$.

**Theorem 9.1** For any unitary representation of $G = \text{SL}_2 \mathbb{R} = KAN$, any $N$-invariant vector is $G$-invariant.

Note: By considering the action on $L^2_0(X)$ this gives another proof of ergodicity of the horocycle flow.

**Proof.** Let $f(g) = \langle gv, v \rangle$ where $v$ is a unit vector. Note that by unitarity, $gv = v$ iff $f(g) = 1$. Suppose $v$ is $N$-invariant; then $f(ngn') = f(g)$. Therefore $f$ is an $N$-invariant function on $\mathbb{R}^2 - \{0\} = G/N$ (since $N$ is the stabilizer of $(1,0)$ for the standard action on $\mathbb{R}^2$).

Now $N$ acts by shearing on $\mathbb{R}^2$; its orbits are the single points on the $x$-axis (with $(1,0)$ corresponding to the identity in $G$), and the horizontal lines other than the $x$-axis. Thus $f$ is constant on horizontal lines, and hence constant on the $x$-axis. But since $f(id) = 1$, $f = 1$ on the whole $x$-axis.

Now $A$ acts on $\mathbb{R}^2$ by real dilations, so $A \cdot id$ corresponds to the positive real axis; therefore $v$ is $A$-invariant. It follows that $f$ is also constant along the orbits of $A$, that is along rays through the origin in $\mathbb{R}^2$. Combined with
constancy along horizontal lines we conclude that $f$ is constant on $\mathbb{R}^2$, and thus $v$ is $G$-invariant.

Remark: In the same spirit, one can give a more formal proof of Hedlund’s theorem on the minimality of the horocycle flow on a compact surface. Namely given a minimal $N$-invariant set $F \subset T^1Y = \Gamma \backslash G$, consider the closed set of $g \in G$ such that $gF \cap F \neq \emptyset$. This set projects to a closed $N$-invariant set $S \subset H = \mathbb{R}^2 - \{0\} = G/N$. Since $F$ is minimal but not just a single orbit, $(1,0) = [\text{id}]$ is not isolated in $S$, and from $N$-invariance this implies (as in the $N$-trick) that $A/N \subset S$. Then $F$ is $A$-invariant and hence dense in $T^1Y$.

The $A$-trick for $\text{SL}_2 \mathbb{R}$.

**Theorem 9.2** An $A$-invariant vector $\xi$ is also $N$-invariant. Hence $\xi$ is $G$-invariant.

**Proof.** For any $n \in \mathbb{N}$ we can find $a_i \in A$ $a_i^{-1}na_i \to \text{id}$. Then if $\xi$ is $A$-invariant and of norm 1 we have

$$\langle n\xi,\xi \rangle = \langle na_i\xi,a_i\xi \rangle = \langle a_i^{-1}na_i\xi,\xi \rangle \to \langle \xi,\xi \rangle = 1.$$ 

Thus $\xi$ is $N$-invariant. By the preceding result, $\xi$ is $G$-invariant.

Note: this passage from $A$ to $N$ invariance is sometimes called the Mauntner phenomenon. An alternative way to finish the proof is to apply the $A$-trick to $N$ and $N^t$ and use the fact that $\text{SL}_2 \mathbb{R} = N^tAN$.

**Theorem 9.3** $G = KAK$ is mixing iff $A$ is mixing.

Note that our deduction of mixing of the horocycle flow from mixing of the geodesic flow is a special case.

**Proof.** Suppose $g_n = k_na_nk'_n \to \infty$ in $G$. Then

$$\langle k_n a_n k'_n x, y \rangle = \langle a_n k'_n x, k_n y \rangle.$$ 

Passing to a subsequence, we can assume $k_n \to k$ and $k'_n \to k'$. But then

$$\langle g_n x, y \rangle \approx \langle a_n k' x, k y \rangle$$

for $n$ large, and the latter tends to zero. Thus $G$ is mixing.
Theorem 9.4 (Howe-Moore, special case) Any unitary representation of $\text{SL}_2 \mathbb{R}$ either has invariant vectors, or is mixing.

Proof. First note that the action $f(x) \mapsto f(x + t)$ of $\mathbb{R}$ on $L^2(\mathbb{R})$ is mixing. Similarly, $f(x) \mapsto f(ax)$ gives a mixing action of $\mathbb{R}^*$ on $L^2(\mathbb{R}, dx/x)$. The idea of the proof is to find this fundamental source of mixing inside any action of $\text{SL}_2 \mathbb{R}$.

Suppose $G = \text{SL}_2 \mathbb{R}$ acts on $H$ with no invariant vectors. Then there are no $N$-invariant vectors. Since $G = KAK$, it suffices to show the action of $A$ is mixing.

First decompose $H$ into irreducible representations of $N \cong \mathbb{R}$. Then

$$H = \int_\mathbb{R} H_x \, d\mu(x)$$

for some spectral measure class $\mu$, where we have identified $\hat{N}$ with $\mathbb{R}$. Since $H$ has no $N$-invariant vectors, $\mu(0) = 0$.

We now view the action of $A \cong \mathbb{R}^*$ on $H$ in terms of the spaces $H_x$. Note that $AN$ is just the semidirect product of $A = \mathbb{R}^*$ acting on $N = \mathbb{R}$ by multiplication. Thus $A$ also acts on $\hat{N} \cong \mathbb{R}$ by multiplication, sending a character $x$ to $ax$. Thus $a(H_x) = H_{ax}$ for $a \in A$.

Now consider $f$ and $g$ in $H$. Let $|f|(x)$ denote the norm of $f$ in $H_x$, and similarly for $g$. Then

$$\langle af, g \rangle \leq \int |af| \cdot |g| \, d\mu = \int |f|(ax) \cdot |g|(x) \, d\mu.$$

Since $\mu$ has no atom at $x = 0$, most of the support of $|f|$ is translated off of most of the support of $|g|$ when $a \to \infty$, and thus $A$ is mixing. (More formally, approximate $f$ and $g$ by vectors whose spectral measures have compact support; then the supports are disjoint from a large enough).

Corollary 9.5 For any noncompact closed subgroup $L$ of $G = \text{SL}_2 \mathbb{R}$, and any lattice $\Gamma$, the action of $L$ on $X = \Gamma \backslash G$ is mixing.

Example: the map $T : X \to X$ given by applying the geodesic flow for time $t > 0$ is mixing.

Semisimple groups. A Lie group $G$ is simple if $G/G^0$ is finite, $G^0$ has finite center, and $G^0$ has no proper normal subgroup of positive dimension. (Here $G^0$ is the connected component of the identity).

A Lie group $G$ is semisimple if $G/G^0$ is finite, and $G^0$ is finitely covered by a product of simple groups.

Examples: $\text{SL}_n \mathbb{R}, n \geq 2$; $\text{SO}(n), n \geq 3$; $\text{SO}(n,1), n \geq 2$; etc.
Theorem 9.6 (Howe-Moore) Let $G = \prod G_i$ be a product of connected, non-compact simple Lie groups. Suppose $\pi$ is a unitary representation of $G$ without $G_i$-invariant vectors. Then $(G, \pi)$ is mixing.

We will prove this theorem for $G = \text{SL}_n \mathbb{R}, n \geq 2$.

Structure of semisimple Lie groups: $\text{KAK}$ and $\text{KAN}$. Let $G$ be a connected semisimple Lie group. Then $G = \text{KAK}$, where $K$ is a maximal compact subgroup and $A$ is a maximal $\mathbb{R}$-torus. This is the polar decomposition.

Examples: for $G = \text{SO}(n, 1)$, the isometries of $\mathbb{H}^n$, this says any two points in $\mathbb{H}^n = G/\text{SO}(n)$ are joined by a geodesic. For $G = \text{SL}_n \mathbb{R}$, this says any two unit volume ellipsoids (points in $\text{SL}_n \mathbb{R}/\text{SO}_n \mathbb{R}$) are related by a rotation and an affine stretch. (Note on the other hand that a typical solvable group cannot be expressed as $\text{KAK}$.)

We can similarly write $G = \text{KAN}$ where $N$ is a maximal unipotent subgroup associated to $A$. This is the Iwasawa decomposition. For $\text{SL}_n \mathbb{R}$, $N$ is the group of upper-triangular matrices with 1’s on the diagonal. For $\text{SO}(n, 1)$, $N$ is the group of horocycle flows for the horocycle at the positive end of the geodesic corresponding to $A$.

To further analyze $\text{SL}_n \mathbb{R}$ it is also useful to consider the parabolic subgroup

$$P = \left\{ \begin{pmatrix} I_{n-1} & x \\ 0 & I_1 \end{pmatrix}, \ x \in \mathbb{R}^{n-1} \right\} \subset \text{SL}_n \mathbb{R}.$$  

Then

$$AP \cong (\mathbb{R}^*)^{n-1} \ltimes \mathbb{R}^{n-1}$$

is a solvable group, with $A$ acting multiplicatively on $P$.

Unitary representations of $\text{SL}_n \mathbb{R}$. The $P$-trick.

Consider a unitary representation of $\text{SL}_n \mathbb{R}$. Recall $P \cong \mathbb{R}^{n-1}$ is the subgroup of shears $(x, x_n) \mapsto (x, x_n + L(x))$.

Theorem 9.7 Any $P$-invariant vector $\xi$ is also $\text{SL}_n \mathbb{R}$-invariant.

Proof. We have seen this for $n = 2$, where $P = N$. Now for any $1 < i \leq n$, we can embed $\text{SL}_2 \mathbb{R}$ as $G_i \subset \text{SL}_n \mathbb{R}$ so it acts on the subspace $\mathbb{R}e_i \oplus \mathbb{R}e_n$. Then $N \subset \text{SL}_2 \mathbb{R}$ maps into $P$, so by the result for $n = 2$ we conclude that $\xi$ is invariant under $G_i$. Since the $G_i$ generate $G$, we conclude that $\xi$ is $G$-invariant. 

93
For completeness we also record:

**Theorem 9.8** Any $A$-invariant vector is also $SL_n \mathbb{R}$-invariant.

**Proof.** As above, $\xi$ is invariant under each $G_i$ by the $A$-trick for $SL_2 \mathbb{R}$, and these generate.$\blacksquare$

**Theorem 9.9** Any unitary representation of $SL_n \mathbb{R}$ without invariant vectors is mixing.

**Proof.** Decompose the representation relative to $P$, obtaining a spectral measure $\mu$ on $\hat{P} \cong \mathbb{R}^{n-1}$. Since there are no invariant vectors, we have $\mu(0) = 0$. For any measurable set $E \subset \mathbb{R}^{n-1}$ we have as before $a(H_E) = H_{aE}$, where $a \in A$ acts multiplicatively on $P$. And by the KAK decomposition, it suffices to show $A$ is mixing.

Now it is clear that for any compact set $E \subset (\mathbb{R}^*)^{n-1}$, the set of $a \in A$ such that $aE \cap E \neq \emptyset$ is compact. Thus by considering the spectral measure of a vector $\xi$, we can conclude that $\langle a\xi, \xi \rangle \to 0$ as $a \to \infty$ in $A$ so long as the measure has no support in a coordinate hyperplanes.

On the other hand, if $\mu$ assigns positive measure to the $i$th hyperplane, then the corresponding subspace of $H$ is invariant under the 1-parameter subgroup $P_i \subset P$. In particular, $P_i$ has an invariant vector $\xi$.

Now $P_i \subset G_i \cong SL_2 \mathbb{R}$, where $G_i$ acts on $\mathbb{R}e_i \oplus \mathbb{R}e_n$. Thus $\xi$ is also $G_i$-invariant, by the $N$-trick (since $P_i = N \subset G_i$). Hence $\xi$ is $A_i$-invariant, where $A_i = G_i \cap A$.

Let $H' \subset H$ be the closed subspace of all vectors fixed by $A_i$. If we can show $H'$ is $G$-invariant, then we are done, since $\pi|H'$ has $A_i$ in its kernel, and thus by simplicity $G$ acts trivially on $H'$ (contrary to our hypothesis that there are no $G$-invariant vectors).

Certainly $H'$ is $A$-invariant, since $A_i$ and $A$ commute. For $j \neq k$ let $B_{jk}$ be the subgroup with one off-diagonal entry in position $(j, k)$. Then $A_i$ normalizes $B_{jk}$; either they commute or $A_iB_{jk}$ is isomorphic to $\mathbb{R}^* \times \mathbb{R}$. In the former case $H'$ is $B_{jk}$-invariant, and in the latter case as well, by the $A$-trick.

Since $A$ and $\bigcup B_{jk}$ generate $G$, $H'$ is $G$-invariant and we are done.$\blacksquare$
10 Amenability

Reference for this section: [Gre].

Amenable groups. Let $G$ be a discrete group. A mean is a linear functional $m : L^\infty(G) \to \mathbb{R}$ on the space of bounded real-valued functions such that $m(1) = 1$ and $f \geq 0 \implies m(f) \geq 0$. It is the same as a finitely additive probability measure.

A group is amenable iff it admits a $G$-invariant mean, for the right (or left) action $(g \cdot f)(x) = f(xg)$.

Basic facts.

• Finite groups are amenable.
  The integers $\mathbb{Z}$ are amenable. To see this, take any weak* limit of the means $m_n(f) = \frac{1}{n} \sum_1^n f(i)$.

• An amenable group is ‘small’: If $G$ is amenable then any quotient $H = G/N$ is amenable. (Pull back functions from $H$ to $G$ and average there.)

• If $H \subset G$ and $G$ is amenable, then so is $H$. (Foliate $G$ by cosets of $H$, choose a transversal and use it to spread functions on $H$ out to $G$.)

• If $0 \to A \to B \to C \to 0$ and $A$ and $C$ are amenable, then so is $B$. (Given a function on $B$, average over cosets of $A$ to obtain a function on $C$, then average over $C$.)

• $G$ is amenable iff every finitely generated subgroup of $G$ is amenable. (For a finitely generated subgroup $H$, we get an $H$-invariant mean on $G$ with the aid of a transversal. In the limit as $H$ exhausts $G$ we get an invariant mean on $G$.)

• Abelian groups are amenable (since finitely generated ones are). Solvable groups are amenable.

Theorem 10.1 The free group $G = \langle a, b \rangle$ is not amenable.

Proof. Since every (reduced) word in $G$ begins with one of $\{a, b, a^{-1}, b^{-1}\}$, we can assume (changing generators) that $m(A) > 0$, where $A$ consists of words beginning with $A$. But $A \supset bA \sqcup b^{-1}A$, so by invariance we would have $m(A) \geq 2m(A)$, a contradiction. \hfill \blacksquare
Corollary 10.2 The fundamental group of a closed surface of genus \( g \geq 2 \) is nonamenable.

Theorem 10.3 The space of means is the closed convex hull of the set of \( \delta \)-functions on \( G \).

Proof. Let \( m \) be a mean. We need to show there exists a net of probability measures \( m_\alpha \) of finite support such that \( m_\alpha(f) \to m(f) \) for every \( f \in L^\infty(G) \).

Consider any finite partition \( \alpha = (G_1, \ldots, G_n) \) of \( G \), and set \( m_\alpha = \sum a_i \delta_i \), where \( a_i = m(G_i) \) (the mean of the indicator function). Then \( m_\alpha(f) = m(f) \) for any simple function which is constant on each blocks \( G_i \) of \( \alpha \). Since simple functions are dense in \( L^\infty(G) \), we have \( m_\alpha(f) \to m(f) \) for all \( f \), where the \( \alpha \) are partially ordered by refinement.

The Cayley graph. Given a finitely generated group \( G \), together with a choice of generating set \( E \), we can build the Cayley graph \( \mathcal{G} \) whose vertices \( V = G \) and whose edges connect elements differing by a generator. We will usually assume \( E^{-1} = E \) in which case the edges are undirected.

If \( V_0 \subseteq V \) is a subset of the Cayley graph, we define its boundary \( \partial V_0 \) to be the vertices connected to, but not lying in, \( V_0 \).

The isoperimetric constant of a graph is given by

\[
\gamma(\mathcal{G}) = \inf \frac{|\partial V_0|}{|V_0|}
\]

where the infimum is over all finite sets.

Theorem 10.4 (Følner’s condition) Let \( G \) be a finitely generated group. Then \( G \) is amenable iff the isoperimetric constant of its Cayley graph is zero.

Proof (Namioka). Suppose we can find finite sets \( G_i \subset G \) such that \( |\partial G_i|/|G_i| \to 0 \). Then any weak* limit of the mean over \( G_i \) gives a \( G \)-invariant mean.

Conversely, suppose \( m \) is a mean on \( G \). Then by density of finite measures we can write \( m = \lim m_\alpha \), where \( m_\alpha(f) = \sum f_\alpha(x) f(x) \) and \( f_\alpha \geq 0 \) is a function with finite support. Since \( m \) is \( G \)-invariant, \( m_\alpha \) is nearly so, and thus for any \( g \in G \), \( \|g f_\alpha - f_\alpha\|_1 \to 0 \) as \( \alpha \to \infty \).

Now comes the trick: imagine slicing the graph of \( f_\alpha \) into horizontal strips. That is, write \( f_\alpha = \sum a_i \chi_{G_i}/|G_i| \), where \( a_i > 0 \), \( \sum a_i = 1 \) and
\[ G_1 \subset G_2 \subset \cdots \subset G_n \] are finite sets. Because of nesting, for any \( x \in G \) the quantities \( \chi_{G_i}(x) - \chi_{G_i}(gx) \) all have the same sign. Thus

\[ \|gf - f\|_1 = \sum a_i \frac{|G_i \Delta gG_i|}{|G_i|}. \]

Summing over a finite generating set \( F \) for \( G \), we conclude that most terms in this convex combination must be small. Thus there exists an \( i \) such that

\[ \sum_{g \in F} \frac{|G_i \Delta gG_i|}{|G_i|} \]

is small. Therefore \( G_i \) is nearly invariant under the generating set \( F \).

**Exponential growth.** By the Følner property, any nonamenable group has exponential growth, since the boundary of any ball has size comparable to the ball itself. But there are also solvable (and hence amenable) groups of exponential growth; e.g. \( \langle a, b : ab = b^2a \rangle \). For such a group the Følner sets cannot be chosen to be balls. (Q. What do the Følner sets look like?)

**Almost invariance.** A unitary representation of \( G \) on a Hilbert space \( H \) has almost invariant vectors if for any compact set \( K \subset G \) and \( \epsilon > 0 \), there is a \( v \in H \) with \( \|v\| = 1 \) and \( \|g \cdot v - v\| < \epsilon \) for all \( g \in K \).

Equivalently, \( \langle g \cdot v, v \rangle \approx 1 \).

**Theorem 10.5** \( G \) is amenable iff \( L^2(G) \) has almost invariant vectors.

**Proof.** If \( G \) is amenable, then by Følner there are finite sets \( G_i \) that are almost invariant. Set \( f_i = \chi_{G_i}/|G_i|^{1/2} \) so \( \|f_i\| = 1 \). Then as \( i \to \infty \),

\[ \langle gf_i, f_i \rangle = \frac{|g \cdot G_i \cap G_i|}{|G_i|} \to 1, \]

so \( L^2(G) \) has almost invariant vectors.

Conversely, if \( f_i \) is almost invariant, then \( m_i(h) = \langle hf_i, f_i \rangle \) for \( h \in L^\infty(G) \) is an almost-invariant mean; taking a weak* limit we conclude that \( G \) is amenable.

**Amenability and Poincaré series.**

**Theorem 10.6** Let \( Y \to X \) be a covering of a compact Riemann surface \( X \) of genus \( g \geq 2 \). Then the Poincaré operator \( \Theta_{Y/X} : Q(Y) \to Q(X) \) satisfies \( \|\Theta_{Y/X}\| < 1 \) if and only if the covering is nonamenable. In particular, \( \|\Theta\| < 1 \) for the universal covering.
This result has applications to the construction of hyperbolic structures on 3-manifolds, following Thurston.

11 The Laplacian

References for this section: [Bus], [Lub], [Sar].

Riemannian manifolds. The Laplacian on functions is defined by

\[ \Delta f = - * d * df. \]

It is the self-adjoint operator associated to the quadratic form \( \langle \nabla f, \nabla f \rangle \), in the sense that

\[ \int_M |\nabla f|^2 = \int_M f \Delta f \]

for any compactly supported function.

For a compact manifold \( M \) the space \( L^2(M) \) has a basis of eigenfunctions of the Laplacian.

For general manifolds we can define the least eigenvalue of the Laplacian by minimizing the Ritz-Rayleigh quotient

\[ \lambda_0(M) = \inf \frac{\int |\nabla f|^2}{\int |f|^2}. \]

If \( M \) has finite volume then \( \lambda_0 = 0 \) and we define \( \lambda_1(M) \) by minimizing the above subject to \( \int f = 0 \).

These definitions agree with the usual eigenvalues on a compact manifold but give only the discrete spectrum in general (even in the case of finite volume).

Behavior under coverings. If \( Y \to X \) is a covering map, where \( X \) and \( Y \) have finite volume, then \( \lambda_1(Y) \leq \lambda_1(X) \), since any test function on \( X \) can be pulled back to \( Y \). It is a challenge to find an infinite tower of coverings with \( \lambda_1(Y_i) \) bounded away from zero.

Note that \( \lambda_1 \) can drop precipitously: if \( X \) is a hyperbolic surface with a very short non-separating geodesic \( \gamma \), it can still happen that \( \lambda_1(X) \) is large; but there is a \( \mathbb{Z}/2 \) covering space \( \pi: Y \to X \) such that \( \pi^{-1}(\gamma) \) cuts \( Y \) into two equal pieces, so \( \lambda_1(Y) \) is small.

