

---

## Chapter 8

# Euclidean Geometry

This chapter begins the second part of the book. It is the first in a series of 3 chapters in which we consider the classical 2 dimensional geometries. In this chapter we will prove some results about Euclidean geometry in the plane. Since Euclidean geometry is so familiar, we will not spend too much time on the basics. Following an introductory first section, we will concentrate on interesting theorems. Most of the theorems revolve around the theme of cutting complicated polygons into simpler ones.

### 8.1. Euclidean Space

The standard dot product on  $\mathbf{R}^n$  is given by the formula

$$(8.1) \quad (x_1, \dots, x_n) \cdot (y_1, \dots, y_n) = x_1y_1 + \dots + x_ny_n.$$

The norm of a vector  $X = (x_1, \dots, x_n)$  is given by

$$(8.2) \quad \|X\| = \sqrt{X \cdot X}.$$

The dot product satisfies the fundamental *Cauchy-Schwarz Inequality*. We will give two proofs of this inequality.

**Lemma 8.1.** *For any vectors  $X$  and  $Y$ , we have*

$$|X \cdot Y| \leq \|X\| \|Y\|.$$

Assuming  $Y$  is nonzero, we get equality if and only if  $X$  is a multiple of  $Y$ .

**First Proof.** To avoid trivialities, assume  $Y$  is nonzero. For any choice of  $t$ , we have

$$\|X\|^2 + t^2\|Y\|^2 + 2t(X \cdot Y) = \|X - tY\|^2 \geq 0.$$

Plugging in  $t = (X \cdot Y)/\|Y\|^2$ , multiplying through by  $\|Y\|^2$ , and simplifying, we get the inequality. The only way to get equality is that  $\|X - tY\| = 0$ . But then  $X = tY$ .

The proof above is the standard proof. Now I will give a second proof which, though more involved, makes the result look less mysterious.

**Second Proof.** If  $c$  and  $s$  are real numbers such that  $c^2 + s^2 = 1$ , then the map

$$(8.3) \quad R_{12} \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} = \begin{pmatrix} cx_1 + sx_2 \\ -sx_1 + cx_2 \\ \dots \\ x_n \end{pmatrix}$$

preserves the dot product. The map  $R_{12}$  changes coordinates 1 and 2 and leaves the rest alone. There is an analogous symmetry  $R_{ij}$  (depending on  $c$  and  $s$ ) which changes coordinates  $i$  and  $j$  and leaves the rest alone. Applying suitable choices of these symmetries, we can reduce to the special case when  $Y = (x_1, 0, \dots, 0)$ . In this case, the inequality is obvious.

The Euclidean distance  $\mathbf{R}^n$  is given by the formula

$$(8.4) \quad d(X, Y) = \|X - Y\|.$$

**Lemma 8.2.**  $d$  satisfies the triangle inequality.

**Proof.** For any vectors  $A$  and  $B$ , we have

$$(8.5) \quad \begin{aligned} \|A + B\|^2 &= (A + B) \cdot (A + B) \\ &= \|A\|^2 + 2(A \cdot B) + \|B\|^2 \\ &\leq^* \|A\|^2 + 2\|A\|\|B\| + \|B\|^2 \leq (\|A\| + \|B\|)^2. \end{aligned}$$

The starred inequality follows from the Cauchy–Schwarz inequality. Hence

$$\|A + B\| \leq \|A\| + \|B\|.$$

Setting  $A = X - Y$  and  $B = Y - Z$ , we see that

$$\begin{aligned} d(X, Y) &= \|X - Z\| = \|A + B\| \leq \|A\| + \|B\| \\ &\leq \|X - Y\| + \|Y - Z\| = d(X, Y) + d(Y, Z). \end{aligned}$$

This holds for any triple  $X, Y, Z$  of vectors, and thereby completes the proof.  $\square$

The angle  $\theta$  between two vectors  $X$  and  $Y$  obeys the equation

$$(8.6) \quad \cos(\theta) = \frac{X \cdot Y}{\|X\| \|Y\|}.$$

To understand this equation, we consider the case  $\|X\| = \|Y\| = 1$ . We can use compositions of the isometries mentioned above to rotate so that  $X = (1, 0, \dots, 0)$  and  $Y = (c, s, 0, \dots, 0)$ , where  $c^2 + s^2 = 1$ . Then, we have

$$(8.7) \quad \cos(\theta) = X \cdot Y = c.$$

This last equation matches our expectation that  $\cos(\theta)$  is the first coordinate of a unit vector in the plane that makes an angle of  $\theta$  with the positive  $x$ -axis.

Now that we have defined distances and angles in Euclidean space, we talk a bit about volumes of solids. Given  $n$  linearly independent vectors  $V_1, \dots, V_n$  in  $\mathbf{R}^n$ , the *parallelepiped* spanned by these vectors is defined as the set of all linear combinations

$$\sum a_j v_j, \quad a_1, \dots, a_n \in [0, 1].$$

The volume of this parallelepiped is given by

$$(8.8) \quad \det(V_1, \dots, V_n) = \sum_{\sigma} (-1)^{|\sigma|} \prod_{i=1}^n V_{i, \sigma(i)}.$$

The sum takes place over all permutations  $\sigma$ . The quantity  $|\sigma|$  is 0 if  $\sigma$  is an even permutation and 1 if  $\sigma$  is an odd permutation. Finally,  $V_{ij}$  is the  $j$ th component of  $V_i$ . If you have not seen the definition of the determinant before, this book is not place to learn it. See any book on linear algebra.

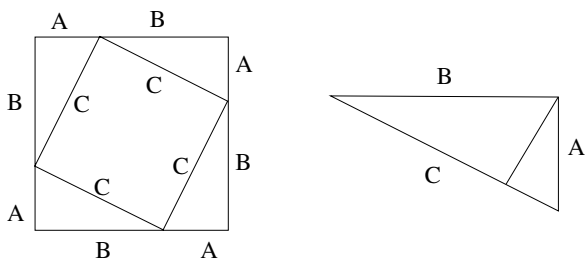
It would be nice if every solid body could be decomposed into finitely many parallelepipeds. Then one could define the volume of an arbitrary solid body by summing up the volumes of the pieces. Unfortunately, this doesn't work, and one must resort to some kind of limiting process. For instance, you fill up a given solid, as best as possible, with increasingly small cubes, and take a limit of the corresponding sums. This is what is typically done in a calculus class. This procedure suffices to give a satisfactory definition of volume for household solids, such as spheres and ellipsoids.

Taking a measure-theoretic approach vastly broadens the number of solid bodies whose volume one can define in a satisfactory way. With the exception of Chapter 22, where we prove the Banach–Tarski Theorem, we will always deal with very simple solids for which all reasonable definitions of volume coincide.

## 8.2. The Pythagorean Theorem

Our definition of distance in  $\mathbf{R}^2$  somewhat has the Pythagorean Theorem built into it. The distance from the point  $(a, b)$  to  $(0, 0)$  is defined to be  $c = \sqrt{a^2 + b^2}$ . So, we automatically have  $a^2 + b^2 = c^2$ . Here  $a, b$  and  $c$  are the side lengths of the right triangle with vertices  $(0, 0)$  and  $(a, 0)$  and  $(a, b)$ . Note that this triangle is rather special: two of its sides are parallel to the coordinate axes.

Here we will prove the Pythagorean Theorem for an arbitrary right triangle in the plane. There are many, many proofs; I'll present my two favorites.



**Figure 8.1.** Two views of the Pythagorean Theorem

Referring to the left half of Figure 8.1, the outer square has area  $(A + B)^2$ . At the same time, the outer square breaks into 4 right triangles, each having area  $AB/2$ , and an inner square having area  $C^2$ . Hence  $(A + B)^2 = 2AB + C^2$ . Simplifying gives  $A^2 + B^2 = C^2$ . That is the first proof.

Here is the second proof. For any right triangle, there is a constant  $k$  such that the distance from the right-angled vertex to the hypotenuse is  $k$  times the length of the hypotenuse. This constant  $k$  only depends on the shape of the triangle, and not on its size. By the base times height formula for area, the area of the triangle is  $kL^2$ , where  $L$  is the length of the hypotenuse. Again, the constant  $k$  only depends on the shape of the triangle and not on its size. The three triangles on the right-hand side of Figure 8.1 have the same shape. The large one has area  $kC^2$ , and the two small ones have area  $kA^2$  and  $kB^2$ . Hence  $kC^2 = kA^2 + kB^2$ . Cancelling the  $k$  (a constant we don't care about) gives  $A^2 + B^2 = C^2$ .

### 8.3. The X Theorem

Here we prove a classic result from high school geometry. Let  $S^1$  be the unit circle in the plane and let  $A$  and  $B$  be two chords of  $S^1$ , as shown on the left-hand side of Figure 8.2. Let  $L(A, B)$  be the length of the region  $R(A, B) \subset S^1$  opposite the two acute angles of  $A \cap B$ . (In case  $A \perp B$  we choose arbitrarily.) Figure 8.2 shows  $R(A, B)$  drawn thickly.

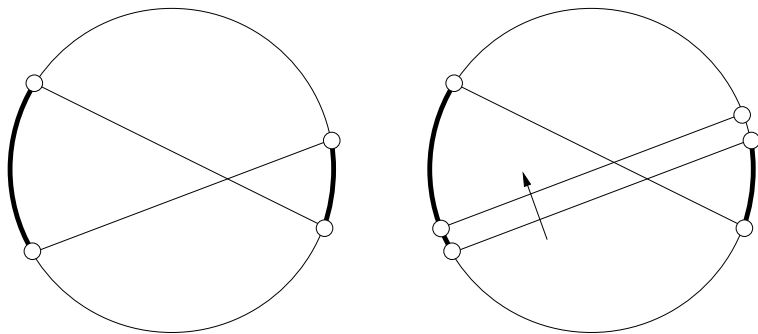


Figure 8.2. The chords  $A$  and  $B$ .

**Theorem 8.3** (The X Theorem).  $L(A, B)$  only depends on the acute angle  $\theta(A, B)$  between  $A$  and  $B$  and not on the positions.

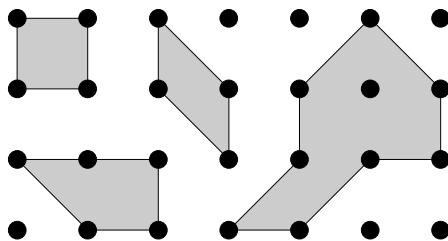
**Proof.** To see this, imagine that  $A$  and  $B$  are toothpicks that we can roll to a new location. The right-hand side of Figure 8.2 shows what happens when roll  $A$  is parallel to itself. By symmetry (about the line perpendicular to the direction of motion) the same length of arc is added to one side of  $R(A, B)$  as is subtracted from the other. Hence, the sum of the lengths does not change. The same goes when we roll  $B$  parallel to itself. At the same time, rotating the disk by any amount changes neither the angle between  $A$  and  $B$  nor  $L(A, B)$ . Rotating and rolling as necessary, we can get to any position without changing  $L(A, B)$ .  $\square$

When  $A$  and  $B$  cross at the center of  $S^1$ , we have  $L(A, B) = 2\theta(A, B)$ . By the X Theorem, this result holds in general.

As a limiting case, the X Theorem applies when  $A \cap B \in S^1$ . In this case, we can reformulate the result. We fix two points  $x_1, x_2 \in S^1$  and consider the angle  $\theta(y)$  between  $\overline{yx_1}$  and  $\overline{yx_2}$  as a function of  $y \in S^1$ . The X Theorem implies that  $\theta(y)$  is independent of  $y$ .

## 8.4. Pick's Theorem

During college I learned Pick's Theorem from a friend and classmate of mine, Sinai Robins. If you want to learn a whole lot about Pick's Theorem and its higher-dimensional generalizations, see the the book [BRO] by Matthias Beck and Sinai Robins.



**Figure 8.3.** Some lattice polygons

Let  $\mathbf{Z}^2 \subset \mathbf{R}^2$  denote the ordinary lattice of integer points. Say that a *lattice polygon* is a polygon in  $\mathbf{R}^2$  whose vertices lie in  $\mathbf{Z}^2$ . That is, the vertices have integer coordinates. Figure 8.3 shows some examples. Let  $P$  be a lattice polygon. We let  $i(P)$  denote the number of vertices contained in the interior of the region bounded by  $P$ . We let  $e(P)$  denote the number of vertices contained on the edges of  $P$ . (The vertices of  $P$  are included in the count for  $e(P)$ .)

**Theorem 8.4** (Pick). *The area of the region bounded by  $P$  is*

$$i(P) + \frac{e(P)}{2} - 1.$$

For the examples in Figure 8.3, you can of course verify the formula directly. During our proof, we will often use the phrase “the area of  $P$ ”, when we really mean to say “the area of the region bounded by  $P$ ”. We hope that this slight abuse of terminology does not cause confusion.

