Mostly Surfaces

Richard Evan Schwartz [*]

August 21, 2011

Abstract

This is an unformatted version of my book *Mostly Surfaces*, which is Volume 60 in the A.M.S. Student Library series. This book has the same content as the published version, but the arrangement of some of the text and equations here is not as nice, and there is no index.

Preface

This book is based on notes I wrote when teaching an undergraduate seminar on surfaces at Brown University in 2005. Each week I wrote up notes on a different topic. Basically, I told the students about many of the great things I have learned about surfaces over the years. I tried to do things in as direct a fashion as possible, favoring concrete results over a buildup of theory. Originally, I had written 14 chapters, but later I added 9 more chapters so as to make a more substantial book.

Each chapter has its own set of exercises. The exercises are embedded within the text. Most of the exercises are fairly routine, and advance the arguments being developed, but I tried to put a few challenging problems in each batch. If you are willing to accept some results on faith, it should be possible for you to understand the material without working the exercises. However, you will get much more out of the book if you do the exercises.

The central object in the book is a surface. I discuss surfaces from many points of view: as metric spaces, triangulated surfaces, hyperbolic surfaces,

[*] Supported by N.S.F. Research Grant DMS-0072607

and so on. The book has many classical results about surfaces, both geometric and topological, and it also has some extraneous stuff that I included because I like it. For instance, the book contains proofs of the Pythagorean Theorem, Pick's Theorem, Green's Theorem, Dehn's Dissection Theorem, the Cauchy Rigidity Theorem, and the Fundamental Theorem of Algebra.

All the material in the book can be found in various textbooks, though there probably isn't one textbook that has it all. Whenever possible, I will point out textbooks or other sources where you can read more about what I am talking about. The various fields of math surrounding the concept of a surface—geometry, topology, complex analysis, combinatorics—are deeply intertwined and often related in surprising ways. I hope to present this tapestry of ideas in a clear and rigorous yet informal way.

My general view of mathematics is that most of the complicated things we learn have their origins in very simple examples and phenomena. A good way to master a body of mathematics is to first understand all the sources that lead to it. In this book, the *square torus* is one of the key simple examples. A great deal of the theory of surfaces is a kind of elaboration of phenomena one encounters when studying the square torus. In the first chapter of the book, I will introduce the square torus and describe the various ways that its structure can be modified and generalized. I hope that this first chapter serves as a good guide to the rest of the book.

I aimed the class at fairly advanced undergraduates, but I tried to cover each topic from scratch. My idea is that, with some effort, you could learn the material for the whole course without knowing too much advanced math. You should be perfectly well prepared for the intended version of the class if you have had a semester each of real analysis, abstract algebra, and complex analysis. If you have just had the first 2 items, you should still be alright, because I embedded a kind of mini-course on complex analysis in the middle of the book.

Following an introductory chapter, this book is divided into 6 parts. The first 5 parts have to do with different aspects of the theory of surfaces. The 6th part is a collection of several topics, loosely related to the rest of the book, which I included because I really like them. Here is an outline of the book.

Part 1: Surfaces and Topology. In this part, we define such concepts as *surface*, *Euler characteristic*, *fundamental group*, *deck group*, and *covering space*. We prove that the deck group of a surface and its fundamental group are isomorphic. We also prove, under some conditions, that a space has a universal cover.

Part 2: Surfaces and Geometry. The first 3 chapters in this part introduce Euclidean, spherical, and hyperbolic geometry, respectively. (In the Euclidean case, which is so well known, we concentrate on nontrivial theorems.) Following this, we discuss the notion of a Riemannian metric on a surface. In the final chapter, we discuss hyperbolic surfaces, as special examples of Riemannian manifolds.

Part 3: Surfaces and Complex Analysis. In this part, we give a rapid primer on the main points taught in the first semester of complex analysis. Following this, we introduce the concept of a Riemann surface and prove some results about complex analytic maps between Riemann surfaces.

Part 4: Flat Surfaces. In this part, we define what is meant by a *flat cone surface*. As a special case, we consider the notion of a *translation surface*. We show how the "affine symmetry group" of a translation surface, known as the Veech group, leads right back to complex analysis and hyperbolic geometry. We end this part with an application to polygonal billiards.

Part 5: The Totality of Surfaces. In this part, we discuss the basic objects one considers when studying the totality of all flat or hyperolic surfaces, namely *moduli space*, *Teichmüller space*, and the *mapping class group*. As a warmup for the flat-surface case, we discuss continued fractions and the modular group in detail.

Part 6: Dessert. In this part, we prove 3 classic results in geometry. The Banach – Tarski Theorem says that—assuming the Axiom of Choice—you can cut up a ball of radius 1 into finitely many pieces and rearrange those pieces into a (solid) ball of radius 2. Dehn's Theorem says that you cannot cut up a cube with planar cuts and rearrange it into a regular tetrahedron. The Cauchy Rigidity Theorem says roughly that you cannot flex a convex polyhedron.

Contents

1 Book Overview

1.1 Behold, the Torus!

The Euclidean plane, denoted $\boldsymbol{R}^2$, is probably the simplest of all surfaces. $\boldsymbol{R}^2$ consists of all points $X = (x_1, x_2)$ where x_1 and x_2 are real numbers. One may similarly define Euclidean 3-space $\boldsymbol{R}^3$. Even though the Euclidean plane is very simple, it has the complicating feature that you cannot really see it all at once: it is unbounded.

Perhaps the next simplest surface is the unit sphere. Anyone who has played ball or blown a bubble knows what a sphere is. One way to define the sphere mathematically is to say that it is the solution set, in $\boldsymbol{R}^3$, to the equation

$$x_1^2 + x_2^2 + x_3^2 = 1.$$

The sphere is bounded and one can, so to speak, comprehend it all at once. However, one complicating feature of the sphere is that it is fundamentally curved. Also, its most basic definition involves a higher-dimensional space, namely $\boldsymbol{R}^3$.

The *square torus* is a kind of compromise between the plane and the sphere. It is a surface that is bounded like the sphere yet flat like the plane. The square torus is obtained by gluing together the opposite sides of a square, in the manner shown in Figure 1.1.

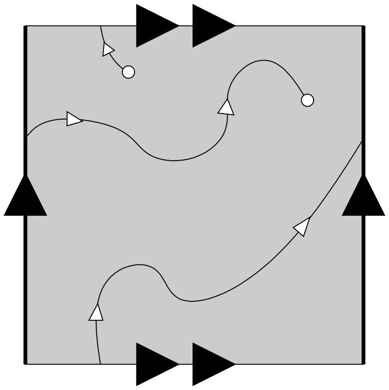

Figure 1.1. The square torus

We will not yet say exactly what we mean by *gluing*, but we say intuitively that a 2-dimensional being–call it a bug–that wanders off the top of the square would reappear magically on the bottom, in the same horizontal position. Likewise, a bug that wanders off the right side of the square would magically reappear on the left side at the same vertical position. We have drawn a continuous curve on the flat torus to indicate what we are talking about. In §3.1 we give a formal treatment of the gluing construction.

At first it appears that the square torus has an edge to it, but this is an illusion. Certainly, points in the middle of the square look just look like the Euclidean plane. A myopic bug sitting near the center of the square would not be able to tell he was living in the torus.

Consider what the bug sees if he sits on one of the horizontal edges. First of all, the bug actually sits simultaneously on *both* horizontal edges, because these edges are glued together. Looking "downward", the bug sees a little half-disk. Looking "upward", the bug sees another little half-disk. These 2 half-disks are glued together and make one full Euclidean disk. So, the bug would again think that he was sitting in the middle of the Euclidean plane. The same argument goes for any point on any of the edges.

The only tricky points are the corners. What if the bug sits at one of the corners of the squares? Note first of all that the bug actually sits simultaneously at all 4 corners, because these corners are all glued together.

As the bug looks in various directions, he sees 4 little quarter-disks that glue together to form a single disk. Even at the corner(s), the bug thinks that he is living in the Euclidean plane.

Modulo a ton of details, we have shown that the square torus has no edges at all. At every point it "looks locally" like the Euclidean plane. In particular, it is perfectly flat at every point. At the same time, the square torus is bounded, like the sphere.

The torus is such a great example that it demands a careful and rigorous treatment. The first question that comes to mind is *What do we mean by a surface?* We will explain this in §2. Roughly speaking, a surface a space that "looks like" the Euclidean plane in the vicinity of each point. We do not want to make the definition of "looks like" too restrictive. For instance, a little patch on the sphere does not look exactly like the Euclidean plane, but we still want the sphere to count as a surface. We will make the definition of "looks like" flexible enough so that the sphere and lots of other examples all count.

1.2 Gluing Polygons

In §3 we give many examples of surfaces and their higher-dimensional analogues, manifolds. One of the main tools we use is the gluing construction. The square torus construction above is the starting point for a whole zoo of related constructions.

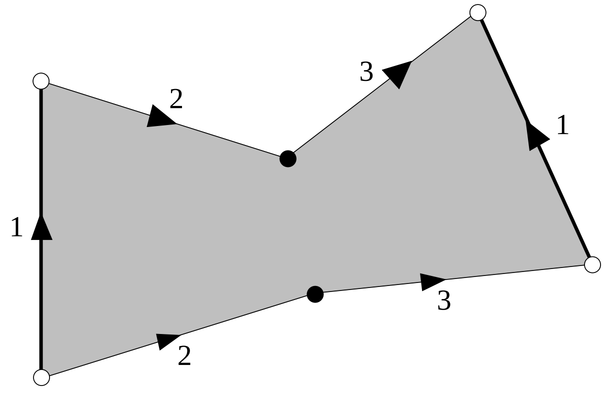

Figure 1.2. Another torus

Imagine, for example, that we take the hexagon shown in Figure 1.2 and glue the sides in the pattern shown. What we mean is that the 2 edges labelled 1 are glued together, according to the direction given by the arrows, and likewise for the edges labelled 2 and 3. We can think of Figure 1.2 as a distorted version of Figure 1.1. The hexagon has a left side, a right side, a top, and a bottom. The top is made from 2 sides and the bottom is made from 2 sides. The left and right sides are glued together and the top is glued to the bottom. The resulting surface retains some of the features of the flat torus: a bug walking around on it would not detect an edge. On the other hand, consider what happens when the bug sits at the point of the surface corresponding to the white dots. Spinning around, the bug would notice that he turns less than 360 degrees before returning to his original position. What is going on is that the sum of the interior angles at the white dots is less than 360 degrees. Similarly, the bug would have to spin around by more than 360 degrees before returning to his original position were he to sit at the point of the surface corresponding to the black points. So, in general, the bug would not really feel like he was living in the Euclidean plane. Our general definition of surfaces and gluing will be such that the example we gave still counts as a surface.

Figure 1.3 shows an example based on the regular octagon, in which the opposite sides of the octagon are glued together.

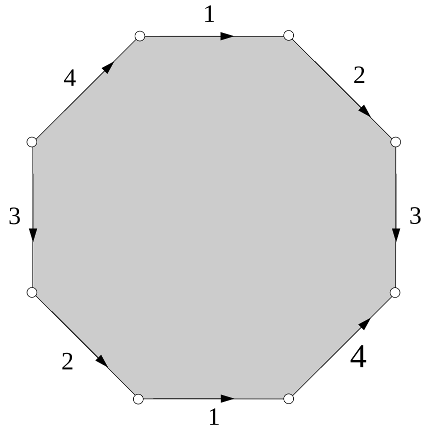

Figure 1.3. Gluing an octagon together

This example is similar to the square torus, except that this time 8 corners, rather than 4, are glued together. A myopic bug sitting anywhere on the surface except at the point corresponding the 8 corners might think that he was sitting in the Euclidean plane. However, at the special point, the bug would have to turn around 720 degrees (or 6π radians) before returning to his original position. We will analyze this surface in great detail. One can view it as the next one in the sequence that starts out sphere, torus, ... At least for this introductory section, we will call it the *octagon surface*. (It is commonly called the *genus* 2 *torus*.) We can construct similar examples based on regular $2n$-gons, for each $n = 5, 6, 7 \ldots$.

1.3 Drawing on a Surface

Once we have defined surfaces and given some examples, we want to work with them to discover their properties. One natural thing we can do is divide a surface up into smaller pieces and then count them. Figure 1.4 shows 2 different subdivisions of the square torus into polygons. We have left off the arrows in the diagram, but we mean for the left/right and top/bottom sides to be glued together.

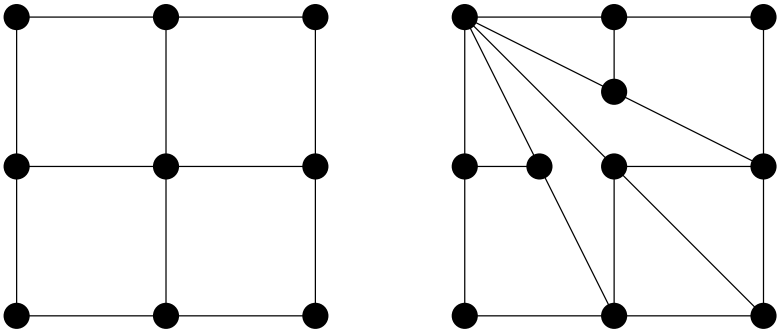

Figure 1.4. Dividing the torus into faces

In the first subdivision, there are 4 faces, 8 edges, and 4 vertices. It first appears that there are more edges, but the edges around the boundary are glued together in pairs. So each edge on the boundary only counts for half an edge. A similar thing happens with the vertices. We make the count

$$\text{faces} - \text{edges} + \text{vertices} = 4 - 8 + 4 = 0.$$

In the second example, we get the count

$$\text{faces} - \text{edges} + \text{vertices} = 8 - 14 + 6 = 0.$$

The same result holds for practically any subdivision of the square torus into polygons. This result is known as the *Euler formula for the torus*. We discuss this formula in more detail in §3.4.

You can probably imagine that you would get the same result for a torus based on a rectangle rather than a square. Likewise, we get the same result for the surface based on the hexagon gluing in Figure 1.2. All these surfaces have an *Euler characteristic* of 0.

Things turn out differently for the sphere. For instance, thinking of the sphere as a puffed-out cube, we get the count

$$\text{faces} - \text{edges} + \text{vertices} = 6 - 12 + 8 = 2.$$

Thinking of the sphere as a puffed-out tetrahedron, we get the count

$$\text{faces} - \text{edges} + \text{vertices} = 4 - 6 + 4 = 2.$$

Thinking of the sphere as a puffed-out icosahedron, we get the count

$$\text{faces} - \text{edges} + \text{vertices} = 20 - 30 + 12 = 2.$$

The Euler formula for the sphere says that the result of this count is always 2, under very mild restrictions. You can probably see that we would get the same result for any of the "sphere-like" surfaces mentioned above.

Were we to make the count for any reasonable subdivision of the octagon surface, we would get an Euler characteristic of -2. Can you guess the Euler characteristic, as a function of n, for the surface obtained by gluing together the opposite sides of a regular $2n$-gon?

Another thing we can do on a surface is draw loops—meaning closed curves—and study how they move around. The left side of Figure 1.5 shows 3 different loops on the square torus.

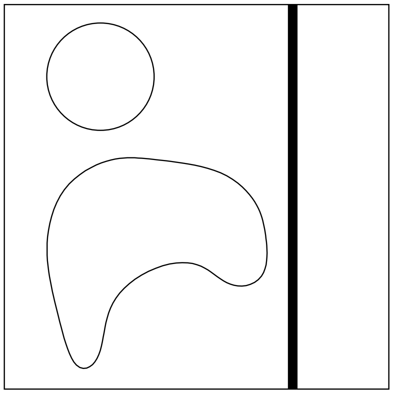

Figure 1.5. Loops on the torus

One of the loops, the one represented by the thick vertical line, is different from the others. Imagine that these loops are made from rubber bands, and are allowed to compress in a continuous way. The first 2 loops can shrink continuously to points, whereas the third loop is "stuck". It can't make itself any shorter no matter how it moves. Such a loop is commonly called *essential*. There are many essential loops on the torus. The right side of Figure 1.5 shows another essential loop. In contrast, the sphere has no essential loops at all.

We will see in §4 that there is an algebraic object we can associate to a surface (and many other kinds of spaces) called the *fundamental group*. The fundamental group organizes all the different ways of drawing loops on the surface into one basic structure. The nice thing about the fundamental group is that it links the theory of surfaces to algebra, especially group theory. Beautifully, it turns out that 2 (compact) surfaces have the same Euler characteristic if and only if they have the same fundamental group. The Euler characteristic and the fundamental group are 2 entry points into the vast subject of algebraic topology.

For the most part, studying algebraic topology is beyond the scope of this book, but we will study the fundamental group and related constructions, in great detail. After defining the fundamental group in §4, we will compute a number of examples in §5.

1.4 Covering Spaces

There is a nice way to unwrap the essential loops on a torus. The idea is that we remember that the square torus is made from a square, which we think of as the unit square with vertices $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$. We draw a line segment in the plane that starts out at the same point as the loop and has the same length. We think of this path starting at the point $(0,0)$. Figure 1.6 shows an example. In this example, the unwrapped path joins $(0,0)$ to $(3,2)$.

The process can be reversed. Starting with a line segment that joins $(0,0)$ to (m,n), a point with integer coordinates, we can wrap the segment around the torus so that it makes an essential loop. In fact, the essential loops that start at $(0,0)$ are, in the appropriate sense, in one-to-one correspondence with the points of $\boldsymbol{Z}^2$, the integer grid in the plane. The basic result is that any 2 essential loops L_1 and L_2, corresponding to points (m_1,n_1) and (m_2,n_2), can be continuously moved, one into the other, if and only if $(m_1,n_1)=(m_2,n_2)$.

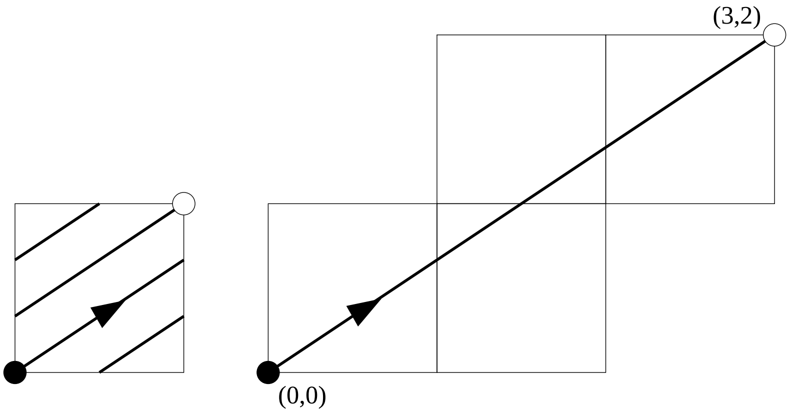

Figure 1.6: Unwrapping a loop on the torus

As we will explain in §6 and §7, this unwrapping construction can be done for any surface. In the case of the torus, we see that the (equivalence classes of) essential simple loops are in exact correspondence with the points of the integer grid in the plane. One might wonder if a similarly nice picture exists in general. The answer is "yes", and in fact the picture becomes more interesting when we consider surfaces, such as the octagon surface. However, in order to "see" the picture in these cases, you have to draw it in the possibly unfamiliar world of hyperbolic geometry. The idea is that hyperbolic geometry does for the octagon surface (and most other surfaces as well) what Euclidean geometry does for the square torus and what spherical geometry does for the sphere.

We will discuss Euclidean, spherical, and hyperbolic geometry in §8, §9, and §10 respectively. Our main goal is to understand how these geometries interact with surfaces, but we will also take time out to prove some classical geometric theorems, such as Pick's Theorem (a relative of the Euler formula) and the angle-sum formula for hyperbolic and spherical triangles.

The Euclidean, spherical, and hyperbolic geometries are the 3 most symmetrical examples of 2-dimensional *Riemannian geometries*. To put the 3 special geometries into a general context, we will discuss Riemannian geometry in §11.

1.5 Hyperbolic Geometry and the Octagon

Now let us return to the question of unwrapping essential loops on the octagon surface. The octagon surface looks a bit less natural than the square torus, thanks to the special point. However, it turns out that the octagon surface "wears" hyperbolic geometry very much in the same way that the square torus "wears" Euclidean geometry.

We already mentioned that we will study hyperbolic geometry in detail in §10. Here we just give the barest of sketches, in order to give you a taste of the beauty that lies in this direction. One of the many models for the hyperbolic plane is the open unit disk. There is a way to measure distances in the open unit disk so that the shortest paths between points are circular arcs that meet the boundary at right angles. These shortest paths are known as *geodesics*. The left-hand side of Figure 1.7 shows some of the geodesics in the hyperbolic plane. The boundary of the unit disk is not part of the hyperbolic plane and the lengths of these geodesics are all infinite. A bug living in the hyperbolic plane would see it as unbounded in all directions.

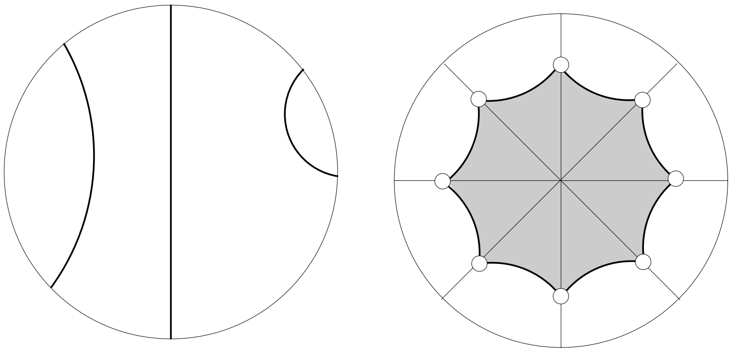

Figure 1.7. Gluing the octagon together

The hyperbolic plane shares many features with the Euclidean plane. There is a unique geodesic joining any 2 distinct points, and any 2 distinct geodesics meet in at most one point. Furthermore, the hyperbolic plane is totally symmetric, in the sense that every point and every direction looks exactly the same. A bug living in an otherwise empty hyperbolic plane would not be able to tell where he was.

On the other hand, the hyperbolic plane and the Euclidean plane have some important differences. For instance, the sum of the angles of a hyperbolic triangle, a shape bounded by 3 geodesic segments, is always less than 180 degrees, or π radians. (When we discuss angles in radians, we will often leave off the word "radians".) Similarly, the individual interior angles of a regular octagon can take on any value less than $3\pi/8$, which is the value in the Euclidean case. The right hand side of Figure 1.4 shows a regular hyperbolic octagon. We decrease the interior angles by making the octagon larger and we increase the interior angles by making the octagon smaller.

In particular, we can adjust the size of the regular octagon so that the interior angles are exactly $\pi/8$. We can then cut the resulting octagon out of the hyperbolic plane and glue the sides together just as in Figure 1.3. From the hyperbolic geometry point of view, the resulting surface would be completely seamless: a myopic bug living on the surface could not tell that he was not living in the hyperbolic plane. With the chosen interior angles, the 8 corners fit together like 8 slices in a pizza to make a perfect hyperbolic disk. We will consider this construction in detail in §12.

A similar construction can be made for the surfaces obtained by gluing together the opposite sides of a regular $2n$-gon, for each $n = 5, 6, 7 \ldots$. All these surfaces "wear" hyperbolic geometry in a seamless way, just like the square torus "wears" Euclidean geometry.

Now, we can tile the Euclidean plane by copies of the unit square. The vertices of this tiling are precisely the integer grid points. In the same way, we can move our hyperbolic octagon around the hyperbolic plane and tile the hyperbolic plane with copies of it. When drawn in the disk model, the picture looks like the drawings in M. C. Escher's *Circle Woodcut* series. To our Euclidean eyes, the octagons appear to get smaller as they move out toward the boundary of the disk. However, in the hyperbolic world, the various octagons all have the same size.

The vertices of this tiling are a kind of hyperbolic geometry version of the integer grid. These points are in one-to-one correspondence with the equivalence classes of essential loops on the octagon surface. The same kind of thing works for the surfaces corresponding to the $(2n)$-gons for $n = 4, 5, 6 \ldots$. In fact, such a construction works for all surfaces that have negative Euler characteristic: one always gets a grid of points in the hyperbolic plane that names the different essential loops on the surface. We will explore this in detail in §12.

1.6 Complex Analysis and Riemann Surfaces

It turns out that there is a single kind of geometry which unifies Euclidean, spherical, and hyperbolic geometry. This geometry, called *Möbius* or *conformal* geometry, takes place in the *Riemann sphere*. The Riemann sphere is the set $\boldsymbol{C} \cup \infty$, where $\boldsymbol{C}$ is the complex plane and ∞ is an extra point that is added. For starters,

- The Euclidean plane is identified with $\boldsymbol{C}$.
- The hyperbolic plane is identified with the open disk $\{z \in \boldsymbol{C} | \ \|z\| < 1\}$.
- The sphere is identified with all of $\boldsymbol{C} \cup \infty$, via *stereographic projection*

  $$(x_1, x_2, x_3) \to \Big(\frac{x_1}{1-x_3}\Big) + \Big(\frac{x_2}{1-x_3}\Big)i, \qquad (0,0,1) \to \infty.$$

  See §9.5 for details on stereograpic projection.

Once these identifications are made, the symmetries of the relevant objects are all given by maps of the form

$$z \to \frac{az+b}{cz+d}, \qquad a,b,c,d \in \boldsymbol{C}, \qquad ad-bc=1. \tag{1}$$

We will discuss these maps in more detail in §10.1. The point ∞ is added so that when the expression in equation (1) looks like "something over 0", we define it to be ∞. Various conditions are placed on the coefficients a,b,c,d to guarantee that the relevant set–e.g., the unit disk–is preserved by the map.

These kinds of transformations are called *linear fractional*, or *Möbius*, transformations. The Möbius transformations are prototypical examples of *complex analytic* functions. These are continuous maps from $\boldsymbol{C}$ to $\boldsymbol{C}$ which have the additional property that their matrix of partial derivatives, at each point, is a similarity–i.e., a rotation followed by a dilation. This constraint on the partial derivatives leads to a surprisingly rich family of functions and this is the subject of complex analysis. In §13, we will give a rapid overview of basic complex analysis, with a view towards its application to surfaces. In §14 and §15 we will discuss some special complex analytic functions in detail.

Going back to our polygon gluing construction, we can view surfaces as being made out of pieces of $\boldsymbol{C}$ that have been glued together. This point of view leads to the notion of a *Riemann surface*, as we explain in §16. One can think of a Riemann surface as a surface that "wears" $\boldsymbol{C}$ in the same seamless way that the square torus "wears" Euclidean geometry or the octagon surface "wears" hyperbolic geometry. Once we have the notion of a Riemann surface, we can "do complex analysis on it" in much the same way that one can do complex analysis in $\boldsymbol{C}$ or in $\boldsymbol{C} \cup \infty$.

The complex analysis point of view on a surface at first seems rather remote from the geometric point of view discussed above, but in fact they are quite similar. The close connection comes from the fact that the Möbius transformations play a distinguished role amongst the complex analytic functions. One example of this is the following result, known as the Schwarz–Pick Theorem:

Theorem 1.1 *Let f be a complex analytic function from the unit disk to itself. If f is one-to-one and onto, then f is a Möbius transformation (and hence a hyperbolic isometry).*

Theorem 1.1 is part of a larger theorem, called the Poincaré Uniformization Theorem. The Uniformization Theorem gives a complete equivalence between the Euclidean/spherical/hyperbolic geometry points of view of surfaces and the Riemann surface point of view. The proof of this result is beyond the scope of our book, but in §16 we will at least explain the result and its ramifications.

1.7 Cone Surfaces and Translation Surfaces

We have mentioned several times that the octagon surface does not "wear" Euclidean geometry as well as the square torus does, and we have taken some pains to explain how one can profitably view the octagon surface with hyperbolic geometry eyes. However, in §17 we come full circle and consider the octagon surface and related surfaces from the Euclidean geometry point of view.

Suppose, as in Figure 1.2 above, we glue together the sides of a polygon in such a way that the sides in each pair of glued sides have the same length. The resulting surface has the property that it is locally indistinguishable from the Euclidean plane, except at finitely many points. At these finitely many points, a bug living in the surface would notice some problem related to spinning around, as we discussed above. These special points are *cone points*. A *Euclidean cone surface* is a surface that is flat except at finitely many cone points.

When we discussed the "torus-like" surface defined in connection with Figure 1.2, we mentioned the spinning-around problem a bug would face when sitting at the 2 special points. At one of the special points, the bug needs to spin more than 2π, say $2\pi + \delta_1$, before returning to his original position. At the other special point, the bug needs to spin less than 2π, say $2\pi - \delta_2$, before returning to his original position. The numbers δ_1 and $-\delta_2$ might be called the *angle error* at the special points.