The Euclidean Laplacian. On \( \mathbb{R}^n \) we have

\[ \Delta f = - * d * df = - \sum \frac{\partial^2 f}{\partial^2 x_i}. \]
It is easy to see $\lambda_0(\mathbb{R}^n) = 0$ for all $n$. In fact the Laplacian has continuous spectrum going down to zero.

On the other hand, on a circle $C_r$ of radius $r$ we have $\lambda_1(C_r) = 1/r^2$ by considering the function $\sin(x/r)$, and the full spectrum is of the form $n^2/r^2$, $n \geq 0$.

The continuous spectrum of $\mathbb{R}$ can be thought of as the limit of the discrete spectra of $C_r$ as $r \to \infty$.

**The hyperbolic Laplacian.** On $\mathbb{H}$ with the metric $ds^2 = (dx^2 + dy^2)/y^2$ of constant curvature $-1$, we have

$$\Delta f = -*d * df = -y^2 \left( \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \right) = (z - \bar{z})^2 \frac{\partial^2 f}{\partial z \partial \bar{z}}.$$  

This is the positive operator associated to the positive definite quadratic form $\int |\nabla f|^2 = \langle \Delta f, f \rangle$. The sign comes from Stokes’ theorem.

Basic examples: $\Delta y^s = s(1 - s)y^s$.

Bounded primitive for volume: for $\omega = -dx/y$ on $\mathbb{H}$ we have $|\omega(v)| \leq 1$ for any unit vector $v$, and $d\omega = dx \, dy / y^2$ is the hyperbolic area element. From this follows that hyperbolic space is nonamenable in a continuous sense. More precisely:

**Theorem 11.1** We have

$\lambda_0(\mathbb{H}) \geq 1/4$, and $\ell(\partial \Omega) \geq \text{area}(\Omega)$ for any compact smooth region $\Omega \subset \mathbb{H}$.

**Proof.** For the isoperimetric inequality just note that $\text{area}(\Omega) = \int_{\partial \Omega} d\omega = \int_{\partial \Omega} \omega \leq \ell(\partial \Omega) \sup |\omega| = \ell(\partial \Omega)$.

For the bound on $\lambda_0$, let $f : \mathbb{H} \to \mathbb{R}$ be compactly supported; then

$$d(f^2 \omega) = f^2 \, d\omega + 2f \, df \, \omega,$$

so by Stokes’ theorem and Cauchy-Schwarz we have

$$\left( \int f^2 d\omega \right)^2 = \left( \int 2f \, df \, \omega \right)^2 \leq 4 \left( \int f^2 \right)^2 \left( \int |df|^2 \right)^2,$$

from which $\int |df|^2 \geq (1/4) \int f^2$. \qed

99
Remark. It can be shown that $\lambda_0(\mathbb{H}) = 1/4$.

Pyramid schemes. On an infinite regular tree of degree $d \geq 3$, it is similarly possible to define a unit speed flow with definite divergence at every vertex. This is directly linked to nonamenability of the graph. Indeed, on an amenable graph, such a pyramid scheme would increase the average wealth per vertex (defined using the mean).

Cheeger’s constant. We will now see that an isoperimetric inequality and a lower bound on $\lambda_0$ always go hand in hand.

Given a Riemannian $n$-manifold $(M, g)$ its Cheeger constant is defined by

$$h(M) = \inf_X \frac{\text{area}(X)}{\min(\text{vol}(A), \text{vol}(B))},$$

where the infimum is over all co-oriented compact separating hypersurfaces $X \subset M$, and where $M - X = A \sqcup B$. (Although $X$ may cut $M$ into more than 2 pieces, the orientation of the normal bundle of $X$ puts them into 2 classes.)

If $\text{vol}(M) = \infty$ then the Cheeger constant reduces to the isoperimetric constant

$$h(M) = \inf_A \frac{\text{area}(\partial A)}{\text{vol}(A)},$$

where the inf is over all compact submanifolds $A$.

The constant $h(M)$ is not a conformal invariant; it is homogeneous of degree $-1$ under scaling the metric.

We will prove the following:

**Theorem 11.2 (Cheeger)** For any Riemannian manifold $M$, we have

$$\lambda_i(M) \geq h(M)^2/4,$$

where $i = 0$ for infinite volume and $i = 1$ for finite volume.

A continuous isoperimetric inequality.

**Theorem 11.3** Suppose $M$ is infinite volume and $f : M \to [0, \infty)$ is a compactly supported smooth function. Then we have:

$$\int |\nabla f| dV \geq h(M) \int f dV.$$
Proof. Let $A_t = f^{-1}[t, \infty)$, let $X_t = \partial A_t = f^{-1}(t)$, and let $V(t) = \text{vol}(A_t)$.

The coarea formula says that for a proper function,

$$\int_{f^{-1}[a,b]} |\nabla f| \, dV = \int_a^b \text{area}(X_t) \, dt.$$

To understand this formula, just note that the width of the region between $X_t$ and $X_{t+\epsilon}$ is approximately $\epsilon/|\nabla f|$. Since $f$ is compactly supported we have

$$\int |\nabla f| \, dV = \int_0^\infty \text{area}(X_t) \, dt \geq h(M) \int_0^\infty \text{vol}(A_t) \, dt = h(M) \int_0^\infty -tV'(t) \, dt = h(M) \int |f| \, dV,$$

where we have integrated by parts and used the fact that $-V'(t) \, dt$ is the push-forward of the volume form from $M$ to $\mathbb{R}$.

Proof of Cheeger’s theorem for infinite volume and $\lambda_0(M)$.

Let $F : M \to \mathbb{R}$ be a compactly supported smooth function on $M$; we need to bound $\inf |\nabla F|^2$ from below. To this end, let $f = F^2$, so $\nabla f = 2F \nabla F$. Then we have

$$h(M)^2 \left( \int |F|^2 \right)^2 = \left( h(M) \int |f| \right)^2 \leq \left( \int |\nabla f| \right)^2 \leq 4 \int |F|^2 \int |\nabla F|^2,$$

and thus

$$\frac{h(M)^2}{4} \leq \frac{\int |\nabla F|^2}{\int |F|^2}.$$

The case of finite volume.

Theorem 11.4 Let $f : M \to \mathbb{R}$ be a smooth function, and assume the level set of 0 cuts $M$ into 2 pieces of equal volume. Then

$$\int_M |\nabla f| \, dV \geq h(M) \int_M |f| \, dV.$$

101
Proof. Let $X_t = f^{-1}(t)$. Let $M - X_t = A_t \cup B_t$ where $A_t = f^{-1}(t, \infty)$. Then for $t > 0$, $A_t$ is the smaller piece, so we have

$$V(t) = \text{vol}(A_t) \leq h(M) \text{area}(X_t).$$

The proof then follows the same lines as for the case of infinite volume, by applying the coarea formula to $f$. That is, we find

$$\int_{A_0} |\nabla f| dV \geq h(M) \int_0^\infty V(t) dt = \int_0^\infty -tV'(t) dt = \int_{A_0} |f|,$$

and similar for $B_0$.

Proof of Cheeger’s inequality for finite volume. Let $f : M \to \mathbb{R}$ satisfy $\int f = 0$ and $\int f^2 = 1$; we need to bound $\int |\nabla f|$ from below. Since $\int (f + c)^2 \geq \int f^2$, we can change $f$ by a constant and trade the assumption $\int f = 0$ for the assumption that $f^{-1}(0)$ cuts $M$ into two pieces of equal size.

Rescaling so $\int f^2 = 1$ only reduces $\int |\nabla f|$.

Now let $F(x) = f(x)|f(x)|$. Then $\int |F| = 1$ and $F^{-1}(0)$ also cuts $M$ into 2 pieces of equal size. Thus

$$\int |\nabla F| \geq h(M) \int |F| = h(M).$$

But $|\nabla F| = 2|f\nabla f|$, so by Cauchy-Schwarz we have

$$h(M)^2 \leq \left( \int |\nabla F| \right)^2 = \left( \int 2|f| |\nabla f| \right)^2 \leq 4 \int |f|^2 \int |\nabla f|^2 = 4 \int |\nabla f|^2,$$

hence the Theorem.

Theorem 11.5 For a closed hyperbolic surface of fixed genus $g$, $\lambda_1(X)$ is small iff $X$ has a collection of disjoint simple geodesics $\gamma_1, \ldots, \gamma_k$, $k \leq 3g - 3$, such that the length of $S = \bigcup \gamma_i$ is small and $S$ separates $X$.

Proof. If $\lambda_1(X)$ is small then by Cheeger there is a separating 1-manifold $S$ such that $\text{length}(S)/\text{area}(A)$ is small, where $A$ is the smaller piece of $X - S$. Some of the curves have to be essential, else we would have $\text{area}(A) \ll \text{length}(S)$. These essential pieces are homotopic to geodesics of the required form.

Conversely, a collection of short geodesics has a large collar, which makes $\lambda_1(X)$ small if it is separating.
The combinatorial Laplacian. On a regular graph $G = (V, E)$ of degree $d$, we define the Laplacian for functions $f : V \rightarrow \mathbb{R}$ by

$$(\Delta f)(x) = f(x) - \frac{1}{d} \sum_{y \sim x} f(y).$$

The sum is over the $d$ vertices $y$ adjacent to $x$. (A different formula is more natural for graphs with variable degree, and this is why our normalization differs from that in [Lub].)

The combinatorial gradient $|\nabla f|$ is a function on the edges of $G$ defined by $|\nabla f|(e) = |f(x) - f(y)|$ if $e = \{x, y\}$.

**Theorem 11.6** For a regular graph $G = (V, E)$ of degree $d$,

$$\frac{1}{2d} \sum_{E} |\nabla f|^2 = \langle f, \Delta f \rangle = \sum_{V} f(x) \Delta f(x).$$

**Proof.** We have

$$\sum |\nabla f|^2 = \sum_{x \sim y} |f(x) - f(y)|^2 = 2d \sum_{x} |f(x)|^2 - 2 \sum_{x \sim y} f(x) f(y)$$

$$= 2d \sum_{x} f(x) \left( f(x) - \frac{1}{d} \sum_{x \sim y} f(x) f(y) \right) = 2d \langle f, \Delta f \rangle.$$

**The maximum principle.** On a finite connected graph, the only harmonic functions are constant; just note that $f(x)$ must be equal to $f(y)$ on the neighbors of any point $x$ where the maximum of $f$ is achieved.

**Random walks.** Let $x_n$ be a random walk on a regular graph $G$. Then the expected value of a function $f_0$ after $n$ steps is given by

$$f_n(x) = E(f(x_n)) = (I - \Delta)f_{n-1}(x) = (I - \Delta)^n f_0(x).$$

If $G = (V, E)$ is finite and connected, we can find a basis $\phi_i$ for $L^2(V)$ with $\Delta \phi_i = \lambda_i \phi_i$, $0 = \lambda_0 < \lambda_1 < \ldots$. If $f_0 = \sum a_i \phi_i$ then

$$f_n = \sum a_i (1 - \lambda_i)^n \phi_i.$$

Thus for $f_0$ orthogonal to the constants we have:

$$\|f_n\|_2 \leq (1 - \lambda^*)^n \|f_0\|_2,$$
where \((1 - \lambda^*) = \sup_{i \neq 0} |1 - \lambda_i|\); frequently \(\lambda^* = \lambda_1\) (a counterexample is when \(G\) is bipartite).

In other words, \(\lambda^*\) measures the rate at which a random walk diffuses towards the uniform distribution on the vertices of \(G\).

**The heat equation.** In the continuum limit we have Brownian motion \(x_t\) on a compact Riemannian manifold, and the random walk equation for \(f_t(x) = E(f_0(x_t))\) becomes

\[
\frac{df_t}{dt} = -\Delta f_t.
\]

If \(f_0 = \sum a_i \phi_i\) then

\[
f_t = \sum a_i \exp(-\lambda_i t) \phi_i
\]

and we see \(\lambda_1\) gives exactly the rate of decay of \(f_t\) towards its mean value.

**The Poisson random walk on a graph.** To mimic the heat equation more closely, we can consider the equation \(df_t/dt = -\Delta f_t\) on a graph. Since \(\exp(-t\Delta) = \lim(I - \Delta/n)^n\), we have a dual diffusion process \(x_t\) that can be thought of as a limit of random walks. In this process, one waits at a given vertex until a Poisson event occurs, then one takes a random step to a new vertex.

For this process, \(\lambda_1\) is exactly the rate of diffusion.

The Cheeger constant for \(G\) is defined by

\[
h(G) = \inf_{A,B} \frac{|E(A,B)|}{\min(|A|,|B|)}
\]

where \(V = A \cup B\) is a partition of \(V\), and \(E(A,B)\) is the set of edges joining \(A\) to \(B\).

**Theorem 11.7** For a regular graph of degree \(d\),

\[
\lambda_i(G) \leq \frac{h(G)}{2d},
\]

where \(i = 0\) when \(G\) is infinite and \(i = 1\) when \(G\) is finite.

**Proof.** Suppose \(G\) is infinite. Consider a partition of the vertices \(V = A \cup B\) with \(|A|\) finite, and let \(f = \chi_A\). Then \(\sum f^2 = |A|, \sum |\nabla f|^2 = |E(A,B)|\), and thus

\[
\lambda_0(G) \leq \frac{\langle f, \Delta f \rangle}{\langle f, f \rangle} = \frac{\sum |\nabla f|^2}{2d\langle f, f \rangle} = \frac{|E(A,B)|}{2d|A|}.
\]

Taking the inf over all partitions yields the theorem. The case of \(G\) finite is similar.
**Theorem 11.8** For a regular graph of degree $d$,

$$\lambda_i(G) \geq \frac{h(G)^2}{2d^2},$$

where $i = 0$ if $|G| = \infty$ and $i = 1$ if $|G| < \infty$.

Cf. [Ch, Chap. 2] or [Lub, Prop. 4.2.4]. Note that the former has a different normalization for Cheeger's constant, while the latter has a different normalization for the Laplacian!

For an infinite regular graph, $h(G)$ is the same (to within a factor depending on $d$) as the expansion constant $\gamma$ introduced earlier in the definition of amenability.

**Theorem 11.9** A finitely-generated group $G$ is amenable iff $\lambda_0(G) = 0$, where $G$ is its Cayley graph.

**Proof.** If $G$ is amenable then the indicator functions of Følner sets show $\lambda_0 = 0$. Conversely, if $\lambda_0(G) = 0$ then $h(G) = 0$ and thus $G$ has Følner sets.

**Theorem 11.10** (Brooks, [Br]) Let $\tilde{X}$ be a Galois covering of a compact Riemannian manifold $X$, with deck group $G$. Then $\lambda_0(\tilde{X}) = 0$ iff $G$ is amenable.

**Expanders.** Let $G_n$ be a collection of finite regular graphs of bounded degree, such that the number of vertices tends to infinity, but $h(G_n) > H > 0$. Then $G_n$ are expanders. By Cheeger’s inequality going both ways we have:

**Theorem 11.11** The graphs $G_n$ are expanders iff $\inf \lambda_1(G_n) > 0$.

**Selberg’s theorem:** $\lambda_1(\mathbb{H}/\Gamma(n)) \geq 3/16$ for all $n$. Selberg’s conjecture replaces $3/16$ by $1/4$ and has the more natural significance that the complementary series do not occur in the decomposition of $L^2(\Gamma \backslash G(n))$ into irreducible representations of $G$.

**Examples of expanding graphs.**

(a) Fix generators $T = (\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix})$ and $S = (\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix})$ for $\mathrm{SL}_2(\mathbb{Z})$ and use them to construct the Cayley graphs for $G_n = \mathrm{SL}_2(\mathbb{Z}/n)$.

(b) For $p$ a prime, consider the projective line $P_p = \mathbb{Z}/p \cup \{\infty\}$, and make it into a trivalent graph by joining $x$ to $x + 1$ and $1/x$. 

105
Theorem 11.12  Both these families of graphs are expanders.

Proof.  (a) Suppose to the contrary that $G_n$ has a low eigenvalue; we will show this implies $X(n) = \mathbb{H}/\Gamma(n)$ has a low eigenvalue, contradicting Selberg’s 3/16 theorem.

Tile $X(n)$ by copies of the standard fundamental domain for $\text{SL}_2(\mathbb{Z})$; we can identify the tiles with the vertices of $G_n$. Let $V \subset G_n$ be a set of vertices such $|\partial V|/|V|$ is small and $|V| < |G_n|/2$. Let $f$ be a function on $X(n)$ with $f(x) = 1$ on the tiles corresponding to $V$, and $f(x) = 0$ on the other tiles.

Now at each edge where a 0-tiles meets a 1-tile, modify $f$ on a strip of definite width around the edge so it makes the transition with bounded gradient. This is possible except near the cusps of $X(n)$, where the transition from 0 to 1 takes place very quickly.

To handle these, go out in each 1-tile to the horocycle of length $L$. Make $f$ taper off to zero between the horocycles of length $L$ and $L/2$; this is possible with bounded gradient.

Now $f$ is supported outside a neighborhood of the cusps of $X(n)$. The final adjustment is made on the edges that run to the cusps between 0 and 1 tiles; there we can smooth out $f$ at the price of a gradient of size $1/L$ (since there is a horocycle of length $L$ on which to make the adjustment).

In the end, $\int |f|^2 \asymp |V|$, and

$$\int |\nabla f|^2 = O(L|V| + |\partial V|(1 + 1/L^2)).$$

These two terms account for the tapering off of the 1-tiles, which requires area $O(L)$ for each such tile; and the gradient squared between 0 and 1 tiles, which is $O(1)$ on finite edges and $O(1/L^2)$ near the cusps.

Since $f$ is supported in at most half the area of $X(n)$, we can adjust it by a constant so $\int f = 0$, while keeping $\int |f|^2 \asymp |V|$. Then the ratio $\int |f|^2 / \int |\nabla f|^2$ gives

$$\lambda_1(X(n)) = O \left( L + \frac{|\partial V|}{|V|} \left( 1 + \frac{1}{L^2} \right) \right).$$

Clearly if the isoperimetric constant of $G_n$ can be made very small, then for a suitable choice of $L$, $\lambda_1(X(n))$ is also small, contradicting Selberg. Thus $\langle G_n \rangle$ are expanders.

(b) The projective plane $P_p$ is covered by $G_p$, so it is at least as expanding as $G_p$.  

\[\blacksquare\]
Embedding the nonamenable tree into the finite universe of the Earth, according to Darwin (*Origin of Species*, Chapter 4):

*If during the long course of ages and under varying conditions of life, organic beings vary at all in the several parts of their organisation, and I think this cannot be disputed; if there be, owing to the high geometrical powers of increase of each species, at some age, season, or year, a severe struggle for life, and this certainly cannot be disputed; then, considering the infinite complexity of the relations of all organic beings to each other and to their conditions of existence, causing an infinite diversity in structure, constitutions, and habits, to be advantageous to them, I think it would be a most extraordinary fact if no variation ever had occurred useful to each being’s own welfare, in the same way as so many variations have occurred useful to man.*

12 All unitary representations of $\text{PSL}_2(\mathbb{R})$

Classification of all irreducible unitary representations of $G = \text{PSL}_2(\mathbb{R})$ [GGP].

**Theorem 12.1** Every such representation occurs in the principal, complementary or discrete series.

First consider the representations that might occur in the regular representation on $L^2(\mathbb{H})$. As is the case for $\mathbb{R}$, the irreducible representations should correspond to eigenfunctions of the Laplacian (like $\exp(it)$ on $\mathbb{R}$), since $\Delta$ commutes with the action of $G$. Also we should not expect the eigenfunctions themselves to be in $L^2$. Finally the eigenvalues should be real and positive since $\Delta$ is a positive operator on $L^2$. (Note $\Delta e^{it} = t^2 e^{it}$ by our sign convention.)

To produce eigenfunctions of $\Delta$, consider a conformal density $\mu = \mu(x)|dx|^s$ on $S^1 = \partial \mathbb{H}$. A point $z$ in $\mathbb{H}$ determines a visual metric on $S^1$, which we can use to convert $\mu$ into a function and take its average, resulting in a function $f(z)$. Put differently, if we normalize so $z = 0$ in the disk model, then $\mu = \mu(\theta)|d\theta|^s$ and $f(0) = \int \mu(\theta) d\theta$. Since this operator is natural, it commutes with the action of $G$.

For $\mu = \delta_\infty(x)|dx|^s$ on $\mathbb{R}$, where $\delta_\infty$ is the delta function at the point at infinity, we find $f(z) = Cy^s$ by naturality. We have thus shown:
conformal densities of dimension \(s\) on \(S^1_\infty\) determine eigenfunctions of \(\Delta\) on \(\mathbb{H}\) with eigenvalue \(\lambda = s(1 - s)\).

The principal series. These are the representations that arise in the decomposition of \(L^2(\mathbb{H})\). Since \(\lambda_0(\mathbb{H}) = 1/4\) we only expect \(\lambda \geq 1/4\) to arise, and this corresponds to \(s = 1/2 + it, t \in \mathbb{R}\). Then for any \(s\)-dimensional density \(\mu\), the quantity \(|\mu|^2\) is naturally a measure (since \(|g'z|^it\) has modulus 1 for any diffeomorphism \(g\) of \(S^1\).) These generalize the half-densities.

For any such \(s = 1/2 + it\), let \(H(s)\) be the Hilbert space of conformal densities such that \(\|\mu\|^2 = \int_{S^1} |\mu|^2 < \infty\).

Then \(G\) acts unitarily on \(H(s)\), and the set of all such representations (as \(t\) ranges in \(\mathbb{R}\)) is the principal series.