**Exercise 1.** Let  $P$  be a parallelogram whose vertices have integer coordinates. Prove that the area of  $P$  is an integer. (*Hint:* Work in  $\mathbf{C}$  and translate so that the vertices are 0 and  $V$  and  $W$  and  $V + W$ . Then establish the formula  $\text{area}(P) = \text{Im}(V\overline{W})$ .)

We say that a lattice parallelogram  $P$  is *primitive* if  $i(P) = 0$  and  $e(P) = 4$ .

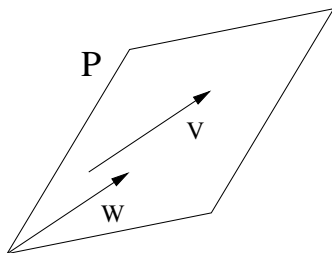
**Lemma 8.5.** *Pick's Theorem holds for primitive parallelograms.*

**Proof.** By Exercise 1, the parallelogram  $P$  has integer area. To finish the proof, we just have to show that  $P$  has area at most 1.

Let  $X$  be the square torus obtained by identifying the opposite sides of the unit square. Note that  $X$  has area 1. Let  $E : \mathbf{R}^2 \rightarrow X$  be the universal covering map. See §6.3. Let  $P^\circ$  denote the interior of the region bounded by the primitive parallelogram  $P$ .

We claim that  $E(P^\circ)$  is embedded in  $X$ . Otherwise, we can find two points  $x_1, x_2 \in P^\circ$  such that  $e(x_1) = e(x_2)$ . But then  $x_1 - x_2 \in \mathbf{Z}^2$ . Let  $V$  be the vector whose tail is  $x_1$  and whose head is  $x_2$ . This is a vector with integer coordinates. Using the convexity of  $P$ , we can

find a vector  $W$  parallel to  $V$  whose tail is a vertex of  $P$  and whose head lies either on the interior of an edge of  $P$  or in  $P_0$ . Figure 8.4 shows the situation.



**Figure 8.4.** Translating a vector

Since  $W \in \mathbf{Z}^2$ , and the vertices of  $P$  are in  $\mathbf{Z}^2$ , the head of  $W$  lies in  $\mathbf{Z}^2$ . But then we either have  $i(P) > 0$  or  $e(P) > 4$ , which is a contradiction. Now we know that  $E(P)$  is embedded. Since  $E(P)$  is embedded, we see that

$$\text{area}(P) = \text{area}(E(P)) \leq \text{area}(X) = 1.$$

This completes the proof.  $\square$

We say that a *primitive triangle* is a lattice triangle  $T$  such that  $i(T) = 0$  and  $e(T) = 3$ .

**Exercise 2.** Prove Pick's Theorem for primitive triangles.

We say that  $P$  *dissects* into two lattice polygons  $P_1$  and  $P_2$  if

- $P_1$  and  $P_2$  bound disjoint open regions, and  $P_1 \cap P_2$  is a connected arc.
- The closed region bounded by  $P$  is the union of the closed region bounded by  $P_1$  and the closed region bounded by  $P_2$ .

**Lemma 8.6.** *Suppose that  $P$  dissects into  $P_1$  and  $P_2$ . If Pick's Theorem holds for  $P_1$  and  $P_2$ , then it also holds for  $P$ .*

**Proof.** Let  $A = \text{area}(P)$  and  $A_1 = \text{area}(P_1)$ , etc. Obviously  $A = A_1 + A_2$ . Let  $n$  denote the number of vertices on  $P_1 \cap P_2$ . Let



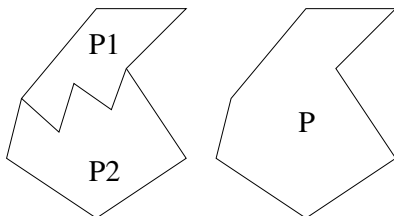
$i = i(P)$  and  $i_1 = i(P_1)$ , etc. We have

$$i = i_1 + i_2 + n - 2, \quad e = e_1 + e_2 - 2n + 2.$$

Therefore,

$$\begin{aligned} & i + e/2 - 1 \\ &= i_1 + i_2 + n - 2 + e_1/2 + e_2/2 - n + 1 - 1 \\ &= (i_1 + e_1/2 - 1) + (i_2 + e_2/2 - 1) =^* A_1 + A_2 = A. \end{aligned}$$

The starred equality comes from Pick's Theorem applied to  $P_1$  and  $P_2$ .  $\square$



**Figure 8.5.** Dissecting a polygon

**Exercise 3.** Suppose that  $P$  is a lattice polygon that is not a primitive triangle. Prove that  $P$  can be dissected into two lattice polygons.

By Exercise 3, any lattice polygon can be written as the finite union of primitive triangles, each of which have area  $1/2$ . Hence, any lattice polygon has area which is a half-integer. The rest of our proof goes by induction on the area.

**Lemma 8.7.** *If  $P$  is a lattice polygon with area at most  $1/2$  then  $P$  is a primitive triangle. In particular, Pick's Theorem holds for  $P$ .*

**Proof.** Applying Exercise 3 iteratively, we see that any lattice polygon can be divided into primitive triangles. If  $P$  is not a primitive triangle, then  $P$  can be divided into at least 2 primitive triangles. But each such triangle has area  $1/2$ . This would force  $P$  to have area at least 1.  $\square$

Now let  $P$  be a general lattice polygon. If  $P$  is not a primitive triangle, we can dissect  $P$  into two lattice polygons  $P_1$  and  $P_2$  having

smaller area. By induction Pick's Theorem holds for  $P_1$  and  $P_2$ . But then Pick's Theorem holds for  $P$  as well. This completes the proof.

### 8.5. The Polygon Dissection Theorem

We continue with the theme of polygon dissections. Here we prove a classic result about polygon dissections. This result is called the *Bolyai–Gerweil Theorem*, but the earliest attribution I have seen is to a work by William Wallace from 1807; see [WAL]. A *dissection* of a polygon  $P$  is a description of  $P$  as the union

$$P_1 \cup \cdots \cup P_n$$

of smaller polygon, no two of which overlap. That is, the polygons have disjoint interiors.

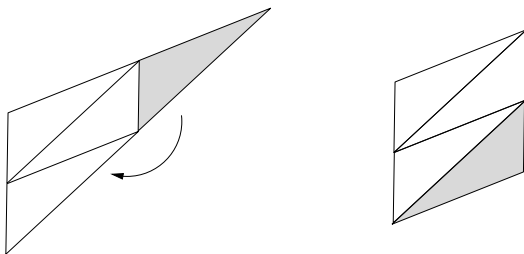
Two polygons  $P$  and  $P'$  are said to be *dissection equivalent* if there are dissections

$$P = \bigcup_{i=1}^n P_i, \quad P' = \bigcup_{i=1}^n P'_i$$

such that  $P_i$  and  $P'_i$  are isometric for all  $i = 1, \dots, n$ . In this case, we write  $P \sim P'$ .

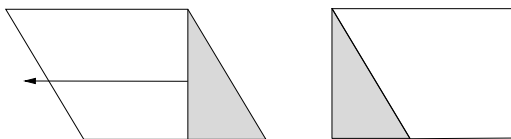
**Exercise 4.** Prove that  $\sim$  is an equivalence relation.

Figure 8.6 illustrates why a triangle is always equivalent to a parallelogram.



**Figure 8.6.** Equivalence between a triangle and a parallelogram

Figure 8.7 illustrates why a parallelogram is always equivalent to a rectangle.

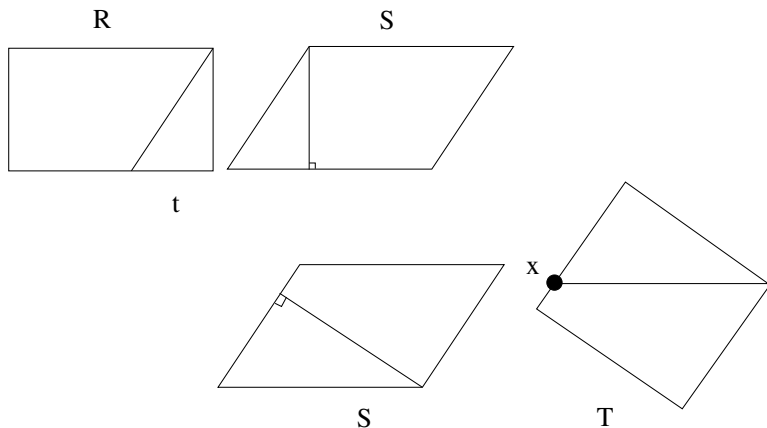


**Figure 8.7.** Equivalence between a parallelogram and a rectangle

Combining the two facts we have just illustrated, we see that a triangle is always equivalent to some rectangle. Let  $R(A, B)$  be a rectangle with side lengths  $A$  and  $B$ . We take  $A < B$ .

**Lemma 8.8.** *Let  $A' \in (A, B)$ . Then  $R(A, B) \sim R(A', B')$ . Here  $B'$  is such that  $A'B' = AB$ . In particular, any rectangle is equivalent to a square.*

**Proof.** Figure 8.8 shows a 2 step construction, based on a real parameter  $t \in (0, B)$ . The first part of the figure shows that  $R \sim S$ , and the second part shows that  $S \sim T$ . The two central figures are both copies of  $S$ , but we have chosen to emphasize a different decomposition in each copy. The shape of the rectangle  $T$  varies continuously with the parameter  $t$ ! The construction works when  $t$  is small, and continues to work until we reach some  $t_0$  so that the point  $x(t_0)$  coincides with a corner of  $T(t_0)$ . But, in this extreme case,  $T$  is a square. As  $t$  varies in  $[0, t_0]$ , the rectangle  $T(t)$  interpolates between  $R(A, B)$  and a square. □



**Figure 8.8.** Two part construction

**Lemma 8.9.** *A triangle of area  $A$  is equivalent to a  $1 \times A$  rectangle.*

**Proof.** First of all, our triangle is equivalent to some rectangle. By the previous result, any two rectangles of the same area are equivalent.  $\square$

Now we can finish the proof. It suffices to prove the result for unit area polygons. Let  $P$  be a polygon of unit area. We first dissect  $P$  into finitely many triangles  $T_1, \dots, T_m$ , having areas  $a_1, \dots, a_m$ . Each  $T_k$  is equivalent to a rectangle  $R(1, a_k)$ . But, when we stack up all these rectangles, we get a rectangle having side lengths 1 and  $\sum a_k = 1$ . That is, any unit area polygon is equivalent to the unit square. The final result is immediate.

You might wonder whether the same result holds for polyhedra in higher dimensions. This turns out to be false, and the result is known as *Dehn's Dissection Theorem*. We will give a proof of Dehn's Dissection Theorem in Chapter 23.

## 8.6. Line Integrals

We now discuss line integrals as a preparation for presenting and proving Green's Theorem. This material can be found in any book on several variable calculus; see, for instance, [SPI].

A *linear functional* is a linear map from  $\mathbf{R}^2$  to  $\mathbf{R}$ . A *1-form* on an open subset  $U \subset \mathbf{R}^2$  is a smooth choice  $p \rightarrow \omega_p$  of a linear functional at each point  $p \in U$ . We mention two special 1-forms,  $dx$  and  $dy$ . These 1-forms are defined on every point of  $\mathbf{R}^2$ , and

$$(8.9) \quad dx(v_1, v_2) = v_1, \quad dy(v_1, v_2) = v_2,$$

for any tangent vector  $(v_1, v_2)$  based at any point. One can write a general 1-form  $\omega$  as a pointwise varying linear combination of these two special ones. That is,

$$(8.10) \quad \omega = f dx + g dy,$$

where  $f, g : U \rightarrow \mathbf{R}$  are smooth functions. At the point  $p$ , we have

$$(8.11) \quad \omega_p(V) = f(p)v_1 + g(p)v_2.$$

Here  $V = (v_1, v_2)$  is some vector based at  $p$ .

Let  $\gamma : [0, 1] \rightarrow \mathbf{R}$  be a smooth curve, and let  $\omega$  be a 1-form. We define

$$\int_{\gamma} \omega = \int_0^1 \omega_{\gamma(t)}(\gamma'(t)) dt.$$

**Exercise 5.** Prove that

$$\int_{\gamma} \omega_1 + \omega_2 = \int_{\gamma} \omega_1 + \int_{\gamma} \omega_2.$$

In other words, the integral is linear.

**Exercise 6.** Prove that

$$\int_{-\gamma} \omega = - \int_{\gamma} \omega.$$

Here  $-\gamma$  is the curve obtained by reversing the direction of  $\gamma$ .