The numbers δ_1 and δ_2 depend on the hexagon in question. As one can see by adding up the interior angles of a hexagon, we have $\delta_1 = \delta_2$. That is, the total angle error is 0. This result holds for any Euclidean cone surface with Euler characteristic 0. More generally, on a surface with Euler characteristic χ, the total angle error is $2\pi\chi$. This result, known as the *combinatorial Gauss–Bonnet Theorem* is one of the main results of §17.

Another topic in §17 is the application of Euclidean cone surfaces to polygonal billiards. It turns out that the contemplation of rolling a frictionless, infinitesimally small billiard ball around inside a polygonal shaped billard table, whose angles are all rational multiples of π, leads naturally to a certain Euclidean cone surface. One can profitably study this surface to get information about how billiards would work out in the polygon.

The Euclidean cone surfaces associated to polygonal billiards have a special structure. They are called *translation surfaces*. A translation surface is a Euclidean cone surface, all of whose angle errors are integer multiples of π. The square torus is the prototypical example of a translation surface, but it is a bit too simple of an example in this case. The octagon surface provides a better example. The octagon surface, considered from the Euclidean geometry perspective, is a translation surface. This surface has a single cone point, and the angle error there is 4π. Translation surfaces are nicer than general Euclidean cone surfaces for a variety of reasons. One reason is that, as it turns out, it is possible to speak about directions (such as due north) on a translation surface without any ambiguity. We will discuss these surfaces in detail in §18.

1.8 The Modular Group and the Veech Group

We have wandered away from hyperbolic geometry and complex analysis, but actually hyperbolic geometry and complex analysis are very closely related to the subject of translation surfaces. Once again, let us consider the square torus. A linear transformation of the form

$$T(x,y) = (ax+by, cx+dy), \qquad a,b,c,d \in \boldsymbol{Z}, \qquad ad-bc=1. \tag{2}$$

acts as transformation of the square torus, via the following 4-step process:

1. Start with a point p in the square torus.

2. Choose a point (x,y) such that p represents the collection of points glued to (x,y).

3. Subtract off integer coordinates of $T(x, y)$ until the result (x', y') lies in the unit square.

4. The image of the map is p', the point that names the collection of points glued to (x', y').

Any ambiguity in the process that takes us from p to p' is absorbed by the gluing process.

So, any integer 2×2 matrix with determinant 1 gives rise to a transformation of the square torus that, on small scales, is indistinguishable from a linear transformation. The set of all such maps forms a group known as *modular group*. The maps in equation (2) have the same form as the Möbius transformations discussed above. Interpreting the maps in Equation 2 as Möbius transformations instead of linear transformations, we can interpret the modular group as a group of symmetries of the hyperbolic plane.

The modular group is an object of great significance in mathematics, and we cannot resist exploring some of its properties that are not, strictly speaking, directly related to surfaces. For instance, in §19 we will discuss continued fractions and their connection to the modular group and hyperbolic geometry. In §22 we will see that the modular group is the main ingredient in the proof of the Banach–Tarski Theorem. The Banach–Tarski Theorem says in particular that, assuming the axiom of choice, one can cut the unit ball in $\boldsymbol{R}^3$ into finitely many pieces and rearrange these pieces so that they make a solid ball of radius 100000. Though this result seems a bit far removed from the theory of surfaces, it is quite beautiful and it shows how objects such as the modular group pop up all over mathematics.

Getting back to translation surfaces, we will see in §18 that one can associate to any translation surface a group of symmetries of the hyperbolic plane. This group is known as the *Veech group* of the translation surface. It often happens that the Veech group is trivial, or very small, but for many special examples the Veech group is large and beautiful. For instance, the Veech group associated to the regular octagon surface is closely related to a tiling of the hyperbolic plane by triangles having angles 0, 0, and $\pi/8$. One of the highlights of §18 is a discussion of (essentially) this example.

1.9 Moduli Space

The square torus is not the only translation surface without any cone points. In §20 we consider the family $\mathcal{M}$ unit area parallelogram, in the same pattern as in Figure 1.1. Essentially the same analysis we made in connection with Figure 1.1 can be made in connection with any surface in our family. All these surfaces are seamlessly flat at each point. A myopic bug on any of these surfaces would not be able to tell that he was not in the Euclidean plane.

On the other hand, these various surfaces are typically not the same geometrically. For instance, a surface made from a long thin rectangle obviously has diameter greater than the diameter of the square torus. Similarly, such a surface has a very short essential loop whereas all essential loops on the square torus have length at least 1. We can consider the family $\mathcal{M}$ as a space in its own right. Each point of $\mathcal{M}$ corresponds to a different flat torus. This space $\mathcal{M}$ is known as the *moduli space* of flat tori. We will discuss $\mathcal{M}$ and related objects in §20.

Amazingly, $\mathcal{M}$ turns out to be a surface in its own right, and (with the exception of 2 special points) this surface is modelled on hyperbolic geometry! Just to repeat: the space of all tori made by gluing together unit area parallelograms turns out to be a surface that naturally wears hyperbolic geometry (away from 2 special points). One of the special points in $\mathcal{M}$ corresponds to the square torus, and the other one corresponds to the surface obtained by gluing together the opposite sides of a rhombus made from 2 equilateral triangles. Referring to the discussion of covering spaces above, we can consider the grid in the hyperbolic plane associated to $\mathcal{M}$. It turns out that the modular group acts as a group of symmetries of this grid. So, when we consider the moduli space $\mathcal{M}$ of unit area flat tori, we get right back to the modular group.

We can play a similar game for the octagon surface. As we discussed above, we can create the octagon surface using a suitable chosen *regular* octagon. However, we can also glue together other hyperbolic octagons to produce a surface that "looks hyperbolic" at each point and has the same Euler characteristic. When we consider the totality of such surfaces, we arrive at a higher-dimensional generalization of $\mathcal{M}$, also called *moduli space*. This higher-dimensional space is not a surface, but it does share some features in common with a hyperbolic surface.

In §20 we also discuss *Teichmüller space*, the space that relates to the higher-dimensional version of $\mathcal{M}$ in the same way that the hyperbolic plane relates to $\mathcal{M}$. Teichmüller space shares some features with the hyperbolic plane, but is much more mysterious and somewhat less symmetric. We will discuss the group of symmetries of Teichmüller space, called the *Mapping* *class group*. The mapping class groups relate to the surfaces of negative Euler characteristic in the same way that the modular group relates to the square torus. We will further explore Teichmüller space in §21.

1.10 Dessert

There are a few topics in this book that I simply threw in because I like them. §22 has a proof of the Banach- Tarski Paradox. One nice thing about the proof is that it involves the modular group in an essential way. So, in a strange way, the Banach–Tarski Paradox has some connection to hyperbolic geometry.

§23 has a proof of Dehn's Dissection Theorem, which says that one cannot cut a cube into finitely many pieces, using planar cuts, and rearrange the result into a regular tetrahedron. This result serves as a kind of foil for the decomposition methods we use to prove the combinatorial Gauss–Bonnet Theorem and other results. Polyhedral decomposition is quite robust in 2 dimensions, but not in higher dimensions.

§24 has a proof of the Cauchy Rigidity Theorem. This result says that there at most one way to snap together a given collection of convex polygons to produce a convex polyhedron. The proof involves some spherical geometry and also the combinatorial Gauss–Bonnet Theorem.

2 Definition of a Surface

We discussed surfaces informally in the previous chapter, and now the time has come to give a formal definition of a surface. Here is the main definition.

Definition 2.1. A *surface* is a metric space X such that every point in X has a neighborhood which is homeomorphic to the plane.

Don't worry if you don't know what some of the words in the above definition mean. The point of this chapter is to explain what they mean. At the end of the chapter, we will say a few words about higher-dimensional surfaces, called manifolds.

2.1 A Word about Sets

A *set* is an undefined notion for us. Informally, a *set* is a collection of things, called *elements*. A book on set theory, such as [DEV], will tell you all about sets. You should be familiar with such sets as

- $\mathbf{Z}$, the integers.
- $\mathbf{N} = \{1, 2, 3, \ldots\}$, the natural numbers.
- $\boldsymbol{R}$, the real numbers.

A *map* between sets A and B is a rule, say f, which assigns to each element $a \in A$, an element $b = f(a) \in B$. This is usually written as $f : A \to B$. The map f is *one-to-one* if $f(a_1) = f(a_2)$ implies that $a_1 = a_2$. The map f is *onto* if the set $\{f(a)|\ a \in A\}$ equals B. The map f is a *bijection* if it is both one-to-one and onto. Two sets are *bijective* if there is some bijection between them. All the sets we consider will be bijective to either a finite set, or $\mathbf{N}$, or $\boldsymbol{R}$.

The product $A \times B$ of sets is the set of ordered pairs (a, b) with $a \in A$ and $b \in B$. In particular, $\boldsymbol{R}^2 = \boldsymbol{R} \times \boldsymbol{R}$ is the plane.

2.2 Metric Spaces

A *metric space* is a set X together with a map $d : X \times X \to \boldsymbol{R}$ satisfying the following properties:

- *Nondegeneracy.* $d(x,y) \geq 0$ for all x, y, with equality iff $x = y$.
- *Symmetry.* $d(x,y) = d(y,x)$ for all x, y.
- *Triangle Inequality* $d(x,z) \leq d(x,y) + d(z,y)$ for all x, y, z.

d is called a metric on X. Note that the same set can have many different metrics.

Here is the most boring example of a metric space. Given any set X define $d(x,y) = 0$ if $x = y$ and $d(x,y) = 1$ if $x \neq y$. This is called the *discrete metric* on X.

Exercise 1. Let $X = \boldsymbol{R}^2$, the plane. Define the *dot product*

$$V \cdot W = v_1 w_1 + v_2 w_2.$$

Here $V = (v_1, v_2)$ and $W = (w_1, w_2)$. Also define

$$\|V\| = \sqrt{V \cdot V}.$$

Finally, define $d(V,W) = \|V - W\|$. Prove that d is a metric on $\boldsymbol{R}^2$. The metric in this exercise is known as the *Euclidean metric* on $\boldsymbol{R}^2$, or else the *standard metric.*

If X is a metric space and $Y \subset X$ is a subset, then the metric on X automatically defines a metric on Y, by restriction. For instance, any subset of the plane automatically can be interpreted as a metric space, using the metric from Exercise 1.

Exercise 2. On $\mathbf{Z}$ define $d(m,n) = 2^{-k}$, where k is such that 2^k divides $|m-n|$ but 2^{k+1} does not. Also define $d(m,m) = 0$. For instance, $d(3,7) = 2^{-2} = 1/4$ because 2^2 divides 4 but $2^3 = 8$ does not. Prove that d is a metric on $\mathbf{Z}$. This metric is called the 2-adic metric. It is quite different from the usual metric on the integers.

2.3 Open and Closed Sets

Let X be a metric space with metric d. An *open ball* in X is a subset of the form

$$\{x \mid d(x,c) < r\}.$$

Here c is the *center* of the ball and r is the radius. Say that a subset $U \subset X$ is *open* if for every point $x \in U$ there is some open ball B_x such that $x \in B_x$ and $B_x \subset U$. Note that open balls are open sets.

Exercise 3. Prove that the intersubsection of two open sets is open. Prove also that the arbitrary union of open sets is open.

Here is some vocabulary, which will be familiar to you if you have had a real analysis class:

- The notation $X - A$ means the complement of A in X, namely the set of points in X which are not in A.
- Given a point $x \in X$, a *neighborhood* of x is any open subset $U \subset X$ such that $x \in U$. For instance, the ball of radius r about x is a perfectly good neighborhood of x.
- The *interior* of a set $A \subset X$ is the union of all open subsets of A. By Exercise 3, the interior of a set is open. Sometimes the interior of A is denoted as A^o. Put another way A^o is the largest open set contained in A.
- A set $C \subset X$ is *closed* if $X - C$ is open.
- The *closure* of a set A is the set
$$\overline{A} = X - (X - A)^o.$$
Put another way, $\overline{A}$ is the smallest closed set which contains A.
- The *boundary* of A is the set
$$\partial A = \overline{A} - A^o.$$
- A set $A \subset X$ is *dense* if $\overline{A} = X$. For instance, the set of rational numbers is dense in the set of real numbers.

2.4 Continuous Maps

A *map* between metric spaces is just a map in the set theoretic sense. There are two equivalent definitions of continuity for maps between metric spaces. The first one is much cleaner but the second one is probably more familiar.

Definition 2.2. The map $f : X \to Y$ is *continuous* if it has the following property: For any open $V \subset Y$ the set

$$U = f^{-1}(V) := \{x| \ f(x) \in V\}$$

is an open set of X.

Definition 2.3. First, f is continuous at $x \in X$ if, for any $\epsilon > 0$, there is some $\delta > 0$ such that $d_X(x, x') < \delta$ implies that $d_Y(f(x), f(x')) < \epsilon$. Here d_X is the metric on X and d_Y is the metric on Y. Then f is continuous on X if f is continuous at each point of X.

Exercise 4. Show that the two definitions of continuity coincide.

Now let X, Y, Z all be metric spaces. Let $f : X \to Y$ be a map, and let $g : Y \to Z$ be map. The composition $h = g \circ f$ is defined as $h(x) = g(f(x))$. So h is a map from X to Z.

Lemma 2.1 *The composition of continuous maps is continuous.*

Proof: Definition 2.2 works much better for this. Let W be an open subset of Z. Our goal is to show that $h^{-1}(W)$ is open in X. Note that $h^{-1}(W) = f^{-1}(V)$, where $V = g^{-1}(W)$. Since g is continuous, V is open. Since f is continuous and V is open, U is open. This works for any choice of open W, so we are done. ♠

Exercise 5. Give an example of metric spaces X and Y, and $f : X \to Y$ such that

- f is a bijection.
- f is continuous.
- f^{-1} (the inverse map) is not continuous.

This is a classic problem.

2.5 Homeomorphisms

Let X and Y be two metric spaces. A map $h : X \to Y$ is a *homeomorphism* if h is a bijection and both h and h^{-1} are continuous. Compare Exercise 5. The spaces X and Y are said to be *homeomorphic* if there is some homeomorphism from X to Y. Intuitively, two sets are homeomorphic if one can be "warped" into the other one. Often we do not care exactly which metric we are using, but we just bring in the metric to be able to talk about things like continuity and open sets. Another way to "throw out the metric" is to introduce the notion of a *topological space*. In some ways topological spaces are easier to work with than metric spaces and more flexible, but they are more abstract. If you are interested in this, check out a book on point-set topology, such as [**MUN**].

Even though sets might look very different to the eye, they might be homeomorphic. The next exercise gives some examples of this.

Exercise 6. Prove that the following subsets of the plane (with the standard metric) are all homeomorphic to each other:

- An open ball.
- The interior of a (filled-in) triangle.
- The plane itself.

Exercise 7. We can give $\boldsymbol{R}$ the standard metric $d(x, y) = |x - y|$. Prove that $\boldsymbol{R}$ is not homeomorphic to $\boldsymbol{R}^2$, with its standard metric.

Exercise 8 (Challenge). You can imitate the construction in Exercise 1 to put a metric on $\boldsymbol{R}^3$, 3-dimensional space. Prove that $\boldsymbol{R}^2$ is not homeomorphic to $\boldsymbol{R}^3$. As it turns out $\boldsymbol{R}^m$ and $\boldsymbol{R}^n$ are homeomorphic if and only if $m = n$. When you try to prove something like this, you start getting into algebraic topology.

2.6 Compactness

We will sometimes use the notion of *compactness*. Say that an *open covering* of a metric space X is a collection $\{U_\alpha\}$ of open sets in X whose union is X. Say that a *subcovering* of a given covering is a subcollection that still covers X. Say that a *finite subcover* is a subcover that only has finitely many elements in it.

Definition 2.4. A metric space X is *compact* if every covering of X has a finite subcover.

The notion of compactness is easier to understand for subsets of Euclidean space. When X is a subset of Euclidean space, X is compact if and only if X is closed and contained in some ball. This result is known as the *Heine–Borel Theorem*.

The original definition of compactness is perfectly adapted to the notion of continuous maps. Suppose that X and Y are homeomorphic. Then X is compact if and only if Y is compact. Here we prove one result which indicates the power of the definition of compactness.

Lemma 2.2 *Suppose that $f : X \to \boldsymbol{R}$ is a continuous function. If X is compact, then f is bounded.*

Proof: Let $U_n = f^{-1}(-n, n)$. Since f is continuous, the set U_n is open. Evidently, the collection $\{U_n\}$ covers X. Since X is compact, there is some finite list of these sets which also covers X. Letting U_N be the largest of these finitely many sets, we see that $|f| \leq N$ on X. ♠

2.7 Surfaces

Now let's go back to Definition 2.1. Let X be a surface. This means, first of all, that X is a metric space. So, it makes sense to talk about open and closed sets on X and also continuous functions from X to other metric spaces. What makes X a surface is that each point $x \in X$ has an open neighborhood U such that U is homeomorphic to $\boldsymbol{R}^2$. You should picture U as a little open disk drawn around x. So X has the property that, around every point, it "looks like" the plane. This is how we make sense of the discussion at the end of §1.1.

Exercise 9. The unit sphere S^2 in $\boldsymbol{R}^3$ is the set $\{(x, y, z)|\ x^2 + y^2 + z^2 = 1\}$. This set inherits a metric from $\boldsymbol{R}^3$. Prove that S is a surface, according to our definition. So, for each point $x \in S$ you need to find an open subset $U_x \subset S$ and also a map $f_x : U_x \to \mathbf{R^2}$ which is a homeomorphism. (*Hint*:

Try to use symmetry to reduce the problem to showing that just one point in S^2 has the desired neighborhood.)

Exercise 10. Consider the following subset of $\boldsymbol{R}^4$:

$$T^2 = \{(x, y, z, w) |\ x^2 + y^2 = 1; \quad z^2 + w^2 = 1\}.$$

This set inherits a metric from $\boldsymbol{R}^4$. You might recognize T^2 as the product of two circles. Prove that T^2 is a surface. This surface is known as a *torus*. (*Hint*: Again, try to use symmetry.) Once we make sense of the gluing construction, we will see that T^2 is homeomorphic to the square torus discussed in the previous chapter.

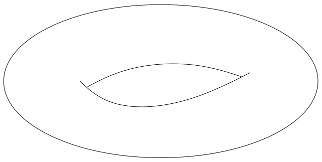

Figure 2.1: A torus

In the coming chapters, we will construct many more examples of surfaces besides the ones in Exercises 9 and 10.

2.8 Manifolds

A manifold is essentially a higher-dimensional surface. Though this book is about surfaces, I am including a subsection about manifolds in case you are curious about them. If you just want to learn about surfaces, you can safely skip this subsection.

Definition 2.5: An *n-dimensional manifold* is a metric space, such that every point has a neighborhood which is homeomorphic to $\boldsymbol{R}^n$.

Technical Comment. This definition of a manifold is slightly nonstandard. The usual definition replaces *metric space* with *Hausdorff topological space*. However, in most cases the metric space definition singles out the same objects as manifolds. The reason we are using the metric space definition is that it is more concrete.

I will give a nice example of a manifold at the end of this section, but first I will introduce a general tool for producing manifolds. The tool is the *Implicit Function Theorem*, a classic result from multivariable calculus. The full Implicit Function Theorem and its proof can be found in practically any book on advanced calculus; e.g., see [SPI]. We will prove a special case below, a case that is fairly easy to prove yet still produces nice examples.

Let $f : \boldsymbol{R}^{n+1} \to \boldsymbol{R}$ be a continuous function. Assume also that the partial derivatives of f exist and are continuous functions. This means that the gradient ∇f exists and is continuous. Say that 0 is a *regular value* for f if it never happens that both $f(x_1, \ldots, x_{n+1}) = 0$ and $\nabla f(x_1, \ldots, x_{n+1}) = 0$ at the same point.

Theorem 2.3 *If 0 is a regular value for f, then $f^{-1}(0)$ is an n dimensional manifold.*

Proof: Let $S = f^{-1}(0)$. First of all, S is a metric space: The distance between any two points in S is defined to be their Euclidean distance in $\boldsymbol{R}^n$. It remains to check that every point in S has a neighborhood that is homeomorphic to $\boldsymbol{R}^n$.

Let $p = (x_1, \ldots, x_{n+1}) \in S$ be an arbitrary point. We know that $\nabla f(p)$ is nonzero. If we rotate and scale space and replace f by a constant multiple of f, we do not change S at all. So, without loss of generality, we can assume that

$$p = (0, \ldots, 0); \qquad \nabla f(p) = (0, \ldots, 0, 1).$$

Let $P = \boldsymbol{R}^n \times \{0\}$. We think of $(0, \ldots, 0, 1)$ as the vertical direction and P as the horizontal directions. See Figure 2.2 below.

Let Q denote the open cube of diameter ϵ centered at $(0, \ldots, 0)$. We call a line segment *special* if it has one endpoint on the bottom face of Q and one endpoint on the top face of Q. Since ∇f varies continuously, we can choose ϵ small enough so that f increases along any special segment, if we move along it from the bottom to the top. Let $U = Q \cap S$. Then U is an open neighborhood of p in S. It suffices to show that U is homeomorphic to an open cube in $\boldsymbol{R}^n$, since an open cube in $\boldsymbol{R}^n$ is homeomorphic to $\boldsymbol{R}^n$ itself.

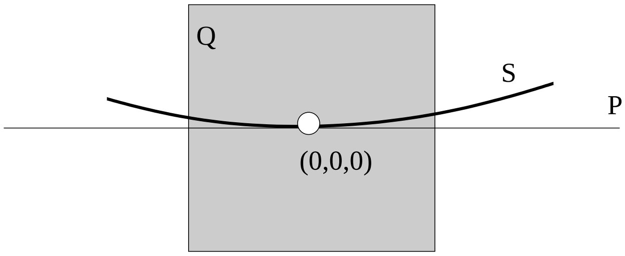

Figure 2.2. Putting a cube around S

Now, $Q \cap P$ is an open cube, and the map $h(x_1, \ldots, x_{n+1}) = (x_1, \ldots, x_n, 0)$ ss a map from U to $Q \cap P$. We just have to show that h is a homeomorphism. Here are the main points.

- h is a distance decreasing map so (using the $\epsilon - \delta$ definition of continuity) h is continuous.
- Each vertical line intersects S at most once, because f increases as we move upward along a vertical line. Hence, h is one-to-one.
- We can connect any point on the top face of Q to $(0, \ldots, 0)$ by "half" of a special segment. Since $f(0, \ldots, 0) = 0$, and f increases along special segments, f is positive on the top face of Q. Similarly, f is negative on the bottom face of Q. Since f increases along vertical segments, we have $f = 0$ somewhere on each vertical segment, by the intermediate value theorem. So, every vertical segment intersects S. Hence h is onto.
- Suppose that X_1 and X_2 are two points in $Q \cap P$ that are very close together. Consider the last coordinates z_1 and z_2 of $h^{-1}(X_1)$ and $h^{-1}(X_2)$, respectively. If z_1 and z_2 are too far apart, then we can join the points (X_1, z_1) and (X_2, z_2) by part of a special segment. Since both these points lie in S, we have a contradiction. This shows, a bit informally, that h^{-1} is continuous.

We have succeeded in showing that an arbitrary point of S has a neighborhood which is homeomorphic to $\boldsymbol{R}^n$. ♠

Now we give a nice example of a 3-dimensional manifold. You can think of the set of 2×2 (real valued) matrices as a copy of $\boldsymbol{R}^4$. There is a nice map from this space into $\boldsymbol{R}$, namely the determinant (minus 1):

$$f\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = ad - bc - 1.$$

Exercise 11. Show that 0 is a regular value for f.

In the above example, $f^{-1}(0)$ is usually denoted by $SL_2(\boldsymbol{R})$. Thus $SL_2(\boldsymbol{R})$ is the set of unit determinant real 2×2 matrices. By Theorem 2.3, the space $SL_2(\boldsymbol{R})$ is a 3-dimensional manifold. A similar argument shows that $SL_n(\boldsymbol{R})$, the set of unit determinant $n \times n$ matrices, is a manifold of dimension $n^2 - 1$. The space $SL_n(\boldsymbol{R})$ is an example of a manifold which is also, and in a compatible way, a group. Such objects are called *Lie groups*. The book [**CHE**] is a classic reference on this subject; see also [**TAP**].

3 The Gluing Construction

The purpose of this chapter is to explain the gluing construction discussed informally in Chapter 1. This construction is usually done for topological spaces, but it can be done for metric spaces as long as we are a bit careful. The advantage to using topological spaces is that the construction always works. The disadvantage to using topological spaces is that it takes a long time to figure out what the construction actually means. For metric spaces, things don't always work out, but whatever happens is more understandable. Also, for our purposes, things always work out.

3.1 Gluing Spaces Together

Before we start, we need to recall the notion of the *infimum* from real analysis. Let $S \subset \boldsymbol{R}$ be a set consisting entirely of non negative numbers. Then $x = \inf S$ denotes the smallest member of the closure of S. Such a number always exists and is unique. The existence (and uniqueness) of the infimum is known as the *completeness axiom* for the reals.

Let X be a set and let $\delta : X \times X \to \boldsymbol{R}$ be a map which satisfies the equation $\delta(x,y) = \delta(y,x) \geq 0$. Note that δ need not satisfy the triangle inequality. The purpose of this section is to show how to replace δ by a new function which sometimes remembers some of the structure of δ and yet satisfies the triangle inequality.

Let $x, y \in X$ be two points. Say that a *chain* from x to y is a finite sequence of points $x = x_0, x_1, \ldots, x_n = y$. Let us call this chain C. We define

$$\delta(C) = \delta(x_0,x_1) + \delta(x_1,x_2) + \ldots + \delta(x_{n-1},x_n).$$

Certainly, $\delta(C) \geq 0$ as long as $x \neq y$. Next, we define

$$d(x,y) = \inf_C \delta(C).$$

The infimum is taken over the set of all possible values $\delta(C)$, where C is a chain from x to y.

This probably looks like an insane definition, but we will try to make it intuitive. Think of $\delta(x,y)$ as the cost of flying from city x to city y—let's say from Providence to Tahiti. Now, you're really desperate to get to Tahiti, and have tons of free time but little money. So, you look on the Internet and try to find all possible flights. You are willing to take any conceivable chain of connecting flights, as long as you start in Providence and end in Tahiti. After searching through all the possiblities, you select the most economical flight. This is $d(x, y)$. The difference between this scenario and the idealized one we're talking about is that X could be an infinite metric space. So, there could be infinitely many chains, and you need to take the infimum rather than just a minimum (which may not exist.) The function d is sometimes called the *pathification* of δ.

Exercise 1. Show that d satisfies the following axioms:

- $d(x, y) \geq 0$.
- $d(x, y) = d(y, x)$.
- $d(x, y) \leq d(x, z) + d(z, y)$.

So it looks like d is a metric. However, note that we are leaving off the part that would say $d(x, y) = 0$ iff $x = y$. In fact give an example of a δ on $X = \boldsymbol{R}^2$ which satisfies the first two axioms for a metric, whose pathification is the zero map.

Again, let X be a set. An *equivalence relation* on X is a relation of the form $\sim$, which satisfies three properties:

- $x \sim x$ for all x.
- $x \sim y$ iff $y \sim x$.
- $x \sim y$ and $y \sim z$ imply $x \sim z$.

An *equivalence class* is a subset $S = \{y \in X |\ y \sim x\}$. So, S is the set of all elements which are equivalent to x. Note that every two equivalence classes are either disjoint or identical. Thus, it makes sense to talk about the set of equivalence classes. This set is denoted $X/\sim$.

Now let's see how $\sim$ interacts with a metric. Let d' be a metric on X. As above, let $X/\sim$ denote the set of equivalence classes of X. Let us define, for $S_1, S_2 \in [X]$, the function

$$\delta(S_1, S_2) = \inf d'(s_1, s_2).$$

The infimum is taken over all possibilities where $s_1 \in S_1$ and $s_2 \in S_2$. In other words the "distance" from S_1 to S_2 is the "minimum" distance between a member of S_1 and a member of S_2.

Let d be the pathification of δ. We call $X/\sim$ a *good quotient* if d is a metric on $X/\sim$. We think of $X/\sim$ as the result of gluing certain points of X together. Moreover, if $x \sim y$ and x' is near x and y' is near y, then the pathification process forces x' to be near y'. So, at least in the case when we get a good quotient, the operation of gluing two points together sort of drags the rest of X along, like a rubber sheet. Before we give concrete examples, we point out that sometimes the gluing process leads to a horrible mess.