We have
\[
L^2(\mathbb{H}) = \int_{1/2-i\infty}^{1/2+i\infty} H(s) \, ds.
\]
(The explicit form of this isomorphism is given by the Plancheral formula, which is in turn related to the Mellin transform, the analogue of the Fourier transform for \(\mathbb{R}^+\).)

Note: the ergodic action of \(G\) on \(S^1\) gives rise, by pure measure theory, to a unitary action on half-densities; this is \(H(1/2)\).

Note: \(H(1/2 + it)\) and \(H(1/2 - it)\) are equivalent; otherwise these representations are distinct.

The complementary series. Next consider conformal densities \(\mu\) of dimension \(s\) in the range \(0 < s < 1, s \neq 1/2\). Then \(\mu\) also determines a density \(\mu \times \mu\) on the space of geodesics. In this range we define the Hilbert space \(H(s)\) by requiring that the norm
\[
\|\mu\|^2 = \int_{S^1 \times S^1} (\mu \times \overline{\mu}) \left( \frac{dx \, dy}{|x - y|^2} \right)^{1-s} = \int_{\mathbb{R} \times \mathbb{R}} \frac{\mu(x)\overline{\mu(y)}}{|x - y|^{2s-2}} \, dx \, dy < \infty.
\]

Note that \(dx \, dy/|x - y|^2\) is the invariant measure on the space of geodesics. The integral requires some regularization when \(s > 1/2\).

As \(s \to 0\), the density \(|dx|^s\) becomes more and more nearly invariant, and thus \(H(s)\) tends to the trivial representation. These representations are the complementary series; no two are equivalent.
The discrete series. The remaining representations have no $K$-invariant vectors. This reflects the fact that they come from sections of nontrivial line-bundles over $\mathbb{H}$ rather than functions on $\mathbb{H}$.

For any integer $n > 0$ we let $H(n)$ be the Hilbert space of holomorphic $n$-forms $f(z)\,dz^n$ on $\mathbb{H}$, with norm

$$\|f\| = \int \rho^{1-n}|f|^2 = \int_{\mathbb{H}} |f(z)|^2 y^{2n-2}|dz|^2.$$ 

Here $\rho = |dz|^2/y^2$ is the hyperbolic area form. For example $H(1)$ is the space of holomorphic 1-forms (whose unitary structure requires no choice of metric).

These representations, together with their antiholomorphic analogues $H(-n)$, form the discrete series. Since they have no $K$-invariant vectors they are isolated from the trivial representation.

The complete picture. The space of all unitary representations of $\text{PSL}_2 \mathbb{R}$ can be organized as a collection $H(s)$ where $s = 1/2 + it$ (the principal series), $0 < s < 1$, $s \neq 1/2$ (the complementary series), or $s = n$, an integer different from zero (the discrete series).

All these representations consist of eigenspaces of the Casimir operator, an invariant differential operator on $C^\infty(G)$, which combines the Laplace operator with $d^2/d\theta^2$ in the $K$-direction. Since Casimir is the same as the Laplacian on $K$-invariant functions, we also denote it by $\Delta$. Now the holomorphic forms that make up the discrete series are sections of line bundles over $\mathbb{H}$; that is, they occur from induced representations of $K$, so they can also be considered as functions on $G$ transforming by a given character $\chi : K \to \mathbb{C}^*$. Then the functions in $H(n)$ satisfy $\Delta f = n(1-n)f$. In other words, elements of the discrete series give negative eigenfunctions of the Casimir operator.

13 Kazhdan’s property T

A topological group $G$ has property T if any unitary representation with almost invariant vectors has invariant vectors.

Example: $\mathbb{Z}$ does not have property T. Acting on $\ell^2(\mathbb{Z})$, $\mathbb{Z}$ has almost invariant vectors but no invariant vectors.

Example: any finite or compact group has property T. If $v$ is almost invariant, then $\int_G gv\,dg$ is invariant.

An infinite discrete group $G$ with property T is nonamenable. Proof.
Otherwise the regular representation of $G$ on $L^2(G)$ has almost invariant vectors.

Any quotient of a group with property $T$ also has property $T$. Thus any quotient is nonamenable or finite.

In particular, the abelianization $G/[G,G]$ is finite.

**Theorem 13.1 (Kazhdan)** Let $G$ be a connected semisimple Lie group with finite center, all of whose factors have $\mathbb{R}$-rank at least 2. Then $G$ has property $T$.

**Corollary 13.2** $\text{SL}_n \mathbb{R}$ has property $T$ for any $n \geq 3$.

**Lemma 13.3** If $\text{SL}_2 \mathbb{R} \ltimes \mathbb{R}^2$ has almost invariant vectors, then $\mathbb{R}^2$ has invariant vectors.

**Proof.** Decompose the representation over $\mathbb{R}^2$ so

$$H = \int_{\mathbb{R}^2} H_t \, d\sigma(t).$$

Then any $f \in H$ determines a measure $\mu_f$ on $\hat{\mathbb{R}}^2$ by

$$\mu_f(E) = \int_E f^2 \, d\sigma = \|\pi_E(f)\|^2.$$

The total mass of this measure is $\|f\|^2$. Also we have

$$|\mu_f(E) - \mu_g(E)| = \left| \int_E (f - g)(f + g) \right| \leq \|f - g\| \cdot \|f + g\|,$$

so $f \mapsto \mu_f$ is continuous in the weak topology.

Now suppose $f_n$ is a sequence of unit vectors, more and more nearly invariant under the action of a compact generating set $K$ for $\text{SL}_2 \mathbb{R}$. Then $\mu_n = \mu_{f_n}$ is nearly invariant under the action of $K$, in the sense that $\mu_n(gE) \approx \mu_n(E)$ for all $g \in K$.

Now if $\mu_n(0) > 0$ then $\mathbb{R}^2$ has an invariant vector, so we are done. Otherwise we can push $\mu_n$ forward by the projection map

$$\mathbb{R}^2 - \{0\} \to \mathbb{RP}^1 \cong S^1,$$

to obtain a sequence of probability measure $\nu_n$ on the circle. But then there is a convergent subsequence, which must be invariant under the action of $\text{SL}_2 \mathbb{R}$ by Möbius transformations. Clearly no such measure exists; hence there is an $\mathbb{R}^2$-invariant vector. 

\[\square\]
The same argument shows a relative form of property T for $\text{SL}_n \mathbb{R} \ltimes \mathbb{R}^n$ for any $n \geq 2$.

Note: for $n = 1$ this false since we obtain the solvable, hence amenable group $\mathbb{R}^* \ltimes \mathbb{R}$.

**Theorem 13.4** $\text{SL}_n \mathbb{R}$ has property T for all $n \geq 3$.

**Proof.** Embed $\text{SL}_{n-1} \mathbb{R} \ltimes \mathbb{R}^{n-1}$ into $\text{SL}_n \mathbb{R}$, so $\mathbb{R}^{n-1} = P$. Then if $\pi$ has almost invariant vectors, we find $\pi$ has a $P$-invariant vector $\xi$. But by the $P$-trick this implies $\xi$ is $\text{SL}_n \mathbb{R}$-invariant. ■

**Induced representations.** Let $X = \Gamma \backslash G$ be a finite volume homogeneous space, and a unitary action $\pi$ of $\Gamma$ on a Hilbert space $H$ is given. Then we can form the flat bundle of Hilbert spaces $H_\pi$ over $X$ and consider the Hilbert space $L^2(X, H_\pi)$ of sections thereof.

Equivalently, we can consider the space of maps $f : G \to H$ such that $f(\gamma^{-1}x) = \pi(\gamma)f(x)$ for all $\gamma \in \Gamma$, normed by integrating over a fundamental domain.

In either case, $G$ acts on the space of sections preserving the norm. Thus we obtain a unitary representation of $G$, the **induced representation** $\text{Ind}_G^G(\pi)$.

**Theorem 13.5** Let $\Gamma \subset G$ be a lattice in a Lie group with property T. Then $\Gamma$ also has property T.

**Proof.** Suppose $\pi$ is a representation of $\Gamma$ on $H$ with almost invariant vectors. We will first show that $\text{Ind}_G^G(\pi)$ has almost invariant vectors.

Choose a reasonable fundamental domain $F \subset G$ for the action of $\Gamma$. Then for $K \subset G$ compact, we can cover almost all of $K \cdot F$ by a finite collection of translates $\langle \gamma_i F; i = 1, \ldots, n \rangle$ of the fundamental domain $F$. (For this it suffices that $K \cdot F$ has finite measure.) Choose an almost invariant vector $v$ for $\Gamma$, such that $\pi(\gamma_i)v \approx v$ for these finitely many $i$. Define $f : G \to H$ by $f(x) = \pi(\gamma)v$ if $x \in \gamma F$ (so $f$ is constant on tiles).

Now compare $f(x)$ and $f(kx)$, $k \in K$, for $x \in F$. Since $kx \in K \cdot F$ is in some tile $\gamma_i F$ for most $x$, we have $f(kx) = \pi(\gamma_i)v \approx v = f(x)$, and thus $f$ is almost invariant.

By property T for $G$, there is a $G$-invariant $f : G \to H$. Then $f$ is constant, and its value $v$ is a $\Gamma$-invariant vector for $\pi$. ■
Corollary 13.6 \( SL_n \mathbb{Z} \) has property \( T \) for all \( n \geq 3 \).

**Proof.** This subgroup is a lattice (by a general result of Borel and Harish-Chandra; see [Rag, Cor 10.5]).

Corollary 13.7 \( SL_2 \mathbb{R} \) does not have property \( T \).

**Proof.** Consider any torsion-free lattice \( \Gamma \) in \( SL_2 \mathbb{R} \). Then \( \mathbb{H}/\Gamma \) is a surface whose first homology \( H^1(\Gamma, \mathbb{Z}) \) is nontrivial; thus \( \Gamma \) maps surjectively to \( \mathbb{Z} \) and so it does not have property \( T \), so neither does \( SL_2 \mathbb{R} \).

Another proof can be given using the fact that the complementary series of unitary representations of \( SL_2 \mathbb{R} \) accumulate at the trivial representation in the Fell topology. (These correspond to eigenfunctions whose least eigenvalue is tending to zero.) A direct sum of such representations has almost invariant vectors but no invariant vector.

Theorem 13.8 Fix a generating set for \( SL_3(\mathbb{Z}) \), and let \( \langle G_p \rangle \) denote the corresponding Cayley graphs for \( SL_3(\mathbb{Z}/p) \). Then \( \langle G_p \rangle \) is a family of expanders.

**Proof.** The action of \( SL_3(\mathbb{Z}) \) on \( L^2_0(G_p) \) has no invariant vectors, so it has no almost invariant vectors, and thus \( \gamma(G_p) > \epsilon > 0 \) where \( \epsilon \) does not depend on \( p \).

Cohomological characterization.

**Theorem 13.9** \( G \) has property \( T \) if and only if every unitary (or orthogonal) affine action of \( G \) on a Hilbert space \( H \) has a fixed-point.

Equivalently, \( H^1(G, H) = 0 \) for every unitary \( G \)-module \( H \).

**Group cohomology.** The group \( H^1(G, H_n) = Z^1(G, H)/B^1(G, H) \) classifies affine actions with a given linear part, modulo conjugation by translations.

An affine action of \( G \) on \( H \) can be written as

\[
g(x) = \pi(g)x + \alpha(g).
\]

112
The linear part $\pi : G \to \text{Aut} H$ is a group homomorphism; the translation part, $\alpha : G \to H$, is a 1-cocycle on $G$. This means $\alpha$ is a crossed-homomorphism:

$$\alpha(gh) = \alpha(g) + \pi(g)\alpha(h).$$

For example, when $\pi$ is the trivial representation, then $\alpha$ is just a homomorphism. The coboundary of $y \in H$ is the cocycle given by

$$\alpha(g) = gy - y;$$

it corresponds to the trivial action conjugated by $x \mapsto x + y$.

The groups $Z^1(G, H) \supset B^1(G, H)$ denote the cocycles and coboundaries respectively. A cocycle $\alpha$ is a coboundary if and only if the affine action of $G$ corresponding to $\alpha$ has a fixed-point $y \in H$.

**Theorem 13.10** $H^1(G, H_\pi) = 0$ implies property $T$ for $\pi$.

**Proof.** Suppose $\pi$ has no fixed vectors. Then the coboundary map $\delta : H \to B^1 = Z^1(G, H)$ is bijective. But $Z^1(G, H)$ is a Banach space. For example, if $G$ is finitely generated, then a cocycle $\alpha$ is determined by its values of generators $g_1, \ldots, g_n$ and we can define

$$\|\alpha\| = \max \|\alpha(g_i)\| = \max \|g_i x - x\|$$

if $\alpha = \delta(x)$. Since $\delta$ is bijective, by the open mapping theorem it is bounded below: that is, there is an $\epsilon > 0$ such that $\|\delta(x)\| > \epsilon\|x\|$. This says exactly that $\pi$ has no almost-invariant vectors. 

**Application of property T to means on the sphere.** Cf. [Lub], [Sar].

**Theorem 13.11 (Margulis, Sullivan, Drinfeld)** Lebesgue measure is the only finitely-additive rotationally invariant measure defined on all Borel subsets of $S^n$, $n \geq 2$.

The result is false for $n = 1$ (Banach). We will sketch the proof of $n \geq 4$ due to Margulis and Sullivan (independently).

The Banach-Tarski paradox states that for $n \geq 2$, $S^n$ can be partitioned into a finite number of (non-measurable) pieces, which can be reassembled by rigid motions to form 2 copies of the sphere.

**Corollary 13.12** There is no finitely-additive rotation invariant probability measure defined on all subsets of $S^n$. 

113
Lemma 13.13 (Cantor-Bernstein) Given injections $\alpha : A \to B$ and $\beta : B \to A$, we can construct a bijection

$$\phi : A \to B$$

by $\phi(x) = \beta^{-1}(x)$ or $\alpha(x)$ according to whether $x \in \beta(C)$ or not. Here

$$C = \bigcup_{0}^{\infty} (\alpha \beta)^n (B - \alpha(A)).$$

Proof. Classify points in $A$ according to their inverse orbits under the combined dynamical system $\alpha \cup \beta : (A \cup B) \to (A \cup B)$. On infinite orbits, $\alpha$ is a bijection. Finite orbits terminate in either $A$ or $B$. If they terminate in $A$, then $\alpha$ gives a bijection; if they terminate in $B$, then $\beta^{-1}$ is a bijection. The set $\beta(C)$ consists of those that terminate in $B$. \qed

Lemma 13.14 The free group $G = \langle a, b \rangle$ can be decomposed into a finite number of sets that can be rearranged by left translation to form $2G$, two copies of $G$.

Proof. Letting $A$ and $A'$ denote the words beginning with $a$ and $a^{-1}$, $B$ and $B'$ similarly for $b$ and $b^{-1}$, and $E$ for the identity element, we have $G = A \cup A' \cup B \cup B' \cup E$, $G = A \cup aA'$, $G = B \cup bB'$, and thus $G$ is congruent to $2G \cup E$. Now applying the Cantor-Bernstein argument to get a bijection. \qed

More explicitly, shifting the set $\{e, a, a^2, \ldots\}$ into itself by multiplication by $a$, we see $G$ is congruent to $G - E$, and thus to $2G$.

Lemma 13.15 For $n \geq 2$, $SO(n+1)$ contains a free group on $2$ generators.

Proof. The form

$$x_0^2 - \sqrt{2} \sum_{1}^{n} x_i^2$$

is Galois equivalent to a definite form, so we get an isomorphism $SO(n, 1, \mathcal{O}) \cong SO(n + 1, \mathcal{O})$, where $\mathcal{O}$ denotes the ring of integers in $\mathbb{Q}(\sqrt{2})$. The group $SO(n, 1, \mathcal{O})$ contains a pair of hyperbolic elements that generate a free group. \qed
Lemma 13.16 For any countable set $E$ in $S^2$, $S^2 - E$ is scissors congruent with $S^2$.

Proof. Find a rotation $R$ such that $R^n(E) \cap E = \emptyset$ if $n \neq 0$. Then by applying $R$, we see $S^2 - \bigcup_0^\infty R^n(E)$ is congruent to $S^2 - \bigcup_1^\infty R^n(E)$. Adding $\bigcup_1^\infty R^n(E)$ to both sides we see $S^2 - E$ is congruent to $S^2$. ■

Proof of the Banach-Tarski paradox. Let $G = \mathbb{Z} * \mathbb{Z}$ act faithfully on $S^2$. Then after deleting a countable set $E$, the action is free. Thus $S^2 - E = G \times F$ for some $F \subset S^2$, where the bijection sends $(g,f)$ to $g \cdot f$. Now the congruence $G = 2G$ gives $(S^2 - E) = 2(S^2 - E)$. On the other hand, $S^2 - E$ is congruent to $S^2$ on both sides, so we are done.

To prove the paradox for $S^n$, $n > 2$, use induction on $n$ and suspension to get a decomposition for $S^{n+1}$ from one for $S^n$. ■

Lemma 13.17 If $\nu$ is a finitely-additive rotation invariant measure on $S^n$, $n \geq 2$, then $\nu$ is absolutely continuous with respect to Lebesgue measure.

Proof. If $D_i$ are disks whose diameters tend to zero, then $\nu(D_i) \to 0$ since many of these disks can be packed in $S^n$. On the other hand, an extension of Banach-Tarski shows that $D_i$ is congruent to $S^n$. So a set of Lebesgue measure zero is congruent to a subset of $D_i$ for any $i$ and thus its $\nu$-measure must also be zero. ■

Lemma 13.18 For $n \geq 4$, $\text{SO}(n+1)$ contains a countable subgroup with property $T$.

Proof. Construct an arithmetic Kazhdan group $\text{SO}(2,n-1,\mathcal{O})$ using the form
$$x_1^2 + x_2^2 - \sqrt{2} \sum_{i=3}^{n+1} x_i^2.$$ Then apply a Galois automorphism to map it into $\text{SO}(n+1,\mathcal{O})$. ■
Note: $SO(p, q)$ has real rank $r = \min(p, q)$ since the form
\[ x_1 y_1 + \ldots + x_r y_r + \sum_{1}^{p+q-2r} z_i^2 \]
has type $(p, q)$. Also $SO(2, p)$ is simple for $p \geq 3$; $SO(2, 2)$, on the other hand, is locally isomorphic to $SO(2, 1) \times SO(2, 1)$ just as $SO(4) \cong SO(3) \times SO(3)$. That is why we need to go up to at least $SO(5)$.

**Theorem 13.19** Lebesgue measure is the only rotation invariant mean on $L^\infty(S^n)$, $n \geq 4$.

**Proof.** Let $G \subset SO(n+1)$ be a finitely-generated group with property $T$. Then there is an $\epsilon > 0$ such that for any $f \in L^2_0(S^n)$, with $\|f\| = 1$, we have $\|f - g_i f\| > \epsilon$ for one of the generators $g_i$ of $G$.

Now suppose there is an invariant mean that is not proportional to Lebesgue measure. The space $L^1(S^n)$ is dense in the space of means, so we can find $F_n \in L^1(S^n)$ converging weakly to the invariant mean. Setting $f_n = \sqrt{F_n}$, we have $\|f_n\| = 1$ and $\|g_i f_n - f_n\| \to 0$ for each generator of $G$. Thus the projection of $f_n$ to the mean-zero functions must tend to zero, which implies that $F_n$ converges to Lebesgue measure on $S^n$.

14 Ergodic theory at infinity of hyperbolic manifolds

Function theory on hyperbolic manifolds.

**Theorem 14.1** $M = \mathbb{H}^n/\Gamma$ admits a non-constant bounded harmonic function $h$ if and only if the action of $\Gamma$ on $S_{\infty}^{n-1}$ is ergodic.

Note: in general there is no invariant measure on $S_{\infty}^{n-1}$!

**Proof.** The boundary values of $h$ give a $\Gamma$-invariant measurable function on $S_{\infty}^{n-1}$, and conversely any bounded $\Gamma$-invariant function on $S_{\infty}^{n-1}$ extends, by visual average, to a harmonic function lifted from $M$. 

116
Theorem 14.2 The geodesic flow on $T_M = \mathbb{H}^n/\Gamma$ is ergodic if and only if the action of $\Gamma$ on $S_{\infty}^{n-1} \times S_{\infty}^{n-1}$ is ergodic.

Example. Let $M \to N$ be a Galois covering of a closed hyperbolic manifold with deck group $\mathbb{Z}^3$. Then $\Gamma_M$ is ergodic on $S_{\infty}^{n-1}$ but not on $S_{\infty}^{n-1} \times S_{\infty}^{n-1}$. This is because random walks in $\mathbb{Z}^3$ are transient (which implies the geodesic flow is not ergodic), but bounded harmonic functions are constant.

For more on function-theory classes of manifolds, see [SN], [LS], [Ly], [MS], [W], [Th, Ch. 9].

Mostow rigidity.

Theorem 14.3 Let $\phi : M \to N$ be a homotopy equivalence between a pair of closed hyperbolic $n$-manifolds, $n \geq 3$. Then $\tilde{\phi} : \mathbb{H}^n \to \mathbb{H}^n$ extends to a quasiconformal map on $S^{n-1}$.