It turns out that the integral only depends on the image and orientation of  $\gamma$ . If

$$s : [0, 1] \rightarrow [0, 1]$$

is an orientation-preserving diffeomorphism, then setting  $\beta = \gamma \circ s$ , we have

**Lemma 8.10.**

$$\int_{\beta} \omega = \int_{\gamma} (\omega).$$

**Proof.** By Exercise 5, it suffices to consider the forms  $f dx$  and  $g dy$ . The proof for  $g dy$  is the same as for  $f dx$ , so we will just consider the case  $\omega = f dx$ . In this case we set  $\gamma(t) = (u(t), v(t))$  and note that

$$\int_{\gamma} \omega = \int_0^1 (f u') dt.$$

Here  $u' = du/dt$ . At the same time

$$\int_{\beta} \omega = \int_0^1 \frac{d(u \circ s)}{dt} f \circ s(t) dt =^* \int_0^1 (f u') \circ s(t) s'(t) dt.$$

The starred equality is the chain rule. The first integral equals the last by the change-of-variables formula for integration.  $\square$

Here is an important observation. Since the line integral only depend on the oriented image of  $\gamma$ , we can specify a line integral just by specifying a curve in the plane and its orientation.

Line integrals can be more generally defined for piecewise smooth curves. To say that  $\gamma$  is a piecewise smooth curve is to say that  $\gamma = \gamma_1 \cup \cdots \cup \gamma_n$ , where each  $\gamma_j$  is a smooth curve, and consecutive curves meet end to end. We define

$$\int_{\gamma} \omega = \sum_{j=1}^n \int_{\gamma_j} \omega.$$

In particular, line integrals make sense for polygonal arcs.

**Exercise 7.** This is a crucial exercise. Let  $P_1$  and  $P_2$  and  $P$  be the polygons from Figure 8.4. Suppose that all these polygons are oriented counterclockwise. Prove that

$$\int_P \omega = \int_{P_1} \omega + \int_{P_2} \omega.$$

## 8.7. Green's Theorem for Polygons

Let  $D$  be a polygon in the plane, and let  $\gamma = \partial D$ , the boundary of  $D$  oriented counterclockwise. Let  $\omega = f dx + g dy$  be a 1-form defined in an open set that contains  $D$  in its interior. Green's Theorem says that

$$(8.12) \quad \int_{\gamma} \omega = \int_D (g_x - f_y) dx dy.$$

Here  $f_y = \partial f / \partial y$  and  $g_x = \partial g / \partial x$ . The integral on the right is a double integral.

In our proof, it is convenient to let  $d\omega$  be the integrand on the right hand side of equation (8.12). We will just use this piece of notation to shorten our equations, but actually  $d\omega$  has a meaning as the exterior derivative of  $\omega$ . See [SPI] if you are curious about this.

We say that a *special triangle* is a right triangle whose sides are parallel to the coordinate axes. The three white triangles in Figure 8.9 below are examples of special triangles.

**Exercise 8.** Let  $D$  be the special triangle with vertices  $(0,0)$  and

$(A, 0)$  and  $(0, B)$  with  $A$  and  $B$  positive. Let  $\gamma$  be the boundary of  $D$ , oriented counterclockwise. Let  $\omega = f dx$ . Prove that

$$\int_{\gamma} \omega = \int_0^A (f(x, 0) - f(x, x')) dx,$$

where  $x'$  (as a function of  $x$ ) is such that  $(x, x')$  lies on the diagonal of  $D$ .

**Lemma 8.11.** *Green's Theorem is true for special triangles.*

**Proof.** Let  $D$  be a special triangle. We can translate the whole picture so that the vertices of  $D$  are as in Exercise 8. By the Fundamental Theorem of Calculus, we get

$$\begin{aligned} \int_D d\omega &= \int_D (-f_y) = \int_{x=0}^A \left( \int_{y=0}^{x'} (-f_y) dy \right) dx \\ &= \int_0^A (f(x, 0) - f(x, x')) dx = \int_{\gamma} \omega. \end{aligned}$$

The last equality comes from Exercise 8. □

Our next result has an easy direct proof, but we will give a rather long-winded proof to illustrate a crucial property of line integrals.

**Lemma 8.12.** *Green's Theorem is true for any rectangle whose sides are parallel to the coordinate axes.*

**Proof.** Let  $R$  be such a rectangle. We write  $R = T_1 \cup T_2$ , where  $T_1$  and  $T_2$  are two special triangles meeting along a diagonal. We certainly have

$$\int_R d\omega = \int_{T_1} d\omega + \int_{T_2} d\omega.$$

On the other hand, by Exercise 7, we have

$$\int_{\partial R} \omega = \int_{\partial T_1} \omega + \int_{\partial T_2} \omega.$$

Here  $\partial R$  denotes the boundary of  $R$  taken counterclockwise, and likewise for the other expressions. Since Green's Theorem holds for special triangles, we can equate the right-hand sides of our last two

equations. But then we can equate the left-hand sides as well. Hence Green's Theorem holds for  $R$ .  $\square$

**Lemma 8.13.** *Green's Theorem is true for any triangle.*

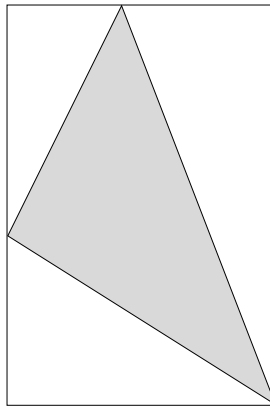
**Proof.** Figure 8.9 shows how we can realize an arbitrary triangle  $D$  as a set of the form  $R - T_1 - T_2 - T_3$ , where  $R$  is a rectangle and  $T_k$  is a special triangle for  $k = 1, 2, 3$ . We have

$$\int_D d\omega + \sum \int_{T_k} d\omega = \int_R d\omega.$$

The same cancellation trick as in the previous lemma shows that

$$\int_{\partial D} \omega + \sum \int_{\partial T_k} \omega = \int_{\partial R} \omega.$$

Green's Theorem, applied to cases we already know, allows us to cancel off all terms, leaving just the one we don't know.  $\square$



**Figure 8.9.** A union of triangles

**Lemma 8.14.** *Green's Theorem is true when the domain  $D$  is an arbitrary polygon.*

**Proof.** Partition  $D$  into triangles and apply the same cancellation trick as above.  $\square$



---

## Chapter 10

# Hyperbolic Geometry

The purpose of this chapter is to give a bare bones introduction to hyperbolic geometry. Most of material in this chapter can be found in a variety of sources, for example [BE1], [KAT], [RAT], or [THU]. The first 2 sections of this chapter might not look like geometry at all, but they turn out to be very important for the subject.

### 10.1. Linear Fractional Transformations

Now we take up the discussion started in §1.6. Suppose that

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

is a  $2 \times 2$  matrix with complex number entries and determinant 1. The set of these matrices is denoted by  $SL_2(\mathbf{C})$ . In fact, this set forms a group under matrix multiplication.

The matrix  $A$  defines a *complex linear fractional transformation*

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

Such maps are also called *Möbius transformations*. Note that the denominator of  $T_A(z)$  is nonzero as long as  $z \neq -d/c$ . It is convenient to introduce an extra point  $\infty$  and define  $T_A(-d/c) = \infty$ . This

definition is a natural one because of the limit

$$\lim_{z \rightarrow -d/c} |T_A(z)| = \infty.$$

The determinant condition guarantees that  $a(-d/c) + b \neq 0$ , which explains why the above limit works. We define  $T_A(\infty) = a/c$ . This makes sense because of the limit

$$\lim_{|z| \rightarrow \infty} T_A(z) = a/c.$$

**Exercise 1.** As in §9.5, we introduce a metric on  $\mathcal{C} \cup \infty$  so that  $\mathcal{C} \cup \infty$  is homeomorphic to the unit sphere  $S^2 \subset \mathbf{R}^3$ . Prove that  $T_A$  is continuous with respect to this metric. (*Hint:* Use the limit formulas above to deal with the tricky points.)

**Exercise 2.** Establish the general formula

$$T_{AB} = T_A \circ T_B,$$

where  $A, B \in SL_2(\mathbf{R})$ . In particular (since  $A^{-1}$  exists) the inverse map  $T_A^{-1}$  exists. By Exercise 1, this map is also a continuous map of  $\mathcal{C} \cup \infty$ . Conclude that  $T_A$  is a homeomorphism of  $\mathcal{C} \cup \infty$ .

## 10.2. Circle Preserving Property

A *generalized circle* in  $\mathcal{C} \cup \infty$  is either a circle in  $\mathcal{C}$  or a set of the form  $L \cup \infty$ , where  $L$  is a straight line in  $\mathcal{C}$ . Topologically, the generalized circles are all homeomorphic to circles. In this section we will prove the following well-known result.

**Theorem 10.1.** *Let  $C$  be a generalized circle and let  $T$  be a linear fractional transformation. Then  $T(C)$  is also a generalized circle.*

One can prove this result by a direct (though tedious) calculation. The book [HCV] has a nice proof involving the geometry of stereographic projection. For fun, I will give a rather unconventional proof. I'll prove 4 straightforward lemmas and then give the main argument.

**Lemma 10.2.** *Let  $C$  be any generalized circle in  $\mathcal{C}$ . Then there exists a linear fractional transformation  $T$  such that  $T(\mathbf{R} \cup \infty) = C$ .*

**Proof.** If  $C$  is a straight line (union  $\infty$ ), then a suitable translation followed by rotation will work. So, consider the case when  $C$  is a circle. The linear fractional transformation

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

maps  $\mathbf{R} \cup \infty$  onto the unit circle  $C_0$  satisfying the equation  $|z| = 1$ . The point is that every point  $z \in \mathbf{R}$  is the same distance from  $i$  and  $-i$ , so that  $|T(z)| = 1$ . Next, one can find a map of the form  $S(z) = az + b$  that carries  $C_0$  to  $C$ . The composition  $S \circ T$  does the job.  $\square$

**Lemma 10.3.** *Suppose that  $L$  is a closed loop in  $\mathbf{C} \cup \infty$ . Then there exists a generalized circle  $C$  that intersects  $L$  in at least 3 points.*

**Proof.** If  $L$  is contained in a straight line (union  $\infty$ ) the result is obvious. Otherwise,  $L$  has 3 noncollinear points and, like any 3 noncollinear points, these lie on a common circle.  $\square$

**Lemma 10.4.** *Let  $(z_1, z_2, z_3) = (0, 1, \infty)$ . Let  $a_1, a_2, a_3$  be a triple of distinct points in  $\mathbf{R} \cup \infty$ . Then there exists a linear fractional transformation that preserves  $\mathbf{R} \cup \infty$  and maps  $a_i$  to  $z_i$  for  $i = 1, 2, 3$ .*

**Proof.** The map  $T(z) = 1/(a_3 - z)$  carries  $a_3$  to  $\infty$ , but does not necessarily do the right thing on the points  $a_1$  and  $a_2$ . However, we can compose  $T$  by a suitable map of the form  $z \rightarrow rz + s$  to fix the images of  $a_1$  and  $a_2$ .  $\square$

**Lemma 10.5.** *Suppose  $T$  is a linear fractional transformation that fixes 0 and 1 and  $\infty$ . Then  $T$  is the identity map.*

**Proof.** Let

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

The condition  $T(0) = 0$  gives  $b = 0$ . The condition  $T(\infty) = \infty$  gives  $c = 0$ . The condition  $T(1) = 1$  gives  $a = d$ . Hence  $T(z) = z$ .  $\square$

Now we can give the main argument. Suppose that there is a linear fractional transformation  $T$  and a generalized circle  $C$  such that  $T(C)$  is not a generalized circle. Composing  $T$  with the map from Lemma 10.2, we can assume that  $C = \mathbf{R} \cup \infty$ . By Lemma 10.3

there is a generalized circle  $D$  such that  $D$  and  $T(\mathbf{R} \cup \infty)$  share at least 3 points. Call these 3 points  $c_1, c_2, c_3$ .

Again by Lemma 10.2, there is a linear fractional transformation  $S$  such that  $S(\mathbf{R} \cup \infty) = D$ . There are points  $a_1, a_2, a_3 \in \mathbf{R} \cup \infty$  such that  $S(a_j) = c_j$  for  $j = 1, 2, 3$ . Also, there are points  $b_1, b_2, b_3 \in \mathbf{R} \cup \infty$  such that  $T(b_j) = c_j$  for  $j = 1, 2, 3$ . By Lemma 10.4 we can find linear fractional transformations  $A$  and  $B$ , both preserving  $\mathbf{R} \cup \infty$  such that  $A(a_j) = z_j$  and  $B(b_j) = z_j$  for  $j = 1, 2, 3$ . Here  $(z_1, z_2, z_3) = (0, 1, \infty)$ . The two maps

$$T \circ B^{-1}, \quad S \circ A^{-1}$$

both map  $(0, 1, \infty)$  to the same 3 points, namely  $(c_1, c_2, c_3)$ . By Lemma 10.5, these maps coincide. However, note that

$$T \circ B^{-1}(\mathbf{R} \cup \infty) = T(\mathbf{R} \cup \infty)$$

is not a generalized circle and  $S \circ A^{-1}(\mathbf{R} \cup \infty) = D$  is a generalized circle. This is a contradiction.