Exercise 2. Let $X = \boldsymbol{R}$ and write $x = y$ iff $x - y$ is rational. Show that $\boldsymbol{R}/\sim$ is not a good quotient.

3.2 The Gluing Construction in Action

In this section, which is mainly a series of exercises, we give you a chance to work through lots of concrete examples of the abstract construction given in the previous section.

Exercise 3. Let $X = X_1 \cup X_2$, where X_1 and X_2 are each copies of the unit disk, equipped with the standard metric, and $d(p_1, p_2) = 1$ if $p_1 \in X_1$ and $p_2 \in X_2$. You should picture two disks hovering, one on top of the other. Define $p_1 \sim p_2$ if and only if either $p_1 = p_2$ or else p_1 and p_2 are corresponding points in the boundaries of X_1 and X_2. Prove that the space $X/\sim$ is a good quotient, and is homeomorphic to the 2-sphere.

Exercise 4. The *projective plane* is the quotient of the sphere S^2 by the equivalence relation $p \sim -p$. The points p and $-p$ are called *antipodal points*. Prove that the projective plane is a surface.

Exercise 5. Let $X = S^1 \times [0,1]$ be a cylinder. Define an equivalence relation by the rule that $(x,0) \sim (x,1)$ and also $(x,y) \sim (x,y)$. Prove that $X/\sim$ is a good quotient, and also a surface, and also homeomorphic to the torus. See Figure 2.1.

Exercise 6. Let X be a metric space of the form

$$T \times \{1, 2, 3, 4, 5, 6, 7, 8\}.$$

So, X is the disjoint union of 8 triangles. Define an equivalence relation on X so that the resulting space is a surface and homeomorphic to a sphere.

Exercise 7. Describe how to glue a finite number of triangles together to make the octagon surface discussed in Chapter 1.

Exercise 8. We have already discussed the square torus in §1.1. Here is another desciption of the same space. On $\boldsymbol{R}^2$ define the equivalence relation $(x_1, y_1) \sim (x_2, y_2)$ iff $x_1 - x_2$ and $y_1 - y_2$ are both integers. Prove that both quotients we have described are good quotients and homeomorphic to each other. Prove also that the resulting space is homeomorphic to the surface of a donut, as in Figure 2.1.

Now we mention the cylinder and the Möbius band. Technically, these are surfaces with boundary. Both surfaces are obtained by gluing together one pair of opposite sides of a rectangle, as shown in Figure 1.

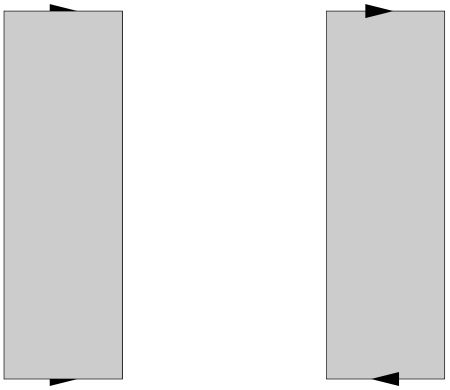

Figure 3.1 The cylinder and the Möbius band

The Möbius band has some amazing properties. First of all, it only has one boundary component. Second of all, consider a bug embedded in the Möbius band that travels from top to bottom. When the bug gets back to its original position, its notions of left and right are reversed.

We call a compact surface *orientable* if it does not contain a Möbius band. Otherwise, we call the surface *nonorientable*. Figure 3.2 shows two prototypical examples of nonorientable surfaces, the projective plane (left) and the Klein bottle (right). We have drawn in Möbius band subsets in both cases.

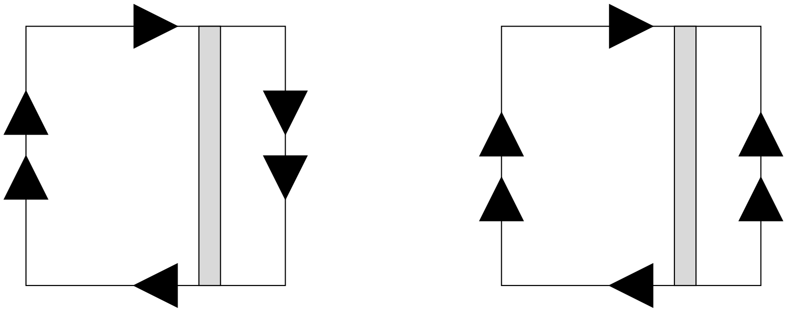

Figure 3.2. The projective plane and the klein bottle

Exercise 9. Prove that the version of the projective plane shown on the left-hand side of Figure 3.2 is homeomorphic to the version described in Exercise 4.

3.3 The Classification of Surfaces

Suppose that S_1 and S_2 are two compact surfaces. Let D_1 and D_2 be small open disks in S_1 and S_2. We assume that the lengths of the boundaries of D_1 and D_2 are the same. We cut out D_1 and D_2 from S_1 and S_2 to produce two new spaces. Finally, we glue the boundary of $S_1 - D_1$ to the boundary of $S_2 - D_2$ by an isometric map. We call the result $S_1 \sharp S_2$. Technically, the result depends on the choice of D_1 and D_2, but any choice of D_1 and D_2 leads to the same surface up to homeomorphism. Figure 3 shows an example of the connect-sum operation applied to two tori.

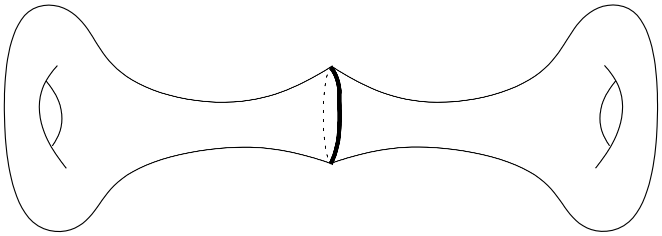

Figure 3.3 The Connected Sum

Exercise 10. Prove that $S_1 \sharp S_2$ is a surface. (*Hint:* The main difficulty is finding coordinate charts along the "seam".)

Letting T^2 stand for the torus, the surface

$$T_2 \sharp \cdots \sharp T_2,$$

made from g connected sum operations, is called the *surface of genus g*. In general, the *genus* of a compact surface is the integer g such that $\chi(S) = 2 - 2g$. A surface of genus g is sometimes denoted Σ_g. We say "it", because any g-fold connected sum of g tori gives rise to the same surface up to homeomorphism. This fact is part of the classification of surfaces.

Theorem 3.1 (Classification of Surfaces) *Let X be a compact surface. If X is orientable, then there is some g such that X is homeomorphic to Σ_g. If X is nonorientable, then there is some g such that Σ is homeomorphic to $\Sigma_g \sharp P$. Here P is the projective plane.*

If we assume that X is made from gluing together finitely many triangles according to the construction above, then the proof of the Theorem 3.1 is elementary. Roughly speaking, you cut X open into one big polygon and analyze the way the sides are glued together. The book [KIN] has a proof along these lines. The proof of the Euler formula that we give in the next section is quite similar to the proof of Theorem 3.1.

The proof for an arbitrary compact surface is to reduce to the special case where X is built from triangles. In other words, one shows that X is homeomorphic to a surface built from triangles. We say in this case that *X has a triangulation*. It turns out that every compact surface does have a triangulation, but the result is quite difficult to prove.

3.4 The Euler Characteristic

We will establish Euler's formula for orientable surfaces. Suppose that Σ_g is decomposed into polygons. We will prove that

$$\chi(\Sigma_g) := \text{faces} - \text{edges} + \text{vertices} = 2 - 2g. \tag{3}$$

The sum in equation (3) defines $\chi(\Sigma_g)$, and the result is the formula for $\chi(\Sigma_g)$.

First of all, we reduce to the case when the decomposition has just a single face. Suppose that Σ_g is decomposed into more than one face. We can find faces F_1 and F_2 joined together along an edge e. We can remove e and set $F = F_1 \cup_e F_2$. By this we mean that we stick F_1 and F_2 together along e and call the union F. We have created a new decomposition with one fewer face and one fewer edge. In particular, we have not changed the Euler characteristic.

It remains to consider the case when there is just one face whose boundary edges are paired in some way. We call this a *gluing pattern* for Σ_g. We say that the gluing pattern *has a cross* if we can find two pairs of glued edges (e_1, e_2) and (f_1, f_2) such that any line segment connecting e_1 to e_2 crosses any line segment connecting f_1 to f_2, as shown in Figure 3.4. In other words, the edges e_1, e_2 separate the edges f_1, f_2 from each other on the boundary of P.

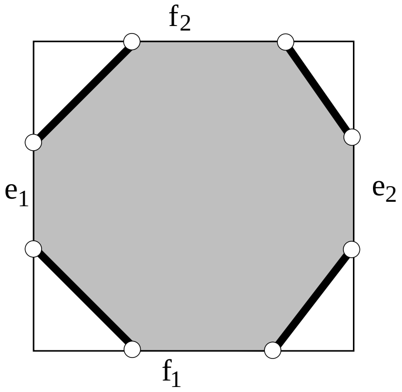

Figure 3.4. A crossing pattern of edges

Lemma 3.2 *If the gluing pattern for Σ_g does not have a cross, then it has a pair of consecutive edges that are glued together.*

Proof: We will assume that the gluing pattern has neither a cross nor a pair of consecutive edges and derive a contradiction. Say that a *special segment* is a line segment in the interior of F that joins the midpoints of a glued pair of edges. Let L_1 be a special segment. We rotate so that L_1 is vertical. Since L_1 does not join consecutive segments, and there are no crosses, we can find a special segment L_2 that lies to the left of L_1. Since L_2 does not join consecutive segments, we can find a special segment L_3, separated from L_1 by L_2. Next, we can find a special segment, L_4, separated from L_1 and

L_2 by L_3. And so on. In this way, we produce an infinite list $L_1, L_2, \ldots$ of distinct special segments. This contradicts the fact that F only has finitely many edges. ♠

Lemma 3.3 *If the gluing pattern for Σ_g has a cross, then Σ_g is not a sphere.*

Proof: Let (e_1, e_2) and (f_1, f_2) be two pairs of edges participating in a cross, as shown in Figure 3.5. Without loss of generality, it suffices to consider the case when the e and f edges are contained in the edges of the unit square. Figure 3.5 below shows the situation. The thick segments between the e edges and the f edges each represent a finite union of edges of F. Though we have not drawn things this way, one more of these segments could be empty.

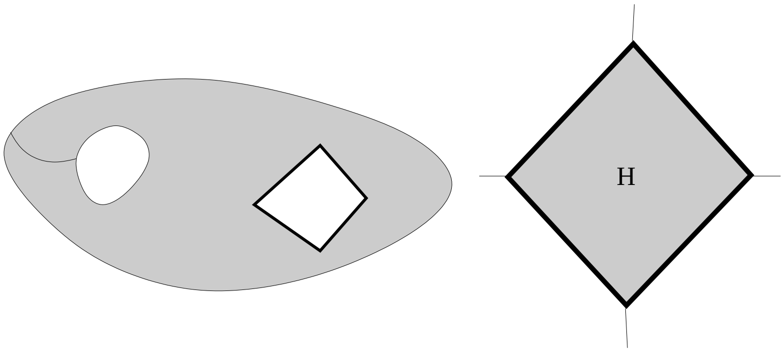

Figure 3.5. A torus with flaps.

Were we to glue the opposite sides of the unit square, we would get a torus, as shown on the left hand-side of Figure 3.5. To obtain Σ_g from this picture, we delete the white "flaps" from the torus, and then glue together the edges of the corresponding boundary, according to the original gluing pattern. The relevant boundary is drawn thickly.

It is convenient to draw the torus in a different way, this time with the handle drawn on the inside. Rather than draw the handle, we have just added the letter H, to denote that the drawn disk is really a disk with a handle attached. The right-hand side of Figure 3.5 shows this. Were we to draw the "flaps", they would be on the outside of the shaded region.

The right-hand side of Figure 3.5 realizes Σ_g as the connected sum of another oriented surface and a torus. In particular, Σ_g cannot be a sphere.

♠

Lemma 3.4 *Euler's formula is true for a sphere.*

Proof: Our proof goes by induction on the number of edges of F. In case F just has 2 edges, the decomposition of Σ has 1 edge and 2 vertices. We have

$$\chi(\Sigma_0) = 2 - 1 + 1 = 2.$$

This takes care of the base case. In general, some pair of consecutive edges of F is glued together. Since F is orientable, these edges point in opposite directions, as shown in Figure 3.6.

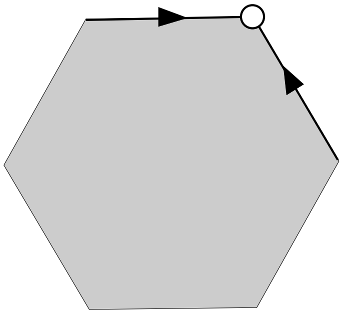

Figure 3.6. Gluing consecutive edges

In this case, we glue up the edges and erase the vertex between them. The result is a gluing pattern for Σ_g that is based on a polygon with 2 fewer edges. This is the induction step. ♠

Now we consider the general case. Our result goes by induction on g. We have already taken care of Σ_0. In light of Lemma 3.4, the converse of Lemma 3.3 is true: If Σ_g is not a sphere, then the gluing pattern for Σ_g does have a cross. Otherwise, we could just "zip up" Σ_g one pair of edges at a time and produce a sphere.

So, we start with a cross and reproduce the construction made in the proof of Lemma 3.3. That is, we arrive at the picture in Figure 3.5. When we replace our disk-with-handle with a disk, we produce a gluing diagram, based on a polygon F', for the surface Σ_{g-1}. Here F' has 4 fewer edges than F does. At the same time, F and F' have the same set of vertices, and they are glued together in the same way.

By induction, Euler's formula holds for Σ'. Hence

$$f' - e' + f' = 2 - 2(g-1).$$

Here $f' = 1$ is the number of faces in the decomposition and e' is the number of edges, and v' is the number of vertices. By construction $f' = f$ and $e' = e - 2$ and $v' = v$. So, we get

$$f - e + v = 2 - 2g,$$

as desired.

Exercise 11. Prove Euler's formula for nonorientable surfaces.

4 The Fundamental Group

The purpose of this chapter is to define the fundamental group, an object we discussed briefly in §1.3. In the next chapter, we will compute some examples. As we mentioned in §1.3, the fundamental group is an object that organizes all the different loops on a surface (or any topological space, for that matter). In this chapter, I will first talk about groups in general, then groups will disappear from the discussion for a while; then they will come back in a really surprising way. For a more formal treatment of the fundamental group, see, e.g., [HAT].

4.1 A Primer on Groups

If you haven't had any group theory, you can find a treatment in any number of abstract algebra books; see, for instance, [HER]. A *group* is a set G, together with an "operation" $*$, which satisfies the following axioms:

- $g_1 * g_2$ is defined and belongs to G for all $g_1, g_2 \in G$.
- $g_1 * (g_2 * g_2) = (g_1 * g_2) * g_3$ for all g_1, g_2, g_3.
- There exists a (unique) $e \in G$ such that $e * g = g * e = g$ for all $g \in G$.
- For each $g \in G$ there is a (unique) element h such that $g * h = h * g = e$. This element is called "g inverse" and is usually written as $h = g^{-1}$.

The group G is called *Abelian* if, additionally, $g_1 * g_2$ and $g_2 * g_1$ are always equal. A *subgroup* of a group is a subset $H \subset G$ which is closed under the group law. So, if $h \in H$ then $h^{-1} \in H$ and if $h_1, h_2 \in H$ then $h_1 * h_2 \in H$.

Here are some examples of groups:

- Z, with the $+$ operation, forms an Abelian group.
- If G_1 and G_2 are groups, then $G_1 \times G_2$ can be made a group using the law $(g_1, g_2) * (h_1, h_2) = (g_1 * h_1, g_2 * h_2)$.
- The set $SL_n(\boldsymbol{Z})$ of $n \times n$ integer matrices with determinant 1 forms a non-Abelian group. The group law is matrix multiplication.

- Let A be a collection of n things, for instance $A = \{1, ..., n\}$. Say that a *permutation* is a bijection $f : A \to A$. There are $n!$ different permutations, and they form a finite group. The $*$ operation is composition of maps. This group is called S_n.

Let G_1 and G_2 be groups. A map $f : G_1 \to G_2$ is a *homomorphism* if

$$f(a * b) = f(a) * f(b)$$

for all $a, b \in G_1$. Here the $*$ on the left-hand side is the rule for G_1 and the $*$ on the right-hand side is the one for G_2. The map f is called an *isomorphism* if f is a bijection and also a homomorphism.

Here is a nice example. Let G be a finite group, and let n be the number of elements in G. We're going to produce a homomorphism from G into S_n, the permutation group on n things. The n things are just the elements of G. So, given an element $g \in G$, how do we permute the elements of G? We define the map $f_g : G \to G$ using the rule $f_g(h) = gh$. It turns out that f_g is a bijection, and $f_{g_1} = f_{g_2}$ only if $g_1 = g_2$. The map $g \to f_g$ is a one-to-one homomorphism from G into S_n. We have essentially given the proof of *Cayley's theorem*: every finite group is isomorphic to a subgroup of a permutation group.

4.2 Homotopy Equivalence

Now we go back to metric spaces. Let X and Y be metric spaces. Let $I = [0, 1]$ be the unit interval. Two maps $f_0, f_1 : X \to Y$ are said to be *homotopic* if there is a continuous map $F : X \times I \to Y$ such that

- $F(x, 0) = f_0(x)$ for all $x \in X$.
- $F(x, 1) = f_1(x)$ for all $x \in X$.

To explain the intuitive idea, it is useful to define $f_t : X \to Y$ by the formula $f_t(x) = F(x, t)$. Then the map f_t interpolates between f_0 and f_1, with f_t being very close to f_0 when t is near 0 and f_t being very close to f_1 when t is near 1. The map F is called *a homotopy* from f_0 to f_1.

If is useful to write $f_0 \sim f_1$ if these maps are homotopic. Let $C(X, Y)$ denote the set of all continuous maps from X to Y. One can think of $\sim$ as a relation on the set $C(X, Y)$.

Exercise 1. Prove that $\sim$ is an equivalence relation on $C(X, Y)$.

Exercise 2. Prove that every two elements of $C(X, \boldsymbol{R}^n)$ are homotopic. (*Hint*: Prove that any map $f : X \to \boldsymbol{R}^n$ is homotopic to the zero-map f_0 defined by the property $f_0(x) = 0$ for all x. Then, use the fact that $\sim$ is an equivalence relation.)

Exercise 3 (Challenge). Let P be a polynomial

$$P(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_0.$$

Let Q be the polynomial $Q(x) = x^n$. So, P and Q have the same leading term. We can think of P as a map from $\boldsymbol{C}$ to $\boldsymbol{C}$. Here $\boldsymbol{C}$ is the complex plane. For any R we can let $X \subset \boldsymbol{C}$ be the circle of radius R centered at 0. That is

$$X = \{z \in \boldsymbol{C} | \ |z| = R\}.$$

First of all, prove that $0 \notin P(X)$ if R is sufficiently large. This means that we can think of P and Q as maps from X to $Y = \boldsymbol{C} - \{0\}$. Prove that $P, Q : X \to Y$ are homotopic if R is sufficiently large.

4.3 The Fundamental Group

From now on we are going to take $X = I$, the unit interval, and we are going to study the space Y by looking at the maps from I into Y. For this entire discussion we choose a special "reference point" $y_0 \in Y$, which we call the *basepoint*.

Say that a *loop* in Y is a continuous map $f : I \to Y$ such that

$$f(0) = f(1) = y_0.$$

The reason for the terminology should be pretty clear. Say that two loops f_0 and f_1 are *loop homotopic* if there is a homotopy F from f_0 to f_1 such that f_t is a loop for all $t \in [0, 1]$. This is to say that $F(0, t) = F(1, t) = y_0$ for all t. We write $f_0 \sim f_1$ in this case. Figure 4.1 shows an example. Just as in Exercise 1, this relation is an equivalence relation. Note that the equivalence relation here is slightly different than the one in the previous section, because of the added constraint that $F(0, t) = F(1, t) = y_0$ for all t.

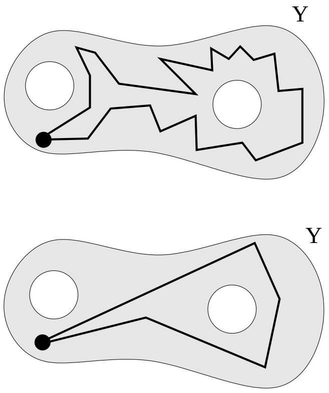

Figure 4.1. homotopic loops

As a set, $\pi_1(Y, y_0)$ is the set of equivalence classes of loops. The really interesting thing is that we can make $\pi_1(Y, y_0)$ into a group. Here is the construction. Suppose that we have two elements $[f]$ and $[g]$ of $\pi_1(Y, y_0)$. We can let f and g be representatives of the equivalence classes $[f]$ and $[g]$, respectively. That is, $f : [0, 1] \to Y$ is a loop and $g[0, 1] \to Y$ are both loops.

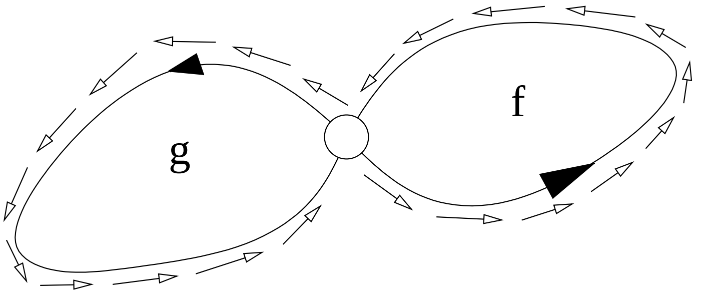

Figure 4.2. Composing loops

We define the new loop $h = f * g$ by the following rule.

- If $x \in [0, 1/2]$, we define $h(x) = f(2x)$. That is, the first half of h traces out all of f, but twice as fast.
- If $x \in [1/2, 1]$, we let $x' = x - 1/2$ and then we define $h(x) = g(2x')$. That is, the second half of h traces out g, but twice as fast.

We write $h = f * g$. See Figure 4.2.

Exercise 4. Suppose that $\widehat{f}$ and $\widehat{g}$ are different representatives for $[f]$ and

$[g]$. That is, f and $\widehat{f}$ are equivalent loops and g and $\widehat{g}$ are equivalent loops. Let $\widehat{h} = \widehat{f} * \widehat{g}$. Prove that $[\widehat{h}] = [h]$. In other words, prove that h and $\widehat{h}$ are equivalent loops. This exercise is pretty easy, but quite tedious.

Given Exercise 4, we can define

$$[f] * [g] = [f * g], \tag{4}$$

and this definition is independent of the equivalence class representatives we used to make the definition.

Exercise 5. Show, for any three loops, f, g, h, that $(f * g) * h$ is equivalent to $f * (g * h)$. This means that $([f] * [g]) * [h] = [f] * ([g] * [h])$. This is the associative law for groups.

Exercise 6. Let e be the loop defined by the rule $e(x) = y_0$ for all $x \in I$. Show that $[e] * [g] = [g] * [e] = [g]$ for all loops g. This means that $[e]$ plays the role of the identity element in $\pi_1(Y, y_0)$.

Exercise 7. Let g be any loop. Define the loop g^* so that it satisfies the equation $g^*(x) = g(1-x)$. In other words, g^* traces out the same loop as g, but in the opposite direction. Prove the following result: If g_1 and g_2 are equivalent, then g_1^* and g_2^* are equivalent. Finally, prove that $[g] * [g^*] = [e]$ and $[g^*] * [g] = [e]$. In other words, the inverse of $[g]$ is given by $[g^*]$.

Combining Exercises 5, 6, and 7, we see that $\pi_1(Y, y_0)$ is a group. So, to each space Y we can pick a basepoint y_0 and then define the group $\pi_1(Y, y_0)$. This group is known as the *fundamental group* of Y. (We will see below that the group you get does not really depend on the basepoint.)

4.4 Changing the Basepoint

Say that two points y_0, y_1 are *connected by a path* if there is a continuous map $f : I \to Y$ such that $f(0) = y_0$ and $f(1) = y_1$. Say that Y is *path connected* if every two points in Y can be connected by a path. For instance $\boldsymbol{R}^n$ is path connected whereas $\boldsymbol{Z}$ is not.

Lemma 4.1 *Suppose that $y_0, y_1 \in Y$ are connected by a path. Then $\pi_1(Y, y_0)$ and $\pi_1(Y, y_1)$ are isomorphic groups. In particular, if Y is path connected,* *then the (isomorphism type of the) group $\pi_1(Y, y)$ is independent of the choice of basepoint y and we can just write $\pi_1(Y)$.*

Proof (Sketch). Let d be a path which joins y_0 to y_1. Let d^* be the reverse path, which connects y_1 to y_0. We want to use d and d^* to define a map from $\pi_1(Y, y_0)$ to $\pi_1(Y, y_1)$. Given any y_0-loop $f_0 : I \to X$ with $f_0(0) = f_0(1) = y_0$, we can form a y_1-loop by the formula

$$f_1 = d * f * d_*.$$

In other words, the first part of f_1 travels backward along d from y_1 to y_0, the second part travels around f_0, and the third part travels back to y_1. You should picture a lasso, as in Figure 4.3.

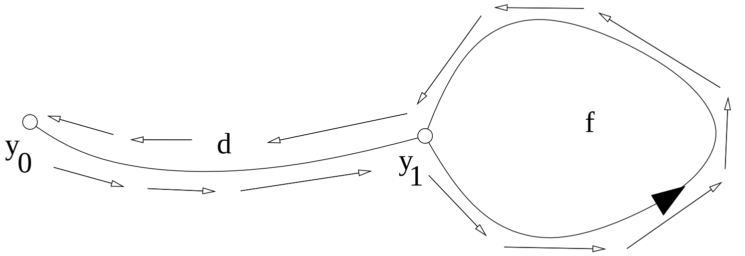

Figure 4.3. A lasso

Using arguments similar to the ones for the exercises above, you can show the following result: If f_0 and $\widehat{f_0}$ are equivalent, then f_1 and $\widehat{f_1}$ are equivalent. In other words, the map H, which sends $[f_0] \in \pi_1(Y, y_0)$ to $[f_1] \in \pi_1(Y, y_1)$ is well defined independent of the equivalence class representative used to define it. So, now we have a well-defined map $H : \pi_1(Y, y_0) \to \pi_1(Y, y_1)$. After this, one shows that H is a homomorphism. That is, $H([f] * [g]) = H([f]) * H([g])$. This is not hard to do once you draw a picture of what is going on.

Rather than show that H is one-to-one and onto directly, one can define a map $H^* : \pi_1(Y, y_1) \to \pi_1(Y, y_0)$ just by reversing the roles of the two points. In other words, the loop f_1 is mapped to

$$f_0^* = d_* * f_1 * d.$$

Note that f_0^* and f_0 are not precisely the same loop. If you draw pictures you will see that there is some extra slack in f_0^*. However, it turns out that $[f_0^*] = [f_0]$. In other words, the two loops are loop homotopic. Thus H and H^* are inverses of each other. Hence H is an isomorphism. ♠

4.5 Functoriality

The word *functoriality* refers to a situation where you are assigning one kind of an object to another in a way which respects the "natural" transformations between the two kinds of objects. This notion is defined precisely in any book on category theory. In our case, we are assigning a group $\pi_1(Y, y_0)$ to a *pointed space* (Y, y_0). (By pointed space we mean a space with a chosen basepoint.) The natural transformations of pointed spaces are basepoint-preserving continuous maps and the natural transformations between groups are homomorphisms.

We would like to see that our transformation (or *functor*) from spaces to groups respects these transformations. Lemmas 4.2 and 4.3 together contain this information.