Proof. By compactness, $\tilde{\phi}$ is a coarse $K$–quasi-isometry. Let us first treat the case of dimension $n = 2$. Let $(a, b, c, d)$ be the cyclically ordered vertices of a regular ideal quadrilateral in $S^1_{\infty}$, mapping to $(a', b', c', d')$ under $\tilde{\phi}$. Note that $d([a, b], [c, d])$ is a constant. Since quasigeodesics are a bounded distance from geodesics, we have $d([a', b'], [c', d']) \leq C_K$. Arranging that $a = a' = \infty$, this condition implies that $\tilde{\phi}|_{\mathbb{R}}$ is a quasi-symmetric map, i.e. there is a $k > 1$ depending only on $K$ such that

$$\frac{1}{k} \leq \frac{|b' - c'|}{|c' - d'|} \leq k.$$

The proof in higher dimensions is similar. Arrange that $\tilde{\phi}$ fixes infinite and consider a sphere $S(x, r) \subset \mathbb{H}^{n-1} \subset S_{\infty}^{n-1}$. Let $x' = \tilde{\phi}(x)$, and similarly for $y$ and $z$. Suppose the image of the sphere contains points $y'$, $z'$. Note that $d([\infty, y], [x, z]) = O(1)$. Thus $d([\infty, y'], [x', z']) = O(1)$. This implies that $d(x', y')/d(x', z') = O(1)$. \hfill $\blacksquare$

Theorem 14.4 Let $\phi : M \to N$ be a homotopy equivalence between a pair of closed hyperbolic $n$-manifolds, $n \geq 3$. Then $\phi$ is homotopic to an isometry.

Proof. Lift $\phi$ to an equivariant map $\tilde{\phi} : \mathbb{H}^n \to \mathbb{H}^n$. Then $\tilde{\phi}$ extends to a quasiconformal map $S_{\infty}^{n-1} \to S_{\infty}^{n-1}$. If $\tilde{\phi}$ is not conformal, then by the theory of quasiconformal maps (when $n > 2$) the directions of maximal stretch define a $\Gamma_M$-invariant $k$-plane field on $S_{\infty}^n$. When $n = 3$ the angle
between the lines at the endpoints of a geodesic descends to a function on \( T_1 M \) invariant under the geodesic flow. This angle must then be constant, since the geodesic flow is ergodic, and one is quickly lead to a contradiction. The argument in dimension \( n > 3 \) is similar.

**Diversion: a cocompact Kleinian group.** Perhaps the easiest way to describe a cocompact Kleinian group is to give the Coxeter diagram of a hyperbolic tetrahedron. A particular nice one is the tetrahedron \( T \) given by:

\[
\begin{array}{c}
O \\
| | \\
O \\
| | \\
O
\end{array}
\]

The group \( \Gamma \) generated by reflections in the sides of \( T \) is cocompact; indeed, \( T \) is a fundamental domain for the group action, so we obtain a tiling of \( \mathbb{H}^3 \) by tetrahedra.

All the faces of \( T \) are congruent. One can compute that the interior angles of a face are not all rational multiples of \( \pi \)! Nevertheless, if we take the plane \( P \) through a face of \( T \), it will intersect each tile \( gT, g \in \Gamma \) just in a face, and these intersections will tile \( H \). The tiling that results is shown in Figure 3.

As a hint to resolving this apparent paradox, we note that the angles \( (\alpha, \beta, \gamma) \) of \( F \) can be ordered so that \( \alpha = \pi/4 \) and \( \beta + \gamma = \pi/2 \).

\[\text{Figure 3. A strange tiling of } \mathbb{H}^2 \text{ by congruent triangles.}\]
15 Lattices: Dimension 1

Let $\mathcal{L}_n = \text{SL}_n(\mathbb{R})/\text{SL}_n(\mathbb{Z})$ be the space of oriented unimodular lattices $L \subset \mathbb{R}^n$. Then $X_n = \text{SO}_n(\mathbb{R}) \backslash \mathcal{L}_n$ is the moduli space of oriented flat tori $X = \mathbb{R}^n/L$ of volume one.

The unique torus $S^1 = \mathbb{R}/\mathbb{Z}$ in dimension one still has a rich structure when one examines its automorphisms and endomorphisms.

Automorphisms and endomorphism. For example, any irrational $t \in S^1$ determines an isometry $f : S^1 \to S^1$ by $f(x) = x + t \mod 1$.

1. Every orbit of $f$ is dense (consider it as a subgroup).
2. Thus $f$ is ergodic (use Lebesgue density).
3. In fact $f$ is uniquely ergodic (consider the action on $L^2(S^1) \cong \ell^2(\mathbb{Z})$).
4. The suspension of $f_t$ determines a foliated 2-torus $X_t$. Two such foliated tori are diffeomorphic iff $t_1 \sim t_2$ under the action of $\text{GL}_2(\mathbb{Z})$ by Möbius transformations.
5. Given a homeomorphism $g : S^1 \to S^1$, define its rotation number by $\rho(g) = \lim \tilde{g}^n(x)/n$. If $t = \rho(g)$ is irrational, then $g$ is semiconjugate to rotation by $t$.
6. If $g$ is $C^2$, then this semiconjugacy is actually a conjugacy. If $g$ is analytic with a good rotation number, then the conjugacy is also analytic (Herman). When $g(z)$ is a rational function preserving $S^1$, this implies $g$ has a ‘Herman ring’. (Such rings were considered by Fatou and Julia in the 1920s but their existence had to wait till much later.)

Endomorphisms. Note that $\text{End}(S^1) = \mathbb{Z}$ has the same rank as $S^1$ and corresponds to the unique order in the field $\mathbb{Q}$.

Now consider $f : S^1 \to S^1$ given by $f(x) = dx$, $|d| > 1$; for concreteness we consider the case $d = 2$. This map is very simple; for example, on the level of binary digits, it is just the shift. At the same time it is remarkably rich. Here are some of its properties.

1. Lebesgue measure is ergodic and invariant — hence almost every orbit is uniformly distributed. (Cf. normal numbers).
2. The map is mixing (e.g. it corresponds to $n \mapsto 2n$ on the level of the dual group $\hat{S}^1 = \mathbb{Z}$.)
3. Periodic cycles are dense and correspond to rationals \(p/q\) with \(q\) odd. The other rationals are pre-periodic.

4. Consequently the map is far from being uniquely ergodic.

5. Theorem. Any degree two LEO covering map \(g : S^1 \to S^1\) is topologically conjugate to \(f\).
   
   Proof: the conjugacy is given by \(h = \lim f^{-n} \circ g^n\), which is easily analyzed on the universal cover. The map \(h\) is a monotone semiconjugacy in general, and a homeomorphism when \(g\) is LEO.
   
   Note: if \(g\) is just continuous then it is still semiconjugate to \(f\). If \(g\) is a covering map, the semiconjugacy is monotone. For a non-monotone example, see Figure 4.

6. Example: \(B(z) = e^{i\theta}z(a)/(1-az), \) for \(a \in \Delta\), leaves invariant Lebesgue measure (and is ergodic). It is therefore expanding! and hence LEO. We get lots of ergodic invariant measures on \((S^1,f)\) this way.

7. Periodic cycles also give invariant measures.

8. Theorem. For any \(p/q\) there exists a unique periodic cycle with rotation number \(p/q\).

9. Theorem. For any irrational \(t\), there exists a monotone semiconjugacy from \(f|K_t\) to rotation by \(t\), where \(K\) is a Cantor set. This gives more ergodic invariant measures (these of entropy zero).
   
   The complementary intervals \((I_1, I_2, \ldots)\) of \(K_t\) have lengths \(1/2, 1/4, 1/8, \ldots\) and thus \(K_t\) is covered by \(n\) intervals each of length at most \(2^{-n}\) (those disjoint from \(I_1, \ldots, I_n\)). Thus \(H.\dim(K_t) = 0\). Cf. simple geodesics.

**Entropy.** Let \(f : X \to X\) be a continuous endomorphism of a compact metric space. A set \(E \subset X\) is \((r, n)\)-separated if for any \(x \neq y\) in \(E\), there exists \(0 \leq k \leq n\) such that \(d(f^k(x), f^k(y)) > r\). The entropy of \(f : X \to X\) is given by

\[
h(f) = \lim_{r \to 0} \limsup_{n \to \infty} \frac{\log N(n,r)}{\log n}
\]

where \(N(r,n)\) is the maximum number of points in \((r,n)\)-separated set.

The entropy of an endomorphism of \(S^1\) of degree \(d\) is given by \(\log d\). For local similarities, like \(f(x) = 3x \mod 1\) on the traditional Cantor set, we
Figure 4. A degree 2 map $f : S^1 \to S^1$, and its conjugacy to $x \mapsto 2x$.

have the rule of thumb:

$$\dim(X) = \frac{\log h(X)}{\log |Df|}.$$  

This explain, in particular, why rotation set for $x \mapsto 2x$ have measure zero.

16 Dimension 2

We now turn to the discussion of lattice $L \subset \mathbb{R}^2$; collecting along the way some important ideas that hold in all dimensions:

1. Mahler’s compactness criterion;
2. The well-rounded spine;
3. Action of $A$: the ring $\text{End}_A(L)$;
4. Compact $A$-orbits and ideals;
5. Norm geometry.
6. Forms, discreteness and integrality.

**Theorem 16.1 (Mahler)** The set of $L \in \mathcal{L}_n$ whose shortest vector is of length $\geq r > 0$ is compact. Similarly, the set of $X \in \mathcal{T}_n$ whose shortest closed geodesic is of length $\geq r > 0$ is compact.

**Proof.** Let $v_1, \ldots, v_n \in L$ be chosen so $v_1$ is the shortest nonzero vector in $\mathcal{L}_n$ and $v_{i+1} \in L$ is the shortest vector in $L/(\mathbb{Z}v_1 \oplus \cdots \oplus \mathbb{Z}v_i)$ (with the quotient
norm). Thus $0 < r = |v_1| \leq \cdots \leq |v_n|$. By construction, the vectors $(v_i)$ form a basis for $L$.

We wish to show that there is a constant $C_n(r)$ such that $|v_i| \leq C_n(r)$ for $i = 1, \ldots, n$. For this case, $(v_1, \ldots, v_n)$ range in a compact set, and span a lattice of bounded covolume, which must contain $L$ with bounded index.

The proof will be by induction on $n$. We can take $C_1(r) = r$. For the inductive step, let $L' = L/\mathbb{Z}v_1$ with the quotient norm. We can think of $L'$ as the projection of $L$ to $v_1^\perp \subset \mathbb{R}^n$.

Let $(v'_2, \ldots, v'_n)$ be the basis for $L'$ obtained by the procedure above. Then there is an $\alpha \in [-1/2, 1/2]$ such that $v'_2 + \alpha v_1 \in L$ and hence

$$r \leq |v'_2|^2 + \alpha^2 |v_1|^2 \leq |v'_2|^2 + r^2/4,$$

which gives $|v'_2| \geq (\sqrt{3}/2)r$. Since $\det(\mathbb{Z}v_1) = r$, we have $\det(L') \leq 1/r$. By induction, $|v'_2|, \ldots, |v'_n|$ are bounded above in terms of $r$. Just as for $v'_2$, each $v'_i$ has a lift to $L$ with

$$|v_i|^2 \leq |v'_i|^2 + r^2/4,$$

so we obtain a uniform bound $C_n(r)$ on $|v_i|$ for all $i \leq n$.

The greedy basis. One might attempt to construct a basis for $L$ by choosing $v_1$ to be the shortest vector in $L$, and then choosing $v_{i+1}$ to be the shortest vector outside the real subspace spanned by $(v_1, \ldots, v_i)$. For this algorithm one can again show $|v_i| \leq C'_n(r)$, but the vectors $(v_i)$ need not form a basis for $L$! However they form a basis for a lattice of bounded index in $L$, which is enough to prove Mahler’s compactness theorem.

Well-rounded lattices. Given a lattice $L \subset \mathbb{R}^n$, we let

$$|L| = \inf\{|y| : y \in L, y \neq 0\}.$$ 

A lattice is well-rounded if the set of vectors with $|y| = |L|$ span $\mathbb{R}^n$. This implies $|L| \geq 1$. Thus by Mahler’s criterion, the set of well-rounded lattices $\mathcal{W}_n \subset \mathcal{L}_n$ is compact.

Example: shortest vectors need not span $L$. In high enough dimensions, the vectors $(v_1, \ldots, v_n)$ need not span $L$. For example, let $L \subset \mathbb{R}^n$ be the lattice spanned by $\mathbb{Z}^n$ and $v = (1/2, 1/2, \ldots, 1/2)$. (Thus $[L : \mathbb{Z}^n] = 2$.) For $n > 4$ we have $|v|^2 = n/4 > 1$, which easily implies that $L$ is well-rounded, but its shortest vectors only span $\mathbb{Z}^n$. 

122
Theorem 16.2 The space of all lattices $\mathcal{L}_n$ retracts onto the compact spine $\mathcal{W}_n$.

Sketch of the proof. Expand the subspace of $\mathbb{R}^n$ spanned by the shortest vectors, and contract the orthogonal subspace, until there is a new, linearly independent shortest vector.

Hyperbolic space. In dimension two, we can regard the space of lattices $G/\Gamma$ as the unit tangent bundle $T_1(M_1)$ to the moduli space of elliptic curves. We can regard the space of marked lattices up to rotation, $K \backslash G$, as the hyperbolic plane. The point $\tau \in \mathbb{H}$ corresponds to $L = \mathbb{Z} \oplus \mathbb{Z}\tau$ (implicitly rescaled to have determinant one).

Retraction to the well-rounded spine. We have $|L| = 1$ iff $\tau$ lies on subarc of the circle $|\tau| = 1$ between $60^\circ$ and $120^\circ$. Thus the well-rounded spine is the orbit of this arc under $\text{SL}_2(\mathbb{Z})$; see Figure 5.

The equivariant retraction of $\mathcal{L} - 2$ to $\mathcal{W}_2$ is given in each horoball-like complementary region in $\mathbb{H}$ by flowing along geodesic rays from the center (at infinity) of the horoball to the spine. It is not given by retraction to the closest point, and indeed the closest point need not be unique because the spine is not convex.

Figure 5. The well-rounded spine for $\text{SL}_2(\mathbb{Z})$.

Closed geodesics.

Theorem 16.3 Closed geodesics in $\mathcal{L}_2$ correspond to integral points on the 1-sheeted hyperboloid, or equivalently to rational points in the dual to $\mathbb{H}$.
(Those with square discriminant connect cusps; the others are closed geodesics.)

**Proof.** The endpoints of a geodesic of either type are at worst exchanged under the Galois group, so the geodesic itself is defined over \( \mathbb{Q} \). Conversely, an integral point has a discrete orbit in \( \mathbb{R}^{2,1} \) so it defines a closed subset of \( \mathbb{H} \).

**Quadratic orders.** A *quadratic order* is a commutative ring with identity, \( R \), which is isomorphic to \( \mathbb{Z}^2 \) as an additive group. Any such ring is isomorphic to one of the form

\[
\mathcal{O}_D = \mathbb{Z}[T]/(T^2 + bT + c),
\]

where \( D = b^2 - 4c \). The *discriminant* \( D \) is an invariant of the ring. This ring has nontrivial nilpotent elements iff \( D = 0 \); otherwise, \( K = \mathcal{O}_D \otimes \mathbb{Q} \) is isomorphic to the field \( \mathbb{Q}(\sqrt{D}) \), when \( D \) is not a square, and to \( \mathbb{Q} \oplus \mathbb{Q} \), when \( D \) is a square. The latter case can of course only occur when \( D > 0 \).

For \( D < 0 \), there is a unique embedding of \( \mathcal{O}_D \) into \( \mathbb{R}[K] \cong \mathbb{C} \subset M_2(\mathbb{R}) \), up to the Galois involution.

For \( D > 0 \), there is a unique embedding of \( \mathcal{O}_D \) into \( \mathbb{R}[A] \cong \mathbb{R} \oplus \mathbb{R} \subset M_2(\mathbb{R}) \), up to the Galois involution.

**Norm geometry and Euclidean geometry.** The quadratic forms \( |x|^2 = x_1^2 + x_2^2 \) and \( N(x) = x_1x_2 \) are preserved by the subgroups \( K \) and \( A \) of \( \text{SL}_2(\mathbb{R}) \) respectively, where \( K = \text{SO}(2, \mathbb{R}) \) is compact and \( A \cong \mathbb{R}^* \) is the diagonal subgroup. Thus \( \mathcal{X}_2 = K \backslash G / \Gamma \) classifies the lattices with respect to their Euclidean geometry; while the more exotic space \( A \backslash G / \Gamma \) classifies the lattices with respect to their norm geometry.

**Complex multiplication.** The endomorphisms of a lattice as a group form a ring \( \text{End}(L) \cong M_2(\mathbb{Z}) \). The subring \( \text{End}_K(L) = \text{End}(L) \cap \mathbb{R} \cdot K \cong \mathbb{C} \) is an invariant of the Euclidean geometry of \( L \); it is constant along the orbit \( K \cdot L \). When this ring is bigger than \( \mathbb{Z} \), it satisfies \( \text{End}_K(L) = \mathcal{O}_D \) for a unique \( D < 0 \), and we say \( L \) admits complex multiplication by \( \mathcal{O}_D \).

**Theorem 16.4** The set of lattices \( \mathcal{L}_2[D] \) admitting complex multiplication by \( \mathcal{O}_D \) is a finite union of \( K \)-orbits. The set of orbits corresponds naturally to the group \( \text{Pic} \mathcal{O}_D \) of (proper) ideal classes for \( \mathcal{O}_D \subset \mathbb{C} \), as well as the set of elliptic curves \( E \in \mathcal{M}_1 \) admitting complex multiplication by \( \mathcal{O}_D \).

**Proof.** Any ideal \( I \subset \mathcal{O}_D \subset \mathbb{C} \cong \mathbb{R}^2 \) can be rescaled to become unimodular; conversely, any lattice \( L \) with CM by \( \mathcal{O}_D \) can be rescaled to contain 1; it
then contains $O_D$ with finite index and can hence be regarded as a fractional ideal.

Let us write $O_D = \mathbb{Z} \oplus \mathbb{Z}T$, where $T(z) = \lambda z$. It suffices to prove the Theorem for the set $L_2[T]$ of lattices such that $T(L) \subset L$, which is obviously closed.

We claim $L_2[T]$ is compact. To see this, consider the shortest vector $v$ in a unimodular lattice $L \subset \mathbb{C}$ admitting complex multiplication by $O_D$. Then $L$ contains the sublattice $L' = \mathbb{Z}v \oplus \mathbb{Z}Tv$, and we have

$$1 \leq \text{area}(\mathbb{R}^2/L) \leq \text{area}(\mathbb{R}^2/L') \leq \|T\| \cdot |v|^2.$$ 

Thus $v$ cannot be too short, so by Mahler’s theorem we have compactness.

Now suppose $L_n \to M \in L_2[T]$; then there are $g_n \to \text{id}$ in $\text{SL}_2(\mathbb{R})$ such that $g_n : M \to L_n$. By assumption, $T_n = g_n^{-1}Tg_n \in \text{End}(M)$ for all $n$, and clearly $T_n \to T$. But $\text{End}(M)$ is discrete, so $T_n = T$ for all $n$ sufficiently large. Thus $g_n$ commutes with $T$, which implies $g_n \in K$ and thus $L_n \in K \cdot M$ for all $n \gg 0$.

Coupled with compactness, this shows $L_2[T]$ is a finite union of $K$-orbits.

**Properness.** An ideal $I$ for $O_D \subset \mathbb{Q}(\sqrt{-D})$ may also be an ideal for a larger order $O_E$ in the same field. If not, we say $I$ is a proper ideal, or a proper $O_D$-module. These proper ideal classes correspond to the lattices with $\text{End}_K(L) \cong O_D$ and form a group $\text{Pic} O_D$ (every proper ideal is invertible, in the quadratic case).

**Examples of CM.** Note that any lattice $L \subset \mathbb{Q} \oplus i\mathbb{Q}$ admits complex multiplication by an order in $\mathbb{Q}(i)$; thus the lattices with CM are dense.

For $D < -1$ odd, consider the lattices $L_i = \mathbb{Z} \oplus \mathbb{Z} \tau_i$ where $\tau_1 = \sqrt{-D}$ and $\tau_2 = (1 + \sqrt{-D})/2$. These points both lie in the fundamental domain for $\text{SL}_2(\mathbb{Z})$ acting on $\mathbb{H}$, so they represent different lattices; and they both admit CM by $\mathbb{Z}[\sqrt{-D}] = O_{4D}$. This shows $h > 1$ for $\mathbb{Z}[\sqrt{-3}]$, $\mathbb{Z}[\sqrt{-5}]$, etc.

For the case of $\sqrt{-5}$, the fact $h > 1$ is related to the failure of unique factorization: $2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$ in $\mathbb{Z}[\sqrt{-5}] = O_{-20}$. Note that $O_{-20}$ is the maximal order in $\mathbb{Q}(\sqrt{-5})$; since $-5 \neq 1 \mod 4$, there is no quadratic order $O_{-5}$.

Although factorization is unique in the Gaussian and Eisenstein integers ($\mathbb{Z}[i] \cong O_{-4}$ and $\mathbb{Z}[\omega] \cong O_{-3}$), the class number of the (non-maximal) order $\mathbb{Z}[\sqrt{-3}] \cong O_{-12}$ is two. In general proper suborder have much larger class groups.
**Discriminant of an order.** For a complex quadratic order, the discriminant $D$ can be alternatively described by:

$$D = 4 \text{area}(\mathbb{C}/\mathcal{O}_D)^2.$$  

For example, $\mathcal{O}_{-20}$ has a fundamental domain with sides of length 1 and $\sqrt{5}$.

**Real multiplication.** Now let $\text{End}_A(L) = \text{End}(L) \cap \mathbb{R}[A]$. This ring is an invariant of the norm geometry of $L$. It is constant along the orbit $A \cdot L$.