### 10.3. The Upper Half-Plane Model

Now we turn to hyperbolic geometry. We are going to imitate the procedure we used in §9.1 to define the round metric on the sphere. Once we define the hyperbolic plane as a set of points, we will define what we mean by the lengths of curves in the hyperbolic plane. Then, we will proceed as in the case of the sphere.

Let  $U \subset \mathbf{C}$  be the upper half-plane, consisting of points  $z$  with  $\text{Im}(z) > 0$ . As a set, the hyperbolic plane is just  $U$ . However, we will describe a funny way of measuring the lengths of curves in  $U$ . Were we to use the ordinary method, we would just produce a subset of the Euclidean plane. So, given a differentiable curve  $\gamma : [a, b] \rightarrow U$ , we define

$$(10.1) \quad L(\gamma) = \int_a^b \frac{|\gamma'(t)|}{\text{Im}(\gamma(t))} dt.$$

In words, the hyperbolic speed of the curve is the ratio of its Euclidean speed to its height above the real axis.

Here is a simple example. Consider the curve  $\gamma : \mathbf{R} \rightarrow U$  defined by

$$\gamma(t) = i \exp(t).$$

Then the length of the portion of  $\gamma$  connecting  $\gamma(a)$  to  $\gamma(b)$ , with  $a < b$ , is given by

$$\int_a^b \frac{\exp(t)}{\exp(t)} dt = \int_a^b dt = b - a.$$

The image of  $\gamma$  is an open vertical ray, but our formula tells us that this ray, measured hyperbolically, is infinite in both directions. Moreover, the formula tells us that  $\gamma$  is a unit speed curve: it accumulates  $b - a$  units of length between time  $a$  and time  $b$ .

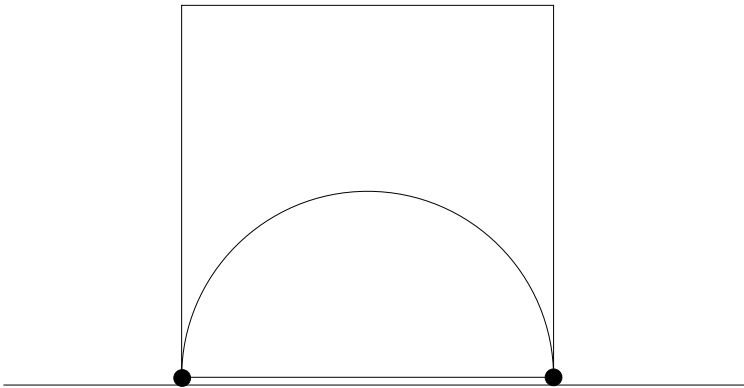
The hyperbolic distance between two points  $p, q \in U$  is defined to be the infimum of the lengths of all piecewise differentiable curves connecting  $p$  to  $q$ . Let us consider informally what these shortest curves ought to look like. Suppose that  $p$  and  $q$  are very near the real axis, say

$$p = 0 + i 10^{-100}, \quad q = 1 + i 10^{-100}.$$

The most obvious way to connect these two points would be to use the path

$$\gamma(t) = t + i 10^{-100}.$$

This curve traces out the bottom of the (Euclidean unit) square shown in Figure 10.1. Our formula tells us that this curve has length  $10^{100}$ .



**Figure 10.1.** Some paths in the hyperbolic plane

Another thing we could do is go around the other three sides of the square. For the left vertical edge, we could use the path  $\gamma$  from our first calculation. This edge has length

$$\log(1) - \log(10^{-100}) = 100.$$

The top horizontal edge has height 1 and Euclidean length 1. So, this leg of the path has length 1. Finally, by symmetry, the length of the right vertical edge is 100. All in all, we have connected  $p$  to  $q$  by a path of length 201. This length is obviously much shorter than the first path. It pays to go upward because, so to speak, unit speed hyperbolic curves cover more ground the farther up they are. Our second path is much better than the first but certainly not the best. For openers, we could save some distance by rounding off the corners. We will show in §10.6 below that the shortest curves, or *geodesics*, in the hyperbolic plane are either arcs of vertical rays or arcs of circles that are centered on the real axis.

When  $U$  is equipped with the metric we have defined, we call  $U$  the *hyperbolic plane* and denote it by  $\mathbf{H}^2$ . So far we have talked about lengths of curves in  $\mathbf{H}^2$ , but we can also talk about angles. The angle between two differentiable and regular (i.e., nonzero speed) curves in  $\mathbf{H}^2$  is defined simply to be the ordinary Euclidean angle between them. That is, the hyperbolic and Euclidean angle between two intersecting curves is just the Euclidean angle between the two tangent vectors at the point of intersection. So, in the upper half-plane model of hyperbolic geometry, the distances are distorted (from the Euclidean model) but the angles are not.

Now that we have talked about hyperbolic length and angles, we discuss hyperbolic area. Given how hyperbolic length relates to Euclidean length, it makes sense to say that the area of a small patch of hyperbolic space is the ratio of its Euclidean area to its height squared. Since the “height” of a patch varies throughout the patch, we really have something infinitesimal in mind. Thus, precisely, we define the hyperbolic area of a region  $D \subset \mathbf{H}^2$  to be the integral

$$(10.2) \quad \int_D \frac{dx \, dy}{y^2}.$$

## 10.4. Another Point of View

An *inner product* on a real vector space  $V$  is a map  $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbf{R}$  which satisfies the following properties:

- $\langle av + w, x \rangle = a\langle v, x \rangle + \langle w, x \rangle$  for all  $a \in \mathbf{R}$  and  $v, w, x \in V$ .
- $\langle x, y \rangle = \langle y, x \rangle$ .
- $\langle x, x \rangle \geq 0$  and  $\langle x, x \rangle = 0$  if and only if  $x = 0$ .

You can remember this by noting that an inner product satisfies the same formal properties as the dot product.

For the moment, we care mainly about inner products on  $\mathbf{R}^2$ . At the point  $z = x + iy$  we introduce the inner product

$$(10.3) \quad \langle v, w \rangle_z = \frac{1}{y^2}(v \cdot w).$$

We mean to apply this to vectors  $v$  and  $w$  that are “based at”  $z$ . We then define the hyperbolic *norm* to be

$$(10.4) \quad \|v\|_{\mathbf{H}^2} = \sqrt{\langle v, v \rangle_z}.$$

With this definition, the length of  $\gamma : [a, b] \rightarrow \mathbf{H}^2$  is given by

$$(10.5) \quad \int_a^b \|\gamma'(t)\|_{\gamma(t)} dt.$$

With this formalism, the notion of hyperbolic length looks much closer to the Euclidean notion. In Chapter 11 we will see that this way of doing things is the beginning of Riemannian geometry.

## 10.5. Symmetries

The hyperbolic metric has more symmetries than you might think. Say that a *real linear transformation* is a linear transformation  $T_A$  based on a matrix with real entries. In this case,  $T_A(z) \in \mathbf{C}$  provided  $z \in \mathbf{C} - \mathbf{R}$ .

**Exercise 3.** Prove that  $z \notin \mathbf{R}$  implies that  $T_A(z) \notin \mathbf{R}$ . Prove also that  $T_A$  maps  $\mathbf{H}^2$  into itself.

The element  $T_A$  is a homeomorphism of  $\mathbf{C} \cup \infty$  which preserves  $\mathbf{H}^2$ .

**Exercise 4.** We say that a real linear fractional transformation is *basic* if it has one of three forms:

- $T(z) = z + 1$ .
- $T(z) = rz$ .
- $T(z) = -1/z$ .

Prove that any real linear fractional transformation is the composition of basic ones.

It turns out that these maps are all hyperbolic isometries. This is pretty obvious for the map  $T(z) = z + 1$ . The hyperbolic metric is built so that the second map is a hyperbolic isometry, and in a moment we will give two proofs of that fact. The really surprising thing is that the third map turns out to be a hyperbolic isometry as well.

**Lemma 10.6.** *The map  $T(z) = rz$  is a hyperbolic isometry.*

**First Proof.** If  $\gamma$  is any curve in  $\mathbf{H}^2$ , then the dilated curve  $T(\gamma)$  moves  $r$  times as fast in the Euclidean sense but is  $r$  times farther from the real axis. Hence  $T(\gamma)$  and  $\gamma$  move at the same hyperbolic speed at corresponding points. So, if we connect points  $p$  and  $q$  by some curve  $\gamma$  we can connect the points  $T(p)$  and  $T(q)$  by the curve  $T(\gamma)$ , which has the same length—and vice versa. This shows that the distance from  $p$  to  $q$  is the same as the distance from  $T(p)$  to  $T(q)$ .

**Second Proof.** Suppose that  $v$  and  $w$  are two vectors based at  $z \in \mathbf{H}^2$ . Then we think of  $dT(v) = rv$  and  $dT(w) = rw$  as two vectors based at  $T(z)$ . Here  $dT$  is linear differential of  $T$ , i.e., the matrix of first partial derivatives. Looking at the formula in equation (10.3), we see that

$$\langle dT(v), dT(w) \rangle_{T(z)} = \langle rv, rw \rangle_{rz} = \frac{1}{r^2 y^2} (rv \cdot rw) = \frac{1}{y^2} (v \cdot w) = \langle v, w \rangle_z.$$



So,  $T$  preserves the hyperbolic inner product at each point. Since the hyperbolic metric is defined entirely in terms of this family of inner products,  $T$  is an isometry.

**Exercise 5.** Prove that the map  $T(z) = -1/z$  is a hyperbolic isometry.

Combining Exercises 4 and 5, we see that any real linear fractional transformation is a hyperbolic isometry of  $\mathbf{H}^2$ . Recall that in §2.8 we proved  $SL_2(\mathbf{R})$  is a 3-dimensional manifold. So,  $\mathbf{H}^2$  has a 3-dimensional group of symmetries!

Say that a *generalized circular arc* is an arc of a generalized circle. We already know that any linear fractional transformation maps generalized circles to circles. Hence, any real linear transformation maps generalized circular arcs to generalized circular arcs.

**Exercise 6.** Prove that a real linear fractional transformation  $T$  has the following property: if  $a$  and  $b$  are two smooth curves in  $\mathbf{H}^2$  which intersect at a point  $x$  and make an angle of  $\theta$ , then  $T(a)$  and  $T(b)$  make the same angle  $\theta$  at the point  $T(x)$ . (*Hint:* If you don't feel like grinding out the calculation, you can assume the result is false and then deduce that the differential  $dT$  fails to map circle to circles. In any case, the result is obvious for all the basic maps except  $z \rightarrow -1/z$ , and so it suffices to consider this one.)

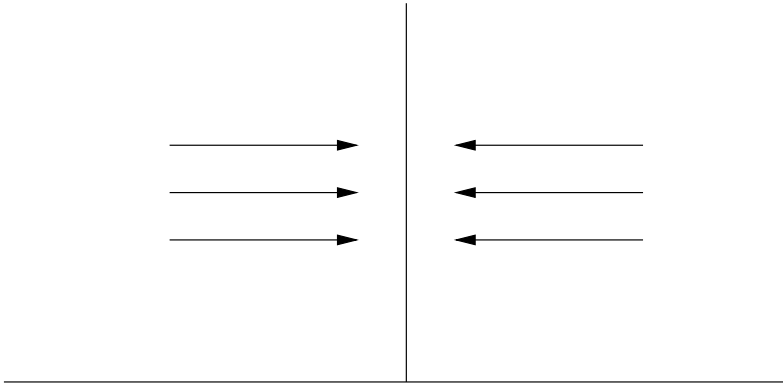
## 10.6. Geodesics

In this section we will describe the shortest curves connecting two points in  $\mathbf{H}^2$ . We first consider the case of points  $p$  and  $q$  that lie on the imaginary axis.

**Lemma 10.7.** *The portion of the imaginary axis connecting  $p$  to  $q$  is the unique shortest curve in  $\mathbf{H}^2$  that connects  $p$  to  $q$ .*

**Proof.** Our proof is very similar to the proof we gave in Lemma 9.1 for the spherical case. Consider the map  $F$  defined by the equation  $F(x + iy) = iy$ ; see Figure 10.2. Looking at the definition of the hyperbolic metric, we see that  $F$  is hyperbolic speed nonincreasing.