Lemma 4.2 *Let (Y, y_0) and (Z, z_0) be two pointed spaces, and let $f : Y \to Z$ be a continuous map such that $f(y_0) = z_0$. Then there is a homomorphism $f_* : \pi_1(Y, y_0) \to \pi_1(Z, z_0)$.*

Proof: Let $[a] \in \pi_1(Y, y_0)$ be an equivalence class of loops, with representative a. So, $a : I \to Y$ is a loop. The composition $f \circ a$ is loop in Z. We define $f_*[a] = [f \circ a]$. If $[a_0] = [a_1]$, then there is a homotopy H from a_0 to a_1. But then $f \circ H$ is a loop homotopy from $f \circ a_0$ to $f \circ a_1$. So, $[f \circ a_0] = [f \circ a_1]$ and our map is well defined. Note that $f \circ (a * b) = (f \circ a) * (f \circ b)$. Hence $f * ([a] * [b]) = (f_*([a])) * (f_*([b]))$. Hence f_* is a homomorphism. ♠

Suppose that $f : Y \to Z$ is a continuous map and $g : Z \to W$ is a continuous map. Let's arrange so that $f(y_0) = z_0$ and $f(z_0) = w_0$. Then $g \circ f$ is a map from Y to W and $(g \circ f)_*$ is a homomorphism from $\pi_1(Y, y_0)$ to $\pi_1(W, w_0)$.

Lemma 4.3 $(g \circ f)_* = g_* \circ f_*$.

Proof: Let $[a] \in \pi_1(Y, y_0)$. Then

$$(g \circ f)_*[a] = [(g \circ f) \circ a] = [g \circ (f \circ a)] = g_*[f \circ a] = g_* f_*[a].$$

That is it. ♠

If $f : Y \to Y$ is the identity map, then f_* is the identity map on $\pi_1(Y, y_0)$. Also, if $h : Y \to Z$ is a homeomorphism, then we have the inverse homeomorphism h^{-1}. But $h \circ h^{-1}$ is the identity. Hence $h_* \circ h_*^{-1}$ is the identity homomorphism. Likewise $h^{-1} \circ h_*$ is the identity homomorphism. In short h_* (and also h_*^{-1}) is a group isomorphism. So

Theorem 4.4 *If $\pi_1(Y, y_0)$ and $\pi_1(Z, z_0)$ are $\underline{\text{not}}$ isomorphic groups, then there is no homeomorphism from Y to Z which maps y_0 to z_0.*

The above is slightly contrived because we don't really care about these basepoints. Recall that $\pi_1(Y, y_0)$ does not depend on the basepoint if Y is path connected. So

Theorem 4.5 *Suppose Y and Z are path connected spaces. If $\pi_1(Y)$ and $\pi_1(Z)$ are not isomorphic, then Y and Z are not homeomorphic.*

What's really great about this result is that we can use it to tell the difference between spaces just by looking at these groups. Of course, the question remains: How do we actually compute these groups? In the next chapter, we will go into much more details about this.

4.6 Some First Steps

Here we will just take some first steps in the computation of fundamental groups. Once we have more theory, these computations will be easy. So, what fundamental groups can we compute? It is easy to see (compare Exercise 2) that any two loops in $\boldsymbol{R}^n$ (based at 0) are equivalent. Hence $\pi_1(\boldsymbol{R}^n, 0)$ is the trivial group.

Exercise 8A (Challenge). Prove that there is a loop in S^2 (the 2-sphere) whose image is all of S^2. (*Hint*: If you know about the Hilbert plane-filling curve from real analysis, you're in good shape for this problem.)

Exercise 8B (Challenge). Prove that $\pi_1(S^2, p)$ is the trivial group. Here $p \in S^2$ is any point. (*Hint*: The intuitive idea is this: If the loop misses some point $q \neq p$, you can just "slide" the loop "down to p" by pushing it away from the missed point. However, you have to deal with the loops which come from Exercise 8A.)

Exercise 9. If (Y, y_0) and (Z, z_0) are two pointed spaces, then the product

$$(Y \times Z, (y_0, z_0)$$

is again a pointed space. Prove that

$$\pi_1(Y \times Z, (y_0, z_0)) = \pi_1(Y, y_0) \times \pi_1(Z, z_0).$$

Exercise 10 (Challenge). Prove that $\pi_1(S^1, p)$ is nontrivial. (*Hint*: Think of S^1 as the unit circle in $\boldsymbol{R}^2$ and consider the loop

$$f(t) = (\cos(2\pi t), \sin(2\pi t)).$$

Show that this loop is inequivalent to the identity loop.)

Let $T = S^1 \times S^1$. Here T^2 is the torus. From Exercises 9 and 10 we know that $\pi_1(T^2)$ is nontrivial. (We don't worry about the basepoint because T is obviously path connected.) On the other hand, by Exercise 8, $\pi_1(S^2)$ is trivial. Hence S^2 and T^2 are not homeomorphic!

5 Examples of Fundamental Groups

The purpose of this chapter is to compute the fundamental group for some familiar objects:

- the circle;
- the torus;
- the 2-sphere;
- the projective plane;
- lens spaces;
- the Poincaré homology sphere.

I will work out the first three in detail and then guide you through the computation for the others. The last section is too advanced for an undergraduate course but I couldn't resist.

5.1 The Winding Number

Let S^1 be the circle. We think of S^1 as the set of unit complex numbers in $\boldsymbol{C}$. We choose 1 for our basepoint of S^1. In this section we will describe how to assign an integer to a continuous loop $g : [0,1] \to S^1$.

First we will explain the idea intuitively and then we will get to the formalities. Think of the loop g as describing a bug crawling around the unit circle. Imagine that you are at the center of the circle watching the bug. You always follow the bug with your eyes, staring straight at it the whole time. (Your rubber neck allows you to do this.) After the bug has completed his trip, you are looking in the same direction as initially. However, your head has been twisted around some number of times. The winding number is the integer, positive for counterclockwise and negative for clockwise, which names how many times your head is twisted around.

Now we come to the formalities. Let $\boldsymbol{R}$ denote the real numbers. There is a natural map $E : \boldsymbol{R} \to S^1$ given by

$$E(t) = \exp(2\pi i t) = \cos(2\pi t) + i \sin(2\pi t).$$

This map is certainly onto and continuous, but it has some other special properties. Say that an *open special arc* in S^1 is a set of the form

$$C(z) = \{w \in S^1 |\ d(z,w) < 1/100\}.$$

Here $d(z,w) = |z-w|$, the usual Euclidean distance. The choice of $1/100$ is convenient but fairly arbitrary. The point is just that open special arcs are smaller than semicircles.

Exercise 1. Let C be an open special arc. Prove that $E^{-1}(C)$ consists of a countably infinite number of disjoint open intervals and that the restriction of E to any of them is a homeomorphism from the interval onto C.

Lemma 5.1 *Let $[a,b] \subset \boldsymbol{R}$ be an interval. Suppose $g : [a,b] \to S^1$ is a map such that $g([a,b])$ is contained in a special arc. Suppose also that there is a map $\widetilde{g} : \{a\} \to \boldsymbol{R}$ such that $E \circ \widetilde{g}(a) = g(a)$. Then we can define $\widetilde{g} : [a,b] \to \boldsymbol{R}$ such that $E \circ \widetilde{g} = g$ on all of $[a,b]$. This extension of $\widetilde{g}$ is unique.*

Proof: If E had an inverse we could define $\widetilde{g} = E^{-1} \circ g$. Also, we would be forced to make this definition and so the extension of $\widetilde{g}$ to $[a,b]$ would be unique. Unfortunately, E is not invertible. Fortunately, we have Exercise 1, which shows that E is "invertible" in some sense. Let C be the special arc which exists by hypothesis. Let $\widetilde{C} \subset E^{-1}(C)$ be the unique interval from Exercise 1 which contains $\widetilde{g}(a)$. By Exercise 1, the map $E : \widetilde{C} \to C$ is a homeomorphism. Let $F : C \to \widetilde{C}$ be the inverse of (the restricted version of) E. Since $g[a,b] \subset C$ we can (and must) define $\widetilde{G} = F \circ g$. ♠

Let $1 = E(\boldsymbol{Z})$ be the basepoint of S^1. Let $I = [0,1]$. Recall that an element of $\pi_1(S^1, 1)$ is a map $g : I \to S^1$ such that $g(0) = g(1) = 1$.

Exercise 2. Given the map g, prove that there exists some N with the following property. If $x, y \in [0,1]$ and $|x-y| < 1/N$ then the set $g([x,y])$ is contained in a special arc. (*Hint*: You might want to use the fact that every infinite sequence in $[0,1]$ has a convergent subsequence. This is basically the Bolzano–Weierstrass theorem.)

Here is an improved version of Lemma 5.1.

Lemma 5.2 *Let $g : [0,1] \to S^1$ be a loop. Then there is a unique map $\widetilde{g} : [0,1] \to \boldsymbol{R}$ such that $\widetilde{g}(0) = 0$ and $E \circ \widetilde{G} = G$ on all of $[0,1]$.*

Proof: From Exercise 2 we can find some N so that the points $t_i = i/N$ have the following property. The image $g([t_i, t_{i+1}])$ is contained in a special arc for $i = 0, \ldots, (N-1)$. Now we go by induction. First of all, by Lemma 5.1 we can define $\widetilde{g}$ uniquely on $[t_0, t_1]$. But then by Lemma 5.1 again, we can define $\widetilde{g}$ uniquely on $[t_1, t_2]$. And so on. ♠

Definition 5.1. We define the *winding number* of g to be the value of $\widetilde{g}(1) \in \boldsymbol{Z}$. We write this as $w(g)$. Note that $\widetilde{g}(1) \in \boldsymbol{Z}$ because $g(1) = E(\widetilde{g}(1)) = 0$.

We would like to see that the winding number only depends on the homotopy class of the loop. Going back to our intuitive notion of the winding number, suppose that two bugs are running around the unit circle, and they stay pretty close to each other. Then you will always be looking in about the same direction if you watch either bug. So, your head will be turned around the same number of times if you watch either bug. Now we give the formal proof.

Lemma 5.3 *Suppose that g_0 and g_1 are homotopic loops in S^1. Then $w(g_1) = w(g_1)$.*

Proof: Let G be the homotopy between g_0 and g_1. Let $g_t(x) = G(x,t)$. The same argument as in Exercise 2 proves that there is some N with the following property: If $s, t \in [0,1]$ are any points such that $|s - t| < 1/N$ and $x \in [0,1]$ is fixed, then

$$|G(x,s) - G(x,t)| < 1/100.$$

Using the other notation, we have $d(g_s(x), g_t(x)) < 1/100$. But then $d(\widetilde{g}_s(x), \widetilde{g}_t(x))$ is either less than $1/100$ or greater than $1/2$. By continuity, the alternative cannot change. Also

$$d(\widetilde{g}_s(0), \widetilde{g}_t(0)) = d(0,0) = 0 < 1/100.$$

This shows that the first alternative always holds and $\widetilde{g}_s(x)$ and $\widetilde{g}_t(x)$ are always within $1/100$ of each other. But then $w(g_s) = w(g_t)$, because both are integers within $1/100$ of each other. From here it is easy to see that $w(g_0) = w(g_1)$. ♠

5.2 The Circle

We will use the winding number to compute $\pi_1(S^1, 1)$, the fundamental group of the circle.

Given a loop g, representing an element of $\pi_1(S^1, 1)$, we define

$$w([g]) = w(g).$$

By Lemma 5.3, this gives us a well-defined map $w : \pi_1(S^1, 1) \to \boldsymbol{Z}$.

Lemma 5.4 *w is onto.*

Proof: Let $g(t) = \exp(2\pi int)$. Then $w(g) = n$. ♠

Execise 3. Prove that w is a homomorphism.

Lemma 5.5 *w is an isomorphism.*

Proof: Since w is a homomorphism, it suffices to prove the following statement. If $w(g) = 0$, then g is homotopic to the constant loop. Now, if $w(g) = 0$, then $\widetilde{g} : [0, 1] \to \boldsymbol{R}$ is a loop. But $\pi_1(\boldsymbol{R}, 0) = 0$. Hence there is a loop homotopy $\widetilde{G}$ from $\widetilde{g}$ to the constant loop $\widetilde{g}_0 : S^1 \to \boldsymbol{R}$. But then $E \circ \widetilde{G}$ is a loop homotopy from g to the constant loop in S^1. This shows that w is an isomorphism. ♠

The last result shows that $\pi_1(S^1, 1)$ is isomorphic to $\boldsymbol{Z}$.

Remark. The main property we used about the circle was the existence and special properties of the map $E : \boldsymbol{R} \to S^1$. We also used the property that $\pi_1(\boldsymbol{R}, 0) = 0$. It turns out that this will be a general method for us when we compute the fundamental groups. All the special properties we established are summarized by the statement that $\boldsymbol{R}$ is the *universal cover* of S^1 and E is the *universal covering map*. In the next chapter I will develop these ideas in great generality.

5.3 The Fundamental Theorem of Algebra

The Fundamental Theorem of Algebra says that every complex polynomial

$$P(z) = a_0 + a_1 z + \cdots + a_n z^n$$

has a root. This result has a nice proof based on the ideas we have been developing. For convenience, we divide through so that $a_n = 1$. We think of P as a continuous map from $\boldsymbol{C}$ to $\boldsymbol{C}$. If P has no roots, then P is a continuous map from $\boldsymbol{C}$ to $\boldsymbol{C} - \{0\}$.

Let C_r denote the circle of radius r centered at the origin. Let S^1 denote the unit complex numbers. Given $r > 0$, consider the map $\gamma_r : S^1 \to S^1$ given by

$$\gamma_r(u) = \frac{P(ru)}{|P(ru)|}.$$

By construction γ_r is a continuous loop, and hence an element of $\pi_1(S_1)$.

When r is small, $P(C_r)$ is just a tiny loop around $f(0) \neq 0$. Hence $[\gamma_r] = 0 \in \pi_1(S_1)$ for r small. But γ_r varies continuously with r. Hence $[\gamma_r] = 0$ for all r. On the other hand, when $z \in C_r$ and r is large, we have

$$P(z) = z^n + f(z), \qquad |f(z)| < \epsilon_r |P(z)|.$$

Here ϵ_r is some constant that tends to 0 as $r \to \infty$. The point is that the highest order term dominates the sum of the remaining terms.

Our estimate tells us that γ_r converges to the loop $z \to z^n$ as $r \to \infty$. Hence $[\gamma_r] = n$ for r large. This is a contradiction. The only way out is that P is not a continuous map from $\boldsymbol{C}$ to $\boldsymbol{C} - \{0\}$. But then 0 must be in the image of P. That is, P has a root.

5.4 The Torus

Exercise 9 of Chapter 4 asked you to show that

$$\pi_1((Y, y) \times (Z, z)) = \pi_1(Y, y) \times \pi_1(Z, z).$$

The torus T^2 is homeomorphic to $S^1 \times S^1$ and also path connected. Hence $\pi_1(T^2) = \boldsymbol{Z} \times \boldsymbol{Z}$. Iterating, we get $\pi_1(T^n) = \boldsymbol{Z}^n$.

5.5 The 2-Sphere

Let $I = [0,1]$ as above. Let $x \in S^2$ be some basepoint. This section, which consists mainly of exercises, will guide you through the proof that $\pi_1(S^2, x) = 0$.

Say that a loop $g : I \to S^2$ (anchored at x) is *bad* if $g(I) = S^2$ and otherwise is *good*.

Exercise 4. Prove that any good loop is homotopic to a point.

Exercise 5. Let $[a,b]$ be an interval, and let H be a hemisphere in $\boldsymbol{R}^2$. Let $f : [a,b] \to H$ be a continuous map. Prove that there is homotopy $F : [a,b] \times [0,1] \to \Delta$ such that

- $F(a,t)$ and $F(b,t)$ are independent of t.
- $F(x,0) = f(x)$ for all x.
- $f_1 : [a,b] \to \Delta$ is contained in a circular arc joining $f(a)$ to $f(b)$.

Exercise 6. Let g be an arbitrary loop on S^2. Prove that there is a finite partition $0 = t_0 < t_1 < \ldots < t_n = 1$ such that g maps each interval $[t_i, t_{i+1}]$ into a hemisphere. Now conclude from Exercise 5 that g is loop homotopic to a good loop.

Since every loop in S^2 is loop homotopic to a good loop, and every good loop is loop homotopic to a point, every loop in S^2 is homotopic to a point. Therefore, $\pi_1(S^2, x) = 0$. The same argument works for S^n, with $n > 2$.

5.6 The Projective Plane

As in §3.2, we think of $\boldsymbol{P}^2$, the projective plane, as the quotient $S^2/\sim$, where $x \in S^2$ is equivalent to itself and to the antipodal point $-x$. There is a nice map $E : S^2 \to \boldsymbol{P}^2$ given by $E(x) = [x]$. As our notation suggests, E plays the same role here that the same-named map played above when we considered the circle.

Let $x_+ = (0,0,1)$, and let $x_- = (0,0,-1)$. Clearly, we have $E(x_+) = E(x_-)$.

Exercise 7. Suppose that $g : [0,1] \to \boldsymbol{P}^2$ is a loop based at x_+. Prove that there is a unique map $\widetilde{g} : [0,1] \to S^2$ such that $\widetilde{g}(0) = x_+$ and $E \circ \widetilde{g} = g$ (*Hint*: Just imitate what was done for the circle.)

Note that either $\widetilde{g}(1) = x_+$ or $\widetilde{g}(1) = x_-$. We define $w(g) = +1$ if $\widetilde{g}(1) = x_+$ and $w(g) = -1$ if $\widetilde{g}(1) = x_-$.

Exercise 8. Prove that $w([g])$ is well defined independent of the loop homotopy equivalence class of g. Prove also that w gives an isomorphism from $\pi_1(\boldsymbol{P}^2)$ to $\boldsymbol{Z}/2$.

In general we have $\boldsymbol{P}^n = S^n/\sim$, where $x \sim -x$. Thus there is always this two-to-one map from S^n to $\boldsymbol{P}^n$. An argument similar to the one given above shows that $\pi_1(\boldsymbol{P}^n) = \boldsymbol{Z}/2$. Here $\boldsymbol{P}^n$ is called *projective n-space*.

5.7 A Lens Space

Before reading this section, you should probably know what a manifold is; see §2.8 for details.

We think of S^3 as the set of the form

$$\{(z,w)|\ |z|^2 + |w|^2 = 1\} \subset \boldsymbol{C}^2 = \boldsymbol{R}^4.$$

This is an exotic way of expressing the fact that S^3, the 3-sphere, is the unit sphere in $\boldsymbol{R}^4$. The equality $\boldsymbol{C}^2 = \boldsymbol{R}^4$ comes from the map

$$(x_1 + iy_1, x_2 + iy_2) \to (x_1, y_1, x_2, y_2).$$

Here is a nice equivalence relation on S^3. Let's define

$$(z,w) \sim (uz, u^2 w)$$

if and only if u is some 5th root of unity. Each equivalence class on $S^3/\sim$ has 5 points. Let's call this space $L(2,5)$. The 2 comes from the u^2 term, and the 5 comes from the fact that we are taking 5th roots of unity. Obviously, you could make this construction for other choices.

Here is a sketch of how to visualize $L(2,5)$. Any point in $L(2,5)$ is equivalent to a point of the form (z,w), where the argument of z lies in the interval $(0, 2\pi/5)$. Let $S \subset S^3$ be this set. We can write

$$S = \bigcup_{\theta \in [0,2\pi/5]} S_\theta,$$

where S_θ consists of points of the form (z, w) where $z = \exp(i\theta)$. The whole sphere S^3 is tiled by 5 copies of S. For instance, one of the adjacent copies consists of those sets S_θ, where $\theta \in [2\pi/5, 4\pi/5]$.

Now we are going to (partially) explain how to visualize S. The "slice" S_θ is a disk, and the boundary of S_θ is the circle

$$C = \{0\} \times \{w : \ |w| = 1\}.$$

All the slices share C as a common boundary, but otherwise they are disjoint. So, S looks something like a circular pillow, or the solid region between two contact lenses stuck boundary to boundary. The left-hand side of Figure 5.1 shows a side view. The two dots represent C. To get a better picture, you could revolve this planar figure about the vertical axis.

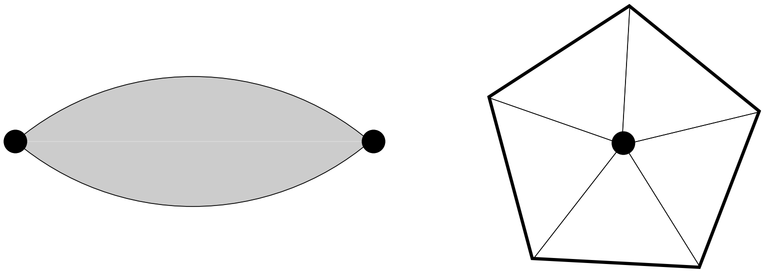

Figure 5.1. The domain S

The right-hand side of Figure 5.1 shows a top view of S. We imagine that we are looking at S_0, and that the rest of S is underneath. The other boundary component is $S_{2\pi/5}$. We have drawn the circle C as a pentagon, to suggest the what is going on. We observe the following things.

- Each point in the interior of S is equivalent only to itself.
- Each point on the interior of the "front" of S, meaning the set $S_0 - C$, is equivalent to one point on $S_{2\pi/5} - C$.
- Each point on one of the edges of C is equivalent to the 4 other points at corresponding positions on the other edges. The triangular subdivision is supposed to serve as a guide to the gluings.

I have not described things completely, because I want to leave you something to think about.

Exercise 9. Prove that $L(2,5)$ is a good quotient in the sense of §3.1, and also a manifold.

There is an obvious map $E : S^3 \to L(2,5)$. Using E we can show that $\pi_1(L(2,5)) = \boldsymbol{Z}/5$. Generalizing this construction in an obvious way, we see that we can produce a 3-manifold whose fundamental group is $\boldsymbol{Z}/n$. These spaces $L(m,n)$ are called *lens spaces*.

5.8 The Poincaré Homology Sphere

Before we give the last example, we need to make a detour and discuss a different way to think about S^3. Let $SO(3)$ denote the group of orientation preserving (i.e., physically possible) rotations of S^2. It turns out that there is an amazing map from S^3 to $SO(3)$ which is really the map from S^3 to $\boldsymbol{P}^3$ in disguise. So, given an element $q \in S^3$, we need to produce a rotation R_q of S^2.

Here is the construction. We think of S^3 as the *unit quaternions*. That is, a point in S^3 can be thought of as a symbol of the form

$$a+bi+cj+dk, \qquad a,b,c,d \in \boldsymbol{R}, \qquad a^2+b^2+c^2+d^2=1.$$

The symbols i,j,k satisfy the following rules:

- $i^2=j^2=k^2=-1$.
- $ij=k$ and $jk=i$ and $ki=j$.

Given these rules, you can multiply quaternions together in a way which is similar to how you multiply complex numbers together.

Given any $q \in S^3$ as above, we define

$$q^{-1} = a-bi-cj-dk.$$

Then you can check that $qq^{-1}=q^{-1}q=1$. In other words, the unit quaternions form a group under multiplication!

We can identify $\boldsymbol{R}^3$ with the pure quaternions, namely those of the form $0+bi+cj+dj$. The isomorphism to $\boldsymbol{R}^3$ is just given by

$$0+ai+bj+ck \to (a,b,c).$$

Thus our special $\boldsymbol{R}^3$ has the usual Euclidean metric on it, coming from the identification with the usual $\boldsymbol{R}^3$.

Given $p \in \boldsymbol{R}^3$ we define

$$R_q(p) = qpq^{-1}.$$

Exercise 10 (Challenge). Show that R_q preserves $\boldsymbol{R}^3$ (the pure quaternions) and is an orientation-preserving rotation.

Multiplication turns out to be associative and so we have

$$R_{q_1} \circ R_{q_2}(p) = q_1(q_2pq_2^{-1})q_1^{-1} = R_{q_1} \circ R_{q_2}(p).$$

This works for any p. Hence the map $q \to R_q$ is a homomorphism. As you might expect, we define $E(q) = R_q$. Note that $E(-q) = E(q)$. It turns out that the kernel of E is precisely $\{1, -1\}$. So, E is both a continuous surjection (with good local inverse properties) and a two-to-one homomorphism from S^3 to $SO(3)$.

Now for our last example. Given the quaternionic picture of S^3, we can define a very interesting 3-dimensional manifold. If $G \subset SO(3)$ is a finite subgroup, then $\widetilde{G} = E^{-1}(G)$ is a subgroup with twice the number of elements. Now we can define an equivalence on S^3 by the rule $q_1 \sim q_2$ iff there exists some $g \in \widetilde{G}$ such that $gq_1 = q_2$. If G has N elements, then $\widetilde{G}$ has $2N$ elements and each equivalance class of $S^3/\sim$ has $2N$ elements. It turns out the quotient space is a manifold with fundamental group $\widetilde{G}$.

As a special case, let G be the orientation-preserving symmetries of the icosahedron, the most interesting finite subgroup of $SO(3)$. Then $\widetilde{G}$ is an order 120 group known as the *binary icosahedral group*. The quotient in this case is called the *Poincaré homology sphere*, and its fundamental group is $\widetilde{G}$.

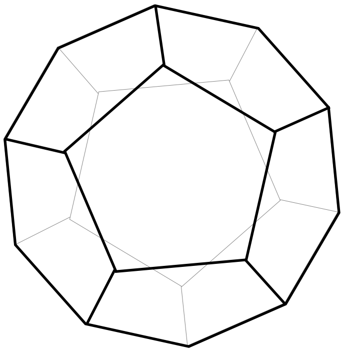

Figure 5.2. A Dodecahedron

The Poincaré homology sphere is one of the great examples in geometry. In the lens space example, S^3 is tiled by 5 copies of a kind of "double lens". In the Poincaré homology sphere example, it turns out that S^3 is tiled by 120 spherical dodecahedra. The spherical dodecahedra look combinatorially the same as Euclidean dodecahedra, but they are "puffed-out" much in the same way that spherical triangles are. Figure 5.2 shows a dodecahedron, drawn so that the thick lines represent visible edges and the thin lines represent hidden edges.

Any point in S^3 is equivalent to a point in one of these dodecahedra, and no two points in the interior of a dodecahedron are equivalent to each other. Thus, analyzing the Poincaré homology sphere boils down to understanding how points on the boundary of one of the dodecahedra are glued together. What happens is that each face of the dodecahedron is glued to the opposite face, with a $2\pi/5$ twist.

6 Covering Spaces and the Deck Group

In §1.4 we discussed how the process of "unwrapping" the essential loops on the square torus leads naturally to the integer grid in the plane. We also mentioned that something similar can be done for the octagon surface and its relatives. The purpose of this chapter and the next one is to make the unwrapping process precise, and to consider it in much greater generality. The central objects in this chapter are covering spaces and the deck group, objects which play the role that the plane and the integer grid, respectively, played in §1.4. Along the way, we will relate covering spaces and the deck group to the fundamental group.

6.1 Covering Spaces

Let $\widetilde{X}$ and X be path connected metric spaces. Let $E : \widetilde{X} \to X$ be a continuous map. An open set $U \subset X$ is said to be *evenly covered* if the preimage $E^{-1}(U)$ consists of a countable disjoint union of sets $\widetilde{U}_1, \widetilde{U}_2, \ldots$ such that the restriction $E : \widetilde{U}_j \to U$ is a homeomorphism. (This makes sense because $\widetilde{U}_j$ is a metric space in its own right.) It is customary to require that U is path connected in this definition. The sets $\widetilde{U}_j$ are called *components* of the pre-image. The map E is said to be a *covering map* if every point in X has a neighborhood that is evenly covered. In this case, $\widetilde{X}$ is said to be a covering space of X.

The "mother of all examples" is the map $E : \boldsymbol{R} \to S^1$ discussed in §5.1. Here we will describe this map in another way. We still think of the line as $\boldsymbol{R}$, but now we think of the circle as the space X obtained from $[0,1]$ by gluing 0 to 1. This time, our map E is given by $E(x) = [x - \text{floor}(x)]$. Here floor$(x)$ is the greatest integer less or equal to x. So, $x - \text{floor}(x)$ is the fractional part of x. Finally, $E(x)$ is the equivalence class of the fractional part of x. The map E is continuous even though it does not appear to be so. If x_1 is sligtly smaller than an integer and x_2 is slightly larger than the same integer, then $E(x_1)$ and $E(x_2)$ are on opposite sides of $[0,1]$. However, the gluing brings them close together in X.

Exercise 1. Verify that $E : \boldsymbol{R} \to S^1$ is indeed a covering map in the example(s) given above. Reconcile the two examples and see that essentially they are the same thing.

6.2 The Deck Group

We are going to give more examples of covering spaces below, but the whole idea of a covering space is enhanced by another concept–the deck group. So, we will bring up the deck group before talking more about covering spaces. We have already associated one group to a (pointed) metric space, namely the fundamental group. Now we are going to assign a group in a second way. Let $E : \widetilde{X} \to X$ be a covering map as above. Say that a *deck transformation* is a homeomorphism $h : \widetilde{X} \to \widetilde{X}$ such that $E \circ h = E$.