Either $\text{End}_A(L) = \mathbb{Z}$ or $\text{End}_A(L) = \mathcal{O}_D \subset A$ for a unique $D > 0$. In the latter case we say $L$ admits real multiplication by $\mathcal{O}_D$. (The terminology is not standard, and is borrowed from the theory of Abelian varieties.)

Suppose $L$ lies on a closed geodesic in $\mathcal{L}_2$. Then there is a diagonal matrix $a \in A$ such that $a(L) = L$, so certainly $\text{End}_A(L) = \mathcal{O}_D$ for some $D > 0$.

**Theorem 16.5** The set $\mathcal{L}_2[D]$ of closed geodesics coming from lattices with real multiplication by $\mathcal{O}_D$ is finite, and corresponds naturally to the group of proper ideal classes for $\mathcal{O}_D \subset \mathbb{R}^2$. Thus there are $h(D)$ such geodesics, all of the same length.

**Corollary 16.6** The class number $h(D)$ of $\mathcal{O}_D$ is finite, and $\mathcal{O}_D^* \cong \mathbb{Z} \times (\mathbb{Z}/2)$.

**Regulator.** A fundamental unit $\epsilon \in \mathcal{O}_D^*$ is a unit that generates $\mathcal{O}_D^*/(\pm 1)$. We can choose this unit so $\epsilon > 1$ for a given real embedding $\mathcal{O}_D \hookrightarrow \mathbb{R}$; then the regulator of $\mathcal{O}_D$ is given by

$$R_D = \text{vol}((\mathbb{R}_+\epsilon^T)) = \log \epsilon > 0.$$  

If $\epsilon' > 0$ (equivalently, if $N(\epsilon) = 1$) the fundamental unit is said to be positive. In this case $\epsilon$ generates $A_L$ for any $L \in \mathcal{L}_2[D]$. Since the corresponding matrix $\begin{pmatrix} 0 & 1 \\ \epsilon & 0 \end{pmatrix}$ acts by $\tau \mapsto \epsilon^2 \tau$ on $\mathbb{H}$, we find that the closed geodesic $\gamma_L \subset \mathcal{M}_1$ has length

$$\ell(\gamma_L) = 2 \log \epsilon = 2R_D.$$  

On the other hand, if $N(\epsilon) = -1$, then $A_L$ is generated by $\epsilon^2$, and hence

$$\ell(\gamma_L) = 4R_D.$$  

126
**Discriminant and area.** For a real quadratic order, the discriminant $D$ can be alternatively described by:

$$D = \text{area}(\mathbb{R}^2 / \mathcal{O}_D)^2.$$  

For example, $\mathcal{O}_{20} \subset \mathbb{R}^2$ is spanned by $(1, 1)$ and $(-\sqrt{5}, \sqrt{5})$, and hence  

$$\text{area}(\mathbb{R}^2 / \mathcal{O}_{20}) = \det \begin{pmatrix} 1 & 1 \\ -\sqrt{5} & \sqrt{5} \end{pmatrix} = 2\sqrt{5}.$$

**Lorentz tori.** A marked torus $E = \mathbb{C}/L$ lies on the closed geodesic corresponding to $g \in \text{SL}_2(\mathbb{Z})$ iff the linear action of $g$ on $E$ has orthogonal foliations. These foliations are the zero sets of the form $xy$ which can be regarded as a Lorentz metric on $E$. Thus $E$ admits Lorentzian isometries.

**Examples.** Given $D = 4n$, the ring $\mathcal{O}_D = \mathbb{Z}[\sqrt{n}]$ embeds in $\mathbb{R}^2$ to give the lattice generated by the two orthogonal vectors

$$L = \mathbb{Z}(1, 1) \oplus \mathbb{Z}(\sqrt{n}, -\sqrt{n}).$$

It is challenging to find a generator $g = \begin{pmatrix} \epsilon & 0 \\ 0 & 1/\epsilon \end{pmatrix}$ of $A_L$! Such a number must satisfy  

$$g \cdot (1, 1) = (a + b\sqrt{n}, a - b\sqrt{n}) = (\epsilon, 1/\epsilon),$$

and thus $a^2 - nb^2 = 1$. That is, we want a solution to Pell’s equation for a given $n$, with of course $b \neq 0$. Some minimal solutions are shown in Table 6.

We remark that the largest solution — for $n = 61$ — comes from $\epsilon^6$, where $\epsilon = (39 + 5\sqrt{61})/2$ is a fundamental unit for $\mathcal{O}_{61}$. Although $\epsilon$ is fairly small in height, we must pass to $\epsilon^3$ to get a unit in $\mathbb{Z}[\sqrt{61}]$, and then square this again to get a unit with norm 1.

**Siegel’s theorem.** For any $\epsilon > 0$, there exists a constant $C(\epsilon)$ such that

$$h(D) > C(\epsilon)D^{1/2-\epsilon}$$

for $D < 0$, and

$$h(D)R_D > C(\epsilon)D^{1/2-\epsilon}$$

for $D > 0$. Both cases measure the total volume of the $K$ or $A$ orbits for complex or real multiplication by $\mathcal{O}_D$. Upper bounds of $O(D^{1/2+\epsilon})$ are also known.

**Continued fractions.** Now consider the tiling of $\mathbb{H}$ generated by reflections in the sides of the ideal triangle $T$ with vertices $\{0, 1, \infty\}$. Given $t \in [1, \infty)$
<table>
<thead>
<tr>
<th>(n)</th>
<th>((a, b))</th>
<th>(n)</th>
<th>((a, b))</th>
<th>(n)</th>
<th>((a, b))</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>(3,2)</td>
<td>27</td>
<td>(26,5)</td>
<td>50</td>
<td>(99,14)</td>
</tr>
<tr>
<td>3</td>
<td>(2,1)</td>
<td>28</td>
<td>(127,24)</td>
<td>51</td>
<td>(50,7)</td>
</tr>
<tr>
<td>5</td>
<td>(9,4)</td>
<td>29</td>
<td>(9801,1820)</td>
<td>52</td>
<td>(649,90)</td>
</tr>
<tr>
<td>6</td>
<td>(5,2)</td>
<td>30</td>
<td>(11,2)</td>
<td>53</td>
<td>(66249,9100)</td>
</tr>
<tr>
<td>7</td>
<td>(8,3)</td>
<td>31</td>
<td>(1520,273)</td>
<td>54</td>
<td>(485,66)</td>
</tr>
<tr>
<td>8</td>
<td>(3,1)</td>
<td>32</td>
<td>(17,3)</td>
<td>55</td>
<td>(89,12)</td>
</tr>
<tr>
<td>10</td>
<td>(19,6)</td>
<td>33</td>
<td>(23,4)</td>
<td>56</td>
<td>(15,2)</td>
</tr>
<tr>
<td>11</td>
<td>(10,3)</td>
<td>34</td>
<td>(35,6)</td>
<td>57</td>
<td>(151,20)</td>
</tr>
<tr>
<td>12</td>
<td>(7,2)</td>
<td>35</td>
<td>(6,1)</td>
<td>58</td>
<td>(19603,2574)</td>
</tr>
<tr>
<td>13</td>
<td>(649,180)</td>
<td>37</td>
<td>(73,12)</td>
<td>59</td>
<td>(530,69)</td>
</tr>
<tr>
<td>14</td>
<td>(15,4)</td>
<td>38</td>
<td>(37,6)</td>
<td>60</td>
<td>(31,4)</td>
</tr>
<tr>
<td>15</td>
<td>(4,1)</td>
<td>39</td>
<td>(25,4)</td>
<td>61</td>
<td>(1766319049,226153980)</td>
</tr>
<tr>
<td>17</td>
<td>(33,8)</td>
<td>40</td>
<td>(19,3)</td>
<td>62</td>
<td>(63,8)</td>
</tr>
<tr>
<td>18</td>
<td>(17,4)</td>
<td>41</td>
<td>(2049,320)</td>
<td>63</td>
<td>(8,1)</td>
</tr>
<tr>
<td>19</td>
<td>(170,39)</td>
<td>42</td>
<td>(13,2)</td>
<td>65</td>
<td>(129,16)</td>
</tr>
<tr>
<td>20</td>
<td>(9,2)</td>
<td>43</td>
<td>(3482,531)</td>
<td>66</td>
<td>(65,8)</td>
</tr>
<tr>
<td>21</td>
<td>(55,12)</td>
<td>44</td>
<td>(199,30)</td>
<td>67</td>
<td>(48842,5967)</td>
</tr>
<tr>
<td>22</td>
<td>(197,42)</td>
<td>45</td>
<td>(161,24)</td>
<td>68</td>
<td>(33,4)</td>
</tr>
<tr>
<td>23</td>
<td>(24,5)</td>
<td>46</td>
<td>(24335,3588)</td>
<td>69</td>
<td>(7775,936)</td>
</tr>
<tr>
<td>24</td>
<td>(5,1)</td>
<td>47</td>
<td>(48,7)</td>
<td>70</td>
<td>(251,30)</td>
</tr>
<tr>
<td>26</td>
<td>(51,10)</td>
<td>48</td>
<td>(7,1)</td>
<td>71</td>
<td>(3480,413)</td>
</tr>
</tbody>
</table>

Table 6. Minimal solutions to Pell’s equation \(a^2 - nb^2 = 1\)
irrational, we can consider the geodesic ray $\gamma(i, t)$ joining $i$ to $t$. (Any point $iy$ will do as well as $i$). As this geodesic crosses the tiles, it makes a series of left or right turns, giving rise to a word

$$w = L^{a_1} R^{a_2} L^{a_3} \cdots .$$

We will need to use the map $R(t) = 1/t$. Notice that this map does not send $\mathbb{H}$ to itself; however it does if we compose with complex conjugation. Thus we define $R(z) = 1/\bar{z}$ on $\mathbb{H}$. The maps $R(t)$ and $T(t) = t + 1$ generate an action of $\text{PGL}_2(\mathbb{Z})$ on $\mathbb{H}$, preserving the tiling. Moreover $R(T) = T$, reversing orientation, and $\text{PGL}_2(\mathbb{Z})$ is the full group of isometries preserving the tiling. (In particular, the stabilizer of $T$ itself is isomorphic to $S_3$).

One can think of $\text{PGL}_2(\mathbb{R})$, since it commutes with complex conjugation, as acting naturally on

$$\hat{\mathbb{C}}/(z \sim \bar{z}) = \mathbb{H} = \mathbb{H} \cup \mathbb{R} \cup \{\infty\}.$$ 

**Theorem 16.7** We have $t = a_1 + 1/a_2 + 1/a_3 + \cdots$.

**Proof.** We have $t \in (a_1, a_1 + 1)$ and thus the first integer in its continued fraction expansion agrees with the first exponent in the word $w$. Now let $t' = G(t) = R(T^{-a_1}(t)) \in (1, \infty)$. Letting $p$ denote the point where $\gamma(i, t)$ first crosses $\gamma(a_1, a_1 + 1)$ and begins its first right turn. Then $G$ sends $\gamma(a_1, a_1 + 1)$ to $\gamma(0, \infty)$ and $g(p) = iy$ for some $y$. The map $G$ reverses orientation, and thus it sends $\gamma(p, t) \subset \gamma(i, t)$ to $\gamma(iy, t')$ with associated word

$$w' = L^{a_2} R^{a_3} L^{a_4} \cdots .$$
Thus \( t \rightarrow t' = 1/\{t\} \) acts as the shift on both its geodesic word and its continued fraction expansion, and hence all the exponents agree.

**Corollary 16.8** Two points \( \alpha, \beta \in (0, \infty) \) are in the same orbit of \( \text{PGL}_2(\mathbb{Z}) \) iff the tails of their continued fractions agree.

We say \( t \in \mathbb{R} \) is of **bounded type** if the continued fraction expansion \(|t| = a_1 + 1/a_2 + \cdots \) has bounded \( a_i \)'s. By convention, rational numbers do not have bounded type.

**Theorem 16.9** The geodesic \( \gamma(iy, t) \) is bounded in \( \mathbb{H}/\text{SL}_2(\mathbb{Z}) \) iff \( t \) has bounded type.

**Proof.** Clearly a large value of \( a_1 \) implies a deep excursion into the cusp by the preceding picture. Conversely, a deep excursion into the cusp that ultimately returns must, near its deepest point, make many turns in the same directly (all \( L \) or all \( R \)).

![Figure 8. A neighborhood of the cusp: balls of diameter \( 1/q^2 \) resting on \( p/q \).](image)

**Horoballs and Diophantine approximation.** Noting that the horoball \((y \geq 1)/\mathbb{Z}\) embeds in \( \mathbb{H}/\text{SL}_2(\mathbb{Z}) \), we see the ball of diameter \( 1/q^2 \) resting on \( p/q \) also embeds, which leads to:

**Corollary 16.10** The real number \( t \) has bounded type iff \( t \) is Diophantine of exponent two: that is, there is a \( C > 0 \) such that

\[
|t - \frac{p}{q}| \geq \frac{C}{q^2}
\]

for all \( p/q \in \mathbb{Q} \).
Examples. Given $u, v \in \mathbb{R}$, How can we find a lattice whose $A$-orbit gives the geodesic $\gamma(u, v)$? Answer: just take $L = L(u, v)$ to be the lattice spanned by $v_1 = (1, 1)$ and $v_2 = (u, v)$.

To see this works, notice that as a lattice with basis $v_1$ and $v_2$ degenerates, in the limit the basis elements $v_1$ and $v_2$ become parallel. The number $u \in \hat{\mathbb{R}}$ satisfies $|u| = \lim |v_2|/|v_2|$, i.e. it records the ratio of the limiting positions of $v_1$ and $v_2$ on their 1-dimensional space. (This is clear in the case $v_1 = (1, 0)$ and $v_2 = (x, y)$, $y$ small, corresponding to $\tau = x + iy \in \mathbb{H}$ close to the real point $x$.)

In the case at hand, as we apply the action of $A$, the lattice $L$ is pushed towards either the $x$ or $y$ axis. Thus the limiting points in $\hat{\mathbb{R}}$ for a basis $v_i = (x_i, y_i)$ are $x_2/x_1$ and $y_2/y_1$. In particular $L(u, v)$ degenerates to $u, v \in \hat{\mathbb{R}}$.

Theorem 16.11 The geodesic $\gamma(u, v)$ is periodic iff $u \neq v$ are conjugate real quadratic numbers.

Proof. Periodicity is equivalent to $L(u, v) \otimes \mathbb{Q} = t\mathbb{Q}(\sqrt{D})$ with its usual embedding into $\mathbb{R}^2$, for some real $t$. Since $L(u, v)$ already contains $(1, 1)$, $t$ is rational, and hence $(u, v) = (u, u')$ where the prime denotes Galois conjugation.

Corollary 16.12 The continued fraction expansion of $t$ is pre-periodic iff $t$ is a quadratic irrational.

Proof. If $t$ is a quadratic irrational, then $\gamma(t', t)$ is periodic and hence the bi-infinite word in $R$ and $L$ representing this geodesic is periodic; hence its tail is preperiodic. Conversely, the preperiodic word for $\gamma(i, t)$ can be replaced by a periodic one by changing only one endpoint; thus $\gamma(u, t)$ is periodic for some $u$, and hence $t$ is a quadratic irrational.

Remark: complete separable metric spaces. The space of all irrationals in $(1, \infty)$ is homeomorphic, by the continued fraction expansion, to $\mathbb{N}^\mathbb{N}$. On the other hand, for any separable complete metric space $X$ there is a continuous surjective map $f : \mathbb{N}^\mathbb{N} \to X$. This shows (cf. [Ku, §32, II]):

Theorem 16.13 Every complete separable metric space is a quotient of the irrational numbers.
Intersection with the well-rounded spine. We remark that the tiling of $\mathbb{H}$ by ideal triangles is precisely the dual of the well-rounded spine. Thus the continued fraction expansion of a geodesic also records its intersections with the well-rounded spine, and can be thought of as a record of the shortest vectors along $a_t \cdot L$ as $t$ varies.

Higher Diophantine exponents. Khinchin’s theorem, as presented by Sullivan, says that if the size of $a(q)$ only depends on the size of $q$, then almost every real number satisfies

$$\left| t - \frac{p}{q} \right| \leq \frac{a(q)}{q^2}$$

infinitely often iff $\int a(x) \, dx/x$ diverges. (For example, $a(q) = 1/\log(q)$ works and strengthens the usual result where $a(q) = 1$, which holds for all $x$.)

One direction is obvious: the subset $E_q \subset [0,1]$ where a given $q$ works has measure $|E_q| = O(a(q)/q)$, and if $\sum |E_q| < \infty$ then $\lim \sup E_q$ has measure zero. The other direction uses a version of the Borel-Cantelli lemma (approximate independence of the events $E_q$).

One can also consider the geodesic $\gamma(s)$ to the point $t$, parameterized by hyperbolic arc length. Then $t$ is Diophantine of optimal exponent exactly $\alpha$ if and only if

$$\lim \sup \frac{d(\gamma(0), \gamma(s))}{s} = \frac{\alpha - 2}{\alpha}.$$  

(Almost every $t$ satisfies $d(\gamma(0), \gamma(s)) = O((\log s)^{1+\epsilon})$ and thus the lim sup is zero and $\alpha = 2$.)

For more on this theorem and generalizations to approximating complex numbers by $p/q \in \mathbb{Q}(\sqrt{-D})$ (using the hyperbolic orbifold $\mathbb{H}^3/\text{SL}_2(\mathbb{Z}[\sqrt{-D}])$), see [Sul].

17 Lattices, norms and totally real fields.

Class numbers and units. From Mahler’s Theorem we can deduce some important results in number theory.

Let $K$ be a totally real field of degree $n$, and let $O_K$ denote its ring of integers. Consider $K$ as a subgroup of $\mathbb{R}^n$, using an ordering of its real places. Then $O_K$ becomes a lattice, with

$$\text{vol}(\mathbb{R}^n / O_K)^2 = \text{disc}(O_K).$$

The multiplicative action of $K$ on itself gives a natural embedding $K \hookrightarrow \mathbb{R}[A] \subset M_n(\mathbb{R})$. 

132
We say $M \subset K$ is a full module if $M \cong \mathbb{Z}^n$ (and hence $\mathbb{Q} \cdot M = K$). This implies $M \cap \mathcal{O}_K$ has finite index in both, and thus $M$ also becomes a lattice under the embedding $K \hookrightarrow \mathbb{R}^n$.

The set of $x \in K$ such that $xM \subset M$ form an order $R \subset \mathcal{O}_K$, naturally isomorphic to $\text{End}_A(L)$. Let $\mathcal{L}_n(R)$ denote the set of all lattices with $\text{End}_A(L) = R$.

**Theorem 17.1** The locus $\mathcal{L}_n(R)$ is a finite union of compact $A$-orbits.

**Proof.** Pick an element $\theta \in \mathcal{O}_M$ that generates $K$ over $\mathbb{Q}$. Let $\theta_1, \ldots, \theta_n$ be images of $\theta$ under the $n$ distinct real embeddings of $K$. Let $A \subset \text{SL}_n(\mathbb{R})$ be the diagonal subgroup, and let $T \in A$ be the matrix $\text{diag}(\theta_1, \ldots, \theta_n)$. Let

$$X = \{L \in \mathcal{L}_n : T(L) \subset L\}.$$ 

Clearly $\mathcal{L}_n(R) \subset X$, and $X$ is closed. But it is also compact, since the very short vectors of $L$ must be permuted by $T$ and span a lattice of rank smaller than $n$, contradicting the fact that $T$ is of degree $n$ over $\mathbb{Q}$.

Now suppose $L_n \to L \in X$. Then $L_n = g_n(L)$ where $g_n \to \text{id} \in \text{SL}_n(\mathbb{R})$. Thus $T_n = g_n^{-1}Tg_n \to T \in \text{End}(L)$. But $\text{End}(L)$ is discrete, so $T_n = T$ for all $n \gg 0$. This implies $g_n$ and $T$ commute. Since $T \in A$ has distinct eigenvalues (i.e., it is a regular element), its connected centralizer in $\text{SL}_n(\mathbb{R})$ is the group of diagonal matrices, and thus $g_n \in A$ for all $n \gg 0$. Thus $X/A$ is finite, and hence $\mathcal{L}_n(R)$ is a finite union of compact $A$-orbits.

**Corollary 17.2** The group of units in $R$ has rank $(n - 1)$, and its set of ideal classes is finite.

**Corollary 17.3** Any full module $M$ in $K$ gives rise to a lattice in $\mathbb{R}^n$ whose $A$-orbit is compact.

**Norm and compact $A$-orbits.** We now define the norm $N : \mathbb{R}^n \to \mathbb{R}$ by

$$N(x) = |x_1 \cdot x_2 \cdots x_n|.$$ 

Clearly $N(x)$ is invariant under the action of $A$, and thus

$$N(L) = \inf\{N(x) : x \in L, x \neq 0\}$$

is an invariant of the $A$-orbit of a lattice $L$. 

133
Theorem 17.4 For any \( x \in \mathbb{R}^n \), we have
\[
\sqrt{n}N(x)^{1/n} \leq |x|.
\]
If \( N(y) \neq 0 \), then equality holds for some \( x \in Ay \).

Theorem 17.5 The orbit \( A \cdot L \) is bounded iff \( N(L) > 0 \).

Proof. If \( N(L) = r > 0 \) then \( |aL| > \sqrt{n}r^{1/n} \) for all \( a \in A \) and hence, by Mahler’s criterion, \( A \cdot L \) is bounded. Conversely, \( A \cdot L \) is bounded then there is an \( r \) such that \( |aL| > r \) for all \( a \in A \); then given \( x \neq 0 \) in \( L \) we can find an \( a \) such that
\[
N(x) = N(ax) = |ax|^n/\sqrt{n} \geq r^n/\sqrt{n} > 0,
\]
so \( N(L) > 0 \).