That is, if  $\gamma$  is a curve in  $\mathbf{H}^2$ , then the hyperbolic speed of  $F(\gamma)$  at any point is at most the hyperbolic speed of  $\gamma$  at the corresponding point. Moreover, if the velocity of  $\gamma$  has any  $x$ -component at all, then  $F(\gamma)$  is slower at the corresponding point. The idea here is that  $F$  does not change the  $y$ -component of the hyperbolic speed, but kills the  $x$ -component. The total hyperbolic length of  $\gamma$  is the integral of its hyperbolic speed. Thus the hyperbolic length of  $F(\gamma)$  is less than the hyperbolic length of  $\gamma$ , unless  $\gamma$  travels vertically the whole time. Our result follows immediately from this.  $\square$



**Figure 10.2.** The map  $F$

It follows from symmetry that the vertical rays in  $\mathbf{H}^2$  are all geodesics. A vertical ray is the unique shortest path in  $\mathbf{H}^2$  connecting any pair of points on that ray.

**Exercise 7.** Let  $p$  and  $q$  be two arbitrary points in  $\mathbf{H}^2$ . Prove that there is a hyperbolic isometry—specifically, some linear fractional transformation—that carries  $p$  and  $q$  to points that lie on the same vertical ray.

**Theorem 10.8.** *Any two distinct points in  $\mathbf{H}^2$  can be joined by a unique shortest path. This path is either a vertical line segment or else an arc of a circle that is centered on the real axis.*

**Proof.** We have already proved this result for points that lie on the same vertical ray. In light of Exercise 7, it suffices to prove, in general,

that the image of a vertical ray under a linear fractional hyperbolic isometry is one of the two kinds of curves described in the theorem.

Let  $\rho$  be a vertical ray, and let  $T$  be a linear fractional transformation that is also a hyperbolic isometry. From the work in §10.2 we know that  $T(\rho)$  is an arc of a circle. Since  $T$  preserves  $\mathbf{R} \cup \infty$ , both endpoints of this circular arc lie on  $\mathbf{R} \cup \infty$ . Finally, since  $T$  preserves angles,  $T(\rho)$  meets  $\mathbf{R}$  at right angles at any point where  $T(\rho)$  intersects  $\mathbf{R}$ . If  $T(\rho)$  limits on  $\infty$ , then  $T(\rho)$  is another vertical ray. Otherwise,  $T(\rho)$  is a semicircle, contained in a circle that is centered on the real axis.  $\square$

## 10.7. The Disk Model

Now that we have defined geodesics in the hyperbolic plane, we can go forward and define geodesics polygons. Before we do this, we would like to have another model in which to draw pictures. This other model is sometimes more convenient.

Let  $\Delta$  be the open unit disk. There is a (complex) linear fractional map  $M : \mathbf{H}^2 \rightarrow \Delta$  given by

$$(10.6) \quad M(z) = \frac{z - i}{z + i}.$$

This map does the right thing because  $z \in \mathbf{H}^2$  is always closer to  $i$  than to  $-i$  and so  $|M(z)| < 1$ . Since  $M$  maps circles to circles and preserves angles,  $M$  maps geodesics in  $\mathbf{H}^2$  to circular arcs in  $\Delta$  that meet the unit circle at right angles.

Sometimes it is convenient to draw pictures of geodesics in the unit disk rather than in the hyperbolic plane. So, when it comes time to draw pictures, we will be drawing circular arcs that meet the unit circle at right angles. The geodesics that go through the Euclidean center of  $\Delta$  are just unit line segments. The rest of them “bend inward” toward the origin.

**Exercise 8.** Draw pictures of 10 geodesics in the disk model.

Rather than just think of  $\Delta$  as a convenient place to draw pictures, we can also think of  $\Delta$  as another model of  $\mathbf{H}^2$ . The cheapest

way to do this is to say that the distance the two points  $p, q \in \Delta$  is defined to be the hyperbolic distance between the points  $M^{-1}(p)$  and  $M^{-1}(q)$  in  $\mathbf{H}^2$ .

A more direct approach is to define a new inner product at each point  $z \in \Delta$ . The formula is given by

$$(10.7) \quad \langle v, w \rangle_z = \frac{4v \cdot w}{(1 - |z|)^2}.$$

Once we have this inner product, we can directly define lengths of curves in  $\Delta$  as in equation (10.5). Then we can define distances in  $\Delta$  as in the upper half-plane model. It turns out that this new method produces the same result as the cheap method. The proof is a calculation similar to our second proof of Lemma 10.6. We just prove that  $M$  is an isometry relative to the inner product on  $\mathbf{H}^2$  and the inner product on  $\Delta$ .

**Exercise 9.** Prove that the map  $M$  is an isometry from  $\mathbf{H}^2$  and  $\Delta$ , when lengths are defined in terms of the inner product in equation (10.7). That is, prove that

$$\langle v, w \rangle_z = \langle dM(v), dM(w) \rangle_{M(z)}$$

for any pair of vectors  $v$  and  $w$  based at  $z \in \mathbf{H}^2$ .

The disk  $\Delta$ , equipped with its metric, is known as the *Poincaré disk model* of the hyperbolic plane. When  $T$  is a real linear fractional transformation, the map  $M \circ T \circ M^{-1}$  is an isometry of  $\Delta$ . Since  $M$  preserves angles, the hyperbolic angle between two curves in  $\Delta$  is the same as the Euclidean angle between them. Thus, in both our models, Euclidean and hyperbolic angles coincide.

Before we continue, we mention one more piece of terminology. The *ideal boundary* of  $\mathbf{H}^2$  is defined to be  $\mathbf{R} \cup \infty$  in the upper half-plane model and the unit circle in the disk model. Points on the ideal boundary are called *ideal points*. The ideal points are not points in  $\mathbf{H}^2$ . They are considered “limit points” of geodesics in  $\mathbf{H}^2$ .

## 10.8. Geodesic Polygons

Now that we have our two models of the hyperbolic plane, and we know that the geodesics are, we are ready to consider geodesic polygons in the hyperbolic plane. To save words, we will use the term  $\mathbf{H}^2$  rather loosely to refer to either of our two models of the hyperbolic plane. Since there is an isometry, namely  $M$ , carrying one model to the other, there doesn't seem to be much harm in doing this.

Say that a *geodesic polygon* in  $\mathbf{H}^2$  is a simple closed path made from geodesic segments. Here, "simple" means that the path does not intersect itself. Say that a *solid geodesic polygon* is the region in  $\mathbf{H}^2$  bounded by a geodesic polygon. It is convenient to allow some of the "vertices" of the polygon to be ideal points. We call such "vertices" by the name *ideal vertices*. The interior angle of a polygon at an ideal vertex is 0: the two geodesics both meet the ideal point perpendicular to the ideal boundary.

We point out a special geodesic triangle, called an *ideal triangle*. An ideal triangle is a geodesic triangle having 3 infinite geodesic sides and 3 ideal vertices; see Figure 10.3 below. The main result in this section, the Gauss–Bonnet formula for hyperbolic geodesic triangles, is the hyperbolic analogue of the result in §9.3. The proof is very similar, too.

**Theorem 10.9.** *Let  $T$  be a geodesic triangle in the hyperbolic plane. The area of  $T$  equals  $\pi$  minus the sum of the interior angles of  $\pi$ . In particular, the sum of these interior angles is less than  $\pi$ .*

We will give the same kind of proof that we gave for the analogous result in §9.3.

**Lemma 10.10.** *Theorem 10.9 holds for ideal triangles.*

**Proof.** We are trying to prove that any ideal triangle has area  $\pi$ . You can move any one ideal triangle to any other using an isometry of  $\mathbf{H}^2$ . So, it suffices to prove this result for a single triangle. Let us prove this for the triangle  $T$ , in the upper half-plane model, with vertices  $-1$  and  $1$  and  $\infty$ . We first observe that

$$\int_{y=y_0}^{\infty} \frac{1}{y^2} dy = 1/y_0.$$

Now we compute our area, using equation (10.2). Integrating in the  $y$  direction, we have

$$\text{area}(T) = \int_{x=-1}^1 \int_{y=\sqrt{1-x^2}}^{\infty} \frac{1}{y^2} dy = \int_{-1}^1 \frac{1}{\sqrt{1-x^2}} dx = \pi.$$

The last integral is most easily done making the trigonometric substitution  $x = \sin(t)$  and  $dx = \cos(t)$ .  $\square$

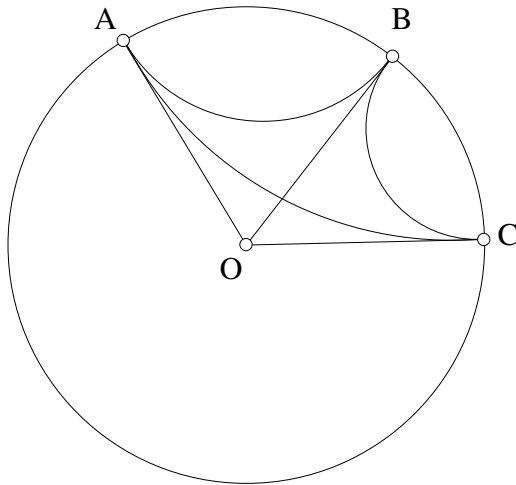
Let  $T(\theta)$  denote a geodesic triangle having two vertices on the ideal boundary of  $\mathbf{H}^2$  and one interior vertex having interior angle  $\theta$ .

**Lemma 10.11.** *Theorem 10.9 holds for  $T(\theta)$ .*

**Proof.** Any two such triangles are isometric to each other. We first match up the interior vertices and then suitably rotate one triangle so that the sides emanating from the common vertex match. In particular, any incarnation of  $T(\theta)$  has the same area. Let

$$f(\theta) = \pi - \text{area}(T(\theta)).$$

We want to show that  $f(\theta) = \theta$  for all  $\theta \in [0, \pi)$ . We already know that  $f(0) = 0$ , by the previous result.



**Figure 10.3.** Two dissections

To analyze the general situation, we work in the disk model and choose  $T(\theta)$  so that it has an interior vertex  $O$  at 0. Figure 10.3 shows a dissection proof that

$$f(\theta_1 + \theta_2) = f(\theta_1) + f(\theta_2),$$

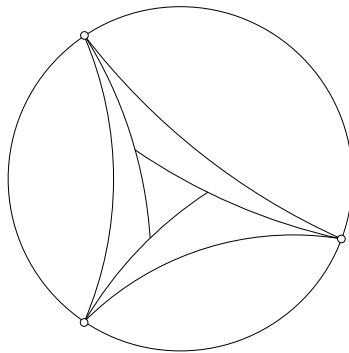
as long as  $\theta_1 + \theta_2 \leq \pi$ . Just to make the picture clear, we point out the following:

- The triangle  $T(\theta_1)$  has vertices  $O, A, B$ .
- The triangle  $T(\theta_2)$  has vertices  $O, B, C$ .
- The triangle  $T(\theta_1 + \theta_2)$  has vertices  $O, A, C$ .
- The triangle with vertices  $A, B, C$  is an ideal triangle.

To make this formula work even when  $\theta_1 + \theta_2 = \pi$ , we set  $f(\pi) = \pi$ . The quadrilateral we have drawn can be dissected in two ways. One way gives  $A_1 + A_2$ . The other way gives  $A + \pi$ . Here  $A_k$  is the area of  $T(\theta_k)$  and  $A$  is the area of  $T(\theta_1 + \theta_2)$ .

Since  $f(\pi) = \pi$ , we can use our formula inductively to show  $f(r\pi) = r\pi$  for any rational  $r \in (0, 1)$ . But the function  $f$  is pretty clearly continuous. Since  $f$  is the identity on a dense set,  $f$  is the identity everywhere.  $\square$

Now we take an arbitrary geodesic triangle and extend the sides so that they hit the ideal boundary of  $\mathbf{H}^2$ . Then we consider the dissection of the ideal triangle defined by the (ideal) endpoints of these sides, as shown in Figure 10.4.



**Figure 10.4.** A dissected ideal triangle

The ideal triangle and also the three outer triangles are of the kind we have already considered. Theorem 10.9 holds true for these. The ideal triangle has area  $\pi$ , and the three outer triangles have areas  $\alpha$ ,  $\beta$ , and  $\gamma$ , the three interior angles of the inner triangle. Hence, the inner triangle has area  $\pi - \alpha - \beta - \gamma$ , as desired. This completes the proof.

A solid geodesic polygon  $P$  is *convex* if it has the following property: if  $p, q \in P$  are two points then the geodesic segment joining  $p$  and  $q$  is also contained in  $P$ . It is easy to prove, inductively, that any convex geodesic polygon can be decomposed into geodesic triangles.

**Lemma 10.12.** *The area of a convex geodesic  $n$ -gon is  $(n - 2)\pi$  minus the sum of the interior angles.*

**Proof.** Just decompose into triangles and then apply the triangle theorem multiple times.  $\square$

**Exercise 10 (Challenge).** Suppose that  $\theta_1, \theta_2, \theta_3$  are three numbers whose sum is less than  $\pi$ . Prove that there is a hyperbolic geodesic triangle with angles  $\theta_1, \theta_2, \theta_3$ .

**Exercise 11 (Challenge).** Say that a geodesic triangle is  $\delta$ -thin if every point in the interior of the (solid version of) triangle is within  $\delta$  of a point on the boundary. Note that there is no universal  $\delta$  so that all Euclidean triangles are  $\delta$ -thin. Prove that all hyperbolic geodesic triangles are 10-thin. (The value  $\delta = 10$  is far from optimal.)

## 10.9. Classification of Isometries

Let  $T$  be a real linear fractional transformation. If  $T(\infty) = \infty$ , then we have  $T(z) = az + b$ . If  $T(\infty) \neq \infty$ , then the equation  $T(z) = z$  leads to a quadratic equation  $az^2 + bz + c$ , with  $a, b, c \in \mathbf{R}$ . If  $T$  is not the identity, then there are 3 possibilities:

- $T$  fixes one point in  $\mathbf{H}^2$  and no other points.
- $T$  fixes no points in  $\mathbf{H}^2$  and one point in  $\mathbf{R} \cup \infty$ .
- $T$  fixes no points in  $\mathbf{H}^2$  and two points in  $\mathbf{R} \cup \infty$ .



$T$  is called *elliptic*, *parabolic*, or *hyperbolic*, according to which possibility occurs. We will discuss these three cases in turn. Before we start, we mention a helpful construction. Given isometries  $g$  and  $T$ , we call  $S = gTg^{-1}$  a *conjugate* of  $T$ . Note that  $g$  maps the fixed points of  $T$  to the fixed points of  $S$ .

Suppose  $T$  is elliptic. Working in the disk model, we can conjugate  $T$  so that the result  $S$  fixes the origin. In this case,  $S$  maps each geodesic through the origin to another geodesic through the origin. Moreover,  $S$  preserves the distances along these geodesics. From here, we see that  $S$  must be a rotation. So, in the disk model, all the elliptic isometries are conjugate to ordinary rotations.

Suppose that  $T$  is parabolic. Working in the upper half-plane model, we can conjugate  $T$  so that the result  $S$  fixes  $\infty$ . In this case  $S(z) = az + b$ . If  $a \neq 1$ , then  $S$  fixes an additional point in  $\mathbf{R}$ . Since this does not happen,  $a = 1$ . Hence  $S(z) = z + b$ . So, in the upper half-plane model, all parabolic isometries are conjugate to a translation.

Suppose that  $T$  is hyperbolic. Working in the upper half-plane model, we can conjugate  $T$  so that the result  $S$  fixes  $0$  and  $\infty$ . But then  $S(z) = rz$  for some  $r \neq 0$ . So, in the upper half-plane model, all hyperbolic isometries are conjugate to dilations (or contractions).

Neither the parabolic elements nor the hyperbolic elements have fixed points in  $\mathbf{H}^2$ , but they still behave in a qualitatively different way. Considering the parabolic map  $S(z) = z + b$ , we see that there is no  $\epsilon > 0$  such that  $S$  moves all points of  $\mathbf{H}^2$  more than  $\epsilon$ . For example, the hyperbolic distance between  $iy$  and  $S(iy)$  tends to  $0$  as  $y \rightarrow \infty$ . On the other hand, if we examine the map  $S(z) = rz$ , we see that there is some  $\epsilon > 0$  such that  $S$  moves all points of  $\mathbf{H}^2$  by at least  $\epsilon$ . Indeed,  $\epsilon = |\log(r)|$ .

---

## Chapter 12

# Hyperbolic Surfaces

In this chapter we will take up the informal discussion from §1.5. We will first explain what a hyperbolic surface is, and then we will show how the gluing construction discussed informally in §1.5 leads to a hyperbolic surface; see [RAT] for a much more general treatment. In fact, we will present a general method of constructing hyperbolic surfaces out of convex geodesic hyperbolic polygons. At the end, we will prove that every complete hyperbolic surface is covered by the hyperbolic plane.

### 12.1. Definition

We will give two definitions of a hyperbolic surface. The first definition requires the material in the last chapter while the second definition does not.

**Definition 12.1.** A hyperbolic surface is a smooth surface with a Riemannian metric, such that each point on the surface has a neighborhood that is isometric to an open disk in the hyperbolic plane.

Our second definition is more elementary and does not require the material on Riemannian manifolds discussed in the previous chapter.

On the other hand, the second definition requires a few preliminaries of its own. Let  $U$  and  $V$  be two open subsets of  $\mathbf{H}^2$ . Say that a *disk-like set* is a subset of the plane that is homeomorphic to an open disk. Say that a map  $f : U \rightarrow V$  is a *local hyperbolic isometry* if the restriction of  $f$  to each open component of  $U$  agrees with the restriction of a hyperbolic isometry. The easiest case to think about is when  $U$  and  $V$  are both connected. Then  $f : U \rightarrow V$  is a local isometry iff  $f$  is the restriction of a hyperbolic isometry to  $U$ .

**Definition 12.2.** A *hyperbolic structure* on  $\Sigma$  is an atlas of coordinate charts on  $\Sigma$  such that the following holds:

- The image of every coordinate chart is a disk-like subset of  $\mathbf{H}^2$ .
- The overlap functions are local hyperbolic isometries.
- The atlas is maximal.

Now we reconcile the two definitions. Suppose that  $\Sigma$  is a hyperbolic surface according to Definition 12.1. Then the local isometries mentioned in Definition 12.1 give rise to an atlas of coordinate charts in which the overlap functions are local isometries. This atlas is not maximal, but then we can complete it to a maximal atlas using Zorn's lemma. (See any book on set theory, such as [DEV], for a discussion of Zorn's lemma.) In this way, we see that  $\Sigma$  is a hyperbolic surface according to Definition 12.2.

**Exercise 1.** Prove that a local hyperbolic isometry is a smooth map. This amounts to showing that a linear fractional is infinitely differentiable.

Suppose that  $\Sigma$  is a hyperbolic surface according to Definition 12.2. According to Exercise 1, the system of coordinate charts on  $\Sigma$  has smooth overlap functions. Therefore,  $\Sigma$  is a smooth surface. We can define a Riemannian metric on  $\Sigma$  as follows. Let  $p \in \Sigma$  be a point. Let  $(U, f)$  be a coordinate chart about  $p$ . This means that  $U$  is an open neighborhood of  $p$  and  $f : U \rightarrow \mathbf{H}^2$  is a homeomorphism onto a disk-like set. Let  $V, W \in T_p(\Sigma)$  be two tangent vectors. This is to say  $V = [\alpha]$  and  $W = [\beta]$  where  $\alpha, \beta : (-\epsilon, \epsilon) \rightarrow \Sigma$  are smooth curves

with  $\alpha(0) = \beta(0) = p$ . We define

$$H_p(V, W) = G_{f(p)}((f \circ \alpha)'(0), (f \circ \beta)'(0)).$$

Here  $G$  is the Riemannian metric on the hyperbolic plane. In other words, we have just used the coordinate chart to transfer the metric on  $\mathbf{H}^2$  to the tangent space  $T_p\Sigma$  of  $\Sigma$  at  $p$ . The fact that the overlap functions are all hyperbolic isometries implies that the above definition of the metric is independent of which coordinate chart is used. This puts a Riemannian metric on  $\Sigma$  with the desired properties. Equipped with this metric,  $\Sigma$  satisfies Definition 12.1.

Now we know that the two definitions pick out the same objects as hyperbolic surfaces.

## 12.2. Gluing Recipes

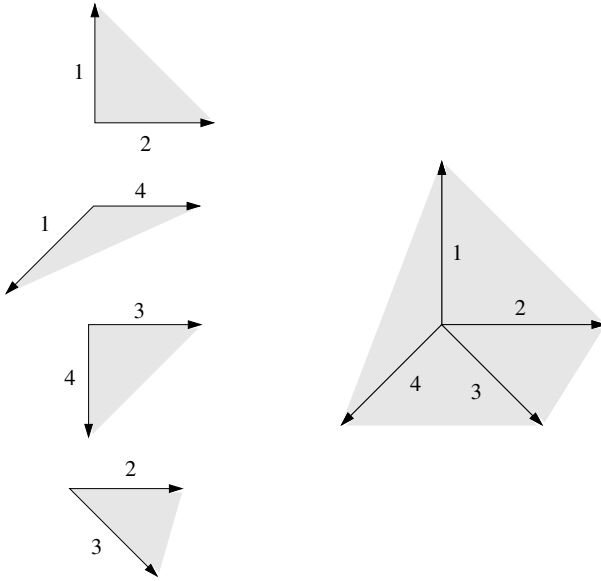
We would like a way to systematically build hyperbolic surfaces. Recall from §10.8 that a *convex geodesic polygon* is a convex subset of  $\mathbf{H}^2$  whose boundary consists of a simple closed path of geodesic segments. The idea is to glue together a bunch of geodesic polygons, taking care to get the angle sums correct.

Let  $P$  be a geodesic polygon. Let  $e \in P$  be an edge. Say a *decoration* of  $e$  is a labelling of  $e$  by both a number and an arrow. Say that a *decoration* of  $P$  is a decoration of every edge of  $P$ . Whenever we have built surfaces by gluing the sides of a polygon together, we have always based the construction on a decoration of the polygon.

We say that a *gluing recipe* for a hyperbolic surface is a finite list  $P_1, \dots, P_n$  of decorated polygons. There are some conditions we want to force:

- If some number appears as a label, then it appears as the label for exactly two edges. This condition guarantees that we will glue the edges together in pairs.
- If two edges have the same numerical label, then they have the same hyperbolic length. This allows us to make our gluing using (the restriction of) a hyperbolic isometry.

- Any *complete circuit* of angles adds up to  $2\pi$ . This condition guarantees that a neighborhood of each vertex is locally isometric to  $H^2$ .



**Figure 12.1.** A complete circuit

The third condition requires some explanation. A *complete circuit* is a collection of edges

$$e_1, e'_1, e_2, e'_2, e_3, e'_3, \dots, e'_k, e_1.$$

with the property, for all  $j$ , that  $e_j$  and  $e'_j$  have the same numerical label and  $e'_j$  and  $e_{j+1}$  are consecutive edges of the same polygon. (Here we are taking the indices cyclically, so that  $k+1$  is set equal to 1.) Figure 12.1 shows what we have in mind.

There is one subtle condition that we need also to require. Let  $v_j$  be the vertex incident to  $e'_j$  and  $e_{j+1}$ . Then the arrow along  $e_{j+1}$  points to  $v_j$  iff the arrow along  $e'_{j+1}$  points to  $v'_{j+1}$ . Figure 12.1 depicts a situation where this holds. The point here is that we want the edges in our chain to emanate from a single vertex in the quotient space. The edges  $e_j$  and  $e'_{j+1}$  subtend an angle  $\alpha_j$  and we want  $\alpha_1 + \dots + \alpha_k = 2\pi$ .

## 12.3. Gluing Recipes Lead to Surfaces

**Theorem 12.1.** *Any gluing recipe gives rise to a hyperbolic surface.*

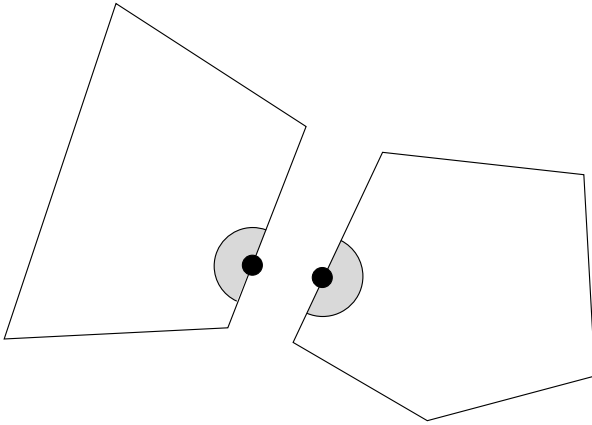
**Proof.** Given a gluing recipe, we can form a surface  $\Sigma$  as follows. First of all, we start out with the metric space  $X$  which is the disjoint union of  $P_1, \dots, P_n$ . We can do this by declaring  $d(p, q) = 1$  if  $p \in P_i$  and  $q \in P_j$  with  $j \neq i$ . For  $p, q \in P_j$  (the same polygon) we just use the hyperbolic metric. So, you should picture  $X$  approximately as a stack of polygons hovering in the air, as on the left-hand side of Figure 12.1.

Now we define an equivalence relation on  $X$  using the rule that  $p \sim p'$  iff  $p$  and  $p'$  are corresponding points on like-numbered edges. Here *corresponding* should be pretty obvious. Suppose  $e$  and  $e'$  are two like-numbered edges, both having length  $\lambda$ . Then there is some  $t$  such that  $p$  is  $t$  units along  $e$  measured in the direction of the arrow. Likewise there is some  $t'$  such that  $p'$  is  $t'$  units along  $e'$ . Then  $p$  and  $p'$  are corresponding points iff  $t = t'$ .