As a mnemonic, think about how the deck group relates to shuffling a deck of cards. There is a natural map E, from your deck of cards to a single card. You can think of holding the deck of cards directly above the single card and then E is vertical projection. If you shuffle the cards and redo the map E there is no change. So, a deck transformation in this case corresponds to shuffling the deck.

In general, you can think of $\widetilde{X}$ as a kind of deck of cards and X as a single card. The analogy isn't perfect because $\widetilde{X}$ is connected, but for an evenly covered neighborhood $U \subset X$, the set $\widetilde{U} = E^{-1}(U)$ really is like a deck of cards. The deck transformation h somehow permutes the disjoint components of $\widetilde{U}$ like shuffling permutes the cards.

If h is a deck transformation, so is h^{-1}. Likewise if h_1 and h_2 are deck transformations, then so is $h_1 \circ h_2$. Thus, the set of deck transformations forms a group under composition. This group is called the deck group of $(\widetilde{X}, X, E)$.

Let's revisit our covering space example considered in the last section. In both examples, the transformation $x \to x + 1$ is a covering transformation. In the first case, this follows from the identity $\exp(2\pi i(x + 1)) = \exp(2\pi i x)$. In the second example, it is obvious from the definition of E.

Exercise 2. Verify that the deck group in the above example is $\boldsymbol{Z}$. In other words, the maps $x \to x + n$ for $n \in \boldsymbol{Z}$ are the only covering transformations.

Note the deck group of $(\boldsymbol{R}, S^1, E)$ is $\boldsymbol{Z}$, the same as $\pi_1(S^1)$; that is, the deck group and the fundamental group are isomorphic. Below we will prove a result that gives general conditions under which this is true.

6.3 A Flat Torus

The next really great example of a covering space is $E : \boldsymbol{R}^2 \to X$, where X is a *flat torus*. As discussed in §1.9, we can make a flat torus X by gluing together the opposite sides of a parallelogram P_0, as shown in Figure 6.1.

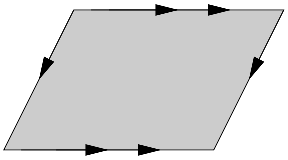

Figure 6.1. The flat torus

The resulting surface X is homeomorphic to $S_1 \times S_1$, and the fundamental group is isomorphic to $\boldsymbol{Z}^2$. There is a nice covering map from $\boldsymbol{R}^2$ to X. We can tile $\boldsymbol{R}^2$ with translates of P_0, as shown in Figure 6.2. Given any point $x \in \boldsymbol{R}^2$, we choose a parallelogram P_x such that $x \in P_x$. There is a unique translation $T_x : P_x \to P_0$ and we define $E(x) = [T_x(x)] \in X$.

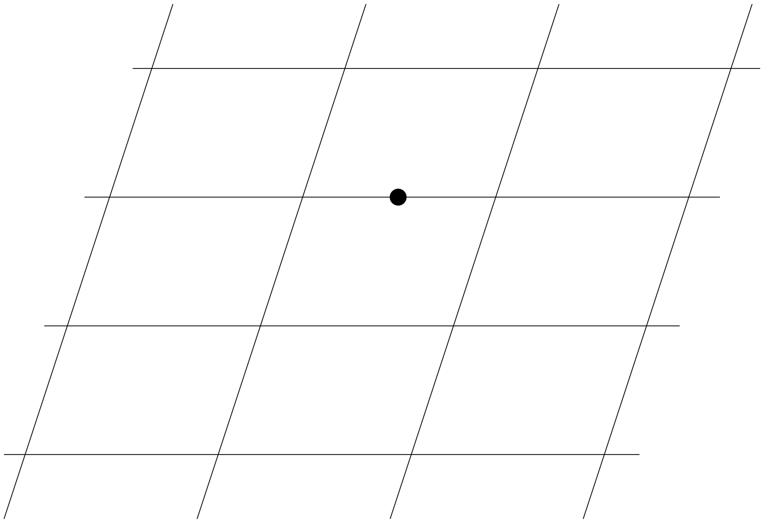

Figure 6.2. The parallelogram tiling

The beautiful thing about this map is that it is well defined even when x lies on the interface between two or more parallelograms. For example, suppose that x lies on a horizontal edge, as shown in Figure 6.2. Then we could take P_x to be the parallelogram either above x or below x. In the one case $T_x(x)$ would like in the middle of the top edge of P_0 and in the other case $T_x(x)$ would lie in the middle of the bottom edge of P_0. However, these two points are identified on X.

Exercise 3. Prove that $E : \boldsymbol{R}^2 \to X$ is a covering map and that the deck group in this case is precisely the group of translation symmetries of the tiling, namely $\boldsymbol{Z}^2$. Once again, the deck group and the fundamental group are isomorphic.

There are a few things about the flat torus example that do not quite represent the general case. For instance, the deck group and fundamental group are both Abelian, and this is rather a special situation. However, in spite of the limitations of the torus example, I would say that it accounts for 80 percent of my intuition about covering spaces. In any case, it is a good example to learn well! The example in Exercise 5 below accounts for another 19 percent of my intuition, and then the last 1 percent comes from more complicated examples.

6.4 More Examples

Here are two more examples of covering spaces and deck groups. In the next example, the fundamental group and the deck group are nontrivial finite groups.

Exercise 4. Let S^2 be the 2-sphere and let $\boldsymbol{P}^2$ be the projective plane, defined as the set of equivalence classes of antipodal points on S^2. Show that the obvious map $S^2 \to \boldsymbol{P}^2$ is a covering map. (*Note*: In order to do this problem, you first have to recall the metric on $\boldsymbol{P}^2$.) Show that the deck group in this example is $\boldsymbol{Z}/2$. Once again, the deck group and the fundamental group are isomorphic.

So far, all the examples we have seen have Abelian deck groups. The next exercise shows an important example in the case when the group is not Abelian.

Exercise 5. Let X be a space that is homeomorphic to an ∞ symbol, as shown on the right-hand side of Figure 6.3 below. Let $\widetilde{X}$ be the 4-valent infinite tree. Exhibit a map $E : \widetilde{X} \to X$ which is a covering map. (The tree in Figure 6.3 is only partially drawn. It is meant to go on forever and have valence 4 at each vertex.) Prove that the deck group for $(\widetilde{X}, X, E)$ is isomorphic to the free group on 2 generators. Once again, the deck group and the fundamental group are isomorphic.

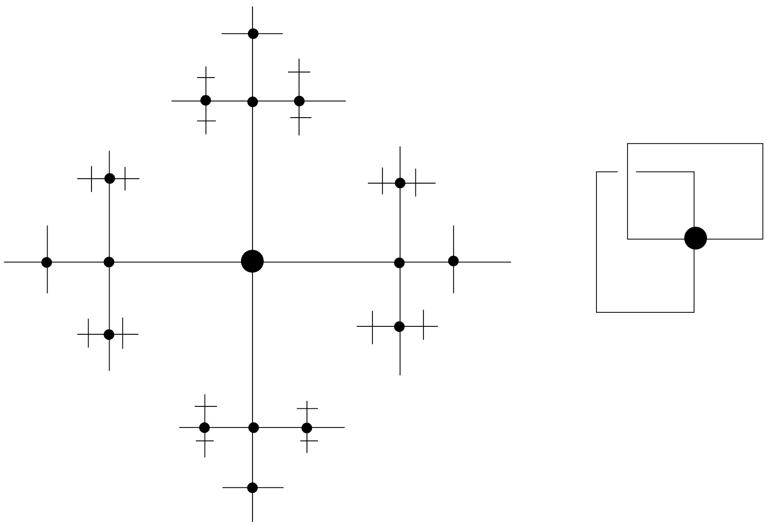

Figure 6.3. The 4-valent tree and the figure 8

6.5 Simply Connected Spaces

Recall that a path connected space is one in which every two points can be joined by a continuous path. Let X be a path connected metric space. X is said to be *simply connected* if $\pi_1(X)$ is trivial. This definition does not depend on the basepoint, because the isomorphism type of the fundamental group is independent of basepoint in path connected spaces. The plane is simply connected and so is a tree.

Suppose that $f_0, f_1 : [0,1] \to X$ are two paths. Suppose also that $f_0(0) = f_1(0)$ and $f_1(0) = f_1(1)$. In other words, the two paths have the same beginning and the same ending. We say that f_0 and f_1 are *path homotopic* if there is a homotopy F from f_0 to f_1 such that $f_t(0)$ and $f_t(1)$ are independent of t. Here, as usual, $f_t(x) = F(x,t)$, where F is a map on the unit square. Intuitively, a path homotopy slides the one path to the other without moving the endpoints. In the case where $f_t(0) = f_t(1)$, the notion of a path homotopy coincides with the notion of a loop homotopy.

The next exercise relates the idea of a path homotopy to the idea of simple connectivity.

Exercise 6. Suppose that X is simply connected. Prove that any two paths, which have the same endpoints as each other, are homotopic. (*Outline*: Let x be the starting point of both loops. Consider the loop g formed by first doing f_0 forward and then doing f_1 backward. Then $[g] \in \pi_1(X, x)$. Hence g is loop homotopic to the identity. Let G be the loop homotopy. Try to modify G slightly so that G becomes a path homotopy from f_0 to f_1. Figure 6.4 shows what we hope is a suggestive picture.)

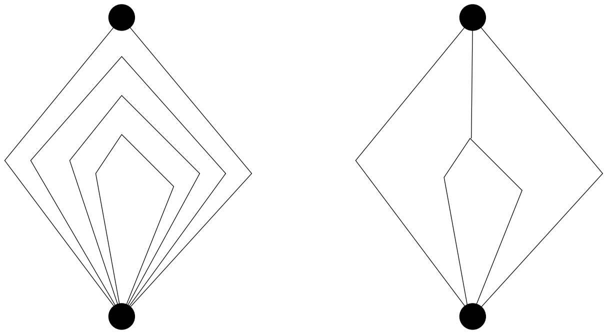

Figure 6.4. altering a homotopy

Exercise 7. Let $\{B_i\}$ denote any countable union of disjoint closed balls in $\boldsymbol{R}^3$. Prove that $\boldsymbol{R}^3 - \bigcup B_i$ is simply connected.

6.6 The Isomorphism Theorem

Here is the main theorem in this chapter, and (in my opinion) one of the best theorems in algebraic topology.

Theorem 6.1 (Isomorphism) *Suppose that*

- $E : \widetilde{X} \to X$ *is a covering map.*
- X *and* $\widetilde{X}$ *are path connected.*
- $\widetilde{X}$ *is simply connected.*

Then $\pi_1(X)$ *is isomorphic to the deck group for* $(\widetilde{X}, X, E)$.

The rest of the chapter is devoted to proving the Isomorphism Theorem.

6.7 The Bolzano–Weierstrass Theorem

A sequence of points $\{c_j\}$ in a metric space X is called *Cauchy* if, for every $\epsilon > 0$, there is some N such that $i, j > N$ implies that $d(c_i, c_j) < \epsilon$. A convergent sequence is automatically Cauchy, and one can ask about the converse. X is said to be *complete* if every Cauchy sequence in X converges to a point in X.

Exercise 8. Prove that $\boldsymbol{Q}$, the rationals, is not complete.

The basic axiom for $\boldsymbol{R}$ is that it is complete. You might ask how one proves that $\boldsymbol{R}$ is complete. The usual way is to construct $\boldsymbol{R}$ from $\boldsymbol{Q}$ in a way that builds in completeness. Here is the barest sketch of the idea. Start with the set X of all Cauchy sequences in $\boldsymbol{Q}$. Define two Cauchy sequences $\{a_i\}$ and $\{b_i\}$ to be *equivalent* if the shuffled sequence $a_1, b_1, a_2, b_2, a_3, b_3, \ldots$ is also a Cauchy sequence. Intuitively, equivalent sequences (were they to converge) have the same limit. $\boldsymbol{R}$ is defined as the set of equivalence classes in X. Cauchy sequences are added, subtracted, multiplied, and (when possible) divided term by term, and you have to check that these operations respect the equivalence relation.

Exercise 9. Using the completeness of $\boldsymbol{R}$ as an axiom, prove the following result. Let $Q_1 \supset Q_2 \supset Q_3 \cdots$ be a nested sequence of cubes in $\boldsymbol{R}^n$ such that the diameter of Q_n tends to 0 as n tends to ∞. Then $\bigcap Q_n$ is one point. (*Hint*: look at the sequence of centers.)

Theorem 6.2 (Bolzano–Weierstrass) *A sequence $\{c_n\}$ contained in the unit cube Q_0 has a convergent subsequence.*

Proof: Note that Q_0 is the union of 2^n cubes having half the size as Q_0. At least one of these subcubes, Q_1, must contain c_j for infinitely many indices. But Q_1 is a union of 2^n subcubes having half the size as Q_1. At least one of these subcubes, Q_2, must contain c_j for infinitely many indices. And so on. The intersection $\bigcap Q_n$, a single point, by Exercise 2, is the limit of some subsequence of $\{c_j\}$. ♠

6.8 The Lifting Property

In this section, $E : \widetilde{X} \to X$ is a covering map. Let Q be a cube, and let $f : Q \to X$ be a continuous map. We say that a *lift* of f is a map $\widetilde{f} : Q \to \widetilde{X}$ such that $E \circ \widetilde{f} = f$. This notion is just a generalization of what we talked about in the previous chapter. The purpose of this section is to prove the formal version of the result we talked about, for some examples, in the previous chapter.

We begin with a technical result.

Lemma 6.3 *There is some N with the following property. If $Q' \subset Q$ is a subcube with side length less than $1/N$, then $f(Q)$ is contained in an evenly covered neighborhood of X.*

Proof: If this result is false, then we can find a sequence of subcubes $\{Q_n\}$, with diameter tending to 0 such that $f(Q_j)$ is not contained in an evenly covered neighborhood. Let $\{c_j\}$ be the center of Q_j. This sequence has a convergent subsequence, by the Bolzano–Weierstrass Theorem. Tossing out everything but the cubes corresponding to this subsequence, we can assume that $\{c_j\}$ converges to some $x \in Q$. Then $f(x)$ is contained in some evenly covered neighborhood U. But then $f(Q_n) \subset U$ for n large, by continuity. This is a contradiction. ♠

Lemma 6.4 *Let Q be a cube, and let $f : Q \to X$ be a continuous map. Let v be a vertex of Q, and let $\widetilde{x} \in X$ be a point such that $E(\widetilde{x}) = f(v)$. Suppose that $f(Q)$ is contained in an evenly covered neighborhood. Then there is a unique lift $\widetilde{f} : Q \to \widetilde{X}$ such that $\widetilde{f}(v) = \widetilde{x}$.*

Proof: Let $U \subset X$ be the evenly covered neighborhood such that $f(Q) \subset U$. Recall that $E^{-1}(U)$ is a disjoint union of sets $\widetilde{U}_1, \widetilde{U}_2, \ldots$ such that the restriction $E : \widetilde{U}_j \to U$ is a homeomorphism. Let $\widetilde{U}_k$ be the component that contains $\widetilde{x}$, and let F be the inverse of the restriction of E to $\widetilde{U}_k$. Then we can and must define $\widetilde{f} = F \circ f$. ♠

Just as we did in the previous chapter, we want to now remove the hypothesis that $f(Q)$ is contained in an evenly covered neighborhood.

Theorem 6.5 *Let Q be a cube and let $f : Q \to X$ be a continuous map. Let v be a vertex of Q, and let $\widetilde{x} \in \widetilde{X}$ be such that $E(\widetilde{x}) = f(v)$. Then there is a unique lift $\widetilde{f} : Q \to \widetilde{X}$ such that $\widetilde{f}(v) = \widetilde{x}$.*

Proof: By Lemma 6.3, we can find some N such that any subcube of diameter less than N is mapped into an evenly covered neighborhood of f. Let's partition Q into such cubes, say $Q = Q_1, \ldots, Q_m$. We can order these cubes so that, for each k, the cube Q_k shares at least one vertex v_k with some Q_j for $j < k$. Also, we set things up so that the initial vertex $v = v_1$ is a vertex of Q_1. We define $\widetilde{f}$ on Q_1 using Lemma 6.4. This tells us the value of $\widetilde{f}$ on v_2 and determines how we define $\widetilde{f}$ on Q_2. The uniqueness guarantees that the definition on Q_2 is compatible with the definition on Q_1. The key point is that $Q_1 \cap Q_2$ is contained in an evenly covered neighborhood. We continue like this, from cube to cube, until we have defined $\widetilde{f}$ in the only way possible on all of Q. ♠

We will only need the above result for the case of the unit interval $[0, 1]$ and the unit square $[0, 1]^2$, but it is nice to know in general.

6.9 Proof of the Isomorphism Theorem

The proof comes in 4 steps:

1. Define the isomorphism.

2. Prove that it is a homomorphism.

3. Prove that the homomorphism is injective.

4. Prove that the homomorphism is surjective.

6.10 Define the Isomorphism

Since X is path connected, the isomorphism type of $\pi_1(X, x)$ is independent of the choice of basepoint. Let $x \in X$ be a basepoint. Let $G = \pi_1(X)$. Let D be the deck transformation group. Let $\widetilde{x} \in \widetilde{X}$ be some point such that $E(\widetilde{x}) = x$. We make this choice once and for all. Suppose that $h \in D$ is a deck transformation. Then $\widetilde{y} = h(\widetilde{x})$ is some other point. Note that

$$E(\widetilde{y}) = E \circ h(\widetilde{x}) = E(\widetilde{x}) = x.$$

Since $\widetilde{X}$ is path connected, there is some path $\widetilde{f} : [0,1] \to \widetilde{X}$ such that $\widetilde{f}(0) = \widetilde{x}$ and $\widetilde{f}(1) = \widetilde{y}$. Let $f = E \circ \widetilde{f}$. By construction, f is a loop based at f. Define

$$\Phi(h) = [f] \in G. \tag{5}$$

To see that Φ is well defined, suppose that $\widetilde{f}_0$ and $\widetilde{f}_1$ are two loops connecting $\widetilde{x}$ to $\widetilde{y}$. Since $\widetilde{X}$ is simply connected, there is a path homotopy $\widetilde{F}$ from $\widetilde{f}_0$ to $\widetilde{f}_1$. But then $F = E \circ \widetilde{F}$ is a loop homotopy from f_0 to f_1. Hence $[f_0] = [f_1]$ and Φ is well defined. Φ is our map from D to G.

6.11 Homomorphism

This step looks quite mysterious, but it is fairly obvious if you draw pictures. Let $h_1, h_2 \in D$ be two deck transformations. We want to prove that

$$\Phi(h_1 \circ h_2) = \Phi(h_1)\Phi(h_2).$$

Let $\widetilde{y}_j = h_j(\widetilde{x})$ for $j = 1, 2$. Let $\widetilde{f}_j$ be a path joining $\widetilde{x}$ to $\widetilde{y}_j$. Let $f_j = E \circ \widetilde{f}_j$. Then $\Phi(h_j) = [f_j]$.

Let $\widetilde{z} = h_1 \circ h_2(\widetilde{x})$. Note that $h_1 \circ \widetilde{f}_2$ is a path joining the points

$$h_1(\widetilde{x}) = \widetilde{y}_1 \qquad \text{and} \qquad h_1(\widetilde{y}_2) = h_1 \circ h_2(\widetilde{x}).$$

Therefore, the concatenated path $\widetilde{f}_1 * (h_1 \circ \widetilde{f}_2)$ joins $\widetilde{x}$ to $\widetilde{z}$. But then

$$\Phi(h_1 \circ h_2) = [E \circ (\widetilde{f}_1 * (h_1 \circ \widetilde{f}_2))] = [(E \circ \widetilde{f}_1) * (E \circ h_1 \circ \widetilde{f}_2)]$$

$$=^* [(E \circ \widetilde{f}_1) * (E \circ \widetilde{f}_2)] = [f_1 * f_2] = [f_1][f_2] = \Phi(h_1)\Phi(h_2).$$

The starred equality comes from the fact that $E \circ h_1 = E$.

Exercise 10. Choose the example of the flat torus, given above, and go through the above argument step by step, illustrating the proof with pictures.

6.12 Injectivity

Since Φ is a homomorphism, we can show that Φ is injective just by showing that Φ has a trivial kernel. So, suppose that $\Phi(h)$ is the trivial element in $\pi_1(X, x)$.

Lemma 6.6 $h(\widetilde{x}) = \widetilde{x}$.

Proof: Let $\widetilde{y} = h(\widetilde{x})$. We want to show that $\widetilde{y} = \widetilde{x}$. Let $\widetilde{f}$ be a path which joins $\widetilde{x}$ to $\widetilde{y}$. It suffices to show that f is path homotopic to the constant path. Let $f = E \circ \widetilde{f}$. Then $\Phi(h) = [f]$. By hypothesis, there is a loop homotopy F from f to the trivial loop. Let Q be the unit square. By construction, $F : Q \to X$ is a continuous map such that $f_0 = f$ and f_1 is the constant map. From the lifting theorem, there is a lift $\widetilde{F} : Q \to \widetilde{X}$ such that $\widetilde{F}(0,0) = \widetilde{x}$ and $E \circ \widetilde{F} = F$. Here are 3 properties of $\widetilde{F}$:

- $\widetilde{f}_0$ is a lift of $f_0 = f$. From the uniqueness of lifts, $\widetilde{f}_0 = \widetilde{f}$.
- $\widetilde{f}_1$ is the constant path since f_1 is the constant path.
- $F(0,t)$ and $F(1,t)$ are the basepoint in X, independent of t. Hence $\widetilde{F}(0,t)$ and $\widetilde{F}(1,t)$ are constant maps. That is, the endpoints of $\widetilde{f}_t$ do not change with t.

From the first item, the endpoints of $\widetilde{f}_0$ are $\widetilde{x}$ and $\widetilde{y}$. From the second item, the endpoints of $\widetilde{f}_1$ are $\widetilde{x}$ and $\widetilde{x}$. From the third item, we see that the two sets of endpoints coincide, forcing $\widetilde{x} = \widetilde{y}$. ♠

The following lemma finishes our injectivity proof.

Lemma 6.7 *If h is a deck transformation such that $h(\widetilde{x}) = \widetilde{x}$, then h is the identity.*

Proof: Let $\widetilde{y}$ be some other point of $\widetilde{X}$. We want to show that $h(\widetilde{y}) = \widetilde{y}$. Let $\widetilde{f}$ be a path joining $\widetilde{x}$ to $\widetilde{y}$. Let $x = E(\widetilde{x})$ and $y = E(\widetilde{y})$. Let $f = E \circ \widetilde{f}$. Then $f : [0,1] \to X$ is a path which joins x to y.

The paths $\widetilde{f}$ and $h \circ \widetilde{f}$ are both lifts of f which agree at 0. That is, $\widetilde{f}(0) = \widetilde{x}$ and $h \circ \widetilde{f}(0) = h(\widetilde{x}) = \widetilde{x}$. By uniqueness of lifts, these two lifts are the same. In particular, $\widetilde{y} = \widetilde{f}(1) = h \circ \widetilde{f}(1) = h(\widetilde{y})$. ♠

6.13 Surjectivity

Let $[g] \in \pi_1(X, x)$ be some element. We want to produce a deck transformation h such that $\Phi(h) = [g]$. Let $\widetilde{y} \in \widetilde{X}$ be any point. We need to define $h(\widetilde{y})$. So, let $\widetilde{f}$ be a path joining $\widetilde{x}$ to $\widetilde{y}$. Let $f = E \circ \widetilde{f}$. Then f is a path in X joining x to $y = E(\widetilde{y})$. Consider the concatenated path $\gamma = g * f$. From the lifting property we can find a lifted path $\widetilde{\gamma}$ which joins $\widetilde{x}$ to some other point, which we define as $h(\widetilde{y})$. Figure 6.5 illustrates the construction in case $\widetilde{X} = \boldsymbol{R}^2$ and $X = T^2$, the torus.

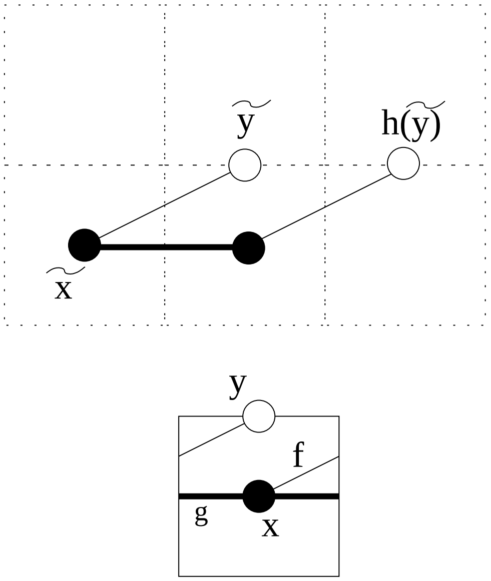

Figure 6.5. Lifted paths

Exercise 11. Show that the definition of $h(\widetilde{y})$ is independent of the choices of f and g. (*Hint*: imitate the proof given in the previous section.)

To compute $\Phi(h)$, we consider the case that $\widetilde{y} = \widetilde{x}$. Then we can take $\widetilde{f}$ to be the trivial path. In this case $\widetilde{\gamma}$ is a path joining $\widetilde{x}$ to $h(\widetilde{x})$ and $E \circ \widetilde{\gamma}$ differs from $g = E \circ \widetilde{g}$ just by concatenating the constant loop. Assuming that h is a deck transformation, we have $\Phi(h) = [\gamma] = [g]$.

To finish the proof, we just have to show that h is a deck transformation.

Lemma 6.8 *$E \circ h = h$.*

Let's compute $E \circ h(\widetilde{y})$. By construction, both γ and f connect x to y. We have

$$E \circ h(\widetilde{y}) =^1 E \circ \widetilde{\gamma}(1) = \gamma(1) = y = f(1) = E \circ \widetilde{f}(1) =^2 E(\widetilde{y}).$$

Equality 1 comes from the fact that $\widetilde{\gamma}(1) = h(\widetilde{y})$ by definition. Equality 2 comes from the fact that $\widetilde{f}(1) = \widetilde{y}$, by definition. ♠

Lemma 6.9 *h is continuous.*

Proof: Let $\widetilde{y} \in \widetilde{X}$ be a point. Let $y = E(\widetilde{y})$. There is an evenly covered neighborhood $U \subset X$ of y. Let $\widetilde{U} - 1$ be the component of $h^{-1}(U)$ which contains $\widetilde{y}$. Let $\widetilde{U}_2 = h(\widetilde{U}_1)$. Then $\widetilde{U}_2$ is another component of $h^{-1}(U)$ because $E \circ h = E$. Let F_j be the inverse of the restriction of E to $\widetilde{U}_j$. Then $h = F_2 \circ E$ on $\widetilde{U}_1$. Being the composition of continuous maps, h is continuous. ♠

Were we to make the above construction for the element $[g]^{-1}$, we would produce the map h^{-1}. Hence h is invertible. The same argument as above shows that h^{-1} is continuous. Hence h is a homeomorphism. Now we know that h belongs to the deck group. This completes our proof.

7 Existence of Universal Covers

In the previous chapter, we proved the Isomorphism Theorem, a result which relates the triple $(\widetilde{X}, X, E)$ to the fundamental group $\pi_1(X)$. Here $\widetilde{X}$ is a simply connected covering space of X and $E : \widetilde{X} \to X$ is a covering map. $\widetilde{X}$ is known as *the universal cover* of X. We use the word "the" because, as it turns out, any two universal covering spaces of X are homeomorphic to each other.

In this chapter we will prove the existence (but not uniqueness) of a universal cover $\widetilde{X}$ under certain assumptions on X. The conditions we place on X are somewhat contrived, but we want to streamline the existence proof. Our main interest in this result is the case when X is a compact surface, and any compact surface satisfies the conditions we impose.

The reader interested in seeing the fully general existence and uniqueness proof should consult an algebraic topology book such as [**HAT**]. The exact condition on X that guarantees the existence of $\widetilde{X}$ is that X is *semilocally simply connected*, and in all such cases $\widetilde{X}$ is unique.

7.1 The Main Result

Given a metric space M and two continuous paths $f_0, f_1 : [0,1] \to M$, we define

$$D(f_0, f_1) = \sup_{t \in [0,1]} d(f_0(t), f_1(t)). \tag{6}$$

Let $x \in M$. We say that the pair (M, x) is *conical* if, for each $y \in M$ there is a continuous path $\gamma_y : [0,1] \to \boldsymbol{M}$ such that $\gamma_y(0) = x$ and $\gamma_y(1) = y$. We insist that γ_x is the trivial path, and also we make the following continuity requirement. For any $y \in M$ and any $\epsilon > 0$, there is some $\delta > 0$ such that $d(y, z) < \delta$ implies that $D(\gamma_y, \gamma_z) < \epsilon$.