Theorem 17.6 Let \( N \) be the norm form on the ring of integers \( \mathcal{O}_K \) in a totally real number field \( K \). Then \( (N, \mathcal{O}_K) \) is equivalent to \((f, \mathbb{Z}^n)\) where \( f \) is an integral form.

Proof. Let \( \epsilon_i, i = 1, \ldots, n \) be an integral basis for \( \mathcal{O}_K \). Then the coefficients of \( N(\sum a_i \epsilon_i) \) are rational, as well as algebraic integers.

Theorem 17.7 For any unimodular lattice \( L \subset \mathbb{R}^n \), the following conditions are equivalent:

1. \( A \cdot L \) is compact.
2. \( L \) arises from a full module \( M \) in a totally real field \( K/\mathbb{Q} \).
3. We have \( N(L) > 0 \), and \( \{N(y) : y \in L\} \) is a discrete subset of \( \mathbb{R} \).
4. The pair \((L, N)\) is equivalent to \((\mathbb{Z}^n, \alpha f)\) where \( \alpha \in \mathbb{R} \) and \( f \) is an integral form that does not represent zero.

Proof of Theorem 17.7. (1) \( \implies \) (2). Suppose \( A \cdot L \) is compact, and let \( A_L \) denote the stabilizer of \( L \) in \( A \). Then \( L \otimes \mathbb{Q} \) is a module over the commutative algebra \( K = \mathbb{Q}[A_L] \subset M_n(\mathbb{R}) \). The matrices in \( A \) have only real eigenvalues, so \( K \) is a direct sum of \( m \) totally real fields, and therefore the rank of \( \mathcal{O}_K \) is \( n - m \). But the matrix group \( A_L \cong \mathbb{Z}^{n-1} \) embeds in the
unit group $\mathcal{O}_K^*$, so $m = 1$ and $K$ itself is a totally real field. Thus $L \otimes \mathbb{Q}$ is a 1-dimensional vector space over $K$, so the lattice $L$ itself is obtained from a full module $M \subset K$ by the construction above.

The implication $(2) \implies (3)$ is immediate from discreteness of the norm $N^K(x)$ on $M$.

To see $(3) \implies (1)$, observe that the map $g \mapsto N(g^{-1}x) = \phi(x)$ gives a proper embedding of $A \backslash G$ into the space of degree $n$ polynomials on $\mathbb{R}^n$. There is a finite set $E \subset \mathbb{Z}^n$ such that $\phi|E$ determines $\phi$. Consequently, if the values of $N(x)$ on $L = g \cdot \mathbb{Z}^n$ are discrete, then $[g] \cdot \Gamma$ is closed in $A \backslash G$, and therefore $A \cdot L$ is closed in $G/\Gamma$. Since $N(L) > 0$, by Theorem 17.5 the orbit $A \cdot L$ is actually compact.

$(2)$ and $(4)$ are equivalent by the preceding result. □

An formulation purely in terms of forms is:

**Theorem 17.8** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a product of linear forms that does not represent zero. Then the following are equivalent:

1. $f(\mathbb{Z}^n)$ is discrete;
2. $f$ is proportional to an integral form; and
3. $(\mathbb{Z}^n, f)$ is equivalent to the norm form on a proper ideal $I$ for an order in a totally real number field.

**The case $n = 2$: bounded $A$-orbits that are not compact.** Recall $L(u, v) = \mathbb{Z}(1, 1) \oplus \mathbb{Z}(u, v) \subset \mathbb{R}^2$. Theorem 16.9 implies:

**Theorem 17.9** We have $N(L(u, v)) > 0$ iff $u$ and $v$ are numbers of bounded type. In particular, there are plenty of lattice in $\mathbb{R}^2$ with $N(L) > 0$ that do not come from number fields/closed geodesics.

We will shortly examine Margulis’s conjecture which implies to the contrary:

**Conjecture 17.10** A lattice $L \subset \mathbb{R}^n$, $n \geq 3$, has $N(L) > 0$ iff $L$ comes from an order in a number field. Equivalently, the set of $L$ with $N(L) > 0$ is a countable union of compact $A$-orbits.

**Remark.** In $\mathbb{R}^2$ the region $N(x) < 1$ has infinite area, since $\int_0^\infty dx/x = \infty$. For the same reason, the region $N(x) < \epsilon$ is a neighborhood of the coordinate planes of infinite volume. The conjecture says that in dimensions 3 or more, it is very hard to construct a lattice that avoids this region. The only possible construction uses arithmetic, i.e. the integrality of the norm on algebraic integers in a totally real field.
18 Dimension 3

The well-rounded spine. Although $\mathcal{X}_3$ has dimension 5, its spine $\mathcal{W}_3$ is only 3-dimensional.

A well-rounded lattice can be rescaled so its Gram matrix $m_{ij} = v_i \cdot v_j$ satisfies $m_{ii} = 2$. Then by consider the vectors $v_i \pm v_j$ we see $|m_{ij}| \leq 1$ for the three off-diagonal elements. Thus we describe $\mathcal{W}_3$ by the coordinates

$$m(x, y, z) = \begin{pmatrix} 2 & x & y \\ x & 2 & z \\ z & y & 2 \end{pmatrix}.$$ 

Since the off-diagonal elements are bounded by one, $\mathcal{W}_3$ is a subset of a cube. (Actually it is a quotient of this set by permutations of coordinates and changing the signs of any pair of coordinates.)

The lattice $m(1, 1, 1)$ gives the densest sphere packing $L \subset \mathbb{R}^3$, the face-centered cubic (fcc) packing with Voronoi cell the rhombic dodecahedron. Thus there are twelve vectors with $|x| = |L|$.

We can think of $L$ as a laminated lattice: start with the hexagonal lattice, and center the second hexagonal over the deep holes of the first. (Alternatively, one can start with a square packing, then again add the next layer in the deep holes. In these coordinates, the lattice is the subgroup of index two in $\mathbb{Z}^3$ where the coordinates have even sum.)

As we roll the second lattice between two adjacent deep holes, the Gram matrix is given by $m(1, 1, t)$, with $t$ going from 1 to 0. Thus $m(1, 1, 0)$ also represents the fcc lattice. The corresponding lattice $L_t$ has determinant $d(t) = 2(1 + t)(2 - t)$, which achieves its minimum value 4 along $[0, 1]$ at the endpoints.

For $t \in [-1, 0)$ the matrix $m(1, 1, t)$ no longer corresponds to a well-rounded lattice. Indeed, the vector $-v_1 + v_2 + v_3$ has squared length $2(1 + 1 + 1 - x - y + z) = 2(3 - 2 - t) < 2$ in this range. Thus this segment must be removed from the edge of the cube; the twelve edges forming its orbit under the symmetry group must similarly be removed.

As shown by Soulé, in these coordinates $\mathcal{W}_3$ is the convex hull of the remaining 12 half-edges. That is $\mathcal{W}_3$ is obtained from the cube by trimming off four of its eight corners [So].

Dimension two. Similar considerations show we can identify the well-
rounded spine for $\mathcal{L}_2$ with the set of matrices

$$m(x) = \begin{pmatrix} 2 & x \\ x & 2 \end{pmatrix}$$

where $x \in [-1, 1]$, modulo $x \equiv -x$.

**Margulis and Littlewood.** We can now relate the following two open problems.

**Conjecture 18.1 (Littlewood)** For any $\alpha, \beta \in \mathbb{R}$ we have

$$\inf_{n>0} n \cdot \|n\alpha\| \cdot \|n\beta\| = 0.$$

**Conjecture 18.2 (Margulis)** Every bounded $A$-orbit in $\mathcal{L}_n$, $n \geq 3$ is closed (and hence comes from a totally real field).

We will show that Margulis’s conjecture for $n = 3$ implies Littlewood’s conjecture. Cf. [Mg, §2].

**Oppenheim conjecture.** To give some more context for Margulis’s conjecture, we note that if we replace $A$ by $H = \text{SO}(2, 1, \mathbb{R})$ the corresponding result is known to be true. That is, Ratner’s theorem implies:

**Theorem 18.3** Every $H$-orbit on $\mathcal{L}_3 = \text{SL}_3(\mathbb{R})/\text{SL}_3(\mathbb{Z})$ is either closed and finite volume, or dense.

This result was first proved by Dani and Margulis. It implies:

**Corollary 18.4 (Oppenheim conjecture)** Let $B(x)$ be an indefinite quadratic form on $\mathbb{R}^n$, $n \geq 3$. Then either $B$ is proportional to an integral form, or $B(\mathbb{Z}^n) \subset \mathbb{R}$ is dense.
As shown by Raghunathan, by taking a generic 3-dimensional rational slice of $\mathbb{R}^n$, one can reduce to the case $n = 3$.

**Background.** Let $B(x, y, z)$ be an integral quadratic form of signature $(2, 1)$ that does not represent zero. For example:

$$B(x, y, z) = x^2 + 2y^2 - 5z^2$$

has this property. (If $B$ represents zero, then it does so on a vector where $(x, y, z)$ are relatively prime, hence not all divisible by 5. But then neither $x$ nor $y$ is divisible by 5, and thus $-2 = 3 = (x/y)^2 \mod 5$, a contradiction (the squares mod 5 are 0, 1, 4).

**Theorem 18.5** The group $SO(B, \mathbb{R}) \subset SL_3(\mathbb{R})$ meets $SL_3(\mathbb{Z})$ in a cocompact lattice.

**Proof.** Since $B \in SO(2, 1, \mathbb{R}) \setminus SL_3(\mathbb{R})$ has integral coefficients, its orbit under $SL_3(\mathbb{Z})$ is discrete. Consequently the orbit of $Z^3 \in L_3$ under $SO(B, \mathbb{R})$ is closed. But it is also compact, because any $v \neq 0$ in a lattice in $SO(B, \mathbb{R}) \cdot Z^3$ satisfies $B(v) \in \mathbb{Z} \setminus \{0\}$ and hence $|v|$ is bounded below. Thus the stabilizer of $Z^3$, $SO(B, \mathbb{Z})$, must be cocompact in $SO(B, \mathbb{R})$.

**Theorem 18.6** If $B$ is a real form of signature $(2, 1)$ and $SO(B, \mathbb{Z})$ is a lattice in $SO(B, \mathbb{R})$, then $B$ is proportional to an integral form.

**Proof.** Let $V \subset Sym(\mathbb{R}^3)$ be the subspace of quadratic forms invariant $F$ under $SO(B, \mathbb{Z})$. Since the matrices in $SO(B, \mathbb{Z})$ have integral entries, the linear equations $F(gx) = F(x)$ defining $V$ are rational, i.e. $V$ is defined over $\mathbb{Q}$.

We claim that $V$ is one-dimensional; in other words, that $SO(B, \mathbb{Z})$ determines $B$ up to scale. To see this, one can note that the lattice $SO(B, \mathbb{Z})$ gives a Kleinian group acting on $\mathbb{RP}^2$, whose limit set is the conic defined by $B = 0$. Alternatively, the Borel density theorem states that a lattice in a semisimple Lie group with no compact factor is Zariski dense; thus any group $SO(F, \mathbb{R})$ containing $SO(B, \mathbb{Z})$ must equal $SO(B, \mathbb{R})$, and hence $F \in \mathbb{R} \cdot B$.

Since $V$ is one dimensional and rational, some real multiple of $B \in V$ must have integral coefficients.
Proof of the Oppenheim conjecture. There is no closed Lie subgroup between \( H = \text{SO}(2,1,\mathbb{R}) \) and \( G = \text{SL}_3(\mathbb{R}) \), and \( H^0 \) is generated by unipotents, so Ratner’s theorem implies the \( H \)-orbit Theorem.

For the Corollary, observe that \( H \backslash G \) can be identified with the space of indefinite forms \( B \) on \( \mathbb{R}^3 \) of signature \( (2,1) \) and determinant one. Thus if \( B \cdot \text{SL}_3(\mathbb{Z}) \) is dense then \( B(\mathbb{Z}^3) \) is also dense.

On the other hand, if we have a finite volume \( H \)-orbit then the stabilizer of \( B \) is a lattice \( \text{SO}(B,\mathbb{Z}) \) in \( \text{SO}(B,\mathbb{R}) \) and hence \( B \) is proportional to an integral form.

For more details, see e.g. [Rn], [BM, Ch. VI].

Compact \( A \)-orbits: isolation. To study the case where \( H = \text{SO}(2,1,\mathbb{R}) \) is replaced by the abelian \( A \), we begin with an ‘isolation result’ of Cassels and Swinnerton-Dyer [CaS]. This result shows, for example, that you cannot construct a bounded \( A \) orbit that spirals or oscillates between one or more compact \( A \) orbits.

**Theorem 18.7** Let \( T \subset \mathcal{L}_3 \) be a compact, \( A \)-invariant torus, and suppose \( X = A \cdot L_0 \) meets \( T \). Then either \( X = T \), or \( N(L_0) \) is dense in \( \mathbb{R}_+ \).

**Proof.** Suppose \( X \neq T \). We will show \( N(L_0) \) is dense. It suffices, by semicontinuity, to show \( N(L) \) is dense for some \( L \in X \).

Let \( V \subset G = \text{SL}_3(\mathbb{R}) \) denote the closure of the set of \( g \) such that \( X \cap gT \neq \emptyset \). Since \( T \) is compact, \( V \) is closed. For \( a,b \in A \) we have

\[
X \cap agbT = a(X \cap gT),
\]

so \( V \) is invariant under the action of \( A \times A \).

Now suppose \( X \neq T \). Then \( V \) contains elements of the form

\[
v = (1 + \delta)(I + \epsilon v_{ij})
\]

where \( \epsilon \) and \( \delta \) are arbitrarily small, \( |v_{ij}| \leq 1 \), \( v_{ij} \) vanishes on the diagonal, and some off-diagonal element, say \( v_{12} \), is equal to one.

Conjugating by \( a = \text{diag}(a_1,a_2,a_3) \) sends \( v_{ij} \) to \( v_{ij}a_i/a_j \). In particular, conjugating by \( \text{diag}(a,1/a,1) \) multiplies \( v_{ij} \) by the matrix:

\[
\begin{pmatrix}
1 & a^2 & a \\
 a^{-2} & 1 & a^{-1} \\
 a^{-1} & a & 1
\end{pmatrix}.
\]
This conjugation makes \( v_{12} \) much larger than the remaining elements of \( v_{ij} \).
It can also be used, at the same time (by appropriate choice of \( a \)), to adjust \( \epsilon v_{12} \) so it is close to any prescribed value.
Since \( V \) is closed, this implies \( V \) contains the 1-parameter subgroup
\[
U = \left\{ u_t = \begin{pmatrix} 1 & t & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} : t \in \mathbb{R} \right\}.
\]

Note that \( U \) is normalized by \( A \), i.e. \( aUa^{-1} = U \) for all \( a \in A \).

Now let \( I \subset \mathcal{O} \subset K \) an ideal for an order in a totally real cubic field,
such that \( T \) contains a lattice \( L_1 \) proportional to the standard embedding \( I \subset \mathbb{R}^3 \). By assumption \( u_t a(L_1) \in X \) for some \( t \neq 0 \). But \( U \) is normalized by \( A \), and thus \( u_t(L_1) \in X \) for some \( t \neq 0 \).

Thus to complete the proof, we need only show that \( N(u_tL_1) \) is dense; or equivalently, that \( N(u_tI) \) is dense. To see this, just note that
\[
N(u_tx) = |x_1 + tx_2| x_2 x_3 = |x_1 x_2 x_3(1 + tx_2/x_1)|.
\]
Pick \( x \neq 0 \) in \( I \), and let \( s \in \mathcal{O}^* \) be a positive unit. Then \( sx \in I \) as well, and we have:
\[
N(u_t sx) = |x_1 x_2 x_3(1 + t(s'/s)(x_2/x_1))|.
\]

Now \( \mathcal{O}^*_+ \cong \mathbb{Z}^2 \) maps injectively into \( \mathbb{R}_+ \) via \( s \mapsto s'/s \), so its image is dense.
Thus \( N(u_t x) \) contains a half line. By varying the choice of \( x \) we then see \( N(u_t I) \) is dense in \( \mathbb{R} \).

By semicontinuity of the norm, it follows that the values of \( N \) on \( L_0 \) are also dense in \( \mathbb{R}_+ \).

\[\text{Theorem 18.8} \quad \text{Margulis’s conjecture implies Littlewood’s conjecture.}\]

\[\text{Proof.} \quad \text{Suppose } (\alpha, \beta) \in \mathbb{R}^2 \text{ is a counterexample to Littlewood’s conjecture. Consider the unimodular lattice } L_0 \subset \mathbb{R}^3 \text{ generated by} \]
\[
\{(e_1, e_2, e_3)\} = \{(1, 0, 0), (0, 1, 0), (\alpha, \beta, 1)\},
\]
and let \( M_0 = \mathbb{Z} e_1 \subset \mathbb{Z} e_2 \subset L_0 \). We then have, for any \( (a, b, n) \in \mathbb{Z}^3 \),
\[
N(-ae_1, -be_2, ne_3) = |n\alpha - a| \cdot |n\beta - b| \cdot n \geq n \cdot \|n\alpha\| \cdot \|n\beta\|.
\]
Thus the norm is bounded away from zero on \( L_0 - M_0 \), and vanishes exactly on \( M_0 \); in particular, \( N(L_0) \) is not dense.
Now we get rid of the lattice $M_0$. Let $a_t = \text{diag}(t, t, t^{-2})$, and let $L_t = a_t(L_0) \supset M_t = a_t(M_0)$. Then as $t \to \infty$, the null vectors $M_t = a_t(M) \subset L_t$ are pushed off to infinity. Thus the shortest vector in $L_t$ must lie in $L_t - M_t$; but there the norm is bounded below, so the shortest vector also has length bounded below.

By Mahler's compactness criterion, there is a subsequence such that $L_t \to L_\infty$. The limit has no nontrivial null-vectors, and thus $N(L_\infty)$ is bounded below. By Margulis’s conjecture, $L_\infty \in T$ for some compact torus orbit. But then $X = A \cdot L_0$ contains $T$. Since $N(L_0)$ is not dense, $X = T$ and thus $L_0 \subset T$. This contradicts the fact that $L_0$ has null vectors. 

**Form version.** The preceding result and conjecture can be recast in terms of cubic forms as follows. Let us say $f$ is a cubic norm form if it is a multiple of the integral norm form on an ideal in a totally real cubic field.

**Theorem 18.9** Suppose $f = f_1 f_2 f_3 : \mathbb{R}^3 \to \mathbb{R}$ is a product of linear and $f(\mathbb{Z}^3)$ is not dense in $\mathbb{R}$. Then either $f$ is a cubic norm form, or $\text{SL}_3(\mathbb{Z}) \cdot f$ contains no cubic norm forms.

In fact: if $f$ is an integral form that does not represent zero, and $f_n \to f$ are distinct from $f$, then $f_n(\mathbb{Z}^3)$ becomes denser and denser in $\mathbb{R}$. See [CaS].

**Conjecture 18.10** If $f$ does not represent zero then either $f(\mathbb{Z}^3)$ is dense or $f$ is a cubic norm form.

Now suppose the conjecture above holds. If $\alpha$ and $\beta$ are a counterexample to Littlewood’s conjecture, then the form

$$f(x, y, z) = x(x\alpha - y)(x\beta - z)$$

is bounded away from zero; and it only represents zero on the plane $x = 0$. Thus means we can take $g_n \in \text{SL}_n(\mathbb{Z})$ such that $g_n \cdot f \to h$ where $h$ does not represent zero. But then $h$ is a cubic norm form, contrary to the preceding theorem.

**The case $n = 2$.** The first part of the Cassels and Swinnerton-Dyer argument, that establishes the existence of a unipotent subgroup in $V$, works just as well in any dimension. (This is exactly the same as finding a single unipotent $u$ such that $L \in X$ and $u(L) \in X$, since then $AuA \subset V$ and $AuA$ contains a 1-parameter unipotent subgroup through $u$.)

In dimension two it shows that if $X \subset L_2$ is a compact $A$-invariant set, then either $X$ consists of a finite union of closed geodesics or $X$ contains
two geodesics with the same endpoint (since this is what it means for leaves to be related by the unipotent horocycle flow.)

When \( X \) comes from a simple lamination, these asymptotic leaves are clearly visible: they come from the boundary of a complementary region such as an ideal triangle.

**Gaps of slopes.** The reason one does not have an isolation result in dimension two is that the ratios \( x'/x \) as \( x \) ranges in the units of a real quadratic field form a discrete set.

In dimension three, one finds the following. Suppose \( X = A \cdot L \subset \mathcal{L}_3 \) is compact, but \( A \cdot L \) is not. Then the set of slopes:

\[
S = \{x_1/x_2 : x \in L\}
\]

is nowhere dense in \( \mathbb{R} \). Otherwise (18.1) would give that \( N(L) = 0 \).

**Forms, discreteness and integrality.** A general setting that includes both the actions of \( A \) and \( \text{SO}(p, q, \mathbb{R}) \) is the following.

Let \( G = \text{SL}_n(\mathbb{R}) \), let \( \Gamma = \text{SL}_n(\mathbb{Z}) \). Let \( f : \mathbb{R}^n \to \mathbb{R} \) be a homogeneous form of degree \( d \), and let \( H \subset G \) be the largest connected subgroup of \( G \) leaving \( f \) invariant. Geometrically \( f \) determines a subvariety \( V(f) \subset \mathbb{P}^{n-1} \), and \( H \) determines a continuous group of symmetries of \( V(f) \).

