Introduction to the Geometry and Dynamics of Hyperbolic Surfaces

CHAPTER 1

Foreword and warning

The aim of this small series of note is to give a concise and elementary introduction to hyperbolic surfaces. Starting from a synthetic point of view, we shall give the classification of compact oriented connected hyperbolic surfaces, introduce some elementary dynamical ideas and ergodic theory, count closed geodesics on them, define arithmetic surfaces and finally state some of the major conjectures in the subject.

The text is meant to be accessible to a student with very little background. However a knowledge of covering spaces and simply connected objects would help as well as an exposition to projective line geometry. No knowledge of differential geometry and specifically Riemannian geometry is assumed and the corresponding basic notions are not introduced. These notes are structured in three parts.

The first part is an introduction to planar hyperbolic geometry. They are many excellent and worthy textbooks introducing hyperbolic geometry and it did not seem useful to repeat the standard (and excellent) presentation which usually starts with the upper half-plane model. Instead, I chose a slightly non-standard synthetic introduction: I present planar hyperbolic geometry as an offshoot of projective line geometry. More precisely, after a classical introduction on the projective line geometry emphasizing the cross-ratio, I define a *hyperbolic plane* as a set with a *boundary at infinity* and *lines* satisfying some axioms. Then I exhibit some models satisfying these axioms: involution model, upper half-plane model. This has the advantage that first it does not rely very much on the ground field provided it is real closed, second it uses very little mathematics besides projective line geometry. The emphasis on the boundary at infinity and the cross-ratio is also an important and useful point of when one wants to deal with surface groups in Lie groups. The disadvantage of this presentation is that the very important metric aspects (length, area, asymptotic geodesics, geodesic minimizing length) of the subject come afterwards, and also that we restrict ourselves to planar geometry. After that, I present classical results about hexagons and the Gauß–Bonnet Formula for triangles. In a second chapter, I present hyperbolic surfaces as metric spaces

locally isometric to hyperbolic planes. Then I deal with completeness and prove the relevant version of Hopf–Rinow Theorem, prove that every complete hyperbolic surfaces is a quotient and that every surface is obtained by gluing pair of pants thus stopping very short of the construction of Teichmüller space with the Fenchel–Nielsen coordinates. Not using the basic tools of differential geometry, Riemannian geometry, makes the notes accessible with very little knowledge, but forces to do some gymnastics.

The reader familiar with hyperbolic geometry should skip directly to the **second part**. In this part, I introduce the unit tangent bundle of the hyperbolic plane together with its two fundamental commuting actions: the action on the left of the isometry group, the action on the right of the group $\mathsf{PSL}_2(\mathbb{R})$. I then introduce the major players: the geodesic and horospherical flows. I then prove the Anosov property, the closing lemma, Poincaré Recurrence Theorem, the Statistical Ergodic Theorem, mixing and ergodicity of the flows using the unitary representation methods – although Hopf argument for ergodicity is given. The final chapters are concerned with equidistributions and counting à la Margulis: equidistribution of circles, counting orbit points, equidistribution of horocycles, of geodesics and their asymptotic counting. The proofs in this simple case of hyperbolic surfaces is much easier and less technical than the general cases that could be considered and I hope that this makes the beautiful ideas of the proofs more transparent. However, the proofs are still intricate.

The third and final part is really about tourism: I tour, essentially without proofs, two very important subjects for hyperbolic surfaces: arithmeticity and harmonic functions. I try to emphasize the relation with the geodesic flow as well as common features. Some very elementary proofs are given and thanks to Margulis commensurability criterion we emphasize the point of view that arihtmetics give extra dynamics which can be seen as acting on the hyperbolic solenoid. Using several crucial results as black boxes, I give a proof of the equidistribution of Hecke point. This part should really be seen as encouraging the reader to have a look into more serious references on these fascinating subjects.

Each part is concluded with suggestions of references and further reading. None of the material here is original, although some of the proofs were simplified since we used a rather specific setting. These notes are supposed to be complementary reading, parallel to some serious study of this central theme in Mathematics, or as a "tourist guide" for a curious reader and in the best case as an invitation to learn more.

They are many exercises in the book. And many parts of the proofs are left as exercises. Some of the exercises are easy applications. Those with a (*)

are harder. Those with (**) are definitely difficult and may require knowledge outside of these notes, and for those with (***), well, I actually do not know the answer.

A shorter version of this set of notes was initially intended for a summer class in the CIRM, and the present much expanded state is the result of an experimental testing on graduate students and post-docs during the Fields Institute semester on Random Geometric Structures, while I was on a University of Toronto Dean Professorship's visiting position. I very warmly thank the audience there for their questions and patience, which helped improve the writing up. I also hope no harm came out of my experimentation.

Contents

Part 1

Hyperbolic geometry and hyperbolic surfaces

CHAPTER 2

Hyperbolic plane

1. The projective line

We start by giving without proofs a summary of projective geometry.

Given a vector space V over a field $\mathbb{K}$, the *projective space* $\mathbf{P}(V)$ is the set of non-zero vectors in V up to multiplication by an element of $\mathbb{K}^*$:

$$\mathbf{P}(V) \coloneqq \{L \subset V \mid \dim(L) = 1\} = V \setminus \{0\} \,/\, \mathbb{K}^* \,.$$

For a non-zero element v in V, we denote by $[v]$ its class in $\mathbf{P}(V)$. Geometrically, we interpret $\mathbf{P}(V)$ as the set of lines of V. By construction the group

$$\mathsf{PGL}(V) \coloneqq \mathsf{GL}(V)/\mathbb{K}\,\mathrm{Id}\,,$$

acts transitively on $\mathbf{P}(V)$. The corresponding bijections of $\mathbf{P}(V)$ are called *projective transformations*.

1.1. Basis and coordinates. A basis $(e_1, \ldots, e_n)$ of V identifies V with $\mathbb{K}^n$. We denote every element L in $\mathbf{P}(V)$ as

$$L = [x_1, \ldots, x_n]\,,$$

where $(x_1, \ldots, x_n)$ is a non zero vector in L. The set $(x_1, \ldots, x_n)$ is well defined up to a multiplication by a non zero element of $\mathbb{K}$.

The set $(x_1, \ldots, x_n)$ are – by a slight abuse of language – called the *projective coordinate* of L. Every n-tuple $(x_1, \ldots, x_n)$, with x_i not all zero, is the projective coordinate of some line.

A usual convention, that we will follow, is to replace the comma "," by a colon ":" and replace the above equation by

$$L = [x_1 : \ldots : x_n]\,.$$

Observe that

$$[x_1 : \ldots : x_n] = [y_1 : \ldots : y_n]\,,$$

if and only if there is a non-zero element λ of $\mathbb{K}$ such that for all i, $x_i = \lambda_i$.

1.2. Projective line and homographies. When $\dim(V) = 2$, the projective space $\mathbf{P}(V)$ is called the *projective line* and we will now focus in that situation. Adding a symbol ∞, called infinity, to $\mathbb{K}$ we have a bijection of $\mathbf{P}^1(\mathbb{K}) := \mathbf{P}(\mathbb{K}^2)$ with $\mathbb{K} \sqcup \{\infty\}$ given in projective coordinates by

$$[1 : \lambda] \mapsto \lambda \ , \ [0, 1] \mapsto \infty \ .$$

Given a basis (e_1, e_2), by a slight abuse of language, the element λ of $\mathbb{K} \sqcup \{\infty\}$ associated to a point x in $\mathbf{P}(V)$ is also called the *projective coordinate* of x.

The associated action of the group $\mathsf{PGL}_2(\mathbb{K})$ on $\mathbb{K} \sqcup \{\infty\}$ is then given by

$$\text{for } z \neq -\tfrac{d}{c} \ , \ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z := \frac{az + b}{cz + d} \ ; \quad \text{for } z = -\tfrac{d}{c} \ , \ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z := \infty \ ,$$

$$\text{for } c \neq 0, \ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \infty := \frac{a}{c} \ ; \quad \text{finally, } \ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \infty := \infty \ .$$

The corresponding bijections of $\mathbb{K} \sqcup \{\infty\}$ are called *homographies*. As an exercise, the reader is invited to check the following proposition

Proposition 1.1. *Given three pairwise distinct points (x, y, z) in $\mathbf{P}(V)$, there exists a basis (e_1, e_2) so that x , y and z have coordinates respectively 0, 1 and ∞. Moreover if (f_1, f_2) is another basis of V satisfying the same condition, then there is a non-zero element λ of $\mathbb{K}$ so that $e_i = \lambda f_i$.*

As a corollary we see that $\mathsf{PGL}_2(\mathbb{K})$ acts freely and transitively on the set of pairwise distinct triples of $\mathbf{P}(V)$. We will see in the next paragraph that this is no longer the case for quadruples of points.

1.3. Quadruple of points and the cross-ratio. Let's denote by V^* the dual of V. Since we are in dimension 2, the map from $\mathbf{P}(V)$ to $\mathbf{P}(V^*)$,

$$x \mapsto x^\perp := \{u \in V^* \mid x \subset \ker(u)\} \ ,$$

is a bijection.

Let a, b, c and d be four pairwise distinct points in $\mathbf{P}(V)$, let a_0, b_0, c^0 and d^0 be non-zero vectors in a, b, $c^\perp$ and $d^\perp$ respectively. We immediately observe that the quantity

$$\frac{c^0(a_0)d^0(b_0)}{d^0(a_0)c^0(b_0)} \ ,$$

only depends on the quadruple (a, b, c, d). Following the same notation, we then define the *cross-ratio* of (a, b, c, d) as

$$[a; b; c; d] := \frac{c^0(a_0)d^0(b_0)}{d^0(a_0)c^0(b_0)} \tag{1}$$

From this definition, we immediately see that

Proposition 1.2. *For any element g of* $\mathsf{PGL}(V)$,

$$[a;b;c;d] = [g(a);g(b);g(c);g(d)] .$$

One can then compute the projective cross-ratio in coordinates and we obtain that if the projective coordinates of a, b, c and d are α, β, γ and δ respectively then

Proposition 1.3. *For α, β, γ and δ all different from infinity*

- *For α, β, γ and δ all different from ∞,* $[a;b;c;d] = \frac{(\alpha-\gamma)(\beta-\delta)}{(\alpha-\delta)(\beta-\gamma)}$,
- *Moreover,* $[a;b;c;\infty] = \frac{(\alpha-\gamma)}{(\beta-\gamma)}$.

In particular from this proposition, we have

$$[\lambda;1;0;\infty] = \lambda .$$

From this proposition we see all the symmetries of the cross-ratio.

Proposition 1.4. *We have, whenever it makes sense,*

$$\begin{aligned} [a;b;c;d]\cdot[a;b;d;e] &= [a;b;c;e] , \\ [a;b;c;d] &= [c;d;a;b] , \\ [a;b;c;d] + [a;c;b;d] &= 1 . \end{aligned}$$

We leave as an exercise, easily proved using the previous propositions, the following theorem which is some sort of reciprocal of the previous proposition

Theorem 1.5 (Fundamental Theorem of projective geometry). *Let X be a set and b a $\mathbb{K}$-valued function with values in $\mathbb{K}^*$ on the set of quadruples of pairwise distinct points of X satisfying, whenever it makes sense,*

$$\begin{aligned} b(x,y,z,t)\, b(x,y,t,u) &= b(x,y,z,u) , \\ b(x,y,z;t) &= b(z;t;x;y) , \\ b(x,y,z;t) + b(x;z;y;t) &= 1 . \end{aligned}$$

Then, given any 2-dimensional vector space V over $\mathbb{K}$, there is an injection Φ from X to $\mathbf{P}(V)$ so that

$$b(x,y,z,t) = [\Phi(x);\Phi(y);\Phi(z);\Phi(t)] ,$$

moreover Φ is well defined up to post composition by an element of $\mathsf{PGL}(V)$.

This theorem has the following corollary

Corollary 1.6. *Let Φ be a bijection of $\mathbf{P}(V)$ so that for any quadruple of pairwise distinct points*

$$[\Phi(x);\Phi(y);\Phi(z)\Phi(t)] = [x;y;z;t] .$$

Then Φ belongs to $\mathsf{PGL}(V)$

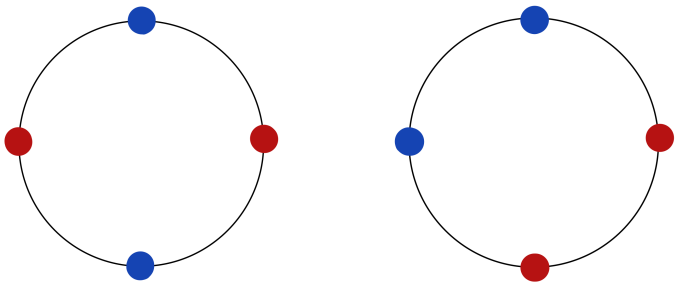
The theorem and it corollary are usually stated as "*the cross-ratio determines the projective geometry of the projective line*".

1.4. The real projective line and orientation. Let us focus on $\mathbb{K} = \mathbb{R}$. In that context, choosing an orientation on V, we distinguish between positively oriented basis and negatively oriented ones.

We saw that every triple (x, y, z) of pairwise distinct points correspond to a basis, up homotheties. Since an homothety has a positive determinant, it follows that we can can speak of *positively oriented triples*, corresponding to positively oriented basis, and *negatively oriented triple* corresponding to *negatively oriented basis*.

We also have another structure that comes from the order structure on $\mathbb{R}$, this time independent on the orientation. Let x, y, z and w be four pairwise distinct points on $\mathbf{P}(V)$. We then say that x, y and z, w *intersect* or *intertwine* if choosing coordinates so that $(x, y, z) = (0, 1, \infty)$, then w belongs to $[0, 1]$.

All this is better seen on pictures. Taking the quotient topology on $\mathbf{P}(V)$, we see that the projective line $\mathbf{P}(V)$ is homeomorphic to the circle. The choice of an orientation on V correspond to an orientation on the circle $\mathbf{P}(V)$. Then if we represent x, y by a pair of blue dots and z, w by red dots, we have figure 1.



(1) Intertwining pairs (2) Non-intertwining pairs

FIGURE 1. Quadruples

1.5. Complex and real projective lines. Let E be a vector space of dimension 2 over $\mathbb{C}$. Recall that a *totally real 2-plane* P is a 2-plane in E seen as 4-dimensional real vector space such that

$$P \oplus iP = E\,.$$

Associated to a totally real plane P we have a *circle* C_P in $\mathbf{P}(E)$ which is the set of complex lines that intersects P.

We can choose a basis in E, which is also a basis for E, thus identifying E with $\mathbb{C}^2$ and P with $\mathbb{R}^2$, we then see

PROPOSITION 1.7. (1) *For every quadruple of points (x, y, z, t) in C_P, then the cross-ratio $[x; y; z; t]$ is real.*
(2) *if (x, y, z) is a triple of pairwise distinct points of P and the cross-ratio $[x; y; z; t]$ is real, then t belongs to C_P.*
(3) *given any triple of pairwise distinct points in $\mathbf{P}(E)$ there exists a unique totally real plane P containing that triple.*
(4) *Let P be a totally real plane, then the map*

$$D \mapsto D \cap P ,$$

from C_P to $\mathbf{P}(P)$ is a bijection preserving the cross-ratio

Let us consider the subgroup G of $\mathsf{PSL}(V)$ defined by

$$\mathsf{G}_P = \{g \in \mathsf{PGL}(V) \mid g(P) = P\} .$$

Since every element of G_P preserves the cross-ratio, we deduce that the restriction of $\mathbf{G}_P$ to C_P is a projective transformation. Moreover, if an element of $\mathbf{G}$ acts trivially on C_P then its is trivial, it then follows that we can identify G_P with $\mathsf{PGL}(P)$, and thus consider routinely $\mathsf{PGL}(P)$ as a subgroup of $\mathsf{PGL}(V)$, the one that preserves globally C_P.

This proposition also shows that we can associate to a totally real plane P, an involution σ_P whose set of fixed points is precisely C_P by the following procedure. if D_P is a complex line, $\sigma(D)$ is characterized by

$$\forall D_1, D_2, D_3, \in C_P,\ [D_1, D_2, D_3, D] = \overline{[D_1, D_2, D_3, \sigma(D)]}.$$

We finally say that two circles C_P and C_Q are orthogonal, if and only if $P \neq Q$ and σ_P and σ_Q commute.

PROPOSITION 1.8. *The circles C_P and C_Q are orthogonal if and only if σ_P leaves C_Q globally invariant. If C_P and C_Q are orthogonal, then they intersect.*

We also have

PROPOSITION 1.9 (CIRCLES IN COORDINATES).
When we use projective coordinates to identify $\mathbf{P}(\mathbb{C})$ to $\mathbb{C} \sqcup \{\infty\}$, the set of circles associated to totally real planes is identified with the set of circles and lines in $\mathbb{C}$.

Moreover, Let P be a totally real plane. Let (x, y, z) be a triple of points in P, thus identifying $\mathbf{P}(V)$ with $\mathbb{C} \sqcup \{\infty\}$ and C_P with $\mathbb{R} \sqcup \infty$. Then if C_Q is orthogonal to P, the image of C_Q in $\mathbb{C} \sqcup \{\infty\}$ is either a line orthogonal to $\mathbb{R}$, or a circle whose center is on $\mathbb{R}$.

Proof. In $\mathbb{C} \sqcup \{\infty\}$, the circle through x_0, y_0 and z_0 is given by

$$C = \{z \mid \Im([z; x_0; y_0; z_0]) = 0\} .$$

Thus we have

$$\frac{(z-y_0)(x_0-z_0)}{(z-z_0)(x_0-y_0)} - \overline{\left(\frac{(z-y_0)(x_0-z_0)}{(z-z_0)(x_0-y_0)}\right)} = 0 .$$

which we see is of the type

$$\lambda\left(\frac{z-y_0}{z-z_0}\right) - \bar{\lambda}\left(\frac{\bar{z}-\bar{y}_0}{\bar{z}-\bar{z}_0}\right) = 0 .$$

Hence using the change of variables $Z = z - z_0$, we obtain an equation of the type

$$a + \frac{b}{Z} - \bar{a} - \frac{\bar{b}}{\bar{Z}} = 0 ,$$

which we rewrite as

$$Z\bar{Z}(a-\bar{a}) + b\,\bar{Z} - \bar{b}\,Z .$$

which is the equation of a circle or a line. Thus the image of every circle C_P is either a circle or a line. Since the line or circle through three points is unique. It follows that every circle or line in $\mathbb{C}$ is the image of a circle or a line.

Finally, in $\mathbb{C}\sqcup\{\infty\}$, σ_P is given by $z \mapsto \bar{z}$. Thus the circles (in $\mathbf{P}(\mathbb{C})$) orthogonal to P are seen in (in $\mathbb{C} \sqcup \{\infty\}$) as circles or lines orthogonal to the rea axis. □

Corollary 1.10. *The group* $\mathsf{PGL}_2(\mathbb{C})$ *acting on* $\mathbb{C} \sqcup \{\infty\}$ *preserves the set of lines and circles. Moreover the group* $\mathsf{PGL}_2(\mathbb{R})$ *preserves the set of lines and circles orthogonal to the real axis.*

2. Axiomatic geometry of the hyperbolic plane

The complete geometry of the hyperbolic plane can be recovered synthetically from several features, namely *lines* and *boundary at infinity* satisfying some axioms.

Let V be a vector space of dimension 2 over $\mathbb{R}$. [1]

Definition 2.1 (Hyperbolic plane). *A* hyperbolic plane $\mathbf{H}^2$ *(over V) is a set equipped a family of* lines *or* geodesics *which are subsets of*

$$\overline{\mathbf{H}^2} := \mathbf{H}^2 \sqcup \mathbf{P}(V)$$

so that, defining $\mathbf{P}(V)$ *as the* boundary at infinity *of* $\mathbf{H}^2$ *and denoting it* $\partial_\infty \mathbf{H}^2$*. we have*

[1] The reader will check that all that which follows make sense whenever we replace $\mathbb{R}$ with any real field, but we will not pursue in that direction

(1) *Given any two distinct points x and y in $\overline{\mathbf{H}^2}$, there exists a unique line containing x and y,*
(2) *Any line l intersects the boundary at infinity in exactly two points called the* ends *or* endpoints *or* extremities *of* point at infinity *of l.*
(3) *If l_1 is a line with end points a_1 and b_1, and l_2 with ends a_2 and b_2, so that (a_1, a_2, b_1, b_2) are distinct and intertwine, then the two lines intersects in $\mathbf{H}^2$.*
(4) *If on the contrary (a_1, a_2, b_1, b_2) are distinct and do not intertwine, then l_1 and l_2 do not intersect.*

This sets of axioms already allows us to define given any point x in $\mathbf{H}^2$ the involution ι_x of $\mathbf{P}(V)$, which exchange the end points of any line through x. We furthermore assume

(1) *for any x in $\mathbf{H}^2$, the involution ι_x preserves the cross-ratio in $\mathbf{P}(V)$.*

The set $\overline{\mathbf{H}^2}$ is the completed hyperbolic plane.

In this case, we use the following notation, we denoted the completed hyperbolic plane by $\overline{\mathbf{H}^2}$, its boundary at infinity by $\partial_\infty \mathbf{H}^2$, and the hyperbolic plane itself is $\mathbf{H}^2 = \overline{\mathbf{H}^2} \setminus \partial_\infty \mathbf{H}^2$.

We finally say

Definition 2.2. *A bijection F between two hyperbolic planes is an* isometry *if it extends to a map between the boundaries at infinity preserving the cross-ratio and end lines to lines.*[2]

2.1. The involution model. Let us consider $\mathbf{H}^2(V)$ to be the set of involutions of $\mathbf{P}(V)$ without any fixed points and preserving the cross-ratio. We will show that this defines a model of the hyperbolic space, that is a set for which the axioms are realized.

We will do that step by step, leaving the details for the reader, and using the interpretations of involutions as elements of $\mathsf{PGL}(V)$

(1) Given two distinct points X and Y in $\mathbf{P}(V)$, the line through x and y is the set of involutions i so that $i(X) = Y$.
(2) Given X in $\mathbf{P}(V)$ and y in $\mathbf{H}^2$, the line through x and y, is the line through X and $y(X)$.
(3) Given distinct involutions i and j, then $i \circ j$ is diagonalizable with eigenlines X and Y, then the *line* though i and j is the line through X and Y. Show that i an j belongs to L.

The less trivial exercise is to show

[2]It goes beyond the scope of these notes to show that, in order to be an isometry, it is enough to send geodesics to geodesics

Proposition 2.3. *Lines for the involution model satisfies axioms (3) and (4).*

From this construction, we obviously have

Proposition 2.4. *The restriction map of the action on* $\overline{\mathbf{H}^2}$ *to* $\mathbf{P}(V)$*, from the group of isometries of* $\mathbf{H}^2(V)$ *to* $\mathsf{PGL}(V)$ *is an isomorphism.*

Proof. Let F be map in $\mathsf{PGL}(V)$, then F extends to an isometry from $\mathbf{H}^2$ to itself by

$$i \mapsto F \circ i \circ F^{-1} .$$

In particular the map from the group of isometries to $\mathsf{PGL}(V)$ is surjective.

Let us prove this restriction map is injective. It is enough to show that an isometry F which restricts to the identity on the boundary is the identity. This follows at once from the axiom: any point x is the unique intersection of two lines L and D (why?). Since these lines are determined by their endpoints in the boundary at infinity, there are both globally invariant, thus the intersection $\{x\}$ is preserved. It follows that $F(x) = x$. □

We now sketch how every hyperbolic plane is isometric to the involution model. In particular

Theorem 2.5. *All hyperbolic planes are isometric. The group of isometries of a hyperbolic plane is isomorphic to* $\mathsf{PGL}_2(\mathbb{R})$.

Exercise 2.1:

(1) Let F be an involution without fixed points of $\mathbf{P}(V)$ preserving the cross-ratio. Show that there exists x in $\mathbf{H}^2$ so that $F = \iota_x$. *Hint* show that if a and b are distinct then $(a, b, F(a), F(b))$ are intertwined. Let x be the intersection of the lines with endpoints a and $F(a)$, and b and $F(b)$ respectively. Show that $i_x \circ F$ has 4 fixed points then that $i_x = F$.
(2) Let y and z two distinct points in a hyperbolic plane $\mathbf{H}^2$. Prove that $i_x \neq i_z$. *Hint:* let D be the geodesic through x and y. Let z be a point in the boundary at infinity of $\mathbf{H}^2$, different from the end points of D. Remark first that $i_x(z)$ is not one of the end points of D: otherwise z would be the other end point of D. Let Z be the geodesic though z and $i_x(z)$ and Y be the geodesic though z and $i_y(z)$. Observe that Z and Y are both different from D, and by counting the intersections with D that $Z \neq Y$. It follows that $i_x(z) \neq i_y(z)$ and hence that $i_x \neq i_y$.
(3) Use the the previous two exercises to show that there is a bijection from any hyperbolic plane with boundary at infinity $\mathbf{P}(V)$ to the set of involutions on $\mathbf{P}(V)$ without fixed points and preserving the cross-ratio.

We can finally prove using the involution model the following theorem.

2.1.1. *Reflexion through geodesics and orthogonality.* The following exercise and subsequent definition are important:

Exercise 2.2: Given any geodesic D, there exists a unique involution σ_D whose set of fixed points on $\mathbb{P}(V)$ is exactly the endpoints on D.

Definition 2.6. *The involution σ_D associated to a geodesic D in the previous exercise is called the* reflexion through *D.*

This allows us to have the following definition

Definition 2.7. *Two geodesics D an L are orthogonal if σ_D and σ_L commutes.*

We then have

Proposition 2.8. *The following are equivalent*

(1) *The geodesics D an L are orthogonal,*
(2) *The reflexion through D preserves L,*
(3) *The end points (a, b) of D and (c, d) of L form an* harmonic division*: $[a; b; c; d] = -1$.*

Moreover, given any point x in D, there exists a unique geodesic orthogonal to D, passing through x.

Proof. *Hint:* use coordinates so that $(a, b) = (0, \infty)$. Write the involution in these coordinates. For the last statement consider the isometry $i_x \circ \sigma_D$. □

2.2. Complex projective lines and the hyperbolic plane. Let us consider a totally real 2-plane P in in 2-dimensional vector space E. We furthermore choose an orientation on P, we say that a pairwise triple (x, y, z) of points in $C_P = \mathbf{P}(P)$ is *oriented* if a basis for which x, y and z have coordinates $1, 0$ and ∞ is oriented.

Proposition 2.9. *Let P be an oriented totally real 2-plane in a complex two-vector space. The* completed hyperbolic plane *is the set of complex lines D so that for any oriented triples (x, y, z) in C_P*

$$\Im([D; x; y; z]) \geqslant 0.$$

The hyperbolic plane $\mathbf{H}^2{}_P$ *associated to P, is the set of complex lines D so that for any oriented triples (x, y, z) in C_P*

$$\Im([D; x; y; z]) > 0.$$

A hyperbolic line *or* geodesic *is the intersection of a circle orthogonal to P. The* boundary at infinity *is C_P.*

We see that the subgroup $\mathsf{PSL}(P)$ preserves $\mathbf{H}^2{}_P$.

We will check in the next paragraph, by choosing a basis, that this data satisfies the axioms of a hyperbolic plane.

2.3. The Poincaré upper half space model. The *Poincaré upper half plane* model of the hyperbolic plane as the upper half plane in $\mathbb{C}$.

- The *completed hyperbolic plane* is
$$\overline{\mathbf{H}^2} = \{z \in \mathbb{C},\ \ \Im(z) \geqslant 0\} \cup \{\infty\}.$$
- The *hyperbolic plane* is
$$\mathbf{H}^2 = \{z \in \mathbb{C},\ \ \Im(z) > 0\}.$$
- The *boundary at infinity* is
$$\partial_\infty \mathbf{H}^2 = \{z \in \mathbb{C},\ \ \Im(z) = 0\} \cup \{\infty\} \sim \mathbf{P}(\mathbb{R}^2)\ .$$

Circles intersecting C_P orthogonally interpreted in the upper half plane model are either a circle whose center is on $\mathbb{R}$ or a line parallel to the imaginary axis. Indeed such a C is characterized by the following characterized by the following two properties

(1) for any quadruple of points (x, y, z, t) in L, then the cross-ratio $[x; y; z; t]$ is real
(2) C is globally invariant by the involution $z \mapsto \bar{z}$.

We define finally a *line* or *hyperbolic geodesic* as the intersection of C with $\overline{\mathbf{H}^2}$. Observe then that axioms (1)–(4) of definition 2.1 are easily satisfied.

Furthermore, the group $\mathsf{PSL}_2(\mathbb{R})$ acts on $\mathbf{H}^2$ is by *homographies*:
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} x = \frac{ax + b}{cx + d}\ .$$
One then check

Proposition 2.10. *The group* $\mathsf{PSL}_2(\mathbb{R})$ *acts transitively on* $\mathbf{H}^2$ *preserves the set of hyperbolic lines and the cross-ratio on* $\partial_\infty \mathbf{H}^2$.

Proof. Let us consider $\mathbf{P}(\mathbb{C}^2)$ as $\mathbb{C} \cup \infty$, where the coordinates are associated to a triple points (x_0, y_0, z_0) identified in these coordinates with $(1, 0, \infty)$. Then $\mathbf{H}^2$ corresponds to those points t, so that
$$\Im([t, x_0, y_0, z_0]) > 0\ .$$
It follows that $\mathsf{PSL}_2(\mathbb{R})$ acting by homographies on $\mathbf{P}(\mathbb{C}^2)$, preserves the complex cross-ratio on $\mathbf{P}(\mathbb{C}^2)$ as well as the real cross-ratio $\partial_\infty \mathbf{H}^2$, which is a subset of $\mathbf{P}(\mathbb{C}^2)$.

Finally $\mathbf{H}^2$ preserves the circles orthogonal to $\partial_\infty \mathbf{H}^2$, hence the geodesics. □

One also sees that the involution associated to the point i is given by the homography

$$z \mapsto -\frac{1}{z}\,.$$

We thus obtain the final axiom of a hyperbolic plane.

PROPOSITION 2.11. *The involution associated to any point x in $\mathbf{H}^2$ preserves the cross-ratio and has no fixed points.*

PROOF. *Hint:* Use an element of $\mathsf{PSL}_2(\mathbb{R})$ to reduce to the case $x = i$. □

EXERCISE 2.3:

(1) Show that the reflexion through the imaginary axis is given by $z \mapsto \frac{1}{\bar{z}}$.
(2) What are the geodesics orthogonal to the imaginary axis?
(3) Given a geodesic D, the set of fixed points of the reflexion σ_D in the completed hyperbolic plane is D.
(4) Show that given two non-intersecting geodesics D_0 and D_1, there exists a unique geodesic intersecting D_0 and D_1 orthogonally. *Hint:* consider the fixed points of $\sigma_{D_0} \circ \sigma_{D_1}$.

Remark that a reflexion is not an homography.

EXERCISE 2.4: Describe the action of $\mathsf{PGL}_2(\mathbb{R})$ on the upper half plane model.

EXERCISE 2.5: Identify the upper-half plane model as a complex projective model: Let us consider a totally real oriented 2-plane P in a complex vector space E. Let us fix three pairwise distinct points (x, y, z) in C_P whose coordinates are $(0, 1, \infty)$. We can now identify $\mathbf{P}(\mathbb{C}^2)$ with $\mathbb{C} \sqcup \{\infty\}$. Show that the upper half model is $\mathbf{H}^2{}_P$.

EXERCISE 2.6: Show that the group $\mathsf{PSL}_2(\mathbb{R})$ is generated as homographies by the involution $z \mapsto -\frac{1}{z}$, translations $z \mapsto z + a$, and hyperbolic elements $z \mapsto \lambda z$, with a and λ real.

3. More geometric features: distances, angles and convex polygons

Now that we have efficient models for hyperbolic planes, we can introduce new concepts and figures.

3.1. Distances and angles. We first define distances between non-intersecting geodesics and then define distances between points. We will later prove that. the latter is indeed a distance.

3.1.1. *Distances between non-intersecting geodesics.* Let D and L be non-intersecting geodesics, with end points (a, b) and (c, d) respectively. We then define

$$d(D, L) := 2\,|\operatorname{arctanh}([a; b; c; d])|\,.$$

Exercise 3.1:

(1) Show that one can always choose coordinates so that $\{a, b\} = \{-1, 1\}$, $\{c, d\} = \{-\alpha, \alpha\}$. Then show that

$$d(D, L) = \left|\log(\alpha)\right|\,.$$

(2) Relate the distance between D and L to the trace of $\sigma_D \circ \sigma_L$.

3.1.2. *Distances between points.* We define the *distance* between two points x and y in the following way. Let D be the geodesic passing through x and y, let L_x the geodesic through x orthogonal to D, and L_y the geodesic through y orthogonal to D, then we define

$$d(x, y) := d(L_x, L_y)\,.$$

It is far from clear that this is a distance, we shall prove that later in some more economic way. For the moment we admit that this is indeed a distance. in the upper half plane model, we see that for $\alpha > \beta > 1$

$$d(i, \alpha i) = \log(\alpha)\,,\ d(\beta i, \alpha i) = \log\left(\frac{\alpha}{\beta}\right)$$

By construction, the distance is invariant under $\mathsf{PSL}(V)$

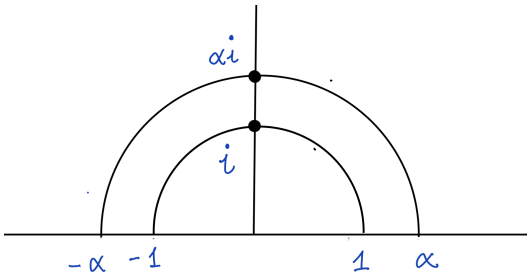


Figure 2. distance between points on the imaginary axis

Exercise 3.2: Show that in the upper half plane model, if D is the geodesic through x an y, if X and Y are endpoints of D, then

$$d(x, y) = |\log([x, y, X, Y])|\,.$$

Hint: reduce to a standard situation using homographies.

EXERCISE 3.3: Relate the distance between x and y, associated to the involutions ι_x and ι_y to the trace of $\iota_x \circ \iota_y$.

3.1.3. *Geodesic arc.* Let x and y be two distinct points in $\mathbf{H}^2$ and D the geodesic through x and y.

DEFINITION 3.1. *The geodesic arc through* $[x, y]$ *is the subset of* D*, given by*

$$[x, y] = \{z \in D \mid d(z, x) + d(z, y) = d(x, y) \ .$$

EXERCISE 3.4: Draw in figure 2, the geodesic arc between i and αi

3.1.4. *Angles between geodesics.* We define an *orientation* on a geodesic D as the choice of the pair of endpoints of D. We observe that the angle between the two geodesics at the intersection point make sense in the Euclidean sense. This angle is invariant under homographies (which are holomorphic mapping) hence makes sense for any hyperbolic space.

However we have a more intrinsic point of view

EXERCISE 3.5: Relate the angle between intersecting geodesic D and L to the trace of $\sigma_D \circ \sigma_L$. Furthermore, obtain a formula relating the angle between between D and L to the cross-ration of the end points.

EXERCISE 3.6: Prove that two distinct geodesics D_0 and D_1 are orthogonal, if and only if D_1 is globally invariant by σ_{D_0}.

EXERCISE 3.7: Let x and y be two distinct points in $\mathbf{H}^2$, show that the only non-trivial isometry fixing x and y is the reflexion σ through the geodesic passing through x and y.

3.2. Circles, horocycles. In the upper half plane model, *Hyperbolic circles* – later on called simply circles – are the circles in our geometric model that do not intersect the boundary at infinity, *horocycles* are circles that intersect the boundary at infinity at exactly one point.

Both circles and horocycles are invariant by $\mathsf{PSL}_2(\mathbb{R})$, it follows that circles and horocycles are defined for every hyperbolic space.

EXERCISE 3.8:

(1) Show that circles are exactly the orbits of the subgroups of $\mathsf{PSL}(V)$ stabilizing a point in $\mathbf{H}^2$, while horocycle are orbit of subgroup stabilizing a point in $\partial_\infty \mathbf{H}^2$.

(2) Show that given any x in $\mathbf{H}^2$ and positive real R , then

$$C := \{z \in \mathbf{H}^2 \mid d(x,z) = R\} ,$$

is a circle.

3.3. Convex polygons. A *half space* is a complementary region to a geodesic.

A *wedge* is one of the complementary region of two oriented geodesic intersecting orthogonally.

A *convex polygon* is the intersection of half spaces. Among them are triangles, hexagons etc ... two points in a convex polygons are joined by a geodesic arc inside this polygon

If P is a non-compact convex polygon, we define $\partial_\infty P$, as the set of points y in ∂_∞ such that there exists a geodesic arc $c : [0, \infty[\to P$ such that $c(\infty) = y$.

Exercise 3.9: Let P be a convex polygon

(1) Assume P invariant by an isometry γ then any fixed point of γ in $\partial_\infty \mathbf{H}^2$ belongs to $\partial_\infty P$.
(2) If $P \neq \mathbf{H}^2$, then $\partial_\infty P \neq \partial_\infty \mathbf{H}^2$.

3.4. Triangles and ideal triangles. An *ideal triangle* is a triangle with three points at infinity, a *2/3-ideal triangle* has two points at infinity and a *1/3-ideal triangle* has one. All ideal triangles are congruent meaning that there is an isometry sending one to another/.

Proposition 3.2. *Given any positive numbers a, b and c satisfying the triangles inequalities there exists a unique triangle – up to the action of $\mathsf{PGL}_2(\mathbb{R})$ – whose length are a, b and c.*

Proof. Using the isometry group, let us take $x = i$ and $y = \alpha i$, with $a = \log(\alpha)$. Let us now consider

$$C_1 = \{z \mid d(z,i) = b\} , \; C_2 = \{z \mid d(z,\alpha_i) = c\} .$$

We showed in an exercise above that these are two (euclidean) circles. By the inequalities satisfied by a, b and c, these two circles intersect and, since they are circles, they intersect in precisely two points which are reflection of each other through the vertical axis. Let z be one of these points then the triangle (x, y, z) satisfies

$$d(x,y) = a \; , \; d(x,z) = b \; , \; d(y,z) = c \; .$$

We leave the reader check the uniqueness. □

PROPOSITION 3.3. *Let A and B be two subsets of the hyperbolic plane. Assume A is not a subset of a geodesic. Assume that there exists a distance preserving map φ from A to B. Then there exists a unique isometry Φ of $\mathbf{H}^2$ such that Φ restricted to X is φ.*

PROOF. We can use the previous proposition to reduce to the case when φ fixes 3 points (x, y, z) not on a geodesic. Let Z be the geodesic through (x, y) and similarly X and Y. We then want to show that φ is the identity. Let t be a fourth point in A distinct from x, y or z. We can always assume that t does not belong to Y or Z. Then considering the triangles (x, z, t), (x, y, t) and (y, z, t) we see that

$$\varphi(t) = t \text{ or } \varphi(t) = \sigma_Z(t) ,$$
$$\varphi(t) = t \text{ or } \varphi(t) = \sigma_Y(t) ,$$

Assume that $\varphi(t) = \sigma_Z(t) \neq t$, then $\sigma_Z(t) = \sigma_Y(t)$. Thus t is fixed by $\sigma_Z \circ \sigma_Y$, but the only fixed point of $\sigma_Z \circ \sigma_Y$ is x, hence the contradiction. □

EXERCISE 3.10: What happens when A is a subset of a geodesic?

3.5. Right-angled hexagons.

PROPOSITION 3.4. *Given any positive number a, b and c, there exists a unique right-angled hexagon – up to the action of $\mathsf{PGL}(2, \mathbb{R})$ – whose length of non-intersecting edges are a, b and c.*

PROOF. Let D_0, D_1 and D_2 the orthogonal geodesics to the side, such that $d(D_0, D_1) = a$, $d(D_0, D_2) = b$, $d(D_1, D_2) = c$. Let (a_i, b_i) be the endpoints of D_i. We can reduce to the situation where $(a_0, b_0) = (-1, 1)$, $(a_1, b_1) = (-\alpha, \alpha)$ with $\alpha > 1$ and α a function of a, $(a_1, b_1) = (x, y)$, with

$$1 < x < y < \alpha .$$

Our goal is to show, given β and γ, we can find x and y uniquely so that

$$[-1; 1; x; y] = \beta ,$$
$$[-\alpha; \alpha; x, y] = \gamma .$$

This gives the equations

$$(x + 1)(y - 1) = \beta(x - 1)(y + 1) ,$$
$$(x + \alpha)(y - \alpha) = \gamma(x - \alpha)(y + \alpha) .$$

These equations give a set of two affine equations in $X := x - y$ and $Y := xy$. One can show that there is a unique solution X and Y, thus a unique solution in x and y within our range. □

3.6. Regular polygons. We will not use the following proposition but just state it for cultural reasons related to the tilings by Escher:

PROPOSITION 3.5. *Given any integer $n > 4$, there exist a unique – up to the action of $\mathsf{PSL}_2(\mathbb{R})$ – regular right-angled n-gon.*

PROOF. *Hint:* a continuity argument. □

4. The Riemannian interpretation: length and area

4.1. The length of a curve. We define the length of a parametrized curve $c = (x, y) : [a, b] \to \mathbf{H}^2$ in the hyperbolic plane in the upper half model as

$$\ell(c) = \int_a^b \frac{\sqrt{\dot{x}^2 + \dot{y}^2}}{y} \, \mathrm{d}t.$$

Then, the following facts is true

PROPOSITION 4.1. *The length of the curves is invariant under the action $\mathsf{PSL}_2(\mathbb{R})$: if c is a curve and g an element of $\mathsf{PSL}_2(\mathbb{R})$ then $\ell(g(c)) = \ell(c)$.*

PROOF. The length of the curves is obviously invariant by hyperbolic elements and translation. A simple computation shows that if φ is the involution $z \mapsto -\frac{1}{z}$, then $\ell(\varphi(c)) = \ell(c)$. Then we can conclude using Exercise 2.6. □

4.2. Geodesics minimize length. Recall that we defined an invariant "distance" (without checking the triangular inequality) satisfying two properties

(1) $d(i, \alpha i) = |\log(\alpha)|$
(2) For g in $\mathsf{PGL}_2(\mathbb{R})$, $d(gx, gy) = d(x, y)$

We are going relate the distance to the length of curves, more precisely we prove

PROPOSITION 4.2 (GEODESICS MINIMIZE LENGTH). *Let x and y be two points in $\mathbf{H}^2$, and c a C^1-piecewise curve joining x to y, then*

$$\ell(c) \geqslant d(x, y) \, ,$$

with equality if and only if c is a parametrization of the geodesic arc between x and y.

We deduce immediately a relation between distance and length

COROLLARY 4.3. *The distance between two points is*

$$d(x, y) = \inf\{\ell(c) \mid c \text{ joins } x \text{ to } y\}.$$

A corollary that immediately implies

COROLLARY 4.4. *The distance that we defined in paragraph 3.1.2 satisfy the triangular inequality.*

PROOF OF PROPOSITION 4.2. Using the invariance under $\mathsf{PGL}_2(\mathbb{R})$ it is enough to check the case when $x = i$ and $y = \alpha i$, with $\alpha > 1$.

In that case, the result follows from the following string of inequalities, for a curve $c : [a, b] \to \mathbf{H}^2$, with $c(a) = i$ and $c(b) = \alpha i$

$$\begin{aligned}\ell(c) = \int_a^b \frac{\sqrt{\dot{x}^2(t) + \dot{y}^2(t)}}{y(t)}\,\mathrm{d}t &\geqslant \int_a^b \frac{|\dot{y}(t)|}{y(t)}\,\mathrm{d}t \\ &\geqslant \left|\int_a^b \frac{\dot{y}(t)}{y(t)}\,\mathrm{d}t\right| = [\log(y(t))]_a^b = \log(\alpha)\,.\end{aligned}$$

Tracking the equality case in the above inequalities give you the equality case in the proposition. □

4.3. The boundary at infinity. Finally, one recover the boundary at infinity from this picture. We say two oriented geodesics are *asymptotic* if given two arc lengths parametrization of these geodesics $t \mapsto \gamma_0(t)$ and $t \mapsto \gamma_1(t)$ then

$$\limsup_{t\to+\infty} (d(\gamma_0(t), \gamma_1(t)) < \infty.$$

Then we have the following characterization.

PROPOSITION 4.5. *The following are equivalent*

(1) *The oriented geodesics γ_0 and γ_1 are asymptotic*
(2) *The two geodesics γ_0 and γ_1 have the same endpoint (in the future)*
(3) *There exists a parametrization of γ_0 and γ_1 so that*

$$\limsup_{t\to+\infty} (d(\gamma_0(t), \gamma_1(t)) = 0\,.$$

PROOF. *Hint:* do explicit computation when γ_0 is a vertical geodesic going to infinity. □

Then two oriented geodesics are asymptotic precisely if they have the same end point at $+\infty$.

4.4. Area and the Gauß-Bonnet formula. The *hyperbolic area* of a measurable set A in Poincaré upper-half plane model is

$$\mathrm{Area}(A) = \int_A \frac{1}{y^2}\,\mathrm{d}x\mathrm{d}y.$$

PROPOSITION 4.6. *The hyperbolic area is invariant under isometries.*

PROOF. This is an exercise on the change of variable formula for the involution $z \mapsto -\frac{1}{z}$. Then it is obvious for the translations $z \mapsto z + a$ and hyperbolic isometries $z \mapsto \lambda z$. Finally, conclude using Exercise 2.6. □

Here is another exercise that explains another relation with the distance.

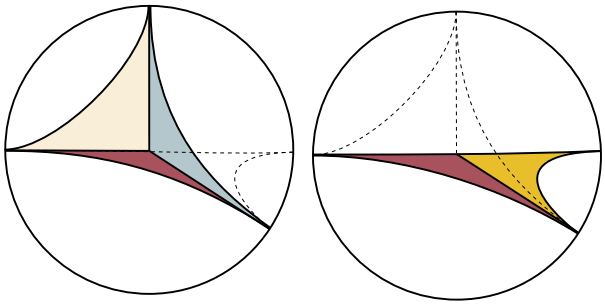


FIGURE 3. Gauss additivity

Our main result is very famous. If T is a triangle, the *angle defect* of T is π minus the sum of its internal angles at the vertices. In Euclidean geometry the angle defect is zero. For hyperbolic geometry, we have:

THEOREM 4.7 (GAUSS–BONNET FORMULA). *The area of a triangle equal its* angle defect.

It turns out that both the Euclidean and hyperbolic result are compatible when one takes in account the curvature of the space. We will not define the curvature, but remark it could be define using the angle defect.

PROOF. We shall follow Gauss approach in several steps

(1) We first check by a direct computation that the area of an ideal triangle is π.
(2) Then let $A(\theta)$ the area of a 2/3-ideal triangle with angle $\pi - \theta$, with θ in $[0, \pi/2]$. Gauss observation is $A(\theta)$ is an additive monotone function,

hence a multiple of θ, hence θ by the normalization of the ideal triangle: $A(\pi) = \pi$. More precisely we show that if $\theta + \pi < \pi$, then

$$A(\theta + \pi) = A(\theta) + A(\pi) .$$

This additivity is done geometrically when $\theta + \pi < \pi$ as follows in Figure 3: the area of the union of the pink and blue triangles – with internal angles $\pi - \theta$ and $\pi - \eta$ respectively – is the area of yellow triangle, since they are both π minus the area of the purple triangle. On the other hand, the internal angle of the yellow triangle is

$$\pi - (2\pi - (\pi - \theta) - (\pi - \eta)) = \pi - (\theta + \eta) .$$

(3) The rest follows: we see a 1/3 ideal triangle as the difference between two 2/3 ideal triangles. Finally, a finite triangle as the difference between one 2/3 ideal triangle and the sum of two 1/3 ideal triangles.

□

EXERCISE 4.1: Show that the area of a right-angled hexagons is 2π and that of of regular right-angled n-gon is $\frac{\pi}{2}n$.

EXERCISE 4.2: (*) Define a measure just using the angle defect: check which properties the angle defect should have to prove them.

5. The Poincaré disk model

We conclude this chapter by describing quickly the Poincaré Disk Model, which is helpful for computations related to balls.

DEFINITION 5.1. *The* Poincaré Disk Model *is given by the disk in* $\mathbb{R}^2$ *of center 0 and radius 1. The* boundary at infinity *is the boundary of the disk with its projective cross-ratio which is the restriction of the complex cross-ratio, the geodesics are circles or lines orthogonal to the boundary.*

The Poincaré Disk Model is obtained after taking a complex homography sending i to 0, 0 to -1 and ∞ to 1, and the real axis to the circle of radius 1 around 0:

$$\psi(z) = \frac{z - i}{z + i} .$$

It follows by Exercise 2.5 that the Poincaré Disk Model satisfies the axioms of the hyperbolic plane. We now define the length of a curve $c(t)$ in the Poincaré disk as

$$\ell_0(c) := \int_a^b \frac{\sqrt{\dot{x}^2 + \dot{y}^2}}{1 - r^2} \mathrm{d}t ,$$

where $r^2 = c^2 + y^2$. A tedious computation then gives

PROPOSITION 5.2. . *The hyperbolic length of a curve c in Poincaré Upper Half Plane Model is the length of its image in the Poincaré Disk Model:*

$$\ell(c) = \ell_0(\psi(c)) \, .$$

PROOF. An alternate proof not using computations, is to show this equality for all horizontal curves through 0, then for any geodesics, then for any curves. □

We then have

COROLLARY 5.3. *Balls in the Poincaré Upper Half Plane Model, or in the Poincaré Disk Model are balls for the Euclidean metric.*

PROOF. From the proposition just before and the rotational invariance of the length, we deduce that the hyperbolic ball around 0 are Euclidean balls. It follows that the hyperbolic balls around i in the upper half plane model are Euclidean balls (with a different center though), since the homography ψ send circle to circles. Thus any hyperbolic ball in the upper half plane model is a Euclidean ball – using the invariance of circles by homographies, and thus the same holds in the Poincaré Disk Model □

A similar computation show that

PROPOSITION 5.4. . *We have*

$$\text{Area}(A) = \int_{\psi(A)} \frac{1}{(1-r^2)^2} \, \mathrm{d}x\mathrm{d}y \, .$$

This leads to the following exercise:

EXERCISE 5.1:

(1) We will need later on an explicit computation: show that

$$\text{Area } B(x,R) = 4\pi \sinh^2\left(\frac{R}{2}\right) , \tag{2}$$

Hint: use Proposition 5.4.

(2) Let x be a point in $\mathbf{H}^2$, R a positive number and $B_x(R)$ the ball of radius R with center x, with respect to the hyperbolic metric. Then

$$\text{Area } B(x,R) \underset{R\to 0}{\sim} \pi R^2 \, , \tag{3}$$

$$\text{Area } B(x,R) \underset{R\to \infty}{\sim} 4\pi e^R \, . \tag{4}$$

(3) Here is a related computation. Let $C(x, R)$ the circle which is the boundary of the ball, then

$$\ell\left(C(x,R)\right) = 2\pi \sinh(R) = \frac{\mathrm{d}}{\mathrm{d}R}\text{Area } B(x,R)\ .$$

Hint: use Proposition 5.2 for the first equality.

EXERCISE 5.2: What is the Euclidean center of the hyperbolic ball centered at i of radius R?

6. Comments, references and further reading

CHAPTER 3

Hyperbolic surfaces

We recall that a map between metric spaces is an *isometry* of it preserves the distance, a map φ is a *local isometry* if for every point x in the source, there exists a neighborhood U of x, such that φ is an isometry from U to $\varphi(U)$.

1. Hyperbolic surfaces and length

1.1. Hyperbolic surfaces. A *hyperbolic surface* is a metric space M such that every point in M has a neighborhood isometric to an open set of the hyperbolic plane.

A *hyperbolic surface with totally geodesic boundary* is a metric space M such that every point in M has a neighborhood isometric to an open set of the hyperbolic plane, or an hyperbolic half plane.

A *hyperbolic surface with totally geodesic boundary and right-angles* is a metric space M such that every point in M has a neighborhood isometric to an open set of the hyperbolic plane, or an hyperbolic half plane, or a right-angled wedge.

To avoid repetition, we call a ball in all cases a ball *in the model*. We immediately have,

PROPOSITION 1.1. *Let x be a point in a hyperbolic surface S, with boundary. Assume there exists a distance preserving map φ from a neighborhood of x to a ball in a wedge such that $\varphi(x)$ is in the boundary of $\varphi(U)$ or a corner. Then, for any a distance preserving map ψ from a neighborhood of x to a ball in a wedge, then $\varphi(x)$ is in the boundary of $\varphi(U)$ or a corner.*

The *boundary* of a hyperbolic surface is the set of points in S satisfying the condition of the previous proposition. The boundary of a hyperbolic surface is made of union of

(1) open geodesic segments,
(2) points (maps in to corner of the wedge)

A hyperbolic surface is *closed* if it has no boundary and is compact.

A *geodesic* in a hyperbolic surface is a curve $c = [a, b]$ to S, so that for every t in $[a, b]$, there exists a neighborhood U of $c(t)$, some positive ε, an isometry

φ from U to an open set in the hyperbolic plane, so that $\varphi([t-\varepsilon, t+\varepsilon]$ is a geodesic arc.

1.2. The hyperbolic length metric. Given a hyperbolic surface (S, d) with a metric d, we can find a better metric on it. Let $c : [a, b] \to S$ be a curve in S, we can first define its *hyperbolic length* $\ell(c)$ as follows. We first find a subdivision

$$a = t_0 < t_1 < \ldots < t_n = b,$$

so that $c[t_i, t_{i+1}] \subset B_i$, where B_i is a ball for d isometric (by a map φ_i) to a hyperbolic ball. Then we define

$$\ell(c) = \sum_{i=0}^{n-1} \ell\left(\varphi_i \circ c|_{[t_i, t_{i+1}]}\right).$$

Exercise 1.1: Prove the following tedious facts

(1) The length $\ell(c)$ does not depend on the subdivision of $[a, b]$.
(2) The length $\ell(c)$ does not depend on the parametrization of c : if φ is a diffeomorphism from $[a, b]$ to $[c, d]$ then $\ell(c) = \ell(c \circ \varphi)$.
(3) If d and d' are two locally isometric metrics on S, both locally isometric to the hyperbolic plane. Then the length for d and the length for d' are equal.

This length allows us to define a new metric on S. For any x and y on S we define the *Riemannian distance* on S by

$$d(x, y) = \inf\{\ell(c) \mid c : [0, 1] \to S,\ \ c(0) = x,\ c(1) = y\}.$$

One now has the following proposition

Proposition 1.2. *The Riemannian distance is a distance on the hyperbolic surface (S, d) which is moreover locally isometric to d. Finally two locally isometric d and d' on S generates the same Riemannian distance.*

From now on, we shall always equip a hyperbolic surface with its Riemannian distance. We finally define: a curve $c : [a, b] \to S$ is *parametrized by arc length* if for any s and t in $[a, b]$,

$$\ell(c|_[s, t]) = |t - s|$$

We then say a curve $c : [a, b] \to S$ is minimizing the length if

$$d(c(a), c(b)) = |a - b|\,.$$

Proposition 1.3. *Let $c : [a, b] \to S$ be a curved parametrized by arc length, then for all s and t in $[a, b]$, we have the* arc-length inequality.

$$d(c(s), c(t)) \leqslant |t - s| \ ,$$

Moreover, if $d(c(s), c(t)) = |t - s|$, then for all u and v in $[t, s]$, we have

$$d(c(u), c(v)) = |u - v| \ .$$

Proof. The first property is an immediate consequence of the definition of the Riemannian distance and the arc length parametrization, while the last one comes from the triangular inequality: it is enough to prove that if $s < u < t$,

$$d(c(u), c(t)) = t - u \ .$$

Now

$$d(c(u), c(t)) \geqslant d(c(t), c(s) - d(c(s), c(u)) \geqslant t - s - (u - s) = t - u \ .$$

Thus the arc-length inequality gives the result □

1.3. Local isometries. Recall that $\varphi : X \to Y$ between two metric spaces is a local isometry, if for every x in X there exists R so that Φ is an isometry from $B(x, R)$ to $B(\varphi(x), R)$, and in particular $\ell(c) = \ell(\Phi(c)$ for any curve. This in particular implies that of X and Y are length spaces, then

$$d(\varphi(x), \varphi(y)) \leqslant d(x, y) \ . \tag{5}$$

We deduce the following proposition

Proposition 1.4 (Local–global). *If φ is a local isometry between two length spaces which is a bijection, then φ preserves the distances.*

Proof. Indeed, φ^{-1} is also a local isometry and thus the reverse inequality to inequality 5 is also satisfied. □

1.4. Geodesics and extension of geodesics. A *geodesic* is a hyperbolic surface (S, d) is a curve

$$\gamma :]a, b[\to S \ ,$$

such that if B is a ball in S, isometric by φ to a ball in the model, then $\varphi(\gamma|_B)$ is a geodesic arc. From the similar property in the hyperbolic plane, we have the following characterization

Proposition 1.5 (Geodesics locally minimize distance). *A curve $\gamma =]a, b[\to S$, is geodesic if an only the following is true. For any t in $]a, b[$, there exist a positive ε, such that if u and s belong to $]t - \varepsilon, t + \varepsilon[$, we obviously have*

$$d(\gamma(u), \gamma(s)) = |s - u| \ .$$

In particular, for any s and u as in the proposition, for any curve c joining s to u, then

$$\ell(c) \geqslant |s-u| = \ell(\gamma|_{[s,u]}) .$$

We also have

PROPOSITION 1.6 (BEYOND A POINT). *Let γ be a geodesic defined on $]b,a[= I_0$. Assume that there exists a sequence $\{t\}_{m\in\mathbb{N}}$ converging to a so that $\{\gamma(t_m)\}_{m\in\mathbb{N}}$ converges to x in S. Then*

Assume first the boundary of S is empty. Then there exists a geodesic γ_0 defined on $]b,c[$ with $c > a$, which coincides with γ_0 on I_0.

In the general case,

1. *if γ is included in the boundary, then either x is not a corner, in which case γ can be extended to $]b,c[$ with $c > a$, or x is a corner then γ can be extended to $]b,a]$.*
2. *if γ is not included in the boundary, then either x is not in the boundary, in which case γ can be extended to $b,c[$ with $c > a$, or x is in the boundary then γ can be extended to $]b,a]$.*

PROOF. Let x be the limit of $\{\gamma(t_m)\}_{m\in\mathbb{N}}$. Let B be a ball round x of radius R, isometric to a hyperbolic ball. By the distance construction of the distance, that $c(]a-R/4,a[)$ lies in the ball of radius R: indeed for s in $]a-R/4,a[$, there exists n as large as we want, with $d(c(s),c(t_n)) \leqslant R/4$. Now the proposition follows from the similar property in the hyperbolic plane.

□

Similarly, we have

PROPOSITION 1.7 (EXTENSION OF GEODESICS). *Let $\gamma_0 : I_0 \to S$ and $\gamma_1 : I_1 \to S$ be two geodesics, where I_i are intervals. Assume there exists a an open interval I in $I_0 \cap I_1$ such that*

$$\gamma_0|_I = \gamma_1|_I ,$$

Then, there exists a geodesic γ defined on $I = I_0 \cup I_1$, such that

$$\gamma|_{I_0} = \gamma_0 , \ \gamma|_{I_1} = \gamma_1 .$$

This proposition allows us to make the following definition: a geodesic $\gamma : I \to S$ is *maximal* if for any geodesic $\gamma_0 : I_0 \to S$ which coincides with γ on a subinterval of $I_0 \cap J$, then I_0 is included on in I. [1]

As an immediate consequence of Proposition 1.7, we have

PROPOSITION 1.8. *Any geodesic can be extended to a maximal geodesic. Such a maximal geodesic is unique.*

[1] In this definition I is no required to be open.

1.5. Completeness. A hyperbolic surface is *complete* if, given any maximal geodesic arc γ defined on I, then if $I =]-\infty, +\infty[$.

In other words, thanks to Proposition 1.8, every geodesic arc can be extended to $]-\infty, \infty[$.

For a hyperbolic surface S with boundary and right-angles, the definition of completeness is slightly different: S is *complete* if given any maximal geodesic arc γ defined on I, then

(1) either $\gamma(I)$ is included in the boundary, in which case if a is a finite extremity of I, then $\gamma(a)$ is a corner.
(2) or there exists a point in γ not in the boundary, then for any finite extremity a of I, we can extend γ, with $\gamma(a)$ in the boundary.

Exercise 1.2: Show that a convex polygon (with right-angles) is complete.

Observe that we have as an immediate consequence of Proposition 1.6:

Proposition 1.9. *Any compact surface is complete.*

We now prove the following result, which is a special case of the important Hopf–Rinow Theorem in Riemannian geometry and which have the same proof

Theorem 1.10 (Hopf–Rinow Theorem). *Any two points x and y in a complete hyperbolic surface can be joined by a geodesic γ such furthermore*

$$\ell(\gamma) = d(x, y) .$$

We will first prove a Lemma, valid for non-complete surfaces.

Lemma 1.11 (Aiming). *Let x and y be two points in a hyperbolic surface. Let R so that $B(x, 2R)$ is isometric to a ball in the model. Let*

$$S_x = \{w \mid d(x, w) = R .$$

Then there exists w in S so that

$$d(x, y) = d(x, w) + d(w, z) = R + d(w, z) .$$

Proof. Let $\{c\}_{m\in\mathbb{N}}$ be a sequence of curves from $[0, 1]$ to S joining x to y such that $\{\ell(c_m)\}_{m\in\mathbb{N}}$ converges to $d(x, y)$. Observe that by the intermediate value theorem, that for all m,there exists t_m, with $w_m := c_m(t_m)$ in S_x. We extract a subsequence w_m converging to some w in S. We now show that w have the required property.

Indeed, from the definition of the Riemannian distance

$$\ell(c_m) \geqslant d(x, w_m) + d(w_m, z) .$$

Taking the limit when m goes to infinity, gives the inequality

$$d(x,z) \geqslant d(x,w) + d(w,z)\ ,$$

which combined with the triangular inequality gives the result. □

Proof of Theorem 1.10. We use the Aiming Lemma 1.11 to find w in S_x with

$$d(x,y) = d(x,w) + d(w,y)\ .$$

Let then $\gamma : I \to S$ be the maximal geodesic with $\gamma(R) = w$. To simplify the argument, let us first assume that S has no boundary in which case the completeness assumption means that $I =]-\infty, \infty[$.

We then consider

$$U = \{t \in I \mid d(x,y) = t + d(\gamma(t), y)\ .$$

By the definition of γ, U is not empty: U contains R. Let us start with the sequence of observations that follows from the triangular inequality,

(1) If t is in U, then $d(\gamma(t), x) = t$. Indeed, since $t \geqslant d(\gamma(t), x)$, we have

$$d(x,y) \geqslant d(x,\gamma(t)) + d(\gamma(t), y)\ ,$$

Then the triangular inequality implies that we actually have the equality

$$d(x,y) = d(x,\gamma(t)) + d(\gamma(t), y)\ ,$$

and thus $t = d(x, \gamma(t))$.

(2) It then follows that for $s \leqslant t$ with t in U, we have by Proposition 1.3

$$d(\gamma(s), \gamma(t)) = t - s\ .$$

(3) if t belongs to U, then $[0,t]$ is included in U: indeed for s smaller than t,

$$\begin{aligned} d(x,y) = t + d(\gamma(t), y) &= s + (t-s) + d(\gamma(t), y) \\ &\geqslant d(\gamma(s), x) + t - s + d(\gamma(t), y) \\ &\geqslant d(\gamma(s), x) + d(\gamma(s), \gamma(t)) + d(\gamma(t), y) \\ &\geqslant d(\gamma(s), x) + d(\gamma(s), x) \\ &\geqslant d(x,y) \end{aligned}$$

Thus all inequalities above are equalities, then $d(\gamma(s), x) = s$ and $d(x,y) = s + d(\gamma(s), x)$

We then consider $t_0 = \sup\{t \in U \mid t \leqslant d(x,y)\}$. We want to prove that $t_0 = d(x,y)$. It is enough to prove that $U = [0, d(x,y]$, and thus that U is open, since U is obviously closed and non-empty by construction.

Let then t in U, with $t < d(y,z)$, $z = \gamma(t)$, which then satisfies

$$d(x,y) = d(z,y) + d(z,x) = t + d(z,y)\ ,$$

We can use again the Aiming Lemma, and obtain $z_1 = \gamma_1(R_1)$, for some geodesic γ_1 such that

$$d(z,y) = R_1 + d(z_1,y) = d(z_1,z) + d(z_1,y)\ ,$$

It follows that

$$d(x,z_1) \geqslant d(x,y) - d(y,z_1) = d(x,z) + d(z,y) - d(z,y) + d(z,z_1) = d(x,z) + d(z,z_1) =$$

Thus using the triangular equality,

$$d(x,z_1) = d(x,z) + d(z,z_1) = t + R_1\ .$$

Let c be the curve defined on $[0, t+R_1]$ such that $c|[0,t] = \gamma$, and for s in $[t, t+R_1]$ $c(s) = \gamma_1(s - R_1)$. Then we have

$$\ell(c) = t + R_1 = d(x,z) + d(z,z_1) = d(x,z_1)\ ,$$

Hence c is length minimizing hence c is a geodesic, coinciding with γ, hence equal to γ. It follows that $z_1 = \gamma(t + R_1)$, hence

$$d(y,\gamma(t+R_1))+t+R_1 = d(y,z_1)+d(z,x)+(d(y,z)-d(y,z_1)) = d(y,z)+d(z,x) = d(y,x)\,.$$

It follows that $t + R_1$ belongs to U and the result follows.

It remains to treat the case of surface with boundary. But a careful reader can see that the above arguments also work in that case.

□

2. Two constructions of hyperbolic surfaces

2.1. Construction of hyperbolic surfaces by quotient. Let Γ be a subgroup of the group $\mathrm{Iso}(\mathbf{H}^2)$ of isometries of an hyperbolic plane $\mathbf{H}^2$. We say that Γ acts *freely and properly discontinuously* on $\mathbf{H}^2$ if for every point x in $\mathbf{H}^2$, there exists a neighborhood U of x so that

$$\{\gamma \in \Gamma \mid \gamma(U) \cap U \neq \emptyset\} = \emptyset\ .$$

Let consider then the map $\pi = \mathbf{H}^2 \to S_\Gamma := \Gamma\backslash\mathbf{H}^2$ This is an important way to construct hyperbolic surfaces. Let us define a metric on S_Γ by

$$d_\Gamma(x,y) := \inf\{d(\bar{x},\bar{y}) \mid \pi(\bar{x}) = x\ ,\ \pi(\bar{y}) = y\}$$

THEOREM 2.1 (QUOTIENTS ARE HYPERBOLIC SURFACES). *Let Γ be a subgroup of* $\mathrm{Iso}(\mathbf{H}^2)$ *acting freely and properly discontinuously on* $\mathbf{H}^2$. *Then d_Γ gives S_Γ the structure of a complete hyperbolic surface. Moreover*

$$\pi : \mathbf{H}^2 \to S_\Gamma ,$$

is a locally isometric map.

PROOF. It is enough to define the distance, we define unambiguously

$$d_\Gamma(\pi(x), \pi(y)) = \inf\{d(\gamma x, \eta y) \mid \gamma,\ \eta \in \Gamma\} .$$

Observe that d_Γ satisfies the triangular inequality, and that moreover

$$d_\Gamma(\pi(x), \pi(y)) \leqslant d(x, y)$$

The only point to prove now is to show that, given x in $\mathbf{H}^2$, there exists an r so that π d an isometry from

$$B_x(r) \coloneqq \{z \in \mathbf{H}^2 \mid d(z, x)) < r\}$$

to

$$B_{\pi(x)}(r) \coloneqq \{z \in S_\Gamma \mid d_\Gamma(z, \pi(x))) < r\} .$$

Observe that by inequality (**??**),

$$\pi(B_x(r)) \subset B_{\pi(x)}(r) .$$

Let x in $\mathbf{H}^2$, by the condition on Γ, there is a positive R, so that for all y in $B_x(R)$, γ in $\Gamma \setminus \{\mathrm{id}\}$, then

$$\gamma(y) \notin B_x(R) .$$

In particular $\varphi : y \mapsto \pi(y)$ is injective from $B_x(R)$ to S_Γ. It then follows that if

$$y\ ,\ z \in B_x\left(\frac{R}{2}\right)$$

and γ and η are different elements of Γ Then

$$d(\gamma(y), \eta(z)) > \frac{R}{2} .$$

And thus

$$d(y, z) = \inf\left\{d(\gamma(y), \eta(z)) \mid \gamma\ ,\ \eta \in \Gamma\right\} = d_\Gamma\left(\pi(y), \pi(z)\right)\ .$$

Thus φ is a measure preserving injection from $B_x\left(\frac{R}{2}\right)$ to $B_{[x]}\left(\frac{R}{2}\right)$. Surjection ? □

2.2. Construction of hyperbolic surfaces by gluing.

2.2.1. *Gluing two length metric spaces.* We can glue two length spaces provided we have a gluing map that preserves the distances, by defining the length of any curves as the sum of the length in each part.

Then one checks that the gluing two hyperbolic half-spaces along their boundary leads to the hyperbolic plane, and gluing two right-angle wedges leads to the hyperbolic half plane.

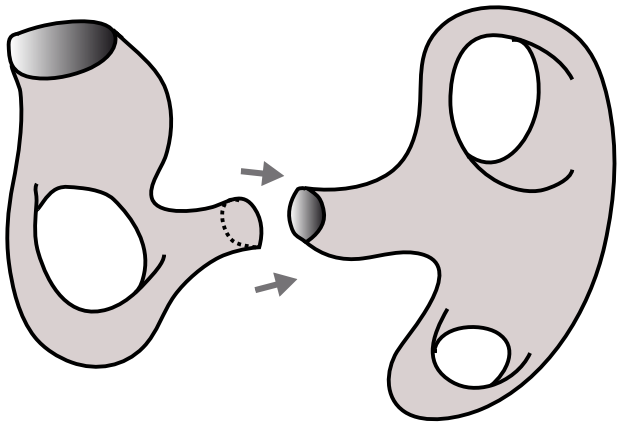


FIGURE 1. Gluing surfaces

2.2.2. *Construction of closed surfaces.* We can therefore construct hyperbolic surfaces of area $4\pi n$ using $3n$-positive real parameters, and $3n$-angles – i.e elements of $\mathbb{R}/\mathbb{Z}$. Moreover, the surface is given together with an extra topological structure namely an *decomposition into pair of pants*. The construction run as follows. First, we construct pairs of right-angled hexagons, fixing the boundary length. Then, gluing two hexagons together we obtain a hyperbolic pair of pants whose boundary length are prescribed. Then we glue these pair of pants together using a prescribed identification of the boundaries. We can isometrically identify each boundary component of length a with $\mathbb{R}/a\mathbb{Z}$, in a canonical way: sending a chosen corner of the previous hexagon to zero. Then the gluing between two boundary components is determined by one parameter in $\mathbb{R}/a\mathbb{Z}$.

2.3. Questions. Two questions remain

(1) Are all compact surfaces obtained this way: by gluing and quotient?

(2) When are two surfaces obtained by gluing isometric? When are two surfaces obtained by quotient isometric?

3. Every complete hyperbolic surface is a quotient

Our goal is to prove the following theorem

THEOREM 3.1 (EVERY HYPERBOLIC SURFACE IS A QUOTIENT). *Every complete connected hyperbolic surface S with possibly boundary and right-angles is the quotient of a convex polygon P by a subgroup Γ of* $\mathrm{Iso}(\mathbf{H}^2)$ *acting freely and properly discontinuously, and preserving P.*
If furthermore S has no boundary then $P = \mathbf{H}^2$.

We call P as in the theorem the universal cover of S and Γ the fundamental group of S. We first give the proof of a corollary.

3.1. Corollary: monodromy group of compact surfaces. We start with a definition: we say two elements γ and η in Γ are *commensurate* if there exists non zero integers p and q such that

$$\gamma^p = \eta^q \ .$$

PROPOSITION 3.2. *Assume S with totally geodesic boundary is compact. Then every element of Γ preserves a unique geodesic. Moreover the following are equivalent:*

(1) *γ and η in Γ fix a common point in $\partial_\infty \mathbf{H}^2$,*
(2) *γ and η in Γ are commensurate,*
(3) *γ and η commute.*
(4) *γ and η are powers of the same element.*

As a classical corollary, we obtain

COROLLARY 3.3. *The fundamental group of a surface S with totally geodesic boundary does not contain a group isomorphic to $\mathbb{Z}^2$ and is not abelian.*

PROOF OF COROLLARY 3.3. Thanks to the proposition, we only have to prove that Γ is different from $\mathbb{Z}$. Assume that $\Gamma = \langle \eta \rangle$
Let us write

$$\partial P = \bigsqcup_{i \in I} \gamma_i \ ,$$

where $(\gamma_i)_{i \in I}$ are the geodesic lifts of the boundary components of S. Each of them is invariant by an element of Γ, hence by η. It follows that ∂P consists in at most one geodesic. Thus P is $\mathbf{H}^2$ or an half-plane. In both cases, $\Gamma \backslash P$ is not compact and we obtain our contradiction. □

Proof. Let us consider the positive function $f : x \mapsto \inf\{d(x, \gamma(x) \mid \gamma \in \Gamma\}$. This function is Γ-invariant and thus defines a function on S. One can prove f is continuous using the fact that Γ acts freely and properly discontinuously. By compactness, it follows that the minimum of f is achieved on P at some x_0. Thus there exists $k(\Gamma) := f(x_0) > 0$ such that $f(x) \geqslant k(\Gamma)$.

Let γ be an element of Γ. Recall that P is Γ, hence γ, invariant. Assume by contradiction that γ has a unique fixed point y_0 in $\partial_\infty \mathbf{H}^2$. It follows that y_0 belongs to $\partial_\infty P$ by exercise 3.9. Let then $c : [0\infty[\to S$ be a geodesic with values in P, such that $c(\infty) = y_0$. Using the upper half plane model so that c is the imaginary axis. The assumption that the unique fixed point of γ – seen in the upper half plane model – is ∞ implies that that

$$\gamma := \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix},$$

for some t. Observe now that

$$\lim_{t\to\infty}(d(c(t), \gamma(c(t)) = 0 .$$

This contradicts the fact that

$$d(c(t), \gamma(c(t) \geqslant k(\Gamma) .$$

Let us prove the that (1) implies (2). Let γ and η be two elements of Γ fixing the same point a at infinity. Let γ_0 and η_0 the geodesics associated to γ and η be the first point. We may assume (possibly after taking inverses) that

$$a = \gamma_0(+\infty) = \eta_0(+\infty)$$

By the contracting property of Proposition 4.5, we have that (after a translation on the parametrization)

$$\lim_{t\to\infty} d(\gamma(t), \eta(t)) = 0 .$$

We want first to show that for all t

$$\gamma_0(t) = \eta_0(t) .$$

Let α and β be the positive real numbers such that

$$\gamma_0(t + \alpha) = \gamma_0(t) , \ \eta_0(t + \beta) = \eta_0(t).$$

We can then find diverging sequences of integers $\{p\}_{m\in\mathbb{N}}$ and $\{q\}_{m\in\mathbb{N}}$ with

$$\lim_{m\to\infty} |p_m\alpha - q_m\beta| = 0 .$$

To prove the existence of these sequences, we just have to remark that $\mathbb{Z}[p, q]$ is a subgroup of $\mathbb{R}$, hence either discrete (in which case $\frac{p}{q}$ is rational and the assertion 6 is obvious) or dense. In this last case, we can find a sequence of

pairwise distinct elements of $\mathbb{Z}[p,q]$ converging to zero. Such a sequence is of the form $\alpha_m p - \beta_m q$ and the assertion follows.

Let us now use the above fact. It follows that for any t,

$$\lim_{m\to\infty} d(\gamma_0(t + q_m\alpha), \eta_0(t + p_m\beta) = 0$$

Let γ_1 and η_1 be the projections of γ_0 and η_0 on S, the previous limits tell us that since, for all integers p and q,

$$d(\gamma_1(t), \eta_1(t)) \leqslant d(\gamma_0(t + p\alpha), \eta_0(1 + q\beta)) \ ,$$

We have that $\gamma_1(t) = \eta_1(t)$. Thus there exists an element ξ of Γ, such that for all t, $\gamma_0(t) = \xi\eta_0(t)$. We now take t large enough so that

$$d(\gamma_0(t), \eta_0(t)) < k(\Gamma) \ .$$

Then for this t large enough,

$$d(\xi(\eta_0(t), \eta_0(t)) = d(\gamma_0(t), \eta_0(t)) < k(\Gamma) \ ,$$

and by the definition of $k(\Gamma)$, we have $\xi = \mathrm{Id}$, hence for all t $\gamma_0(t) = \eta_0(t)$.

Using the same sequences $\{q\}_{m\in\mathbb{N}}$ and $\{p\}_{m\in\mathbb{N}}$ as above we have that, letting $x_0 = \gamma_0(0) = \eta_0(0)$, $c = \gamma_0 = \eta_0$,

$$\begin{aligned} d(x_0, \gamma^{-p_m}\eta^{q_m}x_0) &= d(\gamma^{p_m}x_0, \gamma^{-p_m}\eta^{q_m}) \\ &= d(c(p_m\alpha), c(q_m\beta) \\ &\leqslant |p_m\alpha - q_m\beta| \ . \end{aligned}$$

It follows that

$$\lim_{m\to\infty} d(x_0, \gamma^{-p_m}\eta^{q_m}) = 0 \ ,$$

Thus, by the fact that Γ acts freely and properly discontinuously on $\mathbf{H}^2$, for all m large enough

$$\gamma^{p_m} = \eta^{q_m} \ .$$

We now prove (3) implies (2). Assume that γ and η commutes. Let γ_0 be the unique geodesic invariant by γ. Then $\eta\gamma_0$ is invariant by $\eta\gamma\eta^{-1}$, hence by γ. It follows that γ_0 is preserved by η_0 and the previous implication tells us that γ and η are commensurate.

We now prove (2) implies (3). Assume that γ and η are commensurate. Then they preserve the same geodesic. It follows that γ and η are – as matrices in $\mathsf{PSL}_2(\mathbb{R})$ – diagonalizable in the same basis. Hence γ and η commute.

We finally prove that (3) implies (1). Assume that γ and η commutes. By the implication above, this implies that they preserve the same geodesic c. Let

A be the subgroup of the element of Γ preserving c. Let us consider the map φ from A to $\mathbb{R}$ given by

$$\gamma \mapsto \lambda \text{ where } c(t+\lambda) = \gamma c(t) \, .$$

We observe that φ is an injective morphism. Moreover φ has a discrete image as a consequence of the free and properly discontinuous action of Γ. It follows that $\Phi(A) = \mathbb{Z}(\mu)$, where $\mu = \varphi(\xi)$ for some ξ. Thus every element of A is a power ξ.

All the remaining implications are obvious. □

3.2. Preliminaries: path of balls. Let $c : [a, b] \to S$ be a curve in a hyperbolic space, we can then find a

(1) a sequence of open sets $\{U_i\}_{i\in[1,n]}$ in S, all isometric to balls in $\mathbf{H}^2$,
(2) a partition

$$a = a_0 < \cdots < a_i < a_{i+1} \cdots < a_n = b \, ,$$

with $c[a_{i-1}, a_i] \subset U_i$. Let then V_i be the connected component of $U_i \cap U_{i+1}$ containing a_i. Such a sequence $\{U_i\}_{i\in[1,n]}$ will be called a *sequence of balls associated* to c.

Given a curve $c[a, b] \to S$, an *initial isometry* is an isometry from a neighborhood of $c(a)$ o $\mathbf{H}^2$.

Proposition 3.4. *Given a path of balls $\{U_i\}_{i\in[1,n]}$, and an initial isometry f, there exists a unique family of isometries f_i of U_i in $\mathbf{H}^2$ so that f_i coincide with f_{i+1} on V_i and f_0 coincide with f on a neighborhood of $c(a)$.*

We call f_n the *final isometry*.

Proof. This is an immediate consequence of Proposition 3.3. □

We now have

Proposition 3.5. *Let f and g be the final isometries for two path of balls associated to c, associated to the same initial isometry $f = g$ on some neighborhood of $c(b)$.*

Thus the final isometry only depends on the initial isometry and the curve.

Then, we say two ball paths, corresponding to two curves c_0 and c_1 with fixed extremity i_0 and i_n are *homotopic* if they can be obtained from each other after a succession of the following elementary operation: *deletion* which is just to remove an index (whenever it is possible) or *insertion* which is the converse operation.

As an example one easily sees that two paths of balls for the same curve are homotopic. More generally we leave as an exercise

PROPOSITION 3.6. *If c_0 and c_1 are two homotopic, then the corresponding paths of balls are homotopic.*

We then have a generalization of the previous proposition

PROPOSITION 3.7. *If c_0 and c_1 are two homotopic, with the same initial isometry, then their final isometries coincide.*

PROOF. This is again a consequence of Proposition 3.3. □

Thus the final isometry only depends on the initial isometry and homotopy class of the curve.

3.3. Simply connected hyperbolic surfaces. We now prove the following

THEOREM 3.8. *Let S be a simply connected complete hyperbolic surface then there exists a local isometry φ of S in $\mathbf{H}^2$. If furthermore S is complete, then S is isometric to a convex polygon in $\mathbf{H}^2$.*

PROOF. We fix a base point x_0 and an isometry f_0 from a neighborhood of x_0 to a ball in the hyperbolic plane. If c is any curve from x_0 to x, the final isometry g, starting from f_0 gives an isometry from a neighborhood of x to subset of the hyperbolic plane. Since this final isometry does not depend on the choice of the path, we define

$$\varphi(x) \coloneqq g(x) \,.$$

by construction φ is a local isometry and this proves the first part.

Now if S is complete, then φ is injective: given x and y, we can find a geodesic c joining x to y by Theorem 1.10. Then $\varphi(c)$ is a geodesic in $\mathbf{H}^2$, hence has distinct extremities. Thus $\varphi(x) \neq \varphi(y)$.

The image of φ is a hyperbolic surface with totally geodesic boundary and right-angles, thus a convex polygon P. Then φ^{-1} from P to S is also local isometry. Then φ preserves the distance by Proposition 1.4. □

3.4. Covering by convex polygons. Our goal is to show that every hyperbolic surface is a quotient, namely Theorem 3.1

PROPOSITION 3.9. *Given a complete hyperbolic surface S with geodesic boundary and right-angles, there exists a convex – possibly non-compact – polygon P in the hyperbolic plane, and a locally isometric covering φ from P to S.*

We call P the *universal cover of S*.

Let us start with a proposition interesting in itself

Proposition 3.10. *Let $p : X \to S$ be a covering. Assume that (S, d) is a hyperbolic surface. Then there exists a unique hyperbolic surface structure on X such that p is a local isometry.*

Furthermore if S is complete so is X.

Proof. Let $c : [0,1] \to X$ be a curve . Let us define $\ell(c) \coloneqq \ell(p(c))$. Then for any x and y we define

$$d(x,y) \coloneqq \inf\{\ell(c) \mid c(0) = x\ , c(1) = y\}\ .$$

Let B be a ball of radius R in S of center $z = p(x)$ isometric to a ball in the model, such that

$$p^{-1}(B) = \bigsqcup_{y \in p^{-1}z} U_y\ .$$

with p an homeomorphism from U_y to B, for all y.

It follows that if c is a curve joining x to w, with $p(w) = p(x)$ and $x \neq w$, then $p(c)$ starts from $p(x)$ and leaves B, in particular $\ell(c) = \ell(p(c)) \geqslant R$. Thus, we have that for w different than x, and $p(w) = p(x)$, then $d(x,w) > 0$.

We can now prove that d is a distance. The only non-obvious property to prove is that if $d(x,y) = 0$ then $x = y$. Assume now $d(x,y) = 0$. Since $d(p(x), p(y) \leqslant d(x,y)$, we have that $p(x) = p(y)$, but the property above implies $y = x$.

Let us now prove p is a local isometry. We use the same x, z and R as above.

Let $V_y = p^{-1}(B_0)$, where B_0 is the ball of radius $R/4$ and center z. Assume

We now remark that if w_0 and w_1 belongs to V_x, if c is a curve of length less than $R/2$ joining w_0 to w_1 in V_x, then $p(c)$ is included in B, and thus by the lifting property c is included in U_x. Since there exist a curve joining w_0 to w_1 in B of length strictly less than $R/2$, it follows that

$$\begin{aligned} d(w_0, w_1) &= \inf\inf\{\ell(c) \mid c(0) = x\ , c(1) = y\ , \ell(c) < R/2\} \\ &= \inf\inf\{\ell(c) \mid c(0) = x\ , c(1) = y\ , c[0,1] \subset U_x\} \\ &= \inf\inf\{\ell(\gamma) \mid \gamma(0) = p(w_0)\ , c(1) = p(w_1)\ , \gamma[0,1] \subset B\} \\ &= d(p(w_0), p(w_1))\ . \end{aligned} \tag{6}$$

We have just proved that p is a local isometry.

Observe now that every geodesic X projects to a geodesic in S, conversely if a curve c projects to a geodesic, then c is a geodesic by the local minimizing property. By the lifting property, every maximal geodesic in S can thus be lifted to a geodesic in X. Thus X is complete if S is. □

Obviously this property is true for any length space.

Proof of Theorem 3.1. Let P be the universal cover of S described in Theorem 0.1 Then by Proposition 3.10, P admits the structure of a hyperbolic surface. Moreover since p is a local isometry, we see that every element of Γ, defined in Theorem 0.1, is a local isometry, hence by Proposition 1.4 an isometry. □

Exercise 3.1: (Exponential map)(*) Let S be a complete hyperbolic surface – without boundary. Let x in S and $B(x, 2R)$ a ball of center x and radius $2R$ isometric to a ball in the model. Let

$$S_x := \{z \in S \mid \mathrm{d}(z, x) = R\} .$$

For any z in S_x, let c_z be the geodesic such that $c_z(R) = z$. We define the *exponential map* from

$$\mathsf{T}_x S := \{0\} \cup (S_x \times R^{>0}) ,$$

to S by

$$\exp(z, t) = c_z(t) .$$

(1) Show that exp_x is well defined, is continuous and surjective.
(2) (More difficult) show that $\exp_x$ is a local homeomorphism then a covering: find a proof not using Theorem 3.9 then a proof using Theorem 3.9.
(3) Show that $\mathsf{T}_x S$ equipped with the hyperbolic metric defined by the covering is isometric to $\mathbf{H}^2$.

4. Compact surfaces and pair of pants

We are now interested in compact hyperbolic surfaces.

Theorem 4.1 (Decomposition into hexagons). *Every compact oriented connected hyperbolic surface is of area $4\pi n$ and can be obtained be decomposed into $2n$ hexagons glued as pair of pants. Moreover, this decomposition, the $3n$-length of boundary parameters as well as the $3n$ gluing parameters, is fixed as soon as we fix $3n$ homotopy classes of pairwise non-intersecting simple curves on S.*

Two closed hyperbolic surfaces are isometric if and only if we can find a decomposition of each surface into pair of pants, with the same gluing and length parameter

4.1. Curves on hyperbolic surfaces. We give some properties of curves on complete hyperbolic surfaces with totally geodesic boundary. We refer to the appendix for the homotopy related definitions.

Theorem 4.2 (Curves and arcs). *Let S be a compact surface with totally geodesic boundary*

(1) *Given two points p and q in S and a path c from p to q there exists a unique geodesic γ_c joining p to q and homotopic to c. Moreover this geodesic minimizes the length of all curves joining p to q homotopic to c.*

(2) *Given a closed curve c, there exists a unique closed geodesic γ_c freely homotopic to c. Moreover the length of γ_c minimize the length of all curves freely homotopic to c.*
(3) *Given a closed arc c joining a boundary component a to a boundary component b, then there exists a unique geodesic arc γ_c orthogonal to the boundary and homotopic to c with respect to a and b. Moreover the length of γ_c minimize the length of all arcs homotopic to c relative to a and b.*

Proof. Let P the convex polygon and φ the covering from P to S from Theorem 3.1. The first statement is a consequence of our description of the universal: indeed let p_0 be a lift of p in P, c_0 be the lift of c starting at p_0 and q_0 the end point of c_0. By elementary properties of covers (see appendix A), q_0 only depend on the choice of p_0 and the homotopy class of c. Thus we take γ_c to be the projection of the geodesic from p_0 to q_0, which lies in P since P is convex.

For the second statement, we have to use the compactness of S. let $c : [0,1] \to S$ be a closed curve and $\tilde{c} : \mathbb{R} \to \tilde{S}$ a lift of c in P. If c is not freely homotopic to zero, then there exists an element $\gamma \in \Gamma$ such that for all $n \in \mathbb{N}$.

$$\tilde{c}(x+n) = \gamma^n \tilde{c}(x) \,. \tag{7}$$

Since S is compact, by Proposition 3.2, γ preserves a unique geodesic γ_c whose projection on S we denote by γ_c^0. Let T_0 the length of γ_c^0. Let us choose a parametrization of c so that $c(t+T_0) = c(t)$. It follows from equation that for all t in $\mathbb{R}$,

$$f(t) \coloneqq d(\tilde{c}(t), \gamma_c(t)) \leqslant K \,. \tag{8}$$

It follows that we can choose a homotopy from c to γ_c^0, given by

$$c(s,t) = p(\gamma_t(sf(t)) \,,$$

where $\gamma_t(s)$ is the geodesic arc joining $c(s)$ to $\gamma_c(s)$, and p is the projection from P to S.

The third statement about geodesic arc is an adaptation of the previous argument: Assume that a curve $c : [0,1] \to S$ joins two boundary components a and b (which may coincide). Let c_0 be a lift of c, a_0 the geodesic lifting a passing through $c_0(0)$ and similarly b_0 the geodesic lifting b passing through $c_0(1)$. Observe that any curve homotopic to c (relative to a and b) with initial end point in a_0 has a final end-point in b_0

Since a and b do not intersect so do a_0 and b_0. Moreover by Proposition 3.2 a_0 and b_0 have different point at infinity. It follows that there exists a unique

curve $\gamma_c : [0,1] \to \mathbf{H}^2$ such that

$$\ell(\gamma_c^0) = \inf\{\ell(g) \mid g : [0,1] \to \mathbf{H}^2 \ , \ g(0) \in a_0 \ , g(1) \in b_0\} \ ,$$

Moreover γ_c^0 is a geodesic arc orthogonal to a_0 and a_1. We take γ_c to be the projection of γ_c^0 on $S = \Gamma\backslash P$. □

We have a refinement of the previous construction. We say that two homeomorphisms φ_0 and φ_1 of S are *isotopic* if there exists a continuous map $\Phi : S \times [0,1] \to S$ such that

(1) For all s in S, $\varphi_0(s) = \Phi(s,0)$ and $\varphi_1(s) = \Phi(s,1)$.
(2) For all t in $[0,1]$,
$$\varphi_t : s \mapsto \Phi(s,t) \ ,$$
is a homeomorphism.

Next is a crucial definition: a *pair of pants* is a topological surface homeomorphic to $D \setminus D_1 \cup D_2$, where D is a closed disc in $\mathbb{R}_2$, D_1 and D_2 are open disks whose closures are disjoint and both included in D.

THEOREM 4.3 (EMBEDDED CURVES). (1) *Assume that in the last two items of Theorem 4.2, the curve or arc c are embedded, then the geodesic or arc γ is embedded.*
(2) *Assume that φ is em embedding of a pair of pants U in S, then φ is isotopic to an embedding ψ such that $\psi(U)$ is a hyperbolic surface with totally geodesic boundary.*

THEOREM 4.4. *Let S be a compact hyperbolic surface with boundary which is homeomorphic to a pair of pants. Then there exists two isometric hexagons with right angles H_1 and H_2 such that S is isometric to $H_1 \cup H_2$. Moreover, H_1 and H_2 are unique up to isometries of $\mathbf{H}^2$.*

COROLLARY 4.5. *The area of a hyperbolic pair of pants is 4π.*

Proof to be added

4.2. Finding topological pair of pants.

PROPOSITION 4.6. *Let S be a compact hyperbolic surface. Then S contains a hyperbolic pair of pants.*

COROLLARY 4.7. *Let S be a compact surface, then S contains a pair of pants with totally geodesic boundary.*

Proof to be added

PROPOSITION 4.8. *Let S be a compact hyperbolic surface with a non-empty totally geodesic boundary. Then either there exists an arc from a boundary component to another (or the same) boundary component, not homotopic to the boundary component.*

Proof to be added

4.3. Proof of Theorem. As another consequence, we obtain

THEOREM 4.9. [GAUSS-BONNET FORMULA (BIS)] *Let S be a compact surface admitting a hyperbolic structure. Then two pair of pants decomposition have the same number of pair of pants and this number is even. If $\chi(S)$ is this opposite of this number called the* Euler characteristic of *S, then any hyperbolic structure on the surface has area $-2\pi\chi(S)$.*

4.4. A complement: Dirichlet fundamental domains. Let $S = \Gamma\backslash P$ and x a point of P. The *Dirichlet fundamental domain* (relative to x) is the subset Δ of P given by

$$\Delta = \{y \in P \mid \text{ for all } \gamma \text{ in } \Gamma d(x, \gamma(y) \geqslant d(x, y)\} .$$

We leave the following proposition as an exercise:

EXERCISE 4.1:

(1) A Dirichlet fundamental domain Δ is convex.
(2) Moreover we have

$$\bigsqcup_{\gamma\in\Gamma} \gamma\Delta = P ,$$

$$\gamma\Delta \cap \Delta = \emptyset \quad \text{if} \quad \gamma \neq \mathrm{Id} .$$

(3) If S is compact then Δ is a compact polygon.
(4) If P is different from $\mathbf{H}^2$, then Δ intersects ∂P.

5. Comments, references and further reading

Part 2

Dynamics and ergodic theory

CHAPTER 4

Dynamics

In this chapter, Γ will be a discrete subgroup of $\mathrm{Iso}(\mathbf{H}^2)$ so that $\Gamma\backslash\mathbf{H}^2$ is a compact hyperbolic surface without boundary. We furthermore assume that Γ lies in the connected component of the identity of $\mathrm{Iso}(\mathbf{H}^2)$, isomorphic to $\mathsf{PSL}_2(\mathbb{R})$. The corresponding surface is said to be *oriented*

1. The action on the boundary at infinity

We begin by studying the action of Γ on the boundary at infinity $\partial_\infty\mathbf{H}^2$ of $\mathbf{H}^2$. Every element of γ corresponds to a closed geodesic and will therefore preserves exactly two points at infinity $\{\gamma^-, \gamma^+\}$.

It then follows, that every γ in Γ has *north-south dynamics* meaning that the sequence of iterates of any point x in $\partial_\infty\mathbf{H}^2$ different than γ^- converges to γ^+, as in Figure 1.

$$\lim_{n\to\infty}(\gamma^n(x)) = \gamma^+.$$

We now use the north-south dynamics to show three crucial properties of the action of Γ on the boundary at infinity.

Lemma 1.1 (Minimality). *Every orbit of Γ is dense on $\partial_\infty\mathbf{H}^2$.*

Proof. Let F be a closed Γ invariant set in $\partial_\infty\mathbf{H}^2$ and E be the *convex envelope* of F in $\mathbf{H}^2$, that is the intersection of all hyperbolic half spaces containing F. The set E is a closed convex set which is Γ invariant. Let d be the function on $\mathbf{H}^2$ defined as the distance to E. Then d is Γ invariant. However d is unbounded as we see from taking a geodesic orthogonal to one of the boundary component of the convex set. This contradicts the compactness of $\mathbf{H}^2/\Gamma$. □

Lemma 1.2 (Density of end points). *The set of end points of geodesics $\{(\gamma^+, \gamma^-) \mid \gamma \in \Gamma\}$, is dense in $\partial_\infty\mathbf{H}^2 \times \partial_\infty\mathbf{H}^2$.*

Proof. By the previous lemma the set $\{\gamma^+ \mid \gamma \in \Gamma\}$ is dense in $\partial_\infty\mathbf{H}^2$. Let now (x, y) be pair of points in $\partial_\infty\mathbf{H}^2 \times \partial_\infty\mathbf{H}^2$, we can therefore find a pair of distinct points (η^-, γ^+) associated to elements η and γ. We remark that if two elements α and β of the group are such that $\alpha^+ = \beta^+$ then $\alpha^- = \beta^-$ by Proposition 3.2.

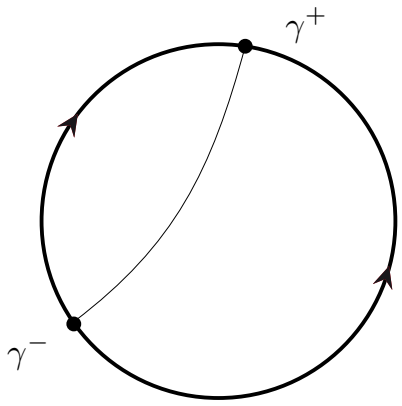


FIGURE 1. North-south dynamics

We therefore assume that all points $\eta^\pm, \gamma^\pm$ are distinct. The final remark is that

$$\lim_{n\to\infty}(\gamma^n\eta^n)^+ = \gamma^+,$$

and symmetrically

$$\lim_{n\to\infty}(\gamma^n\eta^n)^- = \eta^-,$$

The process is described in Figure 2 Let U be a small neighborhood of γ^+. Since γ^+ is different than η^- a high power of η will send U to a very small neighborhood V of η^+. Since η^+ is different than γ^- a high power of γ will send V to a even smaller neighborhood of γ^+. It follows that $\xi_n = \gamma^n\eta^n$ maps U into itself. Therefore it has a fixed point in U. This point is necessarily the attractive fixed point of ξ_n. This is what we wanted to prove.

□

Here is another important consequence of this North-South dynamics.

LEMMA 1.3 (BABY HYPERBOLIC STABILITY). *Let S be a compact surface. Let ρ_1 and ρ_2 be two representations of $\pi_1(S)$ in $\mathsf{PSL}_2(\mathbb{R})$ which are monodromies of hyperbolic structures on S. Then the two corresponding actions on $\partial_\infty\mathbf{H}^2$ are conjugate. More precisely there exists a unique – usually non-smooth – homeomorphism Φ of $\partial_\infty\mathbf{H}^2$ so that*

$$\forall x \in \partial_\infty\mathbf{H}^2,\ \forall\gamma \in \pi_1(S),\ \Phi\left(\rho_1(\gamma)\, x\right) = \rho_2(\gamma)\, \Phi(x)\ .$$

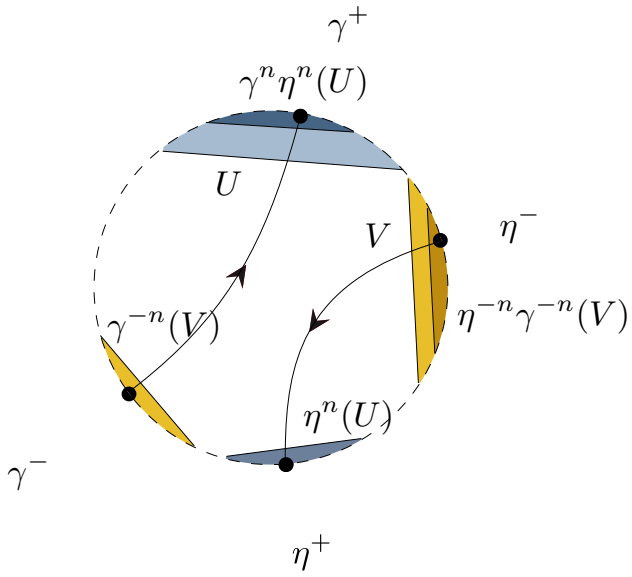


FIGURE 2. Density of pairs of fixed points

For the moment, we just prove the uniqueness of the conjugation and post-pone the proof of the existence.

PROOF OF UNIQUENESS. Let E_1 be E_2 be the set of end points of closed geodesics in ∂_∞ of respectively $\rho_1(\pi_1(S))$ and $\rho(\pi_1(S))$.

Our first remark is that Φ satisfies from $\rho_1(\gamma)^+$ to $\rho_2(\gamma)^+$. Indeed, since Φ conjugate the action it sends attractive fixed points to attractive fixed points.

Then the uniqueness follows from the density of E_1. □

One can actually prove that the conjugacy is Hölder, this is the grown up version of Hyperbolic Stability.

This last lemma leads the an abstract definition of the boundary at infinity of a surface group.

DEFINITION 1.4. *Let S be a closed connected oriented surface of genus greater than 2. The boundary at infinity $\partial_\infty\pi_1(S)$ of a surface group is a topological circle on which $\pi_1(S)$ in a way which is conjugate to the action of $\rho(\pi_1(S)$ on $\partial_\infty\mathbf{H}^2$, where ρ is the monodromy of a hyperbolic structure.*

There is a beautiful theorem by Matsumoto which characterizes the action of $\pi_1(S)$ on $\partial_\infty\pi_1(S)$.

THEOREM 1.5 (MATSUMOTO). *Let S be a closed surface. Let T be a topological space homeomorphic to the circle. Assume that $\pi_1(S)$ acts on T, with the following properties*

- *each non-trivial element has exactly one attractive and one repulsive fixed point,*
- *every orbit is dense*

then there is a homeomorphism conjugating the action of $\pi_1(S)$ between T and $\partial_\infty \pi_1(S)$.

2. The unit tangent bundle and flows

2.1. Three points on view on the unit tangent bundle of $\mathbf{H}^2$. We choose once on for all an orientation on $\partial_\infty \mathbf{H}^2$, and thus talk of positively oriented triples. Then $\mathrm{Iso}_+(\mathbf{H}^2)$ is the subgroup of $\mathrm{Iso}(\mathbf{H}^2)$ preserving positively oriented triples. The group $\mathrm{Iso}_+(\mathbf{H}^2)$ is the connected component of the identity of $\mathrm{Iso}(\mathbf{H}^2)$ and is isomorphic to $\mathsf{PSL}_2(\mathbb{R})$. We have several possible definitions of the unit tangent bundle of $\mathbf{H}^2$ that we denote $\mathsf{U}\mathbf{H}^2$ and we will show that they are equivalent

(1) The set $\mathsf{U}\mathbf{H}^2$ is the set of arc-length parametrized geodesics.
(2) The set $\mathsf{U}\mathbf{H}^2$ is the set of pairs (x, y), where x is a point in $\mathbf{H}^2$ and y a point in $\partial_\infty \mathbf{H}^2$.
(3) The set $\mathsf{U}\mathbf{H}^2$ is the set is the set of positively oriented triples of $\partial_\infty \mathbf{H}^2$
(4) The set of orientation preserving isometries of the upper half plane model with $\mathbf{H}^2$,
(5) The group $\mathsf{PSL}_2(\mathbb{R})$ (this identification requires a choice).

Let see how these different points of view are related.

- Obviously a point x in an oriented geodesic L, defines uniquely a parametrization of L, $\gamma : [-\infty, \infty] \to \mathbf{H}^2$ by having $\gamma(0) = x$, $\gamma(-\infty) = t^-$ and $\gamma(-\infty) = t^-$, where L joins t^- to t^+.
- Given a pair $t = (x, L)$, we consider the triple (t^-, t^+, t^0) of oriented distinct points in $\partial_\infty \mathbf{H}^2$, where L joins t^- to t^+, and the geodesic joining x to t^0 is orthogonal to L.
- Given a triple or positively oriented points (t^-, t^+, t^0) there exists a unique isometry ψ of the upper half plane model with $\bar{\mathbf{H}}^2$, with so that $t^- = \psi(0)$, $t^+ = \psi(+\infty)$ and $t^0 = \psi(1)$.
- Conversely given such an isometry φ, we consider

$$(t^-, t^+, t^0) = (\varphi(0), \varphi(\infty), \varphi(1)) .$$

- Choosing an isometry φ_0 of the upper half plane model with $\mathbf{H}^2$. Every other isometry φ defines an element of $\mathsf{PSL}_2(\mathbb{R})$ by

$$\varphi \mapsto \varphi^{-1} \circ \varphi_0 \, .$$

2.2. Flows and their geometric description. From third point of view on $\mathsf{U}\mathbf{H}^2$, we obtain a left action (by post-composition) of $\mathrm{Iso}(\mathbf{H}^2)$ on $\mathsf{U}\mathbf{H}^2$ as well as a right action (by pre-composition) of $\mathsf{PGL}_2(\mathbb{R})$ on $\mathsf{U}\mathbf{H}^2$. The action of special subgroups of $\mathsf{PGL}_2(\mathbb{R})$ has names

(1) The action of the diagonal subgroup $\varphi_t(x) \coloneqq xa_t$ where

$$a_t = \begin{pmatrix} e^{-\frac{t}{2}} & 0 \\ 0 & e^{\frac{t}{2}} \end{pmatrix}$$

is called the *geodesic flow*. We denote $\mathsf{A} \coloneqq \{a_t\}_{t\in\mathbb{R}}$.

(2) The action of the upper triangular subgroup : $h_t^+(x) \coloneqq x\, n_t^+$, where

$$n_t^+ = \begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix}$$

is called the *stable horocyclic flow*. We denote $\mathsf{N}^+ \coloneqq \{n_t^+\}_{t\in\mathbb{R}}$.

(3) The *time reversion* is given by $\sigma(x) \coloneqq x\, J$ where

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

(4) the *unstable horocyclic flow* is given by the action of the lower triangular group $h_t^-(x) \coloneqq x\, n_t^-$,

$$n_t^- = \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}$$

We denote $\mathsf{N}^- \coloneqq \{n_t^-\}_{t\in\mathbb{R}}$.

Let us consider ${\mathbf{H}^2}_p$ the upper half-plane model, then

Proposition 2.1. *For any element B of $\mathsf{PSL}_2(\mathbb{R})$, the right action of $\mathsf{PSL}_2(\mathbb{R})$ on ${\mathsf{U}\mathbf{H}^2}_p$ is given on the point (i,∞), is*

$$(i,\infty)\, B = (B^{-1}(i), B^{-1}(\infty)) \, .$$

Proof. Indeed in the description of ${\mathsf{U}\mathbf{H}^2}_p$ as the set of oriented isometries from ${\mathbf{H}^2}_p$ to itself on one side, and on ${\mathbf{H}^2}_p \times \partial_\infty {\mathbf{H}^2}_p$. We have

$$(x,y) \leftrightarrow \varphi \, ,$$

if $\varphi(i) = x$ and $\varphi(\infty) = x$. Thus

$$(i,\infty) \leftrightarrow \mathrm{Id},$$

and hence

$$(i,\infty)\ B \leftrightarrow \mathrm{Id}\circ B = B \leftrightarrow (B^{-1}(i), B^{-1}(\infty))\ .$$

This is what we wanted to □

Exercise 2.1: Compute what is $(x,y)\ B$ in general. Notice that it is NOT $(B^{-1}(x), B^{-1}(y))$.

One immediately sees that

$$J\circ\varphi_t = \varphi_t\circ J\ , J\circ h_t^+ = h_t^-\circ J\ . \tag{9}$$

in $\mathsf{PGL}_2(\mathbb{R})$.

We need to give a geometric interpretation of these flows notably in the first point of view

Proposition 2.2 (Geometric interpretation). *Let us interpret* UH^2 *as the set of parametrized geodesic*

1. *The geodesic flow is given by* $\gamma\cdot\varphi_t\ (s) = \gamma(t+s)$.
2. *the time reversion is given by* $\gamma\cdot J(s) = \gamma(-s)$.
3. *For the stable horocyclic flow, it is better to use the interpretation of* UH^2 *as* $\mathbf{H}^2\times\partial_\infty\mathbf{H}^2$*:* $(x,y)\cdot h_t^+ = (z,y)$*, where* $H(t)=z$*,* $H(0)=x$ *and* H *is the unique orientation compatible arc-length parametrization of the horocycle defined by* y *and* x.

Let us remark that if φ is an isometry between two hyperbolic planes ${\mathbf{H}^2}_0$ and ${\mathbf{H}^2}_1$, then φ gives rise to a map ψ from $\mathsf{U}{\mathbf{H}^2}_0$ to $\mathsf{U}{\mathbf{H}^2}_1$ commuting with the left action of $\mathsf{PSL}_2(\mathbb{R})$. We have indeed that (from the point of view (4)) that

$$\psi(\xi) = \xi\circ\psi^{-1}\ .$$

Proof. Let us choose a point (x,y) in $\mathsf{UH}^2 = \mathbf{H}^2\times\partial_\infty\mathbf{H}^2$. We can choose an isometry φ sending (x,y) to (i,∞) in the upper half plane model. Now we observe that, using Proposition 2.1 in the second inequality

$$\varphi_t((i,\infty)) = (i,\infty).a_t = (a_{-t}(i), a_{-t}(\infty)) = (ie^t,\infty)\ .$$

Thus if we consider the parametrized geodesic γ defined by $(\imath,\infty)$:

$$\gamma(t) = ie^t\ ,$$

we get the result. A similar construction shows that

$$(i,\infty)\cdot h_t = (i+t,\infty)\ ,$$

and the result follows by a similar construction. We leave the reader see for themselves the time-reversion property. □

Hence, the geodesic flows and horocycle flows are just the actions of the one-parameter diagonal and upper triangle groups of matrices in $\mathsf{PSL}_2(\mathbb{R})$ on UH^2 respectively.

Closed geodesics can – and will – now be interpreted as closed orbits of the geodesic flow.

EXERCISE 2.2: Write the action of these flows on $\mathcal{T}$ using only the cross-ratio.

2.3. Commuting rule. Observe now that we have the important commuting rule:

$$\varphi_t \circ h_s = h_{e^{ts}} \circ \varphi_t \ , \tag{10}$$

This relation gives the *contraction property* explained in figure 3.

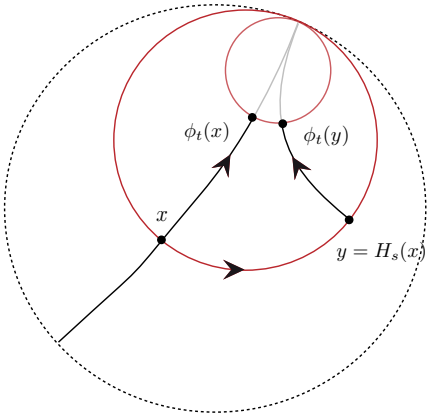


FIGURE 3. Commutation rule

EXERCISE 2.3: Show that the group generated by A, N^+ and N^- is $\mathsf{SL}_2(\mathbb{R})$:

(1) Let B be an invertible matrix. Show that you can find a matrix N_0 in N^+ so that $N_0\, B$ has a non zero top right coefficient.
(2) Show then that there is N_1 in N^- so that $N_1\, N_0\, B$ has a zero bottom left coefficient.
(3) Show then that, $N_1\, N_0\, B = A\, N_3$, where A is in A and N_3 in N^+
(4) Conclude.

2.3.1. *A distance on* UH^2. We prove the following results

PROPOSITION 2.3. *There exists a distance d on UH^2, which is invariant by the left action of* $\mathrm{Iso}(\mathbf{H}^2)$.

As we shall see the distance is not invariant by the left action of $\mathsf{PSL}_2(\mathbb{R})$.

PROOF. We first prove there exists a distance on $\mathsf{U}\mathbf{H}^2$ invariant under the left action of $\mathrm{Iso}(\mathbf{H}^2)$. Let us consider two parametrized geodesics γ and η and define

$$d(\gamma, \eta) := d_{\mathbf{H}^2}(\gamma(0), \eta(0)) + d_{\mathbf{H}^2}(\gamma(1), \eta(1).$$

Then one easily checks that this defines a distance on $\mathsf{U}\mathbf{H}^2$ for which $\mathrm{Iso}(\mathbf{H}^2)$ acts by isometries. □

2.4. The unit tangent bundle of a surface. If now Γ is a subgroup of $\mathrm{Iso}(\mathbf{H}^2)$ such that $S := \Gamma\backslash\mathbf{H}^2$ is a hyperbolic surface, we define the unit tangent bundle $\mathsf{U}S$ by

$$\mathsf{U}S := \Gamma\backslash\mathsf{U}\mathbf{H}^2 .$$

We then define a metric on $\mathsf{U}S$, by

$$d(x, y) := \inf_{\gamma\in\Gamma} d(x_0, \gamma y_0) ,$$

where x_0 and y_0 are elements of $\mathsf{U}S$ that project to x and y respectively. As before we have the following result:

PROPOSITION 2.4. *The projection from $\mathsf{U}\mathbf{H}^2$ to $\mathsf{U}S$ is a local isometry.*

3. The Anosov property and the Closing Lemma

3.1. The Anosov property. Let $\mathsf{U}S$ be the unitary tangent bundle of the surface S, which from the discussion above is a left quotient $\mathsf{U}\mathbf{H}^2$ by a discrete group Γ. We therefore have three flows on $\mathsf{U}S$ and the corresponding foliations

(1) The geodesic flow φ_t
(2) The stable horocycle flow whose orbits we call *stable leaves*.
(3) The unstable horocycle flow whose whose orbits we call *unstable stable leaves*, obtained by interchanging the role of end points.
(4) The *central stable leaf* is the 2-dimensional leaf which is obtained as the orbit under the geodesic flow of the stable leaf.

Then the commutation rules (10) translate into the *Anosov property* of the geodesic flow, which we try to depict in Figure 4

(1) Two points on the same stable leaf get closer under a *positive* action of the geodesic flow.
(2) Two points on the same unstable leaf get closer under a *negative* action of the geodesic flow.

This property is the translation for the geodesic flow of the north-south dynamics of the action of the monodromy group.

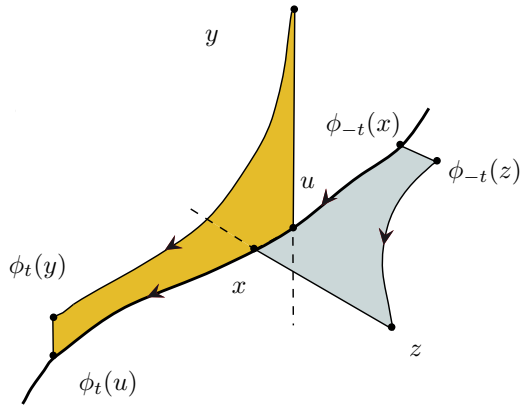


FIGURE 4. Anosov property

3.2. The Closing Lemma. The Anosov property has a crucial consequence.

LEMMA 3.1 (CLOSING LEMMA). *There exists ε_0, such that for any K greater than 1, there exists T_0 such that for all $\varepsilon < \varepsilon_0$, x and T, with $T \geqslant T_0$, satisfying*

$$d(x, \varphi_T(x)) < \varepsilon \ .$$

Then there exists y, with $d(x, y) < K\varepsilon$, a positive number s with $|s - T| \leqslant K\varepsilon$ so that $\varphi_s(y) = y$.

REECRIRE

PROOF. We choose a neighborhood U of x and a parametrization of this neighborhood by $]-\varepsilon, \varepsilon[^3$, given by

$$\psi \ : \ (u, v, w) \mapsto \varphi_u \circ H^+_v \circ H^-_w(x) \ ,$$

Our first step is to prove the following fact:

There exists z such that $d(x, z) < \alpha$, a positive number s with $|s - T| \leqslant \alpha$ so that there exists u, with $|u| \leqslant \varepsilon$ and

$$\varphi_s(z) = H^-_u(z). \tag{11}$$

We can first have

$$\varphi_T(x) = \varphi_{u_0} \circ H^+_{v_0} \circ H^-_{w_0}(x),$$

Let us replace T by $T - u_0$ so that

$$\varphi_T(x) = H^+_{v_0} \circ H^-_{w_0}(x),$$

Let also consider the functions η, ξ, and ζ –depending on w_0 such that

$$H^+_v \circ H^-_{w_0}(x) = \varphi_{\eta(v)} \circ H^-_{\zeta(v)} \circ H^+_{\xi(v)}(x)$$

Let K be a constant (independent on w_0) so that η, ζ and ξ are K-Lipschitz. It follows that for s, $|u| \leqslant \varepsilon$, we have

$$\begin{aligned}\varphi_T(H^+_s(x) &= H^+_{e^{-T}s}\varphi_T(x)\\ &= H^+_{e^{-T}s+v_0}H^-_{w_0}(x)\\ &= \varphi_{\eta(e^{-T}s+v_0)} \circ H^-_{\zeta(e^{-T}s+v_0))} \circ H^+_{\xi(e^{-T}s+v_0))}(x)\,. \end{aligned} \tag{12}$$

Now the map

$$s \mapsto \xi(e^{-T}s + v_0)\,,$$

is a contracting map for T large enough since ζ is K-Lipschitz, hence has a fixed point v_1. Then setting $z = H^+_{v_1}(x)$ and replacing T by $T - \eta(e^{-T}v_1 + v_0)$ we have

$$\varphi_T(z) = H^-_{w_1}(z)\,, \tag{13}$$

where $w_1 = \zeta(e^{-T}v_1 + v_0)$. This proves our first assertion.

Then one obtain the Closing Lemma using the same argument but working backward in time. □

The Closing Lemma implies the density of the reunion of all closed orbits which is also a consequence of Lemma 1.2.

3.3. The Quasi-Orbit Lemma. We say a sequence of points $\{(x_n, T_n)\}_{0\leqslant n\leqslant N}$ in $\mathsf{U}S \times \mathbb{R}$ is an (ε, T)-pseudo-orbit if

(1) for all n, the distance $d(x_n\varphi_T(x_{n-1})) \leqslant \varepsilon$,
(2) for all n, the distance $T_n \geqslant T$,

The following follows from a refinement of the arguments used in the Closing Lemma

Lemma 3.2 (Quasi-Orbit Lemma). *For every α, there exists some ε such that every ε-pseudo orbit is α close to an orbit.*

We only sketch the proof. Assume $T > 1$ to avoid having to take too much care of the constants. We start by a preliminary Lemma

Lemma 3.3. *Let x_1, x_2, x_3 be three points and T_1, T_2 two numbers such that $\varphi_{T_i}(x_i)$ is ε close to x_{i+1}.*

Then there exists y and S_1, with $|S_1 - T_1| \leqslant \varepsilon$

$$\begin{aligned}
\forall s,\ \ 0 \leqslant s \leqslant \frac{T_1}{2},\ \ d(\varphi_s(y), \varphi_s(x_1)) &\leqslant e^{-T/2}\varepsilon,\\
\forall s,\ \ \frac{T_1}{2} \leqslant s \leqslant T_1,\ \ d(\varphi_s(y), \varphi_s(x_1)) &\leqslant \varepsilon,\\
\forall s,\ \ S_1 \leqslant s \leqslant S_1 + \frac{T_2}{2},\ \ d(\varphi_s(y), \varphi_{T+s}(x_2)) &\leqslant \varepsilon,\\
\forall s,\ \ S_1 + \frac{T_2}{2} \leqslant s \leqslant S_1 + T_2,\ \ d(\varphi_s(y), \varphi_{T+s}(x_2)) &\leqslant e^{-T/2}\varepsilon.
\end{aligned}$$

Proof. The proof of this assertion now follows from similar ideas to the proof of the Closing Lemma: we can write

$$x_2 = H_u^- \circ H_v^+ \circ \varphi_{T+w}(x_1),$$

with $|u|$, $|v|$, $|w|$ smaller than ε.

Then we take $y = \varphi_{-T}(z)$ where $z = H_v^+ \circ \varphi_{T+w}(x_1)$. The assertion follows from the contraction property. □

We can now proceed to the proof of the Quasi-Orbit Lemma.

Proof. We give the rough idea. Assume now for simplicity that $N = 2^p$ and let $\{x_n\}_{0\leqslant n\leqslant N}$ in US be an (ε, T) pseudo orbit, with all $T_n = T$. The assertion above tells us that we can produce a is a $(\varepsilon(1+e^{-T}), 2T)$ pseudo orbit $\{y_2 n\}_{0\leqslant n\leqslant N/2}$. Furthermore the orbit arc $\varphi_{[0,T]}(x_i)$ is $(\varepsilon(1 + e^{-T})$-close to $\varphi_{[0,2T]}(y_{[i/2]})$.

We just continue the induction for one more step:

We produce an $(\varepsilon(1 + e^{-T} + e^{-2T} + e^{-3T}), 4T)$ pseudo orbit $\{z_{4n}\}_{0\leqslant n\leqslant N/4}$ where furthermore $\varphi_{[0,T]}(x_i)$ is $(\varepsilon(1 + e^{-T} + e^{-2T} + e^{-3T})$-close to $\varphi_{[0,4T]}(z_{[i/4]}$.

Continuing the induction, we end up with X_0 and X_N which a (α, NT) pseudo-orbit. We ca so that all x_i are α close to the orbit of X_0, where

$$\alpha = \frac{\varepsilon}{1 - e^{-T}}.$$

□

3.4. Hyperbolic stability at last. Now we can come back to the proof of the Hyperbolic Stability using the Shadow Lemma.

We are going to prove this in the case the corresponding hyperbolic metrics g and g' are close enough. The result would then follow using the fact that the space of hyperbolic metrics is connected.

The proof follows from the following Lemma.

LEMMA 3.4. *Let g and g' two close enough hyperbolic metrics on S. Let γ' be a geodesic of g'. Then there exists a unique geodesic γ for g which is at bounded distance of γ'.*

The uniqueness is obvious: two geodesics at a bounded distance coincide up to a reparametrization.

The conjugacy (check the details) is given by

$$\Psi(\gamma'(+\infty)) = \gamma(+\infty),$$

We leave the reader check the details which are easy:

- Ψ is well defined,
- Ψ is continuous.

We now prove the Lemma

PROOF. We denote by φ_t the geodesic flow of the first metric and by φ'_t the geodesic flow of the second metric. We denote by $\mathsf{U}S$ and $\mathsf{U}S$ the unit tangent bundle for g and g'. Observe that we have a natural map F – linear fiber by fiber – sending U_2S to U_1S.

Using F, we now consider φ_t^2 as a flow on $U_1(S)$. Our hypothesis implies that φ'_1 is ε close to φ_1.

Then every geodesic γ' for g' defines a ε-pseudo orbit of φ_t, which is defined by

$$\{\varphi'_n(\dot\gamma'(0)\}_{n\in\mathbb{N}}.$$

By the Quasi-Orbit Lemma, this ε-pseudo orbit is close to a geodesic γ. □

4. Comments, references and further reading

CHAPTER 5

Measures and Ergodic Theory

We give here as baby course on measures and invariant measures properties of the geodesic flow on surfaces. We refer the avid reader to Martine Babillot for a more thorough introduction to the subject.

1. Generalities on measure

1.1. Radon measures. Let X be a compact Hausdorff topological space. A measure μ on Borel sets defines an integration process from the space of positive measurable functions into $\mathbb{R}$

$$f \mapsto \int_X f \, \mathrm{d}\mu \, .$$

Conversely, by Riesz–Kakutani–Markov Theorem, given any linear functional $f \mapsto F(f)$, defined on the space of positive compactly supported and continuous functions with values in the positive numbers , there exists a finite measure μ, such that

$$F(f) = \int_X f \, \mathrm{d}\mu \, .$$

Moreover $\mu$is unique provided we impose some *regularity* relations, namely for any open set U

$$\mu(U) = \inf \{ \mu(K) \mid K \subset U \, , \, K \text{ compact} \} \, .$$

A finite measure μ satisfying these conditions is called a *Radon measure*. A measure on X is a *probability measure* if $\mu(X) = 1$.

A key result is that given a compact space X, the space $\mathcal{M}(X)$ of Radon probability measures on X is *weakly compact*: given any sequence $\{\mu_n\}_{n\in\mathbb{N}}$ in $\mathcal{M}(X)$, there exists a μ_∞ in $\mathcal{M}(X)$ and an increasing sequence of integers $\{p_n\}_{n\in\mathbb{N}}$ such that for any continuous function on X, we have

$$\int_X f \mathrm{d}\mu_n \underset{n\to\infty}{\longrightarrow} \int_X f \, \mathrm{d}\mu_\infty \, .$$

We have defined Radon measure for compact spaces, the definition extends to locally compact spaces without difficulties.

1.2. Support of a measure. Let μ be a (Probability Radon) measure on a compact topological space X. Then the *support* of μ is the closed set $\mathrm{Supp}(\mu)$ whose complementary is the set

$$\bigcup_{U \in V} U\,,$$

where $V := \{U \mid U \text{ open},\ \mu(U) = 0\}$. In other words for every open set V in $\mathrm{Supp}(\mu)$, $\mu(V) \neq 0$.

2. Invariant measures

Let G be a group acting on a measure space. We say that a measure μ is *invariant under* G if for all elements g in G and measurable subset A of X, we have

$$\mu(A) = \mu(g(A)).$$

Equivalently, for a compact topological space a Radon measure is invariant if for all continuous function f and real g in G, we have

$$\int_X f \,\mathrm{d}\mu = \int_X f \circ g \,\mathrm{d}\mu = \int_X f \circ \,\mathrm{d}(g_*\mu)\,.$$

Exercise 2.1: (*) We consider the action of $\mathsf{PSL}_2(\mathbb{R})$ on $\mathbb{P}^1(\mathbb{R})$.

(1) Show that if μ_0 is invariant under the group generated by a non trivial hyperbolic matrix A in $\mathsf{PSL}_2(\mathbb{R})$, then μ_0 is supported on the two fixed points of A.
(2) Show that there is no $\mathsf{PSL}_2(\mathbb{R})$-invariant measure on $\mathbf{P}^1(\mathbb{R})$.

We state now two important elementary results

2.1. Kakutani–Markov theorem. For 1-parameter flow, we always have an invariant measure.

Theorem 2.1 (Kakutani–Markov theorexm). *Let X be a compact space and $\{\varphi_t\}_{t\in\mathbb{R}}$ a flow of homeomorphisms. Then there exists a $\{\varphi_t\}_{t\in\mathbb{R}}$ invariant measure on X.*

Proof. Let ν be any probability measure on X. Let

$$\nu_t = \frac{1}{t}\int_0^t (\varphi_s)_*\nu \,\mathrm{d}s.$$

Since X is compact, the set $\mathcal{M}(X)$ of Radon probability measures on X is weakly compact. In other words, there exists a probability measure μ on X, a sequence

or real numbers $\{t_n\}_{n\in\mathbb{N}}$ converging to infinity such that for any continuous function f

$$\int_X f \,\mathrm{d}\mu = \lim_{n\to\infty} \int_X f \,\mathrm{d}\nu_{t_n} = \lim_{n\to\infty} \frac{1}{t_n} \int_X \int_0^{t_n} f \circ \varphi_s \,\mathrm{d}\mu \,\mathrm{d}s. \tag{14}$$

Thus in particular for any real number u,

$$\begin{aligned}\int_X f \circ \varphi_u \,\mathrm{d}\mu &= \lim_{n\to\infty} \frac{1}{t_n} \int_X \int_0^{t_n} f \circ \varphi_{s+u} \,\mathrm{d}\mu \,\mathrm{d}s \\ &= \lim_{n\to\infty} \frac{1}{t_n} \int_X \int_u^{t_n+u} f \circ \varphi_s \,\mathrm{d}\mu \,\mathrm{d}s. \end{aligned} \tag{15}$$

It follows that

$$\int_X f \circ \varphi_u \,\mathrm{d}\mu - \int_X f \,\mathrm{d}\mu = \lim_{n\to\infty} \frac{1}{t_n} \int_X \left(\int_0^u f \circ \varphi_s \,\mathrm{d}s - \int_{t_n}^{t_n+u} f \circ \varphi_s \,\mathrm{d}s \right) \mathrm{d}\mu \tag{16}$$

since,

$$\left| \int_0^u \left(\int_X f \circ \varphi_s \mathrm{d}\mu \right) \mathrm{d}s - \int_{t_n}^{t_n+u} \left(\int_X f \circ \varphi_s \mathrm{d}\mu \right) \mathrm{d}s \right| \leqslant 2u \|f\|_\infty,$$

it follows that

$$\int_X f \circ \varphi_u \,\mathrm{d}\mu - \int_X f \,\mathrm{d}\mu = 0. \tag{17}$$

The result follows. □

2.2. Poincaré recurrence theorem. The second result is

THEOREM 2.2 (POINCARÉ RECURRENCE THEOREM). *Let X be a compact topological space and $\{\varphi_t\}_{t\in\mathbb{R}}$ a flow of homeomorphisms preserving a Radon probability measure μ. Let x be an element* $\mathrm{Supp}(\mu)$*, then for any neighborhood U of x and positive T, there exists $t > T$ and y U such that*

$$\varphi_t(y) \in U.$$

PROOF. Let x and U such as in the theorem. We know that $\mu(U) \neq 0$. Moreover, since μ is invariant by φ_t, for all t $\mu(\varphi_t(U)) \neq 0$. The key observation is that there exists $n \neq p$ such that

$$\varphi_{T^n}(U) \cap \varphi_{T^p}(U) \neq \emptyset\,. \tag{18}$$

Indeed otherwise,

$$1 = \mu(X) \geqslant \mu\left(\bigcup_{n\in\mathbb{N}} \varphi_{T^n}(U)) \right) = \sum_{n\in\mathbb{N}} \mu\left(\varphi_{T^n}(U)\right) = \infty\,.$$

Let us assume $n > p$, then letting $q = n - p$, we have from equation 18

$$U \cap \varphi_{T^{-q}}(U) \neq \emptyset .$$

Let thus $y \in U \cap \varphi_{T^{-q}}(U)$, then by definition $\varphi_{T^p}(y) \in U$ and $y \in U$. The theorem follows. □

3. Ergodicity

The unit tangent bundle of a compact hyperbolic surface S has a probability measure μ_0 – that we call the *Lebesgue* measure – which comes from the Haar measure of UH^2 and which is invariant under the geodesic flow, as well as the horocyclic flows.

We say a probability measure μ is *ergodic* under a flow $\{\varphi_t\}_{t\in\mathbb{R}}$ if for all flow invariant set A, either $\mu(A) = 0$ or $\mu(A) = 1$.

Equivalently, every L^2 function, invariant by G, is constant.

We state in these notes without proof two important results, the first one is relatively easy to prove.

Theorem 3.1 (Ergodic Decomposition Theorem). *Let $\{\varphi_t\}_{t\in\mathbb{R}}$ be a flow acting on a compact space X. We denote by $\mathcal{M}(X)$ the convex set of probability Radon measures on X, and $\mathcal{M}_0(X)$ the subset of $\{\varphi_t\}_{t\in\mathbb{R}}$ invariant measures. Let μ_0 be an element of $\mathcal{M}_0(X)$, then there exists a probability measure ν_0 on $\mathcal{M}_0(X)$ supported on ergodic measures so that*

$$\mu_0 = \int_{\mathcal{M}_0(X)} \mu \, \mathrm{d}\nu_0(\mu).$$

Proof. We only sketch the construction. The set $\mathcal{M}_0(X)$ is a convex set in the dual of $C^0(X)$. To be ergodic for a measure translate in being an extremal point in $\mathcal{M}(X)$ – that is a point which is in no open segment included in $\mathcal{M}_0(X)$. The result follows from the fact that every point in a compact convex set is a barycenter (with respect to some measure) of extremal points according to Krein–Milman Theorem. □

The second one is a deeper result.

Theorem 3.2 (Birkhoff Ergodic Theorem). *Let $\{\varphi_t\}_{t\in\mathbb{R}}$ be a flow acting on a compact space X. Let μ_0 be an ergodic probability Radon measures on X. Let f be a measurable function. Then there exists a set A of full measure in X so that for all $x \in A$, we have*

$$\lim_{t\to\infty}\left(\frac{1}{t}\int_0^t f(\varphi_s(x))\,\mathrm{d}s\right) = \int_X f(x)\,\mathrm{d}\mu(x).$$

The quantity

$$\mathrm{M}f(x,t) := \frac{1}{t}\int_0^t f(\varphi_s(x))\,\mathrm{d}s,$$

is called a *Birkhoff sum*.

We shall only prove in Theorem 5.4 the much easier *Statistical Ergodic* L^2*-Theorem* due to Von Neumann where we further assume that f is in $\mathrm{L}^2(X,\mu_0)$.

We also leave as an exercise the following proposition which follows from Birkhoff ergodic theorem and the ergodic decomposition theorem

Proposition 3.3. *Let μ be a measure invariant by a flow $\{\varphi_t\}_{t\in\mathbb{R}}$, Let f be a continuous function. Then there exists a set of μ full measure A an invariant function M_f on A, such that for all x in A,*

$$\lim_{t\to+\infty}\frac{1}{t}\int_0^t f\circ\varphi_s(x)\,\mathrm{d}s = \lim_{t\to+\infty}\frac{1}{t}\int_0^t f\circ\varphi_{-s}(x)\,\mathrm{d}s = M_f(x). \tag{19}$$

Moreover, if for every f there exists a subset of full measure $B\subset A$ on which M_f is constant, then μ is ergodic.

4. Invariant measures by the geodesic flow

The unit tangent bundle of the hyperbolic space carries an invariant measure:

4.1. The Lebesgue measure.

More precisely we have

Proposition 4.1. *The unit tangent bundle $\mathsf{U}\mathbf{H}^2$ possess a measure invariant by the left action of $\mathrm{Iso}(\mathbf{H}^2)$ and the right action of $\mathsf{PSL}_2(\mathbb{R})$.*

We call this measure the *Haar measure* or the *Lebesgue measure*. We will use freely the following result: assume that ω is a non vanishing volume on a manifold M, this measure defines a measure μ_ω by

$$\int_X f\,\mathrm{d}\mu_\omega := \int_X f\omega\,,$$

where the right-hand side[1] is the integration of the form $f\omega$ on the manifold M, oriented by ω.

Proof. In order to define a measure, we will first define a volume form on $\mathsf{SL}_2(\mathbb{R})$ invariant by the left and right actions of $\mathsf{SL}_2(\mathbb{R})$ that we consider as a subset of the vector space of 2×2 matrices. Let $\mathfrak{s}$ be the Lie algebra of $\mathsf{PSL}_2(\mathbb{R})$,

[1] By abuse of language, one uses the same notation for two non-equivalent integration procedures

that is the tangent space to $\mathsf{SL}_2(\mathbb{R})$ at the identity. The vector space $\mathfrak{s}$ identifies with the vector space of traceless 2×2 matrices.

Let ω_0 be a non trivial volume form on $\mathfrak{s}$. For instance we may choose

$$\omega_0(A, B, C) = \text{Trace}([A, B]C) .$$

Observe that for any g in $\mathsf{SL}_2(\mathbb{R})$

$$\omega_0(gAg^{-1}, gBg^{-1}, gCg^{-1}) = \text{Trace}([A, B]C) . \tag{20}$$

Observe that the tangent space to $\mathsf{SL}_2(\mathbb{R})$ at a point g can be interpreted in two ways:

$$\mathsf{T}_g\mathsf{SL}_2(\mathbb{R}) = g\mathfrak{s} = \{gA \mid A \in \mathfrak{s}\} = \mathfrak{s}g = \{bg \mid A \in \mathfrak{s}\} .$$

We define now a volume form on $\mathsf{T}_g\mathsf{SL}_2(\mathbb{R})$ by

$$\omega_g(Ag, Bg, Cg) = \omega_0(A, B, C),$$

and thus a volume form ω on $\mathsf{SL}_2(\mathbb{R})$ by

$$g \mapsto \omega_g .$$

Let denote the right (respectively left) multiplication by g, R_g (respectively L_g). We now want to prove that

$$R_g^*\omega = L_g^*\omega = \omega .$$

or more explicitly for g and h in $\mathsf{SL}_2(\mathbb{R})$,

$$\omega_{gh}(Agh, Bgh, Cgh) = \omega_g(Ag, Bg, Cg) , \tag{21}$$

$$\omega_{hg}(hAg, hBg, hCg) = \omega_g(Ag, Bg, Cg) . \tag{22}$$

We now observe that equation (21) derives immediately from the definition of ω, while for equation (22) we have

$$\begin{aligned}\omega_{hg}(hAg, hBg, hCg) =& \omega_{hg}(hAh^{-1}hg, hBh^{-1}hg, hCh^{-1}hg)\\ =& \omega_0(hAh^{-1}, hBh^{-1}, hCh^{-1})\\ =& \omega_0(A, B, C)\end{aligned}$$

The existence of this left and right invariant volume form gives rise to an left and right invariant measure μ_0 on $\mathsf{PSL}_2(\mathbb{R})$ hence a left and right invariant measure μ on UH^2: more precisely we define for a subset A in UH^2,

$$\mu(A) = \mu_0(A\xi_0) ,$$

where ξ_0 is any element of UH^2 seen as the set of isometries from the upper half plane model $\mathbf{H}^2{}_P$ to $\mathbf{H}^2$. This definition is independent on the choice of ξ_0:

indeed any other element ξ_1 of $\mathsf{U}\mathbf{H}^2$ is equal to $\xi_0 g$, and for every subset B of $\mathsf{PSL}_2(\mathbb{R})$

$$\mu_0(Bg) = \mu_0(B) .$$

The left invariance under the action of $\mathrm{Iso}(\mathbf{H}^2)$ and the right invariance under the action of $\mathsf{PSL}_2(\mathbb{R})$ then comes from the left and right in invariance of μ_0 by $\mathsf{PSL}_2(\mathbb{R})$

□

Exercise 4.1: The following exercise is needed in the next one: there is no group homomorphism f from $\mathsf{PSL}_2(\mathbb{R})$ to $\mathbb{R}$. Recall the commutator of a and b in a group is $[a, b] := aba^{-1}b^{-1}$.

(1) by computing the commutator of a diagonal matrix and a unipotent triangular matrix, show that every unipotent matrix is a commutator, hence in $\ker(f)$,
(2) by computing the commutator of a lower triangular matrix with a upper one, show that every diagonal matrix is in $\ker(f)$,
(3) Conclude using exercise 2.3.
(4) *Remark:* This exercise is much easier when you know that every closed subgroup of a Lie group, is a Lie group itself.

Exercise 4.2: (*) Here is another construction of the Lebesgue measure: we identify $\mathsf{U}\mathbf{H}^2$ with $\mathbf{H}^2 \times \partial_\infty \mathbf{H}^2$. For each x in $\mathbf{H}^2$, the stabilizer K_x of x in $\mathrm{Iso}_+(\mathbf{H}^2)$ is a compact group, isomorphic to S^1, acting freely and transitively on $\partial_\infty \mathbf{H}^2$. It follows that, for every x in $\mathbf{H}^2$, there is a unique probability measure ν_x on $\partial_\infty \mathbf{H}^2$ invariant by K_x. We then define the Radon measure μ_0 on $\mathsf{U}\mathbf{H}^2$ by

$$\int_{\mathsf{U}\mathbf{H}^2} f(x, y)\, \mathrm{d}\mu_0(x, y) = \int_{\mathbf{H}^2} \left(\int_{\partial_\infty \mathbf{H}^2} f(x, y)\, \mathrm{d}\nu_x(y) \right) \mathrm{d}\sigma(x) ,$$

where σ is the area on the hyperbolic plane.

(1) Show that μ_0 is a measure on $\mathsf{U}\mathbf{H}^2$, invariant by $\mathrm{Iso}(\mathbf{H}^2)$.
(2) Show that μ_0 is the unique (up to multiplication) measure on $\mathsf{U}\mathbf{H}^2$ invariant by $\mathrm{Iso}(\mathbf{H}^2)$ and in the Lebesgue class – using the transitivity of the action of $\mathrm{Iso}_+(\mathbf{H}^2)$ on $\mathsf{U}\mathbf{H}^2$.
(3) Let A be an element of $\mathsf{PSL}_2(\mathbb{R})$ acting (on the right on $\mathsf{U}\mathbf{H}^2$. Show that $A_*\mu_0 = d(A)\mu_0$, where $d : A \mapsto d(A)$ is a morphism of $\mathsf{PSL}_2(\mathbb{R})$ in $\mathbb{R}$.
(4) Using the fact that there is no non-trivial morphism from $\mathsf{PSL}_2(\mathbb{R})$ to $\mathbb{R}$ (cf exercise 4.2), show that μ_0 is also invariant under the right action of $\mathsf{PSL}_2(\mathbb{R})$.

(5) Observe that for this measure
$$\mu(\mathsf{U}S) = \sigma(S)\ ,$$
where σ is the hyperbolic area measure.

Exercise 4.3:

(1) Prove that any $\mathsf{PSL}_2(\mathbb{R})$-invariant measure on UH^2 in the Lebesgue class is a multiple of the Lebesgue measure.
(2) (**)(Difficult one) Prove that any $\mathsf{PSL}_2(\mathbb{R})$-invariant measure on UH^2 is a multiple of the Lebesgue measure. *Hint:* use the fact that any measure on $\mathbb{R}$ invariant under translation is in the Lebesgue measure.

4.2. Other invariant measures. Any closed geodesic γ defines an invariant measure by the geodesic flow on $\Gamma\backslash\mathsf{UH}^2$. This measure μ_γ is the unique invariant probability measure whose support is that closed geodesic and is sometimes called the *Dirac measure* supported on the closed geodesic. This measure is defined as follows: for any continuous function f for any x in γ, we define
$$\int_X f\ \mathrm{d}\mu_\gamma = \frac{1}{\ell(\gamma)} \int_0^{\ell(\gamma)} f \circ \varphi_s(x)\ \mathrm{d}s.$$

4.3. Hopf argument. The rest of this paragraph is devoted to a purely dynamical proof of the following result.

Theorem 4.2. *Let S be a finite volume surface. Then the Lebesgue measure is ergodic with respect to the geodesic flow.*

In Paragraph 5.1, we shall give a proof of this result using considerations on unitary representations of $\mathsf{SL}_2(\mathbb{R})$ on $\mathrm{L}^2(\mathsf{U}S, \mu)$ and the fact that the geodesic flow comes from an action of $\mathsf{PSL}_2(\mathbb{R})$. However, we feel that another more general argument is due

We are now giving a proof of the Theorem which can be extended (with some extra work) to general Anosov flows preserving a volume form, without assuming the action of a larger group – in the hyperbolic surface case: $\mathsf{PSL}_2(\mathbb{R})$.

Proof of Theorem 4.3. We shall use a weak consequence of the Anosov property. We have these three foliations $\mathcal{L}^+$ the stable foliation, $\mathcal{L}^-$ the unstable foliation, and $\mathcal{L}^0$ the foliation by the orbit of the geodesic flow $\{\varphi_t\}_{t\in\mathbb{R}}$. We denote by $\mathcal{L}^*_x$ the leaf of $\mathcal{L}^*$ passing through x. These foliations are locally a product, meaning that we can find for every x a neighborhood U of x so that we have the identification
$$U = (\mathcal{L}^+_x \cap U) \times (\mathcal{L}^-_x \cap U) \times (\mathcal{L}^0_x \cap U)$$

and that in this identification the three foliations come from the product structure.

Moreover (this is an important feature) these foliations are absolutely continuous with respect to the Liouville measure. This means that at least locally we can decompose the Liouville measure λ can be written in the coordinates that gives the product structure as

$$\lambda = \lambda^+ \otimes \lambda^- \otimes \lambda_0$$

This property, called the absolute continuity of the stable and unstable foliations is obvious in our case. In the general case of Anosov flows, this is a difficult theorem by Anosov. Assuming this theorem, ergodicity follows from the same scheme of ideas.

We now use Proposition 3.3 and consider the function M_f for a continuous function f. By assumption M_f is constant along the leaves of $\mathcal{L}_0$. We now prove that M_f is constant along the leaves of $\mathcal{L}^+$.

Since US is compact, f is uniformly continuous. Thus f is bounded by K, and that for every ε there exists α such that

$$d(u,v) \leqslant \alpha \implies |f(u) - f(v)| \leqslant \varepsilon.$$

Now let x and y belong to A and the same leaf of $\mathcal{L}^+$. In particular by definition, when $t > t_0$

$$d(\varphi_t(x), \varphi_t(y)) \leqslant \alpha.$$

It thus follows that considering the Birkhoff sums for $t > t_0$, we get

$$\begin{aligned} |Mf(x,t) - Mf(y,t)& \\ &\leqslant \frac{1}{t}\int_0^{t_0} \left|f\circ\varphi_s(x) - f\circ\varphi_s(y)\right| \,\mathrm{d}s + \frac{1}{t}\int_{t_0}^{t} \left|f\circ\varphi_s(x) - f\circ\varphi_s(y)\right| \,\mathrm{d}s \\ &\leqslant \frac{K.t_0}{t} + \varepsilon\ . \end{aligned}$$

It follows that for all ε

$$|M_f(x) - M_f(y)| \leqslant \varepsilon.$$

And thus for all x, y in the same leaf of $\mathcal{L}^+$ then $M_f(x) = M_f(y)$. A similar argument works for $\mathcal{L}^-$.

Now we leave as an exercise the proof of the following fact: since A is of full measure and the foliations are absolutely continuous with respect to the Liouville measure, locally there exist three sets of full λ^*-measure B^* in $\mathcal{L}^*_x$ so that

$$B = B^+ \times B^- \times B^0 \subset A.$$

Then M_f is constant on B. Since B has full measure, this means by Proposition 3.3 that λ is ergodic. □

Actually, you can check that it suffices to use the statistical ergodic Theorem.

5. Ergodicity and mixing: spectral approach

Let G be a topological group acting on a compact space X, preserving a measure μ, we say that the action of G is *mixing* if for any measurable sets A and B, and sequence $\{g_m\}_{m\in\mathbb{N}}$ of elements of G leaving any compact set in G, then

$$\lim_{n\to\infty} \mu\left(A \cap g_n(B)\right) = \mu(A)\mu(B)\,.$$

In probabilistic terms, the events "*belonging to A*" and "*the g_n-image belongs to B*" become independent when n goes to infinity.

We have an alternative way of stating the mixing property: if ψ and φ are L^2 functions on X, then for any sequence $\{g_m\}_{m\in\mathbb{N}}$ in G leaving any compact we have

$$\lim_{m\to\infty} \int_X (\varphi \circ g_m)\, \psi \, \mathrm{d}\mu = \int_X \varphi \, \mathrm{d}\mu \int_X \psi \, \mathrm{d}\mu.$$

Mixing is an important concept that comes in different flavors. What is given here is the definition of *strong mixing* of *strong 2-mixing*; but they are other types of related notions: *weak mixing*, *topological mixing*, *strong m-mixing*, *exponential mixing*...

Just a note, here we introduce mixing for a probability measure, obviously this definition extends to finite measures. We just have to to introduce $\mu(X)$, and the corresponding property is that for all measurable sets A and B

$$\lim_{n\to\infty} \mu\left(A \cap g_n(B)\right) = \frac{\mu(A)\mu(B)}{\mu(X)}\,.$$

Being mixing is stronger than being ergodic.

Proposition 5.1. *Every mixing action is ergodic.*

Proof. Let φ be an invariant L^2-function, and g an element of the group. Then we have the equality

$$\int_X (\varphi \circ g)\, \varphi \, \mathrm{d}\mu = \int_X |\varphi|^2 \, \mathrm{d}\mu\,.$$

Assuming mixing we obtain the equality

$$\int_X |\varphi|^2 \, \mathrm{d}\mu = \left| \int_X \varphi \, \mathrm{d}\mu \right|^2.$$

Thus φ is constant: it is enough to check that for function whose integral over X is zero. Applying that to the characteristic function on a set gives the result. □

We now prove the following result

THEOREM 5.2. *The geodesic and horocyclic flows are mixing.*

5.1. The L^2-approach. We recall that a representation π of a topological group G on a Hilbert space $\mathcal{H}$ is *unitary and strongly continuous*, if

(1) for every g in G, $\pi(g)$ is a unitary operator ion $\mathcal{H}$,
(2) and moreover for every f in $\mathcal{H}$ the map $g \mapsto \pi(g) \cdot f$ is continuous.

We have the basic example

PROPOSITION 5.3. *Assume a topological group G acts by homeomorphisms on a compact space X preserving a measure μ. Then there is a unitary and strongly continuous representation π on $L^2(X, \mu)$ given on a continuous function f by*

$$\pi(g)f = f \circ g^{-1} .$$

PROOF. Recall the the space $C^0(X)$ of continuous functions on X is dense in $L^2(X, \mu)$. Then for every continuous function h, we have that

$$\|h \circ g^{-1}\|_2 = \int_X |h \circ g^{-1}| \, \mathrm{d}\mu = \int_X |h| \, \mathrm{d}\mu = \|h\|_2 ,$$

where for the second equality we used that g preserves the measure μ. It follows by density that if sequence $\{h_m\}_{m\in\mathbb{N}}$ in $C^0(X)$ converges to h in $L^2(X, \mu)$, then $\{h_m \circ g^{-1}\}_{m\in\mathbb{N}}$ is a Cauchy sequence, hence converges to an element that we define as $\pi(g)h$.

By density it follows that $\pi(g)$ is unitary for any element g of G. To check the strong continuity we proceed again by density. Let first h be an element of $C^0(X)$, and $\{g_m\}_{m\in\mathbb{N}}$ a sequence of element of G converging to g, then it is a classical fact that $h \circ g_m^{-1}$ converges in $C^0(X)$ – and in particular in $L^2(X, \mu)$ – to $h \circ g^{-1}$. Given now a general element h of $L^2(X, \mu)$, we first find a sequence $\{h_p\}_{p\in\mathbb{N}}$ converging to h in $L^2(X, \mu)$. Then we write

$$\begin{aligned}
&\|\pi(g)h - \pi(g_m)h\|_2 \\
&\leqslant \|\pi(g)h - \pi(g)h_p\|_2 + \|\pi(g)h_p - \pi(g_m)h_p\|_2 + \|\pi(g_m)h - \pi(g_m)h_p\|_2 \\
&\leqslant 2\|h - h_p\|_2 + \|\pi(g)h_p - \pi(g_m)h_p\|_2 .
\end{aligned}$$

Then we let m going to infinity, to obtain that

$$\limsup_{m\to\infty} (\|\pi(g)h - \pi(g_m)h\|_2) \leqslant 2\|h - h_p\|_2 .$$

Since this last inequality is true for all p, we have

$$\limsup_{m\to\infty} (\|\pi(g)h - \pi(g_m)h\|_2) = 0 .$$

Thus the representation is unitary and strongly continuous. □

5.2. The Statistical Ergodic Theorem. As a first utilization of the L^2 approach we have the the Statistical Ergodic Theorem proved by Von Neumann

THEOREM 5.4 (STATISTICAL ERGODIC THEOREM). *Let $\{\varphi_t\}_{t\in\mathbb{R}}$ be a flow acting on a compact space X. Let μ_0 be an ergodic probability Radon measures on X. Let f be an a function in $L^2(X,\mu)$. Then there exists a set A of full measure in X so that for all x in A, we have*

$$\lim_{t\to\infty}\left(\frac{1}{t}\int_0^t f(\varphi_s(x))\,\mathrm{d}s\right) = \int_X f(x)\,\mathrm{d}\mu(x).$$

PROOF. Since the Radon probability measure μ is invariant by $\{\varphi_t\}_{t\in\mathbb{R}}$, then for each φ_t, the operator U_t, which is given by postcompostion by φ_t acts by isometries on $L^2(X,\mu)$. To be ergodic means that the there exists a non-zero s, such that U_s that has no invariant vectors in $L^2_0(X,\mu)$ the subspace of functions whose integral is zero. It follows that

$$\mathrm{Id} - U_{-s} ,$$

has a dense image. Indeed

$$\left(\overline{\mathrm{Image}(\mathrm{Id} - U_{-s})}\right)^{\perp} = \ker(\mathrm{Id} - U_s) = \{0\} .$$

Let then f be a function in $L^2_0(X,\mu)$ and let us consider

$$f_t : x \mapsto \frac{1}{t}\int_0^t f\circ\varphi_s(x)\,\mathrm{d}s .$$

Assume first that $f = g - U_{-s}(g)$. Then

$$\begin{aligned}\frac{1}{t}\int_0^t U_u(f)\,\mathrm{d}u &= \frac{1}{t}\int_0^t U_u(g) - U_{u-s}g)\,\mathrm{d}u\\ &= \frac{1}{t}\left(\int_{-s}^0 U_u(g)\,\mathrm{d}u - \int_t^{t-s} U_u(g)\,\mathrm{d}u\right) \xrightarrow[t\to\infty]{L^2} 0 .\end{aligned}$$

In general, we write $f = \lim_{n\to\infty} f_n$, where f_n belong to the image of $\mathrm{Id} - U_s$. Then

$$\left\|\frac{1}{t}\int_0^t U_u(f)\,\mathrm{d}u - \frac{1}{t}\int_0^t U_u(f_n)\,\mathrm{d}u\right\|_2 \leqslant \|f - f_n\|_2 .$$

Thus for every positive ε, for n large enough,

$$\left\| \frac{1}{t} \int_0^t U_u(f) \, \mathrm{d}u - \frac{1}{t} \int_0^t U_u(f_n) \, \mathrm{d}u \right\|_2 \leqslant \varepsilon$$

But, we have show above that for a fixed n

$$\left\| \frac{1}{t} \int_0^t U_u(f_n) \, \mathrm{d}u \right\|_2 \underset{t\to\infty}{\longrightarrow} 0$$

and thus

$$\limsup_{t\to\infty} \left\| \frac{1}{t} \int_0^t U_u(f) \, \mathrm{d}u \right\|_2 \leqslant \varepsilon \, .$$

The result follows by taking ε as small as we want.

Thus $\{f_t\}_{t\in\mathbb{R}}$ converges to zero in $\mathrm{L}^2(X,\mu)$, hence it converges point-wise to zero almost everywhere.

□

Let X be a space equipped with a probability measure μ. Let $\{\varphi_t\}_{t\in\mathbb{R}}$ be a flow preserving μ acting on X. Let $\mathrm{L}^2_0(X,\mu)$ be the vector subspace of $\mathrm{L}^2(X,\mu)$ consisting of functions whose integral is zero. Observe that any measure preserving mapping f from X to X define a unitary operator A_f on $\mathrm{L}^2_0(X,\mu)$ by $A_f : g \to g \circ f$. Let $U_t = A_{\varphi_t}$.

We now observe the following

Proposition 5.5 (L^2 interpretation). *The flow $\{\varphi_t\}_{t\in\mathbb{R}}$ is ergodic if and only if the one parameter group $\{U_t\}_{t\in\mathbb{R}}$ has no non-trivial invariant vectors on $\mathrm{L}^2_0(X,\mu)$.*

The flow $\{\varphi_t\}_{t\in\mathbb{R}}$ is mixing if and only for any function f and g in $\mathrm{L}^2_0(X,\mu)$ we have

$$\lim_{t\to\infty} \langle U_t.f, g\rangle = 0.$$

Thus Theorem 5.2 (as well as the ergodicity of any non-compact closed subgroup!) follows at once from the following result

Theorem 5.6 (Howe–Moore). *Assume that we have a strongly continuous unitary representation π of $\mathsf{SL}_2(\mathbb{R})$ on a Hilbert space $\mathcal{H}$. Assume that π has no non-trivial invariant vector, then if $\{g_n\}_{n\in\mathbb{N}}$ is a diverging sequence in $\mathsf{SL}_2(\mathbb{R})$ then for any u and v in $\mathcal{H}$,*

$$\lim_{t\to\infty} \langle \pi(g_n)u, v\rangle = 0.$$

Corollary 5.7. *If the group $\mathsf{SL}_2(\mathbb{R})$ acts ergodically on a space X preserving probability measure, then every non-compact subgroup acts ergodically and is mixing.*

In particular since $\mathsf{PSL}_2(\mathbb{R})$ acts transitively on $\mathsf{U}S$, it acts ergodically and Theorem 5.2 follows.

5.3. Proof of Howe–Moore's Theorem. In the whole proof, we will write $n_t \coloneqq n_t^+$. Our first lemma is the following

LEMMA 5.8 (MAUTNER PHENOMENON). *Let u be an element in $\mathcal{H}$ so that $\{a_{t_n}u\}_{n\in\mathbb{R}}$ weakly converges to u_0, where $\{t_n\}_{n\in\mathbb{R}}$ goes to infinity.*
Then u_0 is invariant under the one parameter group $\{n_t\}_{t\in\mathbb{R}}$.

Recall that u_n *weakly converges* to u, if for all z,

$$\lim_{n\to\infty} \langle u_n, z\rangle = \langle u, z\rangle$$

PROOF. We have for any g

$$\begin{aligned}
|\langle \pi(n_s)u_0, v\rangle - \langle u_0, v\rangle| &= \lim_{k\to\infty} \left(|\langle \pi(n_s a_{t_k})u, v\rangle - \langle \pi(a_{t_k})u, v\rangle|\right) \\
&= \lim_{k\to\infty} \left(|\langle \pi(a_{-t_k} n_s a_{t_k})u, \pi(a_{-t_k})v\rangle - \langle u, \pi(a_{-t_k})v\rangle|\right) \\
&\leqslant \lim_{k\to\infty} \left(\|\pi(a_{-t_k} n_s a_{t_k})u - u\|\right).\|v\| = 0,
\end{aligned}$$

where the first equality comes from the hypothesis. Since

$$\lim_{k\to\infty} a_{-t_k} n_s a_{t_k} = \lim_{k\to\infty} n_{e^{-t_k}s} = 1,$$

thus for all v, by the definition of the the fact that the representation of $\mathsf{PSL}_2(\mathbb{R})$ is unitary and strongly continuous.

$$|\langle \pi(n_s)u_0, v\rangle - \langle u_0, v\rangle| = 0\ ,$$

and the result follows. □

Our second lemma is the following

LEMMA 5.9. *Let u be an element in $\mathcal{H}$ invariant under the one parameter group $\{n_t\}_{t\in\mathbb{R}}$. Then u is invariant by $\mathsf{SL}_2(\mathbb{R})$.*

PROOF. Let u be any vector. Let φ_u be the continuous function on $\mathsf{SL}_2(\mathbb{R})$ given by

$$u \mapsto \varphi_u(g) \coloneqq \langle \pi(g)u, u\rangle\ .$$

Then the following are equivalent for any closed subgroup G of $\mathsf{SL}_2(\mathbb{R})$

(1) φ_u is constant on G,
(2) φ_u is G-bi-invariant: for all h in G and g in $\mathsf{PSL}_2(\mathbb{R})$, then $\varphi_u(hg) = \varphi_u(gh) = \varphi_u(g)$
(3) f is $\pi(\mathsf{G})$-invariant.

The implications (3) $\implies$ (2) $\implies$ (1) are obvious. Then (1) implies (3). Indeed, if we assume (1), the first for all g in G, $\langle \pi(g)u, u\rangle = \langle u, u\rangle$ hence

$$\|\pi(g)u - u\|^2 = 2\|\langle u, u\rangle - \langle \pi(g)u, u\rangle)\|^2 = 0\ .$$

Observe that in this last implication, we have used in a crucial manner that $\pi(g)$ is a unitary operator.

We can now proceed to the proof of the Lemma. Let f be a N-invariant vector. Let P be the group generated by N and $\{a_t\}_{t\in\mathbb{R}}$.

Since u is invariant by $\mathsf{N} = \{n_t\}_{t\in\mathbb{R}}$, it follows form the initial remark that φ_u is a left and right N invariant function on $\mathsf{SL}_2(\mathbb{R})$.

Any left and right N invariant function on $\mathsf{SL}_2(\mathbb{R})$ give rise to a left invariant N continuous function on $\mathsf{SL}_2(\mathbb{R})/\mathsf{N}$. However $\mathsf{SL}_2(\mathbb{R})/\mathsf{N}$ together with left action by $\mathsf{SL}_2(\mathbb{R})$ is identified with $\mathbb{R}^2 \setminus \{0\}$: for this identification we identify $\mathsf{SL}_2(\mathbb{R})$ as the set of basis (of determinant 1) of $\mathbb{R}^2$. Then the left N-orbits are points on the horizontal axis, and horizontal lines. It follows by continuity that φ_u is constant on the horizontal axis, which is precisely the image of P.

Similarly, using the remark again, φ_u is now a left and right P invariant function on $\mathsf{SL}_2(\mathbb{R})$, hence a P-invariant function on $\mathsf{SL}_2(\mathbb{R})/\mathsf{P} = \mathbf{P}^1(\mathbb{R})$. Thus since P has a dense orbit on $\mathsf{SL}_2(\mathbb{R})/\mathsf{P}$ we obtain that φ_u is constant. That is what we wanted to prove. □

Proof of Howe–Moore Theorem. Let U and V be two vectors in $\mathcal{H}$, $\{g_m\}_{m\in\mathbb{N}}$ be a divergent sequence in $\mathsf{SL}_2(\mathbb{R})$. We want to prove that

$$\langle \pi(g_m)U, V\rangle \underset{m\to\infty}{\longrightarrow} 0\ .$$

The sequence of real numbers $\{\langle \pi(g_m)U, V\rangle\}_{m\in\mathbb{N}}$ is bounded by $\|U\|\|V\|$, thus we can extract a subsequence so that it converges to a real number λ.

Observe that any matrix B in $\mathsf{PSL}_2(\mathbb{R})$ can be written as

$$B = K^0\, A\, K^1\ ,$$

where K^0 and K^1 are rotations (elements of $\mathsf{SO}(2)$) and A is a diagonal matrix[2]. Thus, we can write

$$g_m = (k_m^0)^{-1}\, a_{t_m} k_m^1,$$

where k_m^0 and k_m^1 belongs to $\mathsf{SO}(2)$. Thus we can as well assume that $\{k_m^0\}_{m\in\mathbb{N}}$ and $\{k_m^1\}_{m\in\mathbb{N}}$ converges to k^0 and k^1 respectively.

Let $u \coloneqq \pi(k^1)U$ and $v \coloneqq \pi(k^0)V$, we first show that we also have.

$$\langle \pi(a_{t_m})u, v\rangle \underset{m\to\infty}{\longrightarrow} \lambda\ .$$

[2]This decomposition is known in general semi-simple Lie groups as the *Cartan* or KAK *-decomposition*.

Indeed, letting $U_m := \pi(k_m^0)U$ and $V_m := \pi(k_m^0)V$ we have, using unitarity and Cauchy–Schwarz Inequality,

$$\begin{aligned}\left|\langle \pi(a_{t_m})u, v\rangle - \langle \pi(g_m)U, V\rangle\right| &= \left|\langle \pi(a_{t_m})u, v\rangle - \langle \pi(a_{t_m})U_m, V_m\rangle\right| \\ &\leqslant \left|\langle \pi(a_{t_m})u, v\rangle - \langle \pi(a_{t_m})U_m, v\rangle\right| + \left|\langle \pi(a_{t_m})U_n, v\rangle - \langle \pi(a_{t_m})U_m, V_m\rangle\right| \\ &\leqslant \|v\|\,\|U_m - U\| + \|V_m - v\|\,\|U\| \\ &= \|v\|\,\|U_m - u\| + \|V_m - v\|\,\|V\|\,,\end{aligned}$$

and we conclude by using the strong continuity of π which implies that

$$U_m \xrightarrow[m\to\infty]{\mathrm{L}^2} U \text{ and } V_m \xrightarrow[m\to\infty]{\mathrm{L}^2} v\,.$$

We can now continue the proof: since $\{\pi(a_{t_m})u\}_{m\in\mathbb{N}}$ is bounded, by the weak compactness theorem, after extracting a subsequence, we can as well assume that $\{\pi(a_{t_m})u\}_{m\in\mathbb{N}}$ converges weakly to u_0.

By the Mautner phenomenon, u_0 is invariant by N. By the second lemma u_0 is invariant by $\mathsf{SL}_2(\mathbb{R})$. Thus $u_0 = 0$ and hence $\lambda = 0$ is achieved. □

Exercise 5.1: Prove the existence of the Cartan decomposition for $\mathsf{PSL}_2(\mathbb{R})$ by considering the (right)-action of $\mathsf{SO}(2)$ and the group of diagonal matrices on $\mathsf{U}\mathbf{H}^2$.

5.4. Unique ergodicity and complements. A flow is said to be *uniquely ergodic* if is possesses a unique invariant measure. By the ergodic decomposition theorem, such a measure is necessarily ergodic. So equivalently a flow is to uniquely ergodic if is possesses a unique ergodic invariant measure. Obviously, the geodesic flow is not uniquely ergodic: all closed geodesics define invariant measures. A deeper result says

Theorem 5.10 (Furstenberg). *The horocyclic flow is uniquely ergodic for a finite volume hyperbolic surface.*

All the previous results have extensions in higher dimensions and for general Anosov flows.

6. Comments, references and further reading

CHAPTER 6

Equidistribution and growth of geodesics

Let $S = \Gamma\backslash\mathbf{H}^2$ be a compact hyperbolic surface, where Γ is a discrete group of $\mathrm{Iso}_+(\mathbf{H}^2)$. We saw that the unit tangent bundle has a probability measure μ_0 – that we call the *Lebesgue* measure – which comes from the Haar measure of $\mathsf{PSL}_2(\mathbb{R})$ and which is invariant under the geodesic flow. Every closed geodesic γ, with *period* of *length* $\ell(\gamma)$ also defines a geodesic flow invariant probability measure μ_γ on $\mathsf{U}S$ by the formula

$$\int f \, \mathrm{d}\mu_\gamma = \frac{1}{\ell(\gamma)} \int_0^{\ell(\gamma)} f(\varphi_t(x)) \, \mathrm{d}t \ .$$

All these measures are related by the following deep result

Theorem 0.1 (Bowen, Margulis). *The closed geodesics are* equidistributed *with respect to the Lebesgue measure:*

$$\lim_{T\to\infty} \frac{1}{\sharp\Gamma_T} \left(\sum_{\gamma\in\Gamma_T} \mu_\gamma \right) = \mu_0,$$

where Γ_T is the set of closed geodesics of length smaller than T.

As one of the first step in the proof, we will show that Γ_T is a finite set. The following theorem will then count asymptotically the number of closed geodesics.

Theorem 0.2 (Margulis). *Let Γ_T be the set of closed geodesics of length smaller than T. Then*

$$\lim_{T\to\infty} 2Te^{-T}\sharp\Gamma_T = 1 \ .$$

We have a similar statement for counting points in an orbit, let us introduce for a point x_0 in $\mathbf{H}^2$,

$$\Gamma_T^0 \coloneqq \{\gamma \in \Gamma \mid d(x_0, \gamma(x_0)) \leqslant T\} \ .$$

Then

THEOREM 0.3 (MARGULIS). *Let x_0 be a point in $\mathbf{H}^2$. Then*

$$\lim_{T\to\infty} \frac{\sharp\Gamma_T^0}{e^T} = \frac{\pi}{\mathrm{Area}(S)},$$

where the area of the hyperbolic surface is computed with respect to the hyperbolic area.

To fix the notation,

(1) Let σ be the hyperbolic measure on S.
(2) Let μ_0 the Lebesgue measure on $\mathsf{U}S$ so that $\mu_0(S) = 1$.
(3) Let x be a point in $\mathbf{H}^2$. Let ν_x be the probability measure on $\partial_\infty\mathbf{H}^2$ such that ν_x is invariant under the stabilizer of x in $\mathrm{Iso}_+(\mathbf{H}^2)$.
(4) Let π be the projection from $\mathsf{U}\mathbf{H}^2$ to $\mathbf{H}^2$, and denote similarly by π be the projection from $\mathsf{U}S$ to S, as well as p the projection from $\mathsf{U}\mathbf{H}^2$ to $\mathsf{U}S$ and by an abuse of language from $\mathbf{H}^2$ to S.

1. Equidistribution of circles

1.1. Equidistribution of circles in the unit bundle and in the surface. Let C_0 be the preimage of x for the projection $\mathsf{U}\mathbf{H}^2 \to \mathbf{H}^2$, in other words $C_0 = \{x\}\times\partial_\infty\mathbf{H}^2$. Let then μ_0 the probability measure on $\mathsf{U}S$ proportional to the Lebesgue measure.

THEOREM 1.1 (EQUIDISTRIBUTION OF CIRCLES I). *Let Y be an interval in C_0. Let f be continuous function on $\mathsf{U}S$,*

$$\frac{1}{\nu_x(Y)}\int_Y f\circ\varphi_T \,\mathrm{d}\nu_x \underset{T\to\infty}{\longrightarrow} \int_{\mathsf{U}S} f\,\mathrm{d}\mu_0\,.$$

One could observe that we could see this as a "mixing property" between a continuous function and a Radon measure – instead of two L^2-functions, with respect to a measure. However, let us warn the reader that such a general mixing property is not true in general and require a theory of *wavefront*.

As a corollary, we have defining for an interval Y in C_0, Y_T the "circle arc at distance T" in $\mathbf{H}^2$,

$$Y_T := \{z\in\mathbf{H}^2 \mid d(x,z) = T\,,\ \exists y\in Y, \text{ such that } z\in[x,y]\}$$

Then, denoting $\mathrm{d}y$ the arc-length measure on Y,

COROLLARY 1.2 (EQUIDISTRIBUTION OF CIRCLES II). *Let Y be an interval in C_0. Let f be continuous function on S,*

$$\frac{1}{\ell(Y_T)}\int_{Y_T} f\circ p\,\mathrm{d}y \underset{T\to\infty}{\longrightarrow} \frac{1}{\mathrm{Area}(S)}\int_S f\,\mathrm{d}\sigma\,,$$

where p is the projection from $\mathbf{H}^2$ to S.

PROOF. It just follows from the remark that $(\varphi_T)_* \nu_x$ is proportional to the arc length measure on Y_T. □

We could interpret these results as saying the 1-dimensional family of curves Y_T becomes equidistributed in the surface as T goes to infinity.

1.2. Yet another description of UH^2. For the purpose of the construction of the next paragraph, we need to express the Lebesgue measure on UH^2 in some coordinates.

Let us consider the group P of triangular matrix in $\mathsf{PSL}_2(\mathbb{R})$. We then have a natural parametrization from

$$\Psi : \left\{ \begin{array}{ccc} \mathbb{R}^2 & \to & \mathsf{P} \\ (s,t) & \mapsto & a_{2s} n_t = \begin{pmatrix} e^{-s} & t \\ 0 & e^s \end{pmatrix} \end{array} \right. .$$

Observe that this map is a group morphism whenever we define the group structure of $\mathbb{R}^2$ as

$$(s,u)\,(t,v) \coloneqq (s+t, e^{-s}v + e^t u)\,.$$

Then one checks that

LEMMA 1.3. *The measure $\lambda = e^t \mathrm{d}t\ \mathrm{d}v$, is invariant under the left action of P on itself.*

PROOF. Indeed for a compactly supported function f

$$\int_{\mathsf{P}} f((s,u)(t,v))\, e^t \mathrm{d}t \mathrm{d}v = \int_{\mathsf{P}} f(s+t, e^{-s}v + e^t v))\ e^t \mathrm{d}t \mathrm{d}v = \int_{\mathsf{P}} f(t',v')\, e^{t'} \mathrm{d}t' \mathrm{d}v'\,,$$

where we performed the change of variables $(t',v') = (s+t, e^{-s}v + e^t u)$. □

Let now fix a point u_0 in UH^2 whose projection in $\mathbf{H}^2$ is x and let K_x be the stabilizer of x and ν_x its bi-invariant measure (well defined up to a multiplicative constant). Let consider the left-action of P on UH^2 given by

$$p\ y \coloneqq y\, p^{-1}\,.$$

We leave the reader check the

LEMMA 1.4. *The left action of $\mathsf{K}_x \times P$ is transitive on UH^2. This action preserves the Lebesgue measure μ_0 on UH^2.*

Let us consider the orbit map

$$\Psi^0 : \left\{ \begin{array}{ccc} \mathsf{K}_x \times \mathsf{P} & \to & \mathsf{UH}^2 \\ v & \mapsto & v u_0 \end{array} \right.$$

We have

Lemma 1.5. *The push-forward measure*

$$\mu_x := \Psi^0_*(\nu_x \otimes \lambda)$$

is a multiple of the Lebesgue measure, by a constant independent on x.

Proof. Indeed, both measures are invariant by the left action of $\mathsf{K}_x \times \mathsf{P}$ and in the same measure class. Since $\mathsf{K}_x \times \mathsf{P}$ acts transitively on $\mathsf{U}\mathbf{H}^2$, their Radon–Nikodym derivative is constant. The fact that this constant $k(x)$ – seen as a function of x – is also constant is a consequence of the equivariance of the construction under the isometry group of $\mathbf{H}^2$: for every g in $\mathrm{Iso}_+(\mathbf{H}^2)$,

$$g_*\mu_x = \mu(g(x)) \, .$$

Thus $k(g(x)) = k(x)$ and the result follows from the transitivity of the action on $\mathrm{Iso}_+(\mathbf{H}^2)$ on $\mathsf{U}S$. □

We choose a multiple ν^0_x of ν_x such that

$$\Psi^0_*(\nu^0_x \otimes \lambda) = \mu_0 \, .$$

1.2.1. *Proof of the equidistribution of circles.* We use the parametrization Ψ^0 describe in the previous paragraph and describe $\mathsf{U}\mathbf{H}^2$ as $\mathsf{K}_x \times \mathbb{R}^2$. In this description $C_0 = \mathsf{K}_x \times \{(0,0)\}$.

Let us consider the following open set, which depends on the choice of a positive number ε

$$U := Y \times]0, \varepsilon[\times]0, \varepsilon[\, .$$

The open set U is some sort of a thickening of C_0 and we have

$$\mu_0(U) = \nu^0_x(Y) \; \alpha \; (e^\varepsilon - 1) \underset{\varepsilon \to 0}{\sim} \nu_x(Y) \; \alpha \; \varepsilon \, .$$

Proposition 1.6. *When T goes to infinity,*

$$\frac{\int_{\mathsf{U}S} (f \circ \varphi_T) \, \chi_U \, \mathrm{d}\mu_0}{\int_{\mathsf{U}S} \chi_U \mathrm{d}\mu_0} \longrightarrow \int_{\mathsf{U}S} f \, \mathrm{d}\mu_0 \, .$$

Proof. This is an immediate consequence of mixing. □

Let also consider the two dimensional set

$$Z_\varepsilon := Y \times]0, \varepsilon[\times \{0\},$$

that we equip with the measure $\mu_1 := \nu^0_x \otimes e^t \mathrm{d}t$ such that for any continuous function

$$\int_{Z_\varepsilon} g \, \mathrm{d}\mu_1 = \int_{Z_\varepsilon} g(y,t) \, e^t \mathrm{d}\nu^0_x(y) \, \mathrm{d}t \, .$$

As a second crucial proposition, we have

Proposition 1.7. *When T goes to infinity,*

$$\left| \frac{\int_{Z_\varepsilon} f \circ \varphi_T \, \mathrm{d}\mu_1}{\mu_1(Z_\varepsilon)} - \frac{\int_U (f_0 \circ \varphi_T) \ \mathrm{d}\mu_0}{\mu_0(U)} \right| \to 0.$$

Proof. We immediately have that $\mu_0(U) = \varepsilon\mu_1(Z_\varepsilon)$. Observe now that Moreover $d(\varphi_T(k,s,t) - d(\varphi_T(k,s,0)$ converges uniformly to zero by the contraction on the stable horospheres for (k,s,t) in U, the uniform continuity of f yields that

$$\left| f \circ \varphi_T(k,s,0) - f \circ \varphi_T(k,s,t) \right| \underset{T\to\infty}{\longrightarrow} 0 \, ,$$

this convergence being uniform for (k,s,t) in U when T goes to infinity. Thus setting for all t in $]-\varepsilon,\varepsilon[$,

$$Z_\varepsilon(t) \coloneqq Y \times]0,\varepsilon[\times \{t\},$$

$$\int_{Z_\varepsilon} g \, \mathrm{d}\mu_1 = \int_{Z_\varepsilon} g(y,t) \, e^t \mathrm{d}\nu_x^0(y) \, \mathrm{d}t \, .$$

Rephrasing, the uniform continuity of f as given through equation (23), we have when T grows to infinity, uniformly in t

$$\left| \int_{Z_\varepsilon(t)} f \circ \varphi_T \, \mathrm{d}\mu_1 - \int_{Z_\varepsilon} f \circ \varphi_T \, \mathrm{d}\mu_1 \right| \to 0.$$

Now it remains to remark that

$$\mu_0(U) = \varepsilon\mu_1(Z_\varepsilon(t)) = 2\varepsilon\mu_1(Z_\varepsilon)$$

and that by Fubini's theorem

$$\int_U f_0 \circ \varphi_T \, \mathrm{d}\mu_0 = \int_{-\varepsilon}^{\varepsilon} \left(\int_{Z_\varepsilon(t)} f \circ \varphi_T \, \mathrm{d}\mu_1 \right) \mathrm{d}t \, ,$$

to conclude the proof of the theorem. □

Finally we have,

Proposition 1.8. *Given a continuous function f and any β and then for any ε with $\varepsilon \leqslant \varepsilon_0$ and any β then*

$$\left| \frac{\int_Y f \circ \varphi_T \, \mathrm{d}\nu_x}{\nu_x(Y)} - \frac{\int_{Z_\varepsilon} f \circ \varphi_T \, \mathrm{d}\mu_1}{\mu_1(Z_\varepsilon)} \right| \leqslant \beta \, .$$

Proof. This is again a consequence of the uniform continuity of f: indeed there exists a positive α, such that

$$d(z,w) \leqslant \alpha \quad \text{implies} \quad |f(z) - f(y)| \leqslant \beta \,.$$

Then there exists ε_0 such that for all z

$$|s-t| \leqslant \varepsilon_0 \quad \text{implies} \quad d(\varphi_s(z), \varphi_t(z))| \leqslant \alpha \,.$$

These two inequalities implies the result. □

Proof of Theorem 1.1. The result follows from a classic decomposition:

$$\left| \frac{\int_Y f\circ\varphi_T \,\mathrm{d}\nu_x}{\nu_x(Y)} - \int_{US} f \,\mathrm{d}\mu_0 \right| \leqslant \underbrace{\left| \frac{\int_Y f\circ\varphi_T \,\mathrm{d}\nu_x}{\nu_x(Y)} - \frac{\int_{Z_\varepsilon} f\circ\varphi_T \,\mathrm{d}\mu_1}{\mu_1(Z_\varepsilon)} \right|}_{(A)}$$
$$+ \underbrace{\left| \frac{\int_{Z_\varepsilon} f\circ\varphi_T \,\mathrm{d}\mu_1}{\mu_1(Z_\varepsilon)} - \frac{\int_{US}(f_0\circ\varphi_T)\,\chi_U \,\mathrm{d}\mu_0}{\mu_0(U)} \right|}_{(B)}$$
$$+ \underbrace{\left| \frac{\int_{US}(f_0\circ\varphi_T)\,\chi_U \,\mathrm{d}\mu_0}{\mu_0(U)} - \int_{US} f \,\mathrm{d}\mu_0 \right|}_{(C)} .$$

Let β be any positive number. We first choose ε small enough so that the term (A) is smaller than β by Proposition 1.8. Then we choose T large enough so that the second term (B) is smaller than β (by Proposition 1.7) and the third term (C) is smaller than β as well (by Proposition 1.6). This concludes the proof. □

2. Counting in the group

Let Γ be the subgroup of $\mathsf{PSL}_2(\mathbb{R})$ such that $S := \Gamma\backslash\mathbf{H}^2$ is compact. We want to count asymptotically the number of points in the Γ-orbit of x, that is get an asymptotics of the following number when T goes to infinity

$$N_T := \sharp\{\gamma \in \Gamma \mid d(\gamma(x), x) \leqslant T\} \,.$$

Our first goal is to obtain Theorem 0.3, namely that when T goes to infinity

$$N_T \sim \frac{\pi}{\mathrm{Area}(S)} e^T \,. \tag{23}$$

We will actually treat a more general and related counting problem. Let Y be an arc in C_0 as in the previous paragraph, and let us define

$$N_{T,Y} := \sharp\{\gamma \in \Gamma \mid \exists y \text{ in } Y_T \text{ such that } d(\gamma(x), y) \leqslant T\} .$$

This amounts geometrically to counting elements of Γ in the "sector" defined by Y. The refined result is

THEOREM 2.1 (COUNTING POINTS IN A SECTOR). *Let ν_x be the K_x invariant probability measure on $\partial_\infty \mathbf{H}^2$, then*

$$N_{T,Y} \sim \nu_x(Y) \frac{\pi}{\text{Area}(S)} e^T , \tag{24}$$

PROOF. Let x_0 be a point in S and let f be a bump function, that is a positive function of integral 1 supported on some ε-neighborhood of x_0. We have that

$$\frac{1}{\ell(Y_T)} \int_{Y_T} f \circ p \, \mathrm{d}y \underset{T\to\infty}{\longrightarrow} \frac{1}{\text{Area}(S)} ,$$

and we saw in exercise 5.1, that

$$\ell(Y_T) = 2\pi \sinh(T) \nu_x(Y) \underset{T\to\infty}{\sim} \frac{\pi e^T}{\text{Area}(S)} \nu_x(Y) ,$$

And thus

$$\int_{Y_T} f \circ p \, \mathrm{d}y \underset{T\to\infty}{\sim} \frac{\pi e^T}{\text{Area}(S)} \nu_x(Y) , \tag{25}$$

It is actually more convenient here to work in the universal cover and that what we will do now.

Let x so that $p(x) = x_0$ and g a bump function in $\mathbf{H}^2$ of integral 1, with respect to the hyperbolic measure, such that g is supported in an ε-neighborhood of x_0. It follows that

$$f_0 := f \circ p = \sum_{\gamma\in\Gamma} g \circ \gamma .$$

For the sake of simplicity, let first treat the case $Y = C_0$, let then

$$B_T = \{z \in \mathbf{H}^2 \mid d(x_0, z) \leqslant T\} .$$

The following inequality is now obvious

$$N_{T-\varepsilon} \leqslant \int_{B_T} f_0 \, \mathrm{d}\sigma \leqslant N_{T+\varepsilon} . \tag{26}$$

It follows that

$$N_{T-\varepsilon} \leqslant \int_0^T \left(\int_{C_t} f_0 \, \mathrm{d}y \right) \mathrm{d}t \leqslant N_{T+\varepsilon} \, . \tag{27}$$

From the asymptotics 25, we get that for any positive α there is T_0, such that for T greater than T_0, we have

$$\frac{(\pi - \alpha)}{\mathrm{Area}(S)} e^t \leqslant \int_{C_t} f_0 \, \mathrm{d}y \leqslant \frac{(\pi + \alpha)}{\mathrm{Area}(S)} e^t \, .$$

Thus we have the *lower inequality*,

$$\frac{(\pi - \alpha)}{\mathrm{Area}(S)} (e^t - e^{-t_0}) - K_0 \leqslant N_{t+\varepsilon} \, .$$

and after a change of variable, there is a positive constant K_1 such that for T large enough,

$$\frac{e^{-\varepsilon}(\pi - \alpha)}{\mathrm{Area}(S)} e^T - K_1 \leqslant N_T \, . \tag{28}$$

Similarly given any α, we have the *upper inequality*: there exists positive constants K and T_0 so that for all $t \geqslant T_0$

$$N_{T-\varepsilon} \leqslant e^T \frac{(\pi + \alpha)}{\mathrm{Area}(S)} + K \, ,$$

thus after a change of variables there is a positive constant K_2 such that for T large enough,

$$N_T \leqslant \frac{e^{\varepsilon}(\pi + \alpha)}{\mathrm{Area}(S)} e^t + K_2 \, .$$

It follows that for all positive ε and α, we have

$$\limsup_{T \to \infty} \frac{N_T}{e^T} \leqslant \frac{1}{\mathrm{Area}(S)} e^{\varepsilon} (\pi + \alpha) \, ,$$

$$\liminf_{T \to \infty} \frac{N_{T,\gamma}}{e^T} \geqslant \frac{1}{\mathrm{Area}(S)} e^{-\varepsilon} (\pi - \alpha) \, ,$$

Thus, since this is true for all positive α and ε, we obtain the upper and lower inequalities for $\frac{N_T}{e^T}$:

$$\limsup_{T \to \infty} \frac{N_T}{e^T} \leqslant \frac{\pi}{\mathrm{Area}(S)} \, , \tag{29}$$

$$\liminf_{T \to \infty} \frac{N_T}{e^T} \geqslant \frac{\pi}{\mathrm{Area}(S)} \, . \tag{30}$$

These two inequalities definitely imply that

$$\lim_{T\to\infty} \frac{N_T}{e^T} = \frac{\pi}{\mathrm{Area}(S)} , \tag{31}$$

as required.

Let us now treat the general case. Let Y and Y^0 be two open intervals such that Y^0 is a proper subinterval in Y:

$$\overline{Y^0} \subset Y .$$

Let similarly

$$B_{T,Y} = \{z \in \mathbf{H}^2 \mid d(x,z) \leqslant T \text{ and } \exists y \in Y,\ z \in [z,y]\} .$$

We have to rethink inequality (26). Since Y^0 is a strict subset of Y, we obtain that given any ε, for T large enough, the ε-neighborhood of Y^0_T (in C_T) is included in Y_T, thus we have to replace inequality (26) by: there exists T_0 and k_0, such that for $T \geqslant T_0$, we have

$$N_{T-\varepsilon,Y^0} - k_0 \leqslant \int_{B_{T,Y}} f_0 \, \mathrm{d}\sigma \leqslant N_{T+\varepsilon,Y} . \tag{32}$$

Since the right inequality has the same nature, the same argument gives a similar *lower inequality* as inequality (30),

$$\liminf_{T\to\infty} \frac{N_{T,Y}}{e^T} \geqslant \frac{\pi}{\mathrm{Area}(S)} \nu_x(Y) . \tag{33}$$

The *upper inequality* is a little different from inequality (29). It says that for all proper subinterval Y^0 of Y,

$$\limsup_{T\to\infty} \frac{N_{T,Y^0}}{e^T} \leqslant \frac{\pi}{\mathrm{Area}(S)} \nu_x(Y) , \tag{34}$$

but since this is true for all Y containing the closure of Y^0, we have

$$\limsup_{T\to\infty} \frac{N_{T,Y^0}}{e^T} \leqslant \frac{\pi}{\mathrm{Area}(S)} \nu_x(Y_0) . \tag{35}$$

Apply that last upper inequality to $Y_0 = Y$, we have

$$\limsup_{T\to\infty} \frac{N_{T,Y}}{e^T} \leqslant \frac{\pi}{\mathrm{Area}(S)} \nu_x(Y) , \tag{36}$$

and thus combining with inequality (33) we obtain the desired asymptotics:

$$\lim_{T\to\infty} \frac{N_{T,Y}}{e^T} = \frac{\pi}{\mathrm{Area}(S)} \nu_x(Y) . \tag{37}$$

This completes the proof of the theorem. □

3. Equidistribution of geodesics and counting geodesics

Theorem 0.1 and 0.2 are interrelated and we will prove in this section

THEOREM 3.1. *Let S be a closed hyperbolic surface, let μ_0 be the right $\mathsf{SL}_2(\mathbb{R})$ invariant probability measure on US. Let $\Gamma(T)$ be the set of closed geodesics of length less then T and $N(T)$ the cardinal of $\Gamma(T)$, then*

$$N(T) \underset{T\to\infty}{\sim} \frac{e^T}{T}\,, \tag{38}$$

$$\frac{1}{N(T)} \sum_{\gamma\in\Gamma(T)} \mu_\gamma \underset{T\to\infty}{\longrightarrow} \mu_0\,. \tag{39}$$

We recall that given a closed geodesic γ, the measure μ_γ is characterized by

$$\mu_\gamma(U) = \frac{\ell(\gamma\cap U)}{\ell(\gamma)}\,.$$

3.1. Cubes. We will start by the following definition

DEFINITION 3.2. *An ε-*cube *or* cube of size ε *at x_0 in UH^2 is an open subset U of UH^2 of the form*

$$U := \Psi\Big(I_\varepsilon \times I_\varepsilon \times I_\varepsilon\Big)\, x_0\,.$$

where $I_\varepsilon :=]-\frac{1}{2}\varepsilon, \frac{1}{2}\varepsilon[$, x_0 is a point in US and

$$\Psi(s,u,t) = \varphi_t \circ h_s^+ \circ h_u^-\,.$$

We define cubes in US as projections of the similar object in UH^2.

We recall that $\{\varphi_t\}_{t\in\mathbb{R}}$ is the geodesic flow while $\{h_s^+\}_{s\in\mathbb{R}}$ and $\{h_u^-\}_{u\in\mathbb{R}}$ are respectively the stable and unstable horocyclic flows. The following is obvious

PROPOSITION 3.3. *There exists some ε_0 such that the projection from an ε_0-cube in UH^2 to US is an embedding.*

All similar cubes have the same volume:

PROPOSITION 3.4. *Let U and V be ε-cubes in UH^2, then $\mu_0(U) = \mu_0(V)$.*

PROOF. This is the consequence of the transitivity of the action of $\mathrm{Iso}_+(\mathbf{H}^2)$ on UH^2 on one hand and the fact that the action of $\mathrm{Iso}_+(\mathbf{H}^2)$ commutes with the action of $\mathsf{PSL}_2(\mathbb{R})$, and hence for every g in $\mathrm{Iso}_+(\mathbf{H}^2)$, the image of the ε-cube centered at x_0 is the ε-cube centered at $g(x_0)$. □

We will finally state the following result as a consequence of Vitali's covering argument and some extra work. Let us denote for any N, given K an ε-cube K^N the corresponding $N\varepsilon$-cube. The following lemma is obvious

PROPOSITION 3.5 (EXTENSION FROM CUBES). *Let ν be a flow invariant Radon measure. Assume there exist a positive number ε such that for every cube C of size less than ε, we have*

$$\nu(C) \leqslant \mu_0(C),$$

where μ_0 is the Lebesgue measure. Then ν is a multiple μ_0.

We start with two Lemmas whose proof – which involves squeezing a cube between two balls – is left to the reader. First given an ε-cube C centered at x_0, we denote by C^N the corresponding cube of size $N\varepsilon$ centered at x_0. Then we have

LEMMA 3.6. *There exists N_0 and ε_0 , such that if $\eta \leqslant \varepsilon$, and two η-cubes C_1 and C_2 intersects then C_2 is included in C_1^N*

For the sake of simplicity, we then write

$$C^0 := C^{N_0} .$$

LEMMA 3.7. *There exist positive constants K_0 and ε_1 such that if C is of size less than ε_1, then*

$$\mu(C^0) \leqslant K_0 \mu(C) .$$

PROOF OF PROPOSITION 3.5. It is enough to prove that $\nu_1 = f\mu_0$ where f is function: indeed by ergodicity of μ_0, f is a constant function.

Thanks to the Radon–Nikodym Theorem, it is enough to show that there exists a constant K_0 such that for every Borel set B:

$$\nu(B) \leqslant K_0 \mu_0(B) .$$

Let ε smaller than ε_1 and ε_2 given in Lemmas 3.6 and 3.7 For any set B, let B_ε be the neighborhood of B given by the reunion of ε-cubes centered on B.

Let finally $f(\varepsilon) = \mu(C)$ where C is any ε-cube.

By an adaptation of Vitali's covering argument given any Borel set B, we can find P disjoints $\frac{1}{N}\varepsilon$-cubes C_i centered on B such that C_i^0 covers B. Thus

$$f\left(\frac{1}{N_0}\varepsilon\right) P = \sum_{i=1}^{p} \mu(C_i) \leqslant \mu_0(B_\varepsilon) .$$

And in particular

$$P \leqslant \frac{1}{f\left(\frac{1}{N_0}\varepsilon\right)} \mu_0(B_\varepsilon) .$$

Then we can write

$$\begin{aligned}\nu_1(B) \leqslant \nu_1(B_\varepsilon) &\leqslant \sum_{i=1}^{P} \nu_1(C_i^0) \\ &\leqslant \sum_{i=1}^{P} \mu_0(C_i^0) = Pf(\varepsilon) \\ &\leqslant \frac{f(\varepsilon)}{f\left(\frac{1}{N_0}\varepsilon\right)} \mu_0(B_\varepsilon) = K_0 \mu_0(B_\varepsilon)\,.\end{aligned}$$

Since μ_0 is a Radon measure, hence outer regular

$$\mu_0(B) = \inf_{\varepsilon>0}(B_\varepsilon)\,.$$

Thus we have

$$\nu(B) \leqslant K_0 \mu(B)\,.$$

for all Borel sets. □

3.2. Hitting a cube. We now introduce the technical quantities needed for our proof. Let then for any positive numbers R_1 and R_2, with $R_2 > R_1$

$$\Gamma(R_1, R_2) := \{\gamma \text{ closed geodesics } \mid R_1 \leqslant \ell(\gamma) \leqslant R_2\}\,, \tag{40}$$

$$N(R_1, R_2) := \sharp\Gamma(R_1, R_2)\,, \tag{41}$$

$$\mu_{[R_1,R_2]} := \sum_{\gamma\in\Gamma(R_1,R_2)} \mu_\gamma\,. \tag{42}$$

Finally to shorten the notation, we denote by $\mathrm{I}(T,\beta)$ the interval centered at T of length β:

$$\mathrm{I}(T,\beta) = \left[T - \frac{\beta}{2}, T + \frac{\beta}{2}\right]\,,$$

Given U be an open set and γ a closed geodesic let

$$n(\gamma, U) = \sharp\,\{\text{connected components of } \gamma \cap U)\,,$$

Finally, for any positive β, let us define the following counting number.

$$N(T,\beta,U) := \sum_{\gamma\in\Gamma(\mathrm{I}(T,\beta))} n(\gamma, U) \tag{43}$$

3.2.1. *Reinterpreting the counting number.* We will need two reinterpretations of this counting number The first proposition is easy and will be used directly in the proof of the Theorem. We have

Proposition 3.8 (Measure reinterpretation). *For any positive ε and ε-cube U, we have*

$$\frac{\varepsilon}{T} N(T, \varepsilon, U) \underset{T\to\infty}{\sim} \mu_{\mathrm{I}(T,\varepsilon)}(U) \, . \tag{44}$$

Proof. The proof follows from the definition of the measure $\mu_{\mathrm{I}(T,\varepsilon)}$. Indeed

$$\mu_\gamma(U) = \frac{\ell(\gamma \cap U)}{\ell(\gamma)} = \frac{\varepsilon}{\ell(\gamma)} n(\gamma, U) \, .$$

It follows that

$$\frac{\varepsilon}{T + \frac{\varepsilon}{2}} n(\gamma, U) \leqslant \mu_\gamma(U) \leqslant \frac{\varepsilon}{T - \frac{\varepsilon}{2}} n(\gamma, U) \, .$$

Thus

$$\frac{\varepsilon}{T + \frac{\varepsilon}{2}} N(T, \varepsilon, U) \leqslant \mu_{\mathrm{I}(T,\varepsilon)}(U) \leqslant \frac{\varepsilon}{T - \frac{\varepsilon}{2}} N(T, \varepsilon, U) \, .$$

Hence

$$\frac{\varepsilon}{T} N(T, \varepsilon, U) \underset{T\to\infty}{\sim} \mu_{\mathrm{I}(T,\varepsilon)}(U) \, ,$$

which is what we wanted to prove. □

The second one will be used in the proof of the crucial Lemma

Lemma 3.9 (Intersecting cubes). *Let V be an ε-cube in $\mathsf{U}S$ and U an ε-cube in $\mathsf{U}\mathbf{H}^2$ such that $p(U) = V$ and*

$$\gamma(U) \cap U \neq \emptyset \text{ implies } \gamma = \mathrm{Id} \, . \tag{45}$$

Let

$$\Gamma(T, \beta, U) = \{\gamma \in \Gamma \mid \exists S \in \mathrm{I}(T, \beta) \, , x_0 \in U \, , \varphi_S(x_0) = \gamma(x_0)\} \, .$$

Then

$$N(T, \beta, V) = \sharp \, \Gamma(T, \beta, U) \, .$$

Proof. Let $C(T, \varepsilon, V)$ be the set of connected components of the intersection of V, with all closed geodesics of length in $\mathrm{I}(T, \varepsilon)$. Let v be an element of $C(T, \varepsilon, V)$, that is a connected component of $g \cap V$, where g is a closed geodesic of length S in the interval $T_T(\varepsilon)$. Let x_0 be the lift of an element in v, by definition there exists γ in Γ such that $\varphi_S(x_0) = \gamma(x_0)$. By construction, γ depends continuously on x_0 and thus actually depends only on v. We write $\gamma = \gamma(v)$. Observe that $\gamma(v)$ belongs to $\Gamma(T, \varepsilon, U)$ by definition.

Thus we have constructed a map $C(T, \varepsilon) \to \Gamma(T, \varepsilon, U)$ given by $v \mapsto \gamma(v)$. We now construct an inverse by the following procedure. Let γ be an element of

$\Gamma(T, \varepsilon, U)$, then by definition there is x_0 in U, S in $\mathrm{I}(T, \varepsilon)$, such that $\varphi_S(x_0) = \gamma(x_0)$. Then we observe that, by construction of a cube, the geodesic g starting at x_0 is so that $U \cap p$ is an interval. We take $v(\gamma)$ to be the projection of that interval. On then checks that $\gamma \to v(\gamma)$ is an inverse of $\gamma \to \gamma(v)$. □

3.3. Measure and counting. We now spend some time proving the following

Proposition 3.10 (Measure and counting). *For any positive ε and ε-cube U, we have*

$$N(T, \varepsilon, U) \underset{T\to\infty}{\sim} e^T \mu(U) \,. \tag{46}$$

We will devote section 3.8 to the proof of Proposition 3.10, then prove Theorem 3.1 in paragraph 3.9.

This is the heart of the construction. It involves both the mixing property and arguments from the Closing Lemma . W need to go through some preliminaries about horospherical arcs, then prove lower and upper bound for the integral of a test function.

3.4. Preliminaries on horospherical arcs. Say an *horospherical arc* Y is an interval in a unstable leave in $\mathsf{U}\mathbf{H}^2$. Such a horospherical arc projects to an horosphere and carries a measure $\mathrm{d}y$ coming from the arc length on the projected horosphere. Given Y an horospherical arc and T a positive number, we defined Y_T to be the horospherical arc $\varphi_T(Y)$.

Let finally p be the projection from $\mathsf{U}\mathbf{H}^2$ to $\mathsf{U}S$.: As a first lemma we have

Lemma 3.11 (Equidistribution of horospheres). *For any continuous function g on $\mathsf{U}S$ and any horospherical arc Y*

$$\frac{1}{\ell(Y)} \int_Y g \circ p \circ \varphi_T \, \mathrm{d}y \underset{T\to\infty}{\sim} \int_{\mathsf{U}S} g \, \mathrm{d}\mu_0 \,.$$

Proof. The proof is a consequence of mixing and is isomorphic to the proof of equidistribution of sectors (Theorem 1.1) . We will not repeat it here. Behind the similarity is a a general wave front Lemma (see [**?**]). □

Our second lemma uses the expansion property and will be used twice in the sequel

Lemma 3.12 (Images of intervals). *Let A and B be respectively η and ζ-cubes at x_0 with $\eta < \zeta$. Given any positive k_0 greater than 1, then there exists T_1 such that the following property holds*

- *Let H be the (unstable) horosphere through x_0.*
- *Let $Y = H \cap B$.*

- *Let γ be in Γ.*

We now assume there exist y_0 in A, T with $T > T_1$, such that $\varphi_T(y_0)$ belongs to $\gamma(A)$, then there exists an interval Z in A containing y_0, such that

(1) *$\varphi_T(Z)$ is a connected component of $\varphi_T(H) \cap \gamma(A)$*
(2) *we have the control:*

$$\frac{1}{k_0} e^{-T} \ell(Y) \leqslant \ell(Z) \leqslant k_0 e^{-T} \ell(Y) \,. \tag{47}$$

Proof. The proof is left to the reader. □

3.5. The test function. Let us now consider positive constants k and k_0 such that

$$0 < k < k_0 < \varepsilon \,.$$

- an ε-cube U_ε at x_0, where ε is small enough so that p restricted to U_ε is an embedding. We then consider the horospherical arc $Y = h^-_{I_\varepsilon} x_0$,
- a positive function f no greater than 1, equal to 1 on the $k_0\varepsilon$-cube $U_{k_0\varepsilon}$ at x_0 and with support in U_ε, where k is a fixed constant less than 1.

Observe that

$$\mu_0(U_{k\varepsilon}) \leqslant \int_{\mathsf{UH}^2} f \, \mathrm{d}\mu_0 \leqslant \mu_0(U_\varepsilon) \,.$$

Let

$$f_0 := \sum_{\gamma \in \Gamma} f \circ \gamma \;\; = g \circ p \,,$$

where g is a continuous function on US. It follows that

$$\int_{\mathsf{UH}^2} f \, \mathrm{d}\mu_0 = \int_{\mathsf{US}} g \, \mathrm{d}\mu_0 \,.$$

Furthermore, we have an immediate corollary of Lemma 3.12

Corollary 3.13. *for any K with $K > 1$, there exists T_1 so that for any γ and T greater than T_1,*

$$\frac{1}{\ell(Y)} \int_Y f \circ \gamma \circ \varphi_T \, \mathrm{d}y \;\leqslant\; K e^{-T} \tag{48}$$

3.6. The lower bound. The lemma of this paragraph is the following

Lemma 3.14. *Let k be a positive constant less than 1. Then there exists T_2 such that for T greater than T_2 Assume that γ belongs to $\Gamma(T, k\varepsilon, U_{k\varepsilon})$, then*

$$\frac{1}{\ell(Y)} \int_Y f \circ \gamma^{-1} \circ \varphi_T \geqslant k e^{-T} \,.$$

We then have as an immediate corollary using Lemma 3.9

Corollary 3.15 (Lower bound). *We have the lower bound: there exists T_3 such that for T greater than T_3,*

$$\frac{1}{\ell(Y)}\int_Y g\circ p\circ\varphi_T\,\mathrm{d}y \geqslant ke^{-T}N(T,k\varepsilon,U_{k\varepsilon})\,.$$

Proof of Lemma 3.14. Let γ be an element of $\Gamma(T,k\varepsilon,U_{k\varepsilon})$. Let $W \coloneqq h^+(I_\varepsilon)Y$, and $W_k = W\cap U_{k\varepsilon}$, $Y_k = Y\cap U_{k\varepsilon}$. Since W is transverse to the geodesic flow, it follows that there exists z_0 in W_k, α in $I_0(k\varepsilon)$ such that $\varphi_T(z_0)=\gamma(\varphi_\alpha z_0)$.

Let us write $z_0 = h^+_u(y_0)$ with y in Y_k. Then

$$\varphi_T(z_0) = h^+_{e^{-T}u}\varphi_T(y_0),$$

Since $\varphi_T(z_0)$ is in $\gamma(U_{k\varepsilon})$ it follows that for T large enough, $\varphi_T(y_0)$ is $\gamma(U_{k_0\varepsilon})$.

We can now apply Lemma 3.12, for $A = U_{k_0\varepsilon}$ and $B = U_\varepsilon$. Let Z be the subinterval of Y produced by the lemma. In particular for T large enough

$$e^{-T}\frac{1}{k}\ell(Y) \ \geqslant \ell(Z) \ \geqslant ke^{-T}\ell(Y)\,, \tag{49}$$

$$\varphi_T(Z)\subset U_{k_0\varepsilon} \tag{50}$$

It follows that

$$\int_Y f\circ\gamma^{-1}\circ\varphi_T\,\mathrm{d}y \geqslant \int_Z f\circ\gamma^{-1}\circ\varphi_T\,\mathrm{d}y = \ell(Z)\,.$$

Thus

$$\frac{1}{\ell(y)}\int_Y f\circ\gamma^{-1}\circ\varphi_T\,\mathrm{d}y \geqslant ke^{-T}\,.$$

The process is summarized in figure 1 □

3.7. The upper bound. The upper bound involve arguments similar to those of the Closing Lemma . Let K be now a constant greater than 1,

Lemma 3.16. *Let K be any constant greater than 1, then there exists T_4 such that if T greater than T_4 and γ in Γ satisfies*

$$\int_Y f\circ\gamma^{-1}\circ\varphi_T\,\mathrm{d}y \neq 0\,,$$

then γ belongs to $\Gamma(T,K\varepsilon,U_K\varepsilon)$

As an immediate corollary using Lemma 3.9, we have

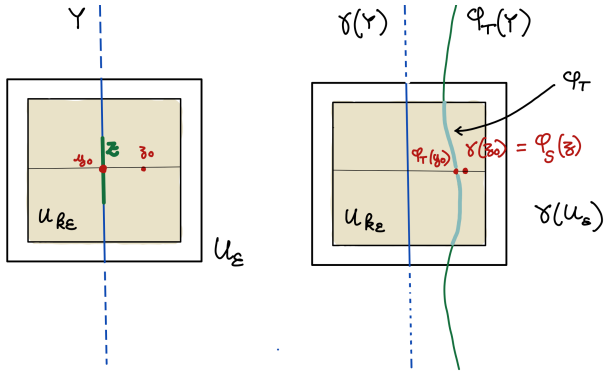


FIGURE 1. lower bound

COROLLARY 3.17 (UPPER BOUND). *We have the upper bound; there exists T_5 such that for T larger than T_5, we have*

$$\frac{1}{\ell(Y)} \int_Y f_0 \circ \varphi_T \, \mathrm{d}y \leqslant K e^{-T} N(T, K\varepsilon, U_{K\varepsilon}) \, .$$

PROOF OF LEMMA 3.16. Let again $W = h^+(I_\varepsilon)Y$. Let H the horosphere containing Y, let $Y_K = H \cap U_{K\varepsilon}$. Assume that γ is such that

$$\int_Y f \circ \gamma^{-1} \circ \varphi_T \, \mathrm{d}y \neq 0 \, .$$

This implies that there exist y_0 in Y, such that $f(\gamma^{-1}\varphi_T(y_0)) \neq 0$ and thus $\varphi_T(y_0)$ belongs to $\gamma(U_\varepsilon)$.

We apply Lemma 3.12, with $A = U_\varepsilon$ and $B = U_{K\varepsilon}$. Then there exist an interval Z, containing y_0, in Y_K such that $\varphi_T(Z)$ is a connected component of $\varphi_T(H)) \cap \gamma(U_{K\varepsilon})$.

Let now Π the projection from $U_{K\varepsilon}$ to Y_K, namely $\Pi(u)$ is the intersection of Y with the orbit of u under the group P^- generated by $\{\varphi_t\}_{t\in\mathbb{R}}$ and $\{h_u^-\}_{u\in\mathbb{R}}$.

Then the map $F : Z \to Y_K$

$$z \mapsto \Pi(\gamma^{-1}\varphi_T(z)) \, ,$$

has a fixed point z_1 in Z – and in particular in $U_{K\varepsilon}$ for T large enough: its inverse is contracting. It follows that

$$\varphi_T(L) = \gamma(L) \, ,$$

where $L = P^{-1}z_1$. Since $G = \gamma^{-1} \circ \varphi_T$ is contracting on L we therefore obtain a fixed point x_0 in L, arbitrarily close to z_1 for T large enough and in particular in $U_{K\varepsilon}$.

We can now conclude: we have S in $I_T(K\varepsilon)$, x_0 in $U_{K\varepsilon}$, such that $\varphi_S(x_0) = \gamma(x_0)$. This completes the proof. The process is summarized in figure 2

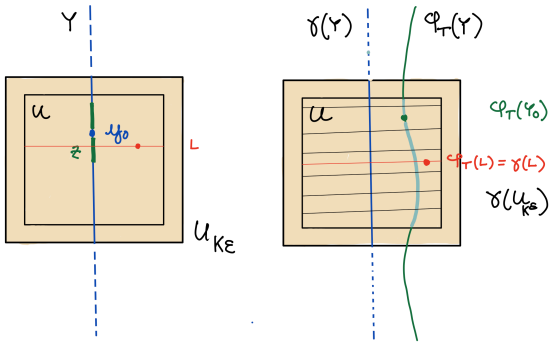


FIGURE 2. upper bound

□

3.8. Proof of Proposition 3.10. We can now put the pieces back together.

PROOF OF PROPOSITION 3.10. It follows from the previous results that for constant K greater than 1 and k smaller than 1, for all ε-cube U

$$\liminf_{T\to\infty}\left(Ke^{-T}N(T, K\varepsilon, U_K)\right) \geqslant \mu_0(U_k)\,. \tag{51}$$

Since this is true for all k we have

$$\liminf_{T\to\infty}\left(Ke^{-T}N(T, K\varepsilon, U_K)\right) \geqslant \mu_0(U_0)\,. \tag{52}$$

We now remark there is a function F_ε, not depending on the choice of such that

$$\mu_0(U) = F_\varepsilon(K)\mu_0(U_K)$$

Moreover F_ε converges to 1, uniformly in ε when K converges to 1. Thus noticing that U_K is an η-cube with $\eta = K\varepsilon$. We obtain that for all η cube V,

$$\liminf_{T\to\infty}\left(e^{-T}N(T, \eta, V)\right) \geqslant \frac{1}{K}F_{\frac{\eta}{K}}\mu_0(V)\,. \tag{53}$$

Hence letting K converge to 1. We obtain

$$\liminf_{T\to\infty}\left(e^{-T}N(T, \eta, V)\right) \geqslant \mu_0(V)\,. \tag{54}$$

A similar argument using the lower bound gives

$$\limsup_{T\to\infty}\left(e^{-T}N(T,\eta,V)\right)\leqslant\mu_0(V)\,. \tag{55}$$

These two inequalities give the result. □

3.9. Proof of Theorem 3.1. Our goal is to prove the following proposition and then Theorem 3.1

Proposition 3.18. *we have*

$$\frac{T}{e^T}\mu_{[0,T]}(U)\underset{T\to\infty}{\longrightarrow}\mu_0(U)\,. \tag{56}$$

Proof. Our starting observation is that combining Proposition 3.10 and Proposition 3.8 yields the following assertion: for every ε-cube U,

$$\mu_{\mathrm{I}(T,\varepsilon)}(U)\underset{T\to\infty}{\sim}\varepsilon\frac{e^T}{T}\mu_0(U)\,. \tag{57}$$

Step 1: *we first show that for any positive ε, any ε-cube U, and k_1 greater than 1, there exists T_6 such that for $T\geqslant T_6$,*

$$\frac{e^{-2\varepsilon}}{k_1^2}\,\mu_0(U)\leqslant\frac{T}{e^T}\mu_{[0,T]}(U)\leqslant k_1^2e^{2\varepsilon}\,\mu_0(U)\,\mathrm{d}s\,. \tag{58}$$

As a consequence of assertion (57), we have that for every k_1, with $k_1>1$, there exists S_1 so that for $T>S_1-1$, the following inequalities holds

$$\frac{\mu_0(U)}{k_1}\int_{T-\varepsilon}^{T}\frac{e^s}{s}\,\mathrm{d}s\leqslant\mu_{\mathrm{I}(T,\varepsilon)}(U)\leqslant k_1\mu_0(U)\int_{T}^{T+\varepsilon}\frac{e^s}{s}\,\mathrm{d}s\,, \tag{59}$$

where we used the inequalities

$$\int_{T-\varepsilon}^{T}\frac{e^s}{s}\leqslant\varepsilon\frac{e^T}{T}\leqslant\int_{T}^{T+\varepsilon}\frac{e^s}{s}\,\mathrm{d}s\,.$$

Thus for all positive integer q,

$$\frac{\mu_0(U)}{k_1}\int_{S_1-\varepsilon}^{S_1+q\varepsilon}\frac{e^s}{s}\,\mathrm{d}s\leqslant\mu_{[S_1,S_1+q\varepsilon]}(U)\leqslant k_1\mu_0(U)\int_{S_1}^{S_1+(q+1)\varepsilon}\frac{e^s}{s}\,\mathrm{d}s\,.$$

Let then p the integer such that

$$S_1+p\varepsilon\leqslant T\leqslant S_1+(p+1)\varepsilon\,. \tag{60}$$

The inequalities

$$\mu_{[S_1,S_1+(p-1)\varepsilon]}(U)\leqslant\mu_{[S_1,T]}(U)\leqslant\mu_{[S_1,S_1+(p+1)\varepsilon]}(U)\,,$$

thus yield

$$\frac{\mu_0(U)}{k_1}\int_{S_1-\varepsilon}^{S_1+(p-1)\varepsilon}\frac{e^s}{s}\,\mathrm{d}s \leqslant \mu_{[S_1,T]}(U) \leqslant k_1\mu_0(U)\int_{S_1}^{S_1+(p+2)\varepsilon}\frac{e^s}{s}\,\mathrm{d}s\,,$$

and in particular using inequalities (60) again

$$\frac{\mu_0(U)}{k_1}\int_{S_1-\varepsilon}^{T-2\varepsilon}\frac{e^s}{s}\,\mathrm{d}s \leqslant \mu_{[S_1,T]}(U) \leqslant k_1\mu_0(U)\int_{S_1}^{T+2\varepsilon}\frac{e^s}{s}\,\mathrm{d}s\,. \tag{61}$$

As a consequence, we remark that

$$\mu_{[S_1,T]}(U) < \infty\,,$$
$$\lim_{T\to\infty}\mu_{[S_1,T]}(U) = \infty\,,$$

and thus

$$\mu_{[S_1,T]}(U) \underset{T\to\infty}{\sim} \mu_{[0,T]}(U)\,.$$

Let us now consider the left and right terms of inequality (61). L'hôpital's rule gives

$$\int_{S_1-\varepsilon}^{T-2\varepsilon}\frac{e^s}{s}\,\mathrm{d}s \underset{T\to\infty}{\sim} e^{-2\varepsilon}\frac{e^T}{T-2\varepsilon} \underset{T\to\infty}{\sim} e^{-2\varepsilon}\frac{e^T}{T}\,,$$
$$\int_{S_1}^{T+\varepsilon}\frac{e^s}{s}\,\mathrm{d}s \underset{T\to\infty}{\sim} e^{2\varepsilon}\frac{e^T}{T+2\varepsilon} \underset{T\to\infty}{\sim} e^{2\varepsilon}\frac{e^T}{T}\,.$$

Thus inequality (61) rereads as: for any positive ε, any ε-cube U and k_1 greater than 1, there exists T_6 such that for $T \geqslant T_6$,

$$\frac{e^{-2\varepsilon}}{k_1^2}\frac{e^T}{T}\,\mu_0(U) \leqslant \mu_{[0,T]}(U) \leqslant k_1^2 e^{2\varepsilon}\frac{e^T}{T}\,\mu_0(U)\,\mathrm{d}s\,.$$

This concludes the proof of step 1.

Step 2: *Let us consider a sequence $\{T_m\}_{m\in\mathbb{N}}$ growing to infinity such that*

$$\mu_m := \frac{T_m}{e^{T_m}}\mu_{[0,T_m]} \underset{m\to\infty}{\longrightarrow} \nu\,,$$

for some geodesic flow invariant Radon measure ν. Then $\nu = \mu_0$.

Let now f be a positive function less than 1 and supported on some ε-cube U. It follows that

$$\int f\mathrm{d}\nu = \lim_{m\to\infty}\int f\mathrm{d}\mu_m \leqslant \mu_m(U) \leqslant e^{2\varepsilon}\mu_0(U)k_1^2\,.$$

Since the previous inequality is true for any k_1 greater than 1, we obtain that

$$\int f \mathrm{d}\nu \leqslant e^{2\varepsilon} \mu_0(U) \,,$$

Thus for any compact K inside U,

$$\nu(K) \leqslant e^{2\varepsilon} \mu_0(U) \,,$$

and since ν is a Radon measure, hence inner regular, for any ε-cube U,

$$\nu(U) \leqslant e^{2\varepsilon} \mu_0(U) \,. \tag{62}$$

Observe this inequality is also true for any cube of size less than ε. Thus applying Proposition 3.5, we have that $\nu(B)$ is a multiple of the Lebesgue measure μ_0 by a constant k:

$$\nu = k\mu_0 \,.$$

The inequality (62) furthermore gives that $k \leqslant e^{2\varepsilon}$, and since this is true for all ε, it follows that

$$k \leqslant 1 \,. \tag{63}$$

Let us now prove the opposite inequality: let f be a function equal to 1 on U,

$$\int f \mathrm{d}\nu \geqslant e^{-2\varepsilon} \mu_0(U) \,.$$

We then obtain, by taking functions furthermore supported on neighborhoods of $\bar{U}$,

$$\nu(\bar{U}) = k_0 \mu(\bar{U}) \geqslant e^{-2\varepsilon} \mu_0(U) \,.$$

It then implies that $k \geqslant e^{-2\varepsilon}$, and this is is true for all ε we have

$$k \geqslant 1 \,. \tag{64}$$

Combining inequalities (63) and (64), we have that

$$\nu = \mu_0 \,.$$

This concludes the proof of step 2.

Conclusion: We remark that the inequality (58) of step 1 implies that $\frac{T}{e^T}\mu_{[0,T]}$ has bounded total measure independent on T: Let us take a covering of $\mathsf{U}S$ by finitely many ε cubes $\{U_i\}_{i\in\{1,\dots,p\}}$ then

$$\frac{T}{e^T}\mu_{[0,T]}(\mathsf{U}S) \leqslant \sum_{i=1}^{p} \left(\frac{T}{e^T}\mu_{[0,T]}(U_i) \right) \leqslant k_1^2 e^{2\varepsilon} \mu_0(K_i) \leqslant p\, k_1^2\, e^{2\varepsilon} < \infty \,.$$

The result then follows from the weak compactness of measures of bounded total measure. □

Proof of Theorem 3.1. Applying Proposition 3.18 to $f = 1$, we get

$$N(T) = \int_{US} \mathrm{d}\mu_{[0,T]} \underset{T\to\infty}{\sim} \frac{e^T}{T} \int_{US} \mathrm{d}\mu_0 = \frac{e^T}{T} ,$$

which is the first part of Theorem 3.1. Secondly, we then have that by the definition of weak convergence

$$\frac{T}{e^T} \sum_{\gamma\in\Gamma(T)} \mu_\gamma \underset{T\to\infty}{\longrightarrow} \mu_0 ,$$

and thus combining with the previous assertion this prove the second half of Theorem 3.1:

$$\frac{1}{N(T)} \sum_{\gamma\in\Gamma(T)} \mu_\gamma \underset{T\to\infty}{\longrightarrow} \mu_0 .$$

This completes the proof. □

4. Comments, references and further reading

Part 3

Tourism around hyperbolic surfaces

CHAPTER 7

Discrete subgroups and closed surfaces

1. Monodromies of hyperbolic structures and the Euler class

Every (oriented) hyperbolic surface gives rise to an embedding of the monodromy group of $\pi_1(S)$ with discrete image, moreover this group has no torsion. Indeed, every torsion element of $\mathsf{PSL}_2(\mathbb{R})$ fixes a point in the hyperbolic plane.

Conversely every torsion free subgroup of $\mathsf{PSL}_2(\mathbb{R})$ acts properly freely on the hyperbolic plane and is the monodromy of a – not necessarily compact – hyperbolic surface. In order to complete the picture, we recall

LEMMA 1.1 (SELBERG). *Every finitely generated linear group possesses a finite index subgroup without torsion.*

We sketch the idea of the proof of Selberg Lemma in Lemma 1.2.

It follows that every faithful representation of the fundamental group of a surface with discrete image is the monodromy of a hyperbolic structure. A little extra work shows that if moreover S is compact then $\rho(\pi_1(S)\backslash H^2$ is homeomorphic to S, so therefore

PROPOSITION 1.2. *Every faithful representation of the fundamental group of a compact surface S with discrete image is the monodromy a hyperbolic structure on S.*

EXERCISE 1.1: (**) Show that this last statement fails for a non-compact surface.

PROPOSITION 1.3 (BOREL DENSITY THEOREM). *Every monodromy of a finite volume hyperbolic surface is Zariski dense.*

PROOF. We prove it for a compact surface $S = \Gamma\backslash\mathbf{H}^2$. From the description of hyperbolic surfaces, it follows that we can find two elements in Γ which generates hyperbolic translations with distinct endpoints. However, the list of algebraic non trivial subgroups of $\mathsf{PSL}_2(\mathbb{R})$ is very short: they either preserve a point in $\mathbf{H}^2$ or a point in $\partial_\infty\mathbf{H}^2$. It follows that the Zariski closure of Γ is $\mathsf{PSL}_2(\mathbb{R})$. □

1.1. Teichmüller space and space of representations. The *Teichmüller space* $\tau(S)$ is the space of all representations of $\pi_1(S)$ with discrete image up to

conjugacy. It is diffeomorphic to a ball, we almost proved it ... . Let us introduce an invariant of representation of $\pi_1(S)$. For that, let us choose a presentation of $\pi_1(S)$

$$\pi_1(S) = \langle a_1, b_1, \dots, a_g, b_g \mid \prod_{i=1} g[a_i, b_i] = 1 \rangle ,$$

where $[c, d] = cdc^{-1}d^{-1}$.

Observe now that $\mathsf{PSL}_2(\mathbb{R})/\operatorname{SO}(2) = \mathbf{H}^2$, hence that $\mathsf{PSL}_2(\mathbb{R})$ has the homotopy type of S^1. We therefore have an exact sequence

$$\mathbb{Z} \to \widetilde{\mathsf{PSL}_2(\mathbb{R})} \to \mathsf{PSL}_2(\mathbb{R}) \to 0.$$

Let us choose a map σ from $\mathsf{PSL}_2(\mathbb{R})$ to $\widetilde{\mathsf{PSL}_2(\mathbb{R})}$ that splits that sequence, σ will actually never be continuous, nor a group morphism. Then we have

Proposition 1.4 (Euler Class). *Let ρ be a representation of $\pi_1(S)$ to $\mathsf{PSL}_2(\mathbb{R})$ The element*

$$e(\rho) = \prod_{i=1} g[\sigma(\rho(a_i)), \sigma(\rho(b_i)],$$

is an element of the center of $\widetilde{\mathsf{PSL}_2(\mathbb{R})}$ which we identify to $\mathbb{Z}$. This number is independent of the choice of σ, of the presentation of $\pi_1(S)$ and is constant under local deformations of ρ. The number $e(\rho)$ is called the Euler class *of the representation.*

Let now $\chi(S)$ be the Euler characteristics of S. Then

Theorem 1.5 (Milnor-Wood inequality). *Let ρ be a representation of $\pi_1(S)$ to $\mathsf{PSL}_2(\mathbb{R})$ Then*

$$|e(\rho)| \leqslant |\chi(S)|.$$

We can use the Euler class to distinguish connected connected components of space of representations. More precisely

Theorem 1.6 (Goldman). *The map from the space of connected components of $\operatorname{Hom}(\pi_1(S), \mathsf{PSL}_2(\mathbb{R}))$ to $\{\chi(S), \chi(S) + 1, \dots, -\chi(S)\}$ is a bijection. Moreover, monodromies of hyperbolic structures are exactly representations such that $|e(\rho)| = |\chi(S)|$.*

It follows from this theorem that we can check whether a representation is the monodromy of a hyperbolic structure just from a presentation of the group.

2. Comments, references and further reading

CHAPTER 8

Arithmetic surfaces

1. Field extensions

A field $\mathbb{K}_1$ containing a field $\mathbb{K}_0$ is said to be a *field extension* of $\mathbb{K}_0$. The *degree* $[\mathbb{K}_1 : \mathbb{K}_0]$ of the field extension is the dimension of $\mathbb{K}_1$ seen as a vector space over $\mathbb{K}_0$. When $[\mathbb{K}_1 : \mathbb{K}_0]$ is finite, we say $\mathbb{K}_1$ is a *finite extension* of $\mathbb{K}_0$.

Exercise 1.1:

(1) $\mathbb{C}$ is a extension of degree 2 of $\mathbb{R}$, while $\mathbb{R}$ is an infinite extension of $\mathbb{Q}$.
(2) Let $\alpha_0, \alpha_1, \dots \alpha_p$ be complex numbers. Let $\mathbb{K} = \mathbb{Q}(\alpha_1, \dots \alpha_p)$. Say α_0 is *algebraically dependent* of $\alpha_1, \dots \alpha_p$ if α_0 is a solution of a non trivial polynomial equation with coefficients in $\mathbb{K}$ and *algebraically independent* other wise. Prove that if the numbers α_0 is algebraically depend from $\alpha_1, \dots \alpha_p$ then $\mathbb{Q}(\alpha_0, \dots \alpha_p)$ is a finite degree extension of $\mathbb{K}$.
(3) (*) Prove that if $\alpha_0, \alpha_1, \dots \alpha_p$ are algebraically independent complex numbers, then the ring $\mathbb{Q}[\alpha_0, \dots \alpha_p]$ generated by $\mathbb{Q}$ and the numbers $(\alpha_0, \alpha_1, \dots \alpha_p)$ is isomorphic to the polynomial ring $\mathbb{Q}(X_0, \dots, X_p)$ over the variables $(X_0, \dots, X_p)$

When $d = [\mathbb{K}_1 : \mathbb{K}_0]$ is finite, it follows that $\mathbb{K}_1 \setminus \{0\}$ acting by multiplication on $\mathbb{K}_1$ can be considered as a subgroup of $\mathsf{GL}_d(\mathbb{K}_0)$.

More generally if $d = [\mathbb{K}_1 : \mathbb{K}_0]$, then we have a group embedding, essentially by replacing coefficients with matrices,

$$\mathsf{GL}_n(\mathbb{K}_1) \to \mathsf{GL}_{dn}(\mathbb{K}_0) \, . \tag{65}$$

Here is a useful proposition

Proposition 1.1. *Let Γ be a finitely generated subgroup of $\mathsf{GL}_n(\mathbb{R})$ then Γ is isomorphic to a subgroup of $\mathsf{SL}_{nM}\left([\mathbb{Z}[1/p, X_1, \dots, X_q]\right)$.*

We give a sketch of the proof in the following exercise.

Exercise 1.2: Let Γ be a finitely generated subgroup of $\mathsf{GL}_n\mathbb{R}$)

(1) prove that there exists some numbers $\alpha_1, \dots, \alpha_q$ so that

$$\Gamma < \mathsf{GL}_n(\mathbb{Z}[\alpha_1, \dots, \alpha_q]) \, .$$

Hint: consider the coefficients of the generating set of Γ

(2) (*) Using the trick developed in the beginning of the section, show that Γ embeds in (for some p) in

$$\mathsf{SL}_{nM}\left(\mathbb{Z}[1/p, X_1, \ldots, X_q]\right) .$$

Hint: Observe that this is obvious when $\alpha_1, \ldots, \alpha_q$ are algebraically independent, then use a finite induction to reduce to this case.

As a consequence of the proposition we have the following classical lemmas

LEMMA 1.2 (SELBERG LEMMA). *Every finitely generated subgroup of* $\mathsf{GL}_n(\mathbb{R})$ *has a finite index subgroup which has no non trivial elements of finite order.*

The next lemma requires a definition: a group Γ of is *residually finite* if for every non trivial element γ in Γ, there is a homomorphism f of Γ in a finite group such that $f(\gamma)$ is non trivial.

LEMMA 1.3 (RESIDUAL FINITENESS). *Every finitely generated subgroup* Γ *of* $\mathsf{GL}_n(\mathbb{R})$ *is residually finite*

It is enough indeed to show both Lemmas for $\mathsf{SL}_N\left(\mathbb{Z}[1/p, X_1, \ldots, X_q]\right)$.

2. Lattices and arithmetic lattices

We start this section by a definition: we say two subgroups Γ_1 and Γ_2 in a group G are *commensurable* if $\Gamma_1 \cap \Gamma_2$ is a subgroup of finite index in both Γ_1 and Γ_2.

2.1. Lattices. Let us now concern ourselves with *lattices* in $\mathsf{SL}_n(\mathbb{R})$, that is discrete subgroups Γ such that $\Gamma\backslash\mathsf{SL}_n(\mathbb{R})$ has a finite volume. When $n = 2$, among those are groups Γ such that $\Gamma\backslash\mathsf{SL}_2(\mathbb{R})$ is a compact hyperbolic surface.

EXERCISE 2.1:

(1) Show that every subgroup commensurable to a lattice is a lattice.
(2) (**) Show that a lattice Γ in $\mathsf{PSL}_2(\mathbb{R})$ that contains no parabolic elements is *cocompact*: $\Gamma\backslash\mathsf{PSL}_2(\mathbb{R})$ is compact.
(3) Show that a lattice Γ in $\mathsf{PSL}_2(\mathbb{R})$ that contains no parabolic elements has a finite index subgroup Γ_0 such that $\Gamma\backslash\mathbf{H}^2$ is a hyperbolic surface.

Lattices are large and we have a generalization of Proposition 1.3 by Armand Borel.

THEOREM 2.1 (BOREL DENSITY THEOREM). *Every lattice in* $\mathsf{SL}_n(\mathbb{R})$ *is Zariski dense in* $\mathsf{SL}_n(\mathbb{R})$

The theorem holds for any semi-simple Lie group.

2.2. Arithmetic lattice.

DEFINITION 2.2. *A subgroup* Γ *in* $\mathsf{G} = \mathsf{SL}_n(\mathbb{R})$*, is* arithmetic *if we have representation* ρ *of* G *in* $\mathsf{SL}_N(\mathbb{R})$ *such that*

(1) *the group* $\rho(\mathsf{G}) \cap \mathsf{SL}_N(\mathbb{Q})$ *is (almost) dense in* $\rho(\mathsf{G})$*,*
(2) *the groups* $\rho(\Gamma)$ *and* $\rho(G) \cap \mathsf{SL}_N(\mathbb{Z})$ *are (almost) equal: they have a common finite index subgroup.*

The expression H *almost dense in* G means that the closure of H contains the connected component of the identity in G. Obviously any subgroup commensurable to an arithmetic lattice is arithmetic. As before the definition can be extended to any non compact semisimple Lie group G.

The condition (1) is equivalent to (or stated as) ρ *is defined over* $\mathbb{Q}$.

THEOREM 2.3 (MINKOWSKI). *The subgroup* $\mathsf{SL}_n(\mathbb{Z})$ *is a lattice in* $\mathsf{SL}_n(\mathbb{R})$*.*

EXERCISE 2.2: (*) Prove Minkowski Theorem for $n = 2$.

A beautiful generalization is due to Armand Borel and Harish-Chandra

THEOREM 2.4. *Any arithmetic subgroup of* $\mathsf{SL}_n(\mathbb{R})$*, or of any semi-simple Lie group, is a lattice.*

We then have

EXERCISE 2.3: (**) The set of arithmetic lattices is countable. (*) There exists non arithmetic surface subgroups.

As an obvious example $\mathsf{SL}_2(\mathbb{Z})$ is an arithmetic lattice. Here is a less obvious construction.

EXERCISE 2.4:

(1) Show that the map $A \mapsto \det(A)$ defines a non degenerate quadratic form q_0 of signature $(1, 2)$ on the space M of $2x2$ matrices with trace zero.
(2) Prove that the map from $\mathsf{PSL}_2(\mathbb{R})$ to $\mathsf{SO}(q_0)$ given by the conjugation action of $\mathsf{PSL}_2(\mathbb{R})$ on M, is an isomorphism.

In other words, $\mathsf{PSL}_2(\mathbb{R})$ is isomorphic to $\mathsf{SO}(1, 2)$

Let now q be the quadratic form on $\mathbb{R}^3$ defined by

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

and let G be the group $\mathsf{SO}(q)$, which is isomorphic to $\mathsf{PSL}_2(\mathbb{R})$ by the previous exercise. The following construction is difficult.

EXERCISE 2.5:

(1) Prove that $\mathsf{SL}_3(\mathbb{Q}) \cap \mathsf{G}$ is dense in G.
(2) Prove that $\Gamma := \mathsf{SL}_3(\mathbb{Z}) \cap \mathsf{G}$ is an arithmetic lattice in G.
(3) (**) Show that Γ contains no parabolic elements.
(4) Show that $\Gamma\backslash\mathbf{H}^2$ is compact.

Let us end this paragraph by a beautiful theorem of Margulis

THEOREM 2.5 (MARGULIS ARITHMETICITY THEOREM). *Any lattice in* $\mathsf{SL}_n(\mathbb{R})$ *with* $n \geqslant 3$ *is arithmetic.*

Again this result is extended to any semi-simple Lie group.

3. Commensurators, arithmeticity and correspondences

3.1. Commensurator group. Let Γ be a lattice in $\mathsf{SL}(\mathbb{R})$ $S = \mathbf{H}^2/\Gamma$. The *commensurator group* of Γ is

$$\mathrm{Comm}(\Gamma) := \{g \in \mathsf{SL}(\mathbb{R}) \mid g\Gamma g^{-1} \cap \Gamma \text{ is of finite index in } \Gamma\}.$$

One can observe that the commensurator group is indeed a group. We check for instance that

$$\mathsf{PSL}_2(\mathbb{Q}) = \mathrm{Comm}(\mathsf{PSL}_2(\mathbb{Z})) .$$

More generally, by the definition of arithmetic groups the commensurator of an arithmetic lattice is dense. The following a deep Theorem by Margulis states the converse.

THEOREM 3.1 (MARGULIS COMMENSURABILITY CRITERION). *A lattice is arithmetic if and only if its commensurator group is dense.*

We will now use this theorem as a non standard definition of arithmetic lattices.

3.2. Arithmetic hyperbolic surfaces. We now concentrate on surfaces. Here is another remark. We say a hyperbolic surface is *arithmetic* if $S = \Gamma\backslash\mathbf{H}^2$, with Γ arithmetic.

LEMMA 3.2. *If a surface is not arithmetic, then* Γ *has finite index in* $\mathrm{Comm}(\Gamma)$.

PROOF. Let H the closure of $\mathrm{Comm}(\Gamma)$. We first prove that H is discrete: otherwise, every element in $\mathrm{Comm}(\Gamma)$ would fix the Lie algebra of the connected component of H of the origin. But this defines a Zariski closed condition. By Borel density theorem, $\mathsf{H} = \mathsf{PSL}_2(\mathbb{R})$ hence S is arithmetic which is a contradiction. It follows that $\mathrm{Comm}(\Gamma)$ is discrete. Then we have a covering map from $\mathsf{PSL}_2(\mathbb{R})/\Gamma$, which is compact, to $\mathsf{PSL}_2(\mathbb{R})/\mathrm{Comm}(\Gamma)$. Hence, the fibers of this map are finite sets which exactly means that Γ has finite index in $\mathrm{Comm}(\Gamma)$. □

Here is an important and far from obvious consequence of the classification of arithmetic surfaces.

PROPOSITION 3.3. *Given a compact surface S, there are only finitely arithmetic hyperbolic surfaces homeomorphic to S.*

We then have

COROLLARY 3.4. *Give a surface S, there is a number $\lambda(S)$, such that for every arithmetic hyperbolic metric g on S, for every closed geodesic γ, then*

$$\ell_g(\gamma) \geqslant \lambda(S)$$

Here $\ell_g(\gamma)$ is the length of γ with respect to g. It is not very difficult tor prove the corollary implies the proposition. Here is an interesting exercise. I actually do not know the solution and would love to hear of one.

EXERCISE 3.1: (***) Give a proof of Corollary 3.4 only using the characterization of arithmetic hyperbolic surfaces using Margulis commensurability criterion.

3.3. Hecke correspondences and arithmetic dynamics. The main feature of arithmetic surfaces are the existence of many correspondences. A *finite correspondence* between two sets X and Y is a subset of Z in the product $X \times Y$ so that the preimage of very point in each of the factor is finite and non-empty. In particular, an element g in the commensurator group of a hyperbolic surface S gives rise to such a correspondence which is furthermore a local isometry see Figure 1. Indeed, we consider the map π_g of $\mathbf{H}^2$ into $S \times S$ given by $x \to (\pi(x), \pi(gx))$ where π is the covering map. To say g is in the commensurator group, is just to say that π_g is a covering map of compact image and that its image is a correspondence Z_g. Actually, the correspondence is just determined by the class of g in $\Gamma \backslash \mathrm{Comm}(\Gamma)/\Gamma$.

Let then furthermore define the *degree* $\deg(T)$ *of a correspondence* T from S to itself, as the infimum of the degree of the covering of S involved in defining the correspondence.

We now define *hyperbolic correspondences* for hyperbolic surfaces to be correspondences which are local isometries. The dichotomy between arithmetic surfaces and non-arithmetic surfaces is then the dichotomy between finitely many and infinitely many hyperbolic self-correspondences.

A self-correspondence gives rise to two types of dynamics. First *quantum dynamics* acting on the space of L^2 functions on S. So if p_1 and p_2 are the two projections – of degree q – of the hyperbolic correspondence $Z \subset S \times S$ on each factor, then

$$H_g(f)(x) := \frac{1}{q} \sum_{z \in p_1(x)} f(p_2(z)),$$

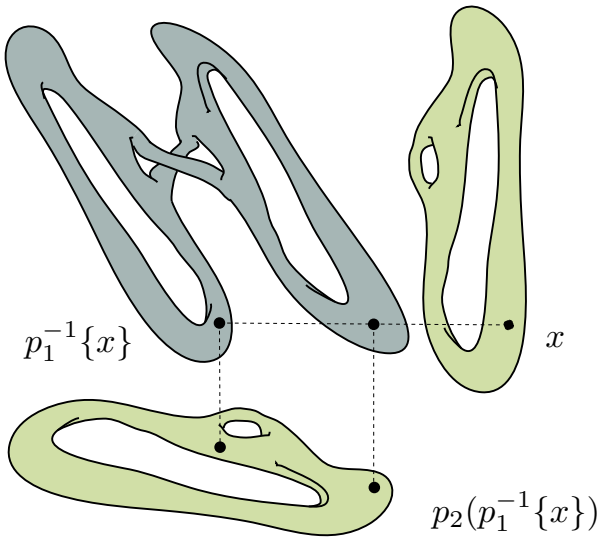


FIGURE 1. A correspondence

is a self adjoint operator called the *Hecke operator* of the correspondence.

3.4. The hyperbolic solenoid. Secondly, we can associate *classical dynamics*. There is a classical dynamical way to turn non-bijective map or more generally correspondence into a bijective map. So, to settle notation, let Z be a self-correspondence of a set S and p_1 and p_2 be the two projections, we say $x\mathcal{R}y$ if $p_1^{-1}(x)$ intersects $p_2^{-1}(y)$. Then we consider the set

$$\mathcal{L}_Z = \{f : \mathbb{Z} \to S \mid f(n)\, \mathcal{R}\, f(n+1)\} \subset S^{\mathbb{Z}}.$$

The *shift* σ is the map from $\mathcal{L}_Z$ to itself given by

$$\sigma(f)(n) = f(n+1)\ .$$

The shift is now a homeomorphism and its dynamics reflect that of the correspondence.

This construction is not sufficient for our purpose. We indeed would like to see all correspondences as acting in the same space. Returning to the case of S being a compact hyperbolic surface, let us first take a look at the space $\mathcal{L}_Z$ in special case. Let $P_{[n,p]}$ be the map from $\mathcal{L}_Z$ to S^{n-p} given by

$$f \to (f(n), \ldots, f(p)).$$

By construction the image of $P_{[n,p]}$ is a compact surface, which we call $S_{[n,p]}$ in the product S^{n-p} and moreover any projection map to a factor – that is $P_{[q]}$ for q

in $[n, p]$ – is a covering map. We can therefore describe $\mathcal{L}_Z$ as a "limit" of some coverings.

We generalize this construction. Let S be a compact hyperbolic surface. The *hyperbolic solenoid* $\mathcal{S}(S)$ is the "limit" of all coverings of S. Let us give a definition as a set. The *hyperbolic solenoid* is the set of sequences $\{(x_n, S_n, p_n)\}_{n\in\mathbb{N}^*}$ so that x_n is a point in S_n, p_n is a covering from S_n to S_{n-1} – where by convention $S_0 = S$ – such that $p_n(x_n) = x_{n-1}$ up to the following *solenoid equivalence*: two sequences $\{(x_n^0, S_n^0, p_n^0)\}_{n\in\mathbb{N}}$ and $\{(x_n^1, S_n^1, p_n^1)\}_{n\in\mathbb{N}}$ are equivalent if there exists a third one $\{(y_n, \Sigma_n, q_n)\}_{n\in\mathbb{N}}$ together with covering maps q_n^i from Σ_n to S_n^i satisfying the commuting diagram conditions

$$q_n^i(y_n) = x_n^i \,,$$
$$p_n^i \circ q_{n+1}^i = q_n^i \circ p_{n+1}^i \,.$$

The equivalence is described in Figure 2.

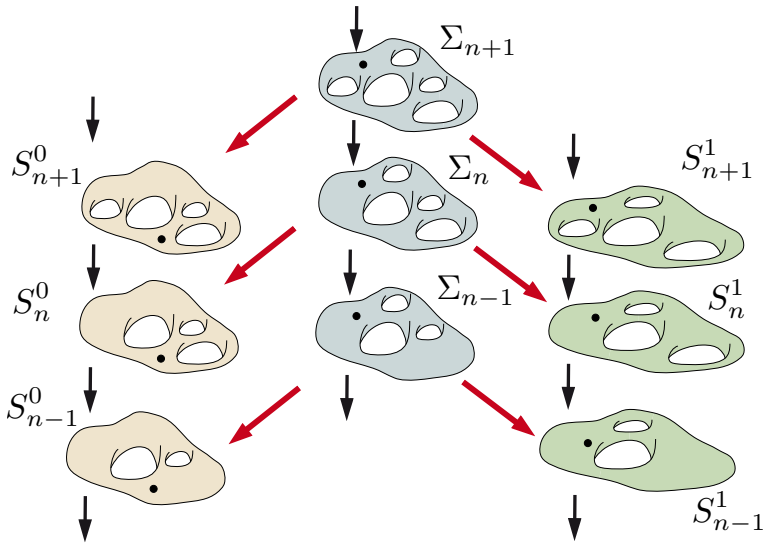


FIGURE 2. Solenoid equivalence

We now give a precise definition that defines the hyperbolic solenoid as a topological space: Let W_S be the 2-dimensional complex, whose vertices are surfaces which are finite covers of S, oriented edges correspond to covering between the extremities, and faces correspond to commuting diagrams of coverings. Let Z_S be the universal cover of this complex, and V_S be the set of vertices of this graph, which we consider as surfaces. If e is an edge of V_S from e^- to e^+, then it gives rise to a covering p_e from e^- to e^+ seen as surfaces. Then

Definition 3.5 (The hyperbolic solenoid). *The* hyperbolic solenoid $\mathcal{S}(S)$ *of a hyperbolic surface* S *is*

$$\mathcal{S}(S) \coloneqq \{(x_\Sigma)_{\Sigma\in V_S} \mid x_\Sigma \in \Sigma,\ \ p_e(x_{e^+}) = x_{e^-}\} \subset \prod_{\Sigma\in V_S} \Sigma.$$

In particular, we have a projection

$$\pi_\Sigma \ : \ \mathcal{S}(S) \to \Sigma$$

for every finite cover Σ of S. Since $\mathcal{S}(S)$ is a subset of the compact space $\prod_{\Sigma\in V_S}\Sigma$ it inherits a topology. Moreover $\mathcal{S}(S)$ is compact thank to the following exercise

Exercise 3.2: (*) The set $\mathcal{S}(S)$ is a closed subset of $\prod_{\Sigma\in V_S}\Sigma$.

Alternatively, we can describe the hyperbolic solenoid as a fiber bundle over S whose structure group is the *profinite completion* of the fundamental group of S. The hyperbolic solenoid is the "universal cover for finite covers", in the sense that it solves a universal problem for finite covers of S.

Let us me more precise, for an infinite group Γ, let us consider the space $\mathcal{N}$ of normal subgroups of Γ of finite index. Denote then for N in $\mathcal{N}$, $\Gamma_\mathsf{N} \coloneqq \Gamma/\mathsf{N}$ and π_N the projection of Γ in Γ_N. Let then consider the compact group

$$\Gamma_0 = \prod_{N\in\mathcal{N}} \Gamma_N$$

Observe that we have a natural morphism φ from Γ to Γ_0 given by

$$\varphi(\gamma) \coloneqq \prod_{\mathsf{N}\in\mathcal{N}} \pi_\mathsf{N}(\gamma)\ .$$

One sees that this morphism is injective if and only if Γ is residually finite. We then define the profinite completion of Γ as the compact group

$$\mathrm{Pro}(\Gamma) = \overline{\Phi(\Gamma)}\ .$$

We immediately observe that the compact group $\mathrm{Pro}(\Gamma)$ is totally disconnected – since Γ_0 is and compact. In particular $\mathrm{Pro}(\Gamma)$ is uncountable and not finitely generated.

Proposition 3.6 (Universality of the profinite completion). *Let g be a morphism of Γ in a finite group F, then there is a morphism π_g of* $\mathrm{Pro}(\Gamma)$ *in F such that $g = \pi_g \circ \varphi$.*

The hyperbolic solenoid and the profinite completion are related by the following result.

Proposition 3.7 (Principal bundle). *The projection π_S from $\mathcal{S}(S)$ is a principal bundle whose structure group is* $\mathrm{Pro}(\Gamma)$.

EXERCISE 3.3: (*)

(1) Give a proof of the previous proposition: show that if we have a hyperbolic ball B that embeds in S, then $\pi^{-1}(B)$ is homeomorphic to

$$B \times \mathrm{Pro}(\Gamma) .$$

(2) Prove that $\mathcal{S}(S)$ is homeomorphic to $\Gamma \backslash (\mathbf{H}^2 \times \mathrm{Pro}(\Gamma))$ where the action of Γ on $\mathrm{Pro}(\Gamma)$ is given by the right action by $\Phi(\Gamma)$ which is a subgroup of $\mathrm{Pro}(\Gamma)$.

3.5. Isometries of the hyperbolic solenoid. From the description of the hyperbolic solenoid as a principal bundle with a totally disconnected fiber, it follows that the *leaf through a point x* – that is the path-wise connected component of x – is a hyperbolic plane – see exercise below..

Then, again from the description of the hyperbolic solenoid as a fiber bundle, $\mathcal{S}(S)$ is a *hyperbolic laminated space* as in Figure 3: every point has a neighborhood – called chart – which is homeomorphic to a product of a ball in $\mathbf{H}^2$ with a topological space – in our case totally disconnected – such that moreover the coordinates changes when we change charts are isometries on the hyperbolic factors.

A *leaf-wise isometry* of a hyperbolic laminated space is then a homeomorphism preserving the laminated structure which is a local isometry on the hyperbolic factors.

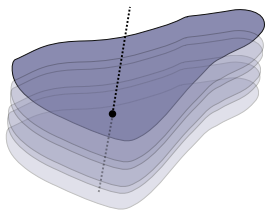


FIGURE 3. A small open set in a laminated space

EXERCISE 3.4: (*) Prove that each hyperbolic leaf of $\mathcal{S}(S)$ is simply connected. *Hint*: use exercise 3.3.

Now, the fundamental though obvious remark is

PROPOSITION 3.8. *if $p : \Sigma_1 \to \Sigma_0$ is a covering, then we have a unique isometry*

$$\hat{p} \ : \ \mathcal{S}(\Sigma_1) \to \mathcal{S}(\Sigma_0)$$

such that $\pi_{S_1} \circ p = \pi_{S_0} \circ \hat{p}$. *Moreover* $\hat{p}$ *is a bijection whose inverse is an isometry.*

Proof. Use the fact that one can induce coverings. □

As a consequence, a hyperbolic self- correspondence on a surface $S = \Gamma\backslash\mathbf{H}^2$, given by an element g of $\mathrm{Comm}(\Gamma)$ and its two covering maps (p_1, p_2) acts as a leaf-wise isometry on the hyperbolic solenoid by

$$\Phi_g = \hat{p}_1 \circ (\hat{p}_2)^{-1} .$$

Therefore, the hyperbolic solenoid of an arithmetic surface has very rich dynamics.

Exercise 3.5:

(1) Prove that the map $g \mapsto \Phi_g$ is a group morphism and relate the group of isometry of $\mathcal{S}(S)$ to $\mathrm{Comm}(\Gamma)$. Describe the action of Γ on $\mathcal{S}(S)$.
(2) (**) (For those who know). Relate the hyperbolic solenoid for a finite index torsion free subgroup of $\mathsf{PSL}_2(\mathbb{Z})$ to the adélic ring.

4. Equidistribution of Hecke points

Let as usual S be a hyperbolic surface associated to an arithmetic hyperbolic group Γ in $\mathsf{PSL}_2(\mathbb{R})$. We saw that every correspondence, which we can associate to an element g of $\mathrm{Comm}(\Gamma)$, defines a Hecke operator H_g from $C^0(S)$ to itself. We can see it also as acting on measure. If the correspondence is of degree q, that is given by two local isometries p_1 and p_2 from a q-cover S_1 of S, then given a probability measure μ on S^1, the corresponding measure is $T_g(\mu)$ given by

$$T_g(\mu) := \frac{1}{q} \ (p_1)_* \circ p_2^*(\mu) .$$

In particular, if we start with a point x in S, denoting by δ_x the probability measure supported at x, then $T_g(\delta_x)$ is the probability measure equidistributed on the set $T_g(x) := p_1(p_2^{-1}\{x\})$.

We then have

Theorem 4.1 (Equidistribution of Hecke points). *Given a sequence* $\{g_i\}_{i\mathbb{N}}$ *in* $\mathrm{Comm}(\Gamma)$ *with* $\{\deg(g_i)\}_{i\in\mathbb{N}}$ *growing to infinity then* $\{T_{g_i}(x)\}_{i\in\mathbb{N}}$ *becomes equidistributed. More precisely, for any* x *in* $\Gamma\backslash\mathbf{H}^2$

$$T_{g_i}(\delta_x) \underset{i\to\infty}{\longrightarrow} \mu_0 ,$$

where μ_0 *is the probability Lebesgue measure on* S, *and* δ_x *the probability measure supported at* x .

Observe that the condition $\{\deg(g_i)\}_{i\in\mathbb{N}}$ growing to infinity, implies that the lattice Γ is arithmetic: this condition can only be satisfied when Γ has infinite index in $\mathrm{Comm}(\Gamma)$. That is the reason why the hypothesis that Γ is arithmetic is not spelled out, even though the Theorem only applies in that case.

This is again part of a long stream of results: it was first noticed and stated without proof by Marc Burger and Peter Sarnak, as a consequence of Marina Ratner Theorem, in the special case the the sequence $\{g_i\}_{i\in\mathbb{N}}$ converges to an element not in $\mathrm{Comm}(\Gamma)$, the proof was then given in Dani–Margulis. Then Alex Eskin and Hee Oh proved it in the generality given here using ergodic theoretic methods, while Clozel–Ullmo and Clozel–Oh–Ullmo have obtained crucial information on the rate of convergence using unitary representation theoretic methods. The result is in general about lattices although we stick here to the case of $\mathsf{PSL}_2(\mathbb{R})$.

We give an idea of the proof of Theorem 4.1 following the initial suggestion by Burger–Sarnak in our special case. The proof relies on the following (immediate) corollary of Ratner's Theorem that we admit for our purpose– see the discussion for more details. First let us denote by $\Delta(\mathsf{PSL}_2(\mathbb{R}))$ the diagonal group in $\mathsf{PSL}_2(\mathbb{R})\times\mathsf{PSL}_2(\mathbb{R})$

Proposition 4.2 (Baby Ratner). *Let Γ be a subgroup of $\mathrm{Iso}_+(\mathbf{H}^2)$ such that $S := \Gamma\backslash\mathbf{H}^2$ is compact. Let ν be an ergodic probability measure on $\mathsf{U}S\times\mathsf{U}S$ invariant under the subgroup $\Delta(\mathsf{PSL}_2(\mathbb{R}))$ in $\mathsf{PSL}_2(\mathbb{R})\times\mathsf{PSL}_2(\mathbb{R})$. Then ν is*

(1) *either $\mu_0\otimes\mu_0$,*
(2) *or supported on a closed orbit of $\Delta(\mathsf{PSL}_2(\mathbb{R}))$.*

In particular the space of ergodic measures for $\Delta(\mathsf{PSL}_2(\mathbb{R}))$ is countable. This proposition is true in more generality than for arithmetic groups, and as an extension to all semi-simple groups.

We shall use a second property which, again, has a more general version due to Shahar Mozes and Nimish Shah:

Proposition 4.3 (Baby Mozes–Shah). *Let Γ be a subgroup of $\mathrm{Iso}_+(\mathbf{H}^2)$ such that $S := \Gamma\backslash\mathbf{H}^2$ is compact. Then the $\Delta(\mathsf{PSL}_2(\mathbb{R}))$-invariant probability measures on $\mathsf{U}S\times\mathsf{U}S$ supported on closed orbits of $\Delta(\mathsf{PSL}_2(\mathbb{R}))$ are* isolated *in the space of ergodic probability measures.*

By *isolated*, we mean that given a measure ν supported on a closed orbit of $\Delta(\mathsf{PSL}_2(\mathbb{R}))$, then for any positive ε, there is a non negative continuous function f such that if μ is another measure supported on a closed orbit of $\Delta(\mathsf{PSL}_2(\mathbb{R}))$,

then

$$\int_{\mathsf{U}S} f \, \mathrm{d}\mu \leqslant \varepsilon \int_{\mathsf{U}S} f \, \mathrm{d}\nu \quad \text{and} \quad 0 < \int_{\mathsf{U}S} f \, \mathrm{d}\nu \, .$$

Proof of Theorem 4.1. Our goal here is to explain how to translate of the Hecke transform of a measure into a $\Delta(\mathsf{PSL}_2(\mathbb{R}))$ invariant measure on $\mathsf{U}S \times \mathsf{U}S$, and thus understanding equidistribution as related to a classification of measures.

Let x be a fixed point in $\mathsf{U}S$ and Γ_x the stabilizer of x in $\mathsf{PSL}_2(\mathbb{R})$ acting on the left on $\mathsf{U}S$. For any topological space X with an action by homeomorphism of a group F, let us denote by $\mathcal{M}(X)^{\mathsf{F}}$ the set of probability Radon measures invariant by F.

The starting point is that we have a homeomorphism Φ from $\mathcal{M}(\mathsf{U}S)^{\Gamma_x}$ to $\mathcal{M}(\mathsf{U}S \times \mathsf{U}S)^{\Delta(\mathsf{PSL}_2(\mathbb{R}))}$. Intuitively we see $\mathsf{U}S$ as the fiber over x of the product $\mathsf{U}S \times \mathsf{U}S$ and we start diffusing a measure ν using the group $\Delta(\mathsf{PSL}_2(\mathbb{R}))$. Formally we have the following construction: if ν is a measure on $\mathsf{U}S$ invariant by Γ_x, we have a well defined map from

$$\lambda_\nu : \mathsf{U}S \to \mathcal{M}(\mathsf{U}S) \, , \; y \mapsto \nu_y$$

given by $\nu_{xg} = g_*^{-1}\nu$. Conversely any map λ from $\mathsf{U}S$ to $\mathcal{M}(\mathsf{U}S)$ such that $\lambda(yg) = g_*^{-1}\lambda(y)$ is obtained as $\lambda = \lambda_\nu$, where $\nu = \lambda(x)$ is in $(\mathsf{U}S)^{\Gamma_x}$.

Then we define $\nu_0 = \Phi(\nu)$ on $\mathsf{U}S \times \mathsf{U}S$ by

$$\int_{\mathsf{U}S \times \mathsf{U}S} f(x,y) \, d\nu_0 = \int_{\mathsf{U}S} \left(\int_{\mathsf{U}S} f(x,y) \, \mathrm{d}\nu_y \right) \mathrm{d}\mu_0(x) \, .$$

We leave the reader check that the corresponding measure is invariant under $\Delta(\mathsf{PSL}_2(\mathbb{R}))$.

This map has an inverse. Let ν_1 be a measure on $\mathsf{U}S \times \mathsf{U}S$ invariant by $\Delta(\mathsf{PSL}_2(\mathbb{R}))$. Let p_1 be the projection on the first factor then $(p_1)_*(\nu)$ is invariant by $\mathsf{PSL}_2(\mathbb{R})$, and thus equal to μ_0. We then decompose the measure ν_1 as

$$\nu_1 = \int_{\mathsf{U}S} \nu_y(y) \, \mathrm{d}\mu_0(y) \, ,$$

where $\nu_y = \lambda(y) \otimes \delta_y$ is a measure on $\mathsf{U}S \times \{y\}$. Since ν_1 is $\Delta(\mathsf{PSL}_2(\mathbb{R}))$ invariant, it follows that $\lambda(yg) = g_*^{-1}\lambda(y)$ and thus by the previous observation, we have $\lambda = \lambda_\nu$. It follows that $\Phi(\nu) = \nu_1$.

Let us now go back to correspondences and Hecke transforms. Given an element a of the commensurator group, a hyperbolic correspondence is equivalent to giving a closed orbit S_a of $\Delta(\mathsf{PSL}_2(\mathbb{R}))$ in $\mathsf{U}S \times \mathsf{U}S$.

Then, since S_a is a quotient of $\Delta(\mathsf{PSL}_2(\mathbb{R}))$, it carries a unique $\Delta(\mathsf{PSL}_2(\mathbb{R}))$ invariant measure μ_a and we leave the reader check that for a given x, we have

$$\Phi(T_a(\delta_x)) = \mu_a \, .$$

Let us now take a weak limit ν of $\nu_i := \{T_{a_i}(\delta_x)\}_{i\in\mathbb{N}}$. The measure ν is obviously invariant by $\Delta(\mathsf{PSL}_2(\mathbb{R}))$. By Baby Ratner Proposition 4.2 and the Decomposition of Ergodic Measures Theorem 3.1, we can find non-zero positive numbers λ_0 and λ_1 ,..., λ_p, with $\lambda_1 \geqslant \lambda_2$ such that

$$\sum_{i=0}^{\infty} \lambda_i = 1$$

and such that

$$\nu_{i \underset{i\to\infty}{} \nu} = \lambda_0(\mu_0 \otimes \mu_0) + \sum_{j=1}^{\infty} \lambda_j \mu_j \, ,$$

where μ_j is supported on a closed orbit of $\Delta(\mathsf{PSL}_2(\mathbb{R}))$ associated to elements g_j in $\mathrm{Comm}(\Gamma)$. We want to show that $\lambda_1 = 0$. It first follows form the previous expression that

$$\nu = (1-\lambda_1)\sigma_1 + \lambda_1\mu_1 \, ,$$

where σ_1 is a probability measure. Since $\{\deg(a_i)\}_{i\in\mathbb{N}}$ grows to infinity, there exists some N_0 so that for $i \geqslant N_0$,

$$a_i \neq g_1 \, .$$

By Proposition 4.3, there is a positive continuous function f, such that for all i greater than N_0

$$\int_{US} f \, \mathrm{d}\nu_i \leqslant \frac{\lambda_1}{2} \int_{US} f \, \mathrm{d}\mu_1 \quad \text{and} \quad 0 < \int_{US} f \, \mathrm{d}\mu_1 \, ,$$

and thus at the limit

$$(1-\lambda_1)\int_{US} f \, \mathrm{d}\sigma_1 + \lambda_1 \int_{US} f \, \mathrm{d}\mu_1 \leqslant \frac{\lambda_1}{2} \int_{US} f \, \mathrm{d}\mu_1 \, .$$

And we have the contradiction using the fact that f is non negative:

$$0 \leqslant (1-\lambda_1)\int_{US} f \, \mathrm{d}\sigma_1 \leqslant -\frac{\lambda_1}{2}\int_{US} f \, \mathrm{d}\mu_1 < 0 \, .$$

It follows that $\Phi(\nu) = \mu_0 \otimes \mu_0$. This implies that $\nu = \mu_0$. This is what we wanted to prove. □

EXERCISE 4.1: (**) Describe the equidistribution of Hecke points using the association $g \mapsto \Phi_g$, which associates to an element g in $\mathrm{Comm}(g)$ the isometry Φ_g of the hyperbolic solenoid $\mathcal{S}(S)$ defined in Paragraph 3.5.

5. Correspondences and the Ehrenpreis conjecture

A correspondence is given by a common cover S of S_1 and S_2 with covering maps p_1 and p_2. One may wonder what happens if one can relax the condition of p_1 and p_2 being local isometries, in other words that the induced metrics by p_1 and p_2 are K-bi-Lipschitz instead of being equal.

The answer, was conjectured as the Ehrenpreis conjecture, and is now a beautiful theorem of Jeremy Kahn and Vladimir Markovic.

THEOREM 5.1 (KAHN–MARKOVIC). *For any ε, for any pair of compact hyperbolic surfaces, there exists a common covering such that the two induced distances are $(1+\varepsilon)$-bi-Lipschitz equivalent.*

The proof involves Margulis counting results explained in these notes as well as ergodicity and mixing.

6. Comments, references and further reading

CHAPTER 9

Harmonic functions

1. Harmonic functions

We finally move to the last topic of these notes. We encourage strongly the reader to have a look at N. Bergeron beautiful set of notes on the Laplacian on hyperbolic surfaces.

The Laplacian on the hyperbolic plane is the map, defined for any smooth function f

$$f \mapsto \Delta(f) = -y^2\left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}\right).$$

Another way to describe the Laplacian is the following: if J is the complex structure of the hyperbolic plane and $\omega_{\mathbf{H}^2}$ its area form, then

$$-\mathrm{d}(\mathrm{d}f \circ J) = \Delta(f)\,\omega_{\mathbf{H}^2}\,. \tag{66}$$

From this last description, of from a verification using the generators of $\mathsf{PSL}_2(\mathbb{R})$ given in Exercise 2.6, one sees that if g is an oriented isometry of $\mathbf{H}^2$, then

$$\Delta(f \circ g) = (\Delta f) \circ g\,.$$

1.1. Laplacian on hyperbolic surfaces. It then follows that we can define the Laplacian on any hyperbolic surface. We could also use directly Equation 66. Moreover

Proposition 1.1. *Given a compact hyperbolic surface S with hyperbolic area μ_0 and two smooth functions f and g on S, then*

$$\int_S g\,\Delta f\,\mathrm{d}\mu_0 = \int_S f\,\Delta g\,\mathrm{d}\mu_0\,,$$
$$\int_S f\,\Delta f\,\mathrm{d}\mu_0 \geqslant 0\,,$$

where the last inequality is an equality if and only if f is constant.

Proof. All these inequalities follow from Equation (66) and the Stokes Formula that yield

$$\int_S g\,\Delta f\,\mathrm{d}\mu_0 = \int_S \mathrm{d}g \wedge \mathrm{d}f \circ J\,.$$

□

We say a function f on S is an *eigenfunction* of the Laplacian of eigenvalue λ if $\Delta(f) = \lambda f$. The multiplicity of the eigenvalue λ is the dimension of the space of eigenfunctions. Eigenfunctions could also have been defined just using integration on small discs and balls.

Since S is assumed to be compact a general theorem (the Spectral Theorem) asserts the following, where we recall that we denote by $\mathrm{L}^2_0(S)$ the closed space of functions of zero integral.

Theorem 1.2 (Spectral theorem for the Laplacian). *There exists an orthonormal Hilbert basis $\{\varphi\}_{i\in\mathbb{N}}$ of $\mathrm{L}^2_0(S)$ of eigenfunctions of the Laplacian with corresponding eigenvalues $\{\lambda\}_{i\in\mathbb{N}}$ satisfying*

$$0 < \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n \leqslant \cdots\,.$$

Moreover the collection of eigenvalues less than a constant is finite.

This allows us to define the *spectrum of* Δ as the infinite collection of eigenvalues $\{\lambda\}_{i\in\mathbb{N}}$.

1.2. The trace formula. The length of closed geodesics and the eigenvalues of the Laplacian are related by many deep results. *Selberg trace formula* is certainly the most striking. Atle Selberg trace formula is a generalization of Poisson summation formula, it reads

Theorem 1.3 (Selberg-trace version). *Let S be a compact hyperbolic surface and h be an even test function satisfying some restriction. Let $\{\lambda_n\}$ be the set of eigenfunctions of the Laplacian. Let $\mu_n^2 + 1/4 = \lambda_n$, with either the imaginary part or the real part of μ_n is positive. Let $\mathcal{G}$ be the set of closed geodesics and $\ell(\gamma)$ be the length of the closed geodesic γ and $m(\gamma)$ be the multiplicity of γ. Then*

$$\sum_{n=0}^{+\infty} h(\mu_n) = -\frac{\chi(S)}{2}\int_{-\infty}^{+\infty} h(s)s\tanh(\pi s)\,\mathrm{d}s + \sum_{\gamma\in\mathcal{G}} R_\gamma \hat{h}(\ell(\gamma)),$$

where

$$R_\gamma = \frac{\ell(\gamma)}{m(\gamma)2\sinh(\ell(\gamma)/2)},\quad \hat{h}(s) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} h(u)e^{-ius}\,\mathrm{d}u.$$

We now state it using a slightly non-standard approach due to Pierre Cartier and André Voros.

Define a *primitive geodesic* as a closed geodesic that does not cover non-trivially another one.We need to introduce two generalized zeta functions.

(1) The *generalized Hurwitz ζ function* reflects the analytic side. It is a defined as

$$\zeta(s,a) = \mathrm{Tr}(\Delta_S + a)^{-s} := \sum_{i=0}^{\infty} \frac{1}{(\lambda_n + a)^s},$$

where (λ_n) is the set of eigenvalues repeated with multiplicities.

(2) The *Selberg zeta function* reflects the dynamical side. Let $\mathcal{P}$ be the set of *primitive* closed geodesics. Let us define

$$Z_S(s) = \prod_{\gamma\in\mathcal{P}} \prod_{k=0}^{\infty} \left(1 - e^{\ell(\gamma)(k+s)}\right).$$

Then we have after taking the analytic continuation of these functions.

Theorem 1.4 (Selberg-determinant version). *Let S be a compact hyperbolic surface. Then for any non negative number u,*

$$\frac{\partial}{\partial s}\Big|_{s=0} \zeta_S\left(s, u^2 - \frac{1}{4}\right) = \psi(u)^{\chi(S)} Z_S\left(\frac{1}{2} + u\right),$$

where $\psi(u)$ is an explicit function only depending on u, which can be interpreted as related to a spectral problem on the two-sphere. The left hand side term is usually interpreted as as the logarithm of a regularized determinant.

Finally we state an much sought after conjecture of Atle Selberg from 1965 with important consequences in number theory

Conjecture 1.5 (Selberg's 1/4 conjecture). *The first eigenvalue of the Laplacian on $\Gamma_0(N)\backslash\mathbf{H}^2$ is greater than* 1/4.

Here

$$\Gamma_0(N) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ in } \mathsf{PSL}_2(\mathbb{Z}) \text{ with } c \text{ divisible by } N \right\}.$$

Selberg obtained the lower bound 3/16, the best lower bound at present is the lower bound $975/4096 \simeq 0.238$ due to Kim–Sarnak.

2. Quantum chaos

We can now state another famous conjectures for hyperbolic surfaces. The *quantum unique ergodicity conjecture* is the quantum pendent of the equidistribution of the orbits of the geodesic flows.

This conjecture due to Peter Sarnak and Steve Rudnick claims the following

Conjecture 2.1 (Quantum unique ergodicity conjecture). *Let $\{\varphi_n\}$ be a sequence of eigenfunctions of the Laplacian on a compact hyperbolic surface S, such that the corresponding eigenvalues go to infinity. Let μ be the probability Lebesgue measure on S, then*

$$\frac{|\varphi_n|^2}{\int_S |\varphi_n|^2 \, \mathrm{d}\mu}\mu \underset{n\to\infty}{\longrightarrow} \mu \, .$$

This conjecture is known to be "almost surely true" as proved by Alexander Schnirelman, Steve Zelditch and Yves Colin de Verdière as the *Quantum Ergodicity Theorem* in the sense that it converges for a – and then plenty – subsequence of density 1 in the sequence of eigenfunctions. This conjecture has natural extensions to higher dimensions, non negative curvature other elliptic problems than on functions. The following breakthrough has been obtained in the arithmetic context.

Theorem 2.2 (Elon Lindenstrauss). *Let S be an arithmetic surface. Then the quantum unique ergodicity conjecture holds if we furthermore assume the sequence of eigenfunctions for the Laplacian are Hecke eigenfunctions.*

The approach is ergodic and uses the extra dynamics coming from Hecke correspondences on a space related to the solenoid.

3. Comments, references and further reading

APPENDIX A

Coverings and curves

We recall the following facts from elementary algebraic topology. Let X be a topological space and Z_0 and Z_1 two closed subset in X. Recall that two continuous curves c_0 and c_1 from $[0,1]$ to X are

(1) *homotopic with extremities x and y, with respect to Z_0 and Z_1*, if there exists a continuous mapping C from $[0,1]\times[0,1]$ to X so that for all s, $C(s,0) = c_0(s)$, $C(s,1) = c_1(s)$, and for all t, $C(0,t) \in Z_0$, $C(1,t) \in Z_1$.
(2) *homotopic with extremities x and y* when we consider $Z = \{x, y\}$.
(3) when c_0 and c_1 are both closed, we say they *homotopic with base point x*, if they are homotopic with extremities x and x.
(4) we say that two closed curves c_0 and c_1 are *freely homotopic* if there exists a continuous mapping C from $[0,1]\times[0,1]$ to X so that for all s, $C(0,s) = c_0(s)$ and $C(1,s) = c_1(s)$.

Recall also that $p : X \to Y$ a continuous surjective map between topological spaces is a covering if every y in Y has an open neighborhood, such that

$$p^{-1}(U) = \bigsqcup_{x\in p^{-1}(y)} U_y \ ,$$

with p an homeomorphisms from U_x to U.

The following exercises are classical propositions from basic homotopy theory. They are not very difficult to prove on their own. Let

Exercise 0.1: Let $p : X \to Y$ be a covering

(1) [Lifting property] Prove that of given any curve $c : [0,1] \to Y$, given any x in X, with $p(x) = c(0)$, there exists a unique curve $\gamma[0,1] \to X$, so that $\gamma(0) = x$ and $c = p \circ \gamma$. We call γ a *lift* of c
(2) Prove moreover that if c_0 and c_1 curves in S with $c_0(0) = c_1(0)$ as well $c_0(1) = c_1(1)$ then c_0 are homotopic if and only if $\gamma_0(1) = \gamma c_0(1)$, where γ_0 and γ_1 are lifts of c_0 and c_1 respectively with $\gamma_0(0) = \gamma_1(1)$.

The more important result is the following. Let X be a "nice" topological space, for instance a metric space, such that every small enough ball is homeomorphic to a ball in a euclidean space. Then

THEOREM 0.1 (EXISTENCE OF THE UNIVERSAL COVER). *There exists a covering $p : Y \to X$, with Y simply connected. Moreover if Γ is defined by*

$$\Gamma = \{\gamma \in \mathrm{Homeo}(Y) \mid p \circ \gamma = p\} ,$$

Then X is homeomorphic to Y/Γ. More precisely there is an homeomorphism φ from Y/Γ to X such that $\varphi(\Gamma \cdot x) = p(x)$.

The space Y is called the *universal cover* of X.