From metrics to moduli space: *preliminary draft*

Christopher J. Leininger

August 8, 2011

Contents

1 Overview, review and first definitions

1.1 Overview

The purpose of these lectures is to provide a geometric–topological introduction to the theory of surfaces. After making precise what a surface is, we will begin the task of understanding all equivalence classes of (closed) surfaces. What notion of equivalence we are interested in depends on what sort of surfaces we are considering, and what kind of information we want to keep track of (the closed assumption imposes a kind of "finiteness" constraint). We will consider two types of surfaces—purely topological surfaces and geometric surfaces and natural equivalence relations on these.

The topological situation forms the backbone of the theory. It aims to understand the most general notion of surface, and hence influences all other notions. Here one of the main objectives will be to develop the necessary tools and prove the Classification Theorem.

The geometric theory is a refinement of the topological one in which additional structure is imposed on a surface. We will focus on a particular notion of "geometric" in terms of locally homogeneous surfaces, but along the way we will point out how this is in fact equivalent to many other notions. Again the objective will be to arrive at a classification-type theorem.

The topological classification is into countably many equivalence classes. The geometric classification is into continuously varying families which are naturally separated from one another according to the topological classification. The pieces of this subdivision are the so-called Riemann moduli spaces. We end with an analysis of the Riemann moduli space using a combination of geometric and topological techniques.

Our goal is to keep the technical machinery to a minimum while still arriving at a description of the general theory. To this end, we work entirely in the world of metric spaces. However, from time to time we will try to indicate how our perspective is situated in the more general context.

1.2 Calculus and linear algebra review

We will denote n–dimensional Euclidean space as

$$\mathbb{R}^n = \{\mathbf{x} = (x_1, \ldots, x_n) \mid x_i \in \mathbb{R}\}.$$

We measure distances in $\mathbb{R}^n$ according to

$$|(x_1, \ldots, x_n) - (y_1, \ldots, y_n)| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

We can also view $\mathbb{R}^n$ as an n–dimensional real vector space, and when we do this we sometimes represent vectors as column vectors. The Euclidean norm $|\cdot|$ used in the measure of distance is induced by the standard Euclidean inner product

$$\mathrm{std}(\mathbf{u}, \mathbf{v}) = \langle \mathbf{u}, \mathbf{v} \rangle = \langle (u_1, \ldots, u_n), (v_1, \ldots, v_n) \rangle = \sum_{i=1}^{n} u_i v_i.$$

The angle $\theta \in [0, \pi]$ between two nonzero vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ satisfies the equation

$$\cos(\theta) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{|\mathbf{u}||\mathbf{v}|}.$$

More generally, on any real vector space with an inner product, we can define the norm, angle and distance using these as definitions.

A path in $\mathbb{R}^n$ is a continuous map $\gamma : [a, b] \to \mathbb{R}^n$, where $[a, b] \subset \mathbb{R}^n$ is an interval (we will also consider paths defined on any type of interval including rays or the entire real line $\mathbb{R}$). We say γ connects $\gamma(a)$ and $\gamma(b)$. The length is defined as the supremum of lengths of polygonal approximations:

$$\ell(\gamma) = \sup \sum_{j=1}^{k} |\gamma(t_j) - \gamma(t_{j-1})|$$

where the supremum is taken over all partitions $a = t_0 < t_1 < \cdots < t_k = b$ of the interval.

Continuously differentiable maps are called C^1. The length of a C^1 path $\gamma : [a, b] \to \mathbb{R}^n$ (or a piecewise C^1 path, meaning a path for which the domain can be partitioned into a finite set of subintervals on which the path is C^1) can be computed as an integral

$$\int_a^b |\gamma'(t)| dt.$$

If we write γ in terms of it s coordinates, then $\gamma'(t) = (\gamma_1'(t), \ldots, \gamma_n'(t))$, and this is the tangent vector to γ at $\gamma(t)$. More generally, at any point $\mathbf{x} \in \mathbb{R}^n$ we consider the vector space of tangent vectors based at $\mathbf{x}$ called the tangent space at $\mathbf{x}$. This is simply $\mathbb{R}^n$ again, but we think of having a distinct copy of $\mathbb{R}^n$ for each point in $\mathbf{x}$. To help distinguish tangent vectors based at different points, we sometimes use a subscript. Thus, the vector $(1, 1, \ldots, 1)_{\mathbf{x}}$ represents the vector $(1, 1, \ldots, 1)$ thought of as a tangent vector at the point $\mathbf{x}$. However, we often omit this subscript when the desired basepoint is clear. In this notation, we might write the tangent vector to γ at $\gamma(t)$ as $(\gamma'(t))_{\gamma(t)}$ though this notation is unnecessarily cumbersome since it is clear that we would like to think of $\gamma'(t)$ as a vector based at $\gamma(t)$.

A subset $U \subset \mathbb{R}^n$ is open if for every point $\mathbf{x} \in U$ there is an $\epsilon > 0$ so that the ϵ–ball about $\mathbf{x}$ is contained in U:

$$B_\epsilon(\mathbf{x}) = \{\mathbf{y} \mid |\mathbf{x} - \mathbf{y}| < \epsilon\} \subset U.$$

Given a C^1 map $F : U \to \mathbb{R}^m$ defined on an open subset $U \in \mathbb{R}^n$, the derivative at $\mathbf{x} \in U$ is a linear map from the tangent space at x to the tangent space at $\mathbf{x}$. With respect to the standard basis vectors this is given by the matrix of partial derivatives

$$dF_{\mathbf{x}} = \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(\mathbf{x}) & \ldots & \frac{\partial F_1}{\partial x_n}(\mathbf{x}) \\ \vdots & & \vdots \\ \frac{\partial F_m}{\partial x_1}(\mathbf{x}) & \ldots & \frac{\partial F_m}{\partial x_n}(\mathbf{x}) \end{pmatrix}.$$

For a C^1 path $\gamma : [a, b] \to \mathbb{R}^n$, the relation with the derivative above is that $d\gamma_t(1) = \gamma'(t)$ where 1 is the standard basis vector for $\mathbb{R} = \mathbb{R}^1$ (= tangent space at any point of $\mathbb{R}$). With this notation, the chain rule for compositions $F \circ G$ takes on a particular simple form

$$d(F \circ G)_{\mathbf{x}} = dF_{G(\mathbf{x})} \circ dG_{\mathbf{x}}$$

and in terms of the matrices, the composition $\circ$ between the linear maps on the right is simply matrix multiplication. In particular, if γ is a C^1 path into an open set $U \subset \mathbb{R}^n$ and $F : U \to \mathbb{R}^m$, then

$$(F \circ \gamma)'(t) = dF_{\gamma(t)}(\gamma'(t)).$$

1.3 Distances in $\mathbb{R}^n$ and straight segments

Given two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we are taught from birth that the most efficient way to get from $\mathbf{x}$ to $\mathbf{y}$ is along the straight line segment from $\mathbf{x}$ to $\mathbf{y}$. This is the content of the following

Theorem 1.1. *If $\gamma : [a, b] \to \mathbb{R}^n$ is a piecewise C^1 path from $\mathbf{x}$ to $\mathbf{y}$, then*

$$\ell(\gamma) \geq |\mathbf{x} - \mathbf{y}|.$$

Moreover, equality holds if and only if γ is a monotone parameterization of the straight line segment from $\mathbf{x}$ to $\mathbf{y}$.

We sketch two proofs of this fact. The first contains an idea which will be very useful later. The second is more elementary, and actually works for any path (not just piecewise C^1 paths).

Proof #1. Let $L \subset \mathbb{R}^n$ denote the straight line through $\mathbf{x}$ and $\mathbf{y}$, and $\pi : \mathbb{R}^n \to L$ the orthogonal projection.

Exercise 1.1. *For any $\mathbf{z} \in \mathbb{R}^n$, prove that $|d\pi_{\mathbf{z}}(\mathbf{v_z})| \leq |\mathbf{v_z}|$ with equality if and only if $\mathbf{v_z}$ is parallel to L.*

Now we have

$$\ell(\gamma) = \int_a^b |\gamma'(t)| dt \geq \int_a^b |d\pi_{\gamma(t)}(\gamma'(t))| dt = \int_a^b |(\pi \circ \gamma)'(t)| dt = \ell(\pi \circ \gamma). \qquad (1)$$

Since $\pi \circ \gamma$ also connects $\mathbf{x}$ to $\mathbf{y}$, this shows that for the first statement it suffices to consider only paths contained in L where the theorem is clear. To see the second statement, observe that any path from $\mathbf{x}$ to $\mathbf{y}$ that is not contained in L must at some point have a tangent vector which is not parallel to L and so the inequality in (1) is strict in this case. □

Proof # 2. By definition of length, it suffices to prove this for polygonal paths connecting $\mathbf{x}$ to $\mathbf{y}$ since the length of a path is the supremum of lengths of such paths. Then we can induct on the number of edges in the polygonal approximation, reducing any polygonal path with $k \geq 2$ edges to a shorter one with $k - 1$ edges by applying the triangle inequality which states that for all $\mathbf{z_1}, \mathbf{z_2}, \mathbf{z_3} \in \mathbb{R}^n$ we have

$$|\mathbf{z_1} - \mathbf{z_2}| \leq |\mathbf{z_1} - \mathbf{z_3}| + |\mathbf{z_3} - \mathbf{z_2}|,$$

with equality if and only if $\mathbf{z_3}$ lies on the line segment between $\mathbf{z_1}$ and $\mathbf{z_2}$. □

1.4 Metric spaces

Given a set X, a function

$$d : X \times X \to \mathbb{R}$$

is called a *metric* if it satisfies the following three axiom for all $x, y, z \in X$.

1. $d(x, y) \geq 0$ with equality if and only if $x = y$, (positive definite)
2. $d(x, y) = d(y, x)$, (symmetry)

3. $d(x, y) \leq d(x, z) + d(z, y)$ (triangle inequality)

A set with a metric (X, d) is called a *metric space*. We sometimes omit d in the notation.

A subset $U \subset X$ is called *open* if for every point $x \in U$ there is an $\epsilon > 0$ so that the ϵ–ball about x is contained in U:

$$B_\epsilon(x) = \{\mathbf{y} \mid d(x, y) < \epsilon\} \subset U.$$

The collection of open sets in a metric space X is called the *topology* of X. An open set containing a point $x \in X$ is called a *neighborhood* of x. A subset $A \subset X$ is called *closed* if the complement $X - A$ is open.

Given two metric spaces (X, d) and (Y, ρ), a map $f : X \to Y$ is said to be continuous if for all $x \in X$ and $\epsilon > 0$ there exists $\delta > 0$ so that

$$\rho(f(x), f(y)) < \epsilon \text{ for all } y \in X \text{ with } d(x, y) < \delta.$$

Examples

1. $\mathbb{R}^n$ together with the function $d(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}|$ is a metric space.

2. Any set X with the function $d(x, y) = 1$ if $x \neq y$ and $d(x, x) = 0$ is a metric space.

3. If (X, d) is a metric space and $A \subset X$, then the restriction $d_A = d|_{A\times A}$ is a metric called the *subspace metric* and (A, d_A) is a metric space. In this case we call A a subspace of X.

4. The product of two metric spaces $(X, d), (Y, \rho)$ becomes a metric space using any of the following (with $p \geq 1$) as metrics

$$(d \times_p \rho)((x_1, y_1), (x_2, y_2)) = \sqrt[p]{d(x_1, x_2)^p + \rho(y_1, y_2)^p}.$$

   For the case $p = 1$, we simply write $d \times \rho$.

5. Consider the n–sphere in $\mathbb{R}^{n+1}$

$$\mathbb{S}^n = \{\mathbf{x} \in \mathbb{R}^{n+1} \mid |\mathbf{x}| = 1\} \subset \mathbb{R}^{n+1}.$$

   (for concreteness, consider $n = 1$ or $n = 2$). This is a subspace of $\mathbb{R}^{n+1}$ so we could give it the subspace metric. However, this does not naturally capture the geometry of creatures who live in $\mathbb{S}^n$. A "better" metric is defined by declaring

$$d(\mathbf{x}, \mathbf{y}) = \inf \ell(\gamma)$$

   where γ ranges over all paths in $\mathbb{S}^n$ connecting $\mathbf{x}$ to $\mathbf{y}$.

Exercise 1.2. *Prove that this defines a metric on $\mathbb{S}^n$, with the same topology as that of the subspace metric. In particular, prove that*

$$|\mathbf{x} - \mathbf{y}| \leq d(\mathbf{x}, \mathbf{y}) \leq \theta(\mathbf{x}, \mathbf{y})$$

where $\theta(\mathbf{x}, \mathbf{y})$ is the angle between $\mathbf{x}$ and $\mathbf{y}$, thought of as vectors in $\mathbb{R}^{n+1}$ (in fact, the second inequality is an equality).

If (X, d) is a metic space, a path into X is a continuous map $\gamma : [a, b] \to X$. The length of a path $\gamma : [a, b] \to X$ is defined as in $\mathbb{R}^n$

$$\ell(\gamma) = \sup \sum_{j=1}^{k} d(\gamma(t_j), \gamma(t_{j-1}))$$

where the supremum is taken over all partitions $a = t_0 < t_1 < \cdots < t_k = b$ of the interval. The length of a path between two points x and y in X is always greater than or equal to the distance between the endpoints by the triangle inequality (compare proof #2 of Theorem 1.1). However, there may be no geodesics between any pair of points. Indeed, there may be no paths at all (e.g. consider the subspace $\{0, 1\} \subset [0, 1]$).

Exercise 1.3. *Even when every two points are connected by a path, there may be no paths of finite length. Prove that on the set* $[0, 1]$*, the function* $d(x, y) = \sqrt{|x - y|}$ *is a metric having the same collection of open sets as the standard metric* $|x - y|$ *on* $[0, 1]$*, but that every nonconstant path has infinite length.*

A metric space for which the distance between any two points is the infimum of lengths of paths between the points is called a *path metric space*. A path between a pair of points whose length is equal to the distance is called a *geodesic* between the points. If any two points of a metric space are connected by a geodesic, then the metric space is called a *geodesic metric space*. By Theorem 1.1, $\mathbb{R}^n$ is a geodesic metric space. By the same theorem, $\mathbb{R}^n - \{\mathbf{0}\}$ is a path metric space, but **not** a geodesic metric space—points on opposite sides of $\mathbf{0}$ are not connected by a geodesic. By construction, $\mathbb{S}^n$ is a path metric space (and is also a geodesic metric space and we'll prove that tomorrow).

1.5 Homeomorphisms and Isometries

There are two natural notions of "equivalence" for metric spaces (X, d) and (Y, ρ).

A continuous map $f : X \to Y$ is a *homeomorphism* if it is a bijection (one-to-one and onto) and f^{-1} is continuous. Observe that the inverse of a homeomorphism is also a homeomorphism. We say (X, d) and (Y, ρ) are *homeomorphic* if there exists a homeomorphism from one to the other. The insistence that f^{-1} be continuous is justified by the example of a continuous bijection

$$f : [0, 1) \to \mathbb{S}^1$$

given by $f(t) = (\cos(2\pi t), \sin(2\pi t))$. The inverse map is not continuous (why?). Indeed, there is no homeomorphism between $[0, 1)$ and $\mathbb{S}^1$, and we will be able to prove this by the end of the next lecture (one is compact and the other is not).

A map $f : X \to Y$ is an *isometry* if it is a bijection and $d(x, y) = \rho(f(x), f(y))$ for all $x, y \in X$. We say (X, d) and (Y, ρ) are *isometric* if there exists an isometry from one to the other.

Exercise 1.4. *Prove that an isometry is a homeomorphism and that the inverse of an isometry is an isometry. Prove that the composition of isometries/homeomorphisms are isometries/homeomorphisms.*

1.6 Some basic facts

We end with a few basic facts about metric spaces whose proofs are left as exercises.

Exercise 1.5. *Prove each of the following.*

Theorem 1.2. *A function $f : X \to Y$ is continuous if and only if for every open set $U \subset Y$, $f^{-1}(U)$ is open in X. Similarly, f is continuous if and only if for every closed set $A \subset Y$, $f^{-1}(A)$ is closed in X.*

□

Corollary 1.3. *Two metrics d and ρ on a set X define the same topologies if and only if the identity $X \to X$ is a homeomorphism with respect to d on the domain and ρ on the range.*

We say that a point x in a metric space X is a *limit point* of a subset $A \subset X$ if every neighborhood U of x has $U \cap A - \{x\} \neq \emptyset$.

Proposition 1.4. *A subset of a metric space $A \subset X$ is closed if and only if it contains all its limit points.*

□

Given a sequence $\{x_n\}_{n=1}^{\infty} \subset X$, we say that $\lim_{n\to\infty} x_n = x \in X$ if for every neighborhood U of x there exists $N > 0$ so that for all $n \geq N$, $x_n \in U$.

Proposition 1.5. *$f : X \to Y$ is continuous if and only if*

$$\lim_{n\to\infty} f(x_n) = f(x)$$

for all sequence $\{x_n\} \subset X$ with $\lim_{n\to\infty} x_n = x$.

□

The *closure* of a subset $A \subset X$, denoted $\overline{A}$, is the intersection of all closed sets containing A.

Proposition 1.6. *Given any subset $A \subset X$ of a metric space, $\overline{A}$ is a closed set equal to A union with its limit points.*

□

The *interior* of a set $A \subset X$, denoted A°, is the union of all open sets contained in A.

Proposition 1.7. *Given any subset $A \subset X$ of a metric space, A° is an open set equal to $X - \overline{X - A}$.*

□

2 Metric and topological properties

Throughout, (X, d) and (Y, ρ) will be metic spaces.

2.1 Compactness

Last time Professor Brendle defined compactness for any subset of a metric space (X, d), just in terms of the topology—A is compact if every open cover contains a finite subcover—and proved that the image of a compact set under a continuous function is compact. In Euclidean space $\mathbb{R}^n$, the Heine–Borel Theorem states that compact sets are precisely the closed and bounded sets. This has the following consequence which is sometimes called the Maximum/Minimum Value Theorem.

Theorem 2.1. *If X is a compact metric space, then any continuous function $f : X \to \mathbb{R}$ attains a minimum and maximum value.*

Proof. $f(X)$ is compact, hence closed and bounded in $\mathbb{R}$. Since it is bounded, the infimum and supremum of $f(X)$ are both finite numbers. If the supremum of $f(X)$ is not in $f(X)$, then it is a limit point which is not contained in $f(X)$ contradicting Proposition 1.4 since $f(X)$ is closed. The same argument works for the infimum. □

Compactness for metric spaces has another description which is often useful. We say that X is *limit point compact* if every infinite subset $A \subset X$ has a limit point in X.

Proposition 2.2. *If X is a metric space, it is compact if and only if it is limit point compact.*

Proof. First, suppose X is compact, and $A \subset X$. We prove that if A has no limit points, then A is finite. For this, just observe that if A has no limit points, then A is closed (since it contains all its limit points), and hence compact. For each $a \in A$, let U_a be a neighborhood that contains a, and no other element of A. This is possible because no a in A is a limit point. The set $\{U_a\}_{a \in A}$ is an open cover of A, so has a finite subcover. Since $U_a = \{a\}$ for each a, this means that A is finite.

Exercise 2.1. *Prove that if X is limit point compact, then every sequence in X has a convergent subsequence.*

The other implication of the proposition requires the following technical fact, often call the Lebesgue number lemma.

Lemma 2.3. *If X is a limit point compact metric space and $\{U_\alpha\}$ is an open cover, then there exists a number $\epsilon > 0$ so that for all $x \in X$, $B_\epsilon(x)$ is contained in some U_α.*

Proof. Suppose that there is a covering $\{U_\alpha\}$ of X and no $\delta > 0$ as in the lemma. Take a sequence $\{x_n\}$ in X for which $B_{1/n}(x_n)$ is not contained in any U_α. Since X is limit point compact, Exercise 2.1 gives us a subsequence $\{x_{n_k}\}$ that converges to some $x \in X$. Let $\epsilon > 0$ be such that $B_\epsilon(x) \subset U_\alpha$ for some α. But then, x_{n_k} lies in $B_\epsilon(x)$ for all sufficiently large k and so $B_{\epsilon/2}(x_{n_k})$ lies in U_α for all sufficiently large k, which is a contradiction as soon as $1/n_k < \epsilon/2$. □

We now resume the proof of Proposition 2.2. We must show that if X is limit point compact, then it is compact. Given an open covering $\{U_\alpha\}$, let $\delta > 0$ be a Lebesgue number for the covering. Cover X by balls of radius δ. Observe that this has a finite subcover: if not, then there is an infinite sequence in X for which the distance between any two points is at least δ, and no such sequence can have a convergent subsequence. Each of these balls is contained in some U_α, and so finitely many of these open sets U_α cover X as required. □

Corollary 2.4. *If X is a compact metric space, then any sequence in X has a convergent subsequence.*

Exercise 2.2. *Prove that the converse of this also holds.*

When X is compact, a continuous function $f : X \to Y$ satisfies a stronger property than continuity.

Proposition 2.5. *If X is compact and $f : X \to Y$ is continuous, then for all $\epsilon > 0$ there exists $\delta > 0$ so that for all $x, y \in X$ with $d(x, y) < \delta$ we have $\rho(f(x), f(y)) < \epsilon$.*

This looks like the definition of continuity, but notice the order of the quantifiers: Given $\epsilon > 0$ we find a $\delta > 0$ that works for *any* $x \in X$, whereas in the definition of continuity, the δ we find may depend on x. Indeed, when X is not compact, this dependence may be necessary: consider the example of $f : \mathbb{R} \to \mathbb{R}$, given by $f(x) = x^2$. A function satisfying the conclusion of the proposition is called *uniformly continuous*. Thus the proposition says that any continuous function defined on a compact set is uniformly continuous.

Proof. Given $\epsilon > 0$, cover Y by balls of radius $\epsilon/2$. The preimage under f gives a covering of X. Let $\delta > 0$ be the Lebesgue number of this covering. For any two points $x, y \in X$ if $d(x, y) < \delta$, then $y \in B_\delta(x)$, so $f(x)$ and $f(y)$ lie in $B_{\epsilon/2}(z)$, for some $z \in Y$. Therefore $\rho(f(x), f(y)) \leq \rho(f(x), z) + \rho(f(y), z) < \epsilon/2 + \epsilon/2 = \epsilon$. □

2.2 Connectivity

Connectivity is another property introduced by Prof. Brendle last time. For subsets of $\mathbb{R}$, this property has a nice description.

Theorem 2.6. *A subset $A \subset \mathbb{R}$ is connected if and only if it is an interval.*

Proof. If A is not an interval, then there is some $x \in \mathbb{R}$ which is not in A so that $(-\infty, x) \cap A \neq \emptyset$ and $(x, \infty) \cap A \neq \emptyset$. These intersections give a separation of A, proving it is not connected.

Conversely, consider first the interval $[a, b]$ with $a \neq b$. Suppose $U, V \subset [a, b]$ are disjoint open sets and we prove $U \cup V \neq [a, b]$. Let $x \in U$ and $y \in V$ and assume without loss of generality that $a < x < y < b$. Set $U_0 = U \cap [x, y]$ and $V_0 = V \cap [x, y]$. We claim that $z = \sup U_0 \notin U_0 \cup V_0$ which suffices since z is a limit point of U_0, hence contained in $[x, y] \subset [a, b]$ but not $U \cup V$ (since $(U \cup V) \cap [x, y] = U_0 \cup V_0$).

To prove the claim, we first observe that $x < z < y$ since U_0 is a neighborhood of x and V_0 is a neighborhood of y and $U_0 \cap V_0 = \emptyset$. Now suppose that $z \in U_0 \cup V_0$. If $z \in V_0$, then $(z - \epsilon, z + \epsilon) \subset V_0$ for some $\epsilon > 0$, and so $z - \epsilon \geq \sup U_0 = z$, a contradiction. Therefore, $z \in U_0$, and there exists $\epsilon > 0$ so that $(z - \epsilon, z + \epsilon) \subset U_0$. But then $\sup U_0 = z < z + \epsilon/2 \in U_0$, another contradiction. Therefore, $z \notin U_0 \cup V_0$, as required.

An arbitrary interval is an increasing union of intervals of the type $[a, b]$ (or is a single point), and such unions of connected sets are connected. □

A consequence of this is the intermediate value theorem from calculus.

Theorem 2.7. *If X is connected, $f : X \to \mathbb{R}$ is continuous and $x, y \in X$, then for all c in the interval between $f(x)$ and $f(y)$, there exists some $z \in X$ so that $f(z) = c$.*

Proof. The image $f(X)$ is connected, hence an interval by the previous theorem. □

Exercise 2.3. *Prove that the product of connected spaces is connected. In particular, $\mathbb{R}^n$ is connected.*

When a metric space X fails to be connected, the notion of connectivity gives rise to a useful equivalence relation on X. Given $x, y \in X$, declare them to be equivalent $x \sim y$ if there exists a connected subset of X containing both x and y. The equivalence classes of this relation are called the *connected components* of X.

Exercise 2.4. *Prove that the relation above does indeed define an equivalence relation, and that the connected components are connected subsets. Further prove that any connected subset is contained in a component.*

Proposition 2.8. *Each component of a surface S is a surface.*

Proof. Let X be a component of S and $x \in X$ an arbitrary point. A neighborhood U of x in S homeomorphic to $\mathbb{R}^2$, being a connected subset, is contained in X. Thus every point of X has a neighborhood homeomorphic to $\mathbb{R}^2$, and X is a surface (second countability of X follows from that of S). □

In general, the components of a metric space X are not necessarily open. Consider, for example, $\mathbb{Q}$ with the subspace metric inside of $\mathbb{R}$. Each one point set is a component, which is not an open set.

2.3 Path connectivity

A more geometric notion of connectivity is the following. Say that a metric space X is *path connected* if for any two points $x, y \in X$, there exists a path in X from x to y.

Proposition 2.9. *If X is path connected, then it is connected.*

Proof. Two points which are connected by a path lie in a connected subset of X (the image of the path), and are therefore in the same component. Since any two points are connected by a path, there is only one connected component to X, and therefore X is connected. □

We can similarly define *path components* of X, and an easy consequence of the previous proposition is the following.

Corollary 2.10. *Each path component of X is contained in a component of X.*

In our primary case of interest, there is no difference between connectivity and path connectivity.

Proposition 2.11. *A surface S is connected if and only if it path connected.*

Proof. According to Proposition 2.9 it suffices to prove that if S is connected, it is path connected. Since any point of S has a neighborhood homeomorphic to $\mathbb{R}^2$, which is a path connected set, the path components of a surface S are open sets. If X is one path component, and Y is the union of all remaining path components, we see that X and Y are disjoint open sets with $X \cup Y = S$. Since S is connected and $X \neq \emptyset$, we must have $Y = \emptyset$ and hence $X = S$ is path connected. □

2.4 Completeness

A sequence $\{x_n\}_{n=1}^{\infty}$ in a metric space X is called a *Cauchy sequence* if for every $\epsilon > 0$ there exists $N > 0$ so that

$$d(x_n, x_m) < \epsilon$$

for every $n, m \geq N$. A convergent sequence is a Cauchy sequence (by the triangle inequality), but not necessarily conversely. For example, in the open interval $(0, 1)$ with the subspace metric, $\{\frac{1}{n}\}$ is a nonconvergent Cauchy sequence.

A metric space X is said to be *complete* if every Cauchy sequence converges. Unlike compactness and connectivity which depend only on the topology, completeness is *not* a topological property. For example, $(0, 1)$ is homeomorphic to $\mathbb{R}$. However, in the usual metrics on these spaces, the former is incomplete, whereas the latter is complete. There is however a relation with topological properties.

Proposition 2.12. *If X is a compact metric space, then it is complete.*

Proof. Given a Cauchy sequence, $\{x_n\} \subset X$, there is a convergent subsequence $\lim_{k\to\infty} x_{n_k} = x \in X$ by Corollary 2.4. Given $\epsilon > 0$, there exists $N_1, N_2 > 0$ so that for all $n_k > N_1$, $d(x_{n_k}, x) < \epsilon/2$ and for all $n, m > N_2$, $d(x_n, x_m) < \epsilon/2$. Set $N = \max\{N_1, N_2\}$ and let $n_k > N$ be an integer with x_{n_k} in the subsequence. Then for all $n \geq N$

$$d(x_n, x) \leq d(x_n, x_{n_k}) + d(x_{n_k}, x) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

□

2.5 Uniform limits

Given a sequence of continuous maps $\{f_n : X \to Y\}$, there are various options for what it means to say that this sequence converges to a map $f : X \to Y$. We are interested in when the limit is continuous, and it turns out that a good notion of convergence is the following. Say that a sequence $\{f_n : X \to Y\}$ *converges uniformly* to $f : X \to Y$ if for all $\epsilon > 0$ there exists N so that if $n \geq N$ and $x \in X$ then

$$\rho(f(x), f_n(x)) < \epsilon.$$

Proposition 2.13. *If a sequence $\{f_n : X \to Y\}$ converges uniformly to $f : X \to Y$, then f is continuous.*

Proof. Given $\epsilon > 0$ and $x \in X$, we must find $\delta > 0$ so that $\rho(f(x), f(y)) < \epsilon$ whenever $d(x, y) < \delta$. First choose $N > 0$ so that for $n \geq N$ we have $\rho(f(z), f_n(z)) < \epsilon/3$ for all $z \in X$. Since f_N is continuous, it is uniformly continuous by Proposition 2.5, so we can choose $\delta > 0$ so that if $d(x, y) < \delta$, then $\rho(f_N(x), f_N(y)) < \epsilon/3$. Therefore, for all $y \in X$ with $d(x, y) < \delta$ we have

$$\rho(f(x), f(y)) \leq \rho(f(x), f_N(x)) + \rho(f_N(x), f_N(y)) + \rho(f_N(y), f(y)) < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$$

as required. □

It would also be useful to have criteria for when a sequence of maps converges, at least up to subsequence. This is given by the following known as the *Arzela-Ascoli Theorem.*

Theorem 2.14. *Let X, Y be a compact metric spaces. Suppose $\{f_n : X \to Y\}$ is a sequence of continuous maps with the property that for all $\epsilon > 0$ there exists δ so that $\rho(f_n(x), f_n(y)) < \epsilon$ whenever $d(x, y) < \delta$. Then there is a subsequence which converges uniformly.*

The hypothesis can be thought of as saying that the sequence of functions are uniformly, uniformly continuous: not only does δ depend only on ϵ, and not the point x, but it's also independent of the particular map in the sequence. This property is often called *equicontinuity*.

Proof sketch. Let $\{x_m\}_{m=1}^{\infty}$ be a countable dense subset of X. We first find a subsequence f_{n_k} so that $\lim_{k\to\infty} f_{n_k}(x_m)$ exists and limits to $y_m \in Y$, for all $m \in \mathbb{Z}_+$. To do this, first use compactness of Y to prove that there is a subsequence so that $\lim_{k\to\infty} f_{n_k}(x_1) = y_1$. We can pass to a further subsequence so that images of x_2 also converge, and then a further subsequence so that the images of x_3 converge, etc. Continuing in this way, passing to further and further subsequences, we could hope to find a subsequence on which the images of all x_m converge. To guarantee that we do not "thin the sequence out too much", we assume that when we choose the j^{th} subsequence we keep the first $j-1$ terms. So far, this has not used the hypothesis of the theorem, other than compactness of X and Y.

Now we claim that all the set of all sequences

$$\{\{f_{n_k}(x)\}_{k=1}^{\infty} \mid x \in X\}$$

are *uniformly Cauchy*, meaning that given $\epsilon > 0$ there exists $N > 0$ so that for all $n_k, n_{k'} \geq N$ we have

$$\rho(f_{n_k}(x), f_{n_{k'}}(x)) < \epsilon$$

for every x. Assuming this, Proposition 2.12 provides a limiting function $f(x) = \lim_{k\to\infty} f_{n_k}(x)$. Moreover, because these are uniformly Cauchy, the proof of that proposition actually shows that f_{n_k} converges uniformly to f.

Given $\epsilon > 0$, let $\delta > 0$ be such that for any n, $\rho(f_n(x), f_n(y)) < \epsilon/3$ whenever $d(x, y) < \delta/2$. Further choose $m > 0$ so that every point of X is within $\delta/2$ of some x_j, for $j = 1, \ldots, m$. Finally, choose N so that if $n_k, n_{k'} \geq N$ we have $\rho(f_{n_k}(x_j), f_{n_{k'}}(x_j))$ for each $j = 1, \ldots, m$. Then given x, let j be such that $d(x, x_j) < \delta/2$, and so

$$\begin{aligned} \rho(f_{n_k}(x), f_{n_{k'}}(x)) &\leq \rho(f_{n_k}(x), f_{n_k}(x_j)) + \rho(f_{n_k}(x_j), f_{n_{k'}}(x_j)) + \rho(f_{n_{k'}}(x_j), f_{n_{k'}}(x)) \\ &\leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon \end{aligned}$$

as required. □

As an application of Theorem 2.14, we have the following.

Exercise 2.5. *Prove that if X is a compact path metric space, then X is a geodesic metric space. For this, you should prove that a sequence of paths connecting x to y whose lengths are converging to $d(x, y)$ limits to a geodesic from x to y, after possibly passing to a subsequence.*

We say X is a *proper metric space* if closed balls $\overline{B}_r(x) = \{y \in X \mid d(x, y) \leq r\}$ are compact for all $r > 0$. These are precisely the metric spaces for which a Heine-Borel-type Theorem holds: compact subsets are precisely the closed and bounded subsets.

Exercise 2.6. *Prove that if X is a proper path metric space, then X is a geodesic metric space.*

3 Riemannian metric spaces

3.1 Riemannian metrics

Let $V \subset \mathbb{R}^2$ be an open set. A *Riemannian structure* on V is a continuous choice of inner product on each of the tangent spaces. We think of this as function from V to the space of inner products on $\mathbb{R}^2$, $\mathbf{x} \mapsto g^{\mathbf{x}}$, where

$$g^{\mathbf{x}} : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$$

is an inner product for each $\mathbf{x} \in V$. Since an inner product is determined by the values on a basis, $g^{\mathbf{x}}$ is determined by the 2×2 matrix

$$(g_{ij}(\mathbf{x})) = \begin{pmatrix} g_{11}(\mathbf{x}) & g_{12}(\mathbf{x}) \\ g_{21}(\mathbf{x}) & g_{22}(\mathbf{x}) \end{pmatrix}$$

where $g_{ij}(\mathbf{x}) = g^{\mathbf{x}}(e_i, e_j)$ for the standard basis $\mathbf{e}_1, \mathbf{e}_2$ on $\mathbb{R}^2$. The continuity can be expressed by requiring the four functions $\mathbf{x} \mapsto g_{ij}(\mathbf{x})$ to be continuous functions on V (symmetry of the inner product means $g_{ij}(\mathbf{x}) = g_{ji}(\mathbf{x})$, so there are really only 3 distinct functions).

Suppose we have a Riemannian structure g on an open set $V \subset \mathbb{R}^2$, we define the g–length of a piecewise C^1 path $\gamma : [a, b] \to V$ to be

$$\ell_g(\gamma) = \int_a^b |\gamma'(t)|_g dt.$$

Here, $|\gamma'(t)|_g$ is shorthand for the norm of $\gamma'(t)$ induced by g at the point $\gamma(t)$

$$|\gamma'(t)|_g = \sqrt{g^{\gamma(t)}(\gamma'(t), \gamma'(t))}.$$

If V is connected (hence path connected by Proposition 2.9) we can define

$$d_g(\mathbf{x}, \mathbf{y}) = \inf \ell_g(\gamma)$$

where γ ranges over all piecewise C^1 paths (an easy exercise proves path connectivity of V implies piecewise C^1 path connectivity).

Proposition 3.1. *The function d_g is a path metric on V defining the standard topology.*

Proof. Since a piecewise C^1 path from $\mathbf{x}$ to $\mathbf{y}$ can be reversed to give a piecewise C^1 path from $\mathbf{y}$ to $\mathbf{x}$ of the same length, we see that d_g is symmetric. Also, a piecewise C^1 path from $\mathbf{x}$ to $\mathbf{z}$ and one from $\mathbf{z}$ to $\mathbf{y}$ can be concatenated to give a piecewise C^1 path from $\mathbf{x}$ to $\mathbf{y}$, and so the triangle inequality also follows easily for d_g. Furthermore, since d_g is defined as an infimum of nonnegative numbers, it is a nonnegative number. To prove that d_g is a metric, the only thing that remains is to prove $d_g(\mathbf{x}, \mathbf{y}) = 0$ implies $\mathbf{x} = \mathbf{y}$.

Consider the following function on $V \times \mathbb{S}^1$ given by

$$K(\mathbf{x}, \cos(\theta)\mathbf{e}_1 + \sin(\theta)\mathbf{e}_2) = |\cos(\theta)\mathbf{e}_1 + \sin(\theta)\mathbf{e}_2|_{g(\mathbf{x})}.$$

As this is constructed from continuous functions, it is continuous (check this). Given $\mathbf{x} \in V$, let $\epsilon > 0$ be such that the Euclidean closed ball $\overline{B}_\epsilon(\mathbf{x})$ is contained in V. Observe that $\overline{B}_\epsilon(\mathbf{x}) \times \mathbb{S}^1$ is compact, and hence K attains a minimum and maximum value, K_0, K_1, respectively, and so

$$0 < K_0 \leq K(\mathbf{x}, \cos(\theta)\mathbf{e}_1 + \sin(\theta)\mathbf{e}_2) \leq K_1.$$

This implies that for any tangent vector $\mathbf{v}$ based at any point $\mathbf{y} \in \overline{B}_\epsilon(\mathbf{x})$ we have

$$K_0|\mathbf{v}| \leq |\mathbf{v}|_g \leq K_1|\mathbf{v}|.$$

From this we see that for any C^1 path $\gamma : [a,b] \to B_\epsilon(\mathbf{x})$

$$K_0\ell(\gamma) \leq \ell_g(\gamma) \leq K_1\ell(\gamma)$$

where $\ell(\gamma)$ is the Euclidean length of γ and hence by Theorem 1.1

$$K_0|\mathbf{x} - \mathbf{y}| \leq d_g(\mathbf{y}, \mathbf{x}) \leq K_1|\mathbf{x} - \mathbf{y}| \tag{2}$$

for all $\mathbf{y} \in B_\epsilon(\mathbf{x})$. On the other hand, if $\mathbf{y} \in V - B_\epsilon(\mathbf{x})$ then

$$d_g(\mathbf{y}, \mathbf{x}) \geq K_0\epsilon$$

since any path from $\mathbf{x}$ to $\mathbf{y}$ must pass through the boundary of $\overline{B}_\epsilon(\mathbf{x})$, so must have length at least $K_0\epsilon$. In particular, $d_g(\mathbf{x}, \mathbf{y}) > 0$ unless $\mathbf{y} = \mathbf{x}$, as required, so that d_g is a metric. By (2), this defines the standard topology.

By construction, the distance is the infimum of lengths of C^1 paths, and since the distance between a pair of points is always no greater than the length of a path between the points, it follows that d_g is a path metric. □

We call d_g the *Riemannian metric* determined by the Riemannian structure. We note that what we are calling a Riemannian structure is often called a Riemannian metric in the literature, and the function d_g is simply called the associated distance function. To avoid the confusion, we will stick with the terminology we have chosen.

We write $B^g_\epsilon(\mathbf{x})$ for the ϵ–ball in the d_g metric, or simply $B_\epsilon(\mathbf{x})$ when no confusion can arise.

Suppose that $U, V \subset \mathbb{R}^2$ are open sets and $F : U \to V$ is a C^1 map with C^1 inverse (so the derivative at every point is nonsingular by the chain rule). Given a Riemannian structure g on V, we can define a Riemannian structure h on U defined by

$$h(\mathbf{u_x}, \mathbf{v_x}) = g(dF_\mathbf{x}(\mathbf{u_x}), dF_\mathbf{x}(\mathbf{v_x})).$$

The standard notation for h is $h = F^*(g)$.

Exercise 3.1. *Prove that the map $F : U \to V$ is an isometry with respect to d_h and d_g.*

Given (U, h) and $\mathbf{x} \in U$, construct a C^1 map $F : U \to V$ for some $V \subset \mathbb{R}^2$, with C^1 inverse, and Riemannian structure g on V so that $F^*(g) = h$, $F(\mathbf{x}) = \mathbf{0}$, and $g^\mathbf{0} = \mathrm{std}$, the standard inner product.

Given a Riemannian structure g on V, we can make sense of the angle between a pair of C^1 curves which intersect at a point $\mathbf{x}$ using the inner product $g^\mathbf{x}$. We can also compute the area of a region $W \subset V$ by

$$\mathrm{Area}_g(W) = \iint_W \sqrt{\det(g_{ij}(x,y))}dxdy.$$

Exercise 3.2. *Prove that for an isometry as in the previous exercise, the notions of angle and area are preserved. That is*

$$\mathrm{Area}_{F^*(g)}(W) = \mathrm{Area}_g(F(W)),$$

and the $F^*(g)$–angle between γ and σ in U is equal to the g–angle between $F(\gamma)$ and $F(\sigma)$ (at the corresponding intersection point).

Exercise 3.3. *To motivate the definition of area, suppose that g is a Riemannian structure on U so that $g^{\mathbf{x}} = \mathrm{std}$ the standard inner product at some $\mathbf{x} \in U$ (compare the previous two exercises to reduce to this case). Prove that if $B_\epsilon(\mathbf{x})$ denotes the Euclidean ball then*

$$\lim_{\epsilon \to 0} \frac{\mathrm{Area}_g(B_\epsilon(\mathbf{x}))}{\pi \epsilon^2} = 1.$$

The point here is that if a Riemannian structure agrees with the standard one to first order, then the areas are the same to first order.

3.2 Examples

Of course, the Euclidean plane $\mathbb{R}^2$ with the standard inner product std (at every point) is an example of a Riemannian structure on $\mathbb{R}^2$. Given a continuous positive function on an open set $f : U \to \mathbb{R}$, we can modify the standard inner product by multiplying it by f to produce a Riemannian structure g_f given by

$$g_f(\mathbf{u}_\mathbf{x}, \mathbf{v}_\mathbf{x}) = f(\mathbf{x})\mathrm{std}(\mathbf{u}, \mathbf{v}).$$

A particular instance of this is the *upper half plane model of the hyperbolic plane*:

$$\mathbb{U}^2 = \{(x, y) \in \mathbb{R}^2 \mid y > 0\} \text{ with Riemannian structure } g_\mathbb{U} := g_{1/y^2} = \frac{1}{y^2}\mathrm{std}.$$

Another example is the *disk model of the hyperbolic plane*:

$$\mathbb{D}^2 = \{\mathbf{x} \in \mathbb{R}^2 \mid |\mathbf{x}| < 1\} \text{ with Riemannian structure } g_\mathbb{D} := g_{(4/(1-|\mathbf{x}|^2)^2)} = \frac{4}{(1 - |\mathbf{x}|^2)^2}\mathrm{std}.$$

Whenever we write $\mathbb{U}^2$ or $\mathbb{D}^2$, it will be assumed that they come equipped with these Riemannian structures $g_\mathbb{U}$ and $g_\mathbb{D}$, respectively. We denote the associated metrics by $d_\mathbb{U}$ and $d_\mathbb{D}$, respectively.

Exercise 3.4. *Prove that the map $F : \mathbb{D}^2 \to \mathbb{U}^2$ given by*

$$F(x, y) = \left(\frac{2x}{x^2 + (y-1)^2}, \frac{1 - x^2 - y^2}{x^2 + (y-1)^2} \right)$$

is an isometry. (Give this computation a try, but we'll see a simpler proof using the complex numbers next time.)

3.3 Riemannian surfaces

A surface S with a metric d which is a path metric when restricted to each component is called a *Riemannian surface* (not to be confused with a Riemann surface, which we will mention later) if for every $\mathbf{x} \in S$, there exists $\epsilon > 0$ so that $B_\epsilon(\mathbf{x})$ in (S, d) is isometric to some (U, d_g) where $U \subset \mathbb{R}^2$ and g is a Riemannian structure on U. We express this by saying that S is *locally isometric to a Riemannian metric.* Any open subset of $\mathbb{R}^2$ with a Riemannian metric is an example of a Riemannian surface.

Consider a connected surface $S \subset \mathbb{R}^n$ that is C^1*–embedded*, meaning that every point $\mathbf{x}$ has a neighborhood U which is the image of a C^1 map $F : V \to U$ defined on an open subset $V \subset \mathbb{R}^2$ with the property that $dF_\mathbf{z}$ is nonsingular at every point $\mathbf{z} \in V$. The distance between two points can be defined as the infimum of lengths of paths between the points, and the C^1 assumption means that any two points can be connected by a piecewise C^1 path, so the distance between any two points is finite. More importantly, the C^1 assumption allows us to see that this metric makes S into a Riemannian surface. Indeed, if $F : V \to U$ is the C^1 map described above mapping to a neighborhood U of $\mathbf{x}$, then defining

$$g(\mathbf{u_z}, \mathbf{v_z}) = \mathrm{std}(dF_\mathbf{z}(\mathbf{u_z}), dF_\mathbf{z}(\mathbf{v_z}))$$

it follows that for a sufficiently small $\epsilon > 0$, $F^{-1} : B_\epsilon(\mathbf{x}) \to B_\epsilon(F^{-1}(\mathbf{x})) \subset V$ is an isometry (where we are using d_g on V). For convenience we denote g here also as $F^*(\mathrm{std})$ because of the similarity with the Riemannian structure defined in the analogous way earlier.

An example is $\mathbb{S}^2 \subset \mathbb{R}^3$. We leave it as an exercise to verify that this is C^1 embedded (hint: cover it by graphs of C^1 functions).

Exercise 3.5. *A* great circle *in $\mathbb{S}^2$ is the intersection of $\mathbb{S}^2$ with a plane in $\mathbb{R}^3$ through the origin. Prove that any two points $\mathbf{x}, \mathbf{y} \in \mathbb{S}^2$ lie on at least one great circle, and exactly one if and only if the points are* not *antipodal, $\mathbf{x} \neq -\mathbf{y}$. Prove that the shorter arc of a great circle containing $\mathbf{x}, \mathbf{y} \in \mathbb{S}^2$ is a geodesic between $\mathbf{x}, \mathbf{y}$, and if $\mathbf{x} \neq -\mathbf{y}$, then this is the unique geodesic. Hint: look at proof # 1 of Theorem 1.1 and define an appropriate projection* locally.

Another example is a *flat torus*: $\mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1 \subset \mathbb{R}^2 \times \mathbb{R}^2 = \mathbb{R}^4$. To see that this is C^1 embedded, consider the C^1 map

$$F(x, y) = (\cos(x), \sin(x), \cos(y), \sin(y)).$$

The restriction of F to any open square $(a, a + 2\pi) \times (b, b + 2\pi) \subset \mathbb{R}^2$ provides the required maps to see that $\mathbb{T}^2$ is C^1 embedded. We observe that in fact, $F^*(\mathrm{std}) = \mathrm{std}$. That is, the Riemannian surface $\mathbb{T}^2$ is locally isometric to $\mathbb{R}^2$.

A modification of these examples is to take any symmetric bilinear function

$$B : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

and a C^1 embedded connected surface $S \subset \mathbb{R}^n$ with the property that for any C^1 path $\gamma : [a, b] \to S$, $B(\gamma'(t), \gamma'(t)) \neq 0$ whenever $\gamma'(t) \neq 0$. This is equivalent to saying that for any $F : V \to U$ which is a C^1 map onto an open subset $U \subset S$ with nonsingular derivative, then the restriction of B to the image of $dF_\mathbf{x}$ is an inner product for any $\mathbf{x} \in U$.

An important example of this for us is the bilinear function

$$B_H((x_0, x_1, x_2), (y_0, y_1, y_2)) = -x_0y_0 + x_1y_1 + x_2y_2$$

on $\mathbb{R}^3$. This is important enough that we will refer to $\mathbb{R}^3$ with this bilinear function as $\mathbb{R}^{2,1}$. The subset of $\mathbb{R}^{2,1}$ defined by

$$\{\mathbf{x} \in \mathbb{R}^{2,1} \mid B(\mathbf{x}, \mathbf{x}) = -1\}$$

is a hyperboloid of 2 sheets. Let $\mathcal{H}^2$ denote the sheet of this hyperboloid containing $\mathbf{e}_0 = (1, 0, 0)$.

Exercise 3.6. *Prove that B defines a Riemannian metric on $\mathcal{H}^2$ as per the discussion above. As we will see later, this is isometric to $\mathbb{U}^2$ (and hence also $\mathbb{D}^2$), and so provides another model for the hyperbolic plane — the* hyperboloid model.

Remark. This notion of Riemannian surface agrees with the usual notion defined in terms of *differentiable (or smooth) structures* on surfaces. We have opted for the given definition to avoid having to develop the machinery of differentiable structures since this is not necessary for our purposes.

4 Homogeneous surfaces

4.1 Isometry groups

We now narrow our focus to a particular class of Riemannian surfaces. To describe these, we recall three particular Riemannian surfaces, namely the Euclidean plane $\mathbb{R}^2$, the 2–sphere $\mathbb{S}^2$, and the hyperbolic plane $\mathbb{H}^2$ (concretely described as either $\mathbb{U}^2$ or $\mathbb{D}^2$, or as we will see $\mathcal{H}^2$). Given a metric surface S, let $\mathrm{Isom}(S)$ denote the group of all isometries of S. A metric surface is *homogeneous* if $\mathrm{Isom}(S)$ acts transitively on S.

Theorem 4.1. *For $\mathbb{X}^2 = \mathbb{S}^2$, $\mathbb{R}^2$ or $\mathbb{H}^2$, the isometry group $\mathrm{Isom}(\mathbb{X}^2)$ acts transitively on $\mathbb{X}^2$ — that is, $\mathbb{X}^2$ is homogeneous. Moreover, the stabilizer of any point $\mathbf{x} \in \mathbb{X}^2$ is isomorphic to the group*

$$O(2) = \{A \in M_{2\times 2}(\mathbb{R}) \mid A^T A = I\}$$

where $M_{2\times 2}(\mathbb{R})$ is the set of all 2×2 matrices with real entries and I is the 2×2 identity matrix.

In fact, we will compute the isometry groups explicitly.

We begin with the proof for $\mathbb{R}^2$ as this is illustrates the main idea. The proof for $\mathbb{S}^2$ is left as an exercise and the proof for $\mathbb{H}^2$ requires a bit more development.

Proof of Theorem 4.1 for $\mathbb{R}^2$. We claim that

$$G = \{F(\mathbf{x}) = A\mathbf{x} + \mathbf{b} \mid A \in O(2) \text{ and } \mathbf{b}\}$$

is the full group $\mathrm{Isom}(\mathbb{R}^2)$.

First observe that $G < \mathrm{Isom}(\mathbb{R}^2)$. That is, every element of G is an isometry. This follows from the fact that $O(2)$ is precisely the set of 2×2 matrices A which preserve the standard inner product

$$\mathrm{std}(A\mathbf{x}, A\mathbf{x}) = \mathrm{std}(\mathbf{x}, \mathbf{x}),$$

and so $\mathbf{x} \mapsto A\mathbf{x}$ preserves the Euclidean distance for all $A \in O(2)$. Since the translations $\mathbf{x} \mapsto \mathbf{x} + \mathbf{b}$ are clearly isometries, every element of G is an isometry.

Next observe that the (normal) subgroup of G consisting of the translations $F(\mathbf{x}) = \mathbf{x} + \mathbf{b}$ acts transitively on $\mathbb{R}^2$, and the stabilizer of $\mathbf{0}$ in G is exactly $O(2)$. Thus, to prove the theorem, we need only prove that $G = \mathrm{Isom}(\mathbb{R}^2)$.

Given an arbitrary isometry $f : \mathbb{R}^2 \to \mathbb{R}^2$ we need to prove that $f \in G$. Since G acts transitively, there exists an element $F_1 \in G$ so that $F_1 \circ f(\mathbf{0}) = \mathbf{0}$. The biinfinite geodesics through $\mathbf{0}$ are precisely the straight lines through $\mathbf{0}$. Since an isometry must take geodesics to geodesics, we see that $F_1 \circ G$ maps all straight lines through $\mathbf{0}$ isometrically to straight lines through $\mathbf{0}$. Since $O(2)$ acts transitively on these lines, we can find $F_2 \in O(2) < G$ so that $F_2 \circ F_1 \circ f$ maps the x–axis to itself and sends the upper half plane to itself (the latter is achieved by further composing with a reflection in the x–axis, also in $O(2)$, if necessary). Since $\mathbb{S}^1$ is precisely the set of points a distance 1 from $\mathbf{0}$ it follows that $F_2 \circ F_1 \circ f$ maps $\mathbb{S}^1$ to itself, fixes the points $(1, 0)$ and $(-1, 0)$, and leaves the upper semi-circle and lower semi-circle invariant. Because any point on the upper semicircle (or lower semicircle) is determined by its distance to $(1, 0)$ and $(-1, 0)$, we see that this must be the identity on $\mathbb{S}^1$. But then $F_2 \circ F_1 \circ f$ is the identity on $\mathbb{R}^2$ and hence $f = (F_2 \circ F_1)^{-1} \in G$ as required. □

Exercise 4.1. *Prove that $O(3)$, the group of 3×3 matrices A with real entries such that $A^T A = I$, preserves $\mathbb{S}^2$ and acts transitively. Further prove that this is precisely the isometry group,* $\mathrm{Isom}(\mathbb{S}^2) = O(3)$*. Finally, prove that the stabilizer of $\mathbf{e}_3$, the third standard basis vector, is isomorphic to $O(2)$.*

4.2 Hyperbolic geometry via complex numbers

The complex numbers $\mathbb{C}$ can be naturally thought of as a plane

$$\mathbb{C} = \{x + iy \mid x, y \in \mathbb{R}\} \leftrightarrow \mathbb{R}^2 = \{(x, y) \mid x, y \in \mathbb{R}\},$$

and we often refer to $\mathbb{C}$ as the *complex plane*. We consider the metric on $\mathbb{C}$ obtained via this bijection, and so for $z \in \mathbb{C}$ the *absolute value* $|z|$ is precisely the Euclidean norm. The norm can be described in terms of complex conjugation $z = x + iy \mapsto \overline{z} = x - iy$ as $|z|^2 = z\overline{z}$.

Given an open subset $U \subset \mathbb{C}$ and a function $f : U \to \mathbb{C}$, we say that f is *holomorphic* if the complex derivative

$$f'(z) = \lim_{h \to 0} \frac{f(z + h) - f(z)}{h}$$

exists at every $z \in U$, and f' is continuous. We can naturally view f as a function from an open set in $\mathbb{R}^2$ to $\mathbb{R}^2$ writing

$$f(x + iy) = u(x, y) + iv(x, y),$$

this is $(x, y) \mapsto (u(x, y), v(x, y))$. Since we already have a notion of differentiability for such maps, we may wonder how this is related to the requirement that f is holomorphic.

Exercise 4.2. *Prove that f is holomorphic if and only if u and v are C^1 and the partial derivative satisfy the* Cauchy-Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \text{ and } \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}.$$

Further, prove that the derivative df_z, thought of as a map from the tangent space at z to $\mathbb{C}$ to the tangent space of $\mathbb{C}$ at $f(z)$, viewed as 1–dimensional complex vector spaces rather than 2–dimensional real vector spaces, is $\mathbb{C}$–linear, and in fact is given by multiplication by $f'(z)$, and so preserves angles.

The rules for computing derivatives of functions $\mathbb{R} \to \mathbb{R}$ are equally valid for functions $\mathbb{C} \to \mathbb{C}$. For example, consider the function

$$f(z) = \frac{az + b}{cz + d}$$

for $a, b, c, d \in \mathbb{C}$ with $ad - bc \neq 0$, which is defined on all of $\mathbb{C}$ when $c = 0$, and $\mathbb{C} - \{d/c\}$ when $c \neq 0$. We can compute

$$f'(z) = \frac{a(cz + d) - c(az + b)}{(cz + d)^2} = \frac{ad - bc}{(cz + d)^2}. \tag{3}$$

So the assumption that $ad - bc \neq 0$ is precisely the statement that the derivative is nonzero everywhere f is defined.

Example. We note that the isometry $F : \mathbb{D}^2 \to \mathbb{U}^2$ is more naturally expressed in these terms as

$$F(z) = \frac{-iz+1}{z-i} = \frac{-iz+1}{z-i}\frac{\overline{z-i}}{\overline{z-i}} = \frac{z+\overline{z}}{|z-i|^2} + i\left(\frac{1-|z|^2}{|z-i|^2}\right) = u(z) + iv(z).$$

Then we compute that

$$|F'(z)|\frac{1}{v(z)} = \left|\frac{-2}{(z-i)^2}\right|\frac{|z-i|^2}{1-|z|^2} = \frac{2}{1-|z|^2}$$

so F is indeed an isometry with respect to $g^z_{\mathbb{D}} = (\frac{2}{1-|z|^2})^2\text{std}$ and $g^{u+iv}_{\mathbb{U}} = (\frac{1}{v})^2\text{std}$.

These functions form an important class of functions called *Möbius transformations*. In fact, the Möbius transformations form a group which is a homomorphic quotient of

$$\mathrm{GL}_2(\mathbb{C}) = \{A \in M_{2\times 2}(\mathbb{C}) \mid \det(A) \neq 0\}$$

by the map

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \left(f_A(z) = \frac{az+b}{cz+d}\right).$$

Exercise 4.3. *Prove that this is indeed a surjective homomorphism and that the kernel is precisely the matrices of the form*

$$\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}.$$

Prove that the restriction of this homomorphism to $\mathrm{SL}_2(\mathbb{C})$ *(the matrices with determinant* 1*) is also surjective and that the kernel has order* 2 *generated by* $-I$*. We accordingly denote the group of Möbius transformations by* $\mathrm{PSL}_2(\mathbb{C})$.

The fact that Möbius transformations are not defined on all of $\mathbb{C}$ is a little annoying, and we fix this as follows. Define a map $\sigma : \mathbb{C} \to \mathbb{S}^2$ called *stereographic projection* as follows. View $\mathbb{S}^2 \subset \mathbb{R}^3 = \mathbb{C} \times \mathbb{R}$ and let $N \in \mathbb{S}^2$ be the "north pole", $N = (0,1) \in \mathbb{C} \times \mathbb{R}$. For any point $z = (z,0) \in \mathbb{C} \times \{0\}$ we can draw the straight line through z and N. This will meet $\mathbb{S}^2$ in exactly one point other than N. Define $\sigma(z)$ to be this point.

Exercise 4.4. *Prove that* σ *is a homeomorphism from* $\mathbb{C}$ *to* $\mathbb{S}^2 - \{N\}$*. Write down an expression for* $\sigma(z)$ *in coordinates showing that it is in fact* C^1*. Prove that* $d\sigma_z$ *preserves angles for every* $z \in \mathbb{C}$.

Consider the disjoint union $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, and define a metric on $\widehat{\mathbb{C}}$ so that the bijection $\sigma : \widehat{\mathbb{C}} \to \mathbb{S}^2$ extending σ by $\sigma(\infty) = N$, is an isometry. By the previous exercise, $\mathbb{C} \subset \widehat{\mathbb{C}}$ as a subspace has the same topology as the standard topology on $\mathbb{C}$. Then every Möbius transformation $f(z) = \frac{az+b}{cz+d}$ extends to a homeomorphism $f : \widehat{\mathbb{C}} \to \widehat{\mathbb{C}}$ by defining

$$f(\infty) = \begin{cases} \infty & \text{if } c = 0 \\ a/c & \text{if } c \neq 0 \end{cases}$$

and

$$f(d/c) = \infty \text{ if } c \neq 0.$$

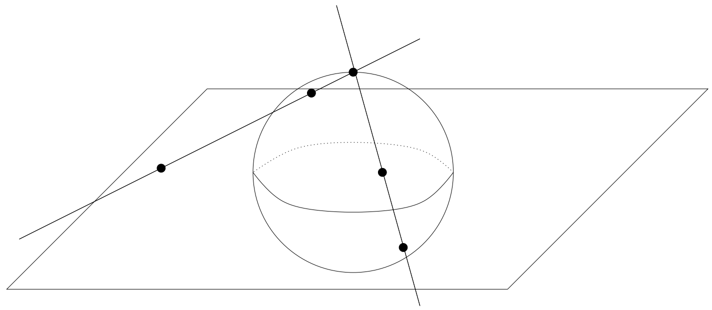

Figure 1: Stereographic projection.

Exercise 4.5. *Lines in* $\mathbb{C}$ *determine simple closed curves in* $\widehat{\mathbb{C}}$*, that is, the line union* $\{\infty\}$ *is a simple closed curve in* $\widehat{\mathbb{C}}$*. Let* $\mathcal{C}$ *denote the set of all such simple closed curves, together with the set of simple closed curves which are round circles in* $\mathbb{C}$*. We refer to* $\mathcal{C}$ *as the set of* round circles in $\widehat{\mathbb{C}}$*. Prove that* $\mathrm{PSL}_2(\mathbb{C})$ *acts on* $\mathcal{C}$*, and that the action is transitive.*

The subgroup consisting of Möbius transformations that preserve the upper half plane $\mathbb{U}^2$ is precisely the group

$$\mathrm{PSL}_2(\mathbb{R}) = \{f_A(z) \mid A \in \mathrm{SL}_2(\mathbb{R})\}$$

and the homomorphism $\mathrm{SL}_2(\mathbb{R}) \to \mathrm{PSL}_2(\mathbb{R})$ has kernel $\{\pm I\}$. The computation in (3) shows that each f_A is an isometry, so $\mathrm{PSL}_2(\mathbb{R}) < \mathrm{Isom}(\mathbb{U}^2)$.

Observe that for any $u + iv \in \mathbb{U}^2$, the element $f(z) = vz + u \in \mathrm{PSL}_2(\mathbb{R})$ sends i to $u + iv$, so $\mathrm{PSL}_2(\mathbb{R})$ acts transitively on $\mathbb{U}^2$. Furthermore, to decide when $f_A(i) = i$, with

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

we observe that

$$\frac{ai+b}{ci+d} = i \Leftrightarrow ai + b = -c + di \Leftrightarrow A = \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \in SO(2)$$

where $SO(2) < O(2)$ is the index 2 subgroup consisting of all rotations. Since $SO(2)/\{\pm I\} \cong SO(2)$, it follows that $\mathrm{PSL}_2(\mathbb{R})$ acts transitively by isometries on $\mathbb{U}^2$ and that the stabilizer of $i \in \mathbb{U}^2$ is isomorphic to $SO(2)$.

We can get a larger group of isometries by including maps of the form

$$z \mapsto \frac{a\overline{z} + b}{c\overline{z} + d}$$

where $ad - bc = -1$. We denote the *set* of all such maps by $\mathrm{PSL}_2^-(\mathbb{R})$. Any element of $\mathrm{PSL}_2^-(\mathbb{R})$ such map can be expressed as the composition of an element of $\mathrm{PSL}_2(\mathbb{R})$ with the map $z \mapsto -\overline{z}$, which is obviously an isometry.

Let $\mathrm{PSL}_2^\pm(\mathbb{R}) = \mathrm{PSL}_2(\mathbb{R}) \cup \mathrm{PSL}_2^-(\mathbb{R})$, and we observe that if $\mathrm{SL}_2^\pm(\mathbb{R})$ is the group of 2×2 matrices with real entries and determinant ± 1, then the obvious map $\mathrm{SL}_2^\pm(\mathbb{R}) \to \mathrm{PSL}_2^\pm(\mathbb{R})$ is a surjective homomorphism with kernel $\{\pm I\}$.

The isometry $F : \mathbb{D}^2 \to \mathbb{U}^2$ extends to a homeomorphism $F : \mathbb{D}^2 \cup \mathbb{S}^1 \to \mathbb{U}^2 \cup \widehat{\mathbb{R}}$, where $\widehat{\mathbb{R}} = \mathbb{R} \cup \{\infty\} \subset \widehat{\mathbb{C}}$ (indeed, it extends to all of $\widehat{\mathbb{C}}$). In the latter case, observe that $\mathrm{PSL}_2^{\pm}(\mathbb{R})$ preserves $\widehat{\mathbb{R}}$, and so can be thought of as a group of homeomorphisms of

$$\overline{\mathbb{U}}^2 = \mathbb{U}^2 \cup \widehat{\mathbb{R}}.$$

Conjugating by the extension of F, we get a group of homeomorphisms of $\overline{\mathbb{D}}^2 = \mathbb{D}^2 \cup \mathbb{S}^1$ which acts by isometries on $\mathbb{D}^2$.

Using this fairly large group of isometries, we can determine what the geodesics are.

Proposition 4.2. *The geodesics in $\mathbb{U}^2$ are precisely arcs of lines and circle orthogonal to $\widehat{\mathbb{R}}$ contained in $\mathbb{U}^2$. Similarly, the geodesics in $\mathbb{D}^2$ are precisely the arcs of lines and circles orthogonal $\mathbb{S}^1$ contained in $\mathbb{D}^2$.*

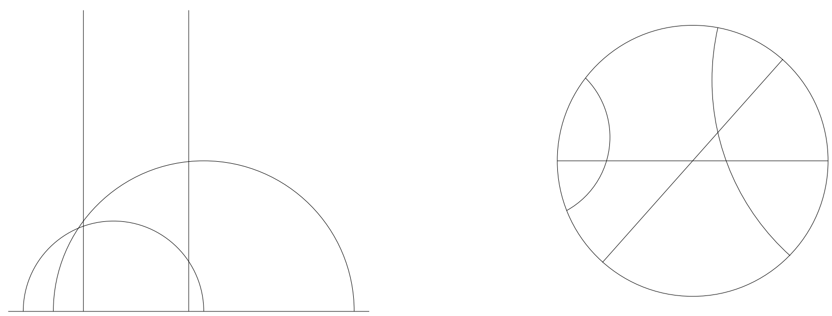

Figure 2: Some geodesics in $\mathbb{U}^2$ and $\mathbb{D}^2$.

Proof. This follows the same idea as proof # 1 of Theorem 1.1, after suitably normalizing. First, let $\mathcal{G}$ denote the set of intersections with $\mathbb{U}^2$ of the set of lines and circles in $\mathbb{C}$ which are orthogonal to $\widehat{\mathbb{R}}$. Because $\mathrm{PSL}_2(\mathbb{R})$ preserves $\mathcal{C}$ by Exercise 4.5, $\widehat{\mathbb{R}}$, and preserves angles by Exercise 4.2, we see that $\mathcal{G}$ is preserved by $\mathrm{PSL}_2(\mathbb{R})$. Furthermore, observe that the elements in $\mathcal{G}$ are completely determined by the intersections of their closures in $\overline{\mathbb{U}}^2$ with $\widehat{\mathbb{R}}$, which is a pair of points.

An explicit computation shows that $\mathrm{PSL}_2(\mathbb{R})$ acts transitively on pairs of points in $\widehat{\mathbb{R}}$—you only need to show that $0, 1$ can be taken to any two points. Therefore, it suffices to prove that any arc of

$$i\mathbb{R}_+ = \{iy \mid y > 0\} \in \mathcal{G}$$

is a geodesic. For this, we mimic proof # 1 of Theorem 1.1 using the projection

$$\pi : \mathbb{U}^2 \to i\mathbb{R}_+$$

given by $\pi(x + iy) = iy$. Since this has

$$|d\pi_z(\mathbf{v}_z)|_{g_{\mathbb{U}}} \leq |\mathbf{v}_z|_{g_{\mathbb{U}}}$$

with equality if and only if $\mathbf{v}$ is a multiple of $\mathbf{e}_2$, the proof can be carried out with only minor changes.

Therefore, the arcs of elements of $\mathcal{G}$ are exactly the set of geodesics in $\mathbb{U}^2$. The statement for $\mathbb{D}^2$ follows from this by applying F. □

As with $\mathbb{R}^2$, any two points in $\mathbb{U}^2$ are connected by a unique geodesic.

We now have the following.

Proposition 4.3. *The isometry group of* $\mathbb{U}^2$ *is precisely* $\mathrm{PSL}_2^{\pm}(\mathbb{R})$.

Proof. We follows the same strategy as for $\mathbb{R}^2$. Because $\mathrm{PSL}_2^{\pm}(\mathbb{R})$ acts transitively on $\mathcal{G}$, given any isometry $f : \mathbb{U}^2 \to \mathbb{U}^2$, we can compose with some $F \in \mathrm{PSL}_2^{\pm}(\mathbb{R})$ so that $F \circ f$ leaves $i\mathbb{R}_+$ invariant. Further composing with an element of the form $(z \mapsto tz) \in \mathrm{PSL}_2(\mathbb{R})$ (and renaming F), we can assume $i\mathbb{R}_+$ is fixed pointwise by $F \circ f$, and that $F \circ f$ preserves the set of all geodesics through i. Finally, modifying F by composing with $z \mapsto -\overline{z}$ if necessary, we can further assume that $F \circ f$ sends the right half plane ($x + iy \in \mathbb{U}^2$ with $x > 0$) to the right half plane. Inspecting the behavior of $F \circ f$ on the set of points a distance 1 from i (which is a circle), we see that $F \circ f$ must fix this circle pointwise, and hence be the identity on all of $\mathbb{U}^2$. So, $f = F^{-1} \in \mathrm{PSL}_2^{\pm}(\mathbb{R})$, as required. □

From the exercises/discussion above, Theorem 4.1 for $\mathbb{U}^2$ now follows.

5 More hyperbolic geometry

5.1 Geodesics and isometries

By the *hyperbolic plane* $\mathbb{H}^2$ we will mean either $\mathbb{U}^2$ or $\mathbb{D}^2$. We let S^1_∞ denote the *circle at infinity*, which is $\widehat{\mathbb{R}}$ in the case of $\mathbb{U}^2$ or $\mathbb{S}^1$ in the case of $\mathbb{D}^2$. We let $\overline{\mathbb{H}}^2 = \mathbb{H}^2 \cup S^1_\infty$ which we identify with a closed disk. There is an intrinsic description of S^1_∞ (independent of the model) which we now describe.

Given a pair of distinct biinfinite geodesics γ_1, γ_2 in $\mathbb{H}^2$, their closures $\overline{\gamma_1}$ and $\overline{\gamma_2}$ in $\overline{\mathbb{H}}^2$ either intersect once in $\mathbb{H}^2$, intersect once in S^1_∞ or are disjoint. Accordingly we say that γ_1 and γ_2 are *intersecting*, *asymptotic* (or *parallel*), or *ultraparallel*, respectively.

In the upper half plane model, we see that by applying an isometry, we can arrange these three possibilities into "standard pictures". Specifically, if we have two intersecting geodesics, then by applying an isometry, we can assume the point of intersection is $i \in \mathbb{U}^2$, and one of the geodesics is $i\mathbb{R}_+$. Given a pair of asymptotic geodesics, we can apply an isometry so that the point in common is ∞, and so the two geodesics are intersections of vertical lines in $\mathbb{C}$ with $\mathbb{U}^2$ (in fact, we can apply an isometry so that one is $i\mathbb{R}_+$ and the other is $1 + i\mathbb{R}_+$). Finally, a pair of ultraparallel geodesics can be taken to the arcs of circles centered at 0 intersected with $\mathbb{U}^2$.

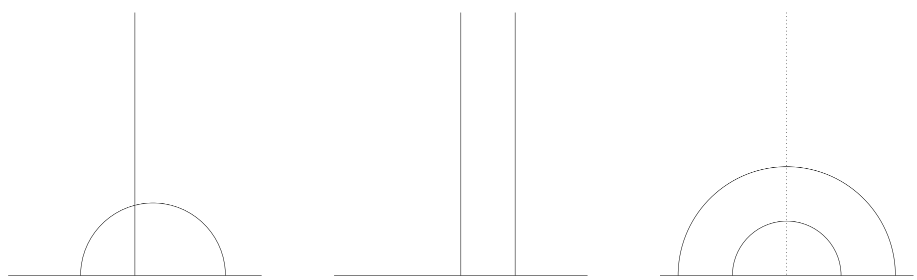

Figure 3: Intersecting, asymptotic and ultraparallel geodesics.

From these standard pictures, we see that the infimum of distances between asymptotic geodesics is 0, indeed, it limits to 0 as we travel out the ends of the geodesics toward their common point on S^1_∞. On the other hand, ultraparallel geodesics have the property that the infimum of distances between the geodesics is positive and realized at exactly one pair of points. In this case, the unique geodesic arc between these closest points is orthogonal (for this, look at the "standard picture" and take the closest point projection to $i\mathbb{R}_+$ which contains the desired arc).

From this we see that the circle at infinity S^1_∞ can be abstractly described as the set of equivalence classes of oriented geodesics (up to orientation preserving reparametrization), where two geodesics are equivalent if the distance between them in the forward direction (appropriately parameterized) tends to zero.

Another "intrinsic" description of S^1_∞ requires a choice of basepoint $x \in \mathbb{H}^2$. Given such a point, then S^1_∞ is in a 1-1 correspondence with the set of all geodesic rays beginning at x. To see this consider the disk model $\mathbb{D}^2$ of $\mathbb{H}^2$. By composing with an isometry if necessary, we can assume $x = 0 \in \mathbb{D}^2$. Then the geodesics rays beginning at 0 are just the arcs of Euclidean lines from 0 intersected with $\mathbb{D}^2$, and the closures of these in $\overline{\mathbb{D}}^2$ is precisely $\mathbb{S}^1$. If we start at a different basepoint, $z \in \mathbb{D}^2$, then S^1_∞ can be described as geodesics rays beginning at z. The natural bijection between the geodesics rays based at x and those based at z (i.e. the one that induces the identity on S^1_∞) is that we identify a geodesic ray based at x with the unique geodesic ray based at z to which it is asymptotic.

Using the geodesics, we can compute formulae for the distances in the two models.

Exercise 5.1. *Prove that for any two points $z = x + iy, w = u + iv \in \mathbb{U}^2$, the distance between them satisfies the equation*

$$\cosh(d_{\mathbb{U}}(z, w)) = 1 + \frac{|z - w|^2}{2yv}$$

and for $z, w \in \mathbb{D}^2$ the distance satisfies

$$\cosh(d_{\mathbb{D}}(z, w)) = 1 + \frac{2|z - w|^2}{(1 - |z|^2)(1 - |w|^2)}.$$

Hint: For the first part, prove that this is true when $y = v$, then prove that the expression is invariant by isometries.

The next two exercises describe what the isometries of $\mathbb{H}^2$ look like.

Exercise 5.2. *Prove each of the following theorems.*

Theorem 5.1. *Given $f_A \in \mathrm{PSL}_2(\mathbb{R})$, not the identity, then*

1. *f_A fixes a unique point of $\mathbb{U}^2$ if and only if $|Tr(A)| \in [0, 2)$,*
2. *f_A fixes a unique point of $\widehat{\mathbb{R}}$ if and only if $|Tr(A)| = 2$,*
3. *f_A fixes exactly two points on $\widehat{\mathbb{R}}$ if and only if $|Tr(A)| > 2$.*

□

The three possibilities in this theorem are called *elliptic*, *parabolic*, and *hyperbolic*, respectively.

We define the *translation length* of an isometry $f_A \in \mathrm{PSL}_2(\mathbb{R})$ to be

$$\tau(f_A) = \inf_{z \in \mathbb{U}^2} d_{\mathbb{U}}(z, f_A(z)).$$

Theorem 5.2. *Given $f_A \in \mathrm{PSL}_2(\mathbb{R})$ then*

- *f_A is elliptic if and only if $\tau(f_A) = 0$ and there exists $z \in \mathbb{U}^2$ with $d_{\mathbb{U}}(z, f_A(z)) = \tau(f_A)$,*
- *f_A is parabolic if and only if $\tau(f_A) = 0$ and there is no $z \in \mathbb{U}^2$ with $d_{\mathbb{U}}(z, f_A(z)) = \tau(f_A)$,*
- *f_A is hyperbolic if and only if $\tau(f_A) > 0$ and there exists $z \in \mathbb{U}^2$ with $d_{\mathbb{U}}(z, f_A(z)) = \tau(f_A)$.*

□

For a hyperbolic isometry f_A, you should see from your proof here that the set of points z for which $d_{\mathbb{U}}(z, f_A(z)) = \tau(f_A)$ is a geodesic in $\mathbb{U}^2$. Then you can compute the translation length exactly.

Proposition 5.3. *If f_A is a hyperbolic isometry, then*

$$\tau(f_A) = 2\cosh^{-1}\left(\frac{|Tr(A)|}{2}\right).$$

Proof. By conjugating, we can assume that the fixed points on $\widehat{R}$ for f_A are 0 and ∞. Then (up to sign)

$$A = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix}$$

for some $\lambda > 0$ with $\lambda \neq 1$. By further conjugating by an isometry swapping 0 and ∞, we can assume that $\lambda > 1$. Then $i\mathbb{R}_+$ is the geodesic of points z with $d_{\mathbb{U}}(z, f_A(z)) = \tau(f_A)$, and hence (from your computations in Exercise 5.1) $\tau(f_A) = 2\log(\lambda)$, so

$$\cosh\left(\frac{\tau(f_A)}{2}\right) = \frac{e^{\log(\lambda)} + e^{-\log(\lambda)}}{2} = \frac{\lambda + \lambda^{-1}}{2} = \frac{|Tr(A)|}{2}.$$

□

5.2 The hyperboloid model

All the hyperbolic geometry so far has taken place in $\mathbb{U}^2$ and $\mathbb{D}^2$. We now return to the hyperboloid $\mathcal{H}^2$ defined in Lecture 3.

Let

$$O(2,1) = \left\{ A \in M_{3\times 3}(\mathbb{R}) \mid A^T \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} A = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right\}.$$

The matrix here is the matrix representation of the bilinear form B_H, and thus $O(2,1)$ is the set of linear transformations of $\mathbb{R}^{2,1}$ that preserve B_H. It follows that $O(2,1)$ preserves the hyperboloid $B_H(\mathbf{x},\mathbf{x}) = -1$. Let $O^+(2,1)$ denote the subgroup that preserves $\mathcal{H}^2$, and since this also preserves B_H, it follows that $O^+(2,1)$ acts by isometries on $\mathcal{H}^2$. Any point $\mathbf{x} \in \mathcal{H}^2$ has B_H–norm -1, and by a Gram–Schmidt type procedure, can be expanded to a basis $\mathbf{x}, \mathbf{v}_1, \mathbf{v}_2$ for which $B_H(\mathbf{x}, \mathbf{v}_i) = 0$ for $i = 1, 2$ and $B_H(\mathbf{v}_i, \mathbf{v}_j) = \delta_{ij}$ ($= 0$ if $i \neq j$ and 1 otherwise). Making these the columns of a matrix,

$$A = (\mathbf{x}\ \mathbf{v}_1\ \mathbf{v}_2)$$

we see that $A \in O^+(2,1)$ and $A(\mathbf{e}_0) = \mathbf{x}$, so that $O^+(2,1)$ acts transitively on $\mathcal{H}^2$. The stabilizer of $\mathbf{e}_0$ takes the simple form

$$\left\{ A = \begin{pmatrix} 1 & 0 \\ 0 & B \end{pmatrix} \mid B \in O(2) \right\}$$

and so the stabilizer is isomorphic to $O(2)$. That is, $\mathcal{H}^2$ is a homogenous surface. We will shortly see that $\mathcal{H}^2$ is indeed another model for $\mathbb{H}^2$.

To construct the isometry from $\mathbb{D}^2$ to $\mathcal{H}^2$, we identify $\mathbb{R}^{2,1}$ with $\mathbb{R} \times \mathbb{C}$ and identify $\mathbb{D}^2 = \{0\} \times \mathbb{D}^2$. We define a projection similar to stereographic projection

$$P : \mathbb{D}^2 \to \mathcal{H}^2$$

as follows. For every $z \in \mathbb{D}^2$, construct the line through $-\mathbf{e}_0$ and $(0, z) \in \mathbb{R} \times \mathbb{C}$, and let $P(z)$ be the intersection of this line with $\mathcal{H}^2$.

Exercise 5.3. *Prove that*

$$P(x+iy) = \left(\frac{1+|z|^2}{1-|z|^2}, \frac{2x}{1-|z|^2}, \frac{2y}{1-|z|^2} \right).$$

Further prove that

$$\cosh(d_{\mathbb{D}}(z,w)) = -B_H(P(z), P(w)).$$

Therefore, given $\mathbf{x}, \mathbf{y} \in \mathcal{H}^2$, $\cosh^{-1}(-B_H(\mathbf{x}, \mathbf{y}))$ defines a metric on $\mathcal{H}^2$ so that P is an isometry, and since P is C^1, this is a Riemannian metric on $\mathcal{H}^2$. Because B_H is invariant by every element of $O^+(2,1)$, this metric is invariant by this entire group. Finally, because $O^+(2,1)$ acts transitively and the stabilizer of any point is $O(2)$, this new metric must be a multiple of the metric $d_{\mathcal{H}}$. Computing the derivative at $0 \in \mathbb{D}^2$, we see that it is exactly this metric. Thus we have proven the following.

Proposition 5.4. *For all* $\mathbf{x}, \mathbf{y} \in \mathcal{H}^2$*,*

$$\cosh(d_{\mathcal{H}}(\mathbf{x}, \mathbf{y})) = -B_H(\mathbf{x}, \mathbf{y}).$$

□

Whenever we refer to $\mathbb{H}^2$, we now allow this to be represented by $\mathbb{D}^2$, $\mathbb{U}^2$ or $\mathcal{H}^2$, whichever is more convenient.

Exercise 5.4. *How do we describe* S^1_∞ *in the hyperboloid model of* $\mathbb{H}^2$*?*

The hyperboloid $\mathcal{H}^2$ and $\mathbb{S}^2$ are similar in several respects. For example, observe that the geodesics in $\mathbb{S}^2$ are intersections of $\mathbb{S}^2$ with planes through 0 and any geodesic in $\mathbb{S}^2$ through $\mathbf{x}$ with unit tangent vector $\mathbf{v}$ can be parameterized with unit speed by

$$\mathbf{x}\cos(t) + \mathbf{v}\sin(t).$$

In comparison, we have the following.

Exercise 5.5. *Prove that the geodesics in* $\mathcal{H}^2$ *are the intersections of planes through* 0 *with* $\mathcal{H}^2$*. Hint: Prove that the image of a geodesic through* 0 *in* $\mathbb{D}^2$ *is sent to such a geodesic, and then use the group* $O^+(2,1)$*.*

Exercise 5.6. *Prove that a geodesic in* $\mathcal{H}^2$ *through* $\mathbf{x}$ *with unit tangent vector* $\mathbf{v}$ *can be parameterized by*

$$\mathbf{x}\cosh(t) + \mathbf{v}\sinh(t)$$

where $\cosh(t)$ *and* $\sinh(t)$ *are the hyperbolic cosine and hyperbolic sine, respectively.*

Exercise 5.7. *Prove that a* B_H*–orthogonal complement to a vector* $\mathbf{v} \in \mathbb{R}^{2,1} \setminus \{\mathbf{0}\}$ *intersects* $\mathcal{H}^2$ *in a geodesic (which we denote* $\gamma_{\mathbf{v}}$*) if and only if* $B_H(\mathbf{v}, \mathbf{v}) > 0$*. Given* $\mathbf{v}, \mathbf{u} \in \mathbb{R}^{2,1}$ *with* $B_H(\mathbf{v}, \mathbf{v}) = 1 = B_H(\mathbf{u}, \mathbf{u})$*, prove that* $|B_H(\mathbf{v}, \mathbf{u})|$ *is less than* 1*, equal to* 1 *or greater than* 1*, respectively, if and only if* $\gamma_{\mathbf{v}}$ *and* $\gamma_{\mathbf{u}}$ *intersect, are asymptotic or are ultraparallel, respectively. In the first case, prove that* $|B_H(\mathbf{v}, \mathbf{u})|$ *is the cosine of the angle (in* $(0, \pi/2)$*) between* $\gamma_{\mathbf{v}}$ *and* $\gamma_{\mathbf{u}}$*. In the third case, prove that* $|B_H(\mathbf{v}, \mathbf{u})|$ *is the hyperbolic cosine of the distance between* $\gamma_{\mathbf{v}}$ *and* $\gamma_{\mathbf{u}}$ *(i.e. the length of the common orthogonal).*

6 Polygons and Gauss–Bonnet part 1

We now set out to study the homogeneous surfaces $\mathbb{X}^2 = \mathbb{S}^2, \mathbb{R}^2$, and $\mathbb{H}^2$. When convenient, we will refer to these as the *model surfaces of curvature* $1, 0$, *and* -1, respectively. Although we will not study curvature, the use of the values $1, 0$ and -1 will become apparent in what follows.

6.1 Polygons

We can develop some intuition for the model surfaces by studying geometric objects inside of them. Of particular interest are polygons. We intuitively understand a polygon in $\mathbb{R}^2$ as the region bounded by some polygonal curve(s). We will simplify the situation by restricting our attention to convex polygons which we now describe.

In the discussion that follows a *geodesic* in $\mathbb{X}^2$ will mean a biinfinite geodesic if $\mathbb{X}^2 = \mathbb{R}^2$ or $\mathbb{H}^2$ and great circle if $\mathbb{X}^2 = \mathbb{S}^2$. A geodesic arc will mean an arc of a geodesic which is the image of an isometric embedding (for $\mathbb{R}^2$ and $\mathbb{H}^2$, any arc of a geodesic is a geodesic arc, but in $\mathbb{S}^2$, a geodesic arc can have length at most π).

For $\mathbb{X}^2 = \mathbb{R}^2$ or $\mathbb{H}^2$, we say that a subset $A \subset \mathbb{X}^2$ is *convex* if for any two points $\mathbf{x}, \mathbf{y} \in A$, the geodesic arc between $\mathbf{x}$ and $\mathbf{y}$ is contained in A. For $\mathbb{X}^2 = \mathbb{S}^2$, we further require that the distance between any two points of A be less than π: this means that any two points are connected by a unique geodesic.

A geodesic γ in $\mathbb{X}^2$ is the common boundary of two subspaces we call the *half-planes* bounded by γ. In $\mathbb{R}^2$ and $\mathbb{H}^2$, half-planes are convex, but not in $\mathbb{S}^2$ by our definition (though this is not entirely standard). A half-plane minus the bounding geodesic is called an *open half-plane* and is convex in $\mathbb{X}^2$ in all three cases.

The intersection of convex sets is convex since given $\mathbf{x}, \mathbf{y}$ in the intersection, the geodesic between these points is in each of the sets, hence also in the intersection. For $\mathbb{R}^2$ and $\mathbb{H}^2$ we define a *(finite sided convex) polygon* to be the intersection of finitely many half-planes. For $\mathbb{S}^2$, we further require that the distance between any two points of the intersection is less than π (equivalently, it lies in an open half-plane).

If

$$P = \bigcap_{j=1}^{n} H_{\gamma_j}$$

is a polygon obtained as the intersection of half-planes bounded by geodesics $\gamma_1, \ldots, \gamma_n$, then the *sides* of P are the maximal geodesic arcs of the γ_i which contain more than one point. A side of a polygon can be isometric to a compact interval, a ray $[0, \infty) \subset \mathbb{R}$ or the entire line $\mathbb{R}$. If P is compact, then the boundary of P is the union of its sides, which are compact, and any two sides are either disjoint or intersect in a unique point. The set of all points of intersection of the sides are called the *vertices* of the polygon.

Compact polygons have finite area, but the converse is only true for $\mathbb{S}^2$ and $\mathbb{R}^2$. We extend the family of polygons we are interested in beyond the compact polygons as follows. A polygon P is of *finite type* if it is compact, or if $P \subset \mathbb{H}^2$ and the closure $\overline{P} \subset \overline{\mathbb{H}}^2$ intersects S^1_∞ in finitely many points. The points in $\overline{P} \cap S^1_\infty$ are intersections with S^1_∞ of the closures of sides of P, and we call these *ideal vertices*. Any ideal vertex is the intersection of the closures of exactly two sides, just as is the case with *ordinary vertices* (i.e. those that lie in $\mathbb{X}^2$). The set of vertices (ordinary and ideal) of a finite type polygon P will be denoted $\mathcal{V}(P)$.

For an ordinary vertex v of a polygon P, the *angle* at v is the interior angle of the sides meeting at that vertex, which is an angle in the interval $(0, \pi)$. The angle at an ideal vertex v of a polygon P is defined to be zero. Note that in $\mathbb{D}^2$ or $\mathbb{U}^2$, the extensions to $\mathbb{S}^1$ or $\widehat{\mathbb{R}}$ are actually C^1 and do make an angle 0. We denote the angle at a vertex v of P by $\text{ang}_P(v)$ (or $\text{ang}(v)$ if P is understood).

A *triangle* is a finite type polygon with exactly 3 sides.

Exercise 6.1. *Prove that for any three points in $\mathbb{S}^2, \mathbb{R}^2$ or $\overline{\mathbb{H}^2}$ that do not all lie on a geodesic (or closure of a geodesic in the case of $\overline{\mathbb{H}^2}$), there exists a unique triangle with these three points as vertices.*

The next theorem relates the angles of a triangle to the area. It is a consequence of the classical Gauss-Bonnet Theorem for Riemannian surfaces.

Theorem 6.1. *If Δ is a triangle in the model surface $\mathbb{X}^2$ of curvature $\kappa \in \{1, 0, -1\}$, then*

$$\kappa \text{Area}(\Delta) = \left(\sum_{v \in \mathcal{V}(\Delta)} \text{ang}(v) \right) - \pi.$$

Proof. For $\mathbb{R}^2$, the statement reduces to the familiar statement that the sum of the interior angles of a triangle is π. We prove the theorem for $\mathbb{H}^2$ and leave the case of $\mathbb{S}^2$ as an exercise.

Figure 4: Preferred position for ideal triangles (left); including a compact triangle in an ideal triangle (right).

First suppose we have a triangle $\Delta \subset \mathbb{U}^2$ with vertices v_0, v_1, v_2 where v_0 is ideal and v_1, v_2 are either ordinary or ideal. By applying an isometry, we can assume that $v_0 = \infty$ in $\widehat{\mathbb{R}}$ and that the side with vertices v_1 and v_2 lies on the Euclidean circle centered at 0 of radius 1 (see Figure 4 on the left). If we write $\alpha_i = \text{ang}(v_i)$, then a little Euclidean geometry shows that the integral computing the area of Δ is the following.

$$\text{Area}(\Delta) = \iint_\Delta \frac{1}{y^2} dx dy = \int_{\cos(\pi - \alpha_1)}^{\cos(\alpha_2)} \int_{\sqrt{1-x^2}}^{\infty} \frac{1}{y^2} dy dx.$$

We now just compute this integral

$$\begin{aligned}
\int_{\cos(\pi-\alpha_1)}^{\cos(\alpha_2)}\int_{\sqrt{1-x^2}}^{\infty}\frac{1}{y^2}dydx &= \int_{\cos(\pi-\alpha_1)}^{\cos(\alpha_2)}\left(\lim_{t\to\infty}\frac{-1}{y}\Big|_{\sqrt{1-x^2}}^{t}\right)dx \\
&= \int_{\cos(\pi-\alpha_1)}^{\cos(\alpha_2)}\frac{dx}{\sqrt{1-x^2}} = \int_{\pi-\alpha_1}^{\alpha_2}\frac{-\sin(\theta)d\theta}{\sqrt{1-\cos^2(\theta)}} \\
&= \int_{\pi-\alpha_1}^{\alpha_2} -d\theta = \pi-\alpha_1-\alpha_2
\end{aligned}$$

Now suppose all three vertices v_0, v_1, v_2 of Δ are ordinary. By extending one side of Δ, we produce a triangle Δ' with one ideal vertex v_0' and two ordinary vertices v_1, v_2, and so that v_0 lies in the side from v_2 to v_0'. A third triangle Δ'' is obtained from this having vertices v_1, v_0, v_0'. We label the angles as in Figure 4 (on the right). Then $\alpha_0 + \beta_0 = \pi$ and since Δ' and Δ'' have one ideal vertex, so that the theorem holds for them, we have

$$\text{Area}(\Delta') = \pi - (\alpha_1 + \beta_1 + \alpha_2) \text{ and } \text{Area}(\Delta'') = \pi - (\beta_1 + \beta_0).$$

Therefore

$$\begin{aligned}
\text{Area}(\Delta) &= \text{Area}(\Delta') - \text{Area}(\Delta'') \\
&= \pi - (\alpha_1 + \beta_1 + \alpha_2) - (\pi - (\beta_1 + \beta_0)) \\
&= \pi - ((\alpha_1 + \beta_1 - \beta_1) + \alpha_2 + (\pi - \beta_0)) \\
&= \pi - (\alpha_1 + \alpha_2 + \alpha_0)
\end{aligned}$$

as required.

This completes the proof for $\mathbb{H}^2$. □

Exercise 6.2. *Prove the theorem for* $\mathbb{S}^2$.

Corollary 6.2. *The model surfaces are not locally isometric. That is, there is no isometry from an open set of one model surface to an open set of another model surface, unless they have the same curvature.*

Although we haven't proven that any isometry between an open set of one model surface to an open set of another model surface must be C^1 with nonsingular derivative, this follows from the same types of arguments from Lecture 4, using existence and uniqueness of geodesics (at least locally). The corollary follows from this, Exercise 3.2, and Theorem 6.1.

6.2 Canonical triangle maps

In the Euclidean plane, there is a "canonical map" between any two triangles (with labeled vertices): the affine map. It will be useful later to have a canonical map between (compact) hyperbolic triangles as well, and we explain this now using the hyperboloid model. To do this, let $C^+ \subset \mathbb{R}^{2,1}$ denote the open subset consisting of points $\mathbf{x} = (x_0, x_1, x_2)$ with $B_H(\mathbf{x},\mathbf{x}) < 0$ and $x_0 > 0$. We can project C^+ onto $\mathcal{H}^2$ radially by normalizing. Specifically, define $\Pi : C^+ \to \mathcal{H}^2$ by

$$\Pi(\mathbf{x}) = \frac{\mathbf{x}}{\sqrt{|B_H(\mathbf{x},\mathbf{x})|}}.$$

Given a compact triangle Δ in $\mathcal{H}^2$ with vertices v_1, v_2, v_3, we construction a canonical parameterization from the standard triangle $\Delta_0 \subset \mathbb{R}^3$ to Δ. Here

$$\Delta_0 = \left\{ \sum_{i=1}^{3} s_i \mathbf{e}_i \mid \sum_{i=1}^{3} s_i = 1 \text{ and } s_i \geq 0 \right\}$$

and $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ is the standard basis for $\mathbb{R}^3$. The numbers (s_1, s_2, s_3) are called the barycentric coordinates of the point $\sum s_i \mathbf{e}_i \in \Delta_0$, and are uniquely determined by the point $\sum s_i \mathbf{e}_i$. Using the ordering of vertices of Δ given by the indices, we construct a canonical C^1 embedding from Δ_0 into $\mathcal{H}^2$ which is a homeomorphism onto Δ according to the formula

$$\psi_\Delta(\sum_{i=1}^{3} s_i \mathbf{e}_i) = \Pi(\sum_{i=1}^{3} s_i v_i).$$

Because the elements of $O^+(2,1)$ are linear, they commute with Π. That is, $\Pi \circ A = A \circ \Pi$, for all $A \in O^+(2,1)$. Furthermore, because the map $\sum s_i \mathbf{e}_i \to \sum s_i v_i$ is the unique linear map extending $\mathbf{e}_i \mapsto v_i$, we see that for any $A \in O^+(2,1)$ we have

$$A \circ \psi_\Delta = \psi_{A(\Delta)}. \tag{4}$$

Given two triangles Δ_1, Δ_2 with ordered vertices, we obtain a canonical map from Δ_2 to Δ_1, namely $\psi_{\Delta_1,\Delta_2} = \psi_{\Delta_1} \circ \psi_{\Delta_2}^{-1}$. Observe that this is canonical in the sense that $\psi_{\Delta_1,\Delta_2}^{-1} = \psi_{\Delta_2,\Delta_1}$, $\psi_{\Delta_1,\Delta_2} \circ \psi_{\Delta_2,\Delta_3} = \psi_{\Delta_1,\Delta_3}$, and by (4), for any isometry $A \in O^+(2,1)$ we have $A \circ \psi_{\Delta_1,\Delta_2} = \psi_{A(\Delta_1),\Delta_2}$.

Exercise 6.3. *Suppose $\Delta_1, \Delta_2, \Delta_3, \Delta_4$ are triangles so that Δ_1 and Δ_2 share an edge and Δ_3 and Δ_4 share an edge. Suppose that the canonical maps ψ_{Δ_3,Δ_1} and ψ_{Δ_4,Δ_2} both send the shared edge of Δ_1 and Δ_2 to the shared edge of Δ_3 and Δ_4 with the same orientation. Prove that ψ_{Δ_3,Δ_1} and ψ_{Δ_4,Δ_2} agree on the shared edge.*

The canonical maps vary continuously in the following sense.

Exercise 6.4. *Suppose $\{\Delta_n\}_{n=1}^{\infty}$ is a sequence of triangles in $\mathbb{H}^2$ with vertices $\{v_1^n, v_2^n, v_3^n\}_{n=1}^{\infty}$ and Δ a triangle with vertices v_1, v_2, v_3. Further let $\{\psi_n : \Delta \to \Delta_n\}$ be the canonical maps. Prove that if $v_i^n \to v_i$ as $n \to \infty$ for each $i = 1, 2, 3$, then ψ_n converges uniformly to the identity on Δ. Hint: It suffices to prove that the canonical parameterizations $\{\psi_{\Delta_n} : \Delta_0 \to \Delta_n\}$ converge uniformly to the canonical parameterization $\psi_\Delta : \Delta_0 \to \Delta$.*

We will want a sharper version which requires a little more work.

Exercise 6.5. *With the set up from the previous exercise, prove that given $K > 1$, there exists N so that for all $n \geq N$ and all $z, w \in \Delta$,*

$$\frac{1}{K} d_{\mathbb{H}}(z, w) \leq d_{\mathbb{H}}(\psi_n(z), \psi_n(w)) \leq K d_{\mathbb{H}}(z, w).$$

Hint: Let $E \subset \mathbb{R}^2$ be any triangle, e.g. the triangle with vertices $\mathbf{0}, \mathbf{e}_1, \mathbf{e}_2$, and $f : E \to \Delta_0$ be an affine parameterization. Then let $\phi_n = \psi_{\Delta_n} \circ f : E \to \Delta_n$ and $\phi = \psi_\Delta \circ f : E \to \Delta$. Show that $d\phi_n$ converges uniformly to $d\phi$, where dF represents the matrix valued function $x \mapsto dF_x$ defined on E, for $F = \phi$ or ϕ_n. Finally, pull back the hyperbolic metric to E by all these maps to produce Riemannian structures $\{g_n\}$ and g on E, then compare the metrics d_{g_n} and d_g as $n \to \infty$; see also the proof of Proposition 3.1.

7 Geometric surfaces and Gauss–Bonnet part 2

7.1 Geometric surfaces

As a special class of Riemannian surfaces, we now define the particular class of surfaces we will study for the rest of this course. A *geometric surface* is a geodesic metric surface locally isometric to one of the model surfaces $\mathbb{S}^2, \mathbb{R}^2$, or $\mathbb{H}^2$. Since these are pairwise not locally isometric, a geometric surface is either a *spherical*, *Euclidean*, or *hyperbolic* surface as it is locally isometric to $\mathbb{S}^2$, $\mathbb{R}^2$ or $\mathbb{H}^2$, respectively. We also call these geometric surfaces of curvature $1, 0$ and -1, respectively. We will focus almost entirely on closed geometric surfaces.

Remark. A consequence of the *Classical Uniformization Theorem* (and the existence of *isothermal coordinates*) is that any Riemannian surface is *conformally equivalent* to a unique geometric surface. We will not elaborate on this at all, but mention it as an explanation for why we focus on the situation of geometric surfaces, rather than arbitrary Riemannian surfaces. The collection of all Riemannian surfaces homeomorphic to a fixed closed surface is unwieldy, though as we will see, the set of geometric surfaces homeomorphic to a given closed surface is much more well-behaved. Another consequence of the Uniformization Theorem is that any *Riemann surface*—a surface with a *complex structure* (something we have not discussed at all)—uniquely determines, and is determined by, a geometric surface.

The only closed geometric surfaces we have seen so far are $\mathbb{S}^2$ and the flat torus $\mathbb{T}^2$. We first explain how to construct many more geometric surfaces. Given a finite collection of polygons $P_1, \ldots, P_k$ with sides paired, from Professor Brendle's lectures we have already seen that we may glue the polygons together to obtain a surface. By requiring the polygons to be geometric and imposing further constraints on the side pairings, we can produce a geometric surface by gluing.

7.2 A hyperbolic example

Fix $g \geq 2$ and consider the set of $2g$ geodesics in $\mathbb{D}^2$ which are the intersections with $\mathbb{D}^2$ of lines in $\mathbb{C}$ through 0 making angles $\frac{k\pi}{2g}$. For every t let $P_{4g}(t)$ denote the $4g$–gon with vertices at distance t from 0 lying on these lines. See Figure 5.

We observe that each polygon $P_{4g}(t)$ is regular, meaning that subgroup of isometries preserving it is exactly the dihedral group of order $8g$. In particular, all the sides have equal length. The interior angles at each vertex are also therefore equal to some number $\alpha_{4g}(t)$ which is a continuous function of t for every g. As $t \to 0$, the polygons look more and more like regular Euclidean $4g$–gons: indeed, the geodesics in $\mathbb{D}^2$ are intersections of circles in $\mathbb{C}$ with larger and larger radii (tending to infinity as $t \to 0$). In particular, $\alpha_{4g}(t) \to \frac{(2g-1)\pi}{2g}$. On the other hand, as $t \to \infty$, the polygons become closer and closer to ideal polygons, and $\alpha_{4g}(t) \to 0$ as $t \to \infty$. By the intermediate value theorem, there exists $0 < t_g < \frac{(2g-1)\pi}{2g}$ for which $\alpha_{4g}(t_g) = \frac{\pi}{2g}$. Let $P_{4g} = P_{4g}(t_g)$.

Gluing opposite sides of P_{4g} produces a closed genus g surfaces. To produce a geometric surface, we choose particular gluing maps. Label the sides in a counterclockwise order $\sigma_1, \ldots, \sigma_{4g}$. For each $j = 1, \ldots, 2g$, let $f_j : \mathbb{D}^2 \to \mathbb{D}^2$ be the unique isometry with $f_j(\sigma_j) = \sigma_{j+2g}$ and $f_j(P_{4g}) \cap P_{4g} = \sigma_{j+2g}$, which exists since all the sides have the same length. Let $\pi : P_{4g} \to P_{4g}/\sim = S$ be the quotient map where $\sim$ is the equivalence relation

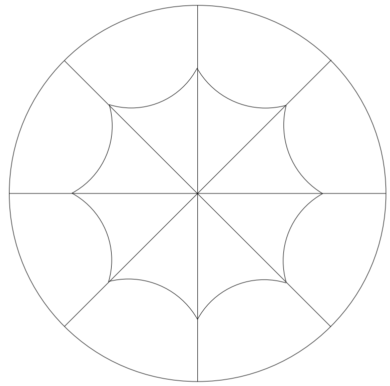

Figure 5: An example of $P_{4g}(t)$ with $g = 2$.

generated by declaring $z \sim f_j(z)$ for all $z \in \sigma_j$. Clearly any point in the interior of P_{4g} has an ϵ–ball that maps down isometrically to an ϵ–ball in S. Next let z be a point in the interior of some edge σ_j. For sufficiently small $\epsilon > 0$, $B_\epsilon(z) \cap P_{4g}$ is a half-ball (i.e. the intersection of $B_\epsilon(z)$ with the half-plane defined by the extension of σ_j) and similarly for $B_\epsilon(f_j(z)) \cap P_{4g}$. We denote these half-balls by $B'(z)$ and $B'(f_j(z))$ respectively. The union projects down to an ϵ–ball about $\pi(z)$. To see that this is locally isometric to an ϵ–ball in $\mathbb{D}^2$, we construct a map

$$h : B'(z) \cup B'(f_j(z)) \to B_\epsilon(z)$$

by

$$h(w) = \begin{cases} w & \text{if } w \in B'(z) \\ f_j^{-1}(w) & \text{if } w \in B'(f_j(z)) \end{cases}$$

Observe that if $w \sim w'$ in $B'(z) \cup B'(f_j(z))$, then $h(w) = h(w')$. It follows that h descends to a map

$$\hat{h} : B_\epsilon(\pi(z)) \to B_\epsilon(z).$$

Because points near the edge of $B_\epsilon(\pi(z))$ may be joined by a path that goes outside $B_\epsilon(\pi(z))$, h may not be an isometry. However, if we restrict to $B_{\frac{\epsilon}{2}}(z)$, then the shortest path between two points stays inside this ball, and hence this restriction of $\hat{h}$ gives the required isometry.

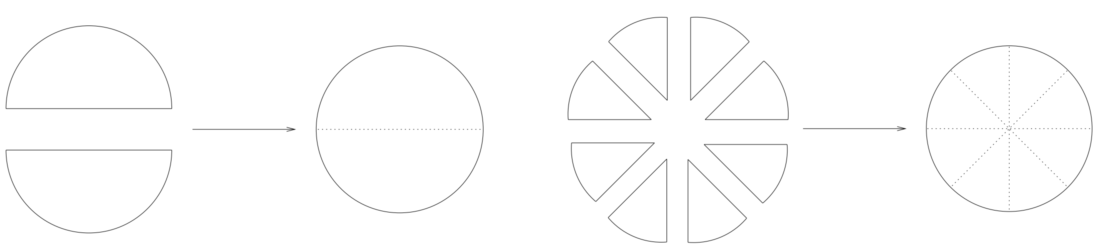

Figure 6: Gluing half-balls to obtain a ball (left) and gluing wedges to obtain a ball (right)

Finally, we note that all the vertices of P_{4g}, call them $z_1, \dots, z_{4g}$, are identified to a single point $z = \pi(z_j)$ in S. For a sufficiently small $\epsilon > 0$ the intersection of $B_\epsilon(z_i)$ is a "wedge" of an ϵ–ball with angle $\frac{\pi}{2g}$. The ϵ–ball about z is obtained by gluing together all these wedges according to $\sim$. Because $4g\frac{\pi}{2g} = 2\pi$, we can construct a local isometry from $B_\epsilon(z)$ to the ball $B_\epsilon(z_1)$ in $\mathbb{D}^2$ by a similar (but more notationally complicated) construction to that given above. Again, the restriction to the $\epsilon/2$–ball is the required isometry. See Figure 6.

More generally, if we glue any finite collection of compact polygons $\mathbb{H}^2$ along their sides in pairs by isometries to produce a connected surface we will produce a hyperbolic surface if and only if for any maximal collection of vertices which are identified to a single point in the quotient, the sum of the angles at those vertices is exactly 2π. We note that we are implicitly assuming that no edge is paired with itself.

There is nothing special about the hyperbolic situation here either. We can carry out the same construction gluing polygons in $\mathbb{S}^2$ or $\mathbb{R}^2$. We will later see that there are no "interesting" examples produced in the case of polygons in $\mathbb{S}^2$. However, given any parallelogram in $\mathbb{R}^2$, we can glue opposite sides by isometries to produce a closed Euclidean surface homeomorphic to a torus.

7.3 Geodesic triangulations

The construction of the previous section for producing geometric surfaces by gluing polygons is not a special one.

Proposition 7.1. *Any closed geometric surface is obtained by gluing sides of a finite collection of triangles in pairs. Said differently, for any geometric surface, there is a triangulation for which each triangle is isometric to a triangle in the model surface.*

Remark. We we use the term triangulation to mean the more general Δ–complex.

First we need a lemma which will also be useful later.

Given a geometric surface S, let $r : S \to (0, \infty]$ denote the function which assigns to each point $z \in S$ the supremum of those ϵ for which $B_\epsilon(z)$ is isometric to $B_\epsilon(z')$ for some z' in the model surface $\mathbb{X}^2$.

Lemma 7.2. *For a closed geometric surface S, the function r is a continuous function to $(0, \infty)$. In particular, r is bounded away from 0 and ∞.*

Proof. Since S is closed, r is a bounded function (bounded by the diameter of S, for example). Now observe that if $z \in S$ and $z' \in \mathbb{X}^2$ and $f : B_\epsilon(z) \to B_\epsilon(z')$ is an isometry, then for all $w \in B_\epsilon(z)$, the restriction

$$f|_{B_{\epsilon - d(z,w)}(w)} : B_{\epsilon - d(z,w)}(w) \to B_{\epsilon - d(z,w)}(f(w))$$

is also an isometry. In particular, if $w \in B_{r(z)}(z)$, then $r(w) \geq r(z) - d(z,w)$. If $d(z,w) < r(z)/2$, then $z \subset B_{r(z)/2}(w) \subset B_{r(z)}(z)$, and so we can reverse the roles of w and z to prove that $r(z) \geq r(w) - d(z,w)$, and hence

$$|r(w) - r(z)| \leq d(z,w).$$

for every $w \in B_{r(z)/2}(z)$. It follows that r is continuous (in the definition of continuity, given z and $\epsilon > 0$, we can let $\delta = \min\{\epsilon, r(z)/2\}$). $\square$

Before the proof of Proposition 7.1, we require the following.

Exercise 7.1. *Given distinct points $z, w \in \mathbb{X}^2$, set*

$$H(z,w) = \{x \in \mathbb{X}^2 \mid d_{\mathbb{X}}(x,z) \leq d_{\mathbb{X}}(x,w)\}.$$

That is, $H(z,w)$ is the set of points in $\mathbb{X}^2$ that are at least as close to z as they are to w. Prove that for any two distinct points z, w, $H(z,w)$ is a half-plane.

Proof of Proposition 7.1. What we describe next is an example of a *Voronoi cell decomposition* of S.

Let $r : S \to \mathbb{R}_+$ be the function described above and $\epsilon > 0$ any number less than $\min(r)$ (and less than π if $\mathbb{X}^2 = \mathbb{S}^2$). By compactness, we can find a set of points $Z = \{z_1, \ldots, z_k\}$ so that S is covered by the $\epsilon/3$–balls with centers at the points in Z. That is, every point is closer than $\epsilon/3$ to some point of Z. For each $z_j \in Z$ let C_j be the set of points in S for which the distance to z_j is no greater than the distance to any other point of Z:

$$C_j = \{w \in S \mid d(w, z_j) \leq d(w, z_i) \text{ for all } z_i \in Z\}.$$

Since every point of S is closer than $\epsilon/3$ to some z_j, it follows that $C_j \subset B_{\epsilon/3}(z_j) \subset B_\epsilon(z_j)$.

We claim that each C_j is isometric to a compact polygon. To see this, we fix some j and let $\{z_{i_0}, \ldots, z_{i_s}\} = Z \cap B_\epsilon(z_j)$ with $z_j = z_{i_0}$. Let $w \in C_j$ be any point. If z_j is not the unique closest point of Z to w, then let $z \in Z$ be any other point which is equally close, so that $d(z,w) = d(z_j, w)$. Observe that z must also lie in $B_\epsilon(z_j)$ by the triangle inequality:

$$d(z, z_j) \leq d(z,w) + d(w,z_j) = d(z_j, w) + d(w, z_j) < \frac{2\epsilon}{3}.$$

Therefore, C_j can also be expressed as

$$C_j = \{w \in B_\epsilon(z_j) \mid d(w, z_j) \leq d(w, z_{i_u}) \text{ for all } 0 \leq u \leq s\}.$$

Take any isometry $f_j : B_\epsilon(z_j) \to B_\epsilon(z_j')$ with $z_j' \in \mathbb{X}^2$, and let $z_{i_u}' = f_j(z_{i_u})$, for each $0 \leq u \leq s$. By the second description of C_j we have

$$f_j(C_j) = B_\epsilon(z_j') \cap \{w' \in \mathbb{X}^2 \mid d_{\mathbb{X}}(w', z_j') \leq d_{\mathbb{X}}(w', z_{i_u}') \text{ for all } 0 \leq u \leq s\}.$$

By Exercise 7.1, $f_j(C_j)$ is the intersection of finitely many half planes with $B_\epsilon(z_j')$

$$f_j(C_j) = \bigcap_{1 \leq u \leq s} H(z_j', z_{i_u}') \cap B_\epsilon(z_j').$$

We claim that the intersection of half-planes here is already contained in $B_\epsilon(z_j')$, and is thus a compact polygon. To see this, connect any point $w' \in \mathbb{X}^2 - B_\epsilon(z_j')$ by a geodesic segment to z_j'. The initial point z_j' of this segment lies in the intersection of half-planes, but by the above, the segment leaves at least one of the half-planes, call it $H(z_j', z_{i_u}')$, by the time the segment is a distance $\epsilon/3$ from z_j'. The remainder of this segment, and hence w', lies in the complementary half-plane, and so not in the given intersection, as required.

From this one can also see that for $z_i \neq z_j$, $C_i \cap C_j$ is either empty, a vertex, or a common side of the polygons C_i and C_j. Therefore, S is obtained by gluing together the polygons $C_1, \ldots, C_k$. To obtain triangles, we can subdivide each C_j into finitely many triangles (indeed, we can subdivide any compact n–gon into $n-2$ triangles). □

We will see another proof of this in the next two lectures.

7.4 Gauss-Bonnet part 2

By Proposition 7.1, every closed geometric surface S admits a *geometric triangulation*: a triangulation in which all the triangles are isometric to triangles in the model surface. Given a geometric surface and a geometric triangulation T we let $\text{Area}(S)$ denote the sum of the areas of the triangles in a geometric triangulation

$$\text{Area}(S) = \sum_{\Delta \in \mathcal{F}(T)} \text{Area}(\Delta).$$

Here $\mathcal{F}(S)$ is the set of triangles (or faces) in the triangulation.

Exercise 7.2. *Given two geometric triangulations of a closed geometric S, prove that there is a common subdivision. From this, prove that the area of S is well defined.*

The next result—our second version of the Gauss-Bonnet Theorem–gives another proof of the second part of this exercise. Recall that $\chi(S)$ is the Euler characteristic of S.

Theorem 7.3. *Given any closed geometric surface S of curvature κ we have*

$$\kappa \text{Area}(S) = 2\pi\chi(S).$$

Proof. This is essentially a computation after Theorem 6.1. Fix a geometric triangulation T for S, and let $\mathcal{V}(T)$, $\mathcal{E}(T)$, and $\mathcal{F}(T)$ denote the vertices, edges, and faces (triangles) of T. The numbers of each are V, E, and F, respectively, and $\chi(S) = V - E + F$. Since each triangle has exactly 3 sides and each edge is a side of exactly two triangles, we have $2E = 3F$ and hence

$$2\chi(S) = 2V - 2E + 2F = 2V - 3F + 2F = 2V - F.$$

Appealing to this equation and Theorem 6.1 we have

$$\begin{aligned}
\kappa \text{Area}(S) &= \sum_{\Delta \in \mathcal{F}(T)} \kappa \text{Area}(\Delta) = \sum_{\Delta \in \mathcal{F}(T)} \left(\left(\sum_{v \in \mathcal{V}(\Delta)} \text{ang}_\Delta(v) \right) - \pi \right) \\
&= \left(\sum_{\Delta \in \mathcal{F}(T)} \sum_{v \in \mathcal{V}(\Delta)} \text{ang}_\Delta(v) \right) - \sum_{\Delta \in \mathcal{F}(T)} \pi = \left(\sum_{v \in \mathcal{V}(T)} 2\pi \right) - \sum_{\Delta \in \mathcal{F}(T)} \pi \\
&= 2\pi V - \pi F = \pi(2V - F) = 2\pi\chi(S).
\end{aligned}$$

□

One corollary of this is that a closed surface can only be homeomorphic to a geometric surface of one type. More precisely, we have the following.

Corollary 7.4. *If S and S' are closed geometric surfaces of curvature κ and κ', respectively, and S is homeomorphic to S', the $\kappa = \kappa'$.*

Exercise 7.3. *You have already seen that any closed orientable surfaces is homeomorphic to some geometric surface. Prove the same is true for the non-orientable surfaces: that is, $\mathbb{R}P^2$ is homeomorphic to a spherical surface, the Klein bottle is homeomorphic to a Euclidean surface, and any other nonorientable closed surface is homeomorphic to a hyperbolic surface.*

8 Geometric surfaces as quotients by group actions

In the previous lecture we described a method for constructing geometric surfaces—gluing polygons. There is another construction which is quite useful which we now describe.

Recall that the action of a group G on a space X is called *properly discontinuous* if for any compact subset $K \subset X$ we have that

$$\{g \in G \mid g(K) \cap K \neq \emptyset\}$$

is a finite set. The action is *free* if no point is fixed by a non-identity element. Our interest lies in the case that G is acting by isometries on some model surface $\mathbb{X}^2$.

Proposition 8.1. *Let $\mathbb{X}^2$ be one of the model spaces and $\Gamma < \mathrm{Isom}(\mathbb{X}^2)$ a group acting properly discontinuously and freely on $\mathbb{X}^2$. Then $\mathbb{X}^2$ is a geometric surface and*

$$p : \mathbb{X}^2 \to \mathbb{X}^2/\Gamma$$

is a local isometry and a covering map.

Before we give the proof, we make a simple observation.

Lemma 8.2. *A subgroup $\Gamma < \mathrm{Isom}(\mathbb{X}^2)$ which acts freely is properly discontinuous if and only if for all $R > 0$ and $x, y \in \mathbb{X}^2$*

$$B_R(x) \cap \Gamma \cdot y$$

consists of finitely many points.

Proof. Since $\overline{B}_R(x)$ is compact in $\mathbb{X}^2$ for every $x \in \mathbb{X}^2$ and $R > 0$, and since compact sets are precisely the closed bounded subsets, this is immediate. □

Proof of Proposition 8.1. Let $p : \mathbb{X}^2 \to \mathbb{X}^2/\Gamma$ denote the quotient map defined by $p(x) = \Gamma \cdot x$. We claim that the quotient distance d between two orbits $\Gamma \cdot x = \Gamma \cdot y$ is given by

$$d(\Gamma \cdot x, \Gamma \cdot y) = \delta(\Gamma \cdot x, \Gamma \cdot y) \tag{5}$$

where

$$\delta(\Gamma \cdot x, \Gamma \cdot y) = \inf_{\gamma_1, \gamma_2 \in \Gamma} d_{\mathbb{X}}(\gamma_1(x), \gamma_2(y))$$

is the intermediate function used in the definition of the quotient metric given in Brendle's lectures.

By definition of d, we see that the left hand side of (5) is less than or equal to the right hand side. So we must prove that the left hand side is greater than or equal to the right hand side.

First observe that since Γ is acting by isometries, for all $\gamma_1, \gamma_2 \in \Gamma$ we have $d_{\mathbb{X}}(\gamma_1(x), \gamma_2(y)) = d_{\mathbb{X}}(x, \gamma_1^{-1}\gamma_2(y))$ and so

$$\delta(\Gamma \cdot x, \Gamma \cdot y) = \inf_{\gamma \in \Gamma} d_{\mathbb{X}}(x, \gamma(y)).$$

Moreover, by Lemma 8.2, it follows that this infimum is actually a minimum, and we can write

$$\delta(\Gamma \cdot x, \Gamma \cdot y) = d_{\mathbb{X}}(x, \gamma_{x,y}(y))$$

where $\gamma_{xy} \in \Gamma$ is an element so that the right hand side of this equation realizes the infimum.

Now, given a chain $\Gamma \cdot x = \Gamma \cdot x_0, \Gamma \cdot x_2, \ldots, \Gamma \cdot x_k = \Gamma \cdot y$, we have

$$\sum_{j=1}^{k} \delta(\Gamma \cdot x_{j-1}, \Gamma \cdot x_j) = \sum_{j=1}^{k} d_{\mathbb{X}}(x_{j-1}, \gamma_{j-1,j}(x_j))$$

where $\gamma_{j-1,j} = \gamma_{x_{j-1},x_j}$. Appealing again to the fact that Γ is acting by isometries we see that

$$\begin{aligned}\sum_{j=1}^{k} d_{\mathbb{X}}(x_{j-1}, \gamma_{j-1,j} x_j) &= d_{\mathbb{X}}(x_0, \gamma_{0,1}(x_1)) + d_{\mathbb{X}}(\gamma_{0,1}(x_1), \gamma_{0,1}\gamma_{1,2}(x_2)) + \cdots \\ &\quad + d_{\mathbb{X}}(\gamma_{0,1}\gamma_{1,2}\cdots\gamma_{k-2,k-1}(x_{k-1}), \gamma_{0,1}\gamma_{1,2}\cdots\gamma_{k-2,k-1}\gamma_{k-1,k}(x_k)).\end{aligned}$$

By the triangle inequality, the right hand side is greater than or equal to

$$d_{\mathbb{X}}(x_0, \gamma_{0,1}\cdots\gamma_{k-1,k}(x_k)) = d_{\mathbb{X}}(x, \gamma_{0,1}\cdots\gamma_{k-1,k}(y)) \geq d_{\mathbb{X}}(x, \gamma_{x,y}(y))$$

where the last inequality comes from the definition of $\gamma_{x,y}$ and the fact that $\gamma_{0,1}\cdots\gamma_{k-1,k} \in \Gamma$.

Putting this all together we see that

$$\sum_{j=1}^{k} d_{\mathbb{X}}(x_{j-1}, \gamma_{j-1,j}(x_j)) \geq d_{\mathbb{X}}(x, \gamma_{x,y}(y)) = \delta(\Gamma \cdot x, \Gamma \cdot y).$$

Therefore, since $d(\Gamma \cdot x, \Gamma \cdot y)$ is defined as the infimum of expressions on the left hand side, we have that

$$d(\Gamma \cdot x, \Gamma \cdot y) \geq \delta(\Gamma \cdot x, \Gamma \cdot y)$$

as required.

With this description of the metric in hand, we easily prove the rest of the proposition. Given $x \in \mathbb{X}^2$, let $\epsilon > 0$ be such that $B_\epsilon(x) \cap \gamma B_\epsilon(x) = \emptyset$ for all $\gamma \in \Gamma - \{1\}$. It follows that $B_\epsilon(x)$ projects down homeomorphically by p to a neighborhood of $\Gamma \cdot x$ in $\mathbb{X}^2/\Gamma$. In fact, we claim that the restriction to $B_{\epsilon/2}(x)$ of this map is an isometry onto its image. That is, for any $y, z \in B_{\epsilon/2}(x)$, we claim that

$$d_{\mathbb{X}}(y, z) = d(\Gamma \cdot y, \Gamma \cdot z).$$

To prove this, first observe that $d_{\mathbb{X}}(y, z) \leq d_{\mathbb{X}}(y, x) + d_{\mathbb{X}}(x, z) < \epsilon$, so we need only prove that for any $\gamma \in \Gamma - \{1\}$, $d_{\mathbb{X}}(y, \gamma(z)) \geq \epsilon$. However, $\gamma(z) \in \gamma B_{\epsilon/2}(x) \subset \gamma B_\epsilon(x)$, and by assumption, this ϵ–ball is disjoint from $B_\epsilon(x)$. Therefore, if we connect y to $\gamma(z)$ by a geodesic, it must first leave $B_{\epsilon/2}(x)$ and then $B_\epsilon(x)$, then enter $\gamma B_\epsilon(x)$, and finally enter $\gamma B_{\epsilon/2}(x)$. However, a path from inside an $\epsilon/2$–ball to outside an ϵ–ball has length at least $\epsilon/2$, and so the geodesic from y to $\gamma(z)$ has length at least ϵ, so $d_{\mathbb{X}}(y, \gamma(z)) \geq \epsilon$, as required.

Therefore, p restricts to an isometry from $B_{\epsilon/2}(x)$ to $B_{\epsilon/2}(\Gamma \cdot x)$, proving that p is a local isometry (since x was arbitrary), and furthermore $p^{-1}(B_{\epsilon/2}(\Gamma \cdot x) = \Gamma \cdot B_{\epsilon/2}$ by the choice of ϵ, and hence p is covering map. Finally, we observe that a geodesic from x to $\gamma_{x,y}(y)$ projects to a geodesic from $\Gamma \cdot x$ to $\Gamma \cdot y$, and so since $\mathbb{X}^2/\Gamma$ is locally isometric to $\mathbb{X}^2$, it is a geometric surface. $\square$

Exercise 8.1. *Prove that there are exactly two subgroups $\Gamma < \mathrm{Isom}(\mathbb{S}^2) = O(3)$ which act properly discontinuously and freely on $\mathbb{S}^2$: the trivial subgroup $\{I\}$ and the order two subgroup $\{\pm I\}$.*

Recall that $\mathrm{Isom}(\mathbb{R}^2) = \{F(x) = Ax + b \mid A \in O(2) \text{ and } b \in \mathbb{R}^2\}$. Since the derivative is constant (with respect to the standard basis $\mathbf{e}_1, \mathbf{e}_2$), by the chain rule we have a homomorphism

$$D : \mathrm{Isom}(\mathbb{R}^2) \to O(2)$$

given by $D(F) = dF_x$ for any $x \in \mathbb{R}^2$. That is, if $F(x) = Ax + b$, then $D(F) = A$. The kernel consists of the subgroup of translations $\mathbb{R}^2 \lhd \mathrm{Isom}(\mathbb{R}^2)$.

Exercise 8.2. *Prove that if $\Gamma < \mathrm{Isom}(\mathbb{R}^2)$ is a group that acts freely, then either $D(\Gamma) = \{I\}$, and hence $\Gamma < \mathbb{R}^2 \lhd \mathrm{Isom}(\mathbb{R}^2)$, or else $D(\Gamma) = \{I, r\}$, where r is a reflection. In the latter case, Γ has an index-two normal subgroup $\Gamma_0 = \Gamma \cap \mathbb{R}^2 \cap \Gamma \lhd \Gamma$. If we further assume that Γ acts properly discontinuously, then in the first case $\Gamma \cong \mathbb{Z}$ or $\mathbb{Z}^2$ (and hence the same is true for Γ_0 in the second case).*

8.1 Dirichlet domains.

In what follows, let $\Gamma < \mathrm{Isom}(\mathbb{X}^2)$ be a subgroup that acts properly discontinuously and freely on $\mathbb{X}^2$. To better understand Γ and the quotient $\mathbb{X}^2/\Gamma$, we relate this to the notion of gluing polygons as in Lecture 7. We assume $\mathbb{X}^2 = \mathbb{R}^2$ or $\mathbb{H}^2$ (which by Exercise 8.1 is not much of a constraint). A *convex fundamental domain* for the action of Γ is a closed convex subset $D \subset \mathbb{X}^2$ with the property that

$$\bigcup_{\gamma \in \Gamma} \gamma(D) = \mathbb{X}^2 \text{ and } D^\circ \cap (\gamma(D))^\circ = \emptyset \text{ for all } \gamma \in \Gamma - \{1\}.$$

Here D° is the interior of D. We now explain one construction of a convex fundamental domain for Γ.

Given a point $x \in \mathbb{X}^2$, define the *Dirichlet domain for* Γ *based at* x to be

$$D_\Gamma(x) = \{z \in \mathbb{X}^2 \mid d_{\mathbb{X}}(z, x) \leq d_{\mathbb{X}}(z, \gamma(x))\} = \{z \in \mathbb{X}^2 \mid d_{\mathbb{X}}(z, x) \leq d_{\mathbb{X}}(\gamma(z), x)\}.$$

Because Γ acts by isometries, the two descriptions are equivalent. The first allows us to see that $D_\Gamma(x)$ is convex. Indeed, recall that we have defined the half-plane $H(x, y)$ as the set of points at least as close to x as to y. Then

$$D_\Gamma(x) = \bigcap_{\gamma \in \Gamma - \{1\}} H(x, \gamma(x)). \tag{6}$$

Being the intersection of convex sets it follows that $D_\Gamma(x)$ is convex.

The second description allows us to think of $D_\Gamma(x)$ as the set of closest orbit representatives to x. Because every orbit has a representative which is closest to x by Lemma 8.2, it follows that

$$\bigcup_{\gamma \in \Gamma} \gamma D_\Gamma(x) = \mathbb{X}^2. \tag{7}$$

These observations make up the bulk of the proof of

Proposition 8.3. *For any group $\Gamma < \mathrm{Isom}(\mathbb{X}^2)$ acting properly discontinuously and freely, and any $x \in \mathbb{X}^2$, $D_\Gamma(x)$ is a fundamental domain for Γ.*

Proof. All that remains is to prove that $(D_\Gamma(x))^\circ \cap (\gamma D_\Gamma(x))^\circ = \emptyset$ for all $\gamma \in \Gamma - \{1\}$. To see this, first observe that

$$\gamma D_\Gamma(x) = D_\Gamma(\gamma(x)),$$

and hence by (6)

$$D_\Gamma(x) \subset H(x, \gamma(x)) \text{ and } D_\Gamma(\gamma(x)) \subset H(\gamma(x), x).$$

Since $H(x,\gamma(x)) \cap H(\gamma(x), x)$ is the geodesic of equidistant points between x and $\gamma(x)$, and in particular has empty interior, we see that $(D_\Gamma(x))^\circ \cap (\gamma D_\Gamma(x))^\circ = \emptyset$. □

The equivalence relation $\sim$ on $\mathbb{X}^2$ from the action of Γ (that is, $y \sim \gamma(y)$ for all $y \in \mathbb{X}^2$ and $\gamma \in \Gamma$) restricts to an equivalence relation on $D_\Gamma(x)$ (for any choice of x).

Proposition 8.4. *Let $\Gamma < \mathrm{Isom}(\mathbb{X}^2)$ be a subgroup that acts properly discontinuously and freely on $\mathbb{X}^2$, let $x \in \mathbb{X}^2$, and let $\sim$ denote the equivalence relation determined by Γ. Then $\mathbb{X}^2/\Gamma$ is isometric to $D_\Gamma(x)/\sim$.*

Proof. The inclusion $D_\Gamma(x) \subset \mathbb{X}^2$ induces a bijection from $D_\Gamma(x)/\sim$ onto $\mathbb{X}^2/\Gamma$. By Lemma 8.2, we see that for any $R > 0$, $D_\Gamma(x) \cap B_R(x)$ is the intersection with $B_R(x)$ of finitely many half-planes. In particular, we can follow the same idea from Lecture 7 to construct ϵ–balls about points in $D_\Gamma(x)/\sim$ built from finitely many half-balls or wedges for which the restriction of the bijection is an isometry. We leave the details as an exercise for the reader. □

Of particular interest to us is the case that $\mathbb{X}^2/\Gamma$ is compact—that is, it is a closed geometric surface. Observe that in this case the diameter of $\mathbb{X}^2/\Gamma$ is finite, and hence so is $D_\Gamma(x)$, which is to say, $D_\Gamma(x)$ is compact. By the discussion in the previous proof, we see that $D_\Gamma(x)$ is itself the intersection of finitely many half-planes. Therefore we have the following.

Corollary 8.5. *With the notation from the proposition, the closed geometric surface $\mathbb{X}^2/\Gamma$ is obtained from the compact polygon $D_\Gamma(x)$ by gluing the sides in pairs by isometries.*

This allows us to derive an alternative description of what it means for a closed quotient geometric surface to be orientable. First, we let $\mathrm{Isom}^+(\mathbb{X}^2) < \mathrm{Isom}(\mathbb{X}^2)$ be the index 2 subgroups consisting of all orientation preserving isometries. These are concretely given as

- $\mathrm{Isom}^+(\mathbb{S}^2) = SO(3)$,
- $\mathrm{Isom}^+(\mathbb{R}^2) = \{F(x) = Ax + b \mid A \in SO(2), b \in \mathbb{R}^2\}$, and
- $\mathrm{Isom}^+(\mathbb{H}^2) = \mathrm{PSL}_2(\mathbb{R})$.

Exercise 8.3. *Suppose $X = \mathbb{X}^2/\Gamma$ is a closed geometric surface. Prove that X is orientable if and only if $\Gamma < \mathrm{Isom}^+(\mathbb{X}^2)$. Hint: Apply the previous corollary and look at the "boundary word" from Brendle's lectures.*

8.2 Discreteness and proper discontinuity

Each of the isometry groups $\mathrm{Isom}(\mathbb{X}^2)$, for $\mathbb{X}^2 = \mathbb{S}^2, \mathbb{R}^2$, and $\mathbb{H}^2$ is also a metric space in such a way that the group operations are continuous. This is clear for $\mathbb{S}^2$ and $\mathbb{H}^2$ whose isometry groups can be realized as groups of matrices $O(3)$ and $O^+(2,1)$ (or $PSL_2^\pm(\mathbb{R})$ which is a good quotient of $\mathrm{SL}_2^\pm(\mathbb{R})$). For $\mathbb{R}^2$, the following exercise explains how $\mathrm{Isom}(\mathbb{R}^2)$ can also be viewed as a group of matrices, and hence a metric space.

Exercise 8.4. *Consider that the map*

$$\Phi : \mathrm{Isom}(\mathbb{R}^2) \to \mathrm{GL}_3(\mathbb{R})$$

given by

$$\Phi(x \mapsto Ax + b) = \begin{pmatrix} A & b \\ 0 & 1 \end{pmatrix}.$$

where we have represented the image as a matrix in block form. Prove that Φ *is an injective homomorphism. Find an affine embedding* $f : \mathbb{R}^2 \to \mathbb{R}^3$ *so that*

$$\Phi(F)(f(x)) = f(F(x)).$$

This is often expressed as saying that f *is* equivariant *with respect to* Φ*.*

Exercise 8.5. *For each of* $\mathbb{X}^2 = \mathbb{S}^2, \mathbb{R}^2$*, and* $\mathbb{H}^2$*, prove that if we view* $Isom(\mathbb{X}^2)$ *as a subset of* $M_{3\times 3}(\mathbb{R}) \cong \mathbb{R}^9$*, then it is a closed subset. Further prove that* $O(3)$ *is a compact subset.*

As an illustration that the metric (or more precisely, the topology) on $\mathrm{Isom}(\mathbb{X}^2)$ is natural, we prove the following.

Proposition 8.6. *Given* $\Gamma < \mathrm{Isom}(\mathbb{X}^2)$*,* Γ *acts properly discontinuously if and only if* Γ *is discrete.*

Here *discrete* means that the topology on Γ is the discrete topology—every subset is an open set. Here are a couple equivalent formulations.

Lemma 8.7. *Suppose* $\Gamma < \mathrm{Isom}(\mathbb{X}^2)$*. Then the following are equivalent.*

1. Γ *is discrete,*
2. *for any sequence* $\{\gamma_n\} \subset \Gamma$*, if* $\lim_{n\to\infty} \gamma_n = \gamma \in \mathrm{Isom}(\mathbb{X}^2)$*, then there exists* $N > 0$ *so that for all* $n \geq N$*,* $\gamma_n = \gamma$*.*
3. *for any sequence* $\{\gamma_n\} \subset \Gamma$*, if* $\lim_{n\to\infty} \gamma_n = 1$*, then there exists* $N > 0$ *so that for all* $n \geq N$*,* $\gamma_n = 1$*.*

Proof. The second condition states that any convergent sequence in Γ is eventually constant, while the third condition states that any sequence that converges to the identity is eventually constant.

To prove that (1) $\Rightarrow$ (3), observe that in a discrete set one point sets are open and hence $\{1\}$ is a neighborhood of Γ. Therefore, if $\{\gamma_n\}$ is a sequence converging to 1, then there exists N so that for all $n \geq N$, γ_n lies in this neighborhood—that is, $\gamma_n = 1$.

Next we prove that (3) $\Rightarrow$ (2). For this, suppose $\gamma_n \to \gamma \in \mathrm{Isom}(\mathbb{X}^2)$. Then since multiplication and taking inverses are continuous operations, we see that

$$\lim_{n\to\infty} \gamma_n \gamma_{n+1}^{-1} = 1(1)^{-1} = 1 \in \Gamma.$$

Since we are assuming (3), this means that there is some $N > 0$ so that for all $n \geq N$ we have $\gamma_n \gamma_{n+1}^{-1} = 1$, and so $\gamma_n = \gamma_{n+1}$. That is, the sequence $\{\gamma_n\}$ is eventually constant and so converges to this constant value $\gamma \in \Gamma$ as required.

Finally, to prove that (2) $\Rightarrow$ (1), we prove the contrapositive: suppose Γ is not discrete and find a convergent sequence $\{\gamma_n\}$ that is not eventually constant. For this, let $\gamma \in \Gamma$ be a point for which $\{\gamma\}$ is not open. Since $B_{1/n}(\gamma)$ is an open set for all n, we see that $B_{1/n}(\gamma)$ contains some point $\gamma_n \neq \gamma$ for all n. Clearly $\gamma_n \to \gamma$, and $\{\gamma_n\}$ is not eventually constant. This completes the proof. □

We need one more lemma, part of which was already implicit in the description of the isometry groups $\mathrm{Isom}(\mathbb{X}^2)$.

Lemma 8.8. *Fix any point $x \in \mathbb{X}^2$ and $R > 0$. Then any $\gamma \in \mathrm{Isom}(\mathbb{X}^2)$ is determined by $\gamma|_{\overline{B}_R(\gamma)}$. Moreover, given $\gamma_n, \gamma \in \mathrm{Isom}(\mathbb{X}^2)$, we have $\gamma_n \to \gamma$ if and only if $\gamma_n|_{\overline{B}_R(x)}$ converges uniformly to $\gamma|_{\overline{B}_R(x)}$.*

Proof. We give the proof for $\mathbb{H}^2$ and leave the other cases as exercises. Recall that in the proofs that the isometry group of $\mathbb{U}^2$ is $\mathrm{PSL}_2(\mathbb{R})$, we showed that any isometry is determined by its behavior on the set of geodesics through a particular point (the choice of point was unimportant). Now just observe that a (biinfinite) geodesic through any point is uniquely determined by its intersection with any ball about that point.

To prove the statement about convergence, it suffices to prove it for a particular point (why?), and we choose $\mathbf{e}_0 \in \mathcal{H}^2 \subset \mathbb{R}^{2,1}$. We can parameterize the closed R-ball about $\mathbf{e}_0$ in polar coordinates (r, θ), $\theta \in [0, 2\pi]$, $r \in [0, R]$ by

$$F(r,\theta) = \cosh(r)\mathbf{e}_0 + \sinh(r)(\cos(\theta)\mathbf{e}_1 + \sin(\theta)\mathbf{e}_2) = (\cosh(r), \sinh(r)\cos(\theta), \sinh(r)\sin(\theta)).$$

Now, for any $\gamma \in O^+(2,1)$, the entries of the 3×3 matrix are

$$\gamma = (\gamma(\mathbf{e}_1)\ \gamma(\mathbf{e}_2)\ \gamma(\mathbf{e}_3)).$$

By the linearity, it follows that $\gamma(F(r,\theta))$ is given by

$$\gamma(\cosh(r)\mathbf{e}_0+\sinh(r)(\cos(\theta)\mathbf{e}_1+\sin(\theta)\mathbf{e}_2)) = \cosh(r)\gamma(\mathbf{e}_0)+\sinh(r)\cos(\theta)\gamma(\mathbf{e}_1)+\sinh(r)\sin(\theta)\gamma(\mathbf{e}_2).$$

Therefore, $\gamma_n \to \gamma$ if and only if $\gamma_n(\mathbf{e}_i) \to \gamma(\mathbf{e}_i)$, for every $i = 0, 1, 2$, which happens if and only if $\gamma_n(F(r,\theta)) \to \gamma(F(r,\theta))$ for all r, θ. Since all γ_n are isometries, we can apply Theorem 2.14, to see that $\gamma_n|_{\overline{B}_R(\mathbf{e}_0)} \to \gamma|_{\overline{B}_R(\mathbf{e}_0)}$ uniformly, as required (strictly speaking, we apply Theorem 2.14 to show that every subsequence converges to γ, and hence the entire sequence does). □

Proof of Proposition 8.6. This is now an easy consequence of the previous two lemmas since for any $x \in \mathbb{X}^2$ and $R > 0$, $\overline{B}_R(x)$ is compact and any compact subset is contained in such a set.

We prove one implication, and leave the other as an exercise. Suppose Γ acts properly discontinuously on $\mathbb{X}^2$ and suppose $\{\gamma_n\} \subset \Gamma$ is a sequence with $\gamma_n \to 1$. By Lemma 8.8, the restriction $\gamma_n|_{\overline{B}_R(x)}$ converges uniformly to the identity on $\overline{B}_R(x)$. In particular, there exists $N > 0$ so that for all $n \geq N$ we have $\gamma_n(\overline{B}_R(x)) \cap \overline{B}_R(x) \neq \emptyset$. Since the action is properly discontinuous, it follows that $\{\gamma_n\}$ consists of only finitely many elements. Since the sequence converges to 1, it must be eventually constant. By Lemma 8.7, it follows that Γ is discrete.

Exercise 8.6. *Prove the converse. Hint: Consider applying Theorem 2.14.*

□

Since the only elements in a group acting properly discontinuously which fix a point are torsion elements, and for $\mathbb{X}^2 = \mathbb{R}^2$ or $\mathbb{H}^2$, every torsion element fixes a point, we have

Corollary 8.9. *A subgroup* $\Gamma < \mathrm{Isom}(\mathbb{X}^2)$ *acts properly discontinuously and freely if and only if* Γ *is discrete and torsion free (or* $\Gamma = \{\pm I\} < O(3)$*).*

9 All geometric surfaces as quotients by actions

The goal of this section is to prove that every closed geometric surface is obtained as in the construction of the previous section.

Theorem 9.1. *If X is a closed geometric surface, then X is isometric to $\mathbb{X}^2/\Gamma$, where $\pi_1(X) \cong \Gamma < \mathrm{Isom}(\mathbb{X}^2)$ is discrete and torsion free or $\Gamma = \{\pm I\} < O(3)$.*

This is true for any complete geometric surface, but for simplicity, we prove it only for closed surfaces here.

9.1 Lifting a geometric metric

In this section we explain how to see that the universal covering of a geometric surface is naturally a geometric surface. This begins with a fairly general construction of lifting a metric.

Proposition 9.2. *If $f : \widetilde{X} \to X$ is a local homeomorphism from a connected metric space $\widetilde{X}$ to a path metric space X then there is a path metric on $\widetilde{X}$ (defining the same topology) making f a local isometry.*

A *local homeomorphism* is a map for which every point in the domain has a neighborhood for which the restriction of the map is a homeomorphism onto an open subset of the range. The new metric on $\widetilde{X}$ is sometimes called the *lifted metric* or *pull-back metric.*

Proof. Given a path $\gamma : [a, b] \to \widetilde{X}$, define

$$\widetilde{\ell}(\gamma) = \ell(f \circ \gamma).$$

We claim that if we define $\widetilde{d} : \widetilde{X} \times \widetilde{X} \to \mathbb{R}$ by

$$\widetilde{d}(x, y) = \inf \widetilde{\ell}(\gamma)$$

where the infimum is taken over all paths connecting x to y, then this gives us the desired metric.

We note that it is not even clear that there is a path between any two points, let alone one with finite $\widetilde{\ell}$–length.

Given any $x \in \widetilde{X}$, let $x \in U \subset \widetilde{X}$ and $V \subset X$ be open sets such that $f|_U : U \to V$ is a homeomorphism. Suppose $\epsilon > 0$ with $B_{2\epsilon}(f(x)) \subset V$. Let $B' = f|_U^{-1}(B_{2\epsilon}(f(x)))$ and $B = f|_U^{-1}(B_\epsilon(f(x))) \subset B'$, which are neighborhoods of x. For any $y \in B$, there is a path in B' of $\widetilde{\ell}$–length less than ϵ connecting x to y: take $f|_U^{-1}$ of such a path connecting $f(x)$ to $f(y)$.

Therefore, any point has a neighborhood of points which it is connected to with finite $\widetilde{\ell}$–length. So, if we look at the set of points A connected to x by a path of finite $\widetilde{\ell}$–length and those that are not connected to x by such a path, these are two disjoint open sets whose union is all of $\widetilde{X}$. Since $\widetilde{X}$ is connected, the latter set must be empty, and hence $\widetilde{d}$ is finite valued for any pair of points.

We also observe from this that for any two points $y, z \in B$, there is path of $\widetilde{\ell}$–length less than 2ϵ contained in B'. On the other hand, the length of any path that leaves B' must have length greater than 2ϵ and so $\widetilde{d}(y, z) = d(f(y), f(z))$. So, if $\widetilde{d}$ is a metric, then f is a local isometry with respect to this metric.

It is straightforward to see that $\widetilde{d}$ is symmetric and satisfies the triangle inequality. As the infimum of a set of nonnegative numbers, it must be nonnegative. To see that $\widetilde{d}(x,y) = 0$ implies $x = y$, we note that if y is in B, then the previous two paragraphs imply $x = y$. On the other hand, if $y \notin B$, then any path from x to y must leave the B, and hence composing with f we obtain a path from $f(x)$ to $f(y)$ that leaves $B_\epsilon(x)$. As such, the path has length at least ϵ. Since this is true for an arbitrary path connecting x to y, we have $\widetilde{d}(x,y) \geq \epsilon$, contradicting the assumption that $\widetilde{d}(x,y) = 0$. This competes the proof that $\widetilde{d}$ is a metric, and the proof of the proposition. □

Corollary 9.3. *If X is a closed geometric surface and $p : \widetilde{X} \to X$ is a connected covering map, then $\widetilde{X}$ admits a metric making it a geometric surface so that p is a local isometry. Moreover, there exists $\epsilon > 0$ so that for all $x \in \widetilde{X}$ there exists an isometric embedding $B_\epsilon(x) \to \mathbb{X}^2$.*

Proof. Suppose X is locally isometric to $\mathbb{X}^2$. According to Lemma 7.2, there is an $\epsilon > 0$ so for that every point $x \in X$ there is an isometric embedding of $B_\epsilon(x)$ into $\mathbb{X}^2$.

Since a covering map is a local homeomorphism, Proposition 9.2 implies the existence of a path metric on $\widetilde{X}$ for which the covering map $p : \widetilde{X} \to X$ is a local isometry. Therefore $\widetilde{X}$ is locally isometric to $\mathbb{X}^2$. To see that this is a geodesic metric, let $x, y \in \widetilde{X}$ be any two points and connect them by a sequence of paths γ_n parameterized proportional to arc length so that $\widetilde{\ell}(\gamma_n) \to d(x,y)$. Pushing these forward by p we have a sequence of paths $p \circ \gamma_n$ into X parameterized proportional to arc length connecting $p(x)$ to $p(y)$. Since X is compact, Theorem 2.14 tells us that a subsequence converges uniformly to a path $\hat{\gamma}$ connecting x to y of length equal to $d(x,y)$. For n sufficiently large, $p \circ \gamma_n$ is homotopic rel endpoints to $\hat{\gamma}$. Therefore, $\hat{\gamma}$ lifts to a path from x to y of length $d(x,y)$, and so $\widetilde{X}$ is a geodesic metric space and hence a geometric surface.

To see this second statement, let $x \in \widetilde{X}$ and $\phi : B_\epsilon(p(x)) \to B_\epsilon(x') \subset \mathbb{X}^2$ be an isometry for some $x' \in \mathbb{X}^2$. The ball $B_\epsilon(x')$ is a disk (if $\mathbb{X}^2 = \mathbb{S}^2$ we must also assume $\epsilon < \pi$). The inverse ϕ^{-1} can be lifted to a map $\widetilde{\phi^{-1}} : B_\epsilon(x') \to B_\epsilon(x)$, which is an isometry (why?). The inverse $\widetilde{\phi^{-1}}^{-1}$ of this gives the required isometry. □

Exercise 9.1. *Prove that the metric on $\widetilde{X}$ from this corollary is complete.*

9.2 The developing map.

We continue to assume that X is a closed geometric surface, locally isometric to $\mathbb{X}^2$ and now let $p : \widetilde{X} \to X$ be the universal covering. By Corollary 9.3 we may endow $\widetilde{X}$ with a geometric metric, also locally isometric to $\mathbb{X}^2$. Observe that $\pi_1(X)$ acts on $\widetilde{X}$ by covering transformations. Given $\gamma \in \pi_1(X)$, write $\gamma : \widetilde{X} \to \widetilde{X}$ for the associated covering transformation, so that $p \circ \gamma = p$. Since γ is a homeomorphism, p is a local isometry, and $\widetilde{X}$ is a geodesic metric space, it follows that γ is in fact an isometry. That is, $\pi_1(X)$ acts on $\widetilde{X}$ by isometries.

All that remains to prove Theorem 9.1 is to prove the following.

Theorem 9.4. *With the notation as above, there is an isometry $D : \widetilde{X} \to \mathbb{X}^2$.*

The isometry D from this theorem is called the *developing map.*

Assuming this, we complete the proof of Theorem 9.1

Proof of Theorem 9.1. The isometry D induces an action of $\pi_1(X)$ on $\mathbb{X}^2$ so that $\gamma \in \pi_1(X)$ acts on $\mathbb{X}^2$ by

$$D \circ \gamma \circ D^{-1}.$$

Since this is the composition of three isometries, it is an isometry. We denote it by $\rho(\gamma) \in \mathrm{Isom}(\mathbb{X}^2)$ so that

$$\rho : \pi_1(X) \to \mathrm{Isom}(\mathbb{X}^2)$$

is a homomorphism. This is called the *holonomy homomorphism.* Set $\Gamma = \rho(\pi_1(X))$.

Since $\pi_1(X)$ acts transitively on the fibers of p, we see that the isometry D descends to a homeomorphism

$$\hat{D} : X \to \mathbb{X}^2/\Gamma.$$

Since p and the covering map $\mathbb{X}^2 \to \mathbb{X}^2/\Gamma$ are local isometries, so is $\hat{D}$. Since the metrics on X and $\mathbb{X}^2/\Gamma$ are geodesic metrics, $\hat{D}$ is in fact an isometry which proves the theorem. □

For the proof of Theorem 9.4 we will need the following addition to Lemma 8.8.

Lemma 9.5. *For $B_\epsilon(x) \subset \mathbb{X}^2$, any isometric embedding $B_\epsilon(x) \to \mathbb{X}^2$ extends to an isometry.*

Proof. For any $\epsilon > 0$ (less than π if $\mathbb{X}^2 = \mathbb{S}^2$) and any $x \in \mathbb{X}^2$, any point in $B_\epsilon(x)$ is connected to x by a unique geodesic arc in $B_\epsilon(x)$. We think of these as geodesics emanating radially from x. An isometric embedding $\phi : B_\epsilon(x) \to \mathbb{X}^2$ sends these to a set of geodesics emanating radially from $\phi(x)$. As in Lecture 4, by examining the image of the circle of points at a distance $\epsilon/2$, say, we see that the angle between these radial geodesic must be preserved. Therefore, we can find an isometry $\gamma : \mathbb{X}^2 \to \mathbb{X}^2$ so that $\gamma \circ \phi$ sends x to itself and is the identity on this set of geodesics. But then, ϕ is the restriction of γ^{-1} to $B_\epsilon(x)$, as required. □

Corollary 9.6. *For X a geometric surface and $x \in X$, if there exists an isometric embedding $\phi : B_\epsilon(x) \to \mathbb{X}^2$, then any isometric embedding $B_\epsilon(x) \to \mathbb{X}^2$ is obtained by composing ϕ with some $f \in \mathrm{Isom}(\mathbb{X}^2)$. In particular, if two isometric embeddings $\phi, \phi' : B_\epsilon(x) \to \mathbb{X}^2$ agree on $B_{\epsilon'}(x)$, for some $0 < \epsilon' \leq \epsilon$ then $\phi = \psi$.*

Proof. Given another isometric embedding $\psi : B_\epsilon(x) \to \mathbb{X}^2$, we have that $\psi \circ \phi^{-1}$ is an isometric embedding from $B_\epsilon(\phi(x))$ to $\mathbb{X}^2$, and so is the restriction of an isometry γ. Therefore $\psi = \psi \circ \phi^{-1} \circ \phi = \gamma \circ \phi$. The second statement then follows by applying Lemma 8.8 to see that the isometry γ is determined by ϕ and ψ on $B_{\epsilon'}(x)$ for any $\epsilon' > 0$. □

The idea for the construction of the map D is as follows. Pick a basepoint $x \in \widetilde{X}$ and first define D on some ball to be any isometric embedding to $\mathbb{X}^2$. For an arbitrary point $z \in \widetilde{X}$, we may connect x to z by a path, and cover the path by balls that admit isometric embeddings to $\mathbb{X}^2$. By compactness, we may choose a subset of these balls that form a chain along the path—so consecutive balls overlap. We arbitrarily choose isometric embeddings for each of these balls, and then start "adjusting" them using Corollary 9.6, one at a time along the chain, so that on the overlap of consecutive balls the isometries agree. When we reach the ball containing z, we have defined $D(z)$.

There are a number of things that must be checked in order to see that we get a well-defined map, and that it is an isometry. For example, we need to check that the map is independent of the collection of balls used in its definition, as well as the choice of path from x to z. The latter property is where we use the fact that $\widetilde{X}$ is simply connected so that any two paths are homotopic rel endpoints. We could argue these directly using ideas similar to those involved in path and homotopy lifting for covering spaces. This suggests a shorter proof using covering spaces which we will give. First we need a fact about covering spaces.

Exercise 9.2. *If Σ_1, Σ_2 are two connected surfaces, Σ_2 simply connected and $p : \Sigma_1 \to \Sigma_2$ is a covering map, then prove that p is a homeomorphism.*

Proof of Theorem 9.4. By Corollary 9.3, there exists $\epsilon > 0$ be so that for all $x \in \widetilde{X}$, there is an isometric embedding $B_{2\epsilon}(x) \to \mathbb{X}^2$. By Corollary 9.6, any isometric embedding $\psi : B_\epsilon(x) \to \mathbb{X}^2$ is the restriction of a unique isometric embedding $B_{2\epsilon}(x) \to \mathbb{X}^2$ which we denote 2ψ. Furthermore, if $\phi : B_\epsilon(x) \to \mathbb{X}^2$ is an isometric embedding and $y \in B_\epsilon(x)$, then another application of Corollary 9.6, implies that there exists a unique isometric embedding $\psi : B_\epsilon(y) \to \mathbb{X}^2$ so that $2\psi|_{B_\epsilon(x)} = \phi$.

Now we define

$$\Omega = \Omega_\epsilon(\widetilde{X}) = \{(\phi, x) \mid x \in \widetilde{X} \text{ and } \phi : B_\epsilon(x) \to \mathbb{X}^2 \text{ is an isometric embedding}\}$$

and construct a metric on this set by

$$d_\Omega((\phi, x), (\psi, y)) = \begin{cases} \widetilde{d}(x, y) & \text{if } \widetilde{d}(x, y) < \epsilon \text{ and } \phi = 2\psi|_{B_\epsilon(x)} \\ \\ \epsilon & \text{otherwise} \end{cases}.$$

By construction, $d_\Omega((\phi, x), (\psi, y)) \geq 0$, with equality if and only if $(\phi, x) = (\psi, y)$ since $\widetilde{d}$ is a metric. Symmetry follows from the fact that if $y \in B_\epsilon(x)$, then $x \in B_\epsilon(y)$ and if $\phi = 2\psi|_{B_\epsilon(x)}$, then $\psi = 2\phi|_{B_\epsilon(y)}$ by Corollary 9.6. Finally, to prove that the triangle inequality holds, observe that since d_Ω is never larger than ϵ, we need only consider a triple $(\phi, x), (\psi, y), (\eta, z)$ in which $d_\Omega((\phi, x), (\eta, z)) + d_\Omega((\eta, z), (\psi, y)) < \epsilon$ (in particular both of these distances are given by the first situation in the definition of d_Ω). In this case, the triangle inequality for $\widetilde{d}$ implies

$$\widetilde{d}(x, y) \leq \widetilde{d}(x, z) + \widetilde{d}(z, y) = d_\Omega((\phi, x), (\eta, z)) + d_\Omega((\eta, z), (\psi, y)) < \epsilon. \tag{8}$$

Furthermore, since ϕ and ψ are the restriction of 2η to the respective ϵ–balls, it follows that ψ agrees with 2ϕ on $B_\epsilon(y)$, so $d_\Omega((\phi, x), (\psi, y)) = \widetilde{d}(x, y)$. Appealing to (8) completes the proof of the triangle inequality for d_Ω, and so d_Ω is a metric.

Next, note that there is a forgetful map $P : \Omega \to \widetilde{X}$ given by $P(\phi, x) = x$. For any $\epsilon > 0$ and $x \in \widetilde{X}$, we observe that by Corollary 9.6

$$P^{-1}(B_\epsilon(x)) = \bigsqcup_{\gamma \in \mathrm{Isom}(\mathbb{X}^2)} B_\epsilon(\gamma \circ \phi, x)$$

where $\phi : B_\epsilon(x) \to \mathbb{X}^2$ is one isometric embedding, and so by the Corollary 9.6, $\gamma \circ \phi$ for $\gamma \in \mathrm{Isom}(\mathbb{X}^2)$ gives all others. From the definition of the metric d_Ω, the restriction of P to each ball $B_\epsilon(\gamma \circ \phi, x) \to B_\epsilon(x)$ is an isometry, in particular, a homeomorphism. It follows that P is a covering map.

Given any $(\phi, x) \in \Omega$, let Ω^ϕ be the component of Ω containing (ϕ, x) and $P^\phi = P|_{\Omega^\phi} : \Omega^\phi \to \widetilde{X}$ the restricted covering map. Since Ω^ϕ is connected and $\widetilde{X}$ is simply connected, P is a homeomorphism by Exercise 9.2. Now we set $(P^\phi)^{-1}(y) = (\phi_y, y)$ and define $D(y) = \phi_y(y)$. Observe that $D|_{B_\epsilon(y)} = \phi_y$, so the restriction of D to any ϵ–ball is an isometry, and in particular D is a local isometry. In fact, every ϵ–ball in $\mathbb{X}^2$ is evenly covered, so D is a covering map. Since $\mathbb{X}^2$ is simply connected, D is a homeomorphism and since $\widetilde{X}$ and $\mathbb{X}^2$ are geodesic metric spaces, D is an isometry. ☐

10 Moduli space and biLipschitz maps

10.1 Definitions

We have now seen that there are many ways to construct geometric surfaces. Our next goal is to describe a way of organizing all geometric surfaces. We will concentrate primarily on closed orientable geometric surfaces. The classification of closed orientable surfaces provides us with the first step toward such an organization, and we thus focus on one homeomorphism type of closed surface at a time. We require the notion of *orientation preserving homeomorphism* which we have not yet defined. We will define this later, so for now, the reader who is unfamiliar with this notion can just insert "homeomorphism" anywhere "orientation preserving homeomorphism" is used without greatly changing the meaning.

Let $S = S_g$ denote a closed orientable surface of genus g (with a reference geometric metric so $S = \mathbb{X}^2/\Gamma$). We define the *moduli space of genus g geometric surfaces*, $\mathcal{M}(S_g) = \mathcal{M}_g$ as a set to be the set of all "area normalized" geometric surfaces $X = \mathbb{X}^2/\Gamma$ homeomorphic to S_g, up to the equivalence relation of orientation preserving isometry. Here the area normalization only refers to the case of $g = 1$, where we require the tori to have area 1 (the Gauss-Bonnet Theorem 7.3 tells us that any two genus $g \neq 1$ geometric surfaces already have the same area, so no normalization is required for the other surfaces). The elements of $\mathcal{M}_g$ are thus equivalence classes of geometric surfaces, where the equivalence relation $X \sim Y$ is that there exists an orientation preserving isometry $f : X \to Y$. We will denote the equivalence class of X by $[X]$, but will usually just write X for points in $\mathcal{M}_g$.

To make this into a metric space, we define a metric on $\mathcal{M}_g$ by $d_{\mathcal{M}}(X, Y) = \log(K(X, Y))$, where $K(X, Y)$ is the infimum of all constants $K \geq 1$ for which there exists an orientation preserving *K–biLipschitz map* $f : X \to Y$, that is f satisfies

$$\frac{1}{K} d_X(x, y) \leq d_Y(f(x), f(y)) \leq K d_X(x, y).$$

We cannot yet prove that this is actually a metric: it may take the value ∞ for some pair of surfaces. We must prove that there exists an orientation preserving biLipschitz map between any two homeomorphic geometric surfaces, and this will be done in the next two lectures.

Exercise 10.1. *Prove that if $d_{\mathcal{M}}$ is finite for any pair of surfaces, then it is indeed a metric. Hint: The main difficulty lies in proving $d_{\mathcal{M}}(X, Y) = 0 \Rightarrow X = Y$, and the hint is to think about Theorem 2.14.*

To better understand the topology of $\mathcal{M}_g$ we describe a space which is a priori more complicated, but which will turn out to be a kind of "universal covering" for $\mathcal{M}_g$. First, we define a *geometric surface marked by S* to be a geometric surface X together with an orientation preserving homeomorphism $f : S \to X$. We can think of the marking as a way for the surface S to "wear" the geometric metric: the homeomorphism $f : S \to X$ tells us how S is going to "put on" the geometric surface X.

We also define an equivalence relation on marked geometric surfaces by declaring $(f : S \to X) \sim (h : S \to Y)$ if there exists an isometry $\sigma : X \to Y$ so that $\sigma \circ f$ is homotopic to h. The equivalence class of $f : S \to X$ is denoted $[f : S \to X]$. As a set, the *Teichmüller space of S*, denoted $\mathcal{T}(S)$, is defined to be the set of equivalence classes of geometric surfaces marked by S. The metic on $\mathcal{T}(S)$ is defined by

$$d_{\mathcal{T}}([f : S \to X], [h : S \to Y]) = \log\left(\inf_{\sigma \simeq h \circ f^{-1}} K(\sigma)\right).$$

Here the infimum is taken over all σ which are biLipschitz maps homotopic to $h \circ f^{-1}$, and $K(\sigma)$ is the infimal biLipschitz constant of σ. As with $d_{\mathcal{M}}$, we postpone the proof of finiteness of $d_{\mathcal{T}}$ for now.

Exercise 10.2. *Prove that if $d_{\mathcal{T}}$ is finite for all pairs in $\mathcal{T}(S)$, then $d_{\mathcal{T}}$ is a metric.*

An alternative definition often used in the literature is the following. A *geometric structure* on S is a geometric metric d on S defining the original topology on S. Two geometric structures d and d' are equivalent if there is an isometry $f : (S, d) \to (S, d')$ which is homotopic to the identity. Then we define $\mathcal{T}'(S)$ as the set of equivalence classes of (area normalized) geometric structures on S with distance $d_{\mathcal{T}'}(d, d')$ given as the logarithm of the infimum of biLipschitz constants for maps $f : (S, d) \to (S, d')$ homotopic to the identity. The next exercise states that $\mathcal{T}(S)$ and $\mathcal{T}'(S)$ are essentially the same.

Exercise 10.3. *Assuming both of $d_{\mathcal{T}}$ and $d_{\mathcal{T}'}$ are finite on all pairs of points and hence define a metric by the previous two exercises, construct an isometry $(\mathcal{T}(S), d_{\mathcal{T}}) \to (\mathcal{T}'(S), d_{\mathcal{T}'})$. Hint: Given $[f : S \to X]$, define a new metric on S by pulling back the metric on X.*

Given any point $[f : S \to X] \in \mathcal{T}(S)$, we can "forget" the marking and just remember X. This defines a map

$$\Phi : \mathcal{T}(S_g) \to \mathcal{M}_g$$

we call the *forgetful map*. To see that we do indeed get a well-defined map, note that if $(f : S \to X) \sim (h : S \to Y)$, then X and Y are isometric by an isometry homotopic to $h \circ f^{-1}$. In particular X and Y are isometric (by an orientation preserving isometry), so $[X] = [Y]$.

Exercise 10.4. *Prove that Φ is 1–Lipschitz. That is,*

$$d_{\mathcal{M}}(X, Y) \leq d_{\mathcal{T}}([f : S \to X], [h : S \to Y]).$$

In particular, Φ is continuous.

10.2 The mapping class group

Let Homeo(S) denote the group of homeomorphisms of S, and Homeo$^+(S)$ the index two subgroup consisting of those that are orientation preserving. The *mapping class group of S* is the group $\mathrm{Mod}(S) = \mathrm{Homeo}^+(S)/\simeq$, that is, the group of homotopy classes of orientation preserving homeomorphisms. We also define the *extended mapping class group of S*, $\mathrm{Mod}^{\pm}(S) = \mathrm{Homeo}(S)/\simeq$. The mapping class group acts on $\mathcal{T}(S)$ by

$$[\phi] \cdot [f : S \to X] = [f \circ \phi^{-1} : S \to X].$$

Well-definedness is a consequence of the equivalence relation defining $\mathcal{T}(S)$. The next Proposition states that Φ is in fact a (set) quotient by the action.

Proposition 10.1. *Given $[f : S \to X], [h : S \to Y]$, then $\Phi([f : S \to X]) = \Phi([h : S \to Y])$ if and only if there exists $[\phi] \in \mathrm{Mod}(S)$ such that $[\phi] \cdot [f : S \to X] = [h : S \to Y]$.*

Proof. By definition of the action

$$\Phi([\phi] \cdot [f : S \to X]) = \Phi([f \circ \phi^{-1} : S \to X]) = [X] = \Phi([f : S \to X])$$

which proves one implication.

Conversely, if $[X] = \Phi([f : S \to X]) = \Phi([h : S \to Y]) = [Y]$, then letting $\sigma : X \to Y$ be an orientation preserving isometry (which exists since $[X] = [Y]$) set

$$\phi = h^{-1} \circ \sigma \circ f : S \to S.$$

since $[\sigma^{-1} \circ h : S \to X] = [h : S \to Y]$ (why?) we have

$$[\phi] \cdot [f : S \to X] = [f \circ f^{-1} \circ \sigma^{-1} \circ h : S \to X] = [\sigma^{-1} \circ h : S \to X] = [h : S \to Y].$$

□

10.3 The torus

The Teichmüller space of the torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ has a very concrete (and familiar!) description which we now explain.

We know that $\pi_1(\mathbb{T}^2) \cong \mathbb{Z}^2$. There are many choices for such an isomorphism, any two of which differ by composing with an element of $\mathrm{GL}_2(\mathbb{Z})$. We declare two such isomorphisms $\pi_1(\mathbb{T}^2) \to \mathbb{Z}^2$ to be equivalent if they differ by an element of $\mathrm{SL}_2(\mathbb{Z})$. There are exactly two equivalence classes, and we call a choice of one of these an *orientation* on $\mathbb{T}^2$ (this agrees with the usual notion). Given two tori T_1, T_2 with orientations $[\psi_i : \pi_1(T_i) \to \mathbb{Z}^2]$, a homeomorphism $f : T_1 \to T_2$ is called *orientation preserving* if $[\psi_2 \circ f_*] = [\psi_1]$.

For the flat torus $\mathbb{T}^2$, the action of $\mathbb{Z}^2$ on $\mathbb{R}^2$ gives a preferred isomorphism $\pi_1(\mathbb{T}^2) \cong \mathbb{Z}^2$. The action of a homeomorphism on the fundamental group induces a homomorphism

$$\widetilde{\Psi} : \mathrm{Homeo}^+(\mathbb{T}^2) \to \mathrm{SL}_2(\mathbb{Z})$$

where $\widetilde{\Psi}(f) = f_*$ and we use the preferred isomorphism $\pi_1(\mathbb{T}^2) \cong \mathbb{Z}^2$.

Theorem 10.2. *The homomorphism $\widetilde{\Psi}$ descends to an isomorphism $\Psi : \mathrm{Mod}(\mathbb{T}^2) \to \mathrm{SL}_2(\mathbb{Z})$.*

Proof. To see that $\widetilde{\Psi}$ is onto, let $A \in \mathrm{SL}_2(\mathbb{Z})$, and view this as a map $A : \mathbb{R}^2 \to \mathbb{R}^2$. Since A preserves $\mathbb{Z}^2$, it follows that A descends to a map $f_A : \mathbb{T}^2 \to \mathbb{T}^2$: to see this, it suffices to observe that for all $\mathbf{x} \in \mathbb{R}^2$ and $\mathbf{n} \in \mathbb{Z}^2$, $A(\mathbf{x} + \mathbf{n}) - A\mathbf{x} = A\mathbf{n} \in \mathbb{Z}^2$, so $A\mathbf{x} \sim A\mathbf{y}$ if and only if $\mathbf{x} \sim \mathbf{y}$. By construction $(f_A)_* = A$, so $\widetilde{\Psi}$ is onto.

To prove that the kernel of $\widetilde{\Psi}$ is precisely the set of homeomorphisms homotopic to the identity, it suffices to prove that for any $f : \mathbb{T}^2 \to \mathbb{T}^2$ we have $f \simeq f_A$, where $A = \widetilde{\Psi}(f) = f_*$. To do this, we observe that by a homotopy we can assume that f sends the image of $\mathbf{0} \in \mathbb{R}^2$ in $\mathbb{T}^2$ to itself. Next we lift f to a map $\widetilde{f} : \mathbb{R}^2 \to \mathbb{R}^2$: that is, $p \circ \widetilde{f} = f \circ p$, where $p : \mathbb{R}^2 \to \mathbb{T}^2$ is the covering map. By translating by an element of $\mathbb{Z}^2$, the covering group, we can assume that $\widetilde{f}(\mathbf{0}) = \mathbf{0}$, and hence $\widetilde{f}(\mathbb{Z}^2) = \mathbb{Z}^2$. Moreover, since $f_* = A$, it follows that the restriction of $\widetilde{f}$ to $\mathbb{Z}^2$ is precisely A.

Exercise 10.5. *Prove that the straight-line homotopy $F(\mathbf{x}, t) = (1 - t)\widetilde{f}(\mathbf{x}) + tA\mathbf{x}$ descends to a homotopy from f to f_A.*

This completes the proof of the theorem. □

According to Theorem 9.1 (and Exercises 8.2 and 8.3), every geometric torus is of the form $X = \mathbb{R}^2/\Lambda$, where $\Lambda < \mathbb{R}^2 < \mathrm{Isom}^+(\mathbb{R}^2)$ where Λ acts properly discontinuously. As in the proof of the previous theorem, any marking $f : \mathbb{T}^2 \to X$ can be taken to have the form $f = f_A$, where $A \in \mathrm{SL}_2(\mathbb{R})$ and $A \cdot \mathbb{Z}^2 = \Lambda_A = \Lambda$. We write $X = X_A = \mathbb{R}^2/\Lambda_A$ and $f_A : \mathbb{T}^2 \to X_A$.

Given two such $f_A : \mathbb{T}^2 \to X_A$ and $f_B : \mathbb{T}^2 \to X_B$, we first want to know when these are equivalent. This requires that we look at the map

$$f_B \circ f_A^{-1} : X_A \to X_B.$$

Proposition 10.3. *The map $f_B \circ f_A^{-1} : X_A \to X_B$ has minimal biLipschitz constant among all biLipschitz maps homotopic to $f_B \circ f_A^{-1}$. In particular, if we denote this constant by $K(BA^{-1})$, then*

$$d_{\mathcal{T}}([f_A : \mathbb{T}^2 \to X_A], [f_B : \mathbb{T}^2 \to X_B]) = \log(K(BA^{-1})).$$

Before we give the proof we need a fact about curves on geometric tori. A closed curve $\gamma : S^1 \to X_A$ is *essential* if it is not homotopic to a constant map. Every essential closed curve determines an element of $\pi_1(X_A)$ and via the canonical isomorphism $\pi_1(X_A) \cong A \cdot \mathbb{Z}^2$, we get a vector $\mathbf{v}_\gamma \in A \cdot \mathbb{Z}^2 \subset \mathbb{R}^2$.

Exercise 10.6. *Given an essential curve $\gamma : S^1 \to X_A$, let $\mathbf{v} = \mathbf{v}_\gamma$. Prove that γ is homotopic to the curve $\gamma_{\mathbf{v}} : S^1 \to X_A$ given by*

$$\gamma_{\mathbf{v}}\left(e^{i\theta}\right) = p\left(\frac{\theta}{2\pi}\mathbf{v}\right)$$

where $p : \mathbb{R}^2 \to X_A$ is the quotient map. Furthermore, prove that the length of $\gamma_{\mathbf{v}}$ is $|\mathbf{v}|$, and that this minimizes length among all closed curves homotopic to γ.

For any $\mathbf{v} \in A \cdot \mathbb{Z}^2$, the curve $\gamma_{\mathbf{v}}$ is called a *closed geodesic* and for $\mathbf{v} = \mathbf{v}_\gamma$, we also call this the *geodesic representative of* γ.

Proof. Lifting $f_B \circ f_A^{-1}$ to $\mathbb{R}^2$ fixing $\mathbf{0}$ we exactly get the linear map $BA^{-1} : \mathbb{R}^2 \to \mathbb{R}^2$. The singular value decomposition for this matrix expresses this as $BA^{-1} = R_1 g_t R_2$, where $R_1, R_2 \in SO(2)$ are rotations, $t \geq 0$ and

$$g_t = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}.$$

If $t = 0$ then $BA^{-1} = R_1 R_2$ is an isometry, and hence so is $f_B \circ f_A^{-1}$.

Now suppose $t > 0$. Since BA^{-1} is linear, it is C^1 and the derivative is given again by BA^{-1}. Given any vector $\mathbf{v} \in \mathbb{R}^2$ we have

$$e^{-t}|\mathbf{v}| \leq |AB^{-1}\mathbf{v}| \leq e^t|\mathbf{v}|.$$

Since the metrics on $\mathbb{R}^2$, X_A and X_B are given in terms of lengths of C^1 paths, this means that BA^{-1} is e^t–biLipschitz, and hence so is $f_B \circ f_A^{-1}$.

We need to see that in the homotopy class of $f_B \circ f_A^{-1}$, we cannot find a biLipschitz map with smaller biLipschitz constant. For this, observe that if there is a K–biLipschitz map homotopic to $f_B \circ f_A^{-1}$, then for any closed curve $\gamma : S^1 \to X_A$, we have

$$\ell_{X_B}(f \circ \gamma) \leq K \ell_{X_A}(\gamma).$$

Now observe that if $\mathbf{v} = R_2^{-1}\mathbf{e}_1$, then $|BA^{-1}\mathbf{v}| = e^t|\mathbf{v}|$. So, let $\{\mathbf{v}_n\}_{n=1}^{\infty} \subset A \cdot \mathbb{Z}^2$ be any sequence of vectors for which

$$\lim_{n\to\infty} \frac{\mathbf{v}_n}{|\mathbf{v}_n|} = \frac{\mathbf{v}}{|\mathbf{v}|}$$

then

$$\lim_{n\to\infty} \frac{\ell_{X_B}(f_B \circ f_A^{-1} \circ \gamma_{\mathbf{v}_n})}{\ell_{X_A}(\gamma_{\mathbf{v}_n})} = \lim_{n\to\infty} \frac{|BA^{-1}\mathbf{v}_n|}{|\mathbf{v}_n|} = \frac{e^t|\mathbf{v}|}{|\mathbf{v}|} = e^t.$$

The curve $f_B \circ f_A^{-1} \circ \gamma_{\mathbf{v}_n} = \gamma_{BA^{-1}\mathbf{v}_n}$ is a geodesic. So, given a map $f \simeq f_B \circ f_A^{-1}$, we must have $\ell_{X_B}(f \circ \gamma_{\mathbf{v}_n}) \geq \ell_{X_B}(f_B \circ f_A^{-1} \circ \gamma_{\mathbf{v}_n})$, and hence such a map f can have no better biLipschitz constant than e^t. This completes the proof. □

Corollary 10.4. *Given $f_A : \mathbb{T}^2 \to X_A$ and $f_B : \mathbb{T}^2 \to X_B$, the following are equivalent.*

1. $(f_A : \mathbb{T}^2 \to X_A) \sim (f_B : \mathbb{T}^2 \to X_B)$,
2. $f_B \circ f_A^{-1}$ *is an isometry,*
3. $BA^{-1} \in SO(2)$.

Therefore, if we let $SO(2)\backslash \mathrm{SL}_2(\mathbb{R})$ denote the set of cosets of $SO(2)$ in $\mathrm{SL}_2(\mathbb{R})$, then we have a bijection

$$F : SO(2)\backslash \mathrm{SL}_2(\mathbb{R}) \to \mathcal{T}(\mathbb{T}^2)$$

given by $F(SO(2)A) = [f_A : \mathbb{T}^2 \to X_A]$. We also have a bijection

$$G : SO(2)\backslash \mathrm{SL}_2(\mathbb{R}) \to \mathbb{U}^2$$

given by $\Psi(SO(2)A) = A^{-1}(i)$ where $A \in \mathrm{SL}_2(\mathbb{R})$ is acting on $\mathbb{U}^2$ by the associated Möbius transformation in $\mathrm{PSL}_2(\mathbb{R})$.

Exercise 10.7. *Prove that these are indeed both bijections.*

Theorem 10.5. $H = G \circ F^{-1} : (\mathcal{T}(\mathbb{T}^2), d_{\mathcal{T}}) \to (\mathbb{U}^2, \frac{1}{2}d_{\mathbb{U}})$ *is an isometry. Moreover, H is equivariant with respect to Ψ:*

$$H([\phi] \cdot [f : \mathbb{T}^2 \to X]) = \Psi([\phi]) \cdot H([f : \mathbb{T}^2 \to X]).$$

Consequently, $\mathcal{M}_1 = \mathbb{U}^2/\mathrm{PSL}_2(\mathbb{Z})$.

Proof. First we check equivariance. Let $[f_A : \mathbb{T}^2 \to X_A]$ and $[f_B] \in \mathrm{Mod}(\mathbb{T}^2)$ where $\Psi([f_B]) = B \in \mathrm{SL}_2(\mathbb{Z})$. Then

$$\begin{aligned} H([f_B] \cdot [f_A : \mathbb{T}^2 \to X_A]) &= H([f_A \circ f_B^{-1} : \mathbb{T}^2 \to X_A]) = H([f_{AB^{-1}} : \mathbb{T}^2 \to X_A]) \\ &= G(AB^{-1}) = B \cdot A^{-1}(i) = \Psi([f_B]) \cdot H([f_A : \mathbb{T}^2 \to X_A]) \end{aligned}$$

as required.

Given $z, w \in \mathbb{U}^2$, if we let $A \in \mathrm{SL}_2(\mathbb{R})$ be an element that takes the geodesic containing z and w to $i\mathbb{R}_+$ with $A(z) = i$ and $A(w) = e^{-2t}i = g_{-t}(i)$ for $t > 0$. Then $d_{\mathbb{U}}(z, w) = d_{\mathbb{U}}(i, e^{-2t}i) = 2t$.

Now we have $G(A) = A^{-1}(i) = z$ and $G(g_tA) = w$ while $F(A) = [f_A : \mathbb{T}^2 \to X_A]$ and $F(g_tA) = [f_{g_tA} : \mathbb{T}^2 \to X_{g_tA}]$. So by Proposition 10.3

$$d_{\mathcal{T}}(H^{-1}(z), H^{-1}(w)) = d_{\mathcal{T}}(F(A), F(g_tA)) = \log(K(g_tAA^{-1})) = \log(K(g_t)) = t = \frac{1}{2}d_{\mathbb{U}}(z, w).$$

□

11 Isomorphisms and biLipschitz homeomorphisms I.

We assume for the next two lectures that $X = \mathbb{H}^2/\Gamma$ and $Y = \mathbb{H}^2/\Lambda$ are orientable hyperbolic surfaces, $\Gamma, \Lambda < \mathrm{Isom}^+(\mathbb{H}^2)$. Our goal for these two lectures is to explain how an arbitrary homeomorphism $f : X \to Y$ can be homotoped to a biLipschitz homeomorphism. In fact, we will prove something stronger.

Let $\mathrm{BiLip}(X, Y) \subset \mathrm{Homeo}(X, Y)$ denote the set of biLipschitz homeomorphisms from X to Y and $\mathrm{Iso}(\Gamma, \Lambda)$ the set of isomorphisms from Γ to Λ. Since $X = \mathbb{H}^2/\Gamma$ and $Y = \mathbb{H}^2/\Lambda$, we have canonical isomorphisms $\pi_1(X) \cong \Gamma$ and $\pi_1(Y) \cong \Lambda$. An element $f : X \to Y \in \mathrm{Homeo}(X, Y)$ induces an isomorphism $f_* : \Gamma \to \Lambda$ up to conjugation (since we have not specified a basepoint). Therefore, we have a map

$$\mathrm{Homeo}(X, Y) \to \mathrm{Iso}(\Gamma, \Lambda)/conj$$

given by $f \mapsto f_*$. We are interested in the restriction of this to $\mathrm{BiLip}(X, Y)$.

Theorem 11.1. *The map* $\mathrm{BiLip}(X, Y) \to \mathrm{Iso}(\Gamma, \Lambda)/conj$ *is surjective. Consequently, the same is true for all of* $\mathrm{Homeo}(X, Y)$*. Moreover, if f_* and h_* differ by conjugation for some $f, h \in \mathrm{Homeo}(X, Y)$, then $f \simeq h$.*

As a corollary, we obtain the *Dehn-Nielsen-Baer Theorem.*

Corollary 11.2. *For any closed orientable surface $S = S_g$ of genus $g \geq 2$, the map*

$$\mathrm{Mod}^{\pm}(S) \to \mathrm{Out}(\pi_1(S))$$

given by $[f] \mapsto [f_*]$ is an isomorphism.

Here $\mathrm{Out}(\pi_1(S)) = \mathrm{Aut}(\pi_1(S))/\mathrm{Inn}(\pi_1(S))$ is the group of automorphisms of $\pi_1(S)$, up to conjugation.

Proof. We view $S = S_g$ as a geometric surface $S_g = \mathbb{X}^2/\Gamma$. Since $g > 1$, $\chi(S_g) < 0$, and so Theorem 7.3 implies $\mathbb{X}^2 = \mathbb{H}^2$. Therefore, setting $X = S_g = Y$ in Theorem 11.1, it follows that $\mathrm{Mod}^{\pm}(S_g) \to \mathrm{Out}(\Gamma)$ is an isomorphism. $\square$

11.1 Straight line homotopy

The second half of Theorem 11.1 is quite easy and we dispense with that now.

Proposition 11.3. *Given continuous maps $f, h : X \to Y$, if f_* and h_* differ by conjugation, then $f \simeq h$.*

Proof. This is much like Exercise 10.5. We lift each of f and h to the universal covers

$$\widetilde{f}, \widetilde{h} : \mathbb{H}^2 \to \mathbb{H}^2$$

so that these are equivariant with respect to f_* and h_*:

$$\widetilde{f}(\gamma(x)) = f_*(\gamma)(\widetilde{f}(x)) \quad \text{and} \quad \widetilde{h}(\gamma(x)) = h_*(\gamma)(\widetilde{h}(x))$$

for all $\gamma \in \Gamma$ and $x \in \mathbb{H}^2$.

Any other other choice of lift $\widetilde{h}'$ for h differs from $\widetilde{h}$ by composing with some covering transformation $\sigma \in \Lambda$, so that $\widetilde{h}' = \sigma\widetilde{h}$. Then observe that

$$\widetilde{h}'(\gamma(x)) = \sigma\widetilde{h}(\gamma(x)) = \sigma h_*(\gamma)\widetilde{h}(x) = \sigma h_*(\gamma)\sigma^{-1}\sigma\widetilde{h}(x) = (\sigma h_*(\gamma)\sigma^{-1})\widetilde{h}'(x).$$

Thus, $\widetilde{h}'$ is equivariant with respect to $\gamma \mapsto \sigma h_*(\gamma)\sigma^{-1}$.

Since h_* and f_* are conjugate, there exists σ so that $\sigma h_*(\gamma)\sigma^{-1} = f_*(\gamma)$ for all $\gamma \in \Gamma$, so $\widetilde{h}'$ is also equivariant with respect to f_*. Therefore the "straight line homotopy" from $\widetilde{f}$ to $\widetilde{h}'$ is also equivariant. This can be done explicitely, for example, using the hyperboloid model and the projection from the cone $\Pi : C^+ \to \mathcal{H}^2 = \mathbb{H}^2$ as in Section 6.2. More precisely, define

$$\widetilde{H}(x,t) = \Pi((1-t)\widetilde{f}(x) + t\widetilde{h}'(x)).$$

Observe that for all $\gamma \in \Gamma$ and $x \in \mathcal{H}^2$ we have

$$(1-t)\widetilde{f}(\gamma(x)) + t\widetilde{h}'(\gamma(x)) = (1-t)f_*(\gamma)\widetilde{f}(x) + tf_*(\gamma)\widetilde{h}'(x) = f_*(\gamma)((1-t)\widetilde{f}(x) + t\widetilde{h}'(x))$$

where the last equality holds since $f_*(\gamma) \in \Lambda < O^+(2,1)$ acts linearly on $\mathbb{R}^{2,1}$. Since any element of $O^+(2,1)$ commutes with Π, we see that $\widetilde{H}$ is equivariant with respect to f_*. It follows that $\widetilde{H}$ descends to a homotopy $H : X \times I \to Y$ from f to h. ☐

11.2 Lipschitz maps

As a first approximation to Theorem 11.1 we prove the following.

Proposition 11.4. *Given an isomorphism $\phi : \Gamma \to \Lambda$, there exists Lipschitz maps $f : X \to Y$ and $h : Y \to X$ which are homotopy inverses to each other so that $f_* = \phi$ and $h_* = \phi^{-1}$.*

Proof. It suffices to construct a Lipschitz map $f : X \to Y$ so that $f_* = \phi$. To see this, observe that by applying the same construction to ϕ^{-1} we can produce a Lipschitz map $h : Y \to X$ so that $h_* = \phi^{-1}$, and since $(h \circ f)_* = h_* \circ f_* = \phi^{-1} \circ \phi = id_\Gamma$, Proposition 11.3 implies $h \circ f$ is homotopic to the identity on X (and similarly $(f \circ h)$ is homotopic to the identity on Y).

To construct the Lipschitz map, take a geodesic triangulation on X, and suppose the basepoint $x \in X$ giving $\pi_1(X,x) \cong \Gamma$ is one of the vertices. Let T be a maximal tree in the 1–skeleton of the triangulation (the 1–skeleton is the graph consisting of all edges and vertices of the triangulation). Let $e_1, \ldots, e_n$ be the set of edges in the 1–skeleton *not* in T which we arbitrarily orient by giving a direction. Associated to each e_i, there is a loop γ_i representing an element in $\pi_1(X,x)(\cong \Gamma)$ given by starting at x, following a path to the initial vertex of e_i inside T, traversing e_i in the designated direction, then following a path back to x inside T.

Now choose a basepoint $y \in Y$ so that $\Lambda \cong \pi_1(Y,y)$ and construct a map $f' : X \to Y$ as follows. First, let f' map the entire tree T to y. Since ϕ is a homomorphism, $\phi(\gamma_i)$ is some element of $\pi_1(Y,y)$, and so we can define f' on the edge e_i to be any representative loop of $\phi(\gamma_i)$ (since both endpoints of e_i are sent to y, this makes sense). We have defined a continuous map f' on the 1–skeleton of X and we just need to extend it over the boundaries of triangles. This is possible because the boundaries of the triangles are null-homotopic loops in X, and ϕ sends such loops to null-homotopic loops. Therefore, we extend over the triangles in any (continuous) way, and then $f' : X \to Y$ is a continuous map with $f'_* = \phi$, by the next exercise.

Exercise 11.1. *To see that $f'_* = \phi$, prove that the loops $\{\gamma_i\}$ generate $\pi_1(X,x)$, so any homomorphism is determined by the values on $\{\gamma_i\}$.*

To make this Lipschitz, we modify f' on each triangle Δ as follows. We can lift the restriction $f'|_\Delta : \Delta \to Y$ to the universal covering of Y, $\widetilde{f'|_\Delta} : \Delta \to \mathbb{H}^2$. As in Section 6.2 there is a canonical triangle map ψ_Δ that agrees with this map on the vertices (the domain is an actual hyperbolic triangle, though the target may be degenerate—this makes no difference in the construction of the canonical triangle map). This map is Lipschitz, and so composing with the universal covering $p : \mathbb{H}^2 \to Y$, we obtain a Lipschitz map $p \circ \psi_\Delta : \Delta \to Y$. We define f to be equal to f' on the vertex set of the triangulation of X and $f|_\Delta = p \circ \psi_\Delta$ for each triangle Δ. The next exercise completes the proof.

Exercise 11.2. *Prove that this does indeed define a Lipschitz map f (you need to check that this defines a map, and that the map is Lipschitz). Also prove that $f_* = \phi$.*

□

11.3 Extensions to S^1_∞

We now suppose $f : X \to Y$ and $h : Y \to X$ are the Lipschitz homotopy inverses given by Proposition 11.4 from an isomorphism $\phi : \Gamma \to \Lambda$. We choose lifts of these to the universal covers

$$\widetilde{f}, \widetilde{h} : \mathbb{H}^2 \to \mathbb{H}^2.$$

Proposition 11.5. *The map $\widetilde{f} : \mathbb{H}^2 \to \mathbb{H}^2$ extends continuously to $\widetilde{f} : \overline{\mathbb{H}}^2 \to \overline{\mathbb{H}}^2$, equivariant with respect to ϕ. Furthermore, the restriction to the boundary is a homeomorphism*

$$\partial \widetilde{f} = \widetilde{f}|_{S^1_\infty} : S^1_\infty \to S^1_\infty.$$

Moreover, $\partial \widetilde{f}$ depends only on ϕ.

Symmetrically, the same is true for $\widetilde{h}$. The proof will occupy the remainder of this lecture.

As lifts of Lipschitz maps, $\widetilde{f}$ and $\widetilde{h}$ are Lipschitz and equivariant with respect to ϕ and ϕ^{-1}, respectively. Not only are these Lipschitz so that distances are increased by a controlled amount, but in fact, the contraction of distances is also controlled. More precisely, we have the following.

Lemma 11.6. *There exists $K \geq 1$ and $C \geq 0$ so that for all $x, y \in \mathbb{H}^2$*

$$\frac{1}{K} d_{\mathbb{H}}(x, y) - C \leq d_{\mathbb{H}}(\widetilde{f}(y), \widetilde{f}(y)) \leq K d_{\mathbb{H}}(x, y).$$

Proof. Let K be the maximum of the Lipschitz constants of $\widetilde{f}$ and $\widetilde{h}$. Then the second inequality in the Lemma holds.

Next, let $\widetilde{H}$ denote the straight line homotopy from $\widetilde{h} \circ \widetilde{f}$ to the identity on $\mathbb{H}^2$. Because $\widetilde{h} \circ \widetilde{f}$ is equivariant with respect to the identity isomorphism $\Lambda \to \Lambda$, and since X is compact, there is a uniform bound $c \geq 0$ for the distance between x and $\widetilde{h} \circ \widetilde{f}(x)$, for all $x \in \mathbb{H}^2$ (consider the distance between x and $\widetilde{h} \circ \widetilde{f}(x)$ for all x in a compact Dirichlet domain). Then we have

$$d_{\mathbb{H}}(x, y) \leq d_{\mathbb{H}}(\widetilde{h} \circ \widetilde{f}(x), \widetilde{h} \circ \widetilde{f}(y)) + 2c \leq K d_{\mathbb{H}}(\widetilde{f}(x), \widetilde{f}(y)) + 2c$$

So setting $C = 2c/K$ completes the proof. □

We also need the following "stability" fact for geodesics.

Lemma 11.7. *There exists a constant $R > 0$ such that for any (biinfinite) geodesic σ in $\mathbb{H}^2$, there exists a geodesic α in $\mathbb{H}^2$ so that*

$$\widetilde{f}(\sigma) \subset N_R(\alpha) \text{ and } \alpha \subset N_R(\widetilde{f}(\sigma)).$$

Here $N_R(\alpha)$ is the R–neighborhood of α. The proof requires the following.

Exercise 11.3. *Given a geodesic $\alpha \subset \mathbb{H}^2$, let $\pi : \mathbb{H}^2 \to \alpha$ denote the "orthogonal projection"; that is, $\pi(x)$ is the intersection point between α and the geodesic through x orthogonal to α. Prove that if β is a C^1 path that lies outside the A–neighborhood of α, then*

$$\ell(\pi \circ \beta) \leq \frac{1}{\cosh(A)} \ell(\beta).$$

Hint: For this, consider the C^1–parameterization $F : \mathbb{R}^2 \to \mathcal{H}^2 \subset \mathbb{R}^{2,1}$ given by

$$F(x, y) = (\cosh(x)\cosh(y), \sinh(x)\cosh(y), \sinh(y)).$$

If α is the geodesic $t \mapsto F(t, 0)$, then π is given by $\pi(F(x, y)) = F(x, 0)$, and the distance from $F(x, y)$ to $F(x, 0) = \pi(F(x, y))$ is exactly $|y|$.

Proof of Lemma 11.7. Parameterize the geodesic by unit speed $\sigma : \mathbb{R} \to \mathbb{H}^2$. For $a < b$ we consider the segment of σ defined on $[a, b]$. Let α be the geodesic in $\mathbb{H}^2$ through $\sigma(a)$ and $\sigma(b)$.

Given $A > 0$, let B be the maximum length of a segment in $[a, b]$ for which the image under $\widetilde{f} \circ \sigma$ lies outside the A–neighborhood.

Claim. If $\cosh(A) > K^2$ then $B \leq \frac{K \cosh(A)}{\cosh(A) - K^2}(2A + C)$.

Proof of Claim. Suppose $a < a_0 < b_0 < b$ with $b_0 - a_0 = B$ such that $d_{\mathbb{H}}(\widetilde{f} \circ \sigma(t), \alpha) \geq A$ for all $a_0 \leq t \leq b_0$. According to Lemma 11.6

$$d_{\mathbb{H}}(\widetilde{f} \circ \sigma(a_0), \widetilde{f} \circ \sigma(b_0)) \geq \frac{1}{K}(b_0 - a_0) - C = \frac{B}{K} - C. \tag{9}$$

From Lemma 11.6 and Exercise 11.3, the orthogonal projection sends $\widetilde{f} \circ \sigma([a_0, b_0])$ to a path along α of length at most $K(b_0 - a_0)/\cosh(A) = KB/\cosh(A)$. So, since the distance from $\widetilde{f} \circ \sigma(a_0)$ to this initial point of this projected path has length A, and likewise for the distance from $\widetilde{f} \circ \sigma(b_0)$ to the terminal point of this path, the triangle inequality implies.

$$d_{\mathbb{H}}(\widetilde{f} \circ \sigma(a_0), \widetilde{f} \circ \sigma(b_0)) \leq 2A + \frac{KB}{\cosh(A)}. \tag{10}$$

Combining (9) and (10) we have

$$\frac{B}{K} - C \leq 2A + \frac{KB}{\cosh(A)}.$$

Since we are assuming $\cosh(A) > K^2$, we can isolate B to find

$$B \leq \frac{K \cosh(A)}{\cosh(A) - K^2}(2A + C).$$

which proves the claim. □

Since the distance from any point of $\widetilde{f} \circ \sigma([a,b])$ to α is at most the A plus the maximal length of a segment that lies outside $N_A(\alpha)$, it follows from Lemma 11.6 that this distance is at most

$$A + KB \leq A + \frac{K^2 \cosh(A)}{\cosh(A) - K^2}(2A + C) \tag{11}$$

Now set R to be the right hand side of (11), and consider the sequence of geodesics α_n in $\mathbb{H}^2$ which connect $\widetilde{f} \circ \sigma(-n)$ to $\widetilde{f} \circ \sigma(n)$. All these geodesics intersect $\overline{B}_R(\widetilde{f} \circ \sigma(0))$ by the claim. Furthermore, $d_{\mathbb{H}}(\widetilde{f} \circ \sigma(0), \widetilde{f} \circ \sigma(\pm n)) \to \infty$ by Lemma 11.6 and so up to subsequence, these converge to a biinfinite geodesic $\alpha_\infty \subset \mathbb{H}^2$ (why?). Since $\widetilde{f} \circ \sigma([-n,n]) \subset N_R(\alpha_m)$ for all n and m with $n < m$, it follows that $\widetilde{f} \circ \sigma(\mathbb{R}) \subset N_R(\alpha_\infty)$.

If we consider the orthogonal projection $\pi : \mathbb{H}^2 \to \alpha_\infty$ from Exercise 11.3, then $\pi(\widetilde{f} \circ \sigma(\mathbb{R})) = \alpha_\infty$ (that is, all of α_∞ is in the image). Therefore, every point of α_∞ is within a distance R of $\widetilde{f} \circ \sigma(\mathbb{R})$, so $\alpha_\infty \subset N_R(\widetilde{f} \circ \sigma(\mathbb{R})) = N_R(\widetilde{f}(\sigma))$, as required. Since no two distinct geodesics in $\mathbb{H}^2$ remain a bounded distance apart, it follows that α_∞ is unique. □

Proof of Proposition 11.5. We define the extension to the boundary $\partial\widetilde{f} : S^1_\infty \to S^1_\infty$ by defining the image of a point $x \in S^1_\infty$ to be the positive endpoint of α where $\sigma \subset \mathbb{H}^2$ is a geodesic whose positive endpoint is x and α is the associated geodesic (from Lemma 11.7). This is well-defined since any two geodesics with the same positive endpoint are asymptotic (see Section 5.1), and hence their images remain a bounded distance from each other in the positive direction, so the associated geodesics are also asymptotic. Since any pair of distinct points in S^1_∞ determines a geodesic, the image of those points are the distinct endpoints of the associated geodesic, and so $\partial\widetilde{f}$ is injective.

To see that $\partial\widetilde{f}$ is continuous consider a sequence $\{x_n\} \subset S^1_\infty$ with $x_n \to x \in S^1_\infty$. For any point $y \in S^1_\infty$, with $y \neq x$ and $y \neq x_n$, the sequence of geodesics σ_n connecting y to x_n converge to the geodesic σ connecting y to x. The geodesics σ_n are closer and closer to σ on longer and longer segments, and so the geodesics α_n associated to σ_n remain a bounded distance from the geodesic α associated to σ on longer and longer segments. From this we see that the endpoints of the α_n converge to the endpoints of α, so that $\partial\widetilde{f}(x_n) \to \partial\widetilde{f}(x)$ and hence $\partial\widetilde{f}$ is continuous. The proof that the extension is continuous on all of $\overline{\mathbb{H}}^2$ is similar.

Exercise 11.4. *Prove that $\partial\widetilde{f} : S^1_\infty \to S^1_\infty$ is surjective. Hint: consider applying the Intermediate Value Theorem 2.7 appropriately.*

All that remains is to show that $\partial\widetilde{f}$ depends only on ϕ. Given any other map $f' : X \to Y$ with $f'_* = \phi$, Proposition 11.3 implies f and f' are homotopic, and we can lift this to a homotopy between $\widetilde{f}$ and $\widetilde{f}'$ (indeed, the proof of that proposition first constructs such a homotopy). Because this is a lifted homotopy defined on $X \times I$, which is compact, it follows that there exists $M > 0$ so that $d_{\mathbb{H}}(\widetilde{f}(x), \widetilde{f}'(x)) \leq M$ for all $x \in \mathbb{H}^2$. In the disk $\mathbb{D}^2 = \mathbb{H}^2$, it follows that if $\{x_n\} \subset \mathbb{D}^2$ is any sequence with $x_n \to x \in S^1_\infty$, then the *Euclidean distance* satisfies $|\widetilde{f}(x_n) - \widetilde{f}'(x_n)| \to 0$, so by continuity $\widetilde{f}(x) = \widetilde{f}'(x)$. □

We now define a homotopy equivalence $f : X \to Y$ to be *orientation preserving* if the induced homeomorphism $\partial\widetilde{f} : S^1_\infty \to S^1_\infty$ preserves the counterclockwise orientation—said differently, $(\partial\widetilde{f})_* : \pi_1(S^1_\infty) \to \pi_1(S^1_\infty)$ is the identity. By the last part of this proposition, this depends only on f_*. For those familiar with the usual notion of orientation, we give $X = \mathbb{H}^2/\Gamma$ the orientation coming from the standard orientation on $\mathbb{H}^2$.

12 Isomorphisms and biLipschitz homeomorphisms II.

In this lecture we complete the proof of Theorem 11.1.

12.1 Hexagon patterns

We have assumed that our "reference surface" $S = S_g$ is geometric (and thus hyperbolic since $g \geq 2$). Now we put a particular geometric structure on S and prove that for any hyperbolic surface $X = \mathbb{H}^2/\Gamma$, and any isomorphism $\phi : \pi_1(S) \to \Gamma$, there is a biLipschitz map $f : S \to X$ with $f_* = \phi$. Let $P \subset \mathbb{H}^2$ be a regular right-angled hexagon. Such a hexagon exists by the construction described in Section 7.2. We glue two of these together along alternating boundaries to produce a hyperbolic *pair of pants* with geodesic boundary; See Figure 7.

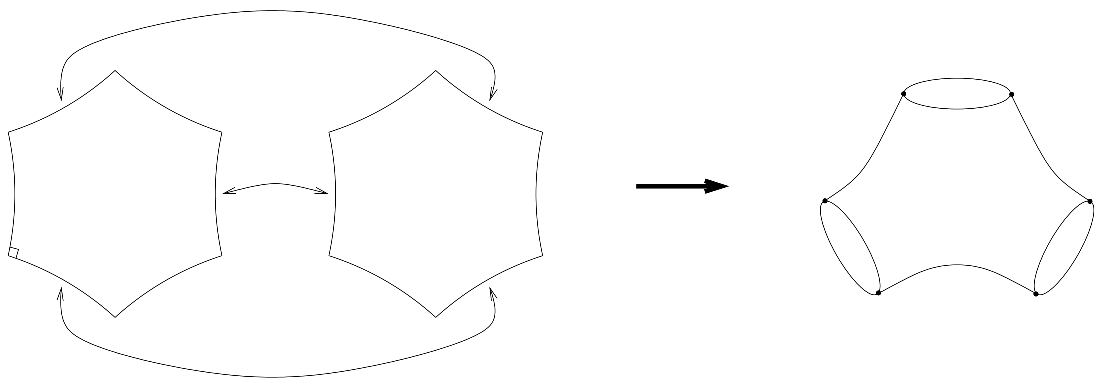

Figure 7: Two regular right-angled hexagons on the left, glued to give a pair of pants on the right.

We glue together $2g-2$ such pairs of pants to produce a genus g hyperbolic surface, and we assume here that the vertices of the hexagons are glued to vertices of hexagons. For concreteness, we can take these to be glued so that we arrive at the genus g surface shown in Figure 8 which we will take to be our reference hyperbolic surface $S = S_g = \mathbb{H}^2/G$ for $G < \mathrm{Isom}^+(\mathbb{H}^2)$.

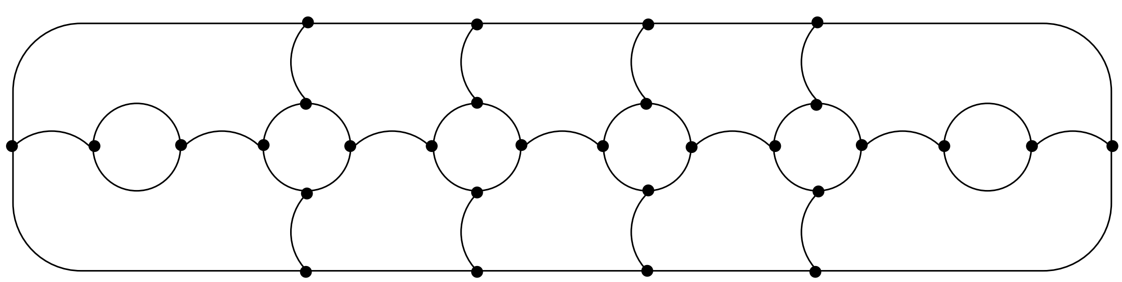

Figure 8: 10 pairs of pants or 20 hexagons glued together to produce a genus 6 surface. The vertices of the hexagons are shown as dots.

The boundaries of the pairs of pants give us $3g-3$ pairwise disjoint simple closed curves on S. This set of curves is called a *pants decomposition* of S. The curves in the pants decomposition can also be obtained by gluing together some of the sides of the hexagons. If we glue the remaining sides of the hexagons we get another set of pairwise disjoint simple closed curves on S which we will call the *transverse curves*. The number of transverse curves depends on the way in which the pants are glued together: In the example of Figure 8 we have 7 such curves.

Exercise 12.1. *Prove that the pants can be glued by isometries (with vertices glued to vertices) so that gluing the remaining edges of the hexagons gives just one simple closed curve.*

Using this geometric metric on S, we prove the following.

Proposition 12.1. *For any hyperbolic surface $X = \mathbb{H}^2/\Gamma$ and isomorphism $\phi : G \to \Gamma$ there exists an biLipschitz homeomorphism $f : S \to X$ for which $f_* = \phi$.*

Proof. Let $p : \mathbb{H}^2 \to S = \mathbb{H}^2/G$ denote the quotient map. Each of the hexagons used in the construction of S lifts to a regular right-angled hexagon in $\mathbb{H}^2$, and we let $\mathcal{R}(S)$ denote the union of all these hexagons in $\mathbb{H}^2$. We can view the hexagons in $\mathcal{R}(S)$ as 2–cells in a cell structure of $\mathbb{H}^2$. The union of the edges of the hexagons in $\mathcal{R}(S)$ piece together to give a union of biinfinite geodesics which we denote $\mathcal{G}(S)$. Conversely, we can reconstruct the cell structure and $\mathcal{R}(S)$ from the geodesics of $\mathcal{G}(S)$. The vertices of the cell structure are precisely the intersections of the geodesics in $\mathcal{G}(S)$, and each hexagon $P \subset \mathcal{R}(S)$ is determined by the six geodesics that contain the sides of P, all of which are contained in $\mathcal{G}(S)$.

Remark. The geodesics in $\mathcal{G}(S)$ all cover the simple closed curves in S described above, showing that these are *simple closed geodesics*.

Now given any geometric surface X and an isomorphism $\phi : G \to \Gamma$, Proposition 11.4 gives us a Lipschitz homotopy equivalence $h : S \to X$ for which $h_* = \phi$. We lift this to a Lipschitz map $\widetilde{h} : \mathbb{H}^2 \to \mathbb{H}^2$ which is equivariant with respect to ϕ. According to Proposition 11.5, this has a continuous extension to $\overline{\mathbb{H}}^2$, and we let $\partial\widetilde{h} : S^1_\infty \to S^1_\infty$ denote the homeomorphism obtained by restricting to the circle, which is also equivariant with respect to ϕ.

Now we recall from Section 5.1 that any (biinfinite) geodesic $\sigma \subset \mathbb{H}^2$ determines and is determined by its endpoints at infinity $\{\partial_\pm\sigma\} \subset S^1_\infty$. Furthermore, geodesics σ_1, σ_2 intersect if and only if $\{\partial_\pm\sigma_1\}$ and $\{\partial_\pm\sigma_2\}$ *link* with each other, meaning that each component of $S^1_\infty \backslash \{\partial_\pm\sigma\}$ contains exactly one point of $\{\partial_\pm\sigma_2\}$.

Define

$$\mathcal{G}(X) = \{\alpha \subset \mathbb{H}^2 \mid \alpha \text{ is a geodesics with } \{\partial_\pm\alpha\} = \partial\widetilde{h}(\{\partial_\pm\sigma\}) \text{ for some } \sigma \in \mathcal{G}(S)\}$$

By construction, the map $\partial\widetilde{h}$ induces a bijection

$$\partial\widetilde{h}_\# : \mathcal{G}(S) \to \mathcal{G}(X)$$

and since $\partial\widetilde{h}$ is equivariant with respect to ϕ, so is $\partial\widetilde{h}_\#$ (in particular, $\mathcal{G}(X)$ is Γ–invariant). Furthermore, because $\partial\widetilde{h}$ is a homeomorphism, it preserves the property of linking/not linking, and so $\partial\widetilde{h}_\#$ preserves intersections between pairs of geodesics.

As with $\mathcal{G}(S)$, there is a cell structure on $\mathbb{H}^2$ whose vertices are points of intersection of the geodesics in $\mathcal{G}(X)$, and whose edges are the arcs between intersection points. Finally, the

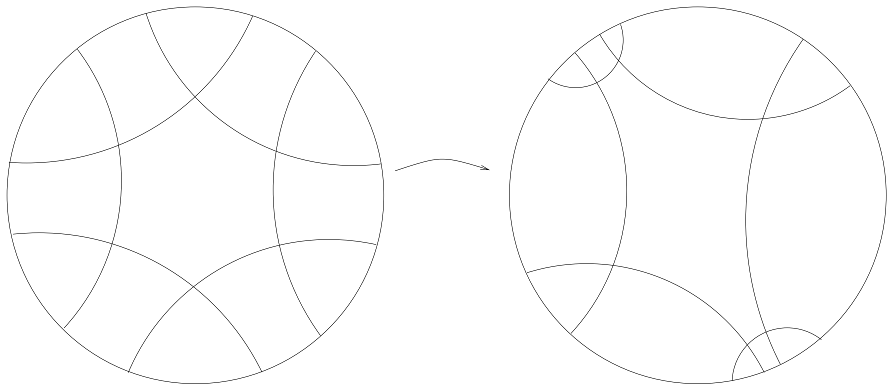

Figure 9: Six geodesics in $\mathcal{G}(S)$ on the left determining a right-angled hexagon and their $\partial\widetilde{h}_{\#}$–images in $\mathcal{G}(X)$ determining the "image" hexagon.

2–cells are (not necessarily right-angled) hexagons bounded by sets of six geodesics which are the image under $\partial\widetilde{h}_{\#}$ of six geodesics in $\mathcal{G}(S)$ that bound a hexagon in $\mathcal{B}(S)$; See Figure 9.

We are ready to construct our biLipschitz map $f : S \to X$. This is done by constructing a ϕ–equivariant biLipschitz map $\widetilde{f} : \mathbb{H}^2 \to \mathbb{H}^2$. Because $\partial\widetilde{h}_{\#}$ preserves intersections of geodesics, we can define $\widetilde{f}$ on the vertices by

$$\widetilde{f}(\sigma \cap \sigma') = \partial\widetilde{h}_{\#}(\sigma) \cap \partial\widetilde{h}_{\#}(\sigma')$$

for each intersecting pair $\sigma, \sigma' \in \mathcal{G}(S)$.

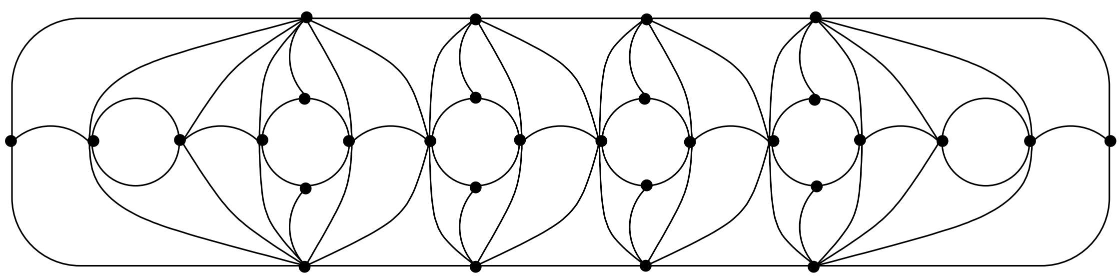

Figure 10: Subdividing the decomposition into hexagons to get a triangulation

Because the cells are completely determined by intersections of geodesics, which are preserved by $\partial\widetilde{h}_{\#}$, there is a ϕ–equivariant bijection from the set of cells of the structure determined by $\mathcal{G}(S)$ to those determined by $\mathcal{G}(X)$ which is compatible with $\widetilde{f}$ on the vertices. To define the map $\widetilde{f}$ on each of the cells, we first subdivide each hexagon into 4 triangles in a G–invariant way. This can be done by subdividing the associated cell structure on S (see Figure 10), and lifting it to $\mathbb{H}^2$. Now we define $\widetilde{f}$ to be the canonical triangle map on each of the triangles as described in Section 6.2. Since $\widetilde{f}$ maps distinct hexagons into distinct hexagons, and is a homeomorphism on each hexagon (since the canonical maps are homeomorphisms—there are no degenerate triangles here), it follows that f is a homeomorphism.

Because the composition of a canonical triangle map with an isometry is another canonical triangle map, $\widetilde{f}$ is ϕ–equivariant and so descends to a homeomorphism $f : S \to X$. Because each triangle map is biLipschitz, the restriction of f to each triangle Δ_i has some biLipschitz constant K_i, and the maximum $K = \max\{K_i\}$ is a biLipschitz constant for f. $\square$

From this, we are able to complete the proof of Theorem 11.1.

Proof of Theorem 11.1. Given two arbitrary closed orientable geometric surfaces $X = \mathbb{H}^2/\Gamma$ and $Y = \mathbb{H}^2/\Lambda$ and an isomorphism $\phi : \Gamma \to \Lambda$, we can realize this by a Lipschitz homotopy equivalence $f_0 : X \to Y$ by Proposition 11.4. Let $S \cong X \cong Y$ and suppose $f_1 : S \to X$ is any biLipschitz homeomorphism and $f_2 : S \to Y$ is a biLipschitz homeomorphism homotopic to $f_0 f_1$, both of which exist by Proposition 12.1. Then $f = f_2 \circ f_1^{-1}$ is homotopic to f_0 and provides the required biLipschitz homeomorphism. $\square$

Corollary 12.2. *The distance in $\mathcal{T}(S)$ and $\mathcal{M}_g$ are both finite.*

The proof of Proposition 12.1 constructs a "canonical" biLipschitz map in the homotopy class, depending on the hexagon decomposition and the subdivision to a triangulation. Given an (orientation preserving) homeomorphism $f : S \to X$ (or even a homotopy equivalence), let $f_{HT} : S \to X$ denote the biLipschitz map homotopic to f from the proposition.

The hexagons determine a set of simple closed geodesics—the pants decomposition and the transverse curves. Observe that given $[f : S \to X], [h : S \to X] \in \mathcal{T}(S)$, the map $h_{HT} f_{HT}^{-1} : X \to Y$ maps this set of simple closed geodesics in X to corresponding set of simple closed geodesics in Y and the triangles of the subdivision to triangles via the canonical triangle map—see Section 6.2. If we only care about the topology, we can treat these maps like the canonical maps f_A for tori.

Define a distance function d_{HT} on $\mathcal{T}(S)$ by

$$d_{HT}([f : S \to X], [h : S \to Y]) = \log(K(h_{HT} \circ f_{HT}^{-1})).$$

So rather than taking the infimum over all maps homotopic to $h \circ f^{-1}$, we just use the biLipschitz constant for $h_{HT} \circ f_{HT}^{-1}$.

Exercise 12.2. *Prove that d_{HT} defines a metric, and this metric gives the same topology as $d_{\mathcal{T}}$. Hint: Proving that this defines the same topology is a slightly subtle. You should look at the proof of Lemma 11.7, and observe that for a K–biLipschitz map, the R from that Lemma tends to 0 as K tends to 1. We let $\mathcal{R}(X)$ and $\mathcal{R}(Y)$ denote the hexagons in $\mathbb{H}^2$ which are image of hexagons in $\mathcal{R}(S)$ under the lifts $\widetilde{f}_{HT}$ and $\widetilde{h}_{HT}$, then you want to show that if $F : X \to Y$ has very small biLipschitz constant and $F \simeq h_{HT} \circ f_{HT}^{-1}$, then the lift of F maps the vertices of the hexagons in $\mathcal{R}(X)$ close to those in $\mathcal{R}(Y)$. Now consider applying Exercise 6.5.*

13 Representation space and Fenchel-Nielsen coordinates

In this lecture we study two concrete descriptions of $\mathcal{T}(S)$. We fix our reference hyperbolic surface $S = S_g = \mathbb{H}^2/G$ of genus $g \geq 2$. The first description is in terms of representations of G into $\mathrm{PSL}_2(\mathbb{R})$ and provides an algebraic framework for studying $\mathcal{T}(S)$. The second gives a global description of the topology of $\mathcal{T}(S)$, proving it is homeomorphic to an appropriate Euclidean space. This latter is via the *Fenchel-Nielsen coordinates* on $\mathcal{T}(S)$ and will take up most of the remainder of the lectures.

13.1 Representation spaces

Given $[f : S \to X = \mathbb{H}^2/\Gamma]$, we have the induced map on $\pi_1(S) = G$,

$$f_* : G \to \Gamma < \mathrm{PSL}_2(\mathbb{R})$$

which is well defined by the element of $\mathcal{T}(S)$ up to conjugacy in $\mathrm{PSL}_2(\mathbb{R})$. To pin down a particular homomorphism within the conjugacy class we do the following. Let $\alpha, \beta \in S$ be a pair of simple closed geodesics that intersect in a single point, and we assume these are one of the curves built from the hexagons described in Lecture 12. After conjugating G if necessary, we can assume that α and β are images of geodesics $\widetilde{a}, \widetilde{b} \subset \mathbb{U}^2 = \mathbb{H}^2$, so that $\widetilde{a} = i\mathbb{R}_+$ and $\widetilde{b}$ ends at $1 \in \widehat{\mathbb{R}}$, and $\widetilde{\alpha} \cap \widetilde{\beta} \neq \emptyset$. There are elements $a, b \in G$ so that $\widetilde{\alpha}$ is the axis of a and $\widetilde{\beta}$ is the axis for b, $\infty \in \widehat{\mathbb{R}}$ is the attracting fixed point for a and 1 is the attracting fixed point for b.

Now, given $[f : S \to X = \mathbb{H}^2/\Gamma] \in \mathcal{T}(S)$, after conjugating if necessary, we can assume $f_*(a)$ has $\infty, 0$ as attracting/repelling fixed points and $f_*(b)$ has 1 as an attracting fixed point. This uses the orientation preserving hypothesis (why?). We call this the *normalized homomorphism.*

Exercise 13.1. *Prove that the normalized homomorphism is unique within the conjugacy class.*

Given $[f : S \to X = \mathbb{H}^2/\Gamma]$, we assume that we have realized $X = \mathbb{H}^2/\Gamma$ so that the induced map $f_* : G \to \Gamma$ is already normalized.

Write $\mathrm{Hom}(G, \mathrm{PSL}_2(\mathbb{R}))$ to denote the set of all homomorphisms from G to $\mathrm{PSL}_2(\mathbb{R})$, then this normalization gives us a map

$$H : \mathcal{T}(S) \to \mathrm{Hom}(G, \mathrm{PSL}_2(\mathbb{R}))$$

given by $H([f : S \to X]) = f_*$. The set $\mathrm{Hom}(G, \mathrm{PSL}_2(\mathbb{R}))$ is in fact a metric space as follows. We recall that $\mathrm{PSL}_2(\mathbb{R})$ is a good quotient of the metric space $\mathrm{SL}_2(\mathbb{R})$, and so is itself a metric space. Let $x_1, \ldots, x_k$ be a generating set for G—for example, you can take the $2g$ standard generators $a_1, b_1, \ldots, a_g, b_g$. Any homomorphism $\phi : G \to \mathrm{PSL}_2(\mathbb{R})$ is determined by its values on $x_1, \ldots, x_k$, and so we have an injective map of sets

$$\mathrm{Hom}(G, \mathrm{PSL}_2(\mathbb{R})) \to (\mathrm{PSL}_2(\mathbb{R}))^k$$

given by $\phi \mapsto (\phi(x_1), \ldots, \phi(x_k))$. Using this injection, we view $\mathrm{Hom}(G, \mathrm{PSL}_2(\mathbb{R}))$ as a subset of $(\mathrm{PSL}_2(\mathbb{R}))^k$ and give it the subspace metric. A more intrinsic description of the topology in terms of convergence is given by the following exercise.

Exercise 13.2. *Prove that a sequence* $\phi_n \in \mathrm{Hom}(G, \mathrm{PSL}_2(\mathbb{R}))$ *converges to* $\phi \in \mathrm{Hom}(G, \mathrm{PSL}_2(\mathbb{R}))$ *if and only if for every element* $x \in G$*,* $\phi_n(x)$ *converges to* $\phi(x)$ *in* $\mathrm{PSL}_2(\mathbb{R})$*.*

Theorem 13.1. *The map* $H : \mathcal{T}(S) \to \mathrm{Hom}(G, \mathrm{PSL}_2(\mathbb{R}))$ *is a homeomorphism onto its image.*

Sketch of proof. We first prove that H is injective. Suppose

$$H([f : S \to X = \mathbb{H}^2/\Gamma]) = H([h : S \to Y = \mathbb{H}^2/\Lambda]).$$

Then $f_* = h_* : G \to \mathrm{PSL}_2(\mathbb{R})$, and so $\Gamma = \Lambda$, and the identity $\mathbb{H}^2 \to \mathbb{H}^2$ descends to an isometry $\sigma : X \to Y$ with $\sigma_* = id_\Gamma$. Then $\sigma_* f_* = h_*$, so $\sigma \circ f \simeq h$ by Proposition 11.3, and so $[f : S \to X] = [h : S \to Y]$ and H is 1-1. Of course, H is onto its image, so H is a bijection onto its image.

To prove that H is a homeomorphism onto its image, it is more convenient to use the metric d_{HT} defined at the end of Lecture 12. We leave the details to the reader, but the idea is as follows. Let $\gamma_i = (f_{HT})_*(x_i)$ and $\lambda_i = (h_{HT})_*(x_i)$. We need to see that $[f_{HT} : S \to X]$ is close to $[h_{HT} : S \to Y]$ in the d_{HT}–metric if and only if γ_i is close to λ_i for each i.

Let $\widetilde{h}_{HT} \circ \widetilde{f}_{HT}^{-1}$ denote a lift of $h_{HT} \circ f_{HT}^{-1}$. Supposing first that $[f_{HT} : S \to X]$ is close to $[h_{HT} : S \to Y]$ in the d_{HT}–metric, then the vertices of all the hexagons of $\mathcal{R}(X)$ in a large ball are close to the $\widetilde{h}_{HT} \circ \widetilde{f}_{HT}^{-1}$–image of the vertices of the hexagons of $\mathcal{R}(Y)$. If we take a large enough ball, then the image of a fixed hexagon $P \in \mathcal{R}(X)$ under the generators of $\gamma_1, \dots, \gamma_k$ also lie inside the ball. The generator $\gamma_i \in \Gamma$ is determined by $\gamma_i(P)$, and similarly the generator $\lambda_i \in \Lambda$ is determined by $\lambda_i(\widetilde{h}_{HT} \circ \widetilde{f}_{HT}^{-1}(P))$. Since P is close to $\widetilde{h}_{HT} \circ \widetilde{f}_{HT}^{-1}(P)$ and $\gamma_i(P)$ is close to $\lambda_i(\widetilde{h}_{HT} \circ \widetilde{f}_{HT}^{-1}(P))$, it follows that λ_i is close to γ_i as required.

Conversely, suppose that $\widetilde{\alpha}_1, \dots, \widetilde{\alpha}_m$ are a collection of geodesics of $\mathcal{G}(X)$ (see Lecture 12) that bound hexagons $P_1, \dots, P_r \in \mathcal{R}(X)$ with the property that the images of the union of the P_i in X cover all of X. For any $[h_{HT} : S \to Y]$, the map $h_{HT} \circ f_{HT}^{-1}$ is determined by the lift $\widetilde{h}_{HT} \circ f_{HT}^{-1}$ on $P_1, \dots, P_r$. Now suppose $(h_{HT})_*$ is close to $(f_{HT})_*$, so that $(h_{HT} \circ f_{HT}^{-1})_*$ is close to the identity on Γ. The geodesics $\widetilde{\alpha}_1, \dots, \widetilde{\alpha}_m$ are axes for elements $\delta_1, \dots, \delta_m \in \Gamma$, and hence the axes $\widetilde{\beta}_1, \dots, \widetilde{\beta}_m$ of $(h_{HT} \circ f_{HT}^{-1})_*(\delta_1), \dots, (h_{HT} \circ f_{HT}^{-1})_*(\delta_m)$ are close to $\widetilde{\alpha}_1, \dots, \widetilde{\alpha}_m$. But then the vertices of the hexagons of $P_1, \dots, P_r$ are close to the vertices of $\widetilde{h}_{HT} \circ \widetilde{f}_{HT}^{-1}(P_1), \dots, \widetilde{h}_{HT} \circ \widetilde{f}_{HT}^{-1}(P_r)$, so $[f_{HT} : S \to X]$ is close to $[h_{HT} : S \to Y]$.

□

Remark. We will not prove it, but in fact the image of $\mathcal{T}(S)$ in $\mathrm{Hom}(G, \mathrm{PSL}_2(\mathbb{R}))$ is an entire component.

13.2 Fenchel-Nielsen coordinates: Hexagons and pants.

The previous arguments have all been based on the use of invariant tilings of $\mathbb{H}^2$ by hexagons. We will use a similar construction to provide a global description of the topology. The main theorem we want to prove is the following.

Theorem 13.2. *For any $g \geq 2$, $\mathcal{T}(S_g) \cong \mathbb{R}^{6g-6}$.*

We will prove this by finding an explicit homeomorphism, the so-called Fenchel-Nielsen coordinates. These are defined via pairs of pants, and to describe this, we need to make a more detailed study of hexagons.

Let $\mathrm{HEX}^\perp$ denote the set of right-angled hexagons with sides labeled $\{1, \dots, 6\}$ in consecutive order, considered up to isometry. Given an element Δ in $\mathrm{HEX}^\perp$, we can measure the length of the i^{th} side, $\ell_i(\Delta)$. There is a natural metric on $\mathrm{HEX}^\perp$ given by

$$d_{\mathrm{HEX}}(\Delta, \Delta') = \inf_\sigma \log(K(\sigma))$$

where $\sigma : \Delta \to \Delta'$ is a $K(\sigma)$–biLipschitz map preserving the labeling of the sides.

Exercise 13.3. *Prove that the map* $\mathrm{HEX}^\perp \to \mathbb{R}^6$ *given by* $\Delta \mapsto (\ell_1(\Delta), \dots, \ell_6(\Delta))$ *is a homeomorphism onto its image.*

In fact, $\Delta \in \mathrm{HEX}^\perp$ is completely determined by only three of its lengths.

Lemma 13.3. *Both of the maps* $(\ell_1, \ell_3, \ell_5) : \mathrm{HEX}^\perp \to \mathbb{R}^3_+$ *and* $(\ell_2, \ell_4, \ell_6) : \mathrm{HEX}^\perp \to \mathbb{R}^3_+$ *are homeomorphisms.*

Proof. We denote the lengths of the sides $1, \dots, 6$ by a, C, b, A, c, B, respectively, so that opposite sides are labeled by the same letter but with different cases, and the alternating sides are ordered alphabetically. By applying an isometry, we can assume our hexagon has its C–side (side 2) contained in the geodesic $i\mathbb{R}_+$, and so that the a–side (side 1) lies to the right of $i\mathbb{R}_+$ on the circle of radius 10, say; see Figure 11.

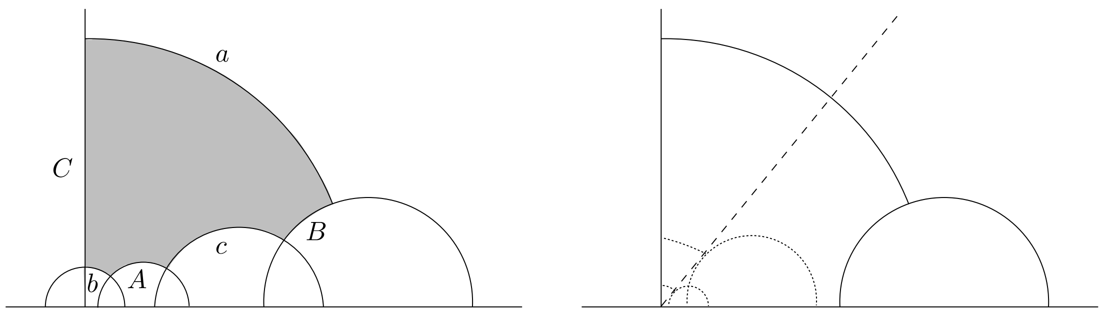

Figure 11: Preferred position of a hexagon on the left. On the right, the a–side is fixed, and the length of b is fixed and as we vary the length of C, the distance c between the B–side and A–side is a strictly increasing function whose image of $\mathbb{R}_+$.

We need to prove that the for any triple of positive real numbers (a, b, c), there exists a unique $\Delta \in \mathrm{HEX}^\perp$ for which these are the lengths of the sides $1, 3, 5$. If we just specify a and b we can construct some hexagon Δ with these as the lengths of sides 1 and 3. In the preferred position all hexagons with these as side lengths 1 and 3 must have the A–side (side 4) contained in a Euclidean circle tangent to a fixed Euclidean line through 0 (this line is the set of points in $\mathbb{U}^2$ to the right of $i\mathbb{R}_+$ whose distance to $i\mathbb{R}_+$ is b).

As we vary the length C of side 2 and let $C \to \infty$, we see that $c \to \infty$, and on the other hand, as C decreases, there is a particular value C_0 for which the hexagons degenerate to a pentagon with four right angles and one ideal vertex (this is when the Euclidean circle containing side 4 becomes tangent to the Euclidean circle containing side 6). We view c as a function $c(C)$, and observe that c is a strictly increasing continuous function from (C_0, ∞) onto $(0, \infty)$. Therefore, C is also a function of c, but since (a, b, C) determines a hexagon uniquely with these as lengths of sides $1, 3, 2$, so does (a, b, c) determine a hexagon uniquely with these as lengths of sides $1, 3, 5$. Moreover these can be taken to be any positive real numbers. □

Now we consider a hyperbolic pair of pants P_0 with geodesic boundary. Formally, this is a geodesic metric surface with boundary, homeomorphic to a sphere minus 3 open disks, so that every point has an ϵ–ball isometric to an ϵ–ball in a half-plane in $\mathbb{H}^2$ (see Lecture 6 for more on half-planes). Given a pair of pants, we can glue two copies together to produce a genus 2 surface, and so all such P can be thought of as sitting inside some closed hyperbolic surface.

Let $\alpha_1, \alpha_2, \alpha_3$ denote the geodesic boundary components of P_0.

Exercise 13.4. *Prove that for each $i, j \in \{1, 2, 3\}$ with $i \neq j$, prove that there exists a unique embedded geodesic segment δ_{ij} from α_i to α_j which is orthogonal to both α_i and α_j. Furthermore, prove that $\delta_{ij} \cap \delta_{jk} = \emptyset$ unless $i = k$, for each $i, j, k \in \{1, 2, 3\}$, $i \neq j$ and $j \neq k$. These arcs are often called the* seams *of the pair of pants.*

The seams cut P_0 up into two right-angled hexagons $H_\pm(P_0)$. We can order the sides $1, \ldots, 6$ as in Figure 12 so that in each hexagon, the arcs of $\alpha_1, \alpha_2, \alpha_3$ in each $H_\pm(P_0)$ are labeled $1, 3, 5$, respectively, and $\delta_{12}, \delta_{23}, \delta_{13}$ in each $H_\pm(P_0)$ are labeled $2, 4, 6$.

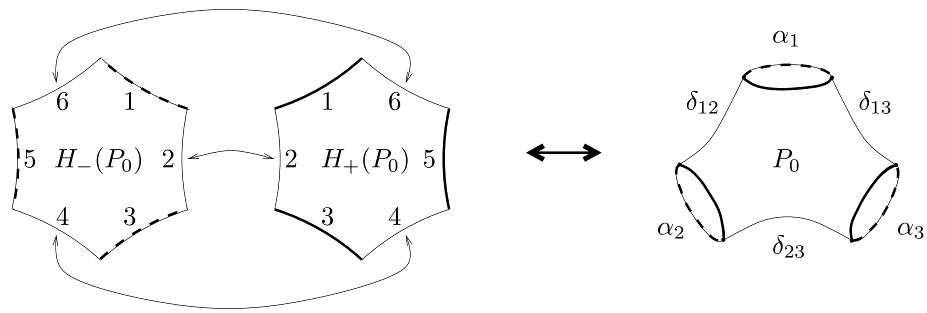

Figure 12: Canonical decomposition of a pair of pants into two right-angled hexagons.

Since $H_-(P_0)$ shares the seams with $H_+(P_0)$ as sides $2, 4, 6$, respectively, it follows that $H_-(P_0)$ and $H_+(P_0)$ are isometric by an isometry preserving the ordering of edges by Lemma 13.3. In particular, the lengths a_1, a_3, a_5, respectively, of the sides $1, 3, 5$ of *both* of $H_-(P_0)$ and $H_+(P_0)$ are exactly half the lengths of the boundary curves: $(a_1, a_3, a_5) = \left(\frac{1}{2}\ell_1(P_0), \frac{1}{2}\ell_2(P_0), \frac{1}{2}\ell_3(P_0)\right)$, where $\ell_i(P_0) = \ell(\alpha_i)$. This leads to the following.

We let $\mathcal{T}(P_0)$ denote the space of hyperbolic pants with ordered geodesic boundary, up to boundary-order-preserving isometry (with the biLipschitz metric). The lengths of the three boundary components defines a map $(\ell_1, \ell_2, \ell_3) : \mathcal{T}(P_0) \to \mathbb{R}^3_+$.

Lemma 13.4. *The map $(\ell_1, \ell_2, \ell_3) : \mathcal{T}(P_0) \to \mathbb{R}^3_+$ is a homeomorphism.*

Proof. First we prove that (ℓ_1, ℓ_2, ℓ_3) is injective. Given P_1, P_2 for which $\ell_i(P_1) = \ell_i(P_2)$ for $i = 1, 2, 3$, by the discussion above, $H_\pm(P_1) = H_\pm(P_2) \in \mathrm{HEX}^\perp$, and the isometries between these hexagons induces an isometry between P_1 and P_2 that preserves the ordering of the boundary components. Therefore $P_1 = P_2$ in $\mathcal{T}(P_0)$ and (ℓ_1, ℓ_2, ℓ_3) is injective.

On the other hand, given any $(a_1, a_2, a_3) \in \mathbb{R}^3_+$, let $H \in \mathrm{HEX}^\perp$ be such that

$$(\ell_1(H), \ell_3(H), \ell_5(H)) = \left(\frac{1}{2}a_1, \frac{1}{2}a_2, \frac{1}{2}a_3\right).$$

Gluing two isometric copies of H along sides $2, 4, 6$ appropriately defines a pair of pants P_1 for which $\ell_i(P_1) = a_i$. Therefore, (ℓ_1, ℓ_2, ℓ_3) is a bijection. Because $\mathrm{HEX}^\perp$ is also given the biLipschitz metric, one can check that this map is a homeomorphism. □

14 Fenchel-Nielsen coordinates II and the action of $\mathrm{Mod}(S)$

14.1 Fenchel-Nielsen coordinates: length and twist

After the setup from the previous lecture, we are now ready to describe the homeomorphism $\mathcal{T}(S_g) \cong \mathbb{R}^{6g-6}$. Given an essential (that is, non-nullhomotopic) closed curve α in S_g and $[f : S \to X] \in \mathcal{T}(S)$, let $\ell_\alpha([f : S \to X])$ denote the length of the geodesic representative of $f(\alpha) \subset X$. When the marking f is understood, we sometimes just write $\ell_\alpha(X)$, although this is in fact vague.

Exercise 14.1. *If $d_{\mathcal{T}}([f : S \to X], [h : S \to Y]) = \delta$, then prove that for any essential closed curve α in S we have*

$$e^{-\delta}\ell_\alpha([f : S \to X]) \leq \ell_\alpha([h : S \to Y]) \leq e^{\delta}\ell_\alpha([f : S \to X]).$$

In particular, ℓ_α is a continuous function for all essential closed curves α.

Now let $\alpha_1, \ldots, \alpha_{3g-3}$ be a pants decomposition of S: that is, a collection of pairwise disjoint essential simple closed curves that cut the surface into $2g-2$ pairs of pants. For example, we can take the pants decomposition from Lecture 12.

We then define

$$L : \mathcal{T}(S) \to \mathbb{R}^{3g-3}$$

by

$$L([f : S \to X]) = (L_1(X), \ldots, L_{3g-3}(X))$$

where $L_i(X) = \log(\ell_{\alpha_i}(X))$.

Given $[f : S \to X]$, we straighten $f(\alpha_1), \ldots, f(\alpha_{3g-3})$ to geodesics which we denote $\alpha_1^X, \ldots, \alpha_{3g-3}^X$. These cut X up into pairs of pants with geodesic boundary which we denote $P_1^X, \ldots, P_{2g-2}^X$. According to Lemma 13.4, the isometry type of each P_i^X is determined by the lengths of its boundary components $(\ell_1(P_i^X), \ell_2(P_i^X), \ell_3(P_i^X))$ which are in fact just numbers from the list $\ell_{\alpha_1}(X), \ldots, \ell_{\alpha_{3g-3}}(X)$. However, the metrics on these pants alone do not determine the metric on X since we also need to know exactly how they are glued together. Moreover, we must also keep track of the marking information.

For this, we introduce a *twisting parameter* T_i for each curve α_i. This is a lengthy definition and we give it over the next few paragraphs. We choose a family of transverse curves on S, $\beta_1, \ldots, \beta_k$. We will take this to be any set of curves so that the geodesic representatives in S of $\alpha_1, \ldots, \alpha_{3g-3}, \beta_1, \ldots, \beta_k$ cut S up into hexagons. For concreteness, we just use the curves from Lecture 12.

In X, we straighten $f(\beta_1), \ldots, f(\beta_k)$ to geodesics which we denote $\beta_1^X, \ldots, \beta_k^X$. These cut the pants obtained from $\alpha_1^X, \ldots, \alpha_{3g-3}^X$ into hexagons, and in particular, the three arcs of $P_i^X \cap (\beta_1^X \cup \ldots \cup \beta_k^X)$ are homotopic (keeping the endpoints on the boundary of P_i^X) to the seams $\delta_{12}, \delta_{13}, \delta_{23}$ of P_i^X. Using this, we choose another representative for each $f(\beta_j)$ which we denote by $\beta_j^{X,tw}$ that runs without backtracking around the geodesics $\alpha_1^X \cup \ldots \cup \alpha_{3g-3}^X$, and when it enters and exits any pair of pants P_i^X, it does so along one of the seams for P_i^X. Furthermore, we assume that these representatives only enter and exit a pair of pants P_i^X as many times as the geodesic representative.

A more concrete way to find this representative is in the universal covering. A lift of some β_j^X is a geodesic $\widetilde{\beta}$. This crosses infinitely many lifts of geodesics in the pants decomposition which we denote $\{\widetilde{\alpha}_j\}_{j\in\mathbb{Z}}$. Moreover, because $\{\widetilde{\alpha}_j\}_{j\in\mathbb{Z}}$ project to a collection of pairwise disjoint simple closed geodesics on X, they have a positive distance apart, and so are connected by a unique orthogonal segment (this is a lift of a seam). We replace $\widetilde{\beta}$ by a concatenation of geodesics arcs in the union of the geodesics $\cup_j \widetilde{\alpha}_j$ together with these orthogonal segments; see Figure 13. A covering transformation that leaves $\widetilde{\beta}$ invariant will also leave this new path invariant, and so this descends to a closed curve in the homotopy class of $f(\beta_j)$ which is $\beta_j^{X,tw}$.

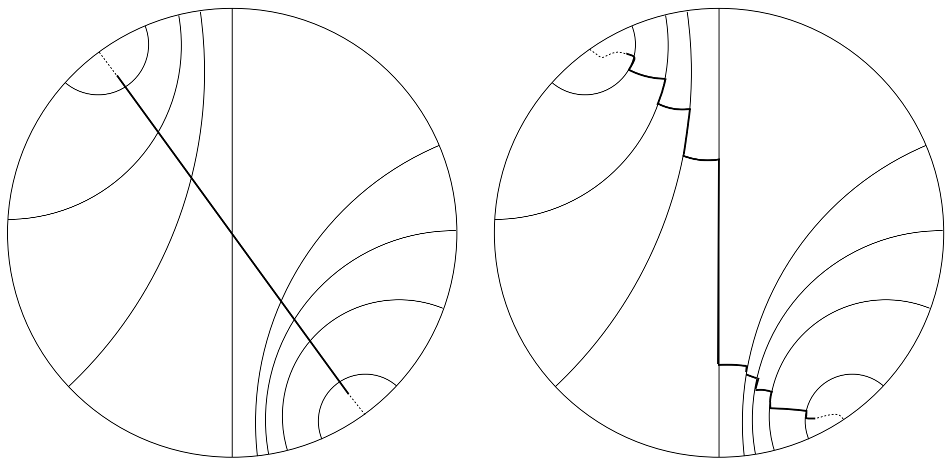

Figure 13: Going from the geodesic to the representative used to define twisting.

Now for each $i = 1, \ldots, 3g-3$ define the twisting $T_i(X)$ by picking a curve β_j that intersects α_i, and setting $T_i(X)$ to be the *signed* distance that $\beta_j^{X,tw}$ traverses along α_i^X. The signed distance is $\pm$ the actual distance, with a positive sign if the $\beta_j^{X,tw}$ makes a "right turn" from a seam onto α_i^X and negative if it makes a "left turn". Notice (for example, from Figure 13) that this is independent of the orientation of β (but does depend on the orientation on X). Note that if every geodesic β_j^X intersects every geodesics α_i^X orthogonally (if it intersects it at all), then $T_i(X) = 0$ for all i.

Exercise 14.2. *Prove that $T_i(X)$ does not depend on which β_j we choose that intersects it, nor does it depend on the choice of intersection point if β_j intersects α_i twice.*

The two maps L and $T = (T_1, \ldots, T_{3g-3})$ determine a map $(L,T) : \mathcal{T}(S) \to \mathbb{R}^{3g-3} \times \mathbb{R}^{3g-3} = \mathbb{R}^{6g-6}$ which are the Fenchel-Nielsen coordinates.

Theorem 14.1. *The Fenchel-Nielsen coordinates map $(L,T) : \mathcal{T}(S) \to \mathbb{R}^{6g-6}$ is a homeomorphism.*

Sketch of proof. To see that (L,T) is surjective, we note that by Lemma 13.4 we can arbitrarily prescribe the lengths of the pants curves by gluing together pants, and appropriately mapping S to the resulting surface. Furthermore, if we glue the seams together appropriately we can arrange for the twisting parameters to all be 0. Now, for any other twisting parameter, we can modify the gluing by a continuous family of gluings, slowly twisting to the right or the left. If we twist around the α_i geodesics a distance t_i to the right, then we see the opposite side twisting to the left. Since this is done continuously, there is a natural way to adjust the markings to obtain a marking of the twisted surface for which $T_i = -t_i$. In this way we can realize any point of $\mathbb{R}^{6g-6}$ in the image of (L,T), and so (L,T) is surjective.

To prove that (L,T) it is injective, observe that given $[f : S \to X = \mathbb{H}^2/\Gamma]$, we can reconstruct $[f : S \to X]$ from the previous paragraph. More precisely, we can "untwist" until we arrive at a new marked surface $f_0 : S \to X_0$ by $L(X) = L(X_0)$ but $T(X_0) = 0$. It is not too hard to see that $[f_0 : S \to X_0]$ is built from the first part of the construction of the previous paragraph, and then $f : S \to X$ is built by adding the twists back in. Another more concrete way to prove this is to look at the right-angled hexagons in the universal covering which can be reconstructed from the data of $L(X)$ and $T(X)$, as well as the homomorphism $f_* : G \to \Gamma$.

To see that (L,T) and $(L,T)^{-1}$ are continuous, it suffices to prove that T is continuous since L is continuous by Exercise 14.1 and that $(L,T)^{-1}$ is continuous. The simplest way to do this is as follows.

Given $[f : S \to X]$, we assume that f_* is normalized (see Lecture 13) and that $\widetilde{f} : \mathbb{H}^2 \to \mathbb{H}^2$ is equivariant with respect to f_*. We could/should have done this exercise earlier.

Exercise 14.3. *Prove that $[f_n : S \to X] \to [f : S \to X]$ if and only if $\partial \widetilde{f_n}$ converges uniformly to $\partial \widetilde{f}$ on S^1_∞.*

Exercise 14.4. *Use Exercise 14.3 to prove that each T_i is continuous and that $(L,T)^{-1}$ is also continuous. Hint: look at Figure 13.*

□

14.2 Proper discontinuity of $\mathrm{Mod}(S)$

We have already seen that $\mathrm{Mod}(S_g)$ acts on $\mathcal{T}(S_g)$ and that the orbits are precisely the fibers of $\Phi : \mathcal{T}(S_g) \to \mathcal{M}_g$ (see Proposition 10.1).

Exercise 14.5. *Prove that $\mathrm{Mod}(S)$ acts by isometries on $\mathcal{T}(S)$.*

Proposition 14.2. *The action of $\mathrm{Mod}(S)$ on $\mathcal{T}(S)$ is properly discontinuous.*

Proof. Let K be a compact set in $\mathcal{T}(S)$. Suppose $\Omega = \{\phi \in \mathrm{Mod}(S) \mid \phi(K) \cap K \neq \emptyset\}$ is infinite. If we let r denote the maximum distance from $[id : S \to S] \in \mathcal{T}(S)$ to K, then $d_{\mathcal{T}}([id : S \to S], \phi \cdot [id : S \to S]) \leq 2r$ for all $\phi \in \Omega$. According to Exercise 14.1 it follows that for any simple closed geodesics $\alpha \subset S$, we have

$$\ell_{\phi^{-1}(\alpha)}(S) = \ell_\alpha(\phi \cdot [id : S \to S]) \leq e^{2r} \ell_\alpha([id : S \to S]) \leq e^{2r} \ell_\alpha(S) \qquad (12)$$

for all $\phi \in \Omega$.

For any $R > 0$, there are only finitely many closed geodesics on S of length at most R: to see this, we can choose axes that cover these geodesics which pass through some fixed Dirichlet domain $D_G(z)$. Then there are elements of G which preserve these axes and translate a distance R, and hence move z a distance at most R plus twice the diameter of $D_G(z)$. By Lemma 8.2 there can be only finitely many such elements of G.

Now take our curves $\alpha_1, \ldots, \alpha_{3g-3}, \beta_1, \ldots, \beta_k$ which cut S up into hexagons. The lengths of these curves are at most R, for some $R > 0$, and by (12) the images of of these have length at most Re^{2r}, and hence there are only finitely many possibilities for their images. Since Ω is infinite, by the pigeonhole principal, we can find an infinite set of distinct elements $\{\phi_n\}_{n=1}^{\infty} \subset \Omega$ so that for all $n, m \geq 1$ and all i, j

$$\phi_n(\alpha_i) = \phi_m(\alpha_i) \text{ and } \phi_n(\beta_j) = \phi_m(\beta_j).$$

Then setting $\phi'_n = \phi_n^{-1} \circ \phi_1$, the set $\{\phi'_n\}$ is an infinite set of distinct elements which preserve all the curves $\alpha_1, \ldots, \alpha_{3g-3}, \beta_1, \ldots, \beta_k$. However, because these curves cut S into regular right-angled hexagons, these are all represented by isometries on S that preserve the hexagon cell structure. There are only finitely many such (why?), which is a contradiction. □

14.3 Beyond Teichmüller space

Theorem 14.1 and Proposition 14.2 is very similar to the case of the torus in Theorem 10.5, except that we have a better description of the metric in that case. The close connection suggests that there might be other analogies between the very well understood case of the torus, and the general case. This is indeed true, and we mention just one more analogy and how it can be used.

Theorem 14.3. *There is a* $\mathrm{Mod}(S)$*–invariant compactification* $\overline{\mathcal{T}(S)}$ *of* $\mathcal{T}(S)$ *and if we denote the boundary by* $\partial\overline{\mathcal{T}(S)}$*, then* $(\overline{\mathcal{T}(S)}, \partial\overline{\mathcal{T}(S)}) \cong (\overline{B}^{6g-6}, \mathbb{S}^{6g-7})$.

This last statement is that the compactification is homeomorphic to a closed ball and in this homeomorphism, the boundary is sent to the boundary sphere.

Given an element $\phi \in \mathrm{Mod}(S)$, the Brower fixed point theorem implies ϕ fixes some point in $\overline{\mathcal{T}(S)}$. If the point lies in $\mathcal{T}(S)$, then this means that ϕ has finite order (by proper discontinuity). On the other hand, if the point lies on the boundary, then a detailed analysis of what the boundary points are, and how they can be fixed, provides the basis for the following which could be compared with the classification of isometries of $\mathbb{H}^2$ from Section 5.1.

Theorem 14.4. *Given any element* $\phi \in \mathrm{Mod}(S)$*, one of the following holds.*

1. *ϕ has finite order,*
2. *ϕ leaves a collection of pairwise disjoint essential simple closed curves invariant,*
3. *ϕ is* pseudo-Anosov*,*

Moreover, case 3 is mutually disjoint from cases 1 and 2.

Case 3 is the analogue in the case of $\mathbb{T}^2$ of a hyperbolic element $A \in \mathrm{SL}_2(\mathbb{Z})$. Observe that there, the map $f_A : \mathbb{T}^2 \to \mathbb{T}^2$ stretches all lines in a particular direction (one of the eigendirections) by a number $\lambda > 1$ (the absolute value of the leading eigenvalue of A), and contracts by a factor $1/\lambda$ in another direction. For a pseudo-Anosov homeomorphism ϕ of S, one has a metric on S which is locally isometric to $\mathbb{R}^2$ at all but a finite set of points (think about gluing together opposite sides of a regular *Euclidean* octagon), and ϕ stretches all lines in some direction by a factor $\lambda > 1$ and contracts in another direction by a factor $1/\lambda$.