Examples: \( n = 2 \) and \( f(x, y) = x^2 + y^2 \) we have \( H = K \); for \( f(x, y) = xy \) we have \( H = A \); and for \( f(x, y) = x \) we have \( H = N \). For general \( n \) and \( f(x) = x_1 \cdots x_n \), we have \( H = A \); for

\[
f(x) = x_1^2 + \cdots + x_p^2 - x_{p+1}^2 - \cdots x_{p+q}^2,
\]

\( p + q = n \), we have \( H = \text{SO}(p, q, \mathbb{R}) \). For a more exotic example, let \( n = m^2 \), so \( \mathbb{R}^n = M_n(\mathbb{R}^m) \), and let \( f(X) = \det(X) \). Then \( H = \text{SL}_n(\mathbb{R}) \times \text{SL}_n(\mathbb{R}) \) acting by \( X \mapsto h_1 X h_2^{-1} \).

A theorem of Jordan asserts that if a form \( f \) of degree \( d \geq 3 \) defines a smooth projective variety, then its stabilizer in \( \text{GL}_n(\mathbb{C}) \) is finite. This means the irreducible factors of ‘exotic examples’ (forms which are not products of linear or quadratic forms) must have vanishing discriminant. See [Bor, §I.6.9]. For the classification of ‘exotic examples’, see [KS].

The space \( F \) can be thought of as the space of homogeneous forms on \( \mathbb{R}^n \) of the same type as \( f \): for example, quadratic forms of signature \( (p, q) \), or products of \( n \) linear forms.

We can study the action of \( H \) on \( G/\Gamma \) dynamically as we have for the geodesic and horocycle flows above. But we can also equivalently study the action of \( \Gamma \) on the space \( F = H \backslash G \). It is easy to see, for example:
Theorem 18.11 The action of \( H \) on \( G/\Gamma \) is ergodic iff the action of \( \Gamma \) on \( F = H \backslash G \) is ergodic. The orbit \( Hx \subset G/\Gamma \) is closed iff the orbit \( x\Gamma \subset H \backslash G \) is closed.

Let us now assume that:

1. \( H \) is a connected reductive group without compact factors; and
2. Any homogeneous form of the same degree as \( f \) that is stabilized by \( H \) is proportional to \( f \).

This holds in all the examples considered above.

Note that \( \Gamma \) preserves the discrete subset \( F(\mathbb{Z}) \) of forms with integral coefficients. We say \( f \) represents zero if there is an \( x \neq 0 \) in \( \mathbb{Z}^n \) such that \( f(x) = 0 \).

Theorem 18.12 (Integral Forms) Suppose \( f \) does not represent zero. Then the following are equivalent.

1. The set \( f(\mathbb{Z}^n) \subset \mathbb{R} \) is discrete.
2. The orbit \( H \cdot \mathbb{Z}^n \subset G/\Gamma \) is compact.
3. A nonzero multiple of \( f \) lies in \( F(\mathbb{Z}) \), and hence takes only integral values on \( \mathbb{Z}^n \).

Proof. (1) implies (2): \( X = H \cdot \mathbb{Z}^n \) is closed, and \( |f(v)| > \epsilon > 0 \) for all nonzero \( v \in L \in X \). By continuity of \( f \), this implies \( |v| > \epsilon' > 0 \) and hence \( X \) is compact by Mahler’s criterion.

(2) implies (3): By assumption, \( H(\mathbb{Z}) = H \cap \Gamma \) is a lattice in \( H \). Invariance of \( f \) under each integral matrix \( g \in H(\mathbb{Z}) \) imposes a rational linear condition on the coefficients of \( f \). These conditions uniquely determine the stabilizer \( H \) of \( f \), by the Borel density theorem (\( H(\mathbb{Z}) \) is Zariski dense in \( H \)). Finally \( H \) determines \( f \) up to a constant multiple by assumption.

(3) implies (1): immediate.

19 Dimension 4, 5, 6

Suppose \( K/\mathbb{Q} \) is a number field with \( r_1 \) real places and \( r_2 \) complex places. Then, a general version of the arguments presented before shows the rank of the unit group, \( \mathcal{O}_K^* \), is \( r_1 + r_2 - 1 \).
Suppose \( u \in K \) is a unit of degree \( d = r_1 + 2r_2 = \deg(K/\mathbb{Q}) > 1 \), and (under some embedding) we also have \( u \in S^1 \). Then \( u \) must be Galois conjugate to \( \overline{u} = 1/u \neq u \), and thus its minimal polynomial is reciprocal and of even degree.

In dimension 2, such a \( u \) must be a root of unity, but in dimension 4 or more it can be a Salem number: its polynomial has two roots outside the circle and two on the circle. By irreducibility those on the circle have infinite multiplicative order.

By considering \( L = \mathcal{O}_K \subset \mathbb{R}^4 \), or an ideal, we obtain a Euclidean lattice with a self-adjoint automorphism \( T : L \to L \) that acts hyperbolically on one \( \mathbb{R}^2 \) and by an irrational rotation on another \( \mathbb{R}^2 \). This action is not structurally stable!

Going to dimension 6, one can obtain a \( \mathbb{Z}^2 \) action on a torus with the same kind of partial hyperbolicity. Damjanović has shown such a \( \mathbb{Z}^2 \) action is structurally stable (e.g. in \( \text{Diff}^\infty(\mathbb{R}^6/\mathbb{Z}^6) \)); see e.g. [DK].

To obtain such an action, following Damjanović, let \( K \) be a totally real field of degree \( d \), and let \( L/K \) be a degree two extension with one complex place and 2\( d - 2 \) real places. We can use this place to regard \( L \) as a subfield of \( \mathbb{C} \). The unit group of a field with \( r_1 \) real places and \( r_2 \) complex places has rank \( r_2 + r_1 - 1 \). Thus \( K^* \) has rank \( d - 1 \) and \( L^* \) has rank \( 2d - 2 \). Consequently the map \( L^* \to K^* \) given by \( u \mapsto u\overline{u} \) has a kernel of rank (at least) \( d - 1 \). Since the rank of \( \mathcal{O}_L \) is 2\( d \), this gives a partially hyperbolic action of \( \mathbb{Z}^{d-1} \) on \((S^1)^{2d}\).

## 20 Higher rank dynamics on the circle

In 1967 Furstenberg showed [Fur1]:

**Theorem 20.1** Let \( X \subset S^1 = \mathbb{R}/\mathbb{Z} \) be a closed set invariant under \( x \mapsto 2x \) and \( x \mapsto 3x \). Then either \( X \) is finite, or \( X = S^1 \).

**Lemma 20.2** Let \( S = \{0 = s_0 < s_1 < s_2 < s_3 \ldots \} \subset \mathbb{R} \) be an infinite discrete set satisfying \( S + S \subset S \). Then either \( S \subset a\mathbb{Z} \) for some \( a > 0 \), or \( s_{n+1} - s_n \to 0 \).

**Proof.** After scaling we can assume \( s_1 = 1 \) and thus \( \mathbb{N} \subset S \). Assume \( S \) is not contained in \((1/q)\mathbb{Z}\) for any \( q \); then the projection then \( S \to S^1 = \mathbb{R}/\mathbb{Z} \) is evidently a dense semigroup \( G \). Now if \( [x] \in G \) then \( x + n \in S \) for some integer \( n \geq 0 \), and thus all but finitely many elements of \( x + \mathbb{N} \) belong to \( S \). The same is true with \( x \) replaced by any finite subset \( X \subset G \).
Given $\epsilon > 0$, choose a finite set $[X] \subset G \subset S^1$ so the complementary gaps have length less than $\epsilon$. Then $X + \mathbb{N}$ also has gaps of length at most $\epsilon$. Since $X + \mathbb{N}$ is eventually contained in $S$, we have $\limsup s_{n+1} - s_n < \epsilon$ and hence the gaps in $S$ tend to zero.

Now let $A \subset \mathbb{N}$ be the multiplicative semigroup generated by 2 and 3. We can write $A = \{t_1 < t_2 < \cdots \}$. The preceding Lemma implies $t_{n+1}/t_n \to 1$ and thus:

**Corollary 20.3** As $x \to 0$, the rescaled sets $xA \cap [0,1]$ converge to $[0,1]$ in the Hausdorff topology.

**Isolation.** We now prove an easy analogue of the ‘isolation result’ of Cassels and Swinnerton-Dyer. Note that the compact $A$-orbits on $S^1$ are just certain finite sets of rational numbers. We will see that any more exotic $A$-invariant sets must stay away from the rationals.

**Lemma 20.4** Let $X \subset S^1$ be a closed set such that $AX = X$, and suppose 0 is not an isolated point of $X$. Then $X = S^1$.

**Proof.** Take $x_n \in X$ tending to zero, and observe that $X$ contains the projection of $x_n A \cap [0,1]$ under $\mathbb{R} \to S^1 = \mathbb{R}/\mathbb{Z}$. By the preceding Corollary, these sets become denser and denser as $n \to \infty$, and thus $X = S^1$.

**Corollary 20.5** If $X$ contains a non-isolated rational point $p/q$ then $X = S^1$.

**Proof.** Then $qX$ accumulates at 0, so $qX = S^1$ which implies $X = S^1$.

**Minimal sets.** Let $X \subset S^1$ be a nonempty closed set that is (forward) invariant under the endomorphisms given by a semigroup $A \subset \mathbb{Z}$ (such as $A = \langle 2,3 \rangle$.) Then $X$ contains, by the Axiom of Choice, a minimal such set $F$.

Note that for any $a \in A$, we have $A(aF) = a(AF) \subset A(F)$. Thus a minimal set satisfies $aF = F$ for all $a \in A$ (else $aF$ would be a smaller invariant set).

We will begin by proving Furstenberg’s theorem for $A$-minimal sets, $A = \langle 2,3 \rangle$. 

145
Lemma 20.6 Let $F \subset S^1$ be a minimal, $A$-invariant set and suppose $X = F - F = S^1$. Then $F = S^1$.

Proof. This is the trickiest step in the proof. Let

$$\pi_n : A \to (\mathbb{Z}/5^n)^*$$

be reduction of the integers in $A$ modulo $5^n$. (Since 2 and 3 are relatively prime to 5, the image lies in the multiplicative group modulo $5^n$.) Let $A_n = \text{Ker} \pi_n \subset A$.

Now let $F_1 \supset F_2 \supset \cdots$ be a sequence of nonempty closed subset of $F$ such that $F_n$ is minimal for the action of $A_n$.

The key point is that if $x_1, x_2 \in A$ and $\pi_n(x_1) = \pi_n(x_2)$ then $x_1 F_n = x_2 F_n$. To see this, choose $z \in A$ (say a power of $x_1$) such that $zx_1 \in A_n$. Then $zx_2 \in A_n$ as well. Consequently $zx_1 F_n = zx_2 F_n = F_n$. But then

$$x_1 F_n = x_1 (zx_2) F_n = x_2 (zx_1) F_n = x_2 F_n$$

as desired.

Now let $x_1, \ldots, x_m \in A$ be a finite set such that $\pi_n(A) = \{\pi_n(x_1), \ldots, \pi_n(x_m)\}$. Then by what we have just shown, $\bigcup x_i F_n$ is invariant under $A$. By minimality, $\bigcup x_i F_n = F$, and thus $\bigcup x_i (F_n - F) = S^1$. Thus one of these closed sets has nonempty interior, which by $A_n$-invariance implies it is equal to $S^1$. That is, $F_n - F = S^1$ for all $n$. Consequently $F_\infty - F = S^1$, where $F_\infty = \bigcap F_n$.

Let $r$ be any point in $F_\infty$, and let $s = p/5^n \in S^1$. Then $s \in F_n - F$; say $s = s_n - f_n$. Now $A_n s_n = F_n \supset F_\infty$, so we can find a sequence $x_i \in A_n$ such that $x_i s_n \to r$. But then $x_i s = x_i (s_n - f_n) \to r - f'_n$, where $f'_n \in F$ as well. Thus $s \in r - F$. Since $s$ was an arbitrary rational of the form $p/5^n$, it follows that $F$ is dense in $S^1$.

Corollary 20.7 If $F$ is a $A$-minimal set, then $F$ is finite.

Proof. If $F$ is infinite then $X = F - F$ is a $A$-invariant set that accumulates at zero, so it is $S^1$; but then $F$ itself is $S^1$ by the preceding result, and $S^1$ is not minimal.

Proof of Theorem 20.1. It suffices to treat the case where $X = \overline{A x}$ and $x$ is irrational. In this case $X$ contains a minimal set and hence $X$ contains a rational point $p/q$. Since $x$ is irrational, $p/q$ is not isolated. Hence 0 is a non-isolated point of $qX$, and hence $qX = S^1$. This implies $X$ contains an open interval and hence $X = S^1$. 

146
The discriminant–regulator paradox

We now turn to conjectures about the equidistribution of periodic orbits, following [ELMV].

**Regulator and discriminant.** There are three basic invariants that can be attached to a compact orbit $A \cdot L \subset L_n$:

1. The *order* $O = \text{End}_A(L) = \text{End}(L) \cap \mathbb{R}[A]$;
2. The *regulator* $R = \text{vol}(A \cdot L) = \text{vol}(A/A_L)$; and
3. The *discriminant* $D = \text{vol}(\mathbb{R}[A]/\text{End}_A(L))^2$.

Here the space of diagonal matrices $\mathbb{R}[A] \cong \mathbb{R}^n$ is given the usual Euclidean volume, which can also be described using the inner product $\langle X,Y \rangle = \text{tr}(XY)$ on matrices.

Let $K = O \otimes \mathbb{Q}$ be the totally real field associated to $O$. Since $O$ is embedded in $\mathbb{R}[A]$ the $n$ real places of $K$, the trace on matrices in $O$ coincides with the usual $\text{tr}_K : K \to \mathbb{Q}$. Consequently if we choose a basis $(e_1, \ldots, e_n)$ for $O$ we see that

$$D = \det(e_{jj}^i)^2 = \det(\text{tr}(e^i e^j)) = \text{disc}(O)$$

coincides with the usual discriminant of $O$ from number theory. In particular $O$, as an abstract ring, determines $D$.

Similarly we use the exponential map $\exp : \mathbb{R}^n \to \mathbb{R}_+ \cdot A$ to identify $A$ with the locus $\sum x_i = 0$ in $\mathbb{R}^n$, with the Euclidean measure normalized so projection to any coordinate plane $\mathbb{R}^{n-1}$ is volume-preserving. Thus if we take a basis $(u_1, \ldots, u_{n-1})$ for $A_L \cong O_+^*$, we find

$$R = \text{vol}(A/A_L) = |\det(\log u_{jj}^i)| = 2^m \text{reg}(O),$$

where $\text{reg}(O)$ is the usual regulator (using the absolute values of all the units) and $2^m = |O^*/(\pm O_+^*)|$.

In particular, $O$ determines the invariants $D$ and $R$.

**Theorem 21.1** For any order $O$ we have:

$$C_n \log(D) < R < C_n \epsilon D^{1/2 + \epsilon}.$$ 

The lower bound is easy to see: e.g. the quadratic case, if $\epsilon > 1$ is a fundamental positive unit for $O$, then we have $R = \log |\epsilon|$; but since $\mathbb{Z}[\epsilon] \subset O$ we also have

$$D \leq \left( \det \left( \begin{array}{cc} 1 & \epsilon \\ \epsilon^{-1} & 1 \end{array} \right) \right)^2 = (\epsilon - \epsilon^{-1})^2 \leq \epsilon^2$$

147
which gives $(1/2) \log D \leq R$. The upper bound in the quadratic case follows from the result of Siegel:

**Theorem 21.2** For any real quadratic order $\mathcal{O}_D$ with class number $h(D)$ and regulator $R(D)$, we have:

$$h(D)R(D) \leq C \varepsilon D^{1/2+\varepsilon}.$$  

Recall that geometrically $h(D)$ is the number of closed geodesics in $\mathcal{L}_1[D]$ and $R(D)$ is the length of each.

**Warning:** Geodesics of the same length can be associated to different discriminants! Equivalently, $\mathbb{Z}[\varepsilon]$ can be (and often is) a proper suborder of $\mathcal{O}_D$, so different orders can have the same group of units.

For example, the fundamental positive unit in $\mathbb{Z}[\sqrt{2}] = \mathcal{O}_8$ is $\varepsilon = 3 + 2\sqrt{2}$, but $\mathbb{Z}[\varepsilon] = \mathbb{Z}[2\sqrt{2}] = \mathcal{O}_{32}$. So $\mathcal{O}_8$ and $\mathcal{O}_{32}$ have the same group of positive units.

Things get worse: for $\mathcal{O} = \mathbb{Z}[\sqrt{13}]$, a fundamental positive unit is $\varepsilon = 649 + 180\sqrt{13}$ which generates a subring of index 180.

**Equidistribution.** In the setting of unipotent orbits we have the following important result [MS]. Let $\Gamma$ be a lattice in a Lie groups $G$. Let us say an algebraic probability measure $\mu$ on $G/\Gamma$ is unipotent if it is ergodic for some one-parameter unipotent subgroup of $G$. Then we have:

**Theorem 21.3** The set of unipotent algebraic probability measures on $G/\Gamma$ is closed.

We have seen a special case of this phenomenon: a long closed horocycle on a Riemann surface of finite volume becomes equidistributed.

One might ask if the same result holds for the ergodic $A$-invariant measures on $\mathcal{L}_n = G/\Gamma = \text{SL}_n(\mathbb{R})/\text{SL}_n(\mathbb{Z})$. Of course it fails badly for $n = 2$: the limits need not be algebraic. In addition, the measure may ‘escape to infinity’: there are closed geodesics that spend most of their time in the cusp of $\mathcal{L}_2$.

Using a result of Duke [Du2], it is shown in [ELMV, §7.2] that measure can also escape to infinity in $\mathcal{L}_n$:

**Theorem 21.4** For any $n \geq 2$ there is a sequence of compact $A$-orbits whose associated probability measures $\mu_n$ converge to an $A$-invariant measure $\nu$ with total mass strictly less than one.
It is still unknown if the limit of such $\mu_n$ can be ‘exotic’, e.g. if it can have non-algebraic support.

**Large orbits.** In these ‘counterexamples’, the compact $A$-orbits are small, in the sense that the regulator $R$ is on the order of $(\log D)^{n-1}$ rather than $D^{1/2}$. Such an orbit represents only a small part of $L_n(\mathcal{O})$. Thus one is led (following [ELMV]) to:

**Conjecture 21.5** As the discriminant of $\mathcal{O}$ goes to infinity, the union $L_n(\mathcal{O})$ of all the associated compact $A$-orbits becomes equidistributed in $L_n$.

**Conjecture 21.6** If $X_i$ is a sequence of compact $A$-orbits in $L_n$ with regulators and discriminants satisfying

$$R_i > D_i^\epsilon \to \infty$$

for a fixed $\epsilon > 0$, then $X_i$ becomes uniformly distributed as $i \to \infty$.

**The case $n = 2$.** These conjectures have interesting content even for $L_2$. For $n = 2$ the first was studied by Linnik and resolved by Duke at least in the case of fundamental discriminants [Du1]. The second is still open.

For $\mathcal{M}_1 = \mathbb{H}/\mathrm{SL}_2(\mathbb{Z})$ we can relate the second conjecture to a general discussion about hyperbolic surfaces $X$ of finite volume. For any such $X$ there is a sequence of closed geodesics $\gamma_n$ which become uniformly distributed in $T_1(X)$ as $n \to \infty$. Equivalently, the length of any closed geodesic $\delta$ on $X$ satisfies

$$L(\delta) = (\pi/2) \text{area}(X) \lim_{X(\gamma_n)} \frac{i(\gamma_n, \delta)}{L_X(\gamma_n)}.$$

How can we distinguish a sequence $g_n \in \mathrm{SL}_2(\mathbb{Z})$ whose associated equidistributed in this sense? The proposed answer is to take any sequence of geodesics that are ‘long’ relative to their discriminants.

**Discriminant and regulator in $\mathrm{SL}_2(\mathbb{Z})$.** Let us now consider a closed geodesic on the modular surface $\mathcal{M}_1$ corresponding to an element

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}).$$

We will compute from $g$ its invariants $\mathcal{O}$, $R$ and $D$. Acting by the center of $\mathrm{SL}_2(\mathbb{Z})$, we can assume

$$t = \text{tr}(g) = a + d > 2.$$
1. The order $\mathcal{O} \subset M_2(\mathbb{Z})$ is just the subgroup of integral matrices commuting with $g$.

Thus $\mathcal{O} \otimes \mathbb{Q} = \mathbb{Q}(g)$, and certainly $\mathcal{O} \supset \mathbb{Z}[g]$, but it may be much larger (just as $\mathcal{O}$ may be larger than $\mathbb{Z}[\epsilon]$).

2. The regulator is easy to compute. The element $g$ corresponds to a fundamental positive unit $\epsilon \in \mathcal{O}$ with

$$\epsilon + \epsilon^{-1} = e^R + e^{-R} = 2 \cosh(R) = \text{tr}(g) = t,$$

and thus

$$R = \log |\text{tr}(g)| + O(1).$$

3. While the regulator is an invariant of the conjugacy class of $g$ in $\text{SL}_2(\mathbb{R})$, the order $\mathcal{O}$ and its discriminant $D$ depend on the conjugacy class in $\text{SL}_2(\mathbb{Z})$.