The nontrivial equivalence classes typically have 2 members, with 1 member coming from each edge. However, for the vertices of the polygons, each of which belongs to two edges, the corresponding equivalence class might be larger. In Figure 12.1, the equivalence class of the relevant vertex has 4 elements.

The surface is defined as  $\Sigma = X/\sim$ . We would like to show that  $\Sigma$  is indeed a surface, so we have to construct an atlas of coordinate charts. Suppose that  $x$  is an interior point of some polygon  $P$ . Then some open neighborhood  $U_x$  of  $x$  remains in the interior of  $P$ . No point in  $U_x$  is equivalent to any other point of  $\Sigma$ . The inclusion map  $U_x \rightarrow P \subset \mathbf{H}^2$  gives a coordinate chart from  $U_x$  to  $\mathbf{H}^2$ . We take  $U_x$  to be a metric disk.

Suppose now that  $p \in \Sigma$  is an equivalence class consisting of two points, in the interiors of a pair of edges, that are glued together when the edges are paired. That is,  $p = \{q, q'\}$ , with  $q \in e$  and  $q' \in e'$ , where  $e$  and  $e'$  are open edges. Let  $P$  and  $P'$  be the polygons containing  $e$  and  $e'$ , respectively. Let  $U$  and  $U'$  be small half-disk neighborhoods of  $q$  and  $q'$  in  $P$  and  $P'$ , respectively, as shown in Figure 12.2.



**Figure 12.2.** Half-disk neighborhoods

We define  $h : U \cup U' \rightarrow \mathbf{H}^2$  so that the following hold.

- The map  $h$ , when restricted to either  $U$  or  $U'$ , is the inclusion map composed with a hyperbolic isometry.
- $h(e \cap U) = h'(e' \cap U')$  and the arrows go the right way.
- $h(U)$  and  $h(U')$  lie on opposite sides of  $h(e) = h(e')$ .

This is pretty obvious. We first define  $h$  as the inclusion map on both halves, and then we compose one half of the map with a suitable isometry to adjust things. The main point here is that  $U \cap e$  and  $U' \cap e'$  are open geodesic segments of the same length.

**Exercise 2.** Prove that  $\Delta = (U \cup U') / \sim$  is homeomorphic to an open disk. More precisely, prove that  $h$  defines a homeomorphism  $\Delta$  to a disk in  $\mathbf{H}^2$ . Finally, prove that  $\Delta$  is an open neighborhood of  $p$  in  $\Sigma$ .

Finally, suppose that  $p$  is the equivalence class coming from some vertices of our polygons. Then we have one of the circuits mentioned above. Let  $\{q_1, \dots, q_k\}$  be the equivalence class of  $p$ . In the example shown in Figure 12.1, we have  $k = 4$ . Let  $P_j$  be the polygon that has  $q_j$  as a vertex. In each  $P_j$  we choose a little wedge-shaped neighborhood consisting of all points of  $P_j$  within  $\epsilon$  of  $q_j$ .

**Exercise 3.** Prove that the union  $(U_1 \cup \cdots \cup U_k)/\sim$  is homeomorphic to a disk.

We define a map  $h : U_1 \cup \cdots \cup U_k \rightarrow \mathbf{H}^2$  in such a way that the following holds.

- The map  $h$ , when restricted to any  $U_j$ , is the inclusion map composed with a hyperbolic isometry.
- $h$  respects the gluing of edges.

Expressing the last condition is a bit clumsy, but I hope that you can see what it means. If two edges are glued together, then  $h$  sends them (or at least the portions inside our little pizza slices) to the same segment in  $\mathbf{H}^2$ .

**Exercise 4.** Prove that  $(U_1 \cup \cdots \cup U_k)/\sim$  is an open neighborhood of  $p$  in  $\Sigma$  and that  $h$  gives a homeomorphism from this set onto an open disk in  $\mathbf{H}^2$ . (*Hint:* The circuit condition guarantees that the images of  $h$  fit together to make a single hyperbolic disk.)

From the way we have defined things, the overlap functions are all local hyperbolic isometries, so we have found an atlas on  $\Sigma$  whose overlap functions are local hyperbolic isometries. We can complete this to a maximal atlas, if we like, using Zorn's lemma.  $\square$

## 12.4. Some Examples

Here are some additional examples for you to work out. The first exercise asks you to work out the discussion in §1.5. The next example points to more flexible and systematic approach.

**Exercise 5.** Prove that there is a regular convex  $4n$ -gon, with angles  $\pi/2n$ , provided that  $n \geq 2$ . Call this polygon  $P_{4n}$ . Decorate  $P_{4n}$  by giving the opposite sides and making the arrows point in the same direction. See Figure 1.7. Prove that  $P_{4n}$ , as decorated, is a gluing diagram for a hyperbolic surface.



**Exercise 6.** Prove that there exists a right angled regular hexagon. Construct a decoration of  $4n$  such hexagons in such a way that it is the gluing diagram for a hyperbolic surface.

**Exercise 7 (Challenge).** If you take  $n = 2$  in Exercises 5 and 6 you get homeomorphic surfaces. Prove that they are not isometric.

**Exercise 8 (Challenge).** Prove that there are uncountably many surfaces, all homeomorphic to the octagon surface from Exercise 5, no two of which are isometric to each other.

## 12.5. Geodesic Triangulations

So far, we have shown how to build some hyperbolic surfaces from gluing diagrams. In this section we will show that every compact hyperbolic surface arises from this construction. We begin with a well-known construction in  $\mathbf{H}^2$ .

Let  $X \subset \mathbf{H}^2$  be a finite collection of points. For each  $p \in X$ , we let  $N_p$  be the set of points that are closer to  $p$  than to any point of  $X$ .

**Lemma 12.2.**  *$N_p$  is convex. If  $N_p$  is bounded, then  $N_p$  is the interior of a convex geodesic hyperbolic polygon.*

**Proof.** Say that a *geodesic half-plane* is a set of points in  $\mathbf{H}^2$  lying to one side of a hyperbolic geodesic. Geodesic half-planes are convex. Given any two points  $p, q \in \mathbf{H}^2$ , the set of points closer to  $p$  is a geodesic half-plane. For this reason,  $N_p$  is the intersection of finitely many geodesic half-planes, and the boundary of  $N_p$  is contained in a finite union of geodesics. Since the intersection of convex sets is convex,  $N_p$  is convex. In case  $N_p$  is bounded, the boundary evidently is a convex geodesic polygon.  $\square$

Say that a *geodesic triangulation* of a hyperbolic surface is a decomposition of the surface as the finite union of geodesic triangles. Every pair of triangles should either be disjoint or share an edge or share a vertex. If a hyperbolic surface has a geodesic triangulation,

then we can cut the surface open along the triangles and thereby obtain a description of the surface in terms of a gluing diagram.

**Lemma 12.3.** *Every compact hyperbolic surface has a geodesic triangulation.*

**Proof.** Let  $S$  be the surface. By compactness, there is some  $d \in (0, 1)$  such that every disk of radius  $d$  on the surface is isometric to a disk of radius  $d$  in  $\mathbf{H}^2$ . Place a finite number of points on  $S$  in such a way that every disk of radius  $D/K$  contains at least one point. The constant  $K$  is yet to be determined. Let  $X$  denote this finite set of points.

Given  $p \in X$ , let  $B_d(p)$  denote the disk of radius  $d$  about  $p$ . Let  $N_p \subset S$  be the set of points in  $S$  that are closer to  $p$  than to any other point in  $X$ . We claim that each  $N_p$  is isometric to the interior of a convex geodesic hyperbolic polygon provided that  $K$  is large enough. (This is not an immediate consequence of the previous result because we are working on a surface and not directly in  $\mathbf{H}^2$ .) The boundary of  $N_p$  consists of points  $q$  such that  $q$  is equidistant between  $p$  and some other point  $p'$  of  $X$ . Let  $X_p$  denote the set of points  $p' \in X$  such that some point of  $N_p$  is equidistant from  $p$  and  $p'$ . We can choose  $K$  large enough so that  $N_p \subset B_d(p)$  and  $X_p$  consists entirely of points in the  $B_d(p)$ . Now we apply the previous result. This shows that  $N_p$  is the interior of a convex geodesic polygon.

We have partitioned  $S$  into convex geodesic polygons. To finish the triangulation, we just add in extra geodesic segments, as needed, to divide each of the convex polygons into triangles.  $\square$

Lemma 12.3 allows us to prove the Gauss–Bonnet Theorem for hyperbolic surfaces.

**Theorem 12.4** (Gauss–Bonnet). *The hyperbolic area of a compact hyperbolic surface  $S$  is  $-2\pi\chi(S)$ , where  $\chi(S)$  is the Euler characteristic of  $S$ . In particular, the area only depends on the Euler characteristic.*

**Proof.** We give  $S$  a geodesic triangulation. From §3.4, we have the formula

$$(12.1) \quad \chi(S) = F - E + V,$$

where  $F$  is the number of faces in the triangulation,  $E$  is the number of edges, and  $V$  is the number of vertices.

Each triangle in the triangulation has 3 edges, and each edge belongs to two triangles. For this reason,  $E = 3F/2$ . At the same time, the total sum of all the interior angles of all the triangles is  $2\pi V$ , because the sum of these angles around any one vertex is  $2\pi$ . Putting these equations together, we get

$$(12.2) \quad \chi(S) = -\frac{F}{2} + V = -\frac{F}{2} + \frac{1}{2\pi} \sum_{\text{angles}} \theta_i.$$

For each triangle  $\tau$ , let  $\theta_i(\tau)$ , for  $i = 1, 2, 3$ , be the three interior angles of  $\tau$ . Hence

$$(12.3) \quad \begin{aligned} -2\pi\chi(S) &= \pi \left( F - \sum_{\text{angles}} \theta_i \right) \\ &= \sum_{\text{triangles}} \left( \pi - \theta_1(\tau) - \theta_2(\tau) - \theta_3(\tau) \right) \\ &=^* \sum_{\text{triangles}} \text{area}(\tau) \\ &= \text{area}(S). \end{aligned}$$

The starred equality comes from Theorem 10.9. □

Theorem 12.4 is a special case of the Gauss-Bonnet Theorem from differential geometry. See [BAL] for a discussion of the proof of this general result.

## 12.6. Riemannian Covers

We say that a *Riemannian cover* of a Riemannian manifold  $X$  is a Riemannian manifold  $\tilde{X}$  such that the covering map  $E : \tilde{X} \rightarrow X$  is a local isometry. We mean that the differential  $dE$  is an isometry

on each tangent plane, measured with respect to the two Riemannian metrics.

**Lemma 12.5.** *Suppose that  $X$  is a Riemannian manifold and  $\tilde{X}$  is a covering space of  $X$ . Then one can make  $\tilde{X}$  into a Riemannian manifold in such a way that the covering map  $E : \tilde{X} \rightarrow X$  is a Riemannian cover.*

**Proof.** First of all,  $\tilde{X}$  inherits the structure of a manifold. We have the covering map  $E : \tilde{X} \rightarrow X$ . Each point  $\tilde{x} \in \tilde{X}$  lies in a small open neighborhood  $\tilde{U}$  such that  $U = E(\tilde{U})$  is an evenly covered neighborhood of  $x = E(\tilde{x})$  and also  $(U, \phi)$  is a coordinate chart for  $x$ . The composition  $\phi \circ E : \tilde{U} \rightarrow \mathbf{R}^n$  gives a coordinate chart for an open neighborhood of  $\tilde{x}$ . The overlap functions for these coordinate charts on  $\tilde{X}$  are the same as for the coordinate charts on  $X$ . Hence  $\tilde{X}$  is a smooth manifold and  $E$  is a smooth map.

There exists a unique Riemannian metric on  $\tilde{X}$  so that  $E : \tilde{X} \rightarrow X$  is an isometry. We define the metric  $\tilde{g}$  such that

$$\tilde{g}_{\tilde{x}}(X, Y) = g_x(dE(X), dE(Y)).$$

Here  $dE$  is the differential of  $E$ . Here  $X$  and  $Y$  are tangent vectors to  $\tilde{X}$  at  $\tilde{x}$ . When measured in the local coordinates we have described, the differential  $dE$  is just the identity map. So, the metric  $\tilde{g}$  is actually an inner product.

There is a second way to think about the Riemannian metric on  $\tilde{X}$  which perhaps is more clear. The Riemannian metric on  $X$  is just a collection of Riemannian metrics on various open sets of  $\mathbf{R}^n$  that are compatible in the sense that all overlap functions are isometries. We may, first of all, restrict our attention to open sets in  $X$  that are evenly covered by the covering map. We can then use the preimages of these open sets as coordinate charts in  $\tilde{X}$ . Since the overlap functions for the charts on  $\tilde{X}$  are the same as on  $X$ , the same collection of compatible metrics defines a Riemannian metric on  $\tilde{X}$ .  $\square$

**Exercise 9.** Show that a Riemannian covering map  $E : \tilde{X} \rightarrow X$  is distance nonincreasing. Also, give an example of a Riemannian covering from a connected space  $\tilde{X}$  to a connected space  $X$  that is

not a global isometry. That is, give an example where there are points  $\tilde{x}, \tilde{y} \in \tilde{x}$  that are farther apart than their corresponding images  $x, y \in X$ .

Recall that a metric space is *complete* if every Cauchy sequence in the space converges. For a Riemannian manifold, there is a different notion, called *geodesic completeness*, which people often mean when they say that a Riemannian manifold is complete. However, the two definitions are the same, thanks to the Hopf–Rinow Theorem. See [DOC] for a proof. We mention this just to keep consistent with other texts. We only care about the metric completeness.

**Lemma 12.6.** *Let  $E : \tilde{X} \rightarrow X$  be a Riemannian covering space. If  $X$  is complete, then so is  $\tilde{X}$ .*

**Proof.** Let  $\{\tilde{x}_n\}$  be a Cauchy sequence in  $\tilde{X}$ . We have constructed things in such a way that the map  $E : \tilde{X} \rightarrow X$  is distance nonincreasing. Setting  $x_n = E(\tilde{x}_n)$ , we now know that  $\{x_n\}$  is a Cauchy sequence in  $X$ . Since  $X$  is complete, there is some limit point  $x_*$ . There is an evenly covered neighborhood  $U$  of  $x_*$  which contains  $x_n$  for  $n$  large. But then all the points  $\tilde{x}_n$  lie in the same component of  $\tilde{E}^{-1}(U)$  for  $n$  large. But  $E : \tilde{U} \rightarrow U$  is a homeomorphism. In particular,  $E$  maps convergent sequences to convergent sequences and so does  $E^{-1}$ . Since  $\{x_n\}$  is a convergent sequence in  $U$ , the sequence  $\{\tilde{x}_n\}$  is a convergent sequence in  $\tilde{U}$ .  $\square$

## 12.7. Hadamard’s Theorem

In this section we prove Hadamard’s Theorem, in two dimensions. See [DOC] for a proof in general. The version of Hadamard’s Theorem we prove is a technical step in our proof that any complete hyperbolic surface is covered by  $\mathbf{H}^2$ . Just for this section, let  $\mathbf{H} = \mathbf{H}^2$  stand for the hyperbolic plane.

**Theorem 12.7** (Hadamard). *Let  $H$  be a complete and simply connected surface that is locally isometric to  $\mathbf{H}^2$ . Then  $H$  is globally isometric to  $\mathbf{H}^2$ .*

A surface is *oriented* if we can make a continuous choice of basis for each tangent plane. Any simply connected surface is oriented. Let  $h \in H$  be a point and let  $\mathbf{h} \in \mathbf{H}$  be a point. Both points have neighborhoods which are isometric to disks in the hyperbolic plane. Thus we can find an isometry  $I$  between a neighborhood  $U \subset H$  of  $h$  and a neighborhood  $U \subset \mathbf{H}$  of  $\mathbf{h}$ . Let  $x \in H$  be any point. We can take  $I$  to be orientation preserving.

Let  $\gamma$  be a continuous path connecting  $h$  to  $x$ .

**Lemma 12.8.**  *$I$  can be extended to a neighborhood of  $\gamma$  in such a way that  $I$  is a local isometry at every point along  $\gamma$ .*

**Proof.** We think of  $\gamma$  as a map from  $[0, 1]$  to  $\mathbf{H}$ , with  $\gamma(0) = h$  and  $\gamma(1) = x$ . Say that a point  $t \in [0, 1]$  is good if this lemma holds for the restriction of  $\gamma$  to the interval  $[0, t]$ . Note that 0 is good. Note also that if  $t$  is good, then so is  $s \in [0, t]$ . Hence the set  $J$  of good points is an interval that contains 0. Moreover, since local isometries are defined on open sets,  $J$  is an open interval.

We claim that  $J$  is a closed interval. Suppose that all points  $t \in [0, s)$  are good. We take a sequence of points  $\{s_n\} \in [0, s)$  such that  $s_n \rightarrow t$ . Then  $\{\gamma(s_n)\}$  is a Cauchy sequence. Since  $I$  is not distance increasing,  $\{I(\gamma(s_n))\}$  is also a Cauchy sequence. Since  $\mathbf{H}$  is complete, this Cauchy sequence converges. We define

$$I(t) = \lim I(\gamma(s_n)).$$

We would like to see that in fact  $I$  is defined and a local isometry in a neighborhood of  $\gamma(t)$ .

There is a local isometry  $I'$  carrying a neighborhood  $U$  of  $\gamma(t)$  to a disk in  $\mathbf{H}^2$ . Since every two points have isometric neighborhoods, we can assume that  $I'$  and  $I$  agree on  $\gamma(t)$ . Once  $n$  is large, we have  $\gamma(s_n) \in U$ . The points  $I(\gamma(s_n))$  and  $I'(\gamma(s_n))$  are the same distance from  $I(\gamma(t))$ . So, we may adjust  $I'$  by a rotation so that  $I$  and  $I'$  agree on some  $\gamma(s_n)$ . But then  $I$  and  $I'$  agree on all of  $\gamma(s_n, t]$ . The point is that two orientation-preserving isometries agree everywhere provided that they agree on two points. This shows that the union map  $I \cup I'$  is a local isometry at all points of  $\gamma[0, t]$ .

Our argument shows that  $t$  is good, and therefore that  $J$  is a closed interval. Since  $J$  is open, closed, and connected, we must have that  $J = [0, 1]$ .  $\square$

Now we have a candidate map  $I : H \rightarrow \mathbf{H}$ . However, we need to see that this map is well defined. That is, we need to see that the point  $I(x)$  is independent of the choice of path  $\gamma$  joining  $h$  to  $x$ . This is where we use the simple connectivity assumption.

Let  $\gamma_0$  and  $\gamma_1$  be two paths joining  $h$  to  $x$ . We think of  $\gamma_0$  and  $\gamma_1$  both as maps from  $[0, 1]$  into  $H$ , with  $\gamma_0(0) = \gamma_1(0) = h$  and  $\gamma_0(1) = \gamma_1(1) = x$ . Since  $H$  is simply connected, there is a path homotopy  $\gamma_t$  from  $\gamma_0$  to  $\gamma_1$ . The point  $\mathbf{x}_t = I(\gamma_t(1))$  varies continuously with  $t$ . On the other hand, note that the same extension in the above lemma works for both  $\gamma_s$  and  $\gamma_t$  as long as  $s$  and  $t$  are close together. Hence  $\mathbf{x}_s = \mathbf{x}_t$  for  $s$  and  $t$  close. But this shows that  $\mathbf{x}_t$  does not move at all.

Our extension gives a local isometry  $I : H \rightarrow \mathbf{H}$ . But the existence of our extension just used the following.

- Completeness of  $\mathbf{H}$ .
- Local homogeneity of  $\mathbf{H}$ , in connection with the map  $I'$  above.
- Path connectivity and simple connectivity of  $H$ .

All these properties hold with the two spaces reversed. Reversing the roles of  $H$  and  $\mathbf{H}$ , we construct the inverse map  $I^{-1}$  using the same method. Hence both  $I$  and  $I^{-1}$  are homeomorphisms and local isometries. Bring local isometries, both maps  $I$  and  $I^{-1}$  are globally distance nonincreasing. This is only possible if both these maps are global isometries.

## 12.8. The Hyperbolic Cover

We are almost done with the proof that every complete hyperbolic surface is covered by the hyperbolic plane. We just need one more technical result.

**Lemma 12.9.** *A complete hyperbolic surface is good in the sense of Chapter 7.*

**Proof.** Let  $X$  be a complete hyperbolic surface. A sufficiently small ball about any  $x \in X$  is isometric to a hyperbolic disk. Such sets are obviously both conical and simply connected. Indeed, we can join each point  $y \in B_\epsilon(x)$  to  $x$  by a geodesic. We just need to see that any path in  $X$  is good.

Consider a continuous path  $f_0 : [0, 1] \rightarrow X$ . Every point  $x \in f_0[0, 1]$  has a neighborhood  $U_x$  that is isometric to a hyperbolic disk. By compactness, there is a single positive constant, say  $2\epsilon$  that works for all points of  $f_0[0, 1]$ . Let  $f_1 : [0, 1] \rightarrow X$  be a path such that  $D(f_0, f_1) < \epsilon$ . This means that distance between  $f_0(t)$  and  $f_1(t)$  is less than  $\epsilon$ . For each  $t \in [0, 1]$  there is a geodesic  $g_t[0, 1] \rightarrow X$  connecting  $f_0(t)$  to  $f_1(t)$  that remains within the  $\epsilon$ -ball about  $f_0(t)$ .

For  $s$  sufficiently near  $t$ , the two paths  $\gamma_s$  and  $\gamma_t$  lie in the  $2\epsilon$  ball about  $\gamma_0(t)$ . Therefore, the path  $\gamma_t$  varies continuously with  $t$ . But then the map  $F(s, t) = \gamma_s(t)$  gives a homotopy from  $f_0 = F(0, *)$  to  $f_1 = F(1, *)$ .  $\square$

**Theorem 12.10.** *A complete hyperbolic surface is universally covered by  $\mathbf{H}^2$ .*

**Proof.** Let  $X$  be a complete hyperbolic surface. We know that  $X$  is a good metric space in the sense of Chapter 7. By Theorem 7.1, there exists a simply connected covering space  $\tilde{X}$  and a covering map  $E : \tilde{X} \rightarrow X$ . The space  $\tilde{X}$  is complete by Lemma 12.6. But then, by Hadamard's Theorem,  $\tilde{X}$  is isometric to  $\mathbf{H}^2$ .  $\square$

What I (and many people) find really great about this result is that it opens the door to beautiful tilings of the hyperbolic plane. These are the kinds of tilings drawn by M. C. Escher in his *Circle Woodcut* series. Here we will sketch the idea behind these tilings. We begin with a general exercise that justifies the construction we give below.

**Exercise 10.** Let  $\tilde{X} \rightarrow X$  be a Riemannian covering of a complete Riemannian manifold  $X$ . Let  $U$  be a simply connected open subset of  $X$ . Let  $\tilde{U} = E^{-1}(U)$ . Prove that  $U$  is evenly covered by  $\tilde{U}$  and that the restriction of  $E$  to any component of  $\tilde{U}$  is an isometry between



that component and  $U$ . (*Hint*: Imitate the proof of Hadamard's Theorem to construct an inverse map that is also a local isometry.)

Now consider a description of a hyperbolic surface as one obtained by gluing together the sides of a hyperbolic polygon. For instance, if we glue together 4 regular right angled hexagons in a suitable pattern, we get a hyperbolic surface of Euler characteristic  $-2$ ; see §12.4. Let  $X$  be a hyperbolic surface obtained by this construction. The interiors of the right angled hexagons are embedded and simply connected in  $X$ . We can consider the preimages of these open hexagons in  $\mathbf{H}^2$  by pulling them back by the map  $E$ . By Exercise 10, the result is an infinite collection of open right angled hexagons  $\mathbf{H}^2$ .

At the same time,  $X$  contains a graph whose edges are embedded geodesic arcs. These arcs are the images of the edges of the hexagons under the gluing maps. The preimages of these arcs in  $\mathbf{H}^2$  are the interfaces between the open hexagons. The whole picture fits together to give a tiling of  $\mathbf{H}^2$  by right angled hexagons. Being right angled, these hexagons necessarily meet 4 per vertex. This is a hyperbolic geometry analogue of the picture we developed in §6.3.

In §6.3, we actually went the other way around. We started with the tiling and then produced the covering map. The situation here is so concrete that we can actually do the same thing. We take an infinite supply of regular right angled hyperbolic hexagons and glue them together so that they meet 4 per vertex. The same argument as the one given in Chapter 12 shows that the result is locally isometric to the hyperbolic plane. With a bit of effort, one can see that the resulting space is both simply connected and complete, and hence globally isometric to the hyperbolic plane. Once we have built this tiling of  $\mathbf{H}^2$  by hexagons, we can imitate the construction in §6.3, directly producing the covering map from  $\mathbf{H}^2$  to the surface.

Given that we can construct the universal cover  $E : \mathbf{H}^2 \rightarrow X$  directly in this case, without resorting to Theorem 12.10, you might wonder why we need this result at all. I suppose that the best answer to this question is that Theorem 12.10 is completely general.

---

We do not have to fool around with the combinatorics of gluing together infinite families of polygons every time we want to construct the universal cover of a hyperbolic surface.