The idea behind our definition is that you are making M into a kind of cone, with x as the apex. The pair $(\boldsymbol{R}^n, 0)$ is a prototypical example of a conical pair. The paths you can use in this example are just line segments traced out at unit speed.

Exercise 1. Prove (M, x) is conical if M is homeomorphic to $\boldsymbol{R}^n$.

Say that the path f_0 in M is *good* if there is some $\epsilon > 0$ with the following property: Suppose that $D(f_0, f_1) < \epsilon$ and f_0 and f_1 have the same endpoints. Then there is a homotopy F from f_0 to f_1 which does not move the endpoints. That is, $f_t(0)$ and $f_t(1)$ are independent of t.

Definition 7.1. A metric space X is *good* if every path in X is good and every point $x \in X$ is such that the ball $B_\epsilon(x)$ is both simply connected and conical for some $\epsilon > 0$. The value of ϵ is allowed to vary with the point and the path.

Here is our main result.

Theorem 7.1 *Any good metric space has a simply connected cover.*

Exercise 2. Prove that a flat torus is good.

Exercise 3. Prove that any finite graph is good.

Exercise 4 (Challenge). Prove that any compact surface is good. (*Hint*: In the proof of Theorem 12.10 we sketch the argument for complete hyperbolic surfaces.)

Exercise 5. Give an example of a metric space that has no nontrivial good paths. (*Hint*: swiss cheese.)

Here is the construction of $\widetilde{X}$ and the map $E : \widetilde{X} \to X$. Choose a basepoint $x \in X$. We define $\widetilde{X}$ to be the set of pairs $(y, [f])$ where $y \in X$ is a point and f is a path which joins x to y. Here $[f]$ denotes the path homotopy equivalence class of f. That is, $[f_1] = [f_2]$ if and only if there is a homotopy from f_1 to f_2 that does not move the endpoints.

So far $\widetilde{X}$ is just a set. We define

$$D([f_0], [f_1]) = \inf D(f_0, f_1). \tag{7}$$

The infimum is taken over all paths f_0 which represent $[f_0]$ and all paths f_1 which represent $[f_1]$. Finally, we define

$$\widetilde{d}((y_0, [f_0]), (y_1, [f_1])) = d(y_0, y_1) + D([f_0], [f_1]). \tag{8}$$

Exercise 6. Prove that $\widetilde{d}$ is a metric on $\widetilde{X}$. (*Hint*: The only hard part of this exercise is showing that $\widetilde{d}(p, q) = 0$ implies $p = q$. Here $p, q \in \widetilde{X}$. This amounts to showing that $D([f_0],[f_1]) = 0$ implies that $[f_0] = [f_1]$. Deduce this from the goodness of X.)

There is an obvious map $E : \widetilde{X} \to X$, given by $E(y,[f]) = y$. There are a few things about E that we can see right away. Since E does not increase distances, E is a continuous map. Also, E is onto because X is path connected.

Exercise 7. Use the fact that X is path connected to prove that $\widetilde{X}$ is also path connected.

It remains to prove that E is a covering map and that $\widetilde{X}$ is simply connected. We will prove these two statements in the next two sections.

7.2 The Covering Property

Let $y \in X$ be a point, and let U be an ϵ-ball about y, chosen to be both simply connected and conical. Let H denote the set of path homotopy classes of curves joining x to y. We first produce a homeomorphism Ψ from $E^{-1}(U)$ to $U \times H$. This is a formal way of saying that $E^{-1}(U)$ is a disjoint union of copies of U.

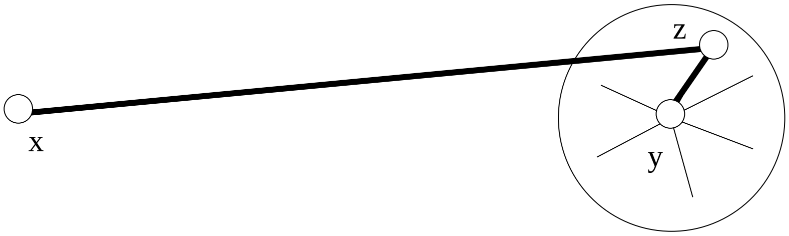

Figure 7.1. The path $f * \gamma(z,y)$

For any $z \in U$, let $\gamma(y,z)$ be the path joining y to z, as specified by the definition of a conical metric space. Let $\gamma(z,y)$ denote the reverse path. Let $(z,[f]) \in E^{-1}(U)$ be a point. We define

$$\Psi((z,[f])) = (z,[f * \gamma(z,y)]). \tag{9}$$

See Figure 7.1. If f_0 and f_1 are both representatives of $[f]$, then a path homotopy from f_0 to f_1 extends to a path homotopy from $f_0 * \gamma$ to $f_1 * \gamma$. Hence $[f_0 * \gamma] = [f_1 * \gamma]$. Hence, our map Ψ is well defined.

Lemma 7.2 Ψ *is a bijection.*

Proof: Suppose $\Psi(z_0,[f_0]) = \Psi(z_1,[f_1])$. Then $z_0 = z_1$. We set $z = z_0 = z_1$. We know that $[f_0 * \gamma(z,y)] = [f_1 * \gamma(z,y)]$. Writing $\gamma = \gamma(z,y)$, we have $[f_0 * \gamma] = [f_1 * \gamma]$ but then

$$[f_0] = [f_0 * \gamma * \gamma^{-1}] = [f_1 * \gamma * \gamma^{-1}] = [f_1].$$

This shows that Ψ is injective.

Now we show that Ψ is surjective. Given any pair $(z,[g]) \in U \times H$, the path $f = g * \gamma(y,z)$ connects x to z. The two paths g and

$$f * \gamma(z,y) = g * \gamma(y,z) * \gamma(z,y)$$

are clearly homotopic. Hence $\Psi(z,[f]) = (z,[g])$. ♠

We put a metric on $U \times H$ by declaring that points in different components are 1 apart. Within a single component, $U \times \{h\}$, we just use the metric we already have on U.

Lemma 7.3 Ψ *is a homeomorphism.*

Proof: We already know that Ψ is a bijection. We just have to show that Ψ and Ψ^{-1} are both continuous. We will consider Ψ. Suppose that $(z_0,[f_0])$ and $(z_1,[f_1])$ are very close. Then $f_0 * \gamma(z_0,y)$ and $f_1 * \gamma(z_1,y)$ are two very nearby paths, both having endpoints x and y. Since X is good, we have

$$[f_0 * \gamma(z_0,y)] = [f_1 * \gamma(z_1,y)]$$

once these paths are sufficiently close. Also z_0 and z_1 are very close. So, the second coordinates of $\Psi(z_0,[f_0])$ and $\Psi(z_1,[f_1])$ agree, and the first coordinates are very close. This shows (a bit informally) that Ψ is continuous.

Now we consider Ψ^{-1}. Using the notation from the proof of the previous lemma, we have

$$\Psi^{-1}(z,[g]) = (z,[f]),$$

where $f = g \circ \gamma^{-1}$. If $(z_0,[g_0])$ and $(z_1,[g_1])$ are less than 1 apart, then $[g_0] = [g_1]$. But then, we can use the same path g to represent both $[g_0]$ and $[g_1]$. But then $f_0 = g * \gamma(z_0,y)^{-1}$ and $f_1 = g * \gamma(z_1,y)^{-1}$ are also close. This shows that Ψ^{-1} is continuous. ♠

Now we know that Ψ is a homeomorhism from $E^{-1}(U)$ to $U \times H$. Let $\pi : U \times H \to U$ be projection onto U. Then the restriction of π to each component of $U \times H$ is clearly a homeomorphism. These components are of the form $U \times \{h\}$, where $h \in H$.

Finally, note that

$$E = \pi \circ \Psi. \tag{10}$$

For each component $\widetilde{U}$ of $E^{-1}(U)$ there is some $h \in H$ so that $\Psi(\widetilde{U}) = U \times \{h\}$ and Ψ is a homeomorphism from $\widetilde{U}$ to $U \times \{h\}$. But then the restriction to $\widetilde{U}$ of $E = \pi \circ \Psi$ is the composition of two homeomorphisms, and hence a homeomorphism. This completes the proof that E is a covering map.

7.3 Simple Connectivity

We take the basepoint $\widetilde{x} \in \widetilde{X}$ to be the pair $(x, *)$ where $*$ is the trivial loop connecting x to x. Suppose $f : [0,1] \to \widetilde{X}$ is a loop. This means that $f(t) = (x_t, [\gamma_t])$, where $x_t \in X$ and γ_t is a path connecting x to x_t. Both $[\gamma_0]$ and $[\gamma_1]$ are trivial elements of $\pi_1(X)$.

Let $\beta(s) = x_t$. Define $\beta_t : [0,1] \to X$ by the formula

$$\beta_t(s) = \beta(st). \tag{11}$$

Note that β_t and γ_t are both paths which join x to x_t.

Lemma 7.4 $[\beta_t] = [\gamma_t]$ *for all* $t \in [0,1]$.

Proof: Let J be the set of parameter values for which $[\beta_t] = [\gamma_t]$. We have $0 \in J$ because β_0 and γ_0 are both trivial in $\pi_1(X, x)$. We show $J = [0,1]$ by showing that J is both closed and open.

Closed: Suppose that $[\beta_t] = [\gamma_t]$ for a sequence of t values converging to s. Since β and f are both continuous,

$$(x_s, [\gamma_s]) = \lim_{t \to s}(x_t, [\gamma_t]) = \lim_{t \to s}(x_t, [\beta_t]) = (x_s, [\beta_s]).$$

Therefore $[\beta_s] = [\gamma_s]$.

Open: Suppose $[\beta_t] = [\gamma_t]$. Let β_{st} denote the restriction of β to $[s,t]$. For s close to t we can take $\gamma_t * \beta_{st}$ as a representative for $[\gamma_s]$. Here we are using the fact that $E : \widetilde{X} \to X$ is a covering map. But then

$$[\gamma_s] = [\gamma_t * \beta_{st}] = [\beta_t * \beta_{st}] = [\beta_s].$$

The central equality comes from the fact that $[\beta_t] = [\gamma_t]$. ♠

By Lemma 7.4 we have

$$f(t) = (\beta(t), [\beta_t]). \tag{12}$$

Since f is a loop in $\widetilde{X}$, the point $f(1) = (x, [\beta])$ is just the basepoint in $\widetilde{X}$. Hence $[\beta]$ is the trivial element in $\pi_1(X, x)$.

For any null loop β, we get the path $f = f_\beta$ defined by equation (12). The loop f_β depends continuoutly on the loop β. As β shrinks down to a point, f_β shrinks down to the constant map. This shows that f is homotopic to a constant map, and hence $\widetilde{X}$ is simply connected.

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 $\boldsymbol{R}^n$ is given by the formula

$$(x_1, \ldots, x_n) \cdot (y_1, \ldots, y_n) = x_1 y_1 + \cdots + x_n y_n. \tag{13}$$

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

$$\|X\| = \sqrt{X \cdot X}. \tag{14}$$

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

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

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

$$R_{12}\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} = \begin{pmatrix} cx_1 + sx_2 \\ -sx_1 + cx_2 \\ \dots \\ x_n \end{pmatrix} \tag{15}$$

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 $\boldsymbol{R}^n$ is given by the formula

$$d(X,Y) = \|X - Y\|. \tag{16}$$

Lemma 8.2 *d satisfies the triangle inequality.*

Proof: For any vectors A and B, we have

$$\|A + B\|^2 = (A + B) \cdot (A + B) =$$

$$\|A\|^2 + 2(A \cdot B) + \|B\|^2 \leq^* \|A\|^2 + 2\|A\|\|B\| + \|B\|^2 \leq (\|A\| + \|B\|)^2.$$

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

$$d(X,Y) = \|X - Z\| = \|A + B\| \leq \|A\| + \|B\|$$

$$\leq \|X - Y\| + \|Y - Z\| = d(X,Y) + d(Y,Z).$$

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

The angle θ between two vectors X and Y obeys the equation

$$\cos(\theta) = \frac{X \cdot Y}{\|X\|\|Y\|}. \tag{17}$$

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

$$\cos(\theta) = X \cdot Y = c. \tag{18}$$

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 θ 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 $V_1, \ldots, V_n$ in $\boldsymbol{R}^n$, the *parallelepiped* spanned by these vectors is defined as the set of all linear combinations

$$\sum a_j v_j, \qquad a_1, \ldots, a_n \in [0, 1].$$

The volume of this parallelepiped is given by

$$\det(V_1, \ldots, V_n) = \sum_{\sigma} (-1)^{|\sigma|} \prod_{i=1}^{n} V_{i,\sigma(i)}. \tag{19}$$

The sum takes place over all permutations σ. The quantity $|\sigma|$ is 0 if σ is an even permutation and 1 if σ is an odd permutation. Finally, V_{ij} is the jth 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 $\boldsymbol{R}^2$ 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.

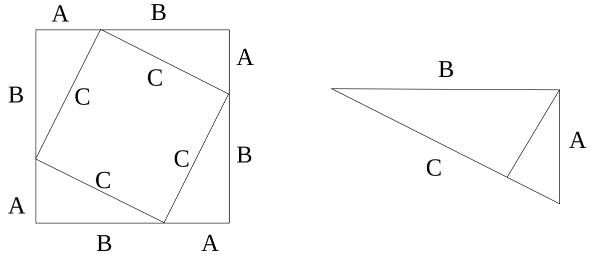

Figure 8.1. Two views of the Pythagorean Theorem

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.

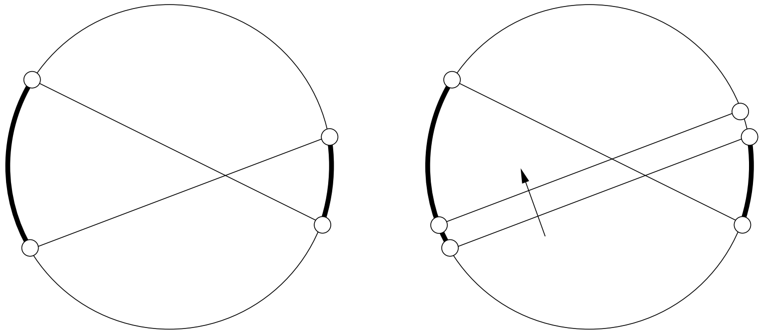

Figure 8.2. The chords A and B.

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 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 $L(A, B)$. ♠

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.

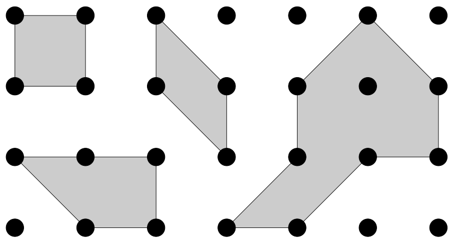

Figure 8.3. Some lattice polygons

Let $\boldsymbol{Z}^2 \subset \boldsymbol{R}^2$ denote the ordinary lattice of integer points. Say that a *lattice polygon* is a polygon in $\boldsymbol{R}^2$ whose vertices lie in $\boldsymbol{Z}^2$. 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 $\boldsymbol{C}$ 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 $E : \boldsymbol{R}^2 \to X$ be the universal covering map. See §6.3. Let P^o denote the interior of the region bounded by the primitive parallelogram P.

We claim that $E(P^o)$ is embedded in X. Otherwise, we can find two points $x_1, x_2 \in P^o$ such that $e(x_1) = e(x_2)$. But then $x_1 - x_2 \in \boldsymbol{Z}^2$. 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.

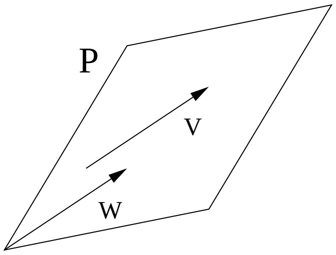

Figure 8.4. Translating a vector

Since $W \in \boldsymbol{Z}^2$, and the vertices of P are in $\boldsymbol{Z}^2$, the head of W lies in $\boldsymbol{Z}^2$. 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. ♠

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

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

Therefore,

$$i + e/2 - 1 = i_1 + i_2 + n - 2 + e_1/2 + e_2/2 - n + 1 - 1$$

$$= (i_1 + e_2/2 - 1) + (i_2 + e_2/2 - 1) =^* A_1 + A_2 = A.$$

The starred equality comes from Pick's Theorem applied to P_1 and P_2. ♠

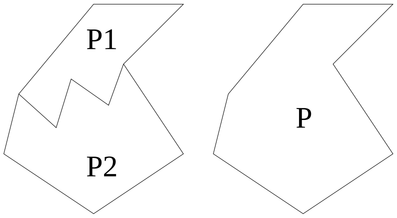

Figure 8.5. Dissecting a polygon

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. ♠

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–Gerwein 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

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

such that P_i and P'_i are isometric for all $i = 1, \ldots, n$. 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.

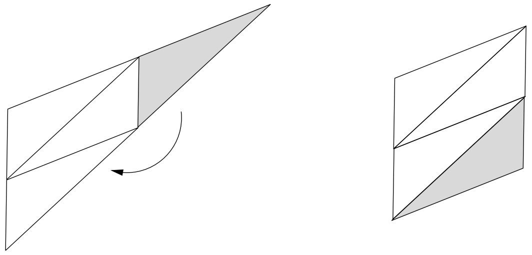

Figure 8.6. Equivalence between a triangle and a parallelogram

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

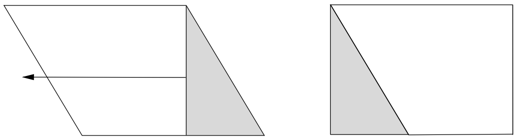

Figure 8.7. Equivalence between a parallelogram and a rectangle

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. ♠

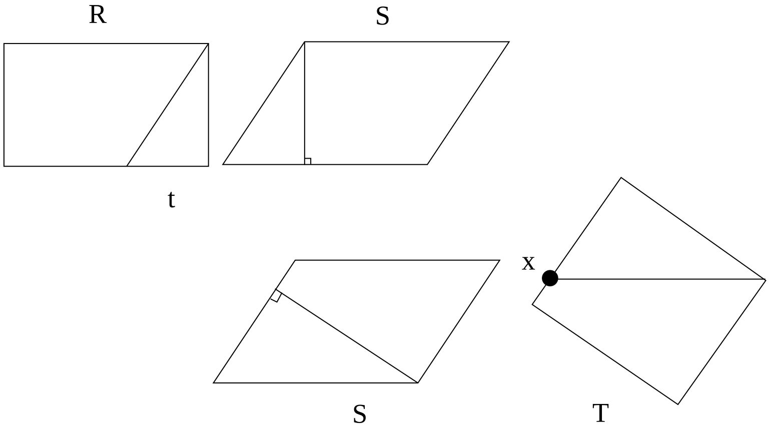

Figure 8.8. Two part construction

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. ♠

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 $T_1, \ldots, T_m$, having areas $a_1, \ldots, a_m$. 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 $\boldsymbol{R}^2$ to $\boldsymbol{R}$. A 1-*form* on an open subset $U \subset \boldsymbol{R}^2$ is a smooth choice $p \to \omega_p$ 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 $\boldsymbol{R}^2$, and

$$dx(v_1, v_2) = v_1, \qquad dy(v_1, v_2) = v_2, \tag{20}$$

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

$$\omega = f dx + g dy, \tag{21}$$

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

$$\omega_p(V) = f(p)v_1 + g(p)v_2. \tag{22}$$

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

Let $\gamma : [0,1] \to \boldsymbol{R}$ be a smooth curve, and let ω 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 γ.

It turns out that the integral only depends on the image and orientation of γ. If

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

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 fdx and gdy. The proof for gdy is the same as for fdx, so we will just consider the case $\omega = fdx$. In this case we set $\gamma(t) = (u(t), v(t))$ and note that

$$\int_\gamma \omega = \int_0^1 (fu')\ dt,$$

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

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

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

Here is an important observation. Since the line integral only depend on the oriented image of γ, 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 γ is a piecewise smooth curve is to say that $\gamma = \gamma_1 \cup \cdots \cup \gamma_n$, where each γ_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 $\omega = fdx + gdy$ be a 1-form defined in an open set that contains D in its interior. Green's Theorem says that

$$\int_\gamma \omega = \int_D (g_x - f_y)\ dxdy. \qquad (23)$$

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 (23). We will just use this piece of notation to shorten our equations, but actually $d\omega$ has a meaning as the exterior derivative of ω. 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 γ 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

$$\int_D d\omega = \int_D (-f_y) = \int_{x=0}^A \Big(\int_{y=0}^{x'} (-f_y)dy\Big)dx$$

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

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

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

Here ∂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. ♠

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. ♠

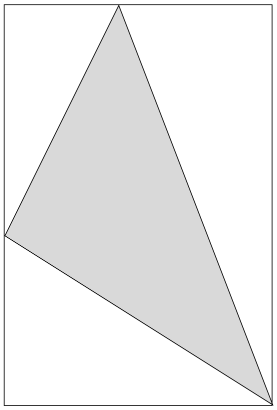

Figure 8.9. A union of triangles

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. ♠

9 Spherical Geometry

The purpose of this chapter is to prove some results about spherical geometry. As usual, S^2 denotes the unit sphere in $\boldsymbol{R}^3$. Most of the results in this chapter can be found in any book on differential geometry; see, for instance, [BAL]. The one topological result, the Hairy Ball Theorem, can be found in most topology books; see, for instance, [GPO].

9.1 Metrics, Tangent Planes, and Isometries

S^2 has two natural metrics on it. The easiest one to define is the *chordal metric*: the distance between $p, q \in S^2$ is $\|p-q\|$. This just uses the Euclidean metric on $\boldsymbol{R}^3$.

The other metric is often called the *round metric*. We define the length of a curve on S^2 to be its length when considered a curve in $\boldsymbol{R}^3$. So, if $\gamma : [a,b] \to S^2$ is a differentiable curve, we have

$$L(\gamma) = \int_a^b \|\gamma'(t)\| \, dt. \tag{24}$$

The distance between two points p and q in the round metric is the infimum of the lengths of all paths on S^2 that join p to q. We will see below that this infimum is realized by a path that is an arc of a great circle. We will see in Chapter 11 that this way of defining a metric is part of a general construction.

In this chapter we will ignore the chordal metric and work with the round metric. Fortunately, any isometry of the chordal metric is an isometry of the round metric and vice versa. The point is that one can give a formula for the one metric in terms of the other. This will become more clear when we work out what the shortest paths are in the round metric.

Later, when we study Riemannian surfaces, we will see that an object called the tangent plane plays a fundamental role in the theory. For the case of the sphere, the tangent plane has a very short and simple definition. The *tangent plane* to S^2 at the point $p \in S^2$ is the plane $T_p(S^2)$ such that $p \in T_p(S^2)$ and $T_p(S^2)$ is perpendicular to the vector pointing from 0 to p. The tangent plane has the following nice property. Any curve $\gamma : [a,b] \to S^2$ is such that the velocity $\gamma'(t)$ lies in the tangent plane $T_{\gamma(t)}(S^2)$.

Any rotation of $\boldsymbol{R}^3$ gives rise to an isometry of S^2. One such rotation is given by the matrix

$$M_t = \begin{bmatrix} \cos(t) & \sin(t) & 0 \\ -\sin(t) & \cos(t) & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

This map rotates by t around the z axis and thus rotates S^2 about the north and south poles. One can find similar maps that rotate around the other two coordinate axes. We call a rotation about one of the coordinate axes a *basic rotation*.

Just by composing the basic rotations, we can move any one point of S^2 to any other point. Moreover, once we know that we can move any point of S^2 to any other, we see that we can find an isometry of S^2 that fixes any given point and rotates through an angle t about that point. Indeed, if $T : S^2 \to S^2$ is an isometry that carries $(0, 0, 1)$ to p, then TM_tT^{-1} is the desired rotation about p.

All the isometries we have described so far come from orientation-preserving linear maps of $\boldsymbol{R}^3$. The other "half" of the isometries come from orientation-reversing linear maps of $\boldsymbol{R}^3$. One such isometry is given by the map $(x, y, z) \to (x, y, -z)$. This map interchanges the north and south poles of S^2 and fixes the equator. More generally, if $v \in S^2$ is any point, the map

$$T_v(w) = -w + 2(v \cdot w)v \tag{25}$$

is an orientation-reversing isometry. The point is that T_v is obviously a linear transformation, and a short calculation shows that

$$T_v(w_1) \cdot T_v(w_2) = w_1 \cdot w_2.$$

Note also that $T_v(v) = -v$, so that T_v swaps v and $-v$. We call the maps in equation (25) *basic reflections*.

9.2 Geodesics

There are many equivalent definitions of a geodesic. To avoid a buildup of terminology, we will give a definition that only relies on what we have already presented. A *geodesic* on S^2 is a curve $\gamma : [a, b] \to S^2$ with the following properties:

- γ has constant speed.
- If t_1 and t_2 are any sufficiently close parameters in $[a, b]$, then the restriction of γ to $[t_1, t_2]$ is the shortest curve on S^2 that joins $\gamma(t_1)$ to $\gamma(t_2)$. In other words, γ is locally a length-minimizing curve.

We will see that a curve is a geodesic if and only if it has constant speed and its image lies in a great circle. A *great circle* is the intersection of a plane through the origin with S^2. The study of geodesics on S^2 is a classical one. It is treated in essentially every book on differential geometry. Here we just establish a few basic facts.

Lemma 9.1 *The shortest differentiable path joining two points on the sphere exists and is an arc of a great circle.*

Proof: Let x and y be two points. We rotate so that x is the north pole. For convenience, we assume that y is not the south pole, so that there is a unique great circle C joining x and y. Let γ be any differentiable curve that joins x to y.

There is a map $\phi : S^2 \to C$. The point $\phi(p)$ is the point of C that lies on the same line of latitude. Geometrically, we think of rotating S^2 around the north and south poles, and watching p rotate around until it sticks on C.

The differential $d\phi_p$ is a map from the tangent plane $T_p(S^2)$ onto the line tangent to C at $\phi(p)$. We note two properties of this map.

- If v is parallel to a line of longitude, then $\|d\phi(v)\| = \|v\|$.
- If w is parallel to a line of latitude, then $d\phi(v) = 0$.

Note also that the lines of longitude and latitude are perpendicular whenever they intersect.

These properties imply that $d\phi$ is a distance nonincreasing map. Moreover, $d\phi$ strictly decreases the length of any tangent vector that is not parallel to a line of longitude. Therefore, the length of $\phi(\gamma)$ is strictly less than the length of γ unless γ traces out a line of longitude. But then γ traces out C, because the lines of longitude are great circles and only one great circle connects x to y (in our case). ♠

Recall that two points $x, y \in S^2$ are called *antipodal* if $x = -y$. Two points are antipodal if and only if they lie on more than one great circle. In case x and y are not antipodal, we define the *geodesic* connecting x to y to be the shorter of the two great circular arcs connecting x to y.

Now that we know about geodesics, we can prove a basic result about isometries of S^2.

Lemma 9.2 *Any isometry of S^2 is a composition of basic reflections.*

Proof: Note that a basic rotation, i.e. a rotation about one of the coordinate axes, is a composition of two basic reflections. So, if we can prove that every isometry is the composition of basic rotations and basic reflections, then we have proved that every isometry is the composition of basic reflections.

Let I be a mystery isometry of S^2. Since we can move any point of S^2 to any other point using compositions of basic rotations, we can compose I with basic rotations so that the result fixes $(0, 0, 1)$. So, without loss of generality, we can assume that I fixes $(0, 0, 1)$.

The *equator* E on S^2 is the set of points of the form $(x, y, 0)$. The equator divides the sphere into the upper hemisphere and the lower hemisphere. Any point on the lower hemisphere is farther from $(0, 0, 1)$ than is any point on the upper hemisphere. For this reason, $I(E) = E$. Now, E is a circle and I is an isometry of E. So, I acts on E either as a rotation or a reflection. Composing I with basic reflections and/or rotations, we can assume that I fixes every point on E.