**Theorem 21.7** We have $D = (t^2 - 4)/(\gcd(a - d, b, c))^2$.

More generally, the order in $M_2(\mathbb{Z})$ given by the commutator of $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{Z})$ has discriminant

$$D = \frac{(a - d)^2 + 4bc}{\gcd(a - d, b, c)^2},$$

and is generated by $(1/g) \begin{pmatrix} a - d & b \\ c & 0 \end{pmatrix}$. This agrees with the formula above when $\det(g) = ad - bc = 1$.

**Lenstra’s heuristics: the paradox.** We now come to the paradox. Recall that for $n = 2$ the discriminant $D$ determines $\mathcal{O}_D$, $h(D)$ and $R(D)$, which satisfy

$$h(D)R(D) \asymp D^{1/2}.$$ 

Now according to heuristics of Cohen and Lenstra there should a definite percentage of $D$ such that $h(D) = 1$ [CL], i.e. such that there is a unique geodesic on $\mathcal{M}_1$ of discriminant $D$. Thus it should be likely for $R(D)$ to be comparable to $D^{1/2}$, and hence with a unique associated geodesic.

(Even if $h(D) = 1$, there need not be a unique geodesic on $\mathbb{H}/\text{SL}_2(\mathbb{Z})$ of length $R(D)$. If the fundamental unit doesn’t generate the maximal order, there can be lots of other discriminants $D$ with the same regulator, and each gives at least one geodesic of the same length.)

On the other hand, it is easy to construct sequences $g_n \in \text{SL}_2(\mathbb{Z})$ such that:
1. $g_n$ gives a primitive geodesic on $\mathcal{M}_1$,

2. $\text{tr}(g_n) \to \infty$,

3. the geodesics $g_n$ do not become equidistributed.

In fact, almost any construction works. For example, let

$$g_n = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^n \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}^n = \begin{pmatrix} 1 + n & 2 \\ n & 1 \end{pmatrix};$$

this geodesic spends most of its time in the cusp. Or take $g_n = A^n B^n$ where $A$ and $B$ are hyperbolic; in the limit this geodesic spirals between the geodesics for $A$ and $B$.

Let $D_n$ be the corresponding discriminants. Since these geodesics do not become equidistributed, we should have $\mathcal{R}(D_n) = O(D_n^\epsilon)$ for every $\epsilon > 0$.

Thus:

$$h(D_n) \gg D_n^{1/2-\epsilon};$$

in particular, the class numbers for any such construction must go rapidly to infinity! How does this fit with the Cohen-Lenstra heuristic?

**Resolution.** An explanation is that when $D$ is chosen ‘at random’, the trace of its fundamental positive unit is not random at all. Namely we have:

**Theorem 21.8** Let $\epsilon \in \mathcal{O}_D$ be unit of norm 1 and trace $t \in \mathbb{Z}$. Then $(t^2 - 4)/D$ is a square.

**Proof.** We can write $\epsilon = a + b(D + \sqrt{D})/2$; then $t = 2a + Db$ and

$$4 = 4N(\epsilon) = (2a + bD)^2 - b^2D = t^2 - b^2D$$

and so $(t^2 - 4)/D = b^2$. ■

Now the good geodesic should have log $t$ of size $D^{1/2}$, i.e. $t$ should be enormous compared to $D$ (as is often the case for a fundamental unit). It should then furthermore have the remarkable property that $t^2 - 4$ is ‘almost a square’.

Thus the set of $t$ that arise from ‘typical $D$’ are very unusual! In other words, we get a totally different ‘measures’ on the set of closed geodesics if we order them by length and if we order them by discriminant.

151
**Example.** Fundamental positive units $\epsilon$ for $\mathbb{Z}[\sqrt{n}]$ (with $D = 4n$) are computed in Table 6. For example, when $n = 29$ we have $\epsilon = 9801 + 1820\sqrt{29}$, with trace $t = 19602$ and

$$t^2 - 4 = 384238400 = 29 \cdot 3640^2.$$  

Turning the resolution, we obtain theorems such as:

**Theorem 21.9** As $t \to \infty$, the class number $h(t^2 - 4) \to \infty$.

**Proof.** The matrix $g = \begin{pmatrix} 0 & 1 \\ -1 & n \end{pmatrix}$ has trace $t$ and discriminant $D = t^2 - 4$, so the class number must be comparable to $D^{1/2}$.  

(The matrix $g$ corresponds to the unit $\epsilon = (t + \sqrt{t^2 - 4})/2$.)

Indeed it can be shown that most fields have large class number, if they are organized by their regulators; see [Sp2], [Sp1].

**Challenge.** Find explicit $g_n \in \text{SL}_2(\mathbb{Z})$ and $\epsilon > 0$ such that $R(g_n) > D(g_n)^\epsilon \to \infty$.

**Circle maps.** We conclude by discussing an analogous phenomenon for the doubling map $f : S^1 \to S^1$. (Note that again there are nontrivial conjectures and results without the need for a rank two action like $\times 2 \times 3$.)

In this setting a compact orbit corresponds to a periodic cycle. Now if $x$ has period $R$, then $x = p/D$ where $D$ is odd, and $2^R = 1 \mod D$. In this setting we regard:

1. the denominator $D$ as the **discriminant** of the orbit; and
2. the period $R$ as the **regulator**.

Note that $R = R(D) = \text{the order of } 2 \text{ in } (\mathbb{Z}/D)^\times$. It does not depend on $x$. As before we have

$$\log_2(D) < R < \phi(D) < D.$$  

We can also let $h(D)$ denote the number of orbits of the doubling map on $(\mathbb{Z}/D)^\times$; we then have:

$$R(D)h(D) = \phi(D).$$

**Binary expansions.** Note that $R(D)$ is the same as the period of the fraction

$$1/D = .x_1 \cdots x_R x_1 \cdots x_R \cdots$$

in base two. The fact that $R$ can be quite long — as long as $D$ itself — is like the fact that the fundamental unit can be very large. We will conjecture that the case where the period is long is dynamically well-behaved.
Examples. We have seen that there are periodic cycles on which \( f \) behaves like a rotation, and that these can only accumulate on subsets of \( S^1 \) of Hausdorff dimension one. In particular they are very badly distributed. We have also seen that there are infinite-dimensional families of smooth expanding mappings that are topologically conjugate to \( f(x) \) and hence give rise to ergodic measures which can in turn be encoded as limits of periodic orbits. Thus there are many sources of orbits which are not uniformly distributed.

Equidistribution of long orbits. The discriminant \( D \) is highly sensitive to the arithmetic of \( S^1 = \mathbb{R}/\mathbb{Z} \). The orbits mentioned above will tend to have large values of \( D \), e.g. \( D = 2^R - 1 \). On the other hand we can formulate:

**Conjecture 21.10** If \( X_n \subset S^1 \) is a sequence of periodic cycles such that \( R_n > D_n^\epsilon \), then \( X_n \) becomes uniformly distributed on the circle.

This conjecture has now been resolved by [Bo]; see below.

The paradox. In this case the paradox assumes the following (more comprehensible) form.

Suppose we pick \( D \) at random. Then there is a good chance that \( R(D) \) is comparable to \( D \). Indeed, the Cohen-Lenstra heuristic is now similar to a conjecture of Artin’s, which asserts that 2 is a generator of \( (\mathbb{Z}/D)^* \) for infinitely many (prime) \( D \). In other words, \( R(D) = D - 1 \) infinitely often.

On the other hand, suppose we construct a point \( x \) of period \( R \) at random. Then \( x = p/(2^R - 1) \) for some \( p \). Now it is likely that there is little cancellation in this fraction, and so \( D \) is comparable to \( 2^R \).

The resolution in this case is again that only very special periodic points \( x = p/D = s/(2^R - 1) \) arise for \( D \) with \( h(D) = O(D^\epsilon) \). Namely for such \( D \) the number
\[
s = x_1 \cdots x_R \asymp 2^R
\]
must almost be a divisor of \( 2^R - 1 \); more precisely we have:
\[
sD = 0 \mod 2^R - 1.
\]

Thus the digits of \( s \) must be carefully chosen so \( s \) accounts for all the divisors of \( 2^R - 1 \) not present in \( D \), so the fraction \( s/(2^R - 1) \) collapses to the much simpler fraction \( p/D \). Such \( s \) are very rare in the range \([0, 2^R]\).

Positive results. From [BGK] we have:

**Theorem 21.11** Fix \( \epsilon > 0 \). Let \( p_n \to \infty \) be a sequence of primes, and let \( G_n \subset S^1 \) be a sequence of multiplicative subgroups of the \( p_n \)th roots of unity such that \( |G_n| > p_n^\epsilon \). Then \( G_n \) becomes uniformly distributed on \( S^1 \).
Even better one has a bound on Gauss sums: for any \( a \neq 0 \mod p_n \),

\[
\left| \sum_{G_n} z^a \right| \leq |G|p_n^{-\delta}.
\]

**Corollary 21.12** Let \( p_n \) be a sequence of primes such that the order of 2 in \((\mathbb{Z}/p_n)^*\) is greater than \( p_n^\epsilon \). Then the orbits of \( 1/p_n \) under \( x \mapsto 2x \mod 1 \) become equidistributed as \( n \to \infty \).

Note: \( \log D \ll N \ll D \). Cases where \( N(x) \propto \log D(x) \) are easily constructed by taking \( x = p/(2^n - 1) \).

The assumption that the denominators \( p_n \) are primes is completely removed in [Bo].

For arithmetic study of \( x \mapsto 2x \), it may be useful to consider the ring \( \mathcal{O} = \mathbb{Z}[1/2] \). Although this ring is infinitely-generated as an additive group, its unit group \((\pm 1) \times 2^\mathbb{Z}\) is essentially cyclic, so it behaves like a real quadratic field.

## A Appendix: The spectral theorem

For reference this section provides the statement of the spectral theorem and a detailed sketch of the proof.

**Hilbert space.** Let \( H \) be a nontrivial separable Hilbert space over the complex numbers. We remark that such a Hilbert space is determined up to isomorphism by its dimension, which can be \( 1, 2, 3, \ldots, \infty \). For example, if \( \dim H = \infty \), then it admits an orthonormal basis \((e_i)_{i=0}^\infty\); then we have an isomorphism

\[
\iota : \ell^2(\mathbb{N}) \cong H
\]

defined by

\[
\iota(a_i) = \sum_{0}^{\infty} a_i e_i.
\]

Here \( \|a_i\|^2 = \sum |a_i|^2 \|\iota(a_i)\| \).

Two other useful models for the same Hilbert space are given by \( L^2(S^1) \) and \( \ell^2(\mathbb{Z}) \). Here the norm on \( L^2(S^1) \) is given by

\[
\|f\|^2 = \frac{1}{2\pi} \int_{S^1} |f(z)|^2 |dz|.
\]
With this norm, the functions $e_i(z) = z^i$, $i \in \mathbb{Z}$, for an orthonormal basis, and give the isomorphism

$$L^2(S^1) \cong \ell^2(\mathbb{Z}).$$

This isomorphism can also be regarded as an instance of Pontryagin duality, i.e. the ‘Fourier transform’ isomorphism $L^2(G) \cong L^2(\hat{G})$ between $L^2$ of a locally compact group and its dual.

**Operators and algebras.** Let $\mathcal{B}(H)$ denote the space of bounded linear operators $A : H \to H$, with the usual operator norm:

$$\|A\| = \sup_{\|x\|=1} \|Ax\| = \sup_{\|x\|=\|y\|=1} |\langle Ax, y \rangle|.$$

The space $\mathcal{B}(H)$ is an Banach algebra: that is, we have $\|AB\| \leq \|A\| \cdot \|B\|$. More importantly, $\mathcal{B}(H)$ is an example of a $C^*$-algebra. That is, the natural map $A \mapsto A^*$ sending an operator to its transpose satisfies the important identity:

$$\|A^*A\| = \|A\|^2.$$

To see this identity, note that $\|A^*\| = \|A\|$ and thus:

$$\|A\|^2 \geq \|A^*A\| = \sup_{\|x\|=\|y\|=1} |\langle A^*Ax, y \rangle| = \sup_{x,y} \sup_{\|x\|=\|y\|=1} |\langle Ax, Ay \rangle| \geq \sup_x |\langle Ax, Ax \rangle| = \|A\|^2.$$

**Spectra and $C^*$-algebras.** The spectrum of $A \in \mathcal{B}(H)$ is defined by

$$\sigma(A) = \{ \lambda \in \mathbb{C} : (\lambda I - A) \text{ is not invertible in } \mathcal{B}(H) \}.$$

(Note: we require a 2-sided inverse.) It is easy to see that the spectrum is a bounded, closed subset of the complex plane. The **spectral radius** of $A$ is defined by

$$\rho(A) = \sup \{ |\lambda| : \lambda \in \sigma(A) \} = \lim \|A^n\|^{1/n}. $$

Now suppose $A$ and $A^*$ commute, i.e. $A$ is a **normal operator**. The most important cases arise when $A$ is self-adjoint ($A = A^*$), and when $A$ is unitary ($A^* = A^{-1}$). Then the norm closure of the polynomial algebra generated by $A$ and $A^*$ gives a **commutative $C^*$–algebra**, $\mathcal{A} = \overline{\mathbb{C}[A]} \subset \mathcal{B}(H)$.

By the general theory of $C^*$ algebras, the algebra $\mathcal{A}$ is isomorphic, as a normed $*$-algebra, to the full algebra of continuous functions on its space of maximal ideals. Moreover a maximal ideal $m$ corresponds to the kernel of a
The multiplicative linear functional \( \phi : A \to C \). Clearly \( \phi \) is uniquely determined by the value \( \lambda = \phi(A) \), and moreover

\[ A - \lambda I \in \text{Ker} \phi = m, \]

so \( \lambda \in \sigma(A) \). The converse is also true — every element of the spectrum gives a unique maximal ideal \( m \). The ideal \( m \) corresponding to \( \lambda \) can be constructed from starting with the principal ideal \( (\lambda I - A) \) and extending it to a maximal ideal; by the preceding argument, the result is unique.

Hence we have an isomorphism

\[ A = \mathbb{C}[A] \cong C(\sigma(A)). \]

This isomorphism sends the \( L^2 \)-norm to the sup–norm, and the adjoint operation to complex conjugation.

To justify these statements, one uses the key fact that \( \|A\| = \rho(A) \) for a normal operator. For example, when \( A \) is self–adjoint, this follows from the equation \( \|A^*A\| = \|A\|^2 = \|A^2\| \), which implies \( \|A\| = \|A^n\|^{1/n} \to \rho(A) \).

**Statements of the spectral theorem.** We are now in a position to formulate the spectral theorem. We wish to give a model for a general normal operator \( A \). The simplest such model is the following. Let \( K \) be a compact subset of the complex plane, let \( \mu \) be a probability measure on \( K \) of full support, and let \( H = L^2(K, \mu) \). Then we have a natural map from \( C(K) \) into \( \mathcal{B}(H) \), given by

\[ T_f(g) = f(x)g(x). \]

This map is a \( C^* \)-algebra isomorphism to its image; in particular, \( \|T_f\|_2 = \sup \|f\|_\infty \). (Here we use the fact that \( \mu \) has full support.) Moreover, \( f \) is invertible iff it does not vanish on \( K \); thus the spectrum of \( T_f \) is given by \( \sigma(A_f) = f(K) \). In particular, if \( f(z) = z \), then \( A = T_f \) satisfies

\[ \sigma(A) = K. \]

Note that the eigenvectors for \( A \) correspond to the atoms of \( \mu \); aside from these, \( A \) has a continuous spectrum. In particular, if \( \mu \) has no atoms then \( A \) has no eigenvalues, properly speaking; but for any measurable subset \( E \subset K \), one can regard the subspace

\[ L^2(E, \mu|E) \subset L^2(K, \mu) \]

as the subspace where the eigenvalues of \( A \) are, morally, in \( E \).

156
We consider the eigenvalues of $A$ to have \textit{multiplicity one}, since the elements of $L^2(\mu|E)$ take values in a 1-dimensional space (namely $\mathbb{C}$).

\textbf{Multiplicity.} It is apparent that we need to incorporate the possibility of \textit{multiple eigenvalues} into our model. To this end, for $n = 1, 2, \ldots, \infty$ we let

$$H_n = \ell^2(\mathbb{Z}/n),$$

where $\mathbb{Z}/\infty = \mathbb{Z}$. These are model Hilbert spaces of dimension $n$. Continuing with the space $(K, \mu)$ above, we let $L^2(K, \mu, H_n)$ denote the space of functions $f : K \to H_n$ with the norm:

$$\|f\|_2^2 = \int_K \|f(x)\|^2 d\mu(x). \quad (A.1)$$

Now the atoms of $\mu$ give $n$–dimensional eigenspaces for the operator $A(f) = zf(z)$.

Finally we must allow the multiplicity of eigenvalues to vary. Thus we introduce a measurable multiplicity function

$$m : K \to \mathbb{Z}_+ \cup \{\infty\},$$

and use it to form a \textit{Hilbert space bundle} $\mathcal{H}(m) \to K$, such that the fiber over $x$ is $H_{m(x)}$. Finally we let $L^2(K, \mu, m)$ denote the space of sections $f : K \to \mathcal{H}(m)$, with the norm again given by (A.1).

Since we are working in the Borel category, the bundle $\mathcal{H}(m)$ be can be trivialized. More precisely, we can partition $K$ into measure sets $E_n = m^{-1}(n)$; then $\mathcal{H}|E_n \cong E_n \times H_n$, and we have

$$L^2(K, \mu, m) = \bigoplus_{n=1}^{\infty} L^2(E_n, \mu|E_n, H_n).$$

We may now finally state:

\textbf{Theorem A.1 (The spectral theorem)} Let $A \in \mathcal{B}(H)$ be a bounded operator on a separable Hilbert space, such that $[A, A^*] = 0$. Then there exists a Borel measure $\mu$ of full support on $\sigma(A)$, and a multiplicity function $m$ on $K$, and an isomorphism

$$H \cong L^2(\sigma(A), \mu, m)$$

sending the action of $A \in \mathcal{B}(H)$ to the operator $f(z) \mapsto zf(z)$ on $L^2(\sigma(A), \mu, m)$. 

157
Complement. The measure class of $\mu$, and the function $m$ (defined a.e.), are complete invariants of $A$.

(Measures $\mu$ and $\nu$ on $\mathbb{C}$ are in the same measure class if they have the same sets of measure zero.)

Self–adjoint operators and quantum theory. To describe the proof as well as some of the physical intuition behind it, consider the case of a self–adjoint operator $A$ on a separable Hilbert space $H$. As above we let $A \subset B(H)$ denote the norm–closed, commutative $C^*$–algebra generated by $A$. We then have an isomorphism

$$A \cong C(\sigma(A)),$$

sending $A$ to the function $f(z) = z$. Since $A = A^*$, we have $f(z) = \overline{f}(z)$, and thus $\sigma(A) \subset \mathbb{R}$. It is natural to write $f(A)$ for the element of $B(H)$ corresponding to $f \in C(\sigma(A))$. Indeed, this isomorphism is part of the ‘functional calculus’, which allows one to compose $A$ with various functions such as $f$ on its spectrum.

Now any vector $\psi \in H$ of norm one determines a state on the algebra $A$, i.e. a positive linear functional, by

$$\phi : f(A) \mapsto \langle f(A)\psi,\psi \rangle.$$

Here positive means that $\phi(aa^*) \geq 0$ for all $a \in A$.

This functional extends by continuity to a bounded operator on $A \cong C(\sigma(A))$. If $f \in C(\sigma(A))$ is non-negative, then it can be written as $f = g^2$ with $g = g$. Then on the level of operators we have $f(A) = g(A)g(A)^*$, so $\phi(f) \geq 0$. This implies that $\phi$ can be regarded as a measure $\mu$ on the spectrum $\sigma(A)$, i.e. there is a unique measure $\mu$ such that

$$\phi(f) = \int_{\sigma(A)} f(x) \, d\mu.$$

Since $\phi(1) = \|\psi\|^2 = 1$, $\mu$ is a probability measure.

Quantum interpretation. In quantum mechanics, self–adjoint operators correspond to observables, which take random values. The expected value of the observable $A$ for a system in the state $\psi$ is given by

$$\langle A \rangle = \langle A\psi,\psi \rangle = \int_{\sigma(A)} \, d\mu.$$

The spectral measure $\mu$ gives much more complete information: it determines the distribution of the random variable $A$. 

158
Using the spectral measure coming from the state \( \psi \), we obtain an isometric injection
\[
A \cdot \psi \rightarrow L^2(\sigma(A), \mu),
\]
which extends by continuity to an isomorphism on the norm closure \( H_1 \subset H \) of left–hand side. Moreover, under this isomorphism, the action of \( A \) becomes \( A(f) = xf(x) \).

If \( H_1 = H \), we are done with the proof of the spectral theorem; otherwise, we apply the same analysis starting with a suitable new stable in \( H_1^+ \). Using separability, we can insure that the process terminates after countably many steps. We then get a spectral decomposition
\[
H \cong \bigoplus_1^\infty L^2(\sigma(A), \mu_n),
\]
for a finite or countable sequence of probability measures \( \mu_n \).

This countable sum can be easily converted to the form given in the spectral theorem as stated above. Namely we let \( \mu = \sum_1^\infty \mu_n/2^n \) (in the case of an infinite sum), we let
\[
E_n = \{ x : d\mu_i/d\mu > 0 \}
\]
and we let \( m(x) = \sum \chi_{E_n}(x) \). It is then easy to construct an isomorphism
\[
\bigoplus_1^\infty L^2(\sigma(A), \mu_n) \cong L^2(\sigma(A), \mu, m),
\]
compatible with the action of \( A \).

The construction makes no real use of the fact that \( A \) is self–adjoint, so it applies to a general normal operator.

References


162


