university of groningen

faculty of science
and engineering

mathematics and applied
mathematics

Geodesic Flow of the Modular Surface and Continued Fractions

Bachelor's Project Mathematics

July 2019

Student: J.S. de Pooter

First supervisor: Dr. M. Seri

Second assessor: Dr. K. Efstathiou

Contents

1 Introduction 4

1.1 Motivation 4

1.2 Geometry 4

1.3 Measure preserving maps and ergodicity 6

1.4 Ergodicity of geodesic flow 8

1.5 Structure 10

2 Poincaré half-plane 11

3 Isometries and congruences 16

3.1 Tranformations of the Poincaré half-plane 16

4 Geodesics 21

4.1 Straightedge and compass construction of geodesics 23

5 Geodesic flow on the hyperbolic half-plane 26

5.1 The horocycle 28

6 The modular surface 31

6.1 Distance on the unit tangent bundle 31

6.2 Area and volume 33

6.3 The quotient 34

6.4 The fundamental domain 35

6.5 Tesselation of Poincaré half-plane 38

7 Ergodicity of geodesic flow 40

7.1 Measure, geodesic flow and horocycle flow on the modular surface 42

8 The Gauss map 49

8.1 The Poincaré section of the modular surface 49

9 Epilogue 55

1 Introduction

1.1 Motivation

The study of geodesic flows on hyperbolic surfaces has a long and rich history. One of the reasons for its rise to the spotlight is that it provides a deep and fascinating link across multiple areas of mathematics, such as dynamical systems, spectral theory, geometry and number theory, and of physics, such as thermodynamics and quantum mechanics.

In this thesis we will focus on the relation between the geodesic flow on the modular surface, a particular example of hyperbolic surfaces, and the Gauss map. We will exploit this relation to transfer results from different mathematical realms and establish some of the interdisciplinary connections mentioned above.

The main investigative method will be a literature study: we will use a selection of articles, textbooks, monographs and lecture material to depict a coherent explanation of some of the theories involved and their relations.

First of all, we will describe the Poincaré half-plane model for hyperbolic geometry and study how Fuchsian groups act on it as Möbius transformations. This will allow us to define arithmetic surfaces, which are finite-volume manifolds obtained as quotients of the half-plane under the group action, and allows us study the properties of their geodesic flows. From this point on, we will focus on a specific surface, the modular surface. This will be the prototypical example to show how properties of its geodesic flow can be studied by studying the dynamical properties of a discrete map: the Gauss map G. To this end, we will define the concept of Poincaré first return map for the geodesic flow on this quotient manifold, show that the Gauss map appears as first return map on a specific section, and prove that the geodesic flow is ergodic. To provide the link, and one of the paths to prove the ergodicity, we will need to study the relation between the Gauss map and continued fractions expansion (for quadratic irrationals).

In this introduction, we will introduce prerequisite mathematical topics and related literature, and we introduce the main connection between the geodesic flow and continued fractions. After this, we outline the content of the thesis.

1.2 Geometry

An introduction to differential geometry, such as the book of Do Carmo [7], discusses the first examples of non-Euclidean geometry. That is to say, the focus is on regular surfaces [7, Definition 2.1] embedded in three dimensional Euclidean space. These surfaces are diffeomorphic to the Euclidean plane, but can be curved. Beyond these two dimensional surfaces, there are smooth manifolds: these are more arbitrary spaces that locally resemble some n dimensional Euclidean space. Formally, they are second countable, Hausdorff spaces that are locally Euclidean of dimension n [16, Definition 5.2]. Roughly speaking, around each point of the manifold, there is an open set in which there is a coordinate system of n dimensions. One can define functions and vector fields on such a manifold, and naturally ask the question what it means to be continuous or to take the derivative on a manifold, for example, or ask other analytic questions. At each point, analytic properties are defined with respect to the local coordinate system, and the definition of an analytic property is then akin to that of their counterparts in Euclidean space. There are extra conditions on the local coordinate systems, in order to define these properties independent of local coordinate system: the different coordinate systems should be *compatible*, which determines the so-called *smooth strucutre*. For a formal introduction to manifolds, we refer to the book by Tu [16].

We will need a little more structure on our manifold, so we will look at Riemannian manifolds. Formally, these are smooth manifolds with a family of inner products on the tangent bundle, that vary smoothly from point to point. The tangent bundle may be thought of as the space where all tangent vectors to the manifold live. On a two dimensional surface, the tangent bundle is a (disjoint) union of tangent planes, one plane for each point on the surface. For our purposes, smooth manifolds may be thought of as regular surfaces. In fact, regular surfaces are smooth, Riemannian manifolds with an inner product on the tangent bundle: the Euclidean inner product. In particular, open subsets of the complex plane (which can be naturally embedded in the Euclidean plane) are smooth manifolds.

On a Riemann manifold, the inner product can be different from point to point, but has to do so in a smooth way. The family of inner products on the tangent bundle changes the local notion of distance: one would have a (smooth) path, denoted $\phi : [0,1] \to M$ between points on the manifold M, and the length of a curve would be given by

$$L(\phi) = \int_0^1 \langle \phi'(t), \phi'(t) \rangle_{\phi(t)}^{\frac{1}{2}} dt.$$

We put emphasis on the fact that the value of the inner product depends on the position on the manifold, so that similar shaped paths might have a completely different length depending on their position on the manifold.

The Riemannian manifold of interest is the upper half of the complex plane with the so-called hyperbolic metric on it, commonly denoted as $\mathbb{H}$ and known as the hyperbolic half-plane or Poincaré half-plane. This surface was originally used to construct an example of a surface where four of the five axioms of Euclidean geometry [9] held, but the last axiom did not. This answered the more than a millennium old question of whether the 'parallel axiom' was always true in any geometry. Specifically, on $\mathbb{H}$, given a line and two other lines that are not parallel (i.e. have different angles at their intersection with the first line), these last two lines will never intersect. This exactly contradicted Euclid's fifth axiom.

The inner product on the tangent bundle of $\mathbb{H}$, roughly speaking, is defined such that a path between points becomes asymptotically long near the real axis, while a similar shaped path between points far away from the real axis becomes asymptotically short. Moreover, the name of the hyperbolic half-plane is not accidental: in Riemannian geometry there is a notion of curvature, and on $\mathbb{H}$ the curvature is negative one everywhere [17, p. 138]. This notion of curvature extends the curvature of Do Carmo (definition 3.6 and 3.7). Moreover, the notion of a path of shortest distance between points, known as a geodesic (definition 4.8), also extends to Riemannian manifolds. On the hyperbolic half-plane, these are straight lines for points lying above each other (i.e. that have the same real value) and circular sections for all other points, where the center of the circle lies on the real axis.

1.3 Measure preserving maps and ergodicity

Measure theory studies a more general notion of size of a set in a space. For example, on the natural numbers $\mathbb{N}$, we can have a so-called counting measure μ as follows: let A be a subset of the natural numbers, then $\mu(A)$ is equal to the number of elements in A. Another intuitive example is the Borel measure on the real numbers $\mathbb{R}$, where the size of an open interval in $(a, b) \subset \mathbb{R}$ is $b - a$. The size of a general subset A is the infimum of the size of all open coverings U of the subset A, made of intervals. The size of the open covering U is then given as the sum of the sizes of the intervals that make up U. For example, the set $[0, 1) \cup (2, 3)$ can be covered by open intervals $U_n = (-1/n, 1) \cup (2, 3)$, and $\inf_{n\in\mathbb{N}} \mu(U_n) = \inf_{n\in\mathbb{N}}(1 + 1/n) + 1 = 2$. So, the Borel measure of $[0, 1) \cup (2, 3)$ is at most 2. It can in fact be shown to be equal to 2. An introduction to Measure Theory can be found in [6]. Measure Theory is a broad and involved topic, so we will not attempt to introduce it in a few paragraphs.

In physics, specifically statistical mechanics, the macroscopic behaviour of different materials is often modeled on a microscopic scale. The microscopic description of the material gives rise to macroscopic properties. Typically, the microscopic states of the system are described and there are some physical grounds for assigning a probability to each specific microscopic state. The macroscopic quantities arise as a mean value over the microscopic states. In other words, if the set of all possible states is X, also called the *phase space*, and for each state $x \in X$ the probability of that state is $f(x)$, then a macroscopic quantity $U(x)$ depending on the microscopic states has a predicted mean of

$$\langle U \rangle = \int_X U(x)f(x)dx.$$

For example, the mean momentum $\langle P \rangle$ of a gas particle could be deduced using this method.

The justification for the probabilistic nature of a mean value, amongst others, is the assumed 'ergodic' nature of the dynamics of microscopic states. More specifically, it is assumed that the microscopic states are continually changing from one to the other, with a probability proportional to their 'size' within the set X. Here, the connection with measure theory becomes apparent. Moreover, it is assumed that eventually all microscopic states are visited. Another way of spelling out the ergodic assumption is that the time average of a macroscopic quantity is equal to a the 'spatial' average over the phase space

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t U(t')dt' = \int_X U(x)f(x)dx \tag{1}$$

Originally an assumption of statistical mechanics, ergodicity has since been studied by many mathematicians. In particular, the focus is on so-called probability spaces $(X, \mathcal{A}, \mu)$, where X is the a set, $\mathcal{A}$ is a *σ-algebra* and μ is the measure. In particular, the measure of the entire space X is 1.

Initially, one is interested in measure preserving maps [8, Definition 2.1]. They are maps between probability spaces that preserve measure in the following sense: if $(X, \mathcal{A}, \mu)$, $(Y, \mathcal{B}, \nu)$ are probability spaces, then $T : X \to Y$ is a measure preserving map if $\mu\left(T^{-1}(B)\right) = \nu\left(B\right)$ for any measurable set $B \in \mathcal{B}$. If T maps X to itself, then the quadruplet $(X, \mathcal{A}, \mu, T)$ is called a *measure preserving system.* In 1890, Poincaré proved an important theorem about measure preserving maps while studying the 'three-body problem' of planetary orbits, even before measure theory was invented [8, p. 26]. The theorem [8, Theorem 2.11] states, that for any measurable set B in a measure preserving system, almost every point in B is mapped back to B infinitely many times. That is, if T is the measure preserving map, then there exists a sequence of integers for almost every point x in B such that $T^{n_i}(x) \in B$ for $0 < n_0 < n_1 < ....$

In terms of statistical mechanics, measure preserving maps on the phase space are maps such that, starting off with almost any point in the phase space (so any initial microscopic state) with nonzero probability, you will get back to a similar state infinitely many times under the time evolution of the system, which is necessary to justify the averaging.

However this is not sufficient. Given an initial state, we want that the evolution map of the phase space 'traverses' the entire phase space (or at least the parts with non-zero probability of occurring). This is what the mathematical definition of ergodicity formalizes [8, Definition 2.13]: let $(X, \mathcal{A}, \mu)$ be a probability space. A measure preserving map $T : X \to X$ is ergodic if for any $A \in \mathcal{A}$ it holds that

$$T^{-1}(B) = B \implies \mu(B) = 0 \text{ or } \mu(B) = 1.$$

In other words, if a part of the phase space is mapped exactly onto itself, it is either a set of measure zero or the entire space minus a set of measure zero.

As a consequence, all sets of measure between one and zero are not mapped exactly onto themselves by an ergodic map, and it can be shown that almost any point visits any subset of positive measure infinitely often. In paragraph 2.5 and 2.6 of [8] multiple theorems are proved, which have a similar interpretation as equation 1: for an ergodic system, the time average of a quantity depending one the current state of the phase space can be computed as a spatial average of the quantity over the phase space.

1.4 Ergodicity of geodesic flow

Armed with the knowledge of the geodesics on the hyperbolic half-plane and ergodic theory, we can turn our attention to *geodesic flow*: pick a point in the half-plane and a unit vector (with respect to the inner product on the tangent bundle) in a direction, then follow the geodesic in that direction at constant speed. It turns out that the geodesic flow on the half plane is not interesting: following a geodesic leads you in a straight line to infinity (in the imaginary direction) or to the real axis via a circular curve. However, there is a quotient manifold of the hyperbolic half-plane, called the modular surface, which has finite volume and much more interesting geodesic flow, which is discussed in chapter 9 of the book by Einsiedler and Ward [8].

The construction of the modular surface follows from considering so-called Möbius transformations of $\mathbb{H}$. The Möbius transformations $g \in SL_2(\mathbb{R})$ are real, two by two matrices with unit determinant, and act on $z \in \mathbb{H}$ by

$$g(z) = \begin{bmatrix} a & b \\ c & d \end{bmatrix} z = \frac{az+b}{cz+d}.$$

They have the special property that they map geodesics to geodesics and are isometries of the hyperbolic metric. A specific subset of the Möbius transformations, namely those elements of $SL_2(\mathbb{R})$ with integer entries, 'generates' the quotient manifold: if we choose the grayed area in figure 1, with $-1/2 < \mathrm{Re}(z) < 1/2$ and $|z| > 1$, as our *fundamental domain*, then there is an element of the subset of Möbius transformations that maps this fundamental domain exactly to one other area enclosed by three geodesics, which are the blue lines in the picture. This is done bijectively, so if we identify a point in the fundamental domain with its image under these specific Möbius transformations, we have our quotient manifold. Pictorially, the modular surface is identified with the fundamental domain.

The geodesic flow on the modular surface allows for the computation of a so-called Poincaré first return map. That is to say, given a point on the imaginary axis and a tangent vector of unit length, if we follow its geodesic flow we can predict when the first return to the imaginary axis will be, and the angle at which it returns. This allows us to switch from a continuous dynamical system to a discrete dynamical system, from the continuous geodesic flow on the quotient manifold to the iteration of the Poincaré first

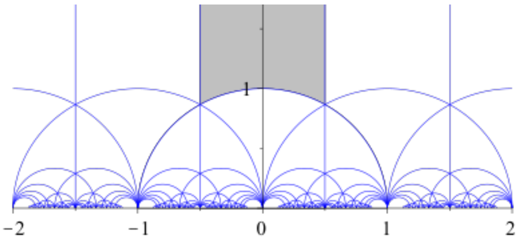

Figure 1: the modular surface [1]

return map. In fact, it turns out that the specific form of the first return map involves the Gauss map $G : \mathbb{R} \to \mathbb{R}$

$$G(x) = \frac{1}{x} - \left\lfloor \frac{1}{x} \right\rfloor .$$

The Gauss map was originally used to study continued fractions. Given a real number s, one can write s as a continued fraction [8, Lemma 3.6]:

$$s = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}}$$

where $a_0 \in \mathbb{Z}$, $a_i \in \mathbb{N}$ for $i > 0$. This is often abbreviated to $s = [a_0; a_1, a_2, a_3, \cdots]$.

The method of finding the coefficients a_i is done using the Gauss map: write s in decimal notation, and let a_0 be the unique integer such that $0 < s - a_0 < 1$. Then, express $\frac{1}{s-a_0}$ in decimal expansion and again find the unique (positive) integer such that $0 < \frac{1}{s-a_0} - a_1 < 1$. This may be written more compactly as $a_1 = \frac{1}{s-a_0} - G\left(\frac{1}{s-a_0}\right)$. Indeed, $G(s)$ can be interpreted as the fractional part of s. Repeating the procedure on

$$\frac{1}{\frac{1}{s-a_0} - a_1}$$

one may find a_2, and so on.

For irrational numbers x, y between zero and one, the leading coefficient is zero and the continued fraction expansion is endless. Moreover, G maps such a number to another irrational number between zero and one, and it can be shown that if $G(x) = y$ and $x = [a_1, a_2, a_3, \cdots]$, then $y = [a_2, a_3, a_4, \cdots]$. Hence, the Gauss map can be seen as a shift on the coefficients [8, p. 83]. This is relevant, because it can be shown [8, theorem 3.7] that the Gauss map is ergodic on the interval $(0, 1)$ with respect to the *Gauss measure* using theory known from general shift maps. The ergodicity of the Gauss map provides interesting results in number theory for continued faction expansions, using the averaging discussed earlier.

Finally, the ergodicity of the Gauss map and its relation to the geodesic flow on the modular surface provides a key connection in mathematics between dynamical systems, geometry, and number theory: it can be shown that the geodesic flow on the modular surface is ergodic if and only if the Gauss map is ergodic on the interval $(0, 1)$ [3], [8, Proposition 9.25]. Historically, Artin first showed the ergodicity of the geodesic flow via the ergodicity of the Gauss map [3]. Einsiedler and Ward show the converse direction and provide a version of Hopf's argument [10] for the geodesic flow instead. Since the work of Artin, geodesic flow on arithmetic surfaces has since become an important topic of study for mathematicians, connecting dynamical systems, number theory and geometry.

1.5 Structure

The thesis starts by introducing the Poincaré half-plane (2) and its isometries, the Möbius transformations, (3), and works towards the description of its geodesics (4) and geodesic flow (5). After that, it investigates the modular surface and its geodesic flow (6). As the goal is to show that the geodesic flow on the modular is ergodic, some more background in Ergodic Theory is provided (7). After proving the ergodic nature of the flow, the Poincaré section on the modular surface is introduced, and it is shown how one can obtain the Gauss map as a first return map for the geodesic flow (8). Up to and including the proof of the ergodicity of the geodesic flow, the author has followed the structure and content of chapter 9 of M. Einsiedler and T. Ward [8]. For the Poincaré section, the author used the work of C. Series [15]. As an epilogue, some generalizations and possible further material are introduced.

The author would like to thank the supervisor for suggesting appropriate materials and giving an thorough introduction to the subject, as well as for giving repeated feedback on drafts.

2 Poincaré half-plane

A model for hyperbolic geometry is given by the pair $(\mathbb{H}, d)$, where the set

$$\mathbb{H} = \{x + yi \in \mathbb{C} |\, x \in \mathbb{R},\ y > 0\}$$

is the upper half of the complex plane, and the metric $d : \mathbb{H} \times \mathbb{H} \to \mathbb{R}$ is known as the hyperbolic metric. Together they are known as the Poincaré half-plane. Throughout this text, we will use the convention that x, y refer respectively to the real and imaginary parts of $z \in \mathbb{H}$ unless specified otherwise.

Before the hyperbolic metric is defined, we first introduce the tangent bundle of $\mathbb{H}$:

$$\mathrm{T}\mathbb{H} := \bigsqcup_{z \in \mathbb{H}} \mathrm{T}_z\mathbb{H} = \bigsqcup_{z \in \mathbb{H}} \{z\} \times \mathbb{C} = \mathbb{H} \times \mathbb{C}$$

which can be thought of as a complex plane for each point in $\mathbb{H}$, where vectors spawning from points in $\mathbb{H}$ can naturally live. The emphasis is on the fact that the tangent bundle is a disjoint union. For each fixed $z \in \mathbb{H}$ we write $\mathrm{T}_z\mathbb{H} = \{z\} \times \mathbb{C}$ for the *tangent space* of z. The tangent bundle is given a family of inner products $\langle \cdot, \cdot \rangle_z : \mathrm{T}_z\mathbb{H} \times \mathrm{T}_z\mathbb{H} \to \mathbb{R}$, for $(z, v), (z, w) \in \mathrm{T}_z\mathbb{H}$

$$\langle (z, v), (z, w) \rangle = \frac{1}{y^2} v \cdot w$$

where $v \cdot w$ denotes the ordinary Euclidean inner product on $\mathbb{C}$. Since the inner product on one tangent space only takes in vectors from that space, we simplify the notation: $\langle v, w \rangle_z \coloneqq \langle (z, v), (z, w) \rangle$. The hyperbolic metric on $\mathbb{H}$ is defined in terms of piecewise smooth curves between two points, which we will call paths, and this inner product on the tangent bundle of $\mathbb{H}$.

For a piecewise smooth curve $\phi : [0, 1] \to \mathbb{H}$ we define the derivative of ϕ at $t \in [0, 1]$ as $D\phi(t) = (\phi(t), \phi'(t))$ and we can conclude that $D\phi(t) \in \mathrm{T}_{\phi(t)}\mathbb{H}$. The derivative of ϕ at t can be thought of as the tangent vector to the curve spawning from the point $\phi(t)$ on the curve, agreeing with our analogy for $\mathrm{T}\mathbb{H}$. In analogy with the length of a path in Euclidean space, the length of a path $\phi : [0, 1] \to \mathbb{H}$ is defined as

$$L(\phi) = \int_0^1 \|D\phi(t)\|_{\phi(t)} dt$$

where the norm is induced by the inner product on $\mathrm{T}_z\mathbb{H}$. Similarly, the speed of the curve at t is $\|D\phi(t)\|_{\phi(t)}$. Note that the speed of a curve grows asymptotically near the real line, regardless of the direction one travels in. The length of paths near the real line are much longer than similar shaped paths far away from it.

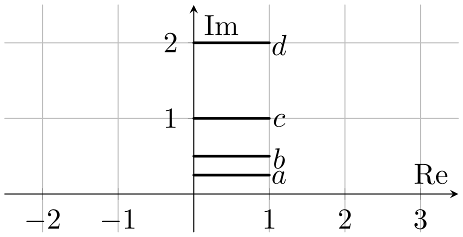

Figure 2: paths for $n = -4, -2, 1, 2$

Example 2.1. Consider the 'straight' paths between $z_0 = i, z_1 = i + 1$, and $z_{-2n} = i/n, z_{-2n+1} = i/n + 1$ and $z_{2n} = ni, z_{2n+1} = ni + 1$, given by

$$\phi_1(t) = i + t, \quad \phi_{-n}(t) = i/n + t, \quad \phi_n(t) = ni + t.$$

The length of the paths are

$$L(\phi_1) = \int_0^1 \|1\|_{i+t} dt = \int_0^1 \frac{1}{1}|1|dt = 1$$

$$L(\phi_{-n}) = \int_0^1 \|1\|_{i/n+t} dt = \int_0^1 n|1|dt = n$$

$$L(\phi_n) = \int_0^1 \|1\|_{ni+t} dt = \int_0^1 \frac{1}{n}|1|dt = \frac{1}{n}$$

Although they look like similarly shaped paths, near the real axis the paths become asymptotically long, whilst the paths become asymptotically short the further one goes in the imaginary direction. △

We prove a Lemma related to the length of paths such as those in figure 3. This will be useful for finding a metric on $\mathbb{H}$.

Lemma 2.2. Suppose there are three points $a, b, c \in \mathbb{H}$ such that $\mathrm{Im}(a) < \mathrm{Im}(b) = \mathrm{Im}(c)$ and $\mathrm{Re}(a) = \mathrm{Re}(b)$. Let $\phi : [0, 1] \to \mathbb{H}$ be a path such that $\phi(0) = a$, $\phi(1) = c$. Then $\psi(t) = \mathrm{Re}(a) + \mathrm{Im}(\phi(t))$ is a path such that $\psi(0) = a, \psi(1) = b$ and $L(\psi) \leq L(\phi)$

Proof. ϕ is piecewise differentiable, hence its imaginary component is too. So ψ is piecewise differentiable. From the assumptions on a, b, c it follows that $\psi(0) = a, \psi(1) = b$. Finally, we may decompose $\phi'(t)$ in its imaginary and real component and deduce that

$$L(\phi) = \int_0^1 \frac{1}{\mathrm{Im}(\phi(t))} \| \mathrm{Re}(\phi'(t)) + \mathrm{Im}(\phi'(t))i\| dt \geq \int_0^1 \frac{1}{\mathrm{Im}(\phi(t))} \| \mathrm{Im}(\phi'(t))\| dt = L(\psi)$$

□

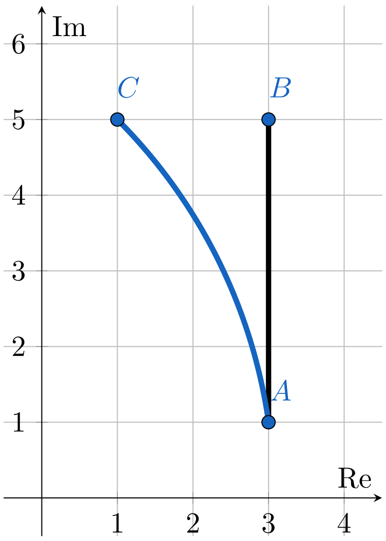

Figure 3: example of Lemma 2.2

The hyperbolic metric can now be defined as follows: for $z_0, z_1 \in \mathbb{H}$ and paths ϕ such that $\phi(0) = z_0$, $\phi(1) = z_1$,

$$d(z_0, z_1) = \inf_{\phi} L(\phi).$$

Proposition 2.3. $d : \mathbb{H} \times \mathbb{H} \to \mathbb{R}$ is a metric.

Proof. We want to check that d is well defined, i.e. the infimum is not taken over an empty set. This is indeed the case, since the straight line connecting two points is always a path: let $\phi(t) = z_0 + t(z_1 - z_0) = x_0 + t(x_1 - x_0) + (y_0 + t(y_1 - y_0))i$. With respect to the Euclidean topology on $\mathbb{C}$ it is differentiable, $\phi(0) = z_0$, $\phi(1) = z_1$ and for $t \in [0, 1]$, $\operatorname{Im}(\phi(t)) > 0$. It remains to check the axioms of a metric.

1. (Non-zero positive distance between unequal points, zero distance between equal points) For any path ϕ, $\|D\phi(t)\|_{\phi(t)}$ is non-negative by definition. Hence $L(\phi) \geq 0$. The infimum respects non-strict inequalities, so $d(z_0, z_1) \geq 0$. Moreover, $\phi(t)$ is continuous, so $1/\operatorname{Im}(\phi(t))$ is too, and $\phi'(t)$ is piecewise continuous. This implies that $\|D\phi(t)\|_{\phi(t)}$ is piecewise continuous. Hence, $L(\phi) = 0 \iff \|D\phi(t)\|_{\phi(t)} \equiv 0$. Recall that in Euclidean geometry, the paths of shortest distance are straight lines. Let $\delta = \|z_0 - z_1\| > 0$. Suppose that $z_0 \neq z_1$ but $d(z_0, z_1) = 0$. Since $z_0 \neq z_1$, we must have for any path ϕ connecting them that $\phi'(t) \neq 0$ for some $t \in [0, 1]$. And, since $d(z_0, z_1) = 0$, $\forall \epsilon > 0\, \exists \phi$ a path, such that $L(\phi) < \epsilon$. Since ϕ is a continuous map from a compact subset of $\mathbb{R}$ to $\mathbb{C}$, $\phi([0, 1])$ is closed and bounded, so there exists a $s \in [0, 1]$ such that $\operatorname{Im}(\phi(t)) \leq \operatorname{Im}(\phi(s)) = y$. It can be deduced that

$$\epsilon > \int_0^1 \|D\phi(t)\|_{\phi(t)} dt \geq \frac{1}{y} \int_0^1 \|\phi'(t)\| dt \geq \frac{1}{y}\delta$$

If we pick $\epsilon = \delta/n$, it follows that a path with length less than ϵ, must at least reach a y coordinate of n. It can be checked that $\tilde{\phi}(t) = \phi(ts)$ is a path from z_0 to $\phi(s)$. As noted above, the integrand is non-negative, so $L(\phi) \geq L(\tilde{\phi})$. But, using Lemma 2.2 we see that $L(\tilde{\phi}) \geq L(\psi)$ if we let $\psi(t)$ be the path from z_0 to $x_0 + ni$. For small enough epsilon, $n > \text{Im}(z_0)$, in which case

$$L(\psi) = \int_0^1 \frac{1}{y_0 + t(n - y_0)}(n - y_0)dt = \ln n - \ln y_0$$

The contradiction now follows for large enough n from the inequalities

$$\frac{\delta}{n} = \epsilon > L(\phi) \geq \ln(n) - \ln(y_0).$$

If $z_0 = z_1$, then $\phi(t) \equiv z_0$ is a path with length zero, so $d(z_0, z_1) = 0 \iff z_0 = z_1$.

2. (Symmetry in the arguments) Symmetry follows since paths are 'reversible' without changing length. Suppose $\phi(t)$ is a path from z_0 to z_1 then $\phi(1-t)$ is a path from z_1 to z_0, using the chain rule to verify piecewise differentiability. Substituting $s = 1 - t$,

$$\int_0^1 \|D\phi(1-t)\|_{\phi(1-t)}dt = -\int_1^0 \|D\phi(s)\|_{\phi(s)}ds = \int_0^1 \|D\phi(s)\|_{\phi(s)}ds$$

Hence we may conclude that d is symmetric in its arguments.

3. (Triangle inequality) Suppose ϕ_{02}, ϕ_{21} are paths between z_0, z_2 and z_2, z_1 respectively. Then

$$\psi(t) = \begin{cases} \phi_{02}(2t) & 0 \leq t \leq 1/2 \\ \phi_{21}(2t + 1/2) & 1/2 < t \leq 1 \end{cases}$$

is a piecewise differentiable function (since its parts are, and ϕ_{02}, ϕ_{21} agree at $t = 1/2$) with $\psi(0) = z_0$ and $\psi(1) = z_1$. It is easy to show that $L(\phi) = L(\phi_{02}) + L(\phi_{21})$ using some substitutions. The set of all paths ψ from z_0 to z_1 constructed in this fashion are a subset of all possible paths from z_0 to z_1. From this we may conclude that

$$d(z_0, z_1) \leq \inf_{\psi} L(\psi) = \inf_{\phi_{02},\phi_{21}} (L(\phi_{02}) + L(\phi_{21})) =$$

$$\inf_{\phi_{02}} L(\phi_{02}) + \inf_{\phi_{21}} L(\phi_{21}) = d(z_0, z_2) + d(z_2, z_1).$$

The triangle inequality holds.

□

Now that we have a metric space $(\mathbb{H}, d)$, we 'complete' it by adding its boundary $\partial\mathbb{H}$: the real line, together with the point at infinity. Adding the point at infinity makes $\mathbb{H} \cup \partial\mathbb{H}$ compact [8, p. 274]. The hyperbolic distance in $\mathbb{H} \cup \partial\mathbb{H}$ is infinite from any point in $\partial\mathbb{H}$ to any other point in the space. One may do this by defining an extension of d to $\mathbb{H} \cup \partial\mathbb{H}$ and showing that this is again a metric, and one may justify it by trying to find a path from any point in $\mathbb{H}$ to a path arbitrarily close to a point in $\partial\mathbb{H}$: such paths become arbitrarily long.

3 Isometries and congruences

In the Euclidean plane, it was possible to construct the torus $\mathbb{T}^2$ by considering the square $[0,1] \times [0,1]$ and identifying the right boundary with the left, and the top with the bottom. The torus has some nice properties compared to the entire plane: it has finite volume (w.r.t. the Euclidean measure of volume) and is compact. Another way of constructing the torus is by tiling the plane into squares, where the boundaries are the lines between integer coordinates $\mathbb{Z}^2 \subset \mathbb{R}^2$. The equivalence relation

$$(x_0, y_0) \sim (x_1, y_1) \in \mathbb{R}^2 \iff (x_0 - x_1, y_0 - y_1) \in \mathbb{Z}^2$$

then identifies a point $p = (x, y)$ in $[0,1] \times [0,1]$ with all the points q in the plane of the form $q = (x + n, y + m)$, $n, m \in \mathbb{Z}$. The quotient space $\mathbb{R}^2/\sim$, denoted by $\mathbb{R}^2/\mathbb{Z}^2$ is isomorphic to the first construction of $\mathbb{T}^2$: one can check that addition and scalar multiplication are preserved under the bijective map $\iota : \mathbb{T}^2 \to \mathbb{R}^2/\mathbb{Z}^2$, $p \mapsto [p]$. By picking the representative of $[p]$ to lie in $[0,1] \times [0,1]$, it becomes clear that $\mathbb{R}^2/\mathbb{Z}^2$ can be thought of as $\mathbb{T}^2$. Note that by adding elements of $\mathbb{Z}^2$ as a vector, we can translate the square at the origin to anywhere in the plane. These translations also preserve distances within the square.

Similar to the second construction, the Poincaré half-plane $\mathbb{H}$ can also be 'tiled' by an equivalence relation, and we will construct a finite volume, compact metric space where one of the tiles will be the representative of the quotient. The 'translations' of the representative tile will preserve distances, which will be of interest when we study the dynamics on the quotient space. We start by studying the *Möbius transformations.*

3.1 Tranformations of the Poincaré half-plane

A specific class of transformations of a space are group actions:

Definition 3.1. Given a group G and a set X, a *group action* is a map $\varphi : G \times X \to X$, $(g, x) \mapsto \varphi(g, x)$ (which will be denoted as $g \cdot x$ or $g(x)$) such that

- $e \cdot x = x, \quad \forall x \in X$
- $(gh) \cdot x = g(h \cdot x), \quad \forall g, h \in G, \forall x \in X$

A group action is called *transitive* if for all $x_1, x_2 \in X$ there exists a $g \in G$ such that $g \cdot x_1 = x_2$. If there is a unique $g \in G$ such that $g \cdot x_1 = x_2$, then the group action is called *simply transitive.*

Simply transitive group actions allow all points y in a space X to be written uniquely as $g \cdot x$ for a single, fixed x in X.

Instead of linear transformations, we want to study certain transformations on the Poincaré half-plane. These are the Möbius transformations

$$g = \begin{bmatrix} a & b \\ c & d \end{bmatrix} : z \mapsto \frac{az+b}{cz+d}$$

where $g \in SL_2(\mathbb{R})$, the special linear group, which consists of 2×2 matrices which have a unit determinant and real entries. This map is well defined: c and d are never both zero, and $cz + d = 0$ implies

1. $z = -d/c \in \mathbb{R}$ if $c \neq 0$
2. $d = 0$ if $c = 0$

which both lead to a contradiction. Finally, we need to check that $g(z) \in \mathbb{H}$, i.e. that $\mathrm{Im}(g(z)) > 0$. In fact,

$$g(z) = \frac{(az+b)(c\bar{z}+d)}{|cz+d|^2}.$$

Hence,

$$\mathrm{Im}(g(z)) = \frac{\mathrm{Im}((ad-bc)z)}{|cz+d|^2} = \frac{\mathrm{Im}(z)}{|cz+d|^2} > 0. \tag{2}$$

The Möbius transformations will allow us to write every element of $\mathbb{H}$ as $g \cdot i$ for some Möbius transformation g. Moreover, we will be able to identify all geodesics in $\mathbb{H}$ by them. We first need a few more results.

Proposition 3.2. The Möbius transformations define a group action on $\mathbb{H}$.

Proof. To show this, consider the complex projective space $\mathbb{P}(\mathbb{C})$. This is $\mathbb{C}^2 \setminus (0,0)$ where a point is identified with all of its complex multiples, i.e. $(z_1, z_2) \sim \lambda(z_1, z_2)$, $\lambda \in \mathbb{C} \setminus \{0\}$. One can verify that $(z, 1)$, $z \in \mathbb{C}$ is a unique class in $\mathbb{P}(\mathbb{C})$ for each unique $z \in \mathbb{C}$ and that these are all the equivalence classes in the space. Hence, there is a bijection between $\mathbb{C}$ and $\mathbb{P}(\mathbb{C})$. Let us define $g \in SL_2(\mathbb{R})$ to act in this space as

$$g((z,1)) = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix} = \begin{bmatrix} az+b \\ cz+d \end{bmatrix} \sim \begin{bmatrix} \frac{az+b}{cz+d} \\ 1 \end{bmatrix}$$

Since this is just matrix multiplication as we know it, it is clear that $SL_2(\mathbb{R})$ is a group action over $\mathbb{P}(\mathbb{C})$. Because of the above equivalence and the bijection, we are done. □

Before we continue, note that if $g = \pm I_2$, then $g \cdot z = z$. Hence, for any $g \in SL_2(\mathbb{R})$ we have that $(\pm g) \cdot z = g \cdot z$, so we identify $\pm g$, where we will write

$$g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

to mean both. The set $SL_2(\mathbb{R})/\{\pm I_2\}$ is called the projective special linear group, denoted by $PSL_2(\mathbb{R})$. The action of $PSL_2(\mathbb{R})$ can be extended to $\mathrm{T}\mathbb{H}$ using the derivative D: define the *derivative action* to be $Dg : \mathrm{T}\mathbb{H} \to \mathrm{T}\mathbb{H}$,

$$Dg(z, v) = \big(g(z), g'(z)v\big)$$

where $g'(z) = 1/(cz+d)^2$. One can verify that $DI_2(z, v) = (z, v)$ and, using the chain rule, that $(Dg)Dh \cdot (z, v) = D(gh) \cdot (z, v)$, so indeed it is an action. Note that for a fixed $z \in \mathbb{H}$, we have that $Dg(z, \cdot)$ is a map from $\mathrm{T}_z\mathbb{H}$ to $\mathrm{T}_{g(z)}\mathbb{H}$. The definition of the derivative action of $PSL_2(\mathbb{R})$ is compatible with the derivative itself: for a path in $\mathbb{H}$ we have that $(Dg)D\phi(t) = (Dg)(\phi(t), \phi'(t)) = (g \cdot \phi(t), g'(\phi(t))\phi'(t)) = D(g \cdot \phi(t))$.

Lemma 3.3. The action of $PSL_2(\mathbb{R})$ on $\mathbb{H}$ is transitive and isometric w.r.t. the hyperbolic metric. Moreover, the derivative action Dg on $\mathrm{T}\mathbb{H}$ preserves the Riemannian metric.

Proof. To show transitivity, it is enough to consider that

$$g = \begin{bmatrix} \sqrt{y} & x/\sqrt{y} \\ 0 & 1/\sqrt{y} \end{bmatrix}$$

belongs to $PSL_2(\mathbb{R})$ and that $g(i) = x + yi$. Since g is invertible and $PSL_2(\mathbb{R})$ is a group, we have that the action is transitive.

To show that the action is isometric, we first prove the last statement. Let $(z, v), (z, w) \in \mathrm{T}_z\mathbb{H}$ and $g \in PSL_2(\mathbb{R})$. Then

$$\langle Dg(z, v), Dg(z, w)\rangle_{g(z)} = \left(\frac{y}{|cz+d|^2}\right)^{-2} \left\langle \frac{1}{(cz+d)^2}v, \frac{1}{(cz+d)^2}w \right\rangle = \frac{1}{y^2}\langle v, w\rangle$$

using equation 2 and scalar multiplication properties of the Euclidean inner product on $\mathbb{C}$. Indeed, the Riemannian metric is preserved.

Finally, we can show that the action is isometric. Suppose that ϕ is a path in $\mathbb{H}$ from z_0 to z_1, then $g \cdot \phi(t)$ is piecewise differentiable by the chain rule, and hence $g \cdot \phi(t)$ is a path from $g(z_0)$ to $g(z_1)$. Similarly, any path from $g(z_0)$ to $g(z_1)$ becomes a path from z_0 to z_1 under the transformation $g^{-1} \in PSL_2(\mathbb{R})$. If we denote $\{\phi\}$ as the set of all paths from z_0 to z_1 and $\{\psi\}$ as the set of all paths from $g(z_0)$ to $g(z_1)$, we can conclude that g is a bijection between them. From the preserved Riemannian metric one can deduce that $\|D(g(\phi(t)))\|_{g(\phi(t))} = \|D\phi(t)\|_{\phi(t)}$ and hence that $L(g \cdot \phi) = L(\phi)$. Finally,

$$d(z_0, z_1) = \inf_{\phi} L(\phi) = \inf_{g \cdot \phi} L(g \cdot \phi) = \inf_{\psi} L(\psi) = d(g(z_0), g(z_1))$$

□

Example 3.4. The action of $PSL_2(\mathbb{R})$ is not simply transitive. Let $g, h \in PSL_2(\mathbb{R})$ be defined as

$$g = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad h = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

One can verify that $g \cdot i = h \cdot i = i$ △

As noted, if we can find a group whose action is simply transitive on $\mathbb{H}$, we can represent every point in $\mathbb{H}$ uniquely as $z = g \cdot i$ for some unique g in the group. Any other 'basepoint' than i would do as well, but for convenience we will choose i. Another point of interest is finding the geodesics on $\mathbb{H}$, parameterized by arclength (or unit speed, equivalently). The geodesics will allow us find paths of minimal distance between two points, giving us a better understanding of $\mathbb{H}$, and in later chapters we will want to study geodesic flow on $\mathbb{H}$. For these reasons, we focus on $\mathrm{T}^1\mathbb{H} = \{(z, v) \in \mathrm{T}\mathbb{H} |\, \|v\|_z = 1\}$, the so called unit tangent bundle. By the above Lemma, the derivative action of $PSL_2(\mathbb{R})$ preserves the Riemannian metric, so the derivative action restricted to $\mathrm{T}^1\mathbb{H}$ is still an action. It would be useful too if the derivative action of $PSL_2(\mathbb{R})$ would be simply transitive on $\mathrm{T}^1\mathbb{H}$. In fact it is, but in order to show it, we first return to the transitivity of $PSL_2(\mathbb{R})$ on $\mathbb{H}$.

Note that in the proof of the above Lemma, we have shown that there is an element of $PSL_2(\mathbb{R})$ mapping i to any z in $\mathbb{H}$. Suppose there are two distinct elements $g, h : i \mapsto z$, then $g^{-1} \cdot h : i \mapsto i$. If we can find all maps $s \in PSL_2(\mathbb{R})$ such that $s : i \mapsto i$, we can write $h = gs$ for some s. Hence, if we can find the so called *stabilizer* of i in $PSL_2(\mathbb{R})$, which is defined as

$$\mathrm{Stab}_{PSL_2(\mathbb{R})}(i) = \{g \in PSL_2(\mathbb{R}) |\, g \cdot i = i\}$$

we can characterize all elements that map i to z. Given any $s \in \mathrm{Stab}_{PSL_2(\mathbb{R})}(i)$ we must have that $|ci + d| = 1$ by equation 2, so for some $\theta \in [0, 2\pi)$, $c = \sin(\theta)$ and $d = \cos(\theta)$. Writing $g \cdot i = i$ out gives

$$\frac{ai + b}{\sin(\theta) + \cos(\theta)i} = i$$

which is equivalent to

$$g = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}.$$

In other words, the stabilizer of i looks like the special orthogonal group, $SO_2(\mathbb{R})$ consisting of matrices with orthonormal vectors as entries. In fact, since we are working over $PSL_2(\mathbb{R})$, the stabilizer is $PSO_2(\mathbb{R})$, the projective special orthogonal group. All elements of $PSL_2(\mathbb{R})$ mapping i to z are $g \cdot PSO_2(\mathbb{R})$ for one $g : i \mapsto z$. With the identification of all these elements, $PSL_2(\mathbb{R})/PSO_2(\mathbb{R})$ is simply transitive on $\mathbb{H}$. Moreover, since $PSL_2(\mathbb{R})$ are isometries, $\mathbb{H}$ is in fact congruent to $PSL_2(\mathbb{R})/PSO_2(\mathbb{R})$ acting on our base-point i, a relation we will denote as

$$\mathbb{H} \cong PSL_2(\mathbb{R})/PSO_2(\mathbb{R}).$$

We can now return to the derivative action of $PSL_2(\mathbb{R})$ on $\mathrm{T}^1\mathbb{H}$. We will determine that the derivative action of $PSL_2(\mathbb{R})$ is simply transitive on the unit tangent bundle, and, similarly, $\mathrm{T}^1\mathbb{H}$ is congruent to $PSL_2(\mathbb{R})$ acting on the base-point (i, i), denoted

$$\mathrm{T}^1\mathbb{H} \cong PSL_2(\mathbb{R}).$$

Lemma 3.5. The derivative action of $PSL_2(\mathbb{R})$ on $T^1\mathbb{H}$ is simply transitive.

Proof. We wish to show that there exists a unique $\tilde{g} \in PSL_2(\mathbb{R})$ such that $(z, v) \in \mathrm{T}^1\mathbb{H}$ can be written as $(z, v) = D\tilde{g}(i, i)$. From the preceding discussion, $PSL_2(\mathbb{R})/PSO_2(\mathbb{R})$ is simply transitive on $\mathbb{H}$, so there certainly exists $g \in PSL_2(\mathbb{R})$ such that $g^{-1}(z) = i$. Elements of $PSO_2(\mathbb{R})$ leave i fixed, so we may multiply g^{-1} with arbitrary $h \in PSO_2(\mathbb{R})$ from the left to find a suitable transformation s.t.

$$D\left(hg^{-1}\right)(z, v) = \left(i, \left(hg^{-1}\right)'(z)v\right) = (i, i).$$

Denote $\tilde{v} = \left(g^{-1}\right)'(z)v$. It follows that $\left(hg^{-1}\right)'(z)v = h'(i)\tilde{v} =$

$$\frac{1}{(\sin(\theta)i + \cos(\theta))^2}\tilde{v} = e^{-2\theta i}\tilde{v}.$$

As v is a unit vector and $PSL_2(\mathbb{R})$ preserves the Riemannian metric, two choices of $\theta \in [0, 2\pi)$ are possible, with π radians difference. But, in $PSL_2(\mathbb{R})$ the corresponding elements $\pm h$ are identified. Hence, $\tilde{g} = gh^{-1}$. □

The established congruence relations on $\mathbb{H}$ and $\mathrm{T}^1\mathbb{H}$ will play a central role in the next chapters: we will use them to find all paths of shortest distance between two points in $\mathbb{H}$, and to define arithmetic surfaces (specifically the modular surface), as well as properties of different flows on $\mathbb{H}$ and the modular surface.

4 Geodesics

The hyperbolic metric d on $\mathbb{H}$ is defined in terms of the infimum of the lengths of paths between two points. It was possible to show that d was indeed a metric, without finding curves of length equal to the infimum. Such curves might not even exist! We will construct one such curve explicitly, and using our congruence relations on $\mathbb{H}$ and $\mathrm{T}^1\mathbb{H}$ we can then characterize all *geodesics* on $\mathbb{H}$, which are the curves of shortest length between points. Note that for our metric space $(\mathbb{H}, d)$ this means that $d(z_0, z_1) = L(\phi)$ if ϕ is a geodesic between points.

Proposition 4.1. *Let $0 < y_0 < y_1$ and let $z_0 = y_0 i, z_1 = y_1 i$. Then there is a geodesic $\phi : [0,1] \to \mathbb{H}$ from z_0 to z_1 which is unique up to a piecewise differentiable, increasing function $f : [0,1] \to [0,1]$, meaning that any such geodesic ψ can be written as $\psi = \phi \circ f$. Moreover, the length of the geodesic is $\ln(y_1) - \ln(y_0)$ and*

$$\phi(t) = y_0 \left(\frac{y_1}{y_0}\right)^t i$$

defines a parameterization of the geodesic with constant speed.

Proof. It is easily verified that $\|D\phi(t)\|_{\phi(t)} = \ln(y_1) - \ln(y_0)$, a constant, and that $L(\phi) = \ln(y_1) - \ln(y_0)$. This shows that $d(z_0, z_1) \leq L(\phi)$. Suppose there is any other path ψ connecting the two points, with $L(\psi) \leq L(\phi)$. We may decompose $\psi(t)$ into its real and imaginary part and find that they are both piecewise differentiable, since $\phi(t)$ is. Note that $\mathrm{Im}(\phi(0)) = y_0$ and $\mathrm{Im}(\phi(1)) = y_1$, so $\mathrm{Im}(\phi(t))$ is itself a path from z_0 to z_1. This yields

$$L(\psi) = \int_0^1 \frac{\|\psi(t)\|}{\mathrm{Im}(\psi(t))} dt \geq \int_0^1 \frac{|\,\mathrm{Im}(\psi(t))'|}{\mathrm{Im}(\psi(t))} dt \geq \int_0^1 \frac{\mathrm{Im}(\psi(t))'}{\mathrm{Im}(\psi(t))} dt = \ln(|\,\mathrm{Im}(\psi(t))|)|_0^1 = L(\phi).$$

This means that $L(\phi)$ is minimal after all, and equality holds if and only if $\mathrm{Im}(\psi(t))' \geq 0$ and $\mathrm{Re}(\psi(t)) = 0$ for all $t \in [0,1]$. In other words, ψ is also a geodesic if and only if its graph is the same as that of ϕ. If $\mathrm{Im}(\psi(t)') > 0$ everywhere, the function f may be computed by inverting

$$s(t) = \frac{1}{L(\phi)} \int_0^t \|D\psi(t)\|_{\psi(t)}$$

which is piecewise invertable, using the inverse function theorem. If not, a similar procedure will yield f after 'cutting out' the parts where $\mathrm{Im}(\psi(t)) = 0$, which won't be elaborated on. □

To use the established congruences, we reparameterize $\phi(t)$ by arclength. We will call the piecewise differentiable curves of shortest distance between two points, parameterized by arclength, the *geodesic paths.* So, by abuse of notation, $\phi : [0, d(z_0, z_1)] \to \mathbb{H}$, $s \mapsto \phi(s)$ and it follows that $D\phi(s) \in \mathrm{T}^1\mathbb{H}$. For the constructed curve above, it yields

$$\phi(s) = y_0 e^s i.$$

As was established in Lemma 3.3, elements of $PSL_2(\mathbb{R})$ are isometries and map curves to new curves with equal length, bijectively. This means that geodesic paths are also mapped to geodesic paths by elements of $PSL_2(\mathbb{R})$. Conversely, all geodesic paths can constructed as a Möbius transformation of the previously constructed geodesic path, where $y_0 = 1$.

Proposition 4.2. Given $z_0, z_1 \in \mathbb{H}$, there exists a unique geodesic path $\phi(s)$ from z_0 to z_1. Moreover, there is a unique $g \in PSL_2(\mathbb{R})$ such that $\phi(s) = g\left(e^s i\right)$.

Proof. The goal is to first find an appropriate g, which will show the existence of the geodesic. By Lemma 3.1, there exists a $f \in PSL_2(\mathbb{R})$ such that $f^{-1}(z_0) = i$. As in Lemma 3.5, we may multiply f^{-1} on the left by some $h \in PSO_2(\mathbb{R})$, leaving i unchanged, so that $hf^{-1}(z_1)$ lies on the imaginary axis, and above i: denote $f^{-1}(z_1)$ by $\tilde{z} = \tilde{x} + \tilde{y}i$, then

$$hf^{-1}(z_1) = \frac{\cos(\theta)\tilde{z} - \sin(\theta)}{\sin(\theta)\tilde{z} + \cos(\theta)} = \frac{\cos(\theta)\sin(\theta)\left(|\tilde{z}|^2 - 1\right) + \cos^2(\theta)\tilde{z} - \sin^2(\theta)\bar{\tilde{z}}}{|\sin(\theta)i + \cos(\theta)|^2}.$$

The real part of the numerator is

$$\frac{1}{2}\sin^2(2\theta)\left(|\tilde{z}|^2 - 1\right) - \cos^2(2\theta)\tilde{x}$$

which is zero if $\tan(2\theta) = \frac{2\tilde{x}}{|\tilde{z}|^2 - 1}$ for $|\tilde{z}| \neq 1$, and $2\theta = \pi/2 + \pi\mathbb{Z}$ otherwise. In either case, there are four solutions lying in $[0, 2\pi)$ of which the first and third, and the second and fourth lie π radians apart. For both pairs, there is a unique $h \in PSL_2(\mathbb{R})$ for which $\tilde{z}$ is mapped onto the imaginary axis. Furthermore, by equation 2 we have that

$$\operatorname{Im}\left(hf^{-1}(z_1)\right) = \frac{\tilde{y}}{|\sin(\theta)\tilde{z} + \cos(\theta)|^2}.$$

Taking the derivative with respect to θ, we find that the extrema lie at

$$\tan(2\theta) = \frac{2\tilde{x}}{|\tilde{z}|^2 - 1} \text{ or } 2\theta = \frac{\pi}{2} + \pi\mathbb{Z} \text{ if } |\tilde{z}| = 1.$$

This implies that for one of the two pairs of solutions, $hf^{-1}(z_1)$ is imaginary and has a maximal imaginary part. Suppose the imaginary part is smaller than i, then

$$K = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

(which belongs to $PSO_2(\mathbb{R})$) will put it above i, contradicting the maximality. So let $g = fh^{-1}$.

By Proposition 4.1 there is a unique geodesic path $\phi(s)$ from i to $g^{-1}(z_1)$, so there exists a unique geodesic path from z_0 to z_1 given by $\psi(s) = g(\phi(s))$: note that the new path is still of unit speed by Lemma 3.5. Moreover, g is unique. Suppose that there is another g_1 such that $g_1(\phi(s))$ is a geodesic path from z_0 to z_1. Then $g_1^{-1}g(\phi(s))$ is a geodesic path (of unit speed) from i to $g^{-1}(z_1)$, too, so in particular $\left(D(g_1^{-1}g)\right)(i,i) = (i,i)$. As $\mathrm{T}^1\mathbb{H} \cong PSL_2(\mathbb{R})$, it follows that $g_1^{-1}g = I_2$, so g is unique. □

4.1 Straightedge and compass construction of geodesics

The congruence relation on $PSL_2(\mathbb{R})$ allowed us to spell out the existence and uniqueness of geodesics between any two points in the hyperbolic half-plane. These geodesics can also be constructed using a straightedge and compass: they are straight lines in the vertical direction and circular sections, where the center of the circle lies on the real axis, as we will show.

We first claim that any Möbius transformation can be written as a product of elements from

$$\mathcal{U} = \left\{ \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix}, b \in \mathbb{R} \right\} \text{ and } K = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Note that $\mathcal{U}$ is a subgroup of $PSL_2(\mathbb{R})$, so we may take inverses in there. Any $\gamma \in SL_2(\mathbb{R})$ may be decomposed in an LU factorization, where L and U are two 2×2 matrices, lower and upper diagonal, respectively. Moreover, L only has 1 on the diagonal, meaning that both $L, U \in SL_2(\mathbb{R})$. Note that K can be used on any upper triangular matrix as follows:

$$KUK^{-1} = K \begin{bmatrix} x & y \\ 0 & z \end{bmatrix} K^{-1} = \begin{bmatrix} z & 0 \\ -y & x \end{bmatrix}.$$

This means in particular that $L = KU_1K^{-1}$ for some $U_1 \in \mathcal{U}$. Note that for any upper triangular matrix $\gamma' \in SL_2(\mathbb{R})$,

$$\gamma' = \begin{bmatrix} a & b \\ 0 & d \end{bmatrix}$$

we have that $d = \frac{1}{a}$, and we can decompose

$$\begin{bmatrix} 1 & -\frac{b}{d} + \frac{1}{d} \\ 0 & 1 \end{bmatrix} \gamma = \begin{bmatrix} a & 1 \\ 0 & d \end{bmatrix} = \begin{bmatrix} a & 1 \\ 0 & \frac{1}{a} \end{bmatrix}.$$

A final observation is that

$$\begin{bmatrix} a & 1 \\ 0 & \frac{1}{a} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{1}{a} & 1 \end{bmatrix} \begin{bmatrix} a & 1 \\ 0 & \frac{1}{a} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{1}{a} & 1 \end{bmatrix} \begin{bmatrix} 1 & -a \\ 0 & 1 \end{bmatrix} K^{-1}.$$

Hence, given an element of $SL_2(\mathbb{R})$ we may decompose it as $\gamma = LU$, where $L = KU_1K^{-1}$, and we may factor $U = U_2M$ so that M is of the form

$$\begin{bmatrix} a & 1 \\ 0 & \frac{1}{a} \end{bmatrix}$$

which may be factored again as $M = KU_3K^{-1}U_4K^{-1}$. So in conclusion

$$\gamma = KU_1K^{-1}U_2KU_3K^{-1}U_4K^{-1} \tag{3}$$

with $U_1, U_2, U_2, U_4 \in \mathcal{U}$.

To show that geodesics are of the desired form, consider the transformation of the geodesic line $\{yi|\, y > 0\}$ by γ factored as above. It is clear that elements of $\mathcal{U}$ translate the line to $\{b + yi|\, y > 0,\ b_2 \in \mathbb{R}\}$, which are the vertical lines. Moreover, K and K^{-1} correspond to the transformation $z \mapsto -1/z$, which maps a vertical line onto itself (if $b = 0$) or a half circle with radius $1/(2b)$ and center $(-1/(2b), 0)$. The latter may be seen as follows: all points $z = x + yi$ on the line satisfy $z = b + yi$, $y > 0$. Hence $K(z) = \frac{b}{b^2+y^2} + \frac{y}{b^2+y^2}i$. It can be verified that $(\mathrm{Re}(z) + 1/(2b))^2 + (\mathrm{Im}(z))^2 = (2b)^{-2}$. Another transformation of an element of $\mathcal{U}$ will shift the circle along the real axis. More generally, a transformation by K will map a circle with center on the real axis to another circle on the real axis, with different radius and center (or a vertical line, if the circle happens to cross the origin). This follows from the following implication, which shows that a circle is mapped to a circle or line by K (depending on the coefficients in the last equation): let $w = -\frac{1}{z}$, then

$$(x-a)^2 + (y-b)^2 - r^2 = |z|^2 + \alpha(z + \bar{z}) + \beta(z - \bar{z}) + \gamma = 0 \implies$$

$$\frac{1}{|w|^2} + \alpha\left(-\frac{1}{w} - \frac{1}{\bar{w}}\right) + \beta\left(-\frac{1}{w} + \frac{1}{\bar{w}}\right) + \gamma = 0 \implies$$

$$\gamma|w|^2 + \beta\,(w - \bar{w}) + \alpha(-w - \bar{w}) + 1 = 0. \tag{4}$$

It is practical to keep track of where the intersection points p_1, p_2 of the half-circle with the real axis lie, as the new circle will have its intersections at $-\frac{1}{p_1}, -\frac{1}{p_2}$. Applying the elements in the decomposition of γ one after another, one will find that all geodesics are indeed half-circles or vertical lines.

Historically, the study of geodesics of $\mathbb{H}$ produced an example of non-Euclidean geometry [8, p. 280-281]. Euclid postulated in his book 'Elements' [9] five constructions that were possible (axiomatically) in any geometrical setting. For example, he postulated that one could draw a straight line from any point to any point, which on $\mathbb{H}$ is indeed possible: these are the geodesics we found. However, his fifth postulate remained debated for nearly two millennia. The fifth postulate stated that, given any two lines that cross a third line, such that the interior angles are on the same side and add up to less than π radiance, the two lines must intersect when extended indefinitely. In other words, the fifth postulate states that two straight lines crossing a third that are not parallel at the moment of intersection, must intersect at some point. The straight lines on $\mathbb{H}$ provide a counterexample for this: one may draw a geodesic with center at the origin and radius one in $\mathbb{H}$, and then draw any geodesic which has its starting point at the origin, and its end point at ∞ or at some point p on the real axis with $|p| > 1$. Almost all of these geodesics will be mutually not parallel, while they never intersect.

5 Geodesic flow on the hyperbolic half-plane

One can study a dynamical system on the hyperbolic half-plane induced by *geodesic flow*, which is roughly the trajectory a point particle on $\mathbb{H}$ (given an initial velocity) would follow. To study this flow, we will exploit the congruence relations established before. Special attention will also be given to the asymptotic behaviour of the flow. We will see that a particular structure of so-called *stable and unstable manifolds* will arise. Geodesic flow is of interest for a large class of problems, an example is the field of *variational calculus in the large* [2].

In the previous section, we showed that for each pair of points z_0, z_1 in the hyperbolic half-plane, there was a unique geodesic connecting them. Moreover, this geodesic can be realized by a unique $g \in PSL_2(\mathbb{R})$ as the graph of $\phi : [0, d(z_0, z_1)] \to \mathbb{H}$,

$$\phi(s) = g(e^s i).$$

Instead of the geodesic between two points, one can also start at a point in $z \in \mathbb{H}$ and 'travel' at constant speed over $\mathbb{H}$ in a direction, given by a unit vector v, over a path of minimal distance. It is then clear that any such path $\psi(s)$ must be a geodesic, by definition. Moreover, by Lemma 3.5 there is a unique $g \in PSL_2(\mathbb{R})$ such that $Dg(i,i) = (z,v)$, which determines that $\psi(s) = g\left(e^s i\right)$.

Traveling at unit speed on a path of minimal distance over any smooth manifold, given a point and a direction, is called geodesic flow. We will denote the geodesic flow over $\mathbb{H}$ by $g_t : \mathrm{T}^1\mathbb{H} \to \mathrm{T}^1\mathbb{H}$, where it is implied that $g_t(z,v)$ is the unique unit speed parameterization of the geodesic spawning from (z,v) at time t. Note that we have gone back to using t instead of s as the parameter, but that the parameterization is still by arclength.

First, we show geodesic flow on $\mathbb{H}$ for a particular point and direction, after which we will characterize geodesic flow from any pair (z,v).

Example 5.1. Perhaps the simplest geodesic flow on $\mathbb{H}$ is the one starting from i in the direction of i. After all, we chose it as the base point for our congruence relations. The unique element of $PSL_2(\mathbb{R})$ for which $Dg(i,i) = (i,i)$ is of course the identity element. It has already been shown that the position in $\mathbb{H}$ of the geodesic flow is given by $e^t i$, from which it follows that $D\left(e^t i\right) = \left(e^t i, e^t i\right)$. Hence, the tangent vector adjoined to the position is given by $v(t) = e^t i$. This shows that the geodesic flow is explicitly given by $g_t(i,i) = \left(e^t i, e^t i\right)$. It may be checked that, for all $t \geq 0$, indeed $g_t(z,v) \in \mathrm{T}^1\mathbb{H}$. Using the derivative action of $PSL_2(\mathbb{R})$ on $\mathrm{T}^1\mathbb{H}$, it follows that the derivative action of the matrix

$$\begin{bmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{bmatrix} \in PSL_2(\mathbb{R})$$

on (i, i) gives

$$D\begin{bmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{bmatrix}(i, i) = \left(e^t i, e^t i\right).$$

So we have two ways of explicitly describing the geodesic flow: one using the derivative acting on the path, the other using the derivative action of $PSL_2(\mathbb{R})$ on the starting point (i, i). △

Our congruence relation on $PSL_2(\mathbb{R})$, together with the example above, allows us to express the geodesic flow from any pair (z, v) explicitly in terms of the derivative action acting on the starting point (i, i): given the matrix

$$a_t = \begin{bmatrix} e^{-t/2} & 0 \\ 0 & e^{t/2} \end{bmatrix}$$

we can express $g_t(i, i) = \left(Da_t^{-1}\right)(i, i)$. Any other geodesic flow is associated to a path ϕ in $\mathbb{H}$, made by a unique transformation g in $PSL_2(\mathbb{R})$ of $e^t i$, so the geodesic flow in $\mathrm{T}^1\mathbb{H}$ is given by

$$D\left(g\left(e^t i\right)\right) = \left(g\left(e^t i\right), g'\left(e^t i\right) e^t i\right) = (Dg)\left(e^t i, e^t i\right) = (Dg)(Da_t^{-1})(i, i) = \left(Dga_t^{-1}\right)(i, i).$$

So indeed, $g_t(z, v) = \left(Dga_t^{-1}\right)(i, i)$. Geodesic flow on $\mathrm{T}^1\mathbb{H}$ is hence fully described by the derivative action of $PSL_2(\mathbb{R})$, and we can identify two maps:

$$g_t : \mathrm{T}^1\mathbb{H} \to \mathrm{T}^1\mathbb{H}, (z, v) \mapsto g_t(z, v) \quad R_{a_t} : PSL^2(\mathbb{R}) \to PSL^2(\mathbb{R}), g \mapsto ga_t^{-1}.$$

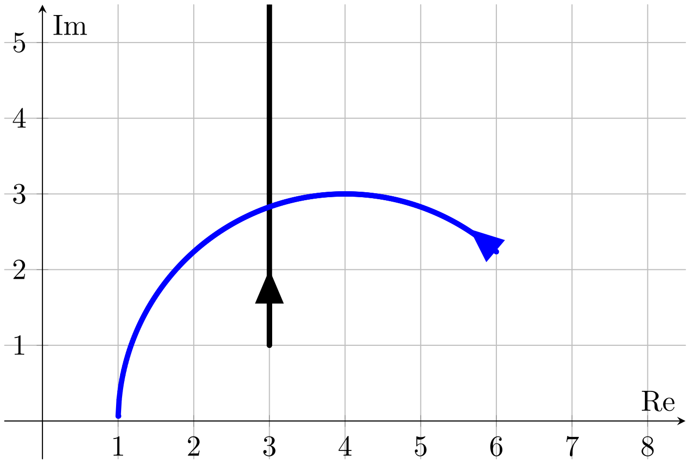

Figure 4: two geodesic flows in $\mathbb{H}$

5.1 The horocycle

Now that we explicitly understand the geodesic flow, both in terms of a transformation on $\mathrm{T}^1\mathbb{H}$ and one on $PSL_2(\mathbb{R})$, we can investigate the asymptotic behaviour of the flow. Following the flow of any pair (z, v) will send us off to infinity in the imaginary direction, or land us on the real axis after infinite time (depending on the direction of v). In that sense, the asymptotic dynamics of the geodesic flow are uninteresting. However, a more interesting question is: for which two pairs $(z_0, v_0), (z_1, v_1)$ do the geodesic flows converge towards one another? I.e. if π is the projection $\pi(z, v) = z$, for which pairs does it hold that

$$d(\pi \circ g_t(z_0, v_0), \pi \circ g_t(z_1, v_1)) \longrightarrow 0?$$

We will again study the case for (i, i) and see which pairs converge to its geodesic flow, and generalize it to all pairs by our congruence relation on $\mathrm{T}^1\mathbb{H}$.

Proposition 5.2. The geodesic flow of the pair (z, v) converges to that of (i, i) if and only if (z, v) is of the form $(i + s, i)$ with $s \in \mathbb{R}$.

Proof. If $(z, v) = (i + s, i)$, then we have that

$$h = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix}$$

is the unique element in $PSL_2(\mathbb{R})$ such that $Dh(i, i) = (i + s, i)$, and so $g_t(z, v) = Dha_t^{-1}(i, i) = \left(e^t i + s, e^t i\right)$. We can find an upper bound for $d\left(e^t i, e^t i + s\right)$ by considering the straight line (in the Euclidean sense) from one to the other: let $\phi : [0, 1] \to \mathbb{H}, \phi(x) = e^t i + sx$ be the straight path connecting the two points in $\mathbb{H}$, then $d\left(e^t i, e^t i + s\right) \leq L(\phi) = s/e^t \longrightarrow 0$ as $t \longrightarrow \infty$.

Conversely, suppose that $g_t(z, v)$ converges to $g_t(i, i)$ as above. Then v is a unit vector (by definition of geodesic flow) and must be in the direction of i, otherwise we have for large enough t that $\mathrm{Im}(\pi(g_t(z, v))) < 1/2$, where we use our knowledge of geodesics. On the other hand, $\mathrm{Im}(\pi(g_t(i, i))) > 1$ for all $t > 0$. This shows that they cannot converge. Suppose that $z = ai + s$ for some $s \in \mathbb{R}$ and $a > 0$. We claim that in this case $g_t(z, v) = g_{t+t_0}(i + s, i)$ for some $t_0 \in \mathbb{R}$. To see this, note that the graph of the geodesic flow is a vertical line with real coordinate s, and so for some t_0 we have $g_{t_0}(i+s, i) = (ai+s, v)$. Moreover, a computation shows that $a_t \cdot a_{t_0} = a_{t+t_0}$, from which it can be deduced that $g_t(g_{t_0}(i+s, i)) = g_{t+t_0}(i+s, i)$. It follows that the imaginary part of $\pi(g_t(z, v))$ is e^{t+t_0} while that of $\pi(g_t(i, i))$ is e^t. Using Lemma 2.2, the length of the geodesic connecting $g_t(z, v)$ to $g_t(i, i)$ will only converge to zero if the imaginary parts converge to one another. Hence, $t_0 = 0$ and indeed $z = i + s$ for some $s \in \mathbb{R}$ and v must then be i. A similar reasoning holds if $a < 0$. □

The proposition shows that the line $i + s$, $s \in \mathbb{R}$, in $\mathbb{H}$ has the property that the geodesic flow starting from this line in the imaginary direction asymptotically converges to that starting from (i, i), and there are no other pairs (z, v) with this property. This line is also called the *stable (sub)manifold* of the pair (i, i). Formally, a stable manifold would be defined as

$$W^s(z', v') = \{(z, v) \in \mathrm{T}^1\mathbb{H} \mid d(g_t(z, v), g_t(z', v')) \longrightarrow 0 \text{ as } t \longrightarrow \infty\}.$$

Note that the definition is in terms of a set. It is in fact a theorem [13, Proposition 6.1] that W^s is a manifold on $\mathbb{H}$. At least for the above example it is clear that this is true. Formally, we are not in a position to describe the stable manifold: we do not have a metric on $\mathrm{T}^1\mathbb{H}$, only on $\mathbb{H}$. For now we will gloss over this, but we will come back to this in the next chapter. The proper stable manifold will be the same.

We described geodesic flow in terms of a map on $PSL_2(\mathbb{R})$, $R_{a_t} : g \mapsto ga_t^{-1}$. Similarly, if we let

$$h_s = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix}$$

we can define the *horocylce flow*: $(i - s, i) = \left(Dh_s^{-1}\right)(i, i) =$

$$D \begin{bmatrix} 1 & -s \\ 0 & 1 \end{bmatrix} (i, i).$$

The equivalent description in terms of a map on $PSL_2(\mathbb{R})$ is $R_{h_s} : g \mapsto gh_s^{-1}$. Although we have not shown it yet, the pairs in $\mathrm{T}^1\mathbb{H}$ that converge to the geodesic flow of any fixed pair $(z, v) = Dg(i, i)$ are given by $\left(Dgh_s^{-1}\right)(i, i)$.

Proposition 5.3. The pairs in $\mathrm{T}^1\mathbb{H}$ that converge to the geodesic flow of any fixed pair $(z, v) = Dg(i, i)$ is given by the horocycle flow.

Proof. The goal is to show that for any two pairs $(z_0, v_0) = Dg_0(i, i)$, $(z_1, v_1) = Dg_1(i, i)$, that

$$d(\pi \circ g_t(z_0, v_0), \pi \circ g_t(z_1, v_1)) \longrightarrow 0$$

if and only if $(z_1, v_1) = Dg_0h_s^{-1}(i, i)$ for some $s \in \mathbb{R}$. Since Möbius transformations are isometries of $\mathbb{H}$, this is true if and only if the geodesic flow of (i, i) and $Dg_0^{-1}g_1(i, i)$ converge. To see this, note that $\pi \circ g_t(z_0, v_0) = g_0a_t^{-1}(i)$ and $\pi \circ g_t(z_1, v_1) = g_1a_t^{-1}(i)$, for two unique $g_1, g_2 \in PSL_2(\mathbb{R})$. From the isometry property it follows that

$$d(\pi \circ g_t(z_0, v_0), \pi \circ g_t(z_1, v_1)) = d\left(a_t^{-1}(i), g_0^{-1}g_1a_t^{-1}(i)\right) =$$

$$d\left(\pi \circ g_t(i, i), \pi \circ g_t\left(Dg_0^{-1}g_1(i, i)\right)\right).$$

From the previous proposition, it follows that (z_1, v_1) is part of the stable manifold of the pair (z_0, v_0) if and only if

$$g_0^{-1}g_1 \in \left\{ \begin{bmatrix} 1 & -s \\ 0 & 1 \end{bmatrix} \mid s \in \mathbb{R} \right\}.$$

In other words, if and only if $(z_1, v_1) = Dg_1(i, i) = Dg_0h_s^{-1}(i, i)$. □

We can construct the image of the horocycle flow in $\mathbb{H}$ using our straightedge and compass constructions of geodesic flows. We saw that all Möbius transformations could be expressed in terms of transformations of the form

$$\mathcal{U} = \left\{ \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix}, \, b \in \mathbb{R} \right\} \text{ and } K = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

The image of the horocycle flow for a pair $(z, v) = Dg(i, i)$ is given by g acting on the line $l : i + t \in \mathbb{H}, t \in \mathbb{R}$. Decomposing g, an element of $\mathcal{U}$ would have the effect of shifting the line in the real direction (leaving the image invariant) and K would give a curve

$$c : \frac{-t}{1+t^2} + \frac{1}{1+t^2} i \quad t \in \mathbb{R}.$$

If $z = x + yi \in c$ then one can show that $x^2 + (y - 1/2)^2 = (1/2)^2$, meaning that c is a circle of radius 1/2 with center $(0, 1/2)$. Another element of $\mathcal{U}$ would shift this circle along the real axis, after which another transformation by K would map c to a new circle. To see this, recall equation (3). One may repeat the procedure as often as necessary, shifting the circle, then applying K to either obtain a new circle with some new center, or, if the center lies above the origin, to obtain a horizontal line l'.

The direction in which the tangent vector points may be derived by applying Dg to (i, i) directly. Using the decomposition of g as above, one may show that the image tangent vectors always point inside the circles, or in the imaginary direction in case we have a horizontal line such as l. If we restrict ourselves to circles, we can observe the following: note that c touches the boundary of $\mathbb{H}$ in the origin, and hence after applying a transformation of $\mathcal{U}$ and a subsequent transformation by K, the horocycle will always share a point with the real axis (unless the image is l').

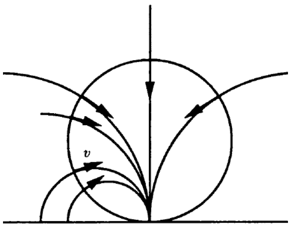

Figure 5: The horocycle flow. Figure 2 from [4]

In conclusion, the horocycle flow on $\mathrm{T}^1\mathbb{H}$ yields a straight line or a circle in $\mathbb{H}$ with tangent vectors pointing in the imaginary direction or pointing inside the circle, respectively. If we have a circle, it touches the real axis in some point. The geodesic flow starting from all points on the circle converge to some point on the real axis (recalling our discussion of geodesic flow), which must then be the point where the circle touches the real axis. This is also depicted in figure 5.

6 The modular surface

In the previous chapters, we made good use of the congruence relation of $PSL_2(\mathbb{R})$ on $\mathrm{T}^1\mathbb{H}$. Proposition (5.3) showed that the horocylce flow of a pair (z, v) gave us exactly all pairs whose geodesic flow converged to that of (z, v). Our notion of convergence merely relied on the convergence of base points in $\mathbb{H}$, but the proof of Proposition 5.2 showed that the tangent vectors must also align if the flows converged. Moreover, the proof of Proposition 5.3 showed implicitly that convergent flows had tangent vectors that had to align in the limit. Notice that this was done by showing that the associated paths in $PSL_2(\mathbb{R})$ converged. We were also able to describe geodesic flow and horocycle flow as paths in $PSL_2(\mathbb{R})$. To study properties of a quotient of the unit tangent bundle, which in general will be dependent on both the base point and tangent vector, we need to give $\mathrm{T}^1\mathbb{H}$ a topology. Motivated by the above, this will be done via $PSL_2(\mathbb{R})$. We will first develop a topology on $PSL_2(\mathbb{R})$, after which we will come back to convergence of geodesic flows. Then, we will develop a notion of area and volume on the unit tangent bundle, after which we move on to the quotient space.

6.1 Distance on the unit tangent bundle

Let $\mathbb{R}^d$ be the d dimensional vector space over $\mathbb{R}$, as usual. If we choose a basis, the space of all mappings from $\mathbb{R}^d$ to itself is $\mathrm{Mat}_{dd}(\mathbb{R})$, the set of all $d \times d$ matrices with real entries. We can give $\mathrm{Mat}_{dd}(\mathbb{R})$ a topology by considering the entries of matrices as the d^2 coordinates of the vector space $\mathbb{R}^{d^2}$ over $\mathbb{R}$, i.e. addition and scalar multiplication work the same in both places, so we can give $\mathrm{Mat}_{dd}(\mathbb{R})$ a topology as though it were $\mathbb{R}^{d^2}$. Here, we will endow $\mathbb{R}^{d^2}$ with the Euclidean topology. The space of all invertible linear transformations of $\mathbb{R}^d$ is called the general linear group of dimension d, denoted $GL_d(\mathbb{R})$. From linear algebra it is known that $GL_d(\mathbb{R})$ consists of all $d \times d$ matrices with nonzero determinant. Using that the determinant is a group homomorphism from $GL_d(\mathbb{R})$ to $\mathbb{R}$, it is easy to show that $GL_d(\mathbb{R})$ is a group. Moreover, it is an open subset of $\mathrm{Mat}_{dd}(\mathbb{R})$: the determinant is a polynomial function of the entries of a matrix, and hence is continuous, so $GL_d(\mathbb{R}) = \det^{-1}(\mathbb{R} \setminus \{0\})$ which is indeed open. We endow $GL_d(\mathbb{R})$ with a subspace topology, and define closed linear groups:

Definition 6.1. A set G is a called a closed linear group if it is an abstract group that can be embedded in $GL_d(\mathbb{R})$ for some $d \in \mathbb{N}$, and the image of the embedding is a closed subset of $GL_d(\mathbb{R})$.

Example 6.2. $SL_2(\mathbb{R})$ is a closed linear group: $SL_2(\mathbb{R})$ is a subgroup of $GL_2(\mathbb{R})$ and hence is a group that is embedded in $GL_2(\mathbb{R})$. Moreover, $SL_2(\mathbb{R}) = \det^{-1}(\{1\})$, and so is closed in $GL_2(\mathbb{R})$. △

It is a bit harder to show, but $PSL_2(\mathbb{R})$ is also a closed linear group, which we won't discuss here [8, p. 284-285]. Our interest in closed linear groups comes from the following proposition:

Proposition 6.3 (Einsiedler and Ward, Corollary 9.11 and Lemma 9.12)**.** For any closed linear group G, one can define a metric $d_G(g_0, g_1)$ that is left invariant under G:

$$d_G(hg_0, hg_1) = d_G(g_0, g_1) \quad \forall h \in G$$

Moreover, the topology induced by this metric is the same as the subspace topology on G, inherited from $GL_d(\mathbb{R})$.

Much like the hyperbolic metric on $\mathbb{H}$, if we let G be $PSL_2(\mathbb{R})$, then d_G is left invariant under Möbius transformations. Now that we have a topology on $PSL_2(\mathbb{R})$, we use the simply transitive (derivative) action of $PSL_2(\mathbb{R})$ to induce a topology on $\mathrm{T}^1\mathbb{H}$: let $(z_0, v_0) = Dg_0(i, i)$ and $(z_1, v_1) = Dg_1(i, i)$ be two pairs in $\mathrm{T}^1\mathbb{H}$, then we define a metric $d_{\mathrm{T}^1\mathbb{H}}((z_0, v_0), (z_1, v_1)) = d_G(g_0, g_1)$.

Remark. From now on, we will occasionally write G for $PSL_2(\mathbb{R})$, for convenience of notation.

We can now return to studying $\mathrm{T}^1\mathbb{H}$. First, we give an extended definition of convergence of geodesic flows, after which we will give the companion of proposition 5.3 for this definition.

Definition 6.4. Let (z_0, v_0), (z_1, v_1) be two pairs in $\mathrm{T}^1\mathbb{H}$, then the geodesic flow of both pairs converge to one another if $d_G(g_t(z_0, v_0), g_t(z_1, v_1)) \longrightarrow 0$ as $t \longrightarrow \infty$.

Proposition 6.5. The stable manifold (as in equation 5.1) of any fixed pair $(z, v) = Dg(i, i)$ is given by the horocycle flow.

Proof. The goal is to show that for any two pairs $(z_0, v_0) = Dg_0(i, i)$, $(z_1, v_1) = Dg_1(i, i)$,

$$d_{\mathrm{T}^1\mathbb{H}}(g_t(z_0, v_0), g_t(z_1, v_1)) = d_G\left(g_0 a_t^{-1}, g_1 a_t^{-1}\right) \longrightarrow 0$$

if and only if $(z_1, v_1) = Dg_0 h_s^{-1}(i, i)$ for some $s \in \mathbb{R}$. The metric on G is left invariant, so $d_G\left(g_0 a_t^{-1}, g_1 a_t^{-1}\right) = d_G\left(I_2, a_t g_0^{-1} g_1 a_t^{-1}\right)$. Let us denote $g_0^{-1} g_1$ by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Then it is clear that

$$d_G\left(I_2, a_t g_0^{-1} g_1 a_t^{-1}\right) \longrightarrow 0 \iff \begin{bmatrix} a & e^{-t}b \\ e^t c & d \end{bmatrix} \longrightarrow I_2$$

which is true if and only if

$$g_0^{-1}g_1 \in \left\{ \begin{bmatrix} 1 & -s \\ 0 & 1 \end{bmatrix} \mid s \in \mathbb{R} \right\}.$$

In other words, if and only if $(z_1, v_1) = Dg_1(i,i) = Dg_0 h_s^{-1}(i,i)$. □

There is also a notion of an unstable manifold. Suppose one picks a pair (z, v) in the unit tangent bundle, one can 'reverse time' by following the geodesic flow of $(z, -v)$. The question is then, which pairs started out arbitrarily close to the geodesic flow of $(z, -v)$? Cast into a definition, the unstable manifold of a pair (z, v) is given by

$$W^u(z', v') = \{(z, v) \in \mathrm{T}^1\mathbb{H} \mid d(g_t(z, v), g_t(z', v')) \longrightarrow 0 \text{ as } t \longrightarrow -\infty\}.$$

Similar to the stable manifold, W^u is in fact a manifold [13, Proposition 6.1]. Let

$$u_s = \begin{bmatrix} 1 & 0 \\ s & 1 \end{bmatrix},$$

and let us define the unstable horocycle flow on $\mathbb{H}$ to be $u_s : (z, v) = Dg(i, i) \mapsto Dgu_s^{-1}(i,i)$. The corresponding path in $PSL_2(\mathbb{R})$ is given by $R_{u_s} : g \mapsto gu_s^{-1}$

Proposition 6.6. The unstable stable manifold of any fixed pair $(z, v) = Dg(i, i)$ is given by the unstable horocycle flow.

It can be verified that $a_{-t} = a_t^{-1}$, and that $u_s = h_s^{-1}$. The proof of the proposition is then almost a copy of the previous one.

6.2 Area and volume

There are some further properties of $\mathrm{T}^1\mathbb{H}$ that need to be discussed before we move to studying a quotient of the unit tangent bundle. Now that we have a notion of distance on $\mathbb{H}$ and $\mathrm{T}^1\mathbb{H}$, d and d_G respectively, the notion of area and volume can be introduced. Here, we will give $\mathbb{H}$ the coordinates x, y and the unit tangent vectors will be parameterized by the angle θ. The hyperbolic area form on $\mathbb{H}$ is defined as $dA = \frac{1}{y^2}dx \wedge dy$, and the hyperbolic volume form on $\mathrm{T}^1\mathbb{H}$ is defined as $dm = \frac{1}{y^2}dx \wedge dy \wedge d\theta$.

Lemma 6.7. The hyperbolic area and volume form are invariant under the action of $PSL_2(\mathbb{R})$.

Proof. We first prove the invariance of the area form. The action of $PSL_2(\mathbb{R})$ is simply that of the Möbius transformation on $\mathbb{H}$, so we may view it as a change of coordinates. From the definition of a group action, it is easily deduced that $g \cdot x = g \cdot y \implies x = y$, so the transformation is injective. Moreover, it is surjective by Lemma 3.3, so it is a bijection. It is clear that a transformation g is continuously differentiable, and its inverse will be too. Recall that for a Möbius transformation $g : z \mapsto \frac{az+b}{cz+d}$, the imaginary part was $\mathrm{Im}(z)/|cz+d|^2$. Let J denote the Jacobian matrix, then

$$J = \begin{bmatrix} \frac{\partial \operatorname{Re}(g)}{\partial x} & \frac{\partial \operatorname{Re}(g)}{\partial y} \\ \frac{\partial \operatorname{Im}(g)}{\partial x} & \frac{\partial \operatorname{Im}(g)}{\partial y} \end{bmatrix} = \begin{bmatrix} \frac{\partial \operatorname{Re}(g)}{\partial x} & \frac{\partial \operatorname{Im}(g)}{\partial y} \\ -\frac{\partial \operatorname{Im}(g)}{\partial y} & \frac{\partial \operatorname{Re}(g)}{\partial x} \end{bmatrix}$$

using the Cauchy-Riemann equations. Hence,

$$g^*(dA) = \frac{1}{\operatorname{Im}(g(z))^2} d(\operatorname{Re}(g(z))) \wedge d(\operatorname{Im}(g(z))) = \frac{1}{y^2}|cz+d|^4 \cdot |\det(J)| \cdot dx \wedge dy = dA.$$

where g^* denotes the pull-back of g. The invariance of the volume form follows the same reasoning. By Lemma 3.5, the derivative action of $PSL_2(\mathbb{R})$ is surjective, and by the same reasoning as above, injective. It is again clear that it is continuously differentiable. Let g again denote an element of $PSL_2(\mathbb{R})$, then $Dg(z,v) = \left(\frac{az+b}{cz+d}, \frac{1}{(cz+d)^2}v\right)$. So, in the (x, y, θ) coordinates, it is clear that the first and second coordinate of $g(z)$ do not depend on θ, while the third is linear in θ. Hence, the Jacobian matrix takes on the form

$$\begin{bmatrix} \frac{\partial \operatorname{Re}(g)}{\partial x} & \frac{\partial \operatorname{Im}(g)}{\partial y} & * \\ -\frac{\partial \operatorname{Im}(g)}{\partial y} & \frac{\partial \operatorname{Re}(g)}{\partial x} & * \\ 0 & 0 & 1 \end{bmatrix}.$$

Hence,

$$Dg^*(dm) = \frac{1}{\operatorname{Im}(g(z))^2} d(\operatorname{Re}(g(z))) \wedge d(\operatorname{Im}(g(z))) \wedge d\left(\frac{1}{(cz+d)^2}v\right) = \frac{1}{y^2} dx \wedge dy \wedge d\theta = dm.$$

□

6.3 The quotient

Now, we are in a position to study quotients of the hyperbolic half-plane: we will focus our attention on a specific example, the modular surface. We return our attention to the analogy of the torus $\mathbb{T}^2$. The real plane was tiled by points in $\mathbb{Z}^2$, i.e. we drew vertical and horizontal lines between points with integer coordinates. Points of $\mathbb{R}^2$ in different tiles were identified by an equivalence relation:

$$(x_0, y_0) \sim (x_1, y_1) \iff (x_0 - x_1, y_0 - y_1) \in \mathbb{Z}^2.$$

The isometries in this analogy are translations by a vector g:

$$g = g_{a,b} : \mathbb{R}^2 \to \mathbb{R}^2, (x, y) \mapsto (x + a, y + b).$$

It is not hard to show that they are isometries, that they form a group under addition, and that, given a base point, say the origin, they are in fact simply transitive on $\mathbb{R}^2$. Compare this to the derivative action of $PSL_2(\mathbb{R})/PSO_2(\mathbb{R})$ on the base point i. With respect to these translations, one can also redefine the equivalence relation: denote by $T\left(\mathbb{Z}^2\right)$ the translations by a vector with integer components, then

$$(x_0, y_0) \sim (x_1, y_1) \iff (x_1, y_1) = g(x_0, y_0) \text{ for some } g \in T\left(\mathbb{Z}^2\right).$$

I.e. the equivalence class of (x_0, y_0) can be realized as $T\left(\mathbb{Z}^2\right)(x_0, y_0)$. In yet other words, if we pick a representative from the unit square $[0,1] \times [0,1]$, its coset is the image of the unit square under translations from $T\left(\mathbb{Z}^2\right)$. The analogy lacks something, namely the tangent bundle on the torus. One could extend the analogy to envelop this, but this is maybe too exhaustive for our purposes.

Returning to the hyperbolic half-plane, we define the 'integer translations' to be $PSL_2(\mathbb{Z})$, the subgroup of $PSL_2(\mathbb{R})$ consisting of all elements with integer entries. This subgroup is also called the modular group. We consider the right cosets of $PSL_2(\mathbb{R})$: for $g \in PSL_2(\mathbb{R})$ the coset is defined as $PSL_2(\mathbb{Z})g$. Letting these cosets act on the chosen base point of $T^1\mathbb{H}$, namely (i, i), we have that the coset is given by

$$D(PSL_2(\mathbb{Z})g)(i, i) = \{D(hg)(i, i) \mid h \in PSL_2(\mathbb{Z})\}.$$

The modular surface X is then defined as the collection of all cosets of $PSL_2(\mathbb{R})$:

$$X = PSL_2(\mathbb{Z}) \setminus PSL_2(\mathbb{R}).$$

Note that the notation of a right quotient is easily confused with that of a complement of sets.

Example 6.8. The coset $PSL_2(\mathbb{Z})I$ is simply $PSL_2(\mathbb{Z})$, and letting it act on (i, i) we find that

$$Dg(i, i) = \left(\frac{ai + b}{ci + d}, \frac{1}{(ci + d)^2} i\right)$$

for integers coefficients a, b, c, d with $ad - bc = 1$. Choosing $b = n$ to be any integer and $c = 0$, we find that necessarily $a = d = 1$, and that the resulting base points in $\mathbb{H}$ are $i + n$. Moreover, using equation 2, any base point has imaginary part of at most i. △

It would be nice if we could think of the modular surface as we did of the torus: pick the representative of each coset to lie in the unit square and study the square instead of the abstract cosets. For the modular surface, we will let it act on (i, i) and pick a triangle in $\mathbb{H}$ which has (almost) exactly one representative of each coset in it.

6.4 The fundamental domain

When constructing quotient spaces X of a space G over a subset Γ in a general setting, it is often convenient to pick one representative from the coset $x = \Gamma g$ in G to represent the class. This can be formalized as follows.

Definition 6.9. A fundamental domain F for a quotient $X = \Gamma \setminus G$ is a measurable subset of G with the property that for almost every $g \in G$ it holds that $|F \cap \Gamma g| = 1$. If it holds for every $g \in G$, we say that F is a strict fundamental domain.

Example 6.10. We have seen that the unit square is morally a fundamental domain for the quotient $T(\mathbb{Z}^2) \setminus \mathbb{R}^2$, but it has some double representatives: for the coset $T(\mathbb{Z}^2)(0, 1/2)$ both $(0, 1/2)$ and $(1, 1/2)$ belong to the unit square. We worked around this by identifying opposite sides of the unit square. A strict fundamental domain would be $F = [0, 1) \times [0, 1)$. △

Before defining the fundamental domain for modular surface, we review some work generalizing the modular surface.

Quotient spaces of the hyperbolic half-plane have been considered more generally. For a notion of distance on the quotient $X = \Gamma \setminus PSL_2(\mathbb{R})$, one can consider two classes $x_0 = \Gamma g_0, x_1 = \Gamma g_1$, and define

$$d_X(x_0, x_1) := \inf_{\gamma_0, \gamma_1} d_G(\gamma_0 g_0, \gamma_1 g_1) = \inf_{\gamma} d_G(g_0, \gamma g_1).$$

This notion of distance is evidently independent of the choice of representatives of x_0, x_1. Note that the second equality follows from the left invariance of d_G. It is not clear that d_X is always a metric. However, if Γ is a so called *discrete subgroup* of $PSL_2(\mathbb{R})$, this can be guaranteed [8, p.296-297].

It turns out that $PSL_2(\mathbb{Z})$ is one such subgroup, much like the integer translations in the real plane form a discrete subgroup of all translations. Formally, a subset of H of a metric space (G, d_G) is discrete if, for each $x \in H$ there exists $\epsilon > 0$ such that the ball of radius ϵ around x only intersects H in $\{x\}$, i.e. $B_\epsilon(x) \cap H = \{x\}$. We will not go into the details of showing that $PSL_2(\mathbb{Z})$ is discrete, since this would involve working out the details of the metric d_G, which is superfluous for our purposes. The study of more general quotients has given rise to some terminology:

Definition 6.11. A Fuchsian group is a discrete subgroup Γ of $PSL_2(\mathbb{R})$. A lattice in $PSL_2(\mathbb{R})$ is a Fuchsian group, such that a fundamental domain for the quotient space $\Gamma \setminus PSL_2(\mathbb{R})$ has finite hyperbolic measure. A lattice is called a uniform lattice if the quotient space is compact.

We end the generalization here, and find a lattice for $\Gamma = PSL_2(\mathbb{Z})$.

Proposition 6.12. The set $E = \{z \in \mathbb{H} \mid |z| \geq 1, |\operatorname{Re}(z)| \leq 1/2\}$ is a fundamental domain for the quotient $PSL_2(\mathbb{Z}) \setminus PSL_2(\mathbb{R})$, in the following sense:

$$A(\gamma E \cap E) = 0 \quad \forall \gamma \in PSL_2(\mathbb{Z}) \setminus I_2,$$
$$\text{and } \mathbb{H} = \bigcup_{\gamma \in PSL_2(\mathbb{R})} \gamma E.$$

Moreover, $PSL_2(\mathbb{Z})$ is a lattice in $PSL_2(\mathbb{R})$.

Proof. Let z be an element of $\mathbb{H}$. The goal is to show that there exists a g in $PSL_2(\mathbb{Z})$ such that $gz = w$ for some $w \in E$. This will show the second claim. Denote

$$\gamma = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad ad - bc = 1, \quad a, b, c, d \in \mathbb{Z}.$$

Recall from equation 2 that $\operatorname{Im}(\gamma z) = \operatorname{Im}(z)/|cz + d|^2$. Writing out the norm in the denominator, one can find that $\operatorname{Im}(\gamma z)$ is strictly, monotonically decreasing in c, d. Hence, either for $(c, d) = \pm(1, 0)$ or $\pm(c, d) = (0, 1)$ the imaginary part is maximal. One may pick any $a, b \in \mathbb{Z}$ such that $\gamma \in PSL_2(\mathbb{Z})$. We turn our attention to two elements of $PSL_2(\mathbb{Z})$,

$$\tau = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad \sigma = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}. \tag{5}$$

Let $\gamma z = x + yi$, then there exists an integer k such that $|x - k| < 1/2$. One can verify that $\tau^{-k}\gamma z = x - k + yi$. We claim that $w = \tau^{-k}\gamma z \in E$. Evidently, $|\operatorname{Re}(w)| \leq 1/2$. Suppose that $|w| < 1$, then $\operatorname{Im}(w) < 1$, but this implies that $\operatorname{Im}(\sigma w) = \operatorname{Im}(-1/w) > 1$, contradicting the maximality of $\operatorname{Im}(\gamma z)$. Hence $w = \tau^{-k}\gamma z \in E$.

To show the first claim, let $z, w \in E$ and suppose $w = \gamma z$ for some $\gamma \in PSL_2(\mathbb{Z})$. The goal is to show that in this case, z or w lie on the boundary of E. We will prove it by exhaustion. Suppose that $\operatorname{Im}(w) < \operatorname{Im}(z)$, then we may replace the triple (w, z, γ) by (z, w, γ^{-1}), since $w = \gamma z \in E$ implies $z = \gamma^{-1}w \in E$. Hence, without loss of generality, assume that $\operatorname{Im}(w) \geq \operatorname{Im}(z)$. Equation 2 shows that $|cz+d| \leq 1$, and since $\operatorname{Im}(z) \geq \frac{1}{2}\sqrt{3}$, $|c| < 2$. Suppose that $c = 0$, then the restrictions on γ show that $a = d = \pm 1$, with b free to choose. In other words, $w = \gamma z = z + b \in E$. This can only be true if $b = 0$, in which case $\gamma = I_2$, or $b = 1$ and z is part of $\{z \in E \mid \operatorname{Re}(z) = -1/2\}$ and w is part of $\{z \in E \mid \operatorname{Re}(z) = 1/2\}$, or $b = -1$ with the sets z and w belong to swapped. Suppose that $c = 1$, then $|z + d| \leq 1$ means that either $d = 0$, or $d = \pm 1$ and $z = \frac{1}{2} \mp \frac{1}{2}\sqrt{3}i$. In the latter case, z lies on the boundary. In the former case $|z| \leq 1$ means that $|z| = 1$, so z lies on the boundary. An identical argument shows that the same holds true for $c = -1$.

Finally, to show that $PSL_2(\mathbb{Z})$ is a lattice, we need to show that $A(E)$ is finite.

$$A(E) = \int_E \frac{1}{y^2} dx \wedge dy = \int_{-1/2}^{1/2} \int_{\sqrt{1-x^2}}^{\infty} \frac{1}{y^2} dy dx =$$

$$\int_{-1/2}^{1/2} \frac{1}{\sqrt{1-x^2}} dx = \arcsin(1/2) - \arcsin(-1/2) = \frac{\pi}{3} < \infty$$

□

Although we will think of E as the fundamental domain for $PSL_2(\mathbb{Z}) \setminus PSL_2(\mathbb{R})$, it is clear that only $\tilde{F} = \{g \in PSL_2(\mathbb{R}) \mid g(i) \in E\}$ could satisfy the definition of a fundamental domain. In fact, even $\tilde{F}$ is not a fundamental domain: $PSO_2(\mathbb{R})(i) = i$, so if $g \in F$, then $PSO_2(\mathbb{R})g \subset F$. However, we have seen that $PSL_2(\mathbb{R})$ is simply transitive on $\mathbb{T}^1\mathbb{H}$, and hence using the above,

$$F = \{g \in PSL_2(\mathbb{R}) \mid Dg(i,i) \in \mathrm{T}^1\mathbb{E}\} \tag{6}$$

is a fundamental domain for $PSL_2(\mathbb{Z}) \setminus PSL_2(\mathbb{R})$.

6.5 Tesselation of Poincaré half-plane

The proof of the previous proposition suggests that finding an element g of $PSL_2(\mathbb{Z})$ such that z is mapped by g to some $w \in E$ is easily done: using τ and σ from equation 5, first find a matrix $\gamma \in PSL_2(\mathbb{R})$ with $(c,d) = \pm(1,0)$ or $(c,d) = \pm(0,1)$, then find an integer k such that $|\operatorname{Re}\left(\tau^{-k}\gamma z\right)| \leq 1/2$. Since we were free to choose a, b of γ, we can choose $\gamma \in \{\sigma, I_2\}$. In other words, any $z \in \mathbb{H}$ is mapped by some $g = \tau^n \sigma^m$ to E, where $n \in \mathbb{Z}$, $m \in \{0,1\}$. One can even show that $PSL_2(\mathbb{Z})$ is generated by τ and σ, but this is superfluous for our purposes.

The action of σ is $z \mapsto -\frac{1}{z}$ and inverts to norm of z, and mirrors the angle in the imaginary axis. The action of τ is a translation by a unit to the right. The image of the fundamental domain by repeated application of τ and σ gives a tessellation of the hyperbolic half-plane, as is depicted in figure 6.

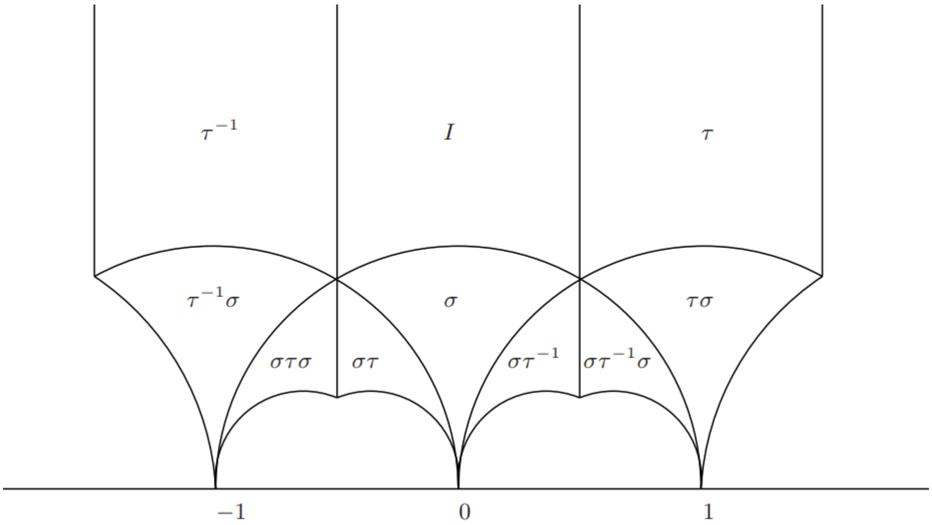

Figure 6: tessellation of $\mathbb{H}$ by E. Figure 9.6 from [8]

The boundaries of E are geodesics, and so the boundaries of all other tiles will be too. To keep track of which geodesics are mapped to which, one can use the straightedge and compass constructions of the geodesics on $\mathbb{H}$. For example, the boundary part $\{z \in \mathbb{H} \mid \operatorname{Re}(z) = -1/2, |z| \geq 1\}$ is mapped by σ to the circular section with radius 1 and center $(1, 0)$, where you start at $\sigma\left(-\frac{1}{2} + \frac{1}{2}\sqrt{3}\right)$ and end up at the origin.

7 Ergodicity of geodesic flow

In this section, we will briefly introduce some ergodic theory. We recall some definitions from the introduction. We also introduce some more background in ergodic theory. The proof of theorems that are given as background are omitted, but references are given to full proofs. To round off the chapter, we will prove ergodicity of the geodesic flow of the modular surface.

Definition 7.1. Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be probability spaces. That is, they are measure spaces with $\mu(X) = \nu(B) = 1$. A map measurable map $f : X \to Y$ is called measure preserving if

$$\mu\left(f^{-1}(B)\right) = \nu(B) \quad \forall B \in \mathcal{B}.$$

If in addition f is bijective almost everywhere (on Y), then f is called an invertible measure preserving map. If $g : X \to X$ is measure preserving, then μ is called g-invariant, and the four-tuple $(X, \mathcal{A}, \mu, g)$ is called a measure preserving system. g is then called a measure preserving transormation.

Ergodic theory often studies finite area or volume quotients of general spaces. We will be studying the quotient $PSL_2(\mathbb{Z}) \setminus PSL_2(\mathbb{R})$ of course, which was shown to be of finite volume, but the torus $\mathbb{R}^2/\mathbb{Z}^2$ is another example. An even easier example is the circle $C = \mathbb{R}/\mathbb{Z}$, which has a strict fundamental domain $[0, 1)$. Similar to the construction of the torus, one may identify the right hand side of the interval $[0, 1]$ with the left hand side, to obtain the quotient. If we use the Borel σ-algebra and measure on $\mathbb{R}$, we can give C the subspace σ-algebra to obtain the probability space $(C, \mathcal{B}_{[0,1)}, m)$. A set E is then measureable in C if $E = [0, 1) \cap B$ for some $B \in \mathcal{B}$.

Example 7.2. A circle rotation $R_\alpha : C \to C, t \mapsto (t + \alpha) \mod 1$ is an invertible measure preserving transformation. It is easy to check that R_α is surjective, and for injectivity it is sufficient to see that $R_\alpha(t_0) = R_\alpha(t_1) \implies t_0 - t_1 \equiv 0 \mod 1 \implies t_0 = t_1 \in C$. Suppose we pick an interval $A = [a, b] \subset C$. Then it is clear that $\mu\left(R_\alpha^{-1}(A)\right) = \mu\left([a - \alpha, b - \alpha]\right) = \mu([a, b]) = \mu(A)$. Of course the interval $[a - \alpha, b - \alpha]$ will in general not be a subset of $[0, 1)$, but computing modulo 1, it will still be an interval in C with same measure. To show that R_α is measure preserving, we need to verify the above for any measureable set $A \in \mathcal{B}_{[0,1)}$. There is a theorem [8, Theorem A.8] stating that it is sufficient here to only consider the intervals. △

The interest in measure preserving systems started with Poincaré. As was stated in the introduction, he proved a theorem concerning the dynamics of a measure preserving systems. It states that for any measureable set E of the system, almost every point in E is mapped back to E by the measure preserving map infinitely many times. An even stronger property of a map on a probability space is ergodicity:

Definition 7.3. A measure preserving transformation $T : X \to X$ of a probability space $(X, \mathcal{A}, \mu)$ is ergodic if for any $A \in \mathcal{A}$ it holds that

$$T^{-1}(A) = A \implies \mu(A) = 0 \text{ or } \mu(A) = 1.$$

We claimed before that ergodic maps on a probability space allowed for interchanging a time average and a spatial average. Suppose we have a function f from the probability space to $\mathbb{R}$, and we are given a starting point x in the space. If T is a measure preserving transformation and moreover ergodic, the average of f evaluated on $x_0 = x, x_1 = T(x), x_2 = T^2(x), \dots$ can be computed by a spatial average of f over the probability space. This is more precisely cast into the following theorem:

Theorem 7.4 (Birkhoff's pointwise ergodic theorem, Einsiedler and Ward, Theorem 2.30)**.** Let $(X, \mathcal{A}, \mu, T)$ be a measure preserving system. If $f \in \mathcal{L}^1_\mu$, the space of all measureable functions with the property that $\int |f| d\mu < \infty$, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f\left(T^j(x)\right) = f^*(x)$$

converges almost everywhere, and it converges in L^1_μ to a T-invariant function $f^* \in \mathcal{L}^1_\mu$. Moreover,

$$\int f^* d\mu = \int f d\mu$$

and if T is ergodic, then

$$f^*(x) = \int f d\mu \quad \text{a.e.}$$

Note that L^1_μ is a space of equivalence classes from $\mathcal{L}^1_\mu$: $f, g \in [f] \in L^1_\mu \iff \int |f - g| d\mu = 0$. Note that the theorem also shows something if T is just a measure preserving transformation, showing another good reason to study them.

Since probability spaces are quite general spaces, ergodic theory can find applications in many areas of mathematics and physics. A particularly nice example is an application in number theory.

Example 7.5. The measure preserving system of interest is $(C, \mathcal{B}_{[0,1)}, m, M_{10})$, where $M_{10} : t \mapsto (10t) \mod 1$. Using intervals again, one may show that M_{10} is indeed measure preserving. To show that M_{10} is ergodic takes some more theory, so we refer to [5]. The following result is due to Borel, which can also be found in [5].

Let x be a real number with decimal expansion $a_0, a_1a_2a_3\dots$. We say x is a *normal* number in base 10, if every block of k digits (e.g. 387, three digits) appears with asymptotic frequency $1/10^k$. We claim that almost every number in $\mathbb{R}$ is normal. The idea is as follows: define the intervals $A(j,k) = [j/10^k, (j+1)/10^k)$ for $j, k \in \mathbb{N}$. If

$x \mod 1 \in A(j,k)$ then $x = a_0, ja_{k+1}, a_{k+2}\ldots$, and if $M_{10}^n(x) \in A(j,k)$, then $x = a_0, a_1 \ldots a_n j a_{n+k+1} \ldots$. Take the measurable functions $\mathbb{1}_{A(j,k)}$, and any $x \in [0,1)$. Then the pointwise ergodic theorem tells us that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}_{A(j,k)}(M_{10}^i(x)) = \int_0^1 \mathbb{1}_{A(j,k)} dm = 10^{-k}.$$

Hence, if we let j be a block of k digits, $j = j_1 10^{k-1} + \ldots + j_{k-1}10 + j_k$, we see that the claim holds. $\triangle$

Ergodicity as given above has some equivalent conditions, one of which we will use in the following section.

Proposition 7.6 (Einsiedler and Ward, Proposition 2.14)**.** Let $(X, \mathcal{A}, \mu, T)$ be a T invariant system, then the following are equivalent:

- T is ergodic.
- For any measurable $f : X \to \mathbb{C}$ it holds that $f \circ T = f$ almost everywhere $\implies$ f is constant almost everywhere

A final loose result we will use is the following proposition on $PSL_2(\mathbb{R})$. This proposition has a much more general counterpart, which we will not go in to here.

Proposition 7.7 ([8], Proposition 8.6)**.** For any two measurable sets A_1, A_2 of $PSL_2(\mathbb{R})$ with $m_G(A_1)m_G(A_2) > 0$, the set

$$\{g \in PSL_2(\mathbb{R}) \mid m_G(A_1 g \cap A_2) > 0\}$$

is open and non-empty.

7.1 Measure, geodesic flow and horocycle flow on the modular surface

On the modular surface, we can still look at geodesic flow as right multiplication by the matrix a_t^{-1}. I.e. let $x \in PSL_2(\mathbb{Z}) \setminus PSL_2(\mathbb{R})$, then define $R_{a_t} : x \mapsto x a_t^{-1}$. Using our fundamental domain E, we may think of a pair in $\mathrm{T}^1 E$, and consider its geodesic flow as though we were in $\mathbb{H}$, until we hit the boundary of E. At that point, we apply either $D(\tau)^{\pm 1}$ if we hit the right or left boundary, or $D(\sigma)^{\pm 1}$ if we hit the lower boundary, so that the tangent vector points inside of E again. The geodesic flow is then continued from the new pair, obtained from this transformation, until we hit a boundary again.

Since multiplication on the right by an element of $PSL_2(\mathbb{R})$ defines a group action on the modular surface X, we can also look at the action of the horocycle and unstable

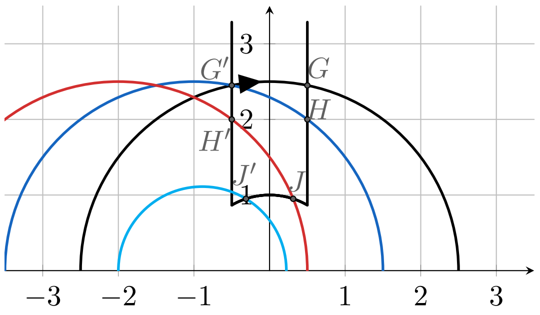

Figure 7: geodesic flow on the modular surface. Starting from G' in the direction of the arrow, we hit right right boundary in G and return to G', then we hit H and return to H', until we finally hit the lower boundary in J and are return in J'. The orientation of the geodesic is flipped when we hit the lower boundary.

horocylce flow. Recalling the defintion of the metric on X,

$$d_X(x_0, x_1) = \inf_{\gamma \in PSL_2(\mathbb{Z})} d_G(g_0, \gamma g_1),$$

it follows that

$$d_X(R_{a_t}(x_0), R_{a_t}(x_1)) \longrightarrow 0 \iff d_G(R_{a_t}(g_0), R_{a_t}(g_1)) \longrightarrow 0 \text{ for some } \gamma \in PSL_2(\mathbb{Z}).$$

From Proposition 5.1 and 6.6 it then follows that the geodesic flow of two pairs in X converges if and only if the second pair belongs to the horocycle flow of the first pair, and a similar property holds for the unstable horocycle flow.

A final ingredient for studying the geodesic flow on the modular surface, or on T^1E, however you like to think of it, is finding a measure on X. We have a measure on $PSL_2(\mathbb{R})$: Take $A \subset PSL_2(\mathbb{R})$, then we say that A is measurable if and only if $DA(i,i)$ is measurable with respect to the measure

$$\int \mathbb{1}_{DA(i,i)} dm = \int \mathbb{1}_{DA(i,i)} \frac{1}{y^2} dx \wedge dy \wedge d\theta.$$

Additionally, it is invariant under multiplication by $PSL_2(\mathbb{R})$ by Proposition 6.7. We can derive a $PSL_2(\mathbb{R})$ invariant measure for X with relative ease.

Proposition 7.8. Let π be the canonical projection map for X, $\pi : PSL_2(\mathbb{R}) \to X$, $g \mapsto x = [g]$, and let F be the fundamental domain for X, as in equation 6. Then

$$m_X(B) = m\left(F \cap \pi^{-1}(B)\right)$$

is a $PSL_2(\mathbb{R})$ invariant measure for X. Moreover, any fundamental domain F' for X has the same measure as F.

Proof. We first make a σ-algebra Σ for X. Note that the σ-algebra on $PSL_2(\mathbb{R})$ is induced by the Borel σ-algebra on $\mathrm{T}^1\mathbb{H}$. We say that $B \subset X$ is belongs to Σ if $F \cap \pi^{-1}(B)$ is measurable. It is clear that $F \cap \pi^{-1}(X) = F$ is measurable, so $X \in \Sigma$. Suppose $A \in \Sigma$, then $F \cap \pi^{-1}(A)$ is measurable, and hence $F \cap \pi^{-1}(X \setminus A) = F \cap \big(\mathbb{H} \setminus \pi^{-1}(A)\big)$ is measurable, so $X \setminus A \in \Sigma$. Finally, if $A_1, A_2, \ldots \in \Sigma$, then $F \cap \pi^{-1}(A_i)$ is measurable for each $i \in \mathbb{N}$, and hence

$$F \cap \pi^{-1}\left(\bigcup_{i \in \mathbb{N}} A_i\right) = F \cap \bigcup_{i \in \mathbb{N}} \pi^{-1}(A_i)$$

is measurable. Hence, $\bigcup_{i \in A_i} \in \Sigma$.

Second, we show that m_X is indeed a measure. By the definition of m, m_X will be non-negative and $m_X(\emptyset) = 0$. Moreover, for a countable set of pairwise disjoint, measureable sets $\{A_i \in \Sigma | i \in \mathbb{N}\}$,

$$m_X\left(\bigcup_{i \in I} A_i\right) = m\left(F \cap \pi^{-1}\left(\bigcup_{i \in I} A_i\right)\right) = m\left(F \cap \left(\bigcup_{i \in I} \pi^{-1}(A_i)\right)\right) = \sum_{i \in \mathbb{N}} m_X(A_i).$$

Third, we show that any fundamental domain F' has the same measure as F. We first prove a more general statement: let F and F' be any two measureable sets in $PSL_2(\mathbb{R})$ with the property that $\pi(F) = \pi(F')$, and $\pi|_F$ and $\pi|_{F'}$ are injective for almost every $g \in F, F'$. This includes fundamental domains. By assumption, for almost every $g \in F$ there exists a unique $\gamma \in PSL_2(\mathbb{Z})$ with $\gamma g \in F'$. Let us denote by $\tilde{F}$ the subset of F for which there does exist such a unique γ. Then $F \setminus \tilde{F}$ is a set of measure zero. One can similarly split F' in $\tilde{F'}$ and $F' \setminus \tilde{F'}$. Note that $g \in \tilde{F} \iff \gamma g \in \tilde{F'}$ for a unique γ. We may decompose $\tilde{F}$ and $\tilde{F'}$ by elements of $PSL_2(\mathbb{Z})$:

$$\tilde{F} = \bigsqcup_{\gamma \in PSL_2(\mathbb{Z})} \tilde{F} \cap \gamma \tilde{F'}$$

$$\tilde{F'} = \bigsqcup_{\gamma \in PSL_2(\mathbb{Z})} \tilde{F'} \cap \gamma \tilde{F}.$$

For a fixed γ we can deduce that

$$\gamma^{-1}\left(\tilde{F} \cap \gamma \tilde{F'}\right) = \tilde{F'} \cap \gamma^{-1} \tilde{F},$$

and moreover $PSL_2(\mathbb{Z})$ is a countable set. Finally, m is invariant under $PSL_2(\mathbb{R})$, and so in particular under $PSL_2(\mathbb{Z})$, hence

$$m(F) = m(\tilde{F}) = \sum_{\gamma \in PSL_2(\mathbb{Z})} m(\tilde{F} \cap \gamma \tilde{F'}) = \sum_{\gamma \in PSL_2(\mathbb{Z})} m(\tilde{F'} \cap \gamma \tilde{F}) = m(\tilde{F'}) = m(F').$$

In particular this holds for fundamental domains.

Finally, the invariance under $PSL_2(\mathbb{R})$ follows from the above and the invariance of m: let B be a measurable set in X, and define $C = \pi^{-1}(B) \cap F$, where F is now the fundamental domain again. Then $Cg = \pi^{-1}(B)g \cap F'$, where $F' = Fg$ is another fundamental domain. Note that $\pi^{-1}(B)g = \pi^{-1}(Bg)$ because B consists of cosets of the form $PSL_2(\mathbb{Z})h$ and Bg consists of cosets of the form $PSL_2(\mathbb{Z})hg$. By definition, $\big(F \cap \pi^{-1}(Bg)\big) \cap \big(F' \cap \pi^{-1}(Bg)\big) = \emptyset$ except on a set of measure zero, $\pi|_{F\cap\pi^{-1}(Bg)}$ and $\pi|_{F'\cap\pi^{-1}(Bg)}$ are injective except on a set of measure zero. Hence,

$$m_X(Bg) = m\left(F \cap \pi^{-1}(Bg)\right) = m\left(F' \cap \pi^{-1}(Bg)\right) = m(Cg) = m(C) = m_X(B).$$

□

We can now prove ergodicity of the geodesic flow. Although historically this was not the first method, this argument due to Hopf [10] fits well in our study of the modular surface.

Theorem 7.9. For any $t > 0$, the geodesic flow R_{a_t} is a measure preserving and ergodic transformation of $X = PSL_2(\mathbb{Z}) \setminus PSL_2(\mathbb{R})$ with respect to the measure m_X.

Proof. For brevity, we will denote $\Gamma = PSL_2(\mathbb{Z})$. We will first show that R_{a_t} is measure preserving. We know that $m_X(X) = \pi/3$, so we normalize the measure m_X by multiplying it with $3/\pi$, so that we are formally in a probability space. We know that R_{a_t} is a bijection on Γ, and we want to show that right multiplication by R_{a_t} is a bijection on X too. Suppose that $x = \Gamma h_0, y = \Gamma h_1$, and that $R_{a_t}(x) = R_{a_t}(y)$, then for some $\gamma_0, \gamma_1 \in \Gamma$, $\gamma_0 h_0 a_t^{-1} = \gamma_1 h_1 a_t^{-1}$ which implies that $\gamma_1^{-1}\gamma_0 h_0 = h_1$, showing that $\Gamma h_0 = \Gamma h_1$. So R_{a_t} is injective. Let $y = \Gamma h$ be an element of X, then $x = \gamma h a_t^{-1} \in X$ and $R_{a_t}(x) = \Gamma h = y$, showing surjectivity. Hence, given a measurable set $B \subset X$, it holds that $R_{a_t}^{-1}(B) = Ba_t$, and since m_X is invariant under $PSL_2(\mathbb{R})$, R_{a_t} is measure preserving.

The rest of the proof is spent on showing ergodicity. We will use proposition 7.6, and show that for any measurable $f : X \to \mathbb{C}$ that is invariant under R_{a_t} almost everywhere, it holds that f is constant almost everywhere. The idea is the following: we want to show that f being invariant under R_{a_t} almost everywhere implies that f is invariant under the horocycle and unstable horocycle flow almost everywhere. We will then show that any element of X can be decomposed as a finite number of stable and unstable horocycle flows, i.e. $x = u_{s_1} h_{s_2} \cdots u_{s_n} h_{s_n}$, and the invariance of f then implies that f must be a constant almost everywhere. We have have to show this for functions f mapping to $\mathbb{C}$ that are invariant under R_{a_t}, but such functions must then also have real and imaginary parts that are invariant under R_{a_t} (and that are of course measurable). Hence we will assume without loss of generality that $f : X \to \mathbb{R}$. This will be important in the last step of the proof.

We will make use of two measure theory related properties of the measure on $PSL_2(\mathbb{R})$: it is an *inner regular measure*, meaning that for every measureable set A it holds that

$m(A) = \sup\{m(K) | K \subset PSL_2(\mathbb{R}), K$ compact and measureable$\}$. Hence, for any $\epsilon > 0$, there exists a measurable, compact set $K \subset A$, with measure $m_G(K) > m_G(K) - \epsilon$. Additionally, we will make use of Lusin's theorem [8, Theorem A.20], which says that for every measurable function f on $PSL_2(\mathbb{R})$ and any $\epsilon > 0$ there exists a continuous function $g : G \to \mathbb{C}$ with the property that

$$m_G(\{x \in G \mid f(x) \neq g(x)\}) < \epsilon.$$

It can be shown that these properties then also hold for m_X, but we do not discuss it further here.

First, we show that f is invariant under the horocycle flow almost everywhere. Fix $\epsilon > 0$. Using the above two properties, we can find a compact set K on which f is continuous and $m_X(K) > 1 - \epsilon$. Roughly speaking, we want to show that under iteration of R_{a_t}, two points $x, y \in X$ are in K over half of the time. In that case, if we compare the lists $(x, R_{a_t}(x), R_{a_t}^2(x), \ldots)$ and $y, R_{a_t}(y), R_{a_t}^2(y), \ldots)$, then there will a sequence $(l_n) \subset \mathbb{N}$ for which simultaneously $R_{a_t}^{l_n}(x)$ and $R_{a_t}^{l_n}(y)$ are in K. Define the set

$$B = \left\{ x \in X \mid \lim_{n\to\infty} \frac{1}{n} \sum_{l=0}^{n-1} \mathbb{1}_K(R_{a_t}^l(x)) > \frac{1}{2} \right\}.$$

Since R_{a_t} is measure preserving, Birkhoff's pointwise ergodic theorem implies that there exists an almost everywhere defined, integrable function

$$g^*(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{l=0}^{n-1} \mathbb{1}_K(R_{a_t}^l(x)) > \frac{1}{2} \in [0, 1]$$

so that $B = g^{-1}(1/2, 1]$ is measurable, and

$$\int g^* dm_X = \int \mathbb{1}_K dm_X = m_X(K).$$

We can now verify that $m_X(B) \geq 1 - 2\epsilon$:

$$1 - \epsilon \leq \int_B g^* dm_X + \int_{X\setminus B} g^* dm_X \leq m_X(B) + \frac{1}{2} m_X(X \setminus B) =$$

$$m_X(B) + \frac{1}{2} - \frac{1}{2} m_X(B) = \frac{1}{2} m_X(B) + \frac{1}{2}.$$

It then follows that indeed $m_X(B) \geq 1 - 2\epsilon$, so for almost every $x, y \in X$ we can find a sequence $(l_n) \subset \mathbb{N}$ such that $R_{a_t}^{l_n}$ maps x and y in to K. Suppose now that $y = R_{h_s}(x)$ for some $s \in \mathbb{R}$. Then f being invariant under R_{a_t} and using that y is a point of the stable manifold of x in X, we can deduce that $f(x) = f(R_{a_t}^{l_n}(x))$, $f(y) = f(R_{a_t}^{l_n}(y))$ and

$$d_X\left(R_{a_t}^{l_n}(x), R_{a_t}^{l_n}(u)\right) \to 0 \text{ as } n \to \infty.$$

So x and y are mapped arbitrarily close to one another in K, a compact set, where f is uniformly continuous. Suppose that $f(x) \neq f(y)$, then for every $\delta > 0$ there exists $\epsilon' = |f(x) - f(y)| > 0$ and $N \in \mathbb{N}$ such that $n \geq N$ implies $d_X\left(R_{a_t}^{l_n}(x), R_{a_t}^{l_n}(y)\right) < \delta$ but $|f(R_{a_t}^{l_n}(x)) - f(R_{a_t}^{l_n}(y))| > \epsilon'$, contradicting uniform continuity of f. Hence, $f(x) = f(y)$. This shows that f is invariant under the horocycle flow on B. Since ϵ was chosen arbitrarily, we may fix $\epsilon_1 < \epsilon$, and find a new set K_1 and B_1 accordingly. We may assume without loss of generality that $K \subset K_1$, since a finite union of compact sets is compact. It then follows that $B \subset B_1$. In other words, if we choose a smaller ϵ, we can find a larger set B on which f is invariant. Hence, if we let $\epsilon = 1/n$ and denote the associated B-set by $B_{1/n}$, then we may find $X' = \bigcup_{n\in\mathbb{N}} B_{1/n}$ of measure 1, on which f is invariant under the horocycle flow.

The same argument applied to Ra_t^{-1} yields the same conclusion, there is a set X'' of measure 1 with the property that f is invariant under the unstable horocycle flow. Hence, if we define $X_1 = X' \cap X''$ we have a set of measure 1 on which f is invariant under R_{a_t}, R_{h_s} and R_{u_s}.

Recall from equation 3 that every element g of $PSL_2(\mathbb{R})$ could be represented as $g = KU_1K^{-1}U_2KU_3K^{-1}U_4K^{-1}$ where

$$K = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad U_1, U_2, U_3, U_4 \in \mathcal{U} = \left\{ \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix} \mid s \in \mathbb{R} \right\}.$$

As was mentioned there, for any $U \in \mathcal{U}$, KUK^{-1} is lower diagonal with 1 on the diagonal. So, every element of $PSL_2(\mathbb{R})$ can be represented by as $g = u_{s_1}h_{s_2}u_{s_3}h_{s_4}K^{-1}$ for some $s_1, s_2, s_3, s_4 \in \mathbb{R}$. Since $K \in \gamma$, on X we can find a representative of $[g]$ that is of the form $u_{s_1}h_{s_2}u_{s_3}h_{s_4}$. We can now show that for every $g \in X$ there is a set X_g of measure 1 with the property that $\forall x \in X_g$ it holds that $f(x) = f(R_g(x))$, and that in particular this holds for $g = a_t$. Define

$$X_g = X_1 \cap R_{h_{s_4}}^{-1}(X_1) \cap R_{u_{s_3}h_{s_4}}^{-1}(X_1) \cap R_{h_{s_2}u_{s_3}h_{s_4}}^{-1}(X_1) \cap R_{u_{s_1}h_{s_2}u_{s_3}h_{s_4}}^{-1}(X_1).$$

The invariance of m_X under $PSL_2(\mathbb{R})$ shows that each set in the intersection is of full measure, and hence that the intersection of sets is of full measure. Moreover, suppose that $x \in X_1 \cap R_{h_{s_4}}^{-1}(X_1)$, then $f(x) = f(R_{h_{s_4}}(x))$ since $x \in X_1$, and $f(x) \in X_1$. Repeating this argument on $y = R_{h_{s_4}}(x)$, we find that $f(x) = f(y) = f(R_{u_{s_3}h_{s_4}}(x)) \in X_1$. Repeating this two more times gives us $f(x) = f(R_g(x))$.

Finally, we can show that f is constant almost everywhere with respect to m_X. Suppose f is not constant almost everywhere. Then there exist two disjoint intervals I_1, I_2 in $\mathbb{R}$, such that

$$C_j = \{h \in PSL_2(\mathbb{R}) \mid f(\gamma h) \in I_j\}$$

for $j = 1, 2$, are both not of measure zero with respect to m_G. By Proposition 7.7 we may conclude that there exists $g \in PSL_2(\mathbb{R})$ such that $m_G(C_1 \cap C_2 g) > 0$. If we denote

$g' = \pi(g) \in X$, then we have the set of measure 1, $X_{g'}$, on which f is invariant under R_g. Hence, the set

$$D_g = \{h \in PSL_2(\mathbb{R}) \mid \Gamma h \in X_{g'}\}$$

has full measure with respect to m_G, i.e. $m_g(G \setminus D_g) = 0$. This also means that there exists some $h \in PSL_2(\mathbb{R})$ that belongs to

$$D_g \cap C_1 g \cap C_2.$$

This is a contradiction, since $f(\Gamma h) = f\left(\Gamma h g^{-1}\right)$ since $h \in D_g$, but $f(\Gamma h) \in I_1$ since $h \in C_1$ and $f(\Gamma h g^{-1}) \in I_2$ since $h \in C_2 g$. □

8 The Gauss map

Geodesic flow on the modular surface can also be 'reduced' via a *Poincaré section.* The idea is to choose a curve C and, given a pair (z, v) on the curve, to follow its geodesic flow on E, the fundamental domain for $PSL_2(\mathbb{Z}) \backslash PSL_2(\mathbb{R})$ as in Proposition 6.12, until we intersect the chosen curve. The curve C is chosen in such a way as to not be tangent to the geodesic flow. We then keep following the flow on E until we cross the curve again. This generates a sequence of intersections with the curve. We can then define a map from this curve to itself, mapping a pair (z, v) to the pair $g_{t'}(z, v)$ where t' is the time at which the curve is intersected first. This map is called the Poincaré map or *first return map.*

It is possible to study properties of the geodesic flow via the returns to the Poincaré section. Historically, Artin [3] used a Poincaré section for the modular surface to show that the geodesic flow is ergodic, by showing that the Poincaré map is ergodic. We will derive a Poincaré section and its first return map, which will turn out to involve the Gauss map, as introduced in the first chapter, relating the geodesic flow on the modular surface and the Gauss map provide a connection between geometry, dynamical systems and number theory.

8.1 The Poincaré section of the modular surface

We will construct a new tessellation of the modular surface, following the paper of Caroline Series [15] in order to construct our Poincaré section. Recall that all elements of $PSL_2(\mathbb{Z})$ could be written in terms of elements τ and σ, as given in equation 5. Given E, we may apply σ to get a triangle in $\mathbb{H}$ with vertices $0, \frac{1 \pm \sqrt{3}i}{2}$. We can slice the quadrilateral $E \cup \sigma E$ in two, along the imaginary axis, and shift the left half by applying τ, so we get a new triangle F with vertices $0, 1, \infty$. Note that F is not a fundamental domain, but contains 2 distinct representatives of each coset of modular surface (for almost every coset of the modular surface). F would therefore also commonly be referred to as a *double cover* of the modular surface. Hence, if we look at the flow of the modular surface as represented in F, we need to be careful. In particular, points with any tangent vector on the imaginary axis of F belong to the same class of representatives of the modular surface if they are one anothers image under $-\frac{1}{z}$. We tile $\mathbb{H}$ by images of F under $PSL_2(\mathbb{Z})$, as in figure 8. This tessellation is called the Farey Tessellation.

We cannot look at every possible geodesic flow in F, so we restrict ourselves to those that do not go to ∞ in forward or backward time, i.e. those geodesics that are not vertical lines. Given an oriented geodesic u in E, the corresponding half-circle in $\mathbb{H}$ will have two end points on the real axis. By applying τ^n for some n, we may translate this circle such that the the starting point of geodesic lies in the interval $[-1, 0)$ or $(0, 1]$, while the end point lies in $[1, \infty)$ or $(-\infty, 0]$, respectively. We will call these points respectively

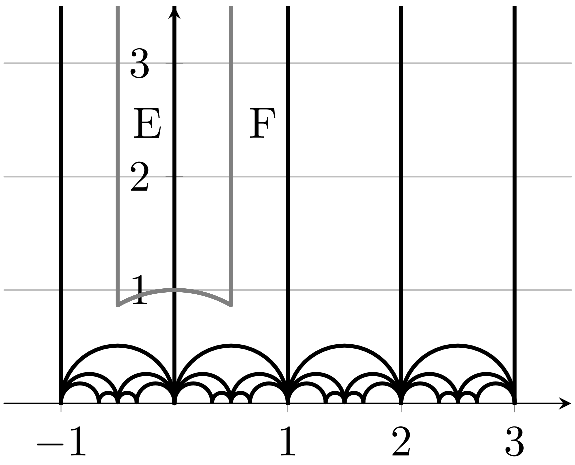

Figure 8: Farey tessellation of $\mathbb{H}$, E is displayed in gray, F is the black tile with vertices $0, 1$ and ∞.

$u_{-\infty}$ and u_∞. The geodesic flow on this translated circle in $\mathbb{H}$ projected to the modular surface will be identical to the flow before the translation, since $\tau \in PSL_2(\mathbb{Z})$.

We will consider the geodesic flow in T^1F as we did on T^1E, by translating the flow by τ and σ to an appropriate power when we hit the boundary of F. Note that $\tau^{\pm 1}$ is used to go from the right boundary to left and vice versa, while hitting the lower boundary is a little more complicated. Given a point and direction on a geodesic, we follow it until we are on the right or left boundary of F, such that the next boundary we hit is the lower boundary. Then we apply $\tau\sigma$ if we are on the left boundary or $\sigma\tau^{-1}$ if we are on the right boundary. The orientation of the new geodesic by applying will have the opposite orientation (i.e. left and right are switched) because we applied σ in either case.

We use the tiling of the hyperbolic half-plane by F to cut an oriented geodesic into slices. We assign to each slice of the geodesic the letter R for right or L for left, depending on the edges the slice enters and leaves through: if the edges meet in a vertex that lies to the left of the oriented geodesic we assign L, if the vertex lies to the right, we assign R. Examples are in the figure below. This assignment is invariant under $PSL_2(\mathbb{Z})$: we only need to check if $\tau^{\pm 1}$, $\sigma\tau^{-1}$ and $\tau\sigma$ leave the assigned R or L in F unchanged, which is readily verified. Given an oriented geodesic that crosses the imaginary axis in some point x where the previous boundary of F hit, was the lower boundary, we can associate to it a cutting sequence $\cdots R^{n_0} x L^{n_1} R^{n_n} \cdots$ or $\cdots L^{n_0} x R^{n_1} L^{n_n} \cdots$ for integers $n_i \in \mathbb{N}$, corresponding to the consecutive number of slices of the same type.

We define the cross section Y of T^1E to be all points on the imaginary axis in E with unit tangent vectors such that the corresponding geodesic in $\mathbb{H}$ has a cutting sequence that changes type at Y. By the discussion above, such a geodesic in F can have its

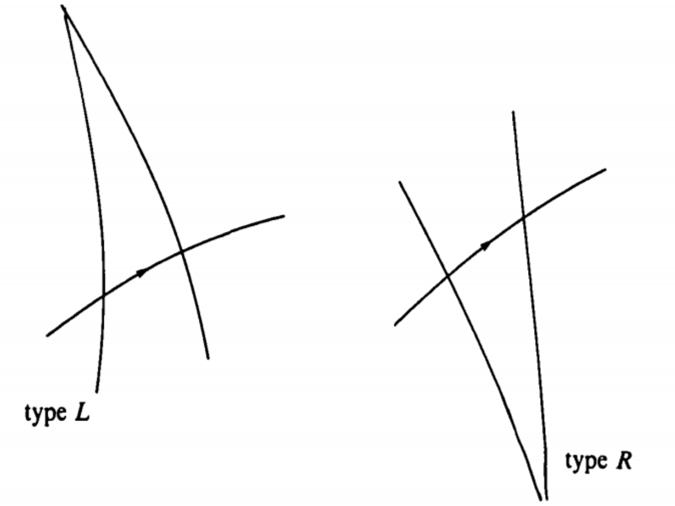

Figure 9: Cutting sequence types, the geodesic flow goes over the line with the arrow on it. To other two black lines represent the edges. Figure 2 from [15]

beginning and end point $u_{-\infty}$ and u_{∞} as follows: $0 < |u_{-\infty}| \leq 1$ and $|u_{\infty}| \geq \infty$ and both are on opposite sides of the imaginary axis. Denote this set of geodesics by A. Moreover, denote for a geodesic $u \in A$ the point of intersection with the imaginary axis and corresponding tangent vector by (ib, v). We will show that the geodesics in A are the only geodesics corresponding to geodesic flow from Y. Note again that every geodesic on the modular surface can be put in such a form, while the cutting sequence is unchanged since we are using only τ, which preserves the orientation of the geodesic. Since F is a double cover, it is not clear whether or not the cutting sequence is unique: there might by another geodesic in F that represents the flow in E with a different cutting sequence.

Remark. There is some ambiguity in the cutting sequence: what if the geodesic under consideration happens to have u_{∞} or $u_{-\infty}$ on one of the vertices $0, 1$? It certainly means that the geodesic flow will never again leave F, so we might assign L^{∞} or R^{∞} in this case, to denote the end of the cutting sequence. The ambiguity of the choice of L or R corresponds to an ambiguity of finite continued fractions: we will shortly see that, starting from the Poincaré section in a point x, if the end point is expanded, $u_{\infty} = [n_0; n_1, n_2, \cdots]$, then cutting sequence around x is $\cdots x L^{n_0} \cdots$. Suppose that $u_{\infty} = n_0 = (n_0 - 1) + \frac{1}{1+0}$, then we have two continued fraction expansions for u_{∞}: $[n_0; 0, 0, \cdots] = [n_0 - 1; 1, 0, \cdots]$. Hence, we cannot unambiguously define the end of cutting sequence.

Proposition 8.1. The map $\iota : A \to Y$, $\iota(u) = \pi(ib, v)$ is surjective, continuous and open. It is injective except for the two oppositely oriented geodesics joining 1 to -1, which have the same image. Moreover, if $(ib, v) \in Y$ defines a geodesic with cutting sequence $\cdots R^{n_0} x L_{n_1} \cdots$ then $\iota^{-1}(u_x)$ has a beginning and end point respectively given by

$$u_\infty = [n_1, n_2, \cdots]$$
$$-\frac{1}{u_{-\infty}} = [n_0, n_{-1}, \cdots].$$

If the cutting sequence R and L are interchanged, the corresponding geodesic has beginning and end point

$$u_\infty = -[n_1, n_2, \cdots]$$
$$\frac{1}{u_{-\infty}} = [n_0, n_{-1}, \cdots]$$

Proof. Let u be any oriented geodesic in $\mathbb{H}$ which intersects the imaginary axis. Since F is convex, $u_\infty \geq 1$ if and only if the segment of u in F is type L, and similarly $-1 \leq u_{-\infty} < 0$ if and only if the segment before F is of type R. Similarly, $u_\infty \leq -1$ and $0 < u_{-\infty} \leq 1$ if and only if the segment in F is of type R and the preceding segment is of type L. This shows that all geodesics that project to Y belong to A. Since any geodesic cutting the imaginary axis in E, in any point with any tangent vector, can be put in A by τ, we have that ι is surjective.

Suppose that there exist distinct $u_1, u_2 \in A$ such that $\iota(u_1) = \iota(u_2) = (ib, v)$, then u_1, u_2 correspond to the geodesic flow of the same class of representatives of E. As was remarked, the only double representatives on the imaginary axis are related by $-\frac{1}{z}$, so $ib = \frac{1}{b} i$ which implies that $b = 1$. Then only two representatives that both point inside of E and are related under $D\sigma$ are $(i, \pm 1)$, whose geodesics indeed correspond to the half-circle joining 1 to -1. Finally, since π is a continuous and open, so is ι. This finishes the first part of the proof.

Suppose that $u \in A$ and $u_\infty > 1$. Let $u_\infty = [n_1, n_2, \cdots]$. Let $p_1 = n_1$ if $u_\infty > n_1$, and let $p_1 = n_1 - 1$ if $u_\infty = n_1$. Starting from x on the imaginary axis, where the cutting sequence changes type, we follow the trajectory along u until we hit the right boundary of F, at which moment we can apply τ^{-1} to stay inside F. This will happen a total of p_1 times. After that, we will hit the lower boundary. meaning that the section in which we hit the lower boundary is of type R. Thus, the cutting sequence around x is $R^{n_0} x L^{p_1} R^{n_2}$. If $u_\infty = 1$, then we assign either R_∞ or L_∞.

If we apply the map $\sigma\tau^{-p_1} : z \mapsto -\frac{1}{z - p_1}$ to u, we will have a new geodesic u' with beginning and end point $u'_{-\infty} = -\frac{1}{u_{-\infty} - p_1} \in (0, 1]$ and $u'_\infty = -\frac{1}{u_\infty - p_1} \leq -1$. Hence, $u' \in A$. Denote by η the point on the imaginary axis that u crossed last before hitting the lower boundary. As was remarked, the cutting sequence of u', around $\sigma(\eta)$, will be the same as that of u around η. If $\sigma(\eta) = -1$ the next element will be either R_∞ or L_∞. Otherwise, we may set $p_2 = n_2$ if $u'_\infty < -n_2$ or $p_2 = n_2 - 1$ if $u_\infty = -n_2$, and we may follow the flow along u' from $\sigma(\eta)$ until we hit the left boundary of F, at which point we apply τ. Note that the integer part of $|u'_\infty|$ is given by n_2 because of the continued fraction expansion of u_∞. We will have hit the left boundary of F a total of p_2 times, so will have a cutting sequence of R^{p_2}, after which the type of the sequence changes and we may apply the map $\sigma\tau^{p_2}$ and repeat the argument. In conclusion, if $u_{-\infty} \in [-1, 0)$ and $u_\infty \geq 1$, then the cutting sequence of the geodesic starting from x will be $\cdots xL^{n_1}R^{n_2}L^{n_3}\cdots$ if $u_\infty = [n_1, n_2, n_2, \cdots]$ has an infinite continued fraction expansion. A similar reasoning yields the same result if $u_\infty \leq -1$ and $u_{-\infty} \in (0, 1]$.

If we reverse the orientation of the geodesic $u \in A$ and apply σ to u, we will get a new geodesic u' with beginning and end point $u'_{-\infty} = -\frac{1}{u_\infty}$ and $u'_\infty = -\frac{1}{u_{-\infty}}$, and it follows that $u' \in A$. Flipping the orientation of u means that in the cutting sequence we switch R and L, and we follow the sequence from right to left. In other words, if the cutting sequence around x was $\cdots L^{n-1}R^{n_0}xL^{n_1}R^{n_2}\cdots$ then the cutting sequence around $\sigma(x)$ will be $\cdots L^{n_2}R^{n_1}\sigma(x)L^{n_0}R^{n-1}\cdots$ by the above. So if $u_\infty = [n_1, n_2, n_2, \cdots]$ and $-\frac{1}{u_{-\infty}} = [n_0, n_{-1}, \cdots]$, we see that the cutting sequence of u around x is indeed $\cdots L^{n-1}R^{n_0}xL^{n_1}R^{n_2}\cdots$. This finishes the proof. □

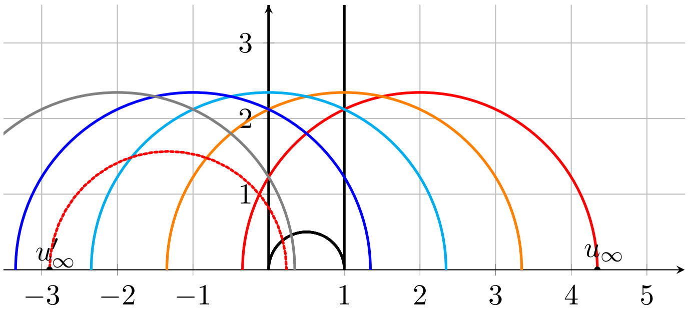

Figure 10: example of the geodesic flow on F (in black). If one starts on the imaginary axis on the red line, one is on the Poincaré section. One consecutively travels over the red, orange, light blue, dark blue line. Then one applies τ^{-1} a final time to obtain the gray line. At this point the next intersection with the boundary of F is in the bottom boundary, so we apply σ and obtain the red dashed line. This is the first return to the Poincaré section. Starting on imaginary axis on the solid red line, the cutting sequence is $\cdots xL^4R^2\cdots$.

The proof of the proposition shows something more. Given an oriented geodesic $u \in A$ with $u_\infty = [n_1, n_2, n_3, \cdots]$ (that does not terminate in one step) and $-\frac{1}{u_\infty} = [n_0, n_{-1}, \cdots]$, the cutting sequence around x will be $\cdots L^{n-1}R^{n_0}xL^{n_1}R^{n_2}\cdots$. Moreover, the first return to A will be on a geodesic $u' = \sigma\tau^{-n_0}$ with cutting sequence $\cdots R^{n_0}L^{n_1}x'R^{n_2}L^{n_3}\cdots$, where $x' = \sigma(\eta)$ with η the last point on the imaginary axis before the sequence changes type. The takeaway is that u' will have a starting and end point $u'_\infty = -[n_2; n_3, \cdots]$

and $\frac{1}{u'_{-\infty}} = [n_0; n_1, \cdots]$. Hence, the first return to X on the modular surface can be viewed as a shift on the continued fraction expansion of beginning and end points of the geodesic associated to (ib, v) in X. This can all be made much more precise, for which we refer to [15, Theorem 2 & 3].

9 Epilogue

The theory presented in this thesis has a modest scope and size, compared to the surrounding material. Therefore, we introduce some generalizations with associated literature. We hope to leave the reader with an appetite for studying further properties of the Poincaré half-plane and its quotients, and with a solid basis to do so.

The construction of the modular surface has been generalized to other quotients of the Poincaré half-plane, as was remarked. Instead of $PSL_2(\mathbb{Z})$ one can take any other Fuchsian group. The fundamental domain will then be another (possibly infinite) polygon. This and more may be found in the book by S. Katok [11]. One may tessellate $\mathbb{H}$ by these polygons, as we did for $PSL_2(\mathbb{Z})$, and study cutting sequences through the boundaries of these polygons. This can lead to the discovery of a Poincaré section for such a system. These cutting sequences will in general have a larger 'alphabet' than L and R, since the number of sides of a polygon can exceed three. On may also extend Hopf's argument to prove that the geodesic flow on these more general quotients is also ergodic [11, Theorem 17.4].

The Poincaré section for the modular surface as presented in this thesis has some weaknesses: one has to use a double cover with a specific fundamental domain, which has frequently led to problems [15]. As it is close to what Artin originally discovered, the author has chosen to include it. However, other Poincaré sections for the modular surface exist, which use the fundamental domain of Proposition 6.12. A number of these sections are discussed in [12], and make use of other number theoretic expansions, such as *negative continued fraction expansions* and *nearest integer fraction expansions.*

The connection between the geodesic flow on the modular surface and the Gauss map, can be used to introduce a *thermodynamic formalism.* As was stated in the introduction, originally ergodic theory came from statistical mechanics, but it has since been mathematically formalized. For example, the book of D. Ruelle [14] introduces the thermodynamic formalism from a mathematical point of view. The main tool in this theory is the *transfer operator method*, an introduction of which can be found in [4, Chapter 7], written by D.H. Mayer. The thermodynamic formalism requires some advanced spectral theory. A rough sketch of the idea is given below.

One can view the dynamical system on modular surface (or other quotients), as though it was a thermodynamic system. For example, the geodesic flow could induce the dynamics via the first return map. One can then also study what the same dynamic looks like on the Poincaré section, which essentially means studying a shift on a sequence of letters from an alphabet (i.e. a shift on the cutting sequence). This type of physical model is well studied in the context of so-called lattice spin systems in statistical mechanics. Of particular interest is the study of the the free energy of the system associated to an observable A. An observable is a continuous function from the Poincaré section to the reals. The free energy associated to an observable has a nice interpretation in a physical context: it roughly describes the amount of work a system can do at constant temperature and pressure. One can then derive properties of observables of the system (for example the distribution of the lengths of geodesics), from the free energy. The free energy of an observable can be expressed as a function of a so-called *partition function* of the system. The transfer operator method has as its purpose to find this partition function, and does so by expressing the partition function as the trace of a certain operator associated to the observable.

References

[1] Wikimedia Commons. https://commons.wikimedia.org/wiki/File:ModularGroup-FundamentalDomain.svg. Accessed: 12-5-2019, figure on geodesics and the fundamental domain of the modular surface.

[2] *Variational calculus in the large*, Encyclopedia of Mathematics. http://www.encyclopediaofmath.org/index.php?title=Variational_calculus_in_the_large&oldid=28277. Accessed: 27-6-2019.

[3] E. Artin, *Ein mechanisches system mit quasiergodischen bahnen*, Abhandlungen aus dem Mathematischen Seminar der Universitaet Hamburg **3** (1924), 170–175.

[4] T. Bedford, M. Keane, and C. Series, *Ergodic theory, symbolic dynamics and hyperbolic spaces*, Oxford University Press, New York, New York, 1991.

[5] É. Borel, *Les probabilités dénombrables et leurs applications arithmétiques*, Rend. Circ. Mat. Palermo **27** (1909), 247–271.

[6] H. De Snoo and H. Winkler, *Measure and integration, an introduction*, 2018. Course syllabus for the course Measure and Integration, taught in Groningen.

[7] M.P. Do Carmo, *Differential geometry of curves and surfaces*, Dover Publications, Inc., Mineola, New York, 2016.

[8] M. Einsiedler and T. Ward, *Ergodic theory, with a view towards number theory*, Springer, Berlin, 2010.

[9] Euclid, *Elements*, 888. http://www.claymath.org/euclid/index/book-1-postulates. Accessed: 08-05-2019, a copy from 888 AD made by Stephen the Clerk, translated by Sir Thomas Heath in 1956.

[10] E. Hopf, *Ergodic theory and the geodesic flow on surfaces of constant negative curvature*, Bulletin of the American Mathematical Society **77** (1971), 863–877.

[11] S. Katok, *Fuchsian groups, geodesic flows on surfaces of constant negative curvature and symbolic coding of geodesics*, Clay Mathematical Proceedings **8** (2008). PDF verison, available at https://pdfs.semanticscholar.org/6e7a/e42c8d9f657cabee452fa7a775ed82e8e3d4.pdf. Accessed 4-7-2019.

[12] S. Katok and I. Ugarcovici, *Arithmetic coding of geodesics on the modular surface via continued fractions* (2005).

[13] J. Jr. Palis and W. de Melo, *Geometric theory of dynamical systems, an introduction*, Springer-Verlag, New York, New York, 1982.

[14] D. Ruelle, *Thermodynamic formalism*, Cambridge University Press, 2004. Online publication date: January 2010.

[15] C. Series, *The modular surface and continued fractions*, J. London Math. Soc. **32** (1985), no. 2, 69–80.

[16] L.W. Tu, *An introduction to manifolds*, Springer, New York, New York, 2011.

[17] C. Wendl, *Lecture notes on bundles and connections*, 2008. Notes from undergraduate course at MIT in Differential Geometery. Available at https://www2.mathematik.hu-berlin.de/ wendl/connections.html. Accessed 27-6-2019.