Any point $p \in E$ is connected to $(0, 0, 1)$ by the arc γ_p, which is one quarter of a great circle. Since I fixes the endpoints of γ_p, and γ_p is the unique shortest path joining p to $(0, 0, 1)$, we have $I(\gamma_p) = \gamma_p$. Moreover, I preserves distances along γ_p. Hence I fixes every point of γ_p. Since p is an arbitary point of E, we see that I fixes every point of the upper hemisphere. A similar argument shows that I fixes every point of the lower hemisphere. Hence, I is the identity. ♠

9.3 Geodesic Triangles

Let d denote the distance on S^2. If x and y are antipodal, then $d(x, y) = \pi$. In general, $d(x, y) = \theta$, where θ is the angle between the vector pointing to x and the vector pointing to y. Familiar formulas in linear algebra give

$$\cos(d(x, y)) = x \cdot y, \qquad \sin(d(x, y)) = \|x \times y\|. \tag{26}$$

Here × is the cross product. What makes these formulas simple is the fact that $\|x\| = \|y\| = 1$.

We measure angles on S^2 using the dot product on $\boldsymbol{R}^3$. Suppose that C_1 and C_2 are two geodesics connecting x to y_1 and y_2. The angle between C_1 and C_2 at x is just the angle between the tangent vectors at x. This is the same as the dihedral angle between the plane Π_1 containing $(0, x, y_1)$ and the plane Π_2 containing $(0, x, y_2)$. As usual, there are two angles we can measure at x, and the sum of these angles is π.

Let x_1, x_2, x_3 be three points, all contained in the same hemisphere. Then there is a unique geodesic C_j joining x_{j-1} and x_{j+1}, with the indices taken cyclically. The union of these geodesics is called a *spherical triangle*. Let θ_j be the interior angle at x_j, and let L_j denote the length of C_j.

There is a beautiful formula for the area of a spherical triangle, known as *Giraud's Theorem*. (Thomas Harriot discovered the result in 1603 but did not publish it.) The area is given by

$$\theta_1 + \theta_2 + \theta_3 - \pi. \tag{27}$$

This result is a special case of the general Gauss–Bonnet Theorem, a result proved much later on. Here we sketch a proof of Giraud's Theorem. The case when the 3 points lie on the same great circle is trivial. In all other cases, the whole triangle lies in an open hemisphere.

Say that a *lune* is a region bounded by two great semicircles. A lune has two vertices. By symmetry, the interior angles at either end of the lune are the same. Any two lunes having the same interior angles are isometric to each other. Let $A(\theta)$ be the area of a lune having angle θ.

Lemma 9.3 *$A(\theta) = 2\theta$ for all $\theta \in [0, \pi]$.*

Proof: If $\theta = \pi$, the lune is precisely a hemisphere. Hence

$$A(\pi) = 2\pi. \tag{28}$$

Moreover, a lune having interior angle θ decomposes into n lunes having interior angle θ/n. Hence

$$A(\theta) = nA(\theta/n). \tag{29}$$

Combining equations (28) and (29) we see that $A(\theta) = 2\theta$ whenever θ is a rational multiple of π. But A is a continuous function of θ. Hence $A(\theta) = 2\theta$

for all θ. ♠

Now let T be a geodesic triangle, contained in a hemisphere, having interior angles θ_1, θ_2, and θ_3. Extending the sides of T, we can cover S^2 by 6 lunes.

Figure 9.1, which needs some interpretation, shows the situation. In Figure 9.1, we have drawn T extremely small, and placed near (say) the north pole. We are looking down on T. The sides of T practically look straight because T is very small. We have extended the sides of T and partially shown them. These sides continue all the way around S^2 and join up again near the south pole, where they form another copy T' of T. Our technical assumption about T lying in a hemisphere guarantees that T and T' are disjoint. We have drawn the boundary of T thickly, and we have shaded two of the lunes. These two lunes meet at both vertices, the other vertex being near the south pole.

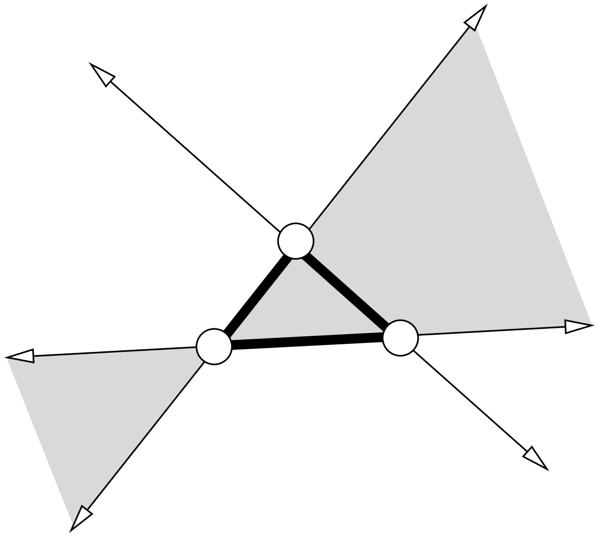

Figure 9.1. Dissected sphere

By Lemma 9.3, the total area of the lunes is

$$4(\theta_1 + \theta_2 + \theta_3). \qquad (30)$$

With the exception of some points on the edges of the lunes, every point of $S^2 - (T \cup T')$ is covered once by a lune. At the same time, every point of

$T \cup T'$ is covered 3 times by the union of the lunes. Letting A be the area of T (and T') we have

$$4(\theta_1 + \theta_2 + \theta_3) = (4\pi - 2A) + 6A = 4\pi + 4A. \tag{31}$$

Simplifying this equation gives $A = \theta_1 + \theta_2 + \theta_3 - \pi$, as desired.

Exercise 1. Try to draw a version of Figure 9.1 that shows the entire sphere, as well as the 6 lunes.

9.4 Convexity

It doesn't really make much sense to talk about the convexity of a general subset of S^2, because some pairs of points on S^2 can be joined by more than one shortest path. However, if $X \subset S^2$ is entirely contained in an open hemisphere H, then any two points of X can be joined by a unique arc of a great circle that has length less than π. This arc remains inside H. We call this the *geodesic segment* joining the points.

We call X *convex* if the geodesic segment joining any pair of points in X remains in X. This definition appears to depend on H, but it does not.

Exercise 2. Prove that the notion of convexity for X does not depend on the hemisphere relative to which it is defined. That is, if X is contained in the intersection $H_1 \cap H_2$ of two open hemispheres, then X is convex relative to either one of them.

If $X \subset H \subset S^2$ is an arbitrary set, we define the *convex hull* of X to be the intersection of all the closed convex subsets of H that contain X. We call this set Hull(X).

Exercise 3. Prove that Hull(X) is well defined, independent of the open hemisphere that contains X. Prove also that Hull(X) is convex relative to any open hemisphere that contains it.

The purpose of the next 2 exercises is to establish some background results needed for the Cauchy Rigidity Theorem, proved in Chapter 24. We say that a *convex spherical polygon* is a simple closed polygonal curve in S^2 made from arcs of great circles that is contained in a hemisphere and bounds a convex set contained in that same hemisphere. We insist that consecutive arcs make an angle which is distinct from π. From Exercise 3, the definition of a convex spherical polygon does not depend the choice of hemisphere that contains it.

Exercise 4. Let Γ be a convex spherical polygon. Let $\widehat{C}$ be a great circle that extends a side C of Γ. Then $\Gamma - C$ is contained in one of the two open hemispheres bounded by $\widehat{C}$.

Exercise 5. Let Q and Q' be convex spherical quadrilaterals. Let the sides of Q be C_1, C_2, C_3, C_4. Let θ_j be the interior angle between C_j and C_{j+1}. Make the same definitions for Q'. Suppose that C_j and C'_j have the same length for all j. Label a vertex of Q by a $(+)$ if $\theta > \theta'$ at that vertex, and by a $(-)$ if the opposite inequality holds. Prove that the labels of the vertices of Q must have the form $(+,-,+,-)$ or $(0,0,0,0)$, up to cyclic ordering.

9.5 Stereographic Projection

Let $\boldsymbol{C}$ denote the complex numbers. We think of ∞ as an extra point and consider $\boldsymbol{C} \cup \infty$. We want to think of $\boldsymbol{C} \cup \infty$ as a sphere. To do this, we want to put a metric on $\boldsymbol{C} \cup \infty$ so that the result is homeomorphic to a sphere. The metric we get on $\boldsymbol{C} \cup \infty$ is not really so natural, but it does allow us to speak of continuous maps from $\boldsymbol{C} \cup \infty$ to itself. This is something we will take up in the next chapter.

One way to put a metric on $\boldsymbol{C} \cup \infty$ is to choose a map from S^2 to $\boldsymbol{C} \cup \infty$ which is a homeomorphism from $\boldsymbol{C}$ to S^2 minus a single point, say $(0,0,1)$. Then, we put a metric on $\boldsymbol{C} \cup \infty$ so that our map is an isometry. One very nice map from S^2 to $\boldsymbol{C} \cup \infty$ is *stereographic projection*.

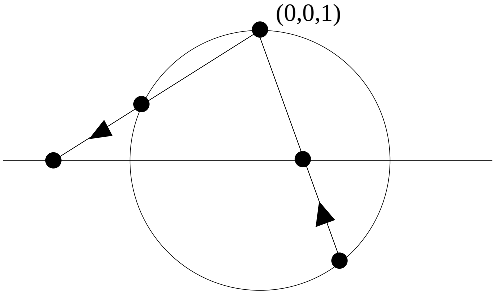

Figure 9.2. Stereographic projection

As Figure 9.2 (drawn 1 dimension down) illustrates, stereographic projection has the following geometric description. We identify $\boldsymbol{C}$ with the horizontal plane $\boldsymbol{R}^2 \times \{0\}$ in $\boldsymbol{R}^3$. Half of S^2 lies above this plane and half below. We map $(0,0,1)$ to ∞. Given any other point $p \in S^2$, we define $\phi(p) \in \boldsymbol{C}$ to be the point such that $(0,0,1)$ and p and $\phi(p)$ are collinear.

The formula is given by

$$\phi(x,y,z) = \Big(\frac{x}{1-z}\Big) + \Big(\frac{y}{1-z}\Big)i. \tag{32}$$

The inverse map is given by

$$\phi^{-1}(x+iy) = \Big(\frac{2x}{1+x^2+y^2}, \frac{2y}{1+x^2+y^2}, 1 - \frac{2}{1+x^2+y^2}\Big).$$

One can check easily that these maps are inverses of each other.

Exercise 6. Check that our formula for stereographic projection matches the geometric description.

Exercise 7. Check that ϕ gives a homeomorphism from $S^2 - (0,0,1)$ to $\boldsymbol{C}$.

One of the nice facts about stereograpic projection is the following. If $C \subset S^2$ is a circle, then $\phi(C)$ is either a circle in $\boldsymbol{C}$ or else a straight line (union ∞). When C contains the point $(0,0,1)$, this result is fairly obvious from the geometric description. The idea is that any circle $C \subset S^2$ has the form $\Pi_C \cap S^2$ for some plane Π_C. When $(0,0,1) \in \Pi_C$, we see from the geometric description that

$$\phi(\Pi_C) = (\boldsymbol{C} \cap \Pi_C) \cup \infty.$$

A general geometric proof, based on conic sections, is given in [HCV]. In §14.3, we will give a proof based on complex analysis.

Exercise 8 (Challenge). Find your own proof that stereographic projection maps circles in S^2 to either circles or straight lines in $\boldsymbol{C}$.

9.6 The Hairy Ball Theorem

Let me end the chapter with the *Hairy Ball Theorem*. This is really a result about the topology of the sphere, and not its geometry, but it is such a great result that I wanted to put it in.

A *unit field* on S^2 is a continuous choice of unit vector tangent to S^2 at each point. The Hairy Ball Theorem says that a unit field on S^2 does not exist. The name of the theorem (which has somewhat fallen out of favor) comes from the following interpretation: If you have a sphere that is completely covered in hair, you cannot comb the hair so that it lies flat and varies continuously. There has to be some kind of cowlick somewhere.

We will suppose that a unit field exists, and derive a contradiction. Suppose we have a unit field $\boldsymbol{U}$ on S^2. Let $\gamma : [0,1] \to S^2$ be the a smooth loop, so that $\gamma(0) = \gamma(1)$. For each $t \in [0,1)$, we let $\theta(t)$ denote the counterclockwise angle between the tangent vector $\gamma'(t)$ and our vector field at $\gamma(t)$. We choose so that $\theta(t)$ varies continuously. As $t \to 1$, the value $\theta(t)$ necessarily tends to an integer multiple of 2π. We let

$$N(\boldsymbol{U}, \gamma) = \lim_{t \to 1} \theta(t) - \theta(0). \tag{33}$$

Intuitively, you are walking along γ, turning your head according the direction of $\boldsymbol{U}$. Once you get back to where you start, you will be looking in the same direction as when you started, except that your head will be turned around N times counterclockwise! Compare this discussion about the winding number given in §5.1,

Exercise 9. Prove that the quantity $N(\boldsymbol{U}, \gamma)$ is independent of the smooth parametrization of γ, as long as the orientation does not switch. Also prove that $N(\boldsymbol{U}, \gamma) = N(\boldsymbol{U}, \gamma')$ when γ and γ' are homotopic loops. (*Hint*: $N(\boldsymbol{U}, \gamma)$ is continuous and integer-valued.)

Consider the case when γ is a small loop that winds once around the north pole. As we walk around γ, keeping our head aligned with the unit field, our head always points in roughly the same direction. So, when we make one complete circuit, our neck is twisted once around. (Don't try this at home.) That is $N(\boldsymbol{U}, \gamma) = \pm 1$. If we replace γ by $-\gamma$, the loop that is oriented in the opposite direction, the sign of $N(\boldsymbol{U}, \gamma)$ switches.

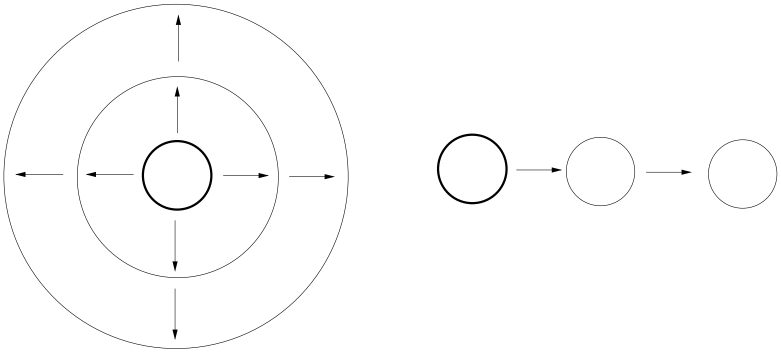

Figure 9.3. Two homotopies

We orient γ so that $N(\boldsymbol{U}, \gamma) = 1$. There are two ways to slide γ to a small loop around the south pole. On the one hand, we can push γ around the side, following a single line of longitude and keeping γ small. On the other hand, we can pull γ down over the whole sphere, moving through the circles of latitude. Figure 9.1 shows a top-down view of the two methods. One method leads to a small loop β about the south pole and the other method leads to $-\beta$, the oppositely oriented loop. By Exercise 1, we have

$$N(\boldsymbol{U}, \beta) = N(\boldsymbol{U}, \gamma) = N(\boldsymbol{U}, -\beta) = -N(\boldsymbol{U}, \beta) = 1.$$

This equation says that $1 = -1$, which is a contradiction. This proves the Hairy Ball Theorem.

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 variery 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 2×2 matrix with complex number entries and determinant 1. The set of these matrices is denoted by $SL_2(\boldsymbol{C})$. 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 ∞ and define $T_A(-d/c) = \infty$. This definition is a natural one because of the limit

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

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

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

Exercise 1. As in §9.5, we introduce a metric on $\boldsymbol{C} \cup \infty$ so that $\boldsymbol{C} \cup \infty$ is homeomorphic to the unit sphere $S^2 \subset \boldsymbol{R}^3$. 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 $A, B \in SL_2(\boldsymbol{R})$. In particular (since A^{-1} exists) the inverse map T_A^{-1} exists. By Exercise 1, this map is also a continuous map of $\boldsymbol{C} \cup \infty$. Conclude that T_A is a homeomorphism of $\boldsymbol{C} \cup \infty$.

10.2 Circle Preserving Property

A *generalized circle* in $\boldsymbol{C} \cup \infty$ is either a circle in $\boldsymbol{C}$ or a set of the form $L \cup \infty$, where L is a straight line in $\boldsymbol{C}$. 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 $\boldsymbol{C}$. Then there exists a linear fractional transformation T such that $T(\boldsymbol{R} \cup \infty) = C$.*

Proof: If C is a straight line (union ∞), 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 $\boldsymbol{R} \cup \infty$ onto the unit circle C_0 satisfying the equation $|z| = 1$. The point is that every point $z \in \boldsymbol{R}$ 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. ♠

Lemma 10.3 *Suppose that L is a closed loop in $\boldsymbol{C} \cup \infty$. 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 ∞) the result is obvious. Otherwise, L has 3 noncollinear points and, like any 3 noncollinear points, these lie on a common circle. ♠

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 $\boldsymbol{R} \cup \infty$. Then there exists a linear fractional transformation that preserves $\boldsymbol{R} \cup \infty$ 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 ∞, 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 $z \to rz + s$ to fix the images of a_1 and a_2. ♠

Lemma 10.5 *Suppose T is a linear fractional transformation that fixes* 0 *and* 1 *and ∞. 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 $T(z) = z$. ♠

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 $C = \boldsymbol{R} \cup \infty$. By Lemma 10.3 there is a generalized circle D such that D and $T(\boldsymbol{R} \cup \infty)$ 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 $S(\boldsymbol{R} \cup \infty) = D$. There are points $a_1, a_2, a_3 \in \boldsymbol{R} \cup \infty$ such that $S(a_j) = c_j$ for $j = 1, 2, 3$. Also, there are points $b_1, b_2, b_3 \in \boldsymbol{R} \cup \infty$ 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 $\boldsymbol{R} \cup \infty$ 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

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

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

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

is not a generalized circle and $S \circ A^{-1}(\boldsymbol{R} \cup \infty) = D$ 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 $U \subset \boldsymbol{C}$ be the upper half-plane, consisting of points z with $\mathrm{Im}(z) > 0$. 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 $\gamma : [a, b] \to U$, we define

$$L(\gamma) = \int_a^b \frac{|\gamma'(t)|}{\mathrm{Im}(\gamma(t))}\, dt. \tag{34}$$

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 $\gamma : \boldsymbol{R} \to U$ defined by

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

Then the length of the portion of γ 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 γ 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 γ 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

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

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}.

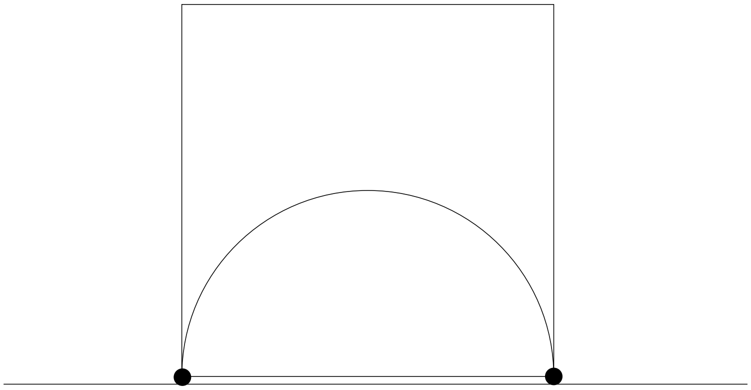

Figure 10.1. Some paths in the hyperbolic plane

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 γ 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 $\boldsymbol{H}^2$. So far we have talked about lengths of curves in $\boldsymbol{H}^2$, but we can also talk about angles. The angle between two differentiable and regular (i.e., nonzero speed) curves in $\boldsymbol{H}^2$ 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 $D \subset \boldsymbol{H}^2$ to be the integral

$$\int_D \frac{dx\ dy}{y^2}. \tag{35}$$

10.4 Another Point of View

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

- $\langle av + w, x\rangle = a\langle v, x\rangle + \langle w, x\rangle$ for all $a \in \boldsymbol{R}$ 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 $\boldsymbol{R}^2$. At the point $z = x + iy$ we introduce the inner product

$$\langle v, w\rangle_z = \frac{1}{y^2}(v \cdot w). \tag{36}$$

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

$$\|v\|_{\boldsymbol{H}^2} = \sqrt{\langle v, v\rangle_z}. \tag{37}$$

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

$$\int_a^b \|\gamma'(t)\|_{\gamma(t)}\ dt. \tag{38}$$

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, $T_A(z) \in \boldsymbol{C}$ provided $z \in \boldsymbol{C} - \boldsymbol{R}$.

Exercise 3. Prove that $z \not\in \boldsymbol{R}$ implies that $T_A(z) \not\in \boldsymbol{R}$. Prove also that T_A maps $\boldsymbol{H}^2$ into itself.

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

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 γ is any curve in $\boldsymbol{H}^2$, 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 γ move at the same hyperblic speed at corresponding points. So, if we connect points p and q by some curve γ 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 $z \in \boldsymbol{H}^2$. 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 (36), we see that

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

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 $\boldsymbol{H}^2$. Recall that in §2.8 we proved $SL_2(\boldsymbol{R})$ is a 3-dimensional manifold. So, $\boldsymbol{H}^2$ 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 $\boldsymbol{H}^2$ which intersect at a point x and make an angle of θ, then $T(a)$ and $T(b)$ make the same angle θ 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 $z \to -1/z$, and so it suffices to consider this one.)

10.6 Geodesics

In this section we will describe the shortest curves connecting two points in $\boldsymbol{H}^2$. 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 $\boldsymbol{H}^2$ 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 γ is a curve in $\boldsymbol{H}^2$, then the hyperbolic speed of $F(\gamma)$ at any point is at most the hyperbolic speed of γ at the corresponding point. Moreover, if the velocity of γ 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 γ is the integral of its hyperbolic speed. Thus the hyperbolic length of $F(\gamma)$ is less than the hyperbolic length of γ, unless γ travels vertically the whole time. Our result follows immediately from this. ♠

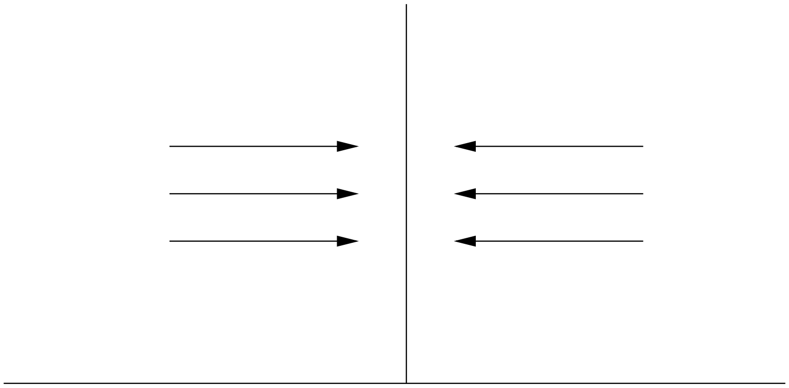

Figure 10.2. The map F

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

Exercise 7. Let p and q be two arbitary points in $\boldsymbol{H}^2$. 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 $\boldsymbol{H}^2$ 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 ρ 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 $\boldsymbol{R} \cup \infty$, both endpoints of this circular arc lie on $\boldsymbol{R} \cup \infty$. Finally, since T preserves angles, $T(\rho)$ meets $\boldsymbol{R}$ at right angles at any point where $T(\rho)$ intersects R. If $T(\rho)$ limits on ∞, then $T(\rho)$ is another vertical ray. Otherwise, $T(\rho)$ is a semicircle, contained in a circle that is centered on the real axis. ♠

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

$$M(z) = \frac{z-i}{z+i}. \tag{39}$$

This map does the right thing because $z \in \boldsymbol{H}^2$ 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 $\boldsymbol{H}^2$ to circular arcs in Δ 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 Δ 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 Δ as a convenient place to draw pictures, we can also think of Δ as another model of $\boldsymbol{H}^2$. 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 $\boldsymbol{H}^2$.

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

$$\langle v, w \rangle_z = \frac{4v \cdot w}{(1-|z|)^2}. \tag{40}$$

Once we have this inner product, we can directly define lengths of curves in Δ as in equation (38). Then we can define distances in Δ 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 $\boldsymbol{H}^2$ and the inner product on Δ.

Exercise 9. Prove that the map M is an isometry from $\boldsymbol{H}^2$ and Δ, when lengths are defined in terms of the inner product in equation (40). 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 $z \in \boldsymbol{H}^2$.

The disk Δ, 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 Δ. Since M preserves angles, the hyperbolic angle between two curves in Δ 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 $\boldsymbol{H}^2$ is defined to be $\boldsymbol{R} \cup \infty$ 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 $\boldsymbol{H}^2$. They are considered "limit points" of geodesics in $\boldsymbol{H}^2$.

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 $\boldsymbol{H}^2$ 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 $\boldsymbol{H}^2$ 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 $\boldsymbol{H}^2$ 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 π minus the sum of the interior angles of π. In particular, the sum of these interior angles is less than π.*

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 π. You can move any one ideal triangle to any other using an isometry of $\boldsymbol{H}^2$. 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 ∞. We first observe that

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

Now we compute our area, using equation (35). 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 $dx = \cos(t)$. ♠

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

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.

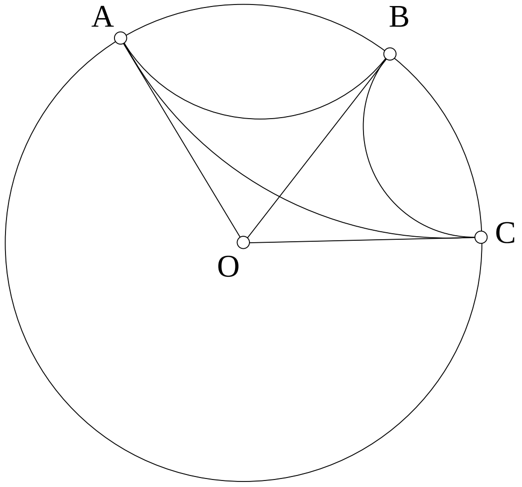

Figure 10.3. Two dissections

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. ♠

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

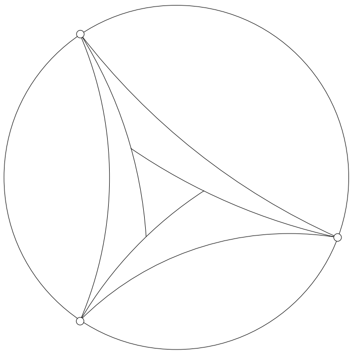

Figure 10.4. A Dissected ideal triangle.

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 π, and the three outer triangles have areas α, β, and γ, 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 propery: 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. ♠

Exercise 10 (Challenge). Suppose that $\theta_1, \theta_2, \theta_3$ are three numbers whose sum is less than π. 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 *δ-thin* if every point in the interior of the (solid version of) triangle is within δ of a point on the boundary. Note that there is no universal δ so that all Euclidean triangles are δ-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 $a, b, c \in \boldsymbol{R}$. If T is not the identity, then there are 3 possibilities:

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

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 ∞. In this case $S(z) = az + b$. If $a \neq 1$, then S fixes an additional point in $\boldsymbol{R}$. 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 ∞. 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 $\boldsymbol{H}^2$, 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 $\boldsymbol{H}^2$ more than ϵ. For example, the hyperbolic distance between iy and $S(iy)$ tends to 0 as $y \to \infty$. 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 $\boldsymbol{H}^2$ by at least ϵ. Indeed, $\epsilon = |\log(r)|$.

11 Riemannian Metrics on Surfaces

The purpose of this chapter is explain what is meant by a *smooth surface with a Riemannian metric*. The main construction generalizes what we did for the sphere in §9.1 and also (especially) what we did for the hyperbolic plane in §10.3. We will give the main definition of a surface with a Riemannian metric at the end, after assembling all the preliminary definitions.

A smooth surface with a Riemannian metric is a special case of a *smooth Riemannian manifold*. Smooth Riemannian manifolds are the subject of differential, or Riemannian, geometry. A book such as [DOC] gives an excellent general account of smooth Riemannian manifolds.

11.1 Curves in the Plane

A *smooth curve* in $\boldsymbol{R}^2$ is a smooth map $f : (a,b) \to \boldsymbol{R}^2$. Such a map is typically given by equations

$$f(t) = (x(t), y(t))$$

such that $x(t)$ and $y(t)$ are smooth functions. This is to say that

$$\frac{d^n f}{dt^n} = \Big(\frac{d^n x}{dt^n}, \frac{d^n y}{dt^n}\Big)$$

exists for all n. We will usually write $f'(t)$ for df/dt.

