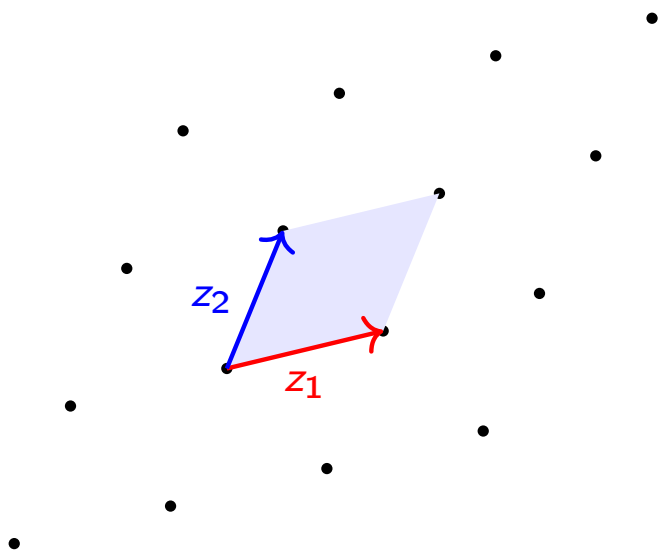
Examples in dynamical systems
Lattices, ergodicity and interval exchanges

Pierre Arnoux

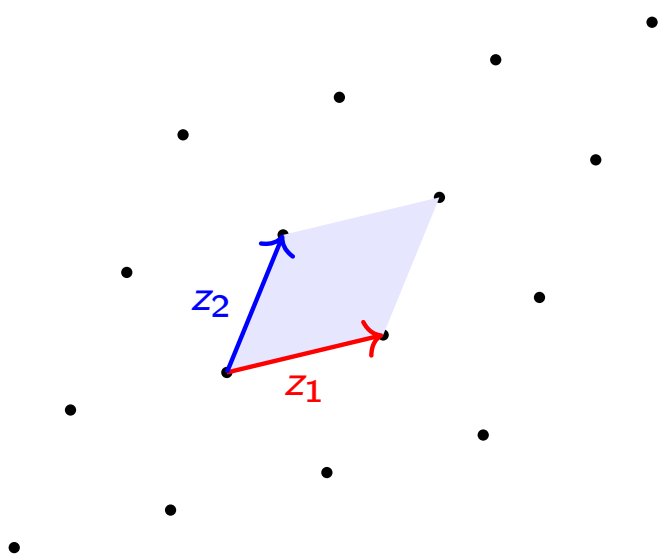
Lecture 8

A basis of a lattice is a pair of complex numbers



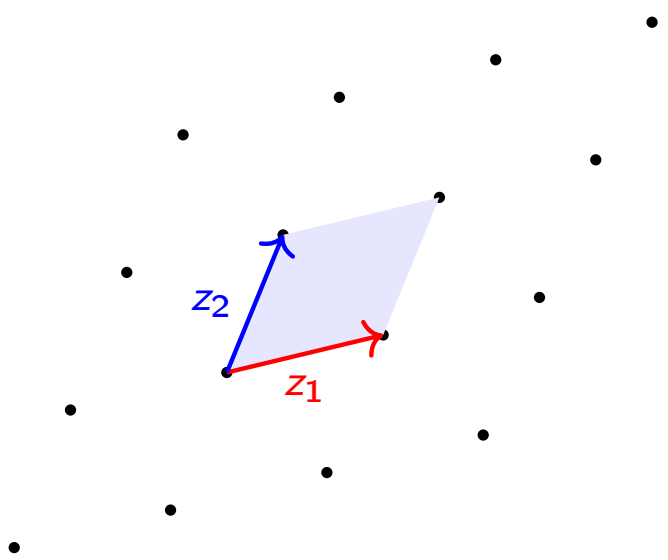
- Identify $\mathbb{R}^2$ with $\mathbb{C}$. A **lattice** is a discrete subgroup $\Lambda = \mathbb{Z}z_1 + \mathbb{Z}z_2$, with z_1, z_2 linearly independent over $\mathbb{R}$. A **basis** is a pair (z_1, z_2).

A basis of a lattice is a pair of complex numbers



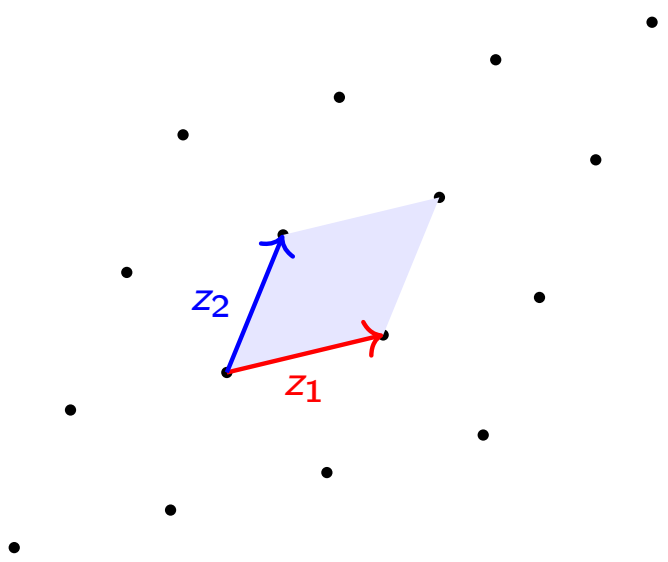
- Identify $\mathbb{R}^2$ with $\mathbb{C}$. A **lattice** is a discrete subgroup $\Lambda = \mathbb{Z}z_1 + \mathbb{Z}z_2$, with z_1, z_2 linearly independent over $\mathbb{R}$. A **basis** is a pair (z_1, z_2).
- To a basis we associate the complex number $\tau = \dfrac{z_2}{z_1} \in \mathbb{C} \setminus \mathbb{R}$.

A basis of a lattice is a pair of complex numbers



- Identify $\mathbb{R}^2$ with $\mathbb{C}$. A **lattice** is a discrete subgroup $\Lambda = \mathbb{Z}z_1 + \mathbb{Z}z_2$, with z_1, z_2 linearly independent over $\mathbb{R}$. A **basis** is a pair (z_1, z_2).
- To a basis we associate the complex number $\tau = \dfrac{z_2}{z_1} \in \mathbb{C} \setminus \mathbb{R}$.
- $\Im\tau = \dfrac{\Im(z_2\bar{z}_1)}{|z_1|^2} = \dfrac{\det(z_1, z_2)}{|z_1|^2}$: so $\tau \in \mathbb{H}$ **if and only if the basis is positively oriented**.

A basis of a lattice is a pair of complex numbers



- Identify $\mathbb{R}^2$ with $\mathbb{C}$. A **lattice** is a discrete subgroup $\Lambda = \mathbb{Z}z_1 + \mathbb{Z}z_2$, with z_1, z_2 linearly independent over $\mathbb{R}$. A **basis** is a pair (z_1, z_2).
- To a basis we associate the complex number $\tau = \dfrac{z_2}{z_1} \in \mathbb{C} \setminus \mathbb{R}$.
- $\Im\tau = \dfrac{\Im(z_2\bar{z}_1)}{|z_1|^2} = \dfrac{\det(z_1, z_2)}{|z_1|^2}$: so $\tau \in \mathbb{H}$ **if and only if the basis is positively oriented**.
- τ is unchanged if we replace (z_1, z_2) by (cz_1, cz_2), $c \in \mathbb{C}^*$: it is invariant by **rotation and homothety**. It only depends on the *shape* of the lattice.

Changing the basis: the action of $\mathrm{SL}(2,\mathbb{Z})$

- Let (z_1, z_2) and (z_1', z_2') be two bases of the same lattice: $z_1' = az_1 + bz_2$, $z_2' = cz_1 + dz_2$, with $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}(2,\mathbb{Z})$.

Changing the basis: the action of $\mathrm{SL}(2,\mathbb{Z})$

- Let (z_1, z_2) and (z_1', z_2') be two bases of the same lattice: $z_1' = az_1 + bz_2$, $z_2' = cz_1 + dz_2$, with $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}(2,\mathbb{Z})$.
- If both bases are positively oriented, $\det M = 1$.

Changing the basis: the action of $\mathrm{SL}(2,\mathbb{Z})$

- Let (z_1, z_2) and (z_1', z_2') be two bases of the same lattice: $z_1' = az_1 + bz_2$, $z_2' = cz_1 + dz_2$, with $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}(2,\mathbb{Z})$.
- If both bases are positively oriented, $\det M = 1$.
- The new invariant is

$$\tau' = \frac{cz_1 + dz_2}{az_1 + bz_2} = \frac{c + d\tau}{a + b\tau} = \gamma \cdot \tau, \quad \gamma = \begin{pmatrix} d & c \\ b & a \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z}).$$

Changing the basis: the action of $\mathrm{SL}(2,\mathbb{Z})$

- Let (z_1, z_2) and (z_1', z_2') be two bases of the same lattice: $z_1' = az_1 + bz_2$, $z_2' = cz_1 + dz_2$, with $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}(2,\mathbb{Z})$.
- If both bases are positively oriented, $\det M = 1$.
- The new invariant is

$$\tau' = \frac{cz_1 + dz_2}{az_1 + bz_2} = \frac{c + d\tau}{a + b\tau} = \gamma \cdot \tau, \quad \gamma = \begin{pmatrix} d & c \\ b & a \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z}).$$

- Conversely, every $\gamma \in \mathrm{SL}(2,\mathbb{Z})$ occurs. So **the invariants of the positively oriented bases of a lattice form exactly one orbit of $\mathrm{SL}(2,\mathbb{Z})$ in $\mathbb{H}$.**

The modular surface is the space of lattices

- **Theorem**: the map which sends a lattice to the $\mathrm{SL}(2,\mathbb{Z})$-orbit of τ is a bijection from the set of lattices of $\mathbb{C}$, up to rotation and homothety, onto the set of orbits $\mathrm{SL}(2,\mathbb{Z})\backslash\mathbb{H}$.

The modular surface is the space of lattices

- **Theorem**: the map which sends a lattice to the $\mathrm{SL}(2,\mathbb{Z})$-orbit of τ is a bijection from the set of lattices of $\mathbb{C}$, up to rotation and homothety, onto the set of orbits $\mathrm{SL}(2,\mathbb{Z})\backslash\mathbb{H}$.
- **Proof**: well defined, by the previous slide. *Onto*: $\tau \in \mathbb{H}$ is the invariant of the lattice $\mathbb{Z}+\mathbb{Z}\tau$. *One-to-one*: if two lattices have bases with the same τ, then $z_1' = cz_1$, $z_2' = cz_2$ with $c = z_1'/z_1 \in \mathbb{C}^*$: $\Lambda' = c\Lambda$.

The modular surface is the space of lattices

- **Theorem**: the map which sends a lattice to the $\mathrm{SL}(2,\mathbb{Z})$-orbit of τ is a bijection from the set of lattices of $\mathbb{C}$, up to rotation and homothety, onto the set of orbits $\mathrm{SL}(2,\mathbb{Z})\backslash\mathbb{H}$.
- **Proof**: well defined, by the previous slide. *Onto*: $\tau \in \mathbb{H}$ is the invariant of the lattice $\mathbb{Z}+\mathbb{Z}\tau$. *One-to-one*: if two lattices have bases with the same τ, then $z_1' = cz_1$, $z_2' = cz_2$ with $c = z_1'/z_1 \in \mathbb{C}^*$: $\Lambda' = c\Lambda$.
- So the modular surface **is** the space of shapes of lattices: a point of the surface is a lattice, up to rotation and homothety.

Topology and distance on the space of lattices

- **Topology**: the quotient topology of $\mathrm{SL}(2,\mathbb{Z})\backslash\mathbb{H}$. Equivalently, $\Lambda_n \to \Lambda$ if there are bases of Λ_n and Λ with the same orientation, which converge up to similarity.

Topology and distance on the space of lattices

- **Topology**: the quotient topology of $\mathrm{SL}(2,\mathbb{Z})\backslash\mathbb{H}$. Equivalently, $\Lambda_n \to \Lambda$ if there are bases of Λ_n and Λ with the same orientation, which converge up to similarity.
- **Distance**: for two lattices with invariants $\tau, \tau' \in \mathbb{H}$,

$$d(\Lambda, \Lambda') = \min_{\gamma \in \mathrm{SL}(2,\mathbb{Z})} d_{\mathbb{H}}(\tau, \gamma \cdot \tau').$$

It is well defined because $\mathrm{SL}(2,\mathbb{Z})$ acts by isometries, and the minimum exists because the action is properly discontinuous. It is the distance of the modular surface.

Topology and distance on the space of lattices

- **Topology**: the quotient topology of $\mathrm{SL}(2,\mathbb{Z})\backslash\mathbb{H}$. Equivalently, $\Lambda_n \to \Lambda$ if there are bases of Λ_n and Λ with the same orientation, which converge up to similarity.
- **Distance**: for two lattices with invariants $\tau, \tau' \in \mathbb{H}$,

$$d(\Lambda, \Lambda') = \min_{\gamma \in \mathrm{SL}(2,\mathbb{Z})} d_{\mathbb{H}}(\tau, \gamma \cdot \tau').$$

  It is well defined because $\mathrm{SL}(2,\mathbb{Z})$ acts by isometries, and the minimum exists because the action is properly discontinuous. It is the distance of the modular surface.
- **Example**: the distance between the square lattice ($\tau = i$) and the hexagonal lattice ($\tau = e^{i\pi/3}$) is $\frac{1}{2}\ln 3$ (computed with $\cosh d(z,w) = 1 + \frac{|z-w|^2}{2\Im z\, \Im w}$).

The shortest basis

- **Algorithm**: take z_1 a shortest nonzero vector of Λ; then take z_2 a shortest vector such that (z_1, z_2) is positively oriented. Let $\tau = z_2/z_1$.

The shortest basis

- **Algorithm**: take z_1 a shortest nonzero vector of Λ; then take z_2 a shortest vector such that (z_1, z_2) is positively oriented. Let $\tau = z_2/z_1$.
- Since $|z_2| \geq |z_1|$, we have $|\tau| \geq 1$.

The shortest basis

- **Algorithm**: take z_1 a shortest nonzero vector of Λ; then take z_2 a shortest vector such that (z_1, z_2) is positively oriented. Let $\tau = z_2/z_1$.
- Since $|z_2| \geq |z_1|$, we have $|\tau| \geq 1$.
- Since $|z_2 \pm z_1| \geq |z_2|$, we have $|\tau \pm 1| \geq |\tau|$, that is $|\Re\tau| \leq \frac{1}{2}$.

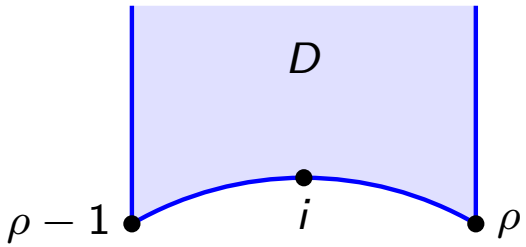
The shortest basis

- **Algorithm**: take z_1 a shortest nonzero vector of Λ; then take z_2 a shortest vector such that (z_1, z_2) is positively oriented. Let $\tau = z_2/z_1$.
- Since $|z_2| \geq |z_1|$, we have $|\tau| \geq 1$.
- Since $|z_2 \pm z_1| \geq |z_2|$, we have $|\tau \pm 1| \geq |\tau|$, that is $|\Re\tau| \leq \frac{1}{2}$.
- So $\tau \in D = \{\tau \in \mathbb{H} : \ |\Re\tau| \leq \frac{1}{2}, \ |\tau| \geq 1\}$: **the usual fundamental domain**.

The shortest basis

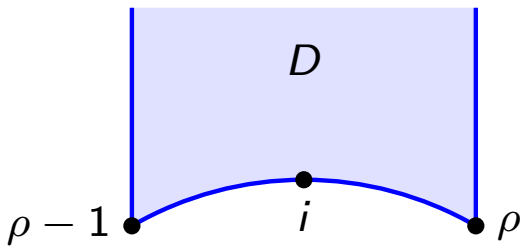
- **Algorithm**: take z_1 a shortest nonzero vector of Λ; then take z_2 a shortest vector such that (z_1, z_2) is positively oriented. Let $\tau = z_2/z_1$.
- Since $|z_2| \geq |z_1|$, we have $|\tau| \geq 1$.
- Since $|z_2 \pm z_1| \geq |z_2|$, we have $|\tau \pm 1| \geq |\tau|$, that is $|\Re\tau| \leq \frac{1}{2}$.
- So $\tau \in D = \{\tau \in \mathbb{H} : \ |\Re\tau| \leq \frac{1}{2}, \ |\tau| \geq 1\}$: **the usual fundamental domain**.
- Conversely, for $\tau \in D$, $(1, \tau)$ is a shortest basis of $\mathbb{Z} + \mathbb{Z}\tau$: $|m + n\tau|^2 \geq m^2 - |mn| + n^2 \geq 1$ for $(m, n) \neq (0, 0)$.

When the algorithm is ambiguous: the boundary of D



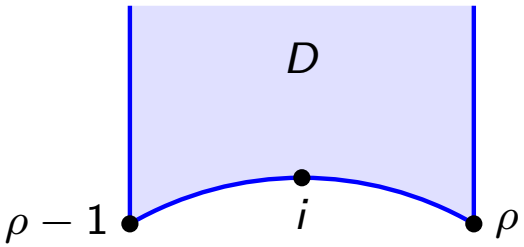
- **Interior of** D: the shortest basis is unique (up to $(z_1, z_2) \mapsto (-z_1, -z_2)$, which does not change τ).

When the algorithm is ambiguous: the boundary of D



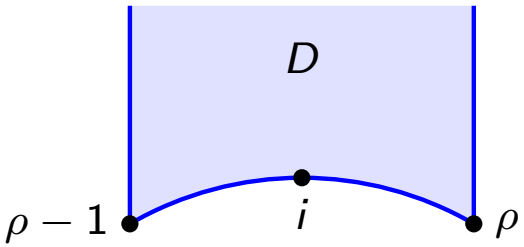
- **Interior of** D: the shortest basis is unique (up to $(z_1, z_2) \mapsto (-z_1, -z_2)$, which does not change τ).
- $|\Re\tau| = \frac{1}{2}$: $|z_2| = |z_2 \mp z_1|$, there are two choices of z_2, giving τ and $\tau \mp 1$: the vertical sides are identified by T.

When the algorithm is ambiguous: the boundary of D



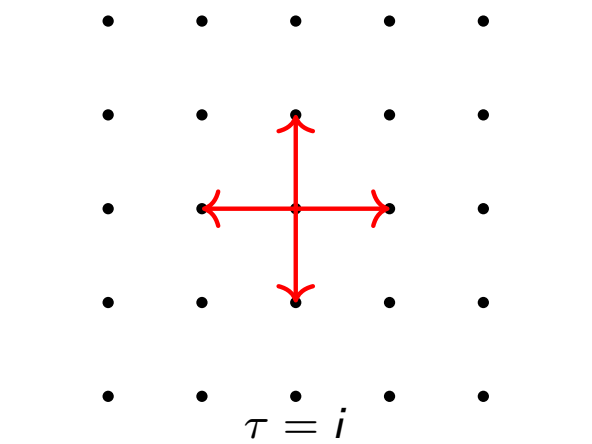
- **Interior of** D: the shortest basis is unique (up to $(z_1, z_2) \mapsto (-z_1, -z_2)$, which does not change τ).
- $|\Re\tau| = \frac{1}{2}$: $|z_2| = |z_2 \mp z_1|$, there are two choices of z_2, giving τ and $\tau \mp 1$: the vertical sides are identified by T.
- $|\tau| = 1$: z_1 and z_2 are both shortest. We can exchange them: $(z_2, -z_1)$ gives $-\frac{1}{\tau}$: the arc is folded by S.

When the algorithm is ambiguous: the boundary of D

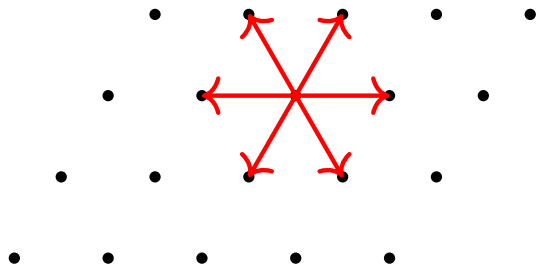


- **Interior of** D: the shortest basis is unique (up to $(z_1, z_2) \mapsto (-z_1, -z_2)$, which does not change τ).
- $|\Re\tau| = \frac{1}{2}$: $|z_2| = |z_2 \mp z_1|$, there are two choices of z_2, giving τ and $\tau \mp 1$: the vertical sides are identified by T.
- $|\tau| = 1$: z_1 and z_2 are both shortest. We can exchange them: $(z_2, -z_1)$ gives $-\frac{1}{\tau}$: the arc is folded by S.
- Singular points: the **corners**, where both ambiguities occur.

The square and the hexagonal lattices



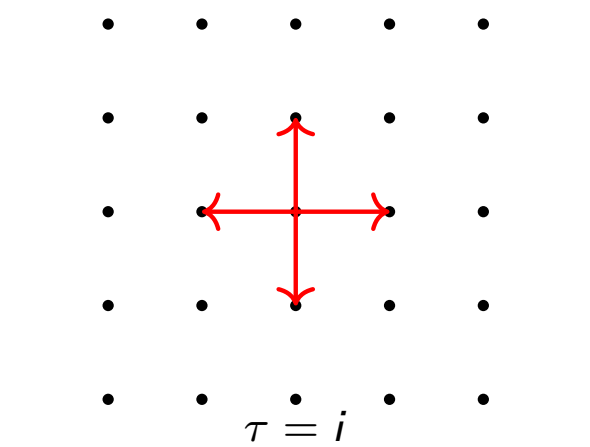
$\tau = i$



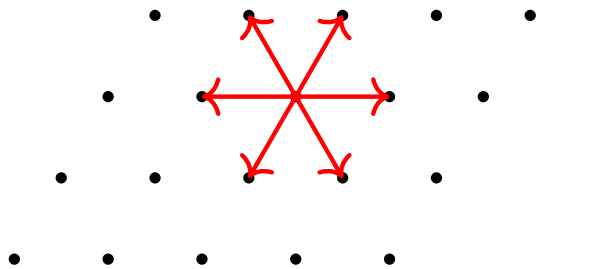
$\tau = \rho = e^{i\pi/3}$

- **Square lattice** $\mathbb{Z} + i\mathbb{Z}$: $z_2 = iz_1$, so $|\tau| = 1$, $\Re\tau = 0$. **Four** shortest vectors $\pm z_1, \pm iz_1$: invariant by the rotation of angle $\frac{\pi}{2}$.

The square and the hexagonal lattices



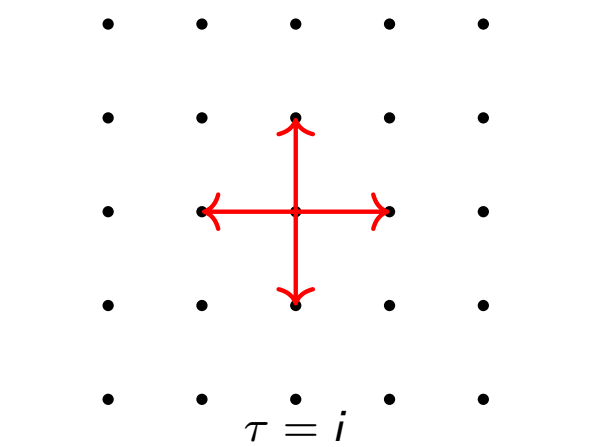
$\tau = i$



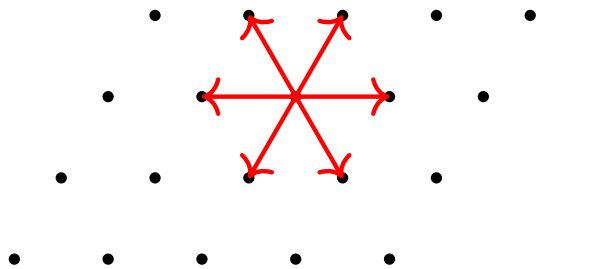
$\tau = \rho = e^{i\pi/3}$

- **Square lattice** $\mathbb{Z} + i\mathbb{Z}$: $z_2 = iz_1$, so $|\tau| = 1$, $\Re\tau = 0$. **Four** shortest vectors $\pm z_1, \pm iz_1$: invariant by the rotation of angle $\frac{\pi}{2}$.
- **Hexagonal lattice** $\mathbb{Z} + \rho\mathbb{Z}$: $|z_1| = |z_2| = |z_2 - z_1|$. **Six** shortest vectors: iinvariant by the rotation of angle $\frac{\pi}{3}$.

The square and the hexagonal lattices



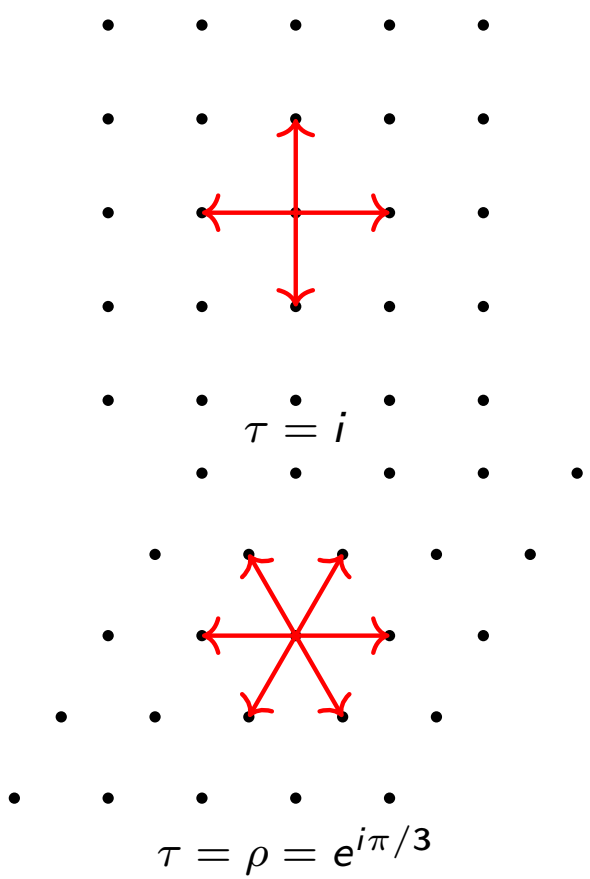
$\tau = i$



$\tau = \rho = e^{i\pi/3}$

- **Square lattice** $\mathbb{Z} + i\mathbb{Z}$: $z_2 = iz_1$, so $|\tau| = 1$, $\Re\tau = 0$. **Four** shortest vectors $\pm z_1, \pm iz_1$: invariant by the rotation of angle $\frac{\pi}{2}$.
- **Hexagonal lattice** $\mathbb{Z} + \rho\mathbb{Z}$: $|z_1| = |z_2| = |z_2 - z_1|$. **Six** shortest vectors: iinvariant by the rotation of angle $\frac{\pi}{3}$.
- A generic lattice is only invariant by the rotation of angle π (that is $-\mathrm{Id}$, trivial in PSL).

The square and the hexagonal lattices



- **Square lattice** $\mathbb{Z} + i\mathbb{Z}$: $z_2 = iz_1$, so $|\tau| = 1$, $\Re\tau = 0$. **Four** shortest vectors $\pm z_1, \pm iz_1$: invariant by the rotation of angle $\frac{\pi}{2}$.
- **Hexagonal lattice** $\mathbb{Z} + \rho\mathbb{Z}$: $|z_1| = |z_2| = |z_2 - z_1|$. **Six** shortest vectors: iinvariant by the rotation of angle $\frac{\pi}{3}$.
- A generic lattice is only invariant by the rotation of angle π (that is $-\mathrm{Id}$, trivial in PSL).
- **The singular points are the lattices with extra rotational symmetry**: the square lattice (order $4/2 = 2$, cone angle π) and the hexagonal lattice (order $6/2 = 3$, cone angle $\frac{2\pi}{3}$).

The cusp, the flows, and Euclid

- **The cusp**: $\Im\tau = \dfrac{\mathrm{Area}(\Lambda)}{|z_1|^2}$. Going up in the cusp means that the lattice has a *very short vector* (Mahler: a set of lattices of covolume 1 is relatively compact iff its shortest vectors are bounded from below).

The cusp, the flows, and Euclid

- **The cusp**: $\Im\tau = \dfrac{\mathrm{Area}(\Lambda)}{|z_1|^2}$. Going up in the cusp means that the lattice has a *very short vector* (Mahler: a set of lattices of covolume 1 is relatively compact iff its shortest vectors are bounded from below).
- **Unit tangent bundle**: the lattices of covolume 1 form $\mathrm{SL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{Z})$, which is T^1 of the modular surface (via $g \mapsto g^{-1}$). The geodesic flow acts by the diagonal matrices, which stretch the lattice in one direction and shrink it in the other; the horocycle flows are shears.

The cusp, the flows, and Euclid

- **The cusp**: $\Im\tau = \dfrac{\mathrm{Area}(\Lambda)}{|z_1|^2}$. Going up in the cusp means that the lattice has a *very short vector* (Mahler: a set of lattices of covolume 1 is relatively compact iff its shortest vectors are bounded from below).
- **Unit tangent bundle**: the lattices of covolume 1 form $\mathrm{SL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{Z})$, which is T^1 of the modular surface (via $g \mapsto g^{-1}$). The geodesic flow acts by the diagonal matrices, which stretch the lattice in one direction and shrink it in the other; the horocycle flows are shears.
- **Euclid**: reducing a basis, $z_2 \mapsto z_2 - kz_1$ and exchanging z_1, z_2 when $|z_2| < |z_1|$, uses only the generators T and S of $\mathrm{SL}(2,\mathbb{Z})$: the same operations as Euclid's algorithm (subtract, invert) and the continued fractions of Lecture 3.

The cusp, the flows, and Euclid

- **The cusp**: $\Im\tau = \dfrac{\mathrm{Area}(\Lambda)}{|z_1|^2}$. Going up in the cusp means that the lattice has a *very short vector* (Mahler: a set of lattices of covolume 1 is relatively compact iff its shortest vectors are bounded from below).
- **Unit tangent bundle**: the lattices of covolume 1 form $\mathrm{SL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{Z})$, which is T^1 of the modular surface (via $g \mapsto g^{-1}$). The geodesic flow acts by the diagonal matrices, which stretch the lattice in one direction and shrink it in the other; the horocycle flows are shears.
- **Euclid**: reducing a basis, $z_2 \mapsto z_2 - kz_1$ and exchanging z_1, z_2 when $|z_2| < |z_1|$, uses only the generators T and S of $\mathrm{SL}(2,\mathbb{Z})$: the same operations as Euclid's algorithm (subtract, invert) and the continued fractions of Lecture 3.
- Three views of the same space: the quotient $\mathrm{SL}(2,\mathbb{Z})\backslash\mathbb{H}$, the space of lattices up to similarity, and (through the pseudosphere) a surface with one cusp.

Exercises on lattices

- **Exercise 1**. Show that $\Im\tau = \mathrm{Area}(\Lambda)/|z_1|^2$, where $\mathrm{Area}(\Lambda) = |\det(z_1, z_2)|$.
- **Exercise 2**. Compute the distance between the square lattice and the hexagonal lattice in the space of lattices.
- **Exercise 3**. Show that the rotations preserving the square lattice form a group of order 4, and those preserving the hexagonal lattice a group of order 6.
- **Exercise 4** ($\star$). Show that the reduction algorithm (replace z_2 by $z_2 - kz_1$, k the nearest integer to $\Re\tau$; exchange if $|z_2| < |z_1|$) terminates, and finds a shortest basis.

A long history (1): quadratic forms and elliptic functions

- **1773–1801 (Lagrange, Gauss)**: reduction of positive definite binary quadratic forms $ax^2 + bxy + cy^2$. A form is a lattice with its metric: $|mz_1 + nz_2|^2$ has $a = |z_1|^2$, $b = 2|z_1|^2\Re\tau$, $c = |z_1|^2|\tau|^2$. The *reduced forms* $|b| \leq a \leq c$ are exactly $\tau \in D$: our fundamental domain and our shortest basis are already in the *Disquisitiones Arithmeticae*.

A long history (1): quadratic forms and elliptic functions

- **1773–1801 (Lagrange, Gauss)**: reduction of positive definite binary quadratic forms $ax^2 + bxy + cy^2$. A form is a lattice with its metric: $|mz_1 + nz_2|^2$ has $a = |z_1|^2$, $b = 2|z_1|^2\Re\tau$, $c = |z_1|^2|\tau|^2$. The *reduced forms* $|b| \leq a \leq c$ are exactly $\tau \in D$: our fundamental domain and our shortest basis are already in the *Disquisitiones Arithmeticae*.
- **1820s–1860s (Abel, Jacobi, Weierstrass)**: *elliptic functions* (inverses of elliptic integrals, as for the period of a pendulum) are the meromorphic functions on $\mathbb{C}$ with periods in a lattice Λ. Weierstrass' function $\wp_\Lambda(z) = \frac{1}{z^2} + \sum_{\omega\in\Lambda\setminus\{0\}}\left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2}\right)$ satisfies $\wp'^2 = 4\wp^3 - g_2\wp - g_3$, with $g_2 = 60\sum\omega^{-4}$ and $g_3 = 140\sum\omega^{-6}$.

A long history (1): quadratic forms and elliptic functions

- **1773–1801 (Lagrange, Gauss)**: reduction of positive definite binary quadratic forms $ax^2 + bxy + cy^2$. A form is a lattice with its metric: $|mz_1 + nz_2|^2$ has $a = |z_1|^2$, $b = 2|z_1|^2\Re\tau$, $c = |z_1|^2|\tau|^2$. The *reduced forms* $|b| \leq a \leq c$ are exactly $\tau \in D$: our fundamental domain and our shortest basis are already in the *Disquisitiones Arithmeticae*.
- **1820s–1860s (Abel, Jacobi, Weierstrass)**: *elliptic functions* (inverses of elliptic integrals, as for the period of a pendulum) are the meromorphic functions on $\mathbb{C}$ with periods in a lattice Λ. Weierstrass' function $\wp_\Lambda(z) = \frac{1}{z^2} + \sum_{\omega\in\Lambda\setminus\{0\}}\left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2}\right)$ satisfies $\wp'^2 = 4\wp^3 - g_2\wp - g_3$, with $g_2 = 60\sum\omega^{-4}$ and $g_3 = 140\sum\omega^{-6}$.
- So the torus $\mathbb{C}/\Lambda$ is the elliptic curve $y^2 = 4x^3 - g_2x - g_3$, and $\mathbb{C}/\Lambda \simeq \mathbb{C}/\Lambda'$ if and only if $\Lambda' = c\Lambda$. **The points of the modular surface are the elliptic curves**, up to isomorphism: the same equivalence relation as in our theorem.

A long history (2): modular functions, moduli spaces, dynamics

- **1870s (Dedekind, Klein)**: the j-invariant $j = 1728 \dfrac{g_2^3}{g_2^3 - 27 g_3^2}$ is invariant by $\mathrm{SL}(2,\mathbb{Z})$ and identifies the modular surface with $\mathbb{C}$; g_2, g_3 are *modular forms*. Poincaré (1880s): Fuchsian groups and the hyperbolic half-plane.

A long history (2): modular functions, moduli spaces, dynamics

- **1870s (Dedekind, Klein)**: the j-invariant $j = 1728 \dfrac{g_2^3}{g_2^3 - 27g_3^2}$ is invariant by $\mathrm{SL}(2,\mathbb{Z})$ and identifies the modular surface with $\mathbb{C}$; g_2, g_3 are *modular forms*. Poincaré (1880s): Fuchsian groups and the hyperbolic half-plane.
- **Algebraic geometry**: the modular surface is the *moduli space of elliptic curves*. Riemann (1857), Siegel (1935): principally polarized abelian varieties of dimension g are parametrized by $\mathrm{Sp}(2g,\mathbb{Z})\backslash\mathbb{H}_g$ (the Siegel half-space); $g = 1$ is our case.

A long history (2): modular functions, moduli spaces, dynamics

- **1870s (Dedekind, Klein)**: the j-invariant $j = 1728 \dfrac{g_2^3}{g_2^3 - 27g_3^2}$ is invariant by $\mathrm{SL}(2,\mathbb{Z})$ and identifies the modular surface with $\mathbb{C}$; g_2, g_3 are *modular forms*. Poincaré (1880s): Fuchsian groups and the hyperbolic half-plane.
- **Algebraic geometry**: the modular surface is the *moduli space of elliptic curves*. Riemann (1857), Siegel (1935): principally polarized abelian varieties of dimension g are parametrized by $\mathrm{Sp}(2g,\mathbb{Z})\backslash\mathbb{H}_g$ (the Siegel half-space); $g = 1$ is our case.
- **Number theory**: modular forms (Ramanujan, Hecke); the modularity of elliptic curves over $\mathbb{Q}$ (Wiles 1995) and Fermat's Last Theorem.

A long history (2): modular functions, moduli spaces, dynamics

- **1870s (Dedekind, Klein)**: the j-invariant $j = 1728 \dfrac{g_2^3}{g_2^3 - 27g_3^2}$ is invariant by $\mathrm{SL}(2,\mathbb{Z})$ and identifies the modular surface with $\mathbb{C}$; g_2, g_3 are *modular forms*. Poincaré (1880s): Fuchsian groups and the hyperbolic half-plane.
- **Algebraic geometry**: the modular surface is the *moduli space of elliptic curves*. Riemann (1857), Siegel (1935): principally polarized abelian varieties of dimension g are parametrized by $\mathrm{Sp}(2g,\mathbb{Z})\backslash\mathbb{H}_g$ (the Siegel half-space); $g = 1$ is our case.
- **Number theory**: modular forms (Ramanujan, Hecke); the modularity of elliptic curves over $\mathbb{Q}$ (Wiles 1995) and Fermat's Last Theorem.
- **Dynamics**: Hadamard (1898); Artin (1924) codes the geodesics of the modular surface by continued fractions; Hedlund (1935): ergodicity; Series, Arnoux: the return map to a cross-section is a version of the natural extension of the Gauss map (Lecture 3).

The Haar measure: definition

- **Definition**: on a locally compact group G, a (left) **Haar measure** is a nonzero regular Borel measure μ, finite on compact sets, invariant by left translations: $\mu(gA) = \mu(A)$ for all $g \in G$ and Borel A (equivalently $\int f(gx)\, d\mu = \int f\, d\mu$).

The Haar measure: definition

- **Definition**: on a locally compact group G, a (left) **Haar measure** is a nonzero regular Borel measure μ, finite on compact sets, invariant by left translations: $\mu(gA) = \mu(A)$ for all $g \in G$ and Borel A (equivalently $\int f(gx)\, d\mu = \int f\, d\mu$).
- Right Haar measures are defined similarly. They may differ (the *modular function* $\Delta : G \to \mathbb{R}_{>0}$ measures the difference); if they coincide, G is **unimodular**.

The Haar measure : examples

- The basic example: counting measure on a discrete group ($\mathbb{Z}$, finite groups) : if a group is finite, and a measure on it is translation invariant, all points have same mass.

The Haar measure : examples

- The basic example: counting measure on a discrete group ($\mathbb{Z}$, finite groups) : if a group is finite, and a measure on it is translation invariant, all points have same mass.
- The fundamental example: Lebesgue measure (the usual length!) on the line $\mathbb{R}$ and on the circle $\mathbb{R}/\mathbb{Z}$.

The Haar measure : examples

- The basic example: counting measure on a discrete group ($\mathbb{Z}$, finite groups) : if a group is finite, and a measure on it is translation invariant, all points have same mass.
- The fundamental example: Lebesgue measure (the usual length!) on the line $\mathbb{R}$ and on the circle $\mathbb{R}/\mathbb{Z}$.
- More generally : Lebesgue measure on $(\mathbb{R}^n, +)$ and $(\mathbb{R}^n, +)/\mathbb{Z}^n$.

The Haar measure : examples

- The basic example: counting measure on a discrete group ($\mathbb{Z}$, finite groups) : if a group is finite, and a measure on it is translation invariant, all points have same mass.
- The fundamental example: Lebesgue measure (the usual length!) on the line $\mathbb{R}$ and on the circle $\mathbb{R}/\mathbb{Z}$.
- More generally : Lebesgue measure on $(\mathbb{R}^n, +)$ and $(\mathbb{R}^n, +)/\mathbb{Z}^n$.
- More exotic : The measure $\frac{dx}{x}$ on $(\mathbb{R}_{>0}, \times)$, which is invariant by multiplication (related to the logarithm).

The Haar measure : examples

- The basic example: counting measure on a discrete group ($\mathbb{Z}$, finite groups) : if a group is finite, and a measure on it is translation invariant, all points have same mass.
- The fundamental example: Lebesgue measure (the usual length!) on the line $\mathbb{R}$ and on the circle $\mathbb{R}/\mathbb{Z}$.
- More generally : Lebesgue measure on $(\mathbb{R}^n, +)$ and $(\mathbb{R}^n, +)/\mathbb{Z}^n$.
- More exotic : The measure $\frac{dx}{x}$ on $(\mathbb{R}_{>0}, \times)$, which is invariant by multiplication (related to the logarithm).
- Or on the group of invertible matrices: $\frac{dX}{|\det X|^n}$ on $\mathrm{GL}(n, \mathbb{R})$.

The Haar measure : examples

- The basic example: counting measure on a discrete group ($\mathbb{Z}$, finite groups) : if a group is finite, and a measure on it is translation invariant, all points have same mass.
- The fundamental example: Lebesgue measure (the usual length!) on the line $\mathbb{R}$ and on the circle $\mathbb{R}/\mathbb{Z}$.
- More generally : Lebesgue measure on $(\mathbb{R}^n, +)$ and $(\mathbb{R}^n, +)/\mathbb{Z}^n$.
- More exotic : The measure $\frac{dx}{x}$ on $(\mathbb{R}_{>0}, \times)$, which is invariant by multiplication (related to the logarithm).
- Or on the group of invertible matrices: $\frac{dX}{|\det X|^n}$ on $\mathrm{GL}(n, \mathbb{R})$.
- A simple not unimodular example: the affine group $\{x \mapsto ax + b\}$, with left Haar measure $\frac{da\,db}{a^2}$ and right Haar measure $\frac{da\,db}{a}$ (exercize!).

The Haar measure : The fundamental property

Theorem

*(Haar, von Neumann): a Haar measure always exists, and it is **unique up to a positive constant**.*

The Haar measure : The fundamental property

Theorem

(Haar, von Neumann): a Haar measure always exists, and it is ***unique up to a positive constant****.*

- **Compact case**: the Haar measure is finite; after normalization, it is the unique left-invariant **probability measure**, and it is also right invariant. Examples: the circle, $\mathrm{SO}(n)$, finite groups (uniform measure).

The Haar measure : The fundamental property

Theorem

*(Haar, von Neumann): a Haar measure always exists, and it is **unique up to a positive constant**.*

- **Compact case**: the Haar measure is finite; after normalization, it is the unique left-invariant **probability measure**, and it is also right invariant. Examples: the circle, $\mathrm{SO}(n)$, finite groups (uniform measure).
- **Quotients**: for G unimodular and Γ discrete, $\Gamma\backslash G$ carries a right-invariant measure, finite if Γ is a lattice (a probability after normalization).

The Haar measure : $\mathrm{SL}(2,\mathbb{R})$

- In the case of $G = \mathrm{SL}(2,\mathbb{R})$: the Haar measure is the Liouville measure $\frac{dx\,dy}{y^2}\,d\theta$ on $T^1\mathbb{H}$ (area $\times$ angle), left invariant since G acts by isometries.

The Haar measure : $\mathrm{SL}(2,\mathbb{R})$

- In the case of $G = \mathrm{SL}(2,\mathbb{R})$: the Haar measure is the Liouville measure $\frac{dx\,dy}{y^2}\,d\theta$ on $T^1\mathbb{H}$ (area $\times$ angle), left invariant since G acts by isometries.
- It has different expressions, depending on the coordinate system you define.

The Haar measure : $\mathrm{SL}(2,\mathbb{R})$

- In the case of $G = \mathrm{SL}(2,\mathbb{R})$: the Haar measure is the Liouville measure $\frac{dx\,dy}{y^2}\,d\theta$ on $T^1\mathbb{H}$ (area $\times$ angle), left invariant since G acts by isometries.
- It has different expressions, depending on the coordinate system you define.
- G is unimodular: $\Delta(h_s^+) = \Delta(h_{se^{-t}}^+) \to 1$ (conjugation, fundamental property), and the horocycle subgroups generate G. So the flows preserve the measure.

Hyperbolicity and the strategy

- **Goal**: the Liouville measure μ on $T^1(\Gamma\backslash\mathbb{H})$ is ergodic for the geodesic flow.

Hyperbolicity and the strategy

- **Goal**: the Liouville measure μ on $T^1(\Gamma\backslash\mathbb{H})$ is ergodic for the geodesic flow.
- **Hyperbolicity** (the fundamental property): $x\, h_s^+\, g_t = x\, g_t\, h_{se^{-t}}^+$ and $x\, h_s^-\, g_t = x\, g_t\, h_{se^t}^-$. The geodesic flow **contracts** the positive horocycles in forward time (two points of the same one are *positively asymptotic*), and **contracts** the negative horocycles in backward time.

Hyperbolicity and the strategy

- **Goal**: the Liouville measure μ on $T^1(\Gamma\backslash\mathbb{H})$ is ergodic for the geodesic flow.
- **Hyperbolicity** (the fundamental property): $x\, h_s^+\, g_t = x\, g_t\, h_{se^{-t}}^+$ and $x\, h_s^-\, g_t = x\, g_t\, h_{se^t}^-$. The geodesic flow **contracts** the positive horocycles in forward time (two points of the same one are *positively asymptotic*), and **contracts** the negative horocycles in backward time.
- **Hopf's idea**: (i) by Birkhoff, the time averages f^+ (forward) and f^- (backward) exist almost everywhere and are equal; (ii) f^+ is constant along positive horocycles, f^- along negative horocycles (for f continuous); (iii) these two foliations, together with the flow, "fill" the space: so the average is a.e. constant.

Hyperbolicity and the strategy

- **Goal**: the Liouville measure μ on $T^1(\Gamma\backslash\mathbb{H})$ is ergodic for the geodesic flow.
- **Hyperbolicity** (the fundamental property): $x\, h_s^+\, g_t = x\, g_t\, h_{se^{-t}}^+$ and $x\, h_s^-\, g_t = x\, g_t\, h_{se^t}^-$. The geodesic flow **contracts** the positive horocycles in forward time (two points of the same one are *positively asymptotic*), and **contracts** the negative horocycles in backward time.
- **Hopf's idea**: (i) by Birkhoff, the time averages f^+ (forward) and f^- (backward) exist almost everywhere and are equal; (ii) f^+ is constant along positive horocycles, f^- along negative horocycles (for f continuous); (iii) these two foliations, together with the flow, "fill" the space: so the average is a.e. constant.
- No spectral theory, and no use of the group structure beyond the contraction: this argument works for *all* hyperbolic systems.

Hopf's argument: a lemma

- **Lemma**: let B_+, B_- be two measurable subsets of $T^1(\Gamma\backslash\mathbb{H})$, invariant by the geodesic flow, with $\mu(B_+ \Delta B_-) = 0$, and such that
  - if $x \in B_+$ and the orbit of y is *positively asymptotic* to the orbit of x, then $y \in B_+$;
  - if $x \in B_-$ and the orbit of y is *negatively asymptotic* to the orbit of x, then $y \in B_-$.

Hopf's argument: a lemma

- **Lemma**: let B_+, B_- be two measurable subsets of $T^1(\Gamma\backslash\mathbb{H})$, invariant by the geodesic flow, with $\mu(B_+ \Delta B_-) = 0$, and such that
  - if $x \in B_+$ and the orbit of y is *positively asymptotic* to the orbit of x, then $y \in B_+$;
  - if $x \in B_-$ and the orbit of y is *negatively asymptotic* to the orbit of x, then $y \in B_-$.
- Then B_+ or its complement has measure 0.

Proof of the lemma (1/2): coordinates

- Lift the sets to the unit tangent bundle of $\mathbb{H}$: the projection is countable-to-one, so a set of measure 0 lifts to a set of measure 0.

Proof of the lemma (1/2): coordinates

- Lift the sets to the unit tangent bundle of $\mathbb{H}$: the projection is countable-to-one, so a set of measure 0 lifts to a set of measure 0.
- **Coordinates**: to a unit vector we associate its geodesic, completely determined by its two extremities (ξ_+, ξ_-) in $\overline{\mathbb{R}} = \partial\mathbb{H}$. Fixing a base point on the geodesic (for instance the point closest to i), we parametrize $T^1\mathbb{H}$ by $\overline{\mathbb{R}} \times \overline{\mathbb{R}} \times \mathbb{R}$ (up to the diagonal, of measure 0).

Proof of the lemma (1/2): coordinates

- Lift the sets to the unit tangent bundle of $\mathbb{H}$: the projection is countable-to-one, so a set of measure 0 lifts to a set of measure 0.
- **Coordinates**: to a unit vector we associate its geodesic, completely determined by its two extremities (ξ_+, ξ_-) in $\overline{\mathbb{R}} = \partial\mathbb{H}$. Fixing a base point on the geodesic (for instance the point closest to i), we parametrize $T^1\mathbb{H}$ by $\overline{\mathbb{R}} \times \overline{\mathbb{R}} \times \mathbb{R}$ (up to the diagonal, of measure 0).
- The lifted set $\tilde{B}_+$ is a union of geodesics: $\tilde{B}_+ = C \times \mathbb{R}$ with $C \subset \overline{\mathbb{R}} \times \overline{\mathbb{R}}$. By the property of B_+, two geodesics with the same *positive* extremity are both in or both out: $\tilde{B}_+ = b_+ \times \overline{\mathbb{R}} \times \mathbb{R}$.

Proof of the lemma (1/2): coordinates

- Lift the sets to the unit tangent bundle of $\mathbb{H}$: the projection is countable-to-one, so a set of measure 0 lifts to a set of measure 0.
- **Coordinates**: to a unit vector we associate its geodesic, completely determined by its two extremities (ξ_+, ξ_-) in $\overline{\mathbb{R}} = \partial\mathbb{H}$. Fixing a base point on the geodesic (for instance the point closest to i), we parametrize $T^1\mathbb{H}$ by $\overline{\mathbb{R}} \times \overline{\mathbb{R}} \times \mathbb{R}$ (up to the diagonal, of measure 0).
- The lifted set $\tilde{B}_+$ is a union of geodesics: $\tilde{B}_+ = C \times \mathbb{R}$ with $C \subset \overline{\mathbb{R}} \times \overline{\mathbb{R}}$. By the property of B_+, two geodesics with the same *positive* extremity are both in or both out: $\tilde{B}_+ = b_+ \times \overline{\mathbb{R}} \times \mathbb{R}$.
- For the same reason $\tilde{B}_- = \overline{\mathbb{R}} \times b_- \times \mathbb{R}$.

Proof of the lemma (2/2): conclusion

- The set $\tilde{B}_- \setminus \tilde{B}_+ = c_+ \times b_- \times \mathbb{R}$, where c_+ is the complement of b_+ in $\overline{\mathbb{R}}$, has measure 0 by hypothesis ($\mu(B_+ \Delta B_-) = 0$).

Proof of the lemma (2/2): conclusion

- The set $\tilde{B}_- \setminus \tilde{B}_+ = c_+ \times b_- \times \mathbb{R}$, where c_+ is the complement of b_+ in $\overline{\mathbb{R}}$, has measure 0 by hypothesis ($\mu(B_+ \Delta B_-) = 0$).
- In these coordinates the Liouville measure is equivalent to the product of Lebesgue measures. So either b_- has measure 0, or c_+ has measure 0.

Proof of the lemma (2/2): conclusion

- The set $\tilde{B}_- \setminus \tilde{B}_+ = c_+ \times b_- \times \mathbb{R}$, where c_+ is the complement of b_+ in $\overline{\mathbb{R}}$, has measure 0 by hypothesis ($\mu(B_+ \Delta B_-) = 0$).
- In these coordinates the Liouville measure is equivalent to the product of Lebesgue measures. So either b_- has measure 0, or c_+ has measure 0.
- If c_+ has measure 0, then $\tilde{B}_+$ has full measure: so does B_+.

Proof of the lemma (2/2): conclusion

- The set $\tilde{B}_- \setminus \tilde{B}_+ = c_+ \times b_- \times \mathbb{R}$, where c_+ is the complement of b_+ in $\overline{\mathbb{R}}$, has measure 0 by hypothesis ($\mu(B_+ \Delta B_-) = 0$).
- In these coordinates the Liouville measure is equivalent to the product of Lebesgue measures. So either b_- has measure 0, or c_+ has measure 0.
- If c_+ has measure 0, then $\tilde{B}_+$ has full measure: so does B_+.
- If b_- has measure 0, then B_- has measure 0, and so does B_+ (the symmetric difference is null). ■

Hopf's theorem

- **Theorem (Hopf, 1936)**: the geodesic flow on a surface of constant negative curvature and finite volume is ergodic.

Hopf's theorem

- **Theorem (Hopf, 1936)**: the geodesic flow on a surface of constant negative curvature and finite volume is ergodic.
- **Proof**: a consequence of Birkhoff's ergodic theorem. Let f be integrable; we must show that its Birkhoff integrals converge a.e. to a constant. By continuity of $f \mapsto f^+$, it is enough to take f continuous with compact support.

Hopf's theorem

- **Theorem (Hopf, 1936)**: the geodesic flow on a surface of constant negative curvature and finite volume is ergodic.
- **Proof**: a consequence of Birkhoff's ergodic theorem. Let f be integrable; we must show that its Birkhoff integrals converge a.e. to a constant. By continuity of $f \mapsto f^+$, it is enough to take f continuous with compact support.
- By Birkhoff's theorem, the limits f^+ (forward) and f^- (backward) exist a.e. and are a.e. equal.

Hopf's theorem

- **Theorem (Hopf, 1936)**: the geodesic flow on a surface of constant negative curvature and finite volume is ergodic.
- **Proof**: a consequence of Birkhoff's ergodic theorem. Let f be integrable; we must show that its Birkhoff integrals converge a.e. to a constant. By continuity of $f \mapsto f^+$, it is enough to take f continuous with compact support.
- By Birkhoff's theorem, the limits f^+ (forward) and f^- (backward) exist a.e. and are a.e. equal.
- Let $B_+ = \{f^+ > c\}$ and $B_- = \{f^- > c\}$. They are invariant by the geodesic flow, they differ by a null set, and, by continuity of f, if two points are positively (resp. negatively) asymptotic, they are both in or both out of B_+ (resp. B_-).

Hopf's theorem

- **Theorem (Hopf, 1936)**: the geodesic flow on a surface of constant negative curvature and finite volume is ergodic.
- **Proof**: a consequence of Birkhoff's ergodic theorem. Let f be integrable; we must show that its Birkhoff integrals converge a.e. to a constant. By continuity of $f \mapsto f^+$, it is enough to take f continuous with compact support.
- By Birkhoff's theorem, the limits f^+ (forward) and f^- (backward) exist a.e. and are a.e. equal.
- Let $B_+ = \{f^+ > c\}$ and $B_- = \{f^- > c\}$. They are invariant by the geodesic flow, they differ by a null set, and, by continuity of f, if two points are positively (resp. negatively) asymptotic, they are both in or both out of B_+ (resp. B_-).
- By the lemma, $\{f^+ > c\}$ is null or conull for every c: f^+ is a.e. constant. ■

Much more can be done

- The same ideas (the commutation relation $h_s^+ g_t = g_t h_{se^{-t}}^+$, hyperbolicity) give much more:
  - weak and strong **mixing** of the geodesic flow (Hedlund, Hopf 1939);
  - **ergodicity** (Hedlund 1936) and **mixing** (Marcus 1978) of the **horocycle flow**;
  - in the compact case, **minimality** (Hedlund 1936) and **unique ergodicity** (Furstenberg 1973) of the horocycle flow.

Much more can be done

- The same ideas (the commutation relation $h_s^+ g_t = g_t h_{se^{-t}}^+$, hyperbolicity) give much more:
  - weak and strong **mixing** of the geodesic flow (Hedlund, Hopf 1939);
  - **ergodicity** (Hedlund 1936) and **mixing** (Marcus 1978) of the **horocycle flow**;
  - in the compact case, **minimality** (Hedlund 1936) and **unique ergodicity** (Furstenberg 1973) of the horocycle flow.
- All this extends to the geodesic flow of any compact manifold of negative curvature (Anosov systems).

Much more can be done

- The same ideas (the commutation relation $h_s^+ g_t = g_t h_{se^{-t}}^+$, hyperbolicity) give much more:
  - weak and strong **mixing** of the geodesic flow (Hedlund, Hopf 1939);
  - **ergodicity** (Hedlund 1936) and **mixing** (Marcus 1978) of the **horocycle flow**;
  - in the compact case, **minimality** (Hedlund 1936) and **unique ergodicity** (Furstenberg 1973) of the horocycle flow.
- All this extends to the geodesic flow of any compact manifold of negative curvature (Anosov systems).
- As we see in the next section

Overview

- We prove the main properties (ergodicity, mixing) of the geodesic and horocycle flows on the modular surface, and sketch the proofs of other properties.

Overview

- We prove the main properties (ergodicity, mixing) of the geodesic and horocycle flows on the modular surface, and sketch the proofs of other properties.
- These proofs are important: they have been much generalized in the theory of **hyperbolic systems**.

Overview

- We prove the main properties (ergodicity, mixing) of the geodesic and horocycle flows on the modular surface, and sketch the proofs of other properties.
- These proofs are important: they have been much generalized in the theory of **hyperbolic systems**.
- We give two proofs of ergodicity: they seem different, but the underlying logic is the same.

Overview

- We prove the main properties (ergodicity, mixing) of the geodesic and horocycle flows on the modular surface, and sketch the proofs of other properties.
- These proofs are important: they have been much generalized in the theory of **hyperbolic systems**.
- We give two proofs of ergodicity: they seem different, but the underlying logic is the same.
- All the proofs rely on the **commutation relations**.

The general structure

1. Ergodicity (Hedlund 1935, Hopf 1936) and weak mixing of the geodesic flow.

The general structure

1. Ergodicity (Hedlund 1935, Hopf 1936) and weak mixing of the geodesic flow.
2. Use (1) to prove ergodicity of the horocycle flow (Hedlund 1936).

The general structure

1. Ergodicity (Hedlund 1935, Hopf 1936) and weak mixing of the geodesic flow.
2. Use (1) to prove ergodicity of the horocycle flow (Hedlund 1936).
3. Use (2) to prove strong mixing of the geodesic flow (Hedlund, Hopf 1939).

The general structure

1. Ergodicity (Hedlund 1935, Hopf 1936) and weak mixing of the geodesic flow.
2. Use (1) to prove ergodicity of the horocycle flow (Hedlund 1936).
3. Use (2) to prove strong mixing of the geodesic flow (Hedlund, Hopf 1939).
4. Use (3) to prove strong mixing of the horocycle flow (Marcus 1978).

The general structure

1. Ergodicity (Hedlund 1935, Hopf 1936) and weak mixing of the geodesic flow.
2. Use (1) to prove ergodicity of the horocycle flow (Hedlund 1936).
3. Use (2) to prove strong mixing of the geodesic flow (Hedlund, Hopf 1939).
4. Use (3) to prove strong mixing of the horocycle flow (Marcus 1978).
5. Use (3) to prove minimality (Hedlund 1936) and unique ergodicity (Furstenberg 1973) of the horocycle flow in the compact case.

The operators

- On $L^2(\mu)$ we define the unitary operators
  $(G_t f)(x) = f(x\, g_t)$, $(H_s f)(x) = f(x\, h_s^+)$,
  $(H_s^- f)(x) = f(x\, h_s^-)$.

The operators

- On $L^2(\mu)$ we define the unitary operators
  $(G_t f)(x) = f(x\, g_t)$, $\quad (H_s f)(x) = f(x\, h_s^+)$,
  $(H_s^- f)(x) = f(x\, h_s^-)$.
- From the fundamental property $h_s^+ g_t = g_t h_{se^{-t}}^+$ we get the **commutation relation**

$$G_{-t}\, H_s\, G_t = H_{se^{-t}},$$

and a symmetric relation $G_{-t} H_s^- G_t = H_{se^t}^-$.

The operators

- On $L^2(\mu)$ we define the unitary operators
  $(G_t f)(x) = f(x\, g_t)$, $\quad (H_s f)(x) = f(x\, h_s^+)$,
  $(H_s^- f)(x) = f(x\, h_s^-)$.
- From the fundamental property $h_s^+ g_t = g_t h_{se^{-t}}^+$ we get the **commutation relation**

$$G_{-t}\, H_s\, G_t = H_{se^{-t}},$$

  and a symmetric relation $G_{-t} H_s^- G_t = H_{se^t}^-$.
- Indeed $(G_{-t} H_s G_t f)(x) = f(x\, g_{-t} h_s^+ g_t)$ and $g_{-t} h_s^+ g_t = h_{se^{-t}}^+$.

Two lemmas

- **Lemma 1**: for every $f \in L^2(\mu)$, $\lim_{s \to 0} H_s f = f$.

Two lemmas

- **Lemma 1**: for every $f \in L^2(\mu)$, $\lim_{s \to 0} H_s f = f$.
- **Proof**: obvious for continuous functions with compact support; these are dense in $L^2(\mu)$, and the H_s are unitary.

Two lemmas

- **Lemma 1**: for every $f \in L^2(\mu)$, $\lim_{s \to 0} H_s f = f$.
- **Proof**: obvious for continuous functions with compact support; these are dense in $L^2(\mu)$, and the H_s are unitary.
- **Lemma 2**: the horocycle flows generate $\mathrm{SL}(2, \mathbb{R})$.

Two lemmas

- **Lemma 1**: for every $f \in L^2(\mu)$, $\lim_{s \to 0} H_s f = f$.
- **Proof**: obvious for continuous functions with compact support; these are dense in $L^2(\mu)$, and the H_s are unitary.
- **Lemma 2**: the horocycle flows generate $\mathrm{SL}(2, \mathbb{R})$.
- **Proof**: left as an exercise. Algebraically: every matrix is a product of unipotent matrices. Geometrically: any tangent vector can be sent to the base vector (i, i) by at most 3 applications of the positive and negative horocycle flows.

Eigenfunctions of the geodesic flow are horocycle invariant

- **Proposition**: let f be an eigenfunction of G_t: there are λ_t with $G_t f = \lambda_t f$ for all t. Then f is invariant by H_s and by H_s^-.

Eigenfunctions of the geodesic flow are horocycle invariant

- **Proposition**: let f be an eigenfunction of G_t: there are λ_t with $G_t f = \lambda_t f$ for all t. Then f is invariant by H_s and by H_s^-.
- **Proof**: we have

$$\lambda_t \, G_{-t}(H_s f - f) = G_{-t} H_s G_t f - f = H_{se^{-t}} f - f \, .$$

Eigenfunctions of the geodesic flow are horocycle invariant

- **Proposition**: let f be an eigenfunction of G_t: there are λ_t with $G_t f = \lambda_t f$ for all t. Then f is invariant by H_s and by H_s^-.
- **Proof**: we have

$$\lambda_t \, G_{-t}(H_s f - f) = G_{-t} H_s G_t f - f = H_{se^{-t}} f - f.$$

- The left side has *constant norm* $\|H_s f - f\|$ (as $|\lambda_t| = 1$ and G_{-t} is unitary).

Eigenfunctions of the geodesic flow are horocycle invariant

- **Proposition**: let f be an eigenfunction of G_t: there are λ_t with $G_t f = \lambda_t f$ for all t. Then f is invariant by H_s and by H_s^-.
- **Proof**: we have

$$\lambda_t \, G_{-t}(H_s f - f) = G_{-t} H_s G_t f - f = H_{se^{-t}} f - f.$$

- The left side has *constant norm* $\|H_s f - f\|$ (as $|\lambda_t| = 1$ and G_{-t} is unitary).
- The right side tends to 0 in $L^2(\mu)$ when $t \to \infty$ (Lemma 1). So $H_s f = f$. The proof is symmetric for H_s^- (take $t \to -\infty$).

Ergodicity and weak mixing

- **Theorem**: the geodesic flow is ergodic, and even weakly mixing, for the Liouville (Haar) measure.

Ergodicity and weak mixing

- **Theorem**: the geodesic flow is ergodic, and even weakly mixing, for the Liouville (Haar) measure.
- **Proof**: every eigenfunction of the geodesic flow is invariant by both horocycle flows, hence (Lemma 2) by the action of $\mathrm{SL}(2,\mathbb{R})$. But this action is transitive: the function is constant.

Ergodicity and weak mixing

- **Theorem**: the geodesic flow is ergodic, and even weakly mixing, for the Liouville (Haar) measure.
- **Proof**: every eigenfunction of the geodesic flow is invariant by both horocycle flows, hence (Lemma 2) by the action of $\mathrm{SL}(2,\mathbb{R})$. But this action is transitive: the function is constant.
- An invariant function is an eigenfunction (eigenvalue 1): this gives ergodicity. A flow with no nonconstant eigenfunction is weakly mixing.

Hopf's argument: a lemma

- **Lemma**: let B_+, B_- be two measurable subsets of $T^1(\Gamma\backslash\mathbb{H})$, invariant by the geodesic flow, with $\mu(B_+ \Delta B_-) = 0$, and such that
  - if $x \in B_+$ and the orbit of y is *positively asymptotic* to the orbit of x, then $y \in B_+$;
  - if $x \in B_-$ and the orbit of y is *negatively asymptotic* to the orbit of x, then $y \in B_-$.

Hopf's argument: a lemma

- **Lemma**: let B_+, B_- be two measurable subsets of $T^1(\Gamma\backslash\mathbb{H})$, invariant by the geodesic flow, with $\mu(B_+ \Delta B_-) = 0$, and such that
  - if $x \in B_+$ and the orbit of y is *positively asymptotic* to the orbit of x, then $y \in B_+$;
  - if $x \in B_-$ and the orbit of y is *negatively asymptotic* to the orbit of x, then $y \in B_-$.
- Then B_+ or its complement has measure 0.

Proof of the lemma (1/2): coordinates

- Lift the sets to the unit tangent bundle of $\mathbb{H}$: the projection is countable-to-one, so a set of measure 0 lifts to a set of measure 0.

Proof of the lemma (1/2): coordinates

- Lift the sets to the unit tangent bundle of $\mathbb{H}$: the projection is countable-to-one, so a set of measure 0 lifts to a set of measure 0.
- **Coordinates**: to a unit vector we associate its geodesic, completely determined by its two extremities (ξ_+, ξ_-) in $\overline{\mathbb{R}} = \partial\mathbb{H}$. Fixing a base point on the geodesic (for instance the point closest to i), we parametrize $T^1\mathbb{H}$ by $\overline{\mathbb{R}} \times \overline{\mathbb{R}} \times \mathbb{R}$ (up to the diagonal, of measure 0).

Proof of the lemma (1/2): coordinates

- Lift the sets to the unit tangent bundle of $\mathbb{H}$: the projection is countable-to-one, so a set of measure 0 lifts to a set of measure 0.
- **Coordinates**: to a unit vector we associate its geodesic, completely determined by its two extremities (ξ_+, ξ_-) in $\overline{\mathbb{R}} = \partial\mathbb{H}$. Fixing a base point on the geodesic (for instance the point closest to i), we parametrize $T^1\mathbb{H}$ by $\overline{\mathbb{R}} \times \overline{\mathbb{R}} \times \mathbb{R}$ (up to the diagonal, of measure 0).
- The lifted set $\tilde{B}_+$ is a union of geodesics: $\tilde{B}_+ = C \times \mathbb{R}$ with $C \subset \overline{\mathbb{R}} \times \overline{\mathbb{R}}$. By the property of B_+, two geodesics with the same *positive* extremity are both in or both out: $\tilde{B}_+ = b_+ \times \overline{\mathbb{R}} \times \mathbb{R}$.

Proof of the lemma (1/2): coordinates

- Lift the sets to the unit tangent bundle of $\mathbb{H}$: the projection is countable-to-one, so a set of measure 0 lifts to a set of measure 0.
- **Coordinates**: to a unit vector we associate its geodesic, completely determined by its two extremities (ξ_+, ξ_-) in $\overline{\mathbb{R}} = \partial\mathbb{H}$. Fixing a base point on the geodesic (for instance the point closest to i), we parametrize $T^1\mathbb{H}$ by $\overline{\mathbb{R}} \times \overline{\mathbb{R}} \times \mathbb{R}$ (up to the diagonal, of measure 0).
- The lifted set $\tilde{B}_+$ is a union of geodesics: $\tilde{B}_+ = C \times \mathbb{R}$ with $C \subset \overline{\mathbb{R}} \times \overline{\mathbb{R}}$. By the property of B_+, two geodesics with the same *positive* extremity are both in or both out: $\tilde{B}_+ = b_+ \times \overline{\mathbb{R}} \times \mathbb{R}$.
- For the same reason $\tilde{B}_- = \overline{\mathbb{R}} \times b_- \times \mathbb{R}$.

Proof of the lemma (2/2): conclusion

- The set $\tilde{B}_- \setminus \tilde{B}_+ = c_+ \times b_- \times \mathbb{R}$, where c_+ is the complement of b_+ in $\overline{\mathbb{R}}$, has measure 0 by hypothesis ($\mu(B_+ \Delta B_-) = 0$).

Proof of the lemma (2/2): conclusion

- The set $\tilde{B}_- \setminus \tilde{B}_+ = c_+ \times b_- \times \mathbb{R}$, where c_+ is the complement of b_+ in $\overline{\mathbb{R}}$, has measure 0 by hypothesis ($\mu(B_+ \Delta B_-) = 0$).
- In these coordinates the Liouville measure is equivalent to the product of Lebesgue measures. So either b_- has measure 0, or c_+ has measure 0.

Proof of the lemma (2/2): conclusion

- The set $\tilde{B}_- \setminus \tilde{B}_+ = c_+ \times b_- \times \mathbb{R}$, where c_+ is the complement of b_+ in $\overline{\mathbb{R}}$, has measure 0 by hypothesis ($\mu(B_+ \Delta B_-) = 0$).
- In these coordinates the Liouville measure is equivalent to the product of Lebesgue measures. So either b_- has measure 0, or c_+ has measure 0.
- If c_+ has measure 0, then $\tilde{B}_+$ has full measure: so does B_+.

Proof of the lemma (2/2): conclusion

- The set $\tilde{B}_- \setminus \tilde{B}_+ = c_+ \times b_- \times \mathbb{R}$, where c_+ is the complement of b_+ in $\overline{\mathbb{R}}$, has measure 0 by hypothesis ($\mu(B_+ \Delta B_-) = 0$).
- In these coordinates the Liouville measure is equivalent to the product of Lebesgue measures. So either b_- has measure 0, or c_+ has measure 0.
- If c_+ has measure 0, then $\tilde{B}_+$ has full measure: so does B_+.
- If b_- has measure 0, then B_- has measure 0, and so does B_+ (the symmetric difference is null). ■

Hopf's theorem

- **Theorem (Hopf, 1936)**: the geodesic flow on a surface of constant negative curvature and finite volume is ergodic.

Hopf's theorem

- **Theorem (Hopf, 1936)**: the geodesic flow on a surface of constant negative curvature and finite volume is ergodic.
- **Proof**: a consequence of Birkhoff's ergodic theorem. Let f be integrable; we must show that its Birkhoff integrals converge a.e. to a constant. By continuity of $f \mapsto f^+$, it is enough to take f continuous with compact support.

Hopf's theorem

- **Theorem (Hopf, 1936)**: the geodesic flow on a surface of constant negative curvature and finite volume is ergodic.
- **Proof**: a consequence of Birkhoff's ergodic theorem. Let f be integrable; we must show that its Birkhoff integrals converge a.e. to a constant. By continuity of $f \mapsto f^+$, it is enough to take f continuous with compact support.
- By Birkhoff's theorem, the limits f^+ (forward) and f^- (backward) exist a.e. and are a.e. equal.

Hopf's theorem

- **Theorem (Hopf, 1936)**: the geodesic flow on a surface of constant negative curvature and finite volume is ergodic.
- **Proof**: a consequence of Birkhoff's ergodic theorem. Let f be integrable; we must show that its Birkhoff integrals converge a.e. to a constant. By continuity of $f \mapsto f^+$, it is enough to take f continuous with compact support.
- By Birkhoff's theorem, the limits f^+ (forward) and f^- (backward) exist a.e. and are a.e. equal.
- Let $B_+ = \{f^+ > c\}$ and $B_- = \{f^- > c\}$. They are invariant by the geodesic flow, they differ by a null set, and, by continuity of f, if two points are positively (resp. negatively) asymptotic, they are both in or both out of B_+ (resp. B_-).

Hopf's theorem

- **Theorem (Hopf, 1936)**: the geodesic flow on a surface of constant negative curvature and finite volume is ergodic.
- **Proof**: a consequence of Birkhoff's ergodic theorem. Let f be integrable; we must show that its Birkhoff integrals converge a.e. to a constant. By continuity of $f \mapsto f^+$, it is enough to take f continuous with compact support.
- By Birkhoff's theorem, the limits f^+ (forward) and f^- (backward) exist a.e. and are a.e. equal.
- Let $B_+ = \{f^+ > c\}$ and $B_- = \{f^- > c\}$. They are invariant by the geodesic flow, they differ by a null set, and, by continuity of f, if two points are positively (resp. negatively) asymptotic, they are both in or both out of B_+ (resp. B_-).
- By the lemma, $\{f^+ > c\}$ is null or conull for every c: f^+ is a.e. constant. ■

Remarks on the two proofs

- We did *not* use here that the horocycle flows generate everything, but the weaker fact that they, together with the geodesic flow, generate everything. And we got a weaker result: ergodicity, not weak mixing.

Remarks on the two proofs

- We did *not* use here that the horocycle flows generate everything, but the weaker fact that they, together with the geodesic flow, generate everything. And we got a weaker result: ergodicity, not weak mixing.
- A very symmetric situation: a group $G = \mathrm{SL}(2, \mathbb{R})$, a discrete subgroup $\Gamma = \mathrm{SL}(2, \mathbb{Z})$, a flow (one-parameter subgroup A) acting on the right on $\Gamma \backslash G$. We can reverse the situation: Γ acts on G/A, the space of geodesics of $\mathbb{H}$.

Remarks on the two proofs

- We did *not* use here that the horocycle flows generate everything, but the weaker fact that they, together with the geodesic flow, generate everything. And we got a weaker result: ergodicity, not weak mixing.
- A very symmetric situation: a group $G = \mathrm{SL}(2, \mathbb{R})$, a discrete subgroup $\Gamma = \mathrm{SL}(2, \mathbb{Z})$, a flow (one-parameter subgroup A) acting on the right on $\Gamma \backslash G$. We can reverse the situation: Γ acts on G/A, the space of geodesics of $\mathbb{H}$.
- **Proposition**: the action of $\mathrm{SL}(2, \mathbb{Z})$ on $\overline{\mathbb{R}}$ is ergodic: an invariant measurable set has measure 0, or its complement has measure 0.

Remarks on the two proofs

- We did *not* use here that the horocycle flows generate everything, but the weaker fact that they, together with the geodesic flow, generate everything. And we got a weaker result: ergodicity, not weak mixing.
- A very symmetric situation: a group $G = \mathrm{SL}(2, \mathbb{R})$, a discrete subgroup $\Gamma = \mathrm{SL}(2, \mathbb{Z})$, a flow (one-parameter subgroup A) acting on the right on $\Gamma \backslash G$. We can reverse the situation: Γ acts on G/A, the space of geodesics of $\mathbb{H}$.
- **Proposition**: the action of $\mathrm{SL}(2, \mathbb{Z})$ on $\overline{\mathbb{R}}$ is ergodic: an invariant measurable set has measure 0, or its complement has measure 0.
- **Duality (Moore)**: let Γ_1, Γ_2 be closed subgroups of a group G. The left action of Γ_1 on G/Γ_2 is ergodic if and only if the right action of Γ_2 on $\Gamma_1 \backslash G$ is ergodic.

Horocycle invariant implies geodesic invariant (1/2)

- We give two proofs. **Lemma**: if $f \in L^2$ is invariant by the horocycle flow, then $(G_t f \mid f) = (f \mid f)$ for all t.

Horocycle invariant implies geodesic invariant (1/2)

- We give two proofs. **Lemma**: if $f \in L^2$ is invariant by the horocycle flow, then $(G_t f \mid f) = (f \mid f)$ for all t.
- **Proof**: let $s = e^{t/2}$, so $g_t = \begin{pmatrix} s & 0 \\ 0 & 1/s \end{pmatrix}$. We want $|(G_t f | f) - (f | f)| < \epsilon$ for ϵ arbitrary.

Horocycle invariant implies geodesic invariant (1/2)

- We give two proofs. **Lemma**: if $f \in L^2$ is invariant by the horocycle flow, then $(G_t f \mid f) = (f \mid f)$ for all t.
- **Proof**: let $s = e^{t/2}$, so $g_t = \begin{pmatrix} s & 0 \\ 0 & 1/s \end{pmatrix}$. We want $|(G_t f|f) - (f|f)| < \epsilon$ for ϵ arbitrary.
- By continuity of the negative horocycle flow, choose a small so that

$$|(G_t f|f) - (G_t H_a^- f|f)| < \tfrac{\epsilon}{2}, \qquad |(H_{a/s}^- f|f) - (f|f)| < \tfrac{\epsilon}{2}.$$

(the first is independent of t because G_t is unitary; the second holds for $s \geq 1$).

Horocycle invariant implies geodesic invariant (2/2)

- A matrix computation: with $s_1 = \dfrac{s - s^2}{a}$ and $s_2 = \dfrac{s-1}{a}$,

$$h^+_{s_1}\, g_t\, h^-_a\, h^+_{s_2} = h^-_{a/s}.$$

Horocycle invariant implies geodesic invariant (2/2)

- A matrix computation: with $s_1 = \dfrac{s - s^2}{a}$ and $s_2 = \dfrac{s-1}{a}$,

$$h^+_{s_1}\, g_t\, h^-_a\, h^+_{s_2} = h^-_{a/s}.$$

- Since f is invariant by H_s, we get
$(G_t H^-_a f \mid f) = (H_{s_1} G_t H^-_a H_{s_2} f \mid f) = (H^-_{a/s} f \mid f)$.

Horocycle invariant implies geodesic invariant (2/2)

- A matrix computation: with $s_1 = \dfrac{s-s^2}{a}$ and $s_2 = \dfrac{s-1}{a}$,
$$h^+_{s_1}\, g_t\, h^-_a\, h^+_{s_2} = h^-_{a/s}.$$
- Since f is invariant by H_s, we get
$(G_t H^-_a f \mid f) = (H_{s_1} G_t H^-_a H_{s_2} f \mid f) = (H^-_{a/s} f \mid f)$.
- So $|(G_t f|f) - (f|f)| < \epsilon$ for all large t and all ϵ. ■

Horocycle invariant implies geodesic invariant (2/2)

- A matrix computation: with $s_1 = \dfrac{s-s^2}{a}$ and $s_2 = \dfrac{s-1}{a}$,
$$h^+_{s_1}\, g_t\, h^-_a\, h^+_{s_2} = h^-_{a/s}.$$
- Since f is invariant by H_s, we get
$(G_t H^-_a f \,|\, f) = (H_{s_1} G_t H^-_a H_{s_2} f \,|\, f) = (H^-_{a/s} f \,|\, f)$.
- So $|(G_t f|f) - (f|f)| < \epsilon$ for all large t and all ϵ. ■
- **Theorem**: the horocycle flow is ergodic.

Horocycle invariant implies geodesic invariant (2/2)

- A matrix computation: with $s_1 = \dfrac{s-s^2}{a}$ and $s_2 = \dfrac{s-1}{a}$,
$$h^+_{s_1}\, g_t\, h^-_a\, h^+_{s_2} = h^-_{a/s}.$$
- Since f is invariant by H_s, we get $(G_t H^-_a f \mid f) = (H_{s_1} G_t H^-_a H_{s_2} f \mid f) = (H^-_{a/s} f \mid f)$.
- So $|(G_t f|f) - (f|f)| < \epsilon$ for all large t and all ϵ. ■
- **Theorem**: the horocycle flow is ergodic.
- **Proof**: let f be invariant, and E the space generated in L^2 by the orbit of f under G_t: every element of E is horocycle invariant (commutation relation), so $(G_t x|x) = (x|x)$ on E. As G_t is unitary, $G_t x = x$: f is geodesic invariant, hence constant by the ergodicity of the geodesic flow. ■

Another identity

- Let R_a be the rotation $\begin{pmatrix} \cos\frac{a}{2} & \sin\frac{a}{2} \\ -\sin\frac{a}{2} & \cos\frac{a}{2} \end{pmatrix}$ (do not forget we are in $\mathrm{PSL}(2,\mathbb{R})$). For $s > 0$ define t and $a \in (0, \frac{\pi}{2}]$ by

$$\frac{s}{2} = \cot a = \sinh\frac{t}{2}.$$

Another identity

- Let R_a be the rotation $\begin{pmatrix} \cos\frac{a}{2} & \sin\frac{a}{2} \\ -\sin\frac{a}{2} & \cos\frac{a}{2} \end{pmatrix}$ (do not forget we are in $\mathrm{PSL}(2,\mathbb{R})$). For $s > 0$ define t and $a \in (0, \frac{\pi}{2}]$ by

$$\frac{s}{2} = \cot a = \sinh\frac{t}{2}.$$

- Then

$$R_a\, h_s^+ = g_t\, R_{\pi-a}, \qquad \text{i.e.} \qquad h_s^+ = R_{-a}\, g_t\, R_{\pi-a}.$$

(a Cartan decomposition of h_s^+).

Another identity

- Let R_a be the rotation $\begin{pmatrix} \cos\frac{a}{2} & \sin\frac{a}{2} \\ -\sin\frac{a}{2} & \cos\frac{a}{2} \end{pmatrix}$ (do not forget we are in $\mathrm{PSL}(2,\mathbb{R})$). For $s > 0$ define t and $a \in (0, \frac{\pi}{2}]$ by

$$\frac{s}{2} = \cot a = \sinh\frac{t}{2}.$$

- Then

$$R_a\, h_s^+ = g_t\, R_{\pi-a}, \qquad \text{i.e.} \qquad h_s^+ = R_{-a}\, g_t\, R_{\pi-a}.$$

(a Cartan decomposition of h_s^+).

- When $s \to \infty$, $t \to \infty$ and $a \to 0$: formally $h_\infty^+ = g_\infty R_\pi$.

Another identity

- Let R_a be the rotation $\begin{pmatrix} \cos\frac{a}{2} & \sin\frac{a}{2} \\ -\sin\frac{a}{2} & \cos\frac{a}{2} \end{pmatrix}$ (do not forget we are in $\mathrm{PSL}(2,\mathbb{R})$). For $s > 0$ define t and $a \in (0, \frac{\pi}{2}]$ by
$$\frac{s}{2} = \cot a = \sinh\frac{t}{2}.$$
- Then
$$R_a\, h_s^+ = g_t\, R_{\pi-a}, \qquad \text{i.e.} \qquad h_s^+ = R_{-a}\, g_t\, R_{\pi-a}.$$
(a Cartan decomposition of h_s^+).
- When $s \to \infty$, $t \to \infty$ and $a \to 0$: formally $h_\infty^+ = g_\infty R_\pi$.
- Here $R_\pi = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ is of order 2 in $\mathrm{PSL}(2,\mathbb{R})$, and preserves the measure.

Second proof of the ergodicity of the horocycle flow

- Let f be continuous with compact support, invariant by h_s^+; we write $s = s(t)$, $a = a(t)$ as above. Using continuity ($a(t) \to 0$), the invariance, and the identity $R_a h_s^+ = g_t R_{\pi - a}$:

$$f(x) = \lim_{T\to\infty} \frac{1}{T} \int_0^T f(x)\, dt = \lim_{T\to\infty} \frac{1}{T} \int_0^T f(x\, R_{a(t)})\, dt$$

$$= \lim_{T\to\infty} \frac{1}{T} \int_0^T f(x\, R_{a(t)} h^+_{s(t)})\, dt = \lim_{T\to\infty} \frac{1}{T} \int_0^T f(x\, g_t\, R_{\pi - a(t)})\, dt$$

$$= \lim_{T\to\infty} \frac{1}{T} \int_0^T f(x\, g_t\, R_\pi)\, dt = \int f(y\, R_\pi)\, d\mu(y) = \int f\, d\mu.$$

Second proof of the ergodicity of the horocycle flow

- Let f be continuous with compact support, invariant by h_s^+; we write $s = s(t)$, $a = a(t)$ as above. Using continuity ($a(t) \to 0$), the invariance, and the identity $R_a h_s^+ = g_t R_{\pi - a}$:

$$f(x) = \lim_{T\to\infty} \frac{1}{T} \int_0^T f(x)\, dt = \lim_{T\to\infty} \frac{1}{T} \int_0^T f(x\, R_{a(t)})\, dt$$
$$= \lim_{T\to\infty} \frac{1}{T} \int_0^T f(x\, R_{a(t)} h^+_{s(t)})\, dt = \lim_{T\to\infty} \frac{1}{T} \int_0^T f(x\, g_t\, R_{\pi - a(t)})\, dt$$
$$= \lim_{T\to\infty} \frac{1}{T} \int_0^T f(x\, g_t\, R_\pi)\, dt = \int f(y\, R_\pi)\, d\mu(y) = \int f\, d\mu.$$

- The last limit is Birkhoff's theorem for the geodesic flow, which is ergodic, applied to $f \circ R_\pi$; and μ is invariant by R_π. So f is a.e. constant; by density, the horocycle flow is ergodic. ■

Strong mixing of the geodesic flow: statement

- We want, for two functions $e, f \in L^2(\mu)$:

$$\lim_{t\to\infty} \int f(x\, g_t)\, e(x)\, d\mu(x) = \int f\, d\mu \int e\, d\mu.$$

Strong mixing of the geodesic flow: statement

- We want, for two functions $e, f \in L^2(\mu)$:

$$\lim_{t\to\infty} \int f(x\, g_t)\, e(x)\, d\mu(x) = \int f\, d\mu \int e\, d\mu.$$

- By density and unitarity of G_t we may suppose e and f continuous with compact support.

Strong mixing of the geodesic flow: statement

- We want, for two functions $e, f \in L^2(\mu)$:

$$\lim_{t\to\infty} \int f(x\, g_t)\, e(x)\, d\mu(x) = \int f\, d\mu \int e\, d\mu.$$

- By density and unitarity of G_t we may suppose e and f continuous with compact support.
- **Idea**: the geodesic flow *expands* the negative horocycles: a short negative horocycle segment becomes, after time t, a very long one. A long horocycle segment equidistributes, by the ergodicity of the horocycle flow.

Strong mixing of the geodesic flow: proof (1/2)

- By invariance of μ under h_s^-, then averaging over $s \in [0, \delta]$, and since $e(x\, h_s^-) = e(x) + O(\epsilon)$ uniformly for small δ:

$$\int f(x\, g_t)\, e(x)\, d\mu = \int e(x)\, \frac{1}{\delta} \int_0^\delta f(x\, h_s^-\, g_t)\, ds\, d\mu(x) + O(\epsilon).$$

Strong mixing of the geodesic flow: proof (1/2)

- By invariance of μ under h_s^-, then averaging over $s \in [0, \delta]$, and since $e(x\, h_s^-) = e(x) + O(\epsilon)$ uniformly for small δ:

$$\int f(x\, g_t)\, e(x)\, d\mu = \int e(x)\, \frac{1}{\delta} \int_0^\delta f(x\, h_s^-\, g_t)\, ds\, d\mu(x) + O(\epsilon).$$

- **Expansion**: $h_s^- g_t = g_t h_{se^t}^-$. With $z = x\, g_t$ and $u = se^t$, the inner average is

$$A_{\delta e^t}(z), \qquad A_T(z) = \frac{1}{T} \int_0^T f(z\, h_u^-)\, du.$$

Strong mixing of the geodesic flow: proof (2/2)

- With $x = z\, g_{-t}$ (invariance of μ):

$$\int f(x\, g_t)\, e(x)\, d\mu = \int e(z\, g_{-t})\, A_{\delta e^t}(z)\, d\mu(z) + O(\epsilon).$$

Strong mixing of the geodesic flow: proof (2/2)

- With $x = z\, g_{-t}$ (invariance of μ):

$$\int f(x\, g_t)\, e(x)\, d\mu = \int e(z\, g_{-t})\, A_{\delta e^t}(z)\, d\mu(z) + O(\epsilon).$$

- The negative horocycle flow is ergodic (by symmetry with the positive one, conjugating by R_π): $A_T \to \int f\, d\mu$ in $L^2(\mu)$ when $T \to \infty$.

Strong mixing of the geodesic flow: proof (2/2)

- With $x = z\,g_{-t}$ (invariance of μ):

$$\int f(x\,g_t)\,e(x)\,d\mu = \int e(z\,g_{-t})\,A_{\delta e^t}(z)\,d\mu(z) + O(\epsilon).$$

- The negative horocycle flow is ergodic (by symmetry with the positive one, conjugating by R_π): $A_T \to \int f\,d\mu$ in $L^2(\mu)$ when $T \to \infty$.
- As $\|e \circ g_{-t}\|_2 = \|e\|_2$ and $\int e(z\,g_{-t})\,d\mu = \int e\,d\mu$, the limit is $\int f\,d\mu \int e\,d\mu$, up to $O(\epsilon)$. ■

Mixing of the horocycle flow

- We use the commutation relation again, for two measurable sets A and B. From $h_s^+ = R_{-a}\, g_t\, R_{\pi-a}$:

$$\mu(A\, h_s^+ \cap B) = \mu\big(A\, R_{-a}\, g_t\, R_{\pi-a} \cap B\big) = \mu\big(A\, R_{-a}\, g_t \cap B\, R_{a-\pi}\big),$$

by invariance of μ under the rotations.

Mixing of the horocycle flow

- We use the commutation relation again, for two measurable sets A and B. From $h_s^+ = R_{-a}\, g_t\, R_{\pi-a}$:

$$\mu(A\, h_s^+ \cap B) = \mu\big(A\, R_{-a}\, g_t\, R_{\pi-a} \cap B\big) = \mu\big(A\, R_{-a}\, g_t \cap B\, R_{a-\pi}\big),$$

  by invariance of μ under the rotations.

- When $s \to \infty$, $a \to 0$: up to ϵ we can replace a by 0, and we get $\mu(A\, g_t \cap B\, R_\pi)$ (here $R_{-\pi} = R_\pi$ in PSL), which tends to $\mu(A)\,\mu(B\, R_\pi) = \mu(A)\,\mu(B)$ by the mixing of the geodesic flow and the invariance of μ under R_π. ■

The non-compact case: the cusp

- **Remark**: in the non-compact case, there are trivial obstructions to minimality because of the cusps.

The non-compact case: the cusp

- **Remark**: in the non-compact case, there are trivial obstructions to minimality because of the cusps.
- The first example is the modular surface: the horizontal horocycles, corresponding to the element $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, give **periodic orbits** of the horocycle flow (they are the closed parallels of the end of the pseudosphere).

The non-compact case: the cusp

- **Remark**: in the non-compact case, there are trivial obstructions to minimality because of the cusps.
- The first example is the modular surface: the horizontal horocycles, corresponding to the element $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, give **periodic orbits** of the horocycle flow (they are the closed parallels of the end of the pseudosphere).
- **Remark**: the geodesic flow is never minimal: it has many periodic orbits and many invariant measures.

Minimality of the horocycle flow (compact case)

- **Theorem**: in the compact case, the horocycle flow is minimal.

Minimality of the horocycle flow (compact case)

- **Theorem**: in the compact case, the horocycle flow is minimal.
- **Proof** (sketch). Cover the space by finitely many balls B_j of diameter ϵ, and by finitely many boxes A_i generated by the geodesic and horocycle flows for time ϵ. It is enough to show that every horocycle orbit meets every B_j.

Minimality of the horocycle flow (compact case)

- **Theorem**: in the compact case, the horocycle flow is minimal.
- **Proof** (sketch). Cover the space by finitely many balls B_j of diameter ϵ, and by finitely many boxes A_i generated by the geodesic and horocycle flows for time ϵ. It is enough to show that every horocycle orbit meets every B_j.
- By mixing and finiteness, there is T such that $A_i\, g_{-T}$ meets B_j for all i, j. Consider the horocycle orbit $c_T = \{x\, h_s^+ :\ 0 \leq s \leq 2\epsilon e^T\}$.

Minimality of the horocycle flow (compact case)

- **Theorem**: in the compact case, the horocycle flow is minimal.
- **Proof** (sketch). Cover the space by finitely many balls B_j of diameter ϵ, and by finitely many boxes A_i generated by the geodesic and horocycle flows for time ϵ. It is enough to show that every horocycle orbit meets every B_j.
- By mixing and finiteness, there is T such that $A_i\, g_{-T}$ meets B_j for all i, j. Consider the horocycle orbit $c_T = \{x\, h_s^+ :\ 0 \leq s \leq 2\epsilon e^T\}$.
- Its image by g_T is a horocycle segment of length 2ϵ (contraction), which covers at least one box A_i (take a box containing $x\, h_{e^T}^+ g_T$).

Minimality of the horocycle flow (compact case)

- **Theorem**: in the compact case, the horocycle flow is minimal.
- **Proof** (sketch). Cover the space by finitely many balls B_j of diameter ϵ, and by finitely many boxes A_i generated by the geodesic and horocycle flows for time ϵ. It is enough to show that every horocycle orbit meets every B_j.
- By mixing and finiteness, there is T such that $A_i\, g_{-T}$ meets B_j for all i, j. Consider the horocycle orbit $c_T = \{x\, h_s^+ :\ 0 \leq s \leq 2\epsilon e^T\}$.
- Its image by g_T is a horocycle segment of length 2ϵ (contraction), which covers at least one box A_i (take a box containing $x\, h_{e^T}^+ g_T$).
- Hence c_T covers $A_i\, g_{-T}$, which meets every ball B_j: c_T passes at distance less than 2ϵ of every point. ■

Unique ergodicity

- **Theorem** (Furstenberg): in the compact case, the horocycle flow is **uniquely ergodic**: Liouville measure is its only invariant probability.

Unique ergodicity

- **Theorem** (Furstenberg): in the compact case, the horocycle flow is **uniquely ergodic**: Liouville measure is its only invariant probability.
- It is proved by the same kind of argument (expansion of the geodesic flow and mixing).

Unique ergodicity

- **Theorem** (Furstenberg): in the compact case, the horocycle flow is **uniquely ergodic**: Liouville measure is its only invariant probability.
- It is proved by the same kind of argument (expansion of the geodesic flow and mixing).
- Summary of the proofs: the commutation relation $h_s^+ g_t = g_t h_{se^{-t}}^+$ is used again and again, in the ergodic theory of $\mathrm{SL}(2,\mathbb{R})$-flows. These ideas generalize to all hyperbolic systems.

Exercises on the ergodic theory of the flows (1/2)

- **Exercise 15**. Check the commutation relations $G_{-t} H_s^+ G_t = H_{se^{-t}}^+$ and $G_{-t} H_s^- G_t = H_{se^t}^-$.
- **Exercise 16**. Verify that $R_a \, h_s^+ = g_t \, R_{\pi - a}$ when $\frac{s}{2} = \cot a = \sinh \frac{t}{2}$.
- **Exercise 17**. Show that, for a measure preserving flow, strong mixing implies weak mixing, which implies ergodicity.
- **Exercise 18**. Show that the horocycle flow on the modular surface has periodic orbits, and compute their periods. Deduce that it is not minimal, and not uniquely ergodic.

Exercises on the ergodic theory of the flows (2/2)

- **Exercise 19** ($\star$). In the coordinates (ξ_+, ξ_-, t) of Hopf's proof, the Liouville measure is $\dfrac{d\xi_+\, d\xi_-}{(\xi_+ - \xi_-)^2}\, dt$. With $x = \xi_+ \in (0,1)$ and $y = -1/\xi_- \in (0,1)$, show that it becomes $\dfrac{dx\, dy}{(1+xy)^2}\, dt$: the invariant measure of the natural extension of the Gauss map (Lecture 3).

- **Exercise 20** ($\star$). Show that the endpoints of the lift of a closed geodesic of the modular surface are conjugate quadratic irrationals. Relate this to the periodic continued fractions of Lecture 3.

- **Exercise 21** ($\star$). Show that the closed geodesics are dense in the unit tangent bundle of the modular surface.