The function f is *regular* if $f'(t) \neq 0$ for all $t \in (a,b)$. As usual, $f'(t)$ is known as the *velocity* of f at t. Sometimes it is useful to talk about smooth curves defined on closed intervals. Thus, if we write $f : [a,b] \to \boldsymbol{R}^2$, we really mean that f is defined on some larger open interval $(a-\epsilon, b+\epsilon)$ and is smooth there. In particular $f : [0,0] \to \boldsymbol{R}^2$ is defined in a neighborhood of 0. This is the usual treatment of the problem with taking derivatives at the endpoints.

11.2 Riemannian Metrics on the Plane

We defined inner products at the top of §10.4. Let $\mathcal{I}$ denote the set of inner products on $\boldsymbol{R}^2$. Let $U \subset \boldsymbol{R}^2$ be an open set. A *Riemannian metric* on U is a smooth map $\Psi : U \to \mathcal{I}$. In other words, a Riemannian metric on U is a choice G_p of inner product for each $p \in U$. This choice gives rise to the functions $g_{ij}(p)$, via the formula

$$g_{ij}(p) = G_p(e_i, e_j). \tag{41}$$

Here $e_1 = (1,0)$ and $e_2 = (0,1)$. We require that the functions g_{ij} are smooth functions on U. So, you can specify a Riemannian metric on U by specifying 4 smooth functions $g_{ij} : U \to \boldsymbol{R}$ subject to the following constraints:

- $g_{12}(p) = g_{21}(p)$ for all $p \in U$.
- For all $p \in U$, the matrix $\{g_{ij}(p)\}$ is positive definite. That is, the matrix has positive eigenvalues.

A *curve in U* is just a curve in $\boldsymbol{R}^2$ which happens to lie entirely in U. We can measure the length of a curve in U relative to the given Riemannian metric, as follows: Let $f : [a,b] \to U$ be a smooth curve. We define

$$\text{Riemannian length}(f) = \int_a^b \sqrt{G_{f(t)}(f'(t), f'(t))}\, dt. \tag{42}$$

The integrand above is called the *Riemannian speed* of f at t. So, we are computing the Riemannian length of f by integrating its Riemannian speed. Of course, these quantities depend on the choice of Riemannian metric. If we choose the standard Riemannian metric, which is to say the ordinary dot product at each point, then we recover the ordinary notions of speed and length.

Exercise 1: Using the material in the previous chapter, describe the Riemannian metric on the upper half-plane which gives rise to the hyperbolic plane.

Exercise 2: Come up with a sensible definition of the Riemannian area of a subset of $\boldsymbol{R}^2$, assuming that $\boldsymbol{R}^2$ has been equipped with a Riemannian metric.

Exercise 3. Give an example of a Riemannian metric, defined on all of $\boldsymbol{R}^2$, which has the following property. Any two points in $\boldsymbol{R}^2$ can be joined by a smooth curve whose Riemannian length is less than 1.

Exercise 4. Let G be a Riemannian metric on the plane and let p, q be two distinct points. Prove that there is some $\epsilon > 0$ such that any curve connecting p to q has length at least ϵ relative to G. Of course, ϵ depends on the metric. (*Hint*: Use the fact that a positive continuous function on a compact set has a positive infimum.)

11.3 Diffeomorphisms and Isometries

Let U and V be two open subsets of $\boldsymbol{R}^2$. A *diffeomorphism* from U to V is a homeomorphism $f : U \to V$ with the following additional properties:

- f is smooth, that is, all orders of partial derivatives of f exist.
- For each $p \in U$, the matrix $df(p)$ of first partial derivatives is nonsingular. That is, df defines a vector space isomorphism at each point. We abbreviate this by saying that f is *regular*.
- f^{-1} is smooth and regular.

Actually, the third condition follows from the other two and the Inverse Function Theorem.

Note that df_p maps a tangent vector based at p to a tangent vector based at $f(p)$. Suppose that U and V are given Riemannian metrics. We say that a diffeomorphism $f : U \to V$ is a *Riemannian isometry* if

$$H_{f(p)}(df_p(v), df_p(w)) = G_p(v, w), \qquad \forall v, w, p.$$

Here v and w are vectors and $p \in U$. Also G is the Riemannian metric defined on U, and H is the Riemannian metric defined on V. We have already encountered this concept in our second proof of Lemma 10.6.

Here is another point of view on Riemannian isometries. A Riemannian metric on $U \subset \boldsymbol{R}^2$ turns U into a metric space in the following way. Given $p, q \in U$ we define $S(p, q)$ to be the set of smooth curves in U which join p to q. We define $d(p, q)$ to be the infimum of the lengths of curves in $S(p, q)$. This is exactly what we did both for the sphere and for the hyperbolic plane in the preceding chapters. A smooth map $f : U \to V$ is a Riemannian isometry if and only if it is a metric isometry relative to the two metric space structures.

Exercise 5. Prove that d really is a metric on U. Prove also that a Riemannian isometry between U and V gives rise to a metric space isometry.

Exercise 6 (Challenge). Prove that there is a Riemannian metric on the plane which makes it isometric to the upper hemisphere of S^2, relative to the round metric. (This part is not so hard.) Now, prove that there is no Riemannian metric on the plane which makes it isometric to the upper hemisphere of S^2 relative to the chordal metric. See §9.1 for definitions.

11.4 Atlases and Smooth Surfaces

Recall that a *surface* is a metric space S such that every point has a neighborhood which is homeomorphic to $\boldsymbol{R}^2$. We say that a collection of such neighborhoods is called an *atlas*. The neighborhoods themselves are called *coordinate charts*. So, each element of the atlas is a pair (U, h), where U is an open subset of Σ and $h : U \to \boldsymbol{R}^2$ is a homeomorphism. We require that the union of all the coordinate charts in the atlas is the entire surface. In other words, each point in the surface is contained in at least one coordinate chart.

Suppose now that (U_1, h_1) and (U_2, h_2) are 2 coordinate charts, and it happens that $V = U_1 \cap U_2$ is not empty. We define $V_1 = h_1(V)$ and $V_2 = h_2(V)$. Being the intersection of two open sets, V is an open subset of both U_1 and U_2. Since h_1 and h_2 are homeomorphisms, V_1 and V_2 are open subsets of $\boldsymbol{R}^2$. On V_1 the map

$$h_{12} = h_2 \circ h_1^{-1}$$

is well defined. We have $h_{12}(V_1) = V_2$. The map

$$h_{21} = h_1 \circ h_2^{-1}$$

is defined on V_2 and evidently $h_{21}(V_2) = V_1$. The two maps h_{12} and h_{21} are inverses of each other. Also, both maps are continuous, since they are the composition of continuous maps. In summary $h_{12} : V_1 \to V_2$ is a homeomorphism and $h_{21} : V_2 \to V_1$ is the inverse homeomorphism. These two functions are called *overlap functions* because they are defined on the overlaps between coordinate charts.

Our atlas on Σ is said to be a *smooth structure* if all its overlap functions are smooth diffeomorphisms. In other words, every time we can produce an overlap function $h_{12} : V_1 \to V_2$, it turns out to be a diffeomorphism. We say that a *smooth surface* is a surface equipped with a smooth structure.

Here is an annoying technical point. Let (U,h) be a pair such that U is an open subset of Σ and $h : U \to \boldsymbol{R}^2$ is a homeomorphism. If (U,h) is not part of our atlas, then we can enlarge our atlas by including (U,h) in it. This will produce possibly some new overlap functions. If all the new overlap functions are diffeomorphisms, then we say that (U,h) is compatible with our atlas. We say that our atlas is *maximal* if it already contains all compatible coordinate charts. It is conventional for us to require that our atlases be maximal. However, this point never actually comes up in practice.

11.5 Smooth Curves and the Tangent Plane

We have already discussed what we mean by a smooth curve in the plane. Now we will generalize the idea, and speak about smooth curves on a smooth surface. If we happen to have a smooth surface embedded in Euclidean space, such as the sphere embedded in $\boldsymbol{R}^3$, then it is easy to talk about smooth curves. For instance, we could say that a smooth curve $f : (a,b) \to S^2$ is smooth if each of the 3 coordinate functions is smooth. When we deal with an abstract smooth surface, the situation is a bit trickier. We always need to refer back to the coordinate charts defining the surface.

Let Σ be a smooth surface. Say that a map $f : (a,b) \to \Sigma$ is *smooth at* t if there is some $\epsilon > 0$ such that the following holds.

- $(t-\epsilon, t+\epsilon) \in (a,b)$.
- $f((t-\epsilon, t+\epsilon))$ is contained in a coordinate chart (U,h) in our atlas.
- The curve $h \circ f : (t-\epsilon, t+\epsilon) \to \boldsymbol{R}^2$ is a smooth curve.

The fact that our overlap functions are all diffeomorphisms means that the notion of smoothness does not depend on which coordinate chart we use. In other words, if $f(t-\epsilon, t+\epsilon) \subset U_1 \cap U_2$ and (U_1,h_1) and (U_2,h_2) are both coordinate charts, then

$$h_2 \circ f = h_{12} \circ (h_1 \circ f).$$

Since h_{12} is smooth, the curve $h_1 \circ f$ is smooth if and only if the curve $h_2 \circ f$ is smooth. Here are using the fact that the composition of smooth maps is again smooth. This fact is in turn a consequence of the chain rule.

We say that $f : (a,b) \to \Sigma$ is *smooth* if f is smooth at each $t \in (a,b)$. We say that $f : [a,b] \to \Sigma$ is smooth if f is defined and smooth on a larger interval $(a-\epsilon, b+\epsilon)$.

Let $p \in \Sigma$ be a point. Suppose that

$$f_1, f_2 : [0, 0] \to \Sigma$$

are two curves such that $f_1(0) = f_2(0) = p$. We write $f_1 \sim f_2$ if there is a coordinate chart (U, h) such that $p \in U$ and $h \circ f_1$ and $h \circ f_2$ have the same velocity at 0. In other words, $(h \circ f_1)'(0) = (h \circ f_2)'(0)$.

Exercise 7. Prove that $\sim$ is well defined, independent of the coordinate chart we use. Prove also that $\sim$ is an equivalence relation.

We define $T_p(\Sigma)$ to be the set of equivalence classes of curves $f : [0, 0] \to \Sigma$ such that $f(0) = p$. We can make $T_p(\Sigma)$ into a vector space as follows. If $[f_1]$ and $[f_2]$ are two equivalence classes of curves, we define $[f_1] + [f_2]$ to be the equivalence class of the curve g such that the velocity of $h \circ g$ is the velocity of $h \circ f_1$ plus the velocity of $h \circ f_2$. That is,

$$(h \circ g)'(0) = (h \circ f_1)'(0) + (h \circ f_2)'(0).$$

Exercise 8. Prove that this notion of addition is well defined. In other words, if we made this definition relative to two different coordinate charts (U_1, h_1) and (U_2, h_2), then we could get the same answer. (*Hint*: Use the fact that

$$h_2 \circ g = h_{12} \circ (h_1 \circ g)$$

(and likewise for f_1 and f_2) and the fact that dh_{12} is a linear transformation at each point. Now use the chain rule.)

We can also define scaling on $T_p(\Sigma)$. We define $r[f]$ to be the equivalence class of the curve which has r times the velocity of f at 0, measured in any coordinate chart. Again, this is well defined because the overlap functions are diffeomorphisms.

All in all, $T_p(\Sigma)$ is a vector space for each $p \in \Sigma$.

Exercise 9. Prove that $T_p(\Sigma)$ is isomorphic to $\boldsymbol{R}^2$.

11.6 Riemannian Surfaces

Suppose that Σ is a smooth surface. This means that we have a (maximal) atlas on Σ whose overlap functions are smooth diffeomorphisms. Suppose, for each coordinate chart (U, h), we choose a Riemannian metric on $\boldsymbol{R}^2$. We say that our choice is *consistent* if all the overlap functions are Riemannian isometries relative to the choices. Thus, the overlap function h_{12} considered above is a Riemannian isometry from V_1 to V_2, when V_1 is equipped with the Riemannian metric associated to (U_1, h_1) and V_2 is equipped with the Riemannian metric associated to (U_2, h_2).

A *Riemannian metric* on Σ is a consistent choice of Riemannian metrics on $\boldsymbol{R}^2$, one per coordinate chart. This definition is pretty abstract, so I will give a second definition at the end of this section which is perhaps more concrete.

Let $f : [a, b] \to \Sigma$ be a smooth curve. We can define the *Riemannian length* of f as follows: First of all, we can find a partition $a = t_0 < \cdots < t_n = b$ such that $f([t_i, t_{i+1}])$ is contained in a coordinate chart (U_i, h_i). Next, we can define L_i to be the Riemannian length of

$$h_i \circ f([t_i, t_{i+1}]).$$

Finally, we define the length of f to be $L_0 + \cdots + L_n$. In other words, we compute the lengths of a bunch of little pieces of f and then add them together.

Lemma 11.1 *The Riemannian length of f is well defined, independent of the choices made in its definition.*

Proof: Suppose first of all that we keep the partition the same but use new coordinate charts (U_i', h_i') such that $f([t_i, t_{i+1}]) \subset U_i'$. Then, on $[t_i, t_{i+1}]$ we have

$$h_i' \circ f = (h_i' \circ h_i) \circ (h_i \circ f).$$

But the map $h_i' \circ h_i$ is an overlap function and is an isometry relative to the two Riemannian metrics. Thus $L_i = L_i'$. This shows that the Riemannian length of f does not change if we use different coordinate charts from our atlas.

Suppose now that $a = s_0 < ... < s_m = b$ is another partition, and we are using a different sequence $\{(U_i', h_i')\}$ of coordinate charts to calculate the length. Then by considering all the s_i and also all the t_j (from our original partition) we can find a *refinement* $a = u_0 < \cdots < u_l = b$ which contains all the s_i and also all the t_j. (Basically, we just take the collection of all the numbers and then resort them.)

We can use the charts (U_i, h_i) to compute the length relative to the u-partition, and we will get the same answer as if we used the t-partition. The point here is just that integration is additive:

$$\int_{t_i}^{t_{i+1}} = \int_{t_1}^{u_{k+1}} + \cdots + \int_{u_{k+h-1}}^{t_{i+1}}.$$

Here $t_i = u_k < \cdots < u_{k+h} = t_{i+1}$. Likewise, we can use the charts (U_i', h_i') to compute the length relative to the u-partition, and we will get the same answer as if we used the s-partition. Thus, we reduce to the previously considered case where the partition is the same but the charts change. ♠

Here is another point of view. The object $T_p(\Sigma)$ is a 2-dimensional real vector space for each point $p \in \Sigma$. We could define a Riemannian metric on Σ to be a smoothly varying choice of inner product G_p on the vector space $T_p(\Sigma)$ for each point $p \in \Sigma$. We just have to make sense of the notion of smoothness. If we fix a coordinate chart (U, h), then a Riemannian metric G on Σ gives rise to a Riemannian metric H on $\boldsymbol{R}^2$ as follows. Suppose we have a point $q \in \boldsymbol{R}^2$ and two vectors v, w. Let $p = h^{-1}(q) \in U$ and $[f_1], [f_2] \in T_p(\Sigma)$ be the two classes so that $(h \circ f_1)'(0) = v$ and $(h \circ f_2)'(0) = w$. Then we define $H_q(v, w) = G_p([f_1], [f_2])$. To say that our Riemannian metric on Σ varies smoothly is to say that H is a smooth Riemannian metric on $\boldsymbol{R}^2$ for any choice of coordinate chart. This other definition is completely equivalent to the one I gave above.

Exercise 10. Make up a plausible definition for what a smooth Riemannian n-manifold ought to be, and develop the theory as far as we have done here for surfaces.

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 $\boldsymbol{H}^2$. Say that a *disk-like set* is a subset of the plane that is homeomorphic to an open disk. Say that a map $f : U \to V$ 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 $f : U \to V$ is a local isometry iff f is the restriction of a hyperbolic isometry to U.

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

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

Now we reconcile the two definitions. Suppose that Σ 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 Σ 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 Σ is a hyperbolic surface according to Definition 12.2. According to Exercise 1, the system of coordinate charts on Σ has smooth overlap functions. Therefore, Σ is a smooth surface. We can define a Riemannian metric on Σ 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 $f : U \to \boldsymbol{H}^2$ 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 $\alpha, \beta : (-\epsilon, \epsilon) \to \Sigma$ 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 $\boldsymbol{H}^2$ to the tangent space $T_p\Sigma$ of Σ 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 Σ with the desired properties. Equipped with this metric, Σ 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 $\boldsymbol{H}^2$ 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 $P_1, \ldots, P_n$ 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 2π. This condition guarantees that a neighborhood of each vertex is locally isometric to $\boldsymbol{H}^2$.

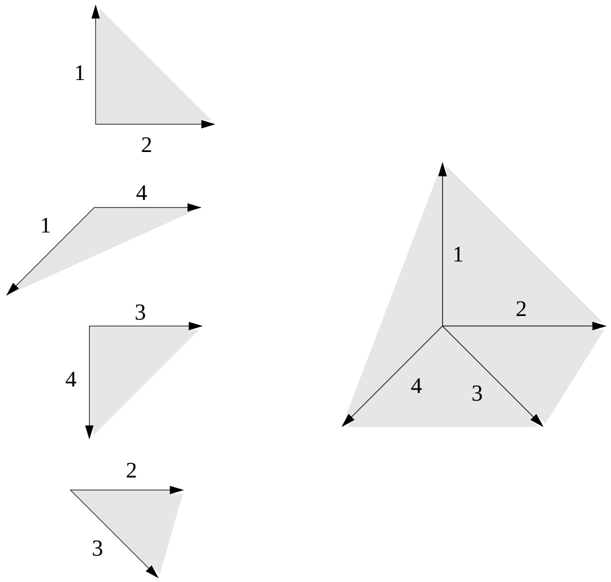

Figure 12.1. A complete circuit

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

$$e_1, e_1', e_2, e_2', e_3, e_3', \ldots, e_k', e_1.$$

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 α_j and we want $\alpha_1+\cdots+\alpha_k=2\pi$.

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 Σ as follows. First of all, we start out with the metric space X which is the disjoint union of $P_1,\ldots,P_n$. 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 λ. 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 Σ 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 Σ. The inclusion map $U_x\to P\subset \boldsymbol{H}^2$ gives a coordinate chart from U_x to $\boldsymbol{H}^2$. 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.

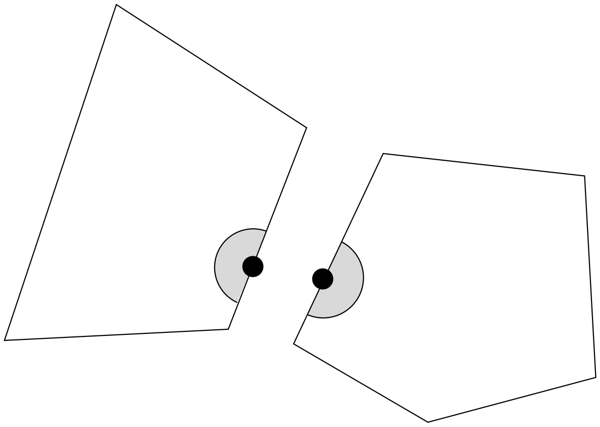

Figure 12.2. Half-disk neighborhoods

We define $h : U \cup U' \to \boldsymbol{H}^2$ so that the following holds.

- 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 Δ to a disk in $\boldsymbol{H}^2$. Finally, prove that Δ is an open neighborhood of p in Σ.

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 $\{q_1, \ldots, q_k\}$ 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 ϵ 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 $h : U_1 \cup \cdots \cup U_k \to \boldsymbol{H}^2$ 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 $\boldsymbol{H}^2$.

Exercise 4. Prove that $(U_1 \cup \ldots \cup U_k)/\sim$ is an open neighborhood of p in Σ and that h gives a homeomorphism from this set onto an open disk in $\boldsymbol{H}^2$. (*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 Σ whose overlap functions are local hyperbolic isometries. We can complete this to a maximal atlas, if we like, using Zorn's lemma. ♠

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 $\boldsymbol{H}^2$.

Let $X \subset \boldsymbol{H}^2$ 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 $\boldsymbol{H}^2$ lying to one side of a hyperbolic geodesic. Geodesic half-planes are convex. Given any two points $p, q \in \boldsymbol{H}^2$, 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. ♠

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 $\boldsymbol{H}^2$. 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 $\boldsymbol{H}^2$.) 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. ♠

Theorem 12.3 allows 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

$$\chi(S) = F - E + V, \tag{43}$$

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 2π. Putting these equations together, we get

$$\chi(S) = -\frac{F}{2} + V = -\frac{F}{2} + \frac{1}{2\pi}\sum_{\text{angles}} \theta_i. \tag{44}$$

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

$$-2\pi\chi(S) = \pi\Big(F - \sum_{\text{angles}} \theta_i\Big) =$$

$$\sum_{\text{triangles}} \Big(\pi - \theta_1(\tau) - \theta_2(\tau) - \theta_3(\tau)\Big) =^*$$

$$\sum_{\text{triangles}} \text{area}(\tau) = \text{area}(S).$$

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 $\widetilde{X}$ such that the covering map $E : \widetilde{X} \to X$ 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 $\widetilde{X}$ is a covering space of X. Then one can make $\widetilde{X}$ into a Riemannian manifold in such a way that the covering map $E : \widetilde{X} \to X$ is a Riemannian cover.*

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

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

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

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

There is a second way to think about the Riemannian metric on $\widetilde{X}$ which perhaps is more clear. The Riemannian metric on X is just a collection of Riemannian metrics on various open sets of $\boldsymbol{R}^n$ 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 $\widetilde{X}$. Since the overlap functions for the charts on $\widetilde{X}$ are the same as on X, the same collection of compatible metrics defines a Riemannian metric on $\widetilde{X}$. ♠

Exercise 9. Show that a Riemannian covering map $E : \widetilde{X} \to X$ is distance nonincreasing. Also, give an example of a Riemannian covering from a connected space $\widetilde{X}$ to a connected space X that is not a global isometry. That is, give an example where there are points $\widetilde{x}, \widetilde{y} \in \widetilde{x}$ 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 completness*, 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 $E : \widetilde{X} \to X$ be a Riemannian covering space. If X is complete, then so is $\widetilde{X}$.*

Proof: Let $\{\widetilde{x}_n\}$ be a Cauchy sequence in $\widetilde{X}$. We have constructed things in such a way that the map $E : \widetilde{X} \to X$ is distance nonincreasing. Setting

$x_n = E(\widetilde{x}_n)$, 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 $\widetilde{x}_n$ lie in the same component of $\widetilde{E}^{-1}(U)$ for n large. But $E : \widetilde{U} \to U$ 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 $\{\widetilde{x}_n\}$ is a convergent sequence in $\widetilde{U}$. ♠

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 $\boldsymbol{H}^2$. Just for this section, let $\boldsymbol{H} = \boldsymbol{H}^2$ stand for the hyperbolic plane.

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

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 $\mathbf{h} \in \boldsymbol{H}$ 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 neigborhood $\boldsymbol{U} \subset \boldsymbol{H}$ of $\mathbf{h}$. Let $x \in H$ be any point. We can take I to be orientation preserving.

Let γ be a continuous path connecting h to x.

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

Proof: We think of γ as a map from $[0,1]$ to $\boldsymbol{H}$, 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 γ 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 $s_n \to t$. Then

$\{\gamma(s_n)\}$ is a Cauchy sequence. Since I is not distance increasing, $\{I(\gamma(s_n)\}$ is also a Cauchy sequence. Since $\boldsymbol{H}$ 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 $\boldsymbol{H}^2$. 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 $J = [0, 1]$. ♠

Now we have a candidate map $I : H \to \boldsymbol{H}$. 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 γ joining h to x. This is where we use the simple connectivity assumption.

Let γ_0 and γ_1 be two paths joining h to x. We think of γ_0 and γ_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 γ_t from γ_0 to γ_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 γ_s and γ_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 $I : H \to \boldsymbol{H}$. But the existence of our extension just used the following.

- Completeness of $\boldsymbol{H}$.

- Local homogeneity of $\boldsymbol{H}$, 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 $\boldsymbol{H}$, 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 $f_0 : [0,1] \to X$. 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 2ϵ that works for all points of $f_0[0,1]$. Let $f_1 : [0,1] \to X$ 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 ϵ. For each $t \in [0,1]$ there is a geodesic $g_t[0,1] \to X$ connecting $f_0(t)$ to $f_1(t)$ that remains within the ϵ-ball about $f_0(t)$.

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

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

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 $\widetilde{X}$ and a covering map $E : \widetilde{X} \to X$. The space $\widetilde{X}$ is complete by Lemma 12.6. But then, by Hadamard's Theorem, $\widetilde{X}$ is isometric to $\boldsymbol{H}^2$. ♠

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 $\widetilde{X} \to X$ be a Riemannian covering of a complete Riemannian manifold X. Let U be a simply connected open subset of X. Let $\widetilde{U} = E^{-1}(U)$. Prove that U is evenly covered by $\widetilde{U}$ and that the restriction of E to any component of $\widetilde{U}$ 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 $\boldsymbol{H}^2$ by pulling them back by the map E. By Exercise 10, the result is an infinite collection of open right angled hexagons $\boldsymbol{H}^2$.

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 $\boldsymbol{H}^2$ are the interfaces between the open hexagons. The whole picture fits together to give a tiling of $\boldsymbol{H}^2$ 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 $\boldsymbol{H}^2$ by hexagons, we can imitate the construction in §6.3, directly producing the covering map from $\boldsymbol{H}^2$ to the surface.

Given that we can construct the universal cover $E : \boldsymbol{H}^2 \to X$ 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.

13 A Primer on Complex Analysis

The purpose of this chapter is to present some of the foundational results in complex analysis. I have tried to write this chapter in such a way that someone who knows no complex analysis could follow along. However, the development here is rather rapid and terse. The ideal reader is a person who has already taken a semester of complex analysis, but who perhaps does not remember the proofs of the main results. This chapter collects the basic results in one place. All the material here can be found in any book on complex analysis; see, e.g., [AHL].

13.1 Basic Definitions

Throughout the chapter U will denote an open subset of $\boldsymbol{C}$, the complex plane. Let $f : U \to \boldsymbol{C}$ be a continuous map. We say that f *has a complex derivative* at $z \in U$ if the quotient

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

exists and is finite. Note that h is allowed to be a complex number. f is said to be *complex analytic* in U if $f'(z)$ exists for all $z \in U$ and the function $z \to f'(z)$ varies continuously in U. Complex analytic functions are sometimes called *holomorphic functions*. The two terms are synonyms.

Complex analysis is mainly the study of complex analytic functions. In this chapter we will discuss complex analytic functions from 3 points of view:

- A complex analytic function is a function that has a complex derivative at each point, as we have just discussed.
- A complex analytic function is a function which satisfies the Cauchy Integral Formula.
- A complex analytic function is a function which agrees with its Taylor series in a neighborhood of each point.

Each of these concepts brings out a different characteristic of a complex analytic function. A major part of an undergraduate complex analysis course is explaining why these three definitions are the same. Among other things, we will establish the equivalence of the 3 definitions in this chapter.

Here is an overview of this chapter. The next several sections lead up to the Cauchy Integral Formula. Once we establish the Cauchy Integral Formula, we will prove a number of results about complex analytic functions. We will consider the connection to power series at the end.

Exercise 1. Suppose that f and g are complex analytic in U and g is never 0 in U. Prove that the functions $f+g$ and $f-g$ and fg and f/g are all complex analytic in U. Conclude that any function $P(z)/Q(z)$, where P and Q are polynomials, is complex analytic away from the roots of Q.

Exercise 2. Suppose that f is complex analytic on U and g is complex analytic on V and $f(U) \subset V$. Prove $g \circ f$ is complex analytic and the complex derivative satisfies $(g \circ f)'(z) = g'(f(z))f'(z)$. This is the *chain rule.*

Being a complex analytic map is rather special. For instance, the function $f(z) = z^2 + 3\overline{z}$ is not complex analytic in $\boldsymbol{C}$. So, not all smooth maps are complex analytic.

We can think of a complex analytic function f as a map from $\boldsymbol{R}^2$ to $\boldsymbol{R}^2$ by writing

$$f(x+iy) = u(x+iy) + iv(x+iy).$$

Recall that f is differentiable at the point (x,y) if the matrix of partial derivatives

$$df = \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix}$$

exists at $p = (x,y)$ and

$$\lim_{t \to 0} \frac{f(p+tv) - f(v)}{t} = df|_p(v).$$

Here $t \in \boldsymbol{R}$. To say that f has a complex derivative at $z = x+iy$ is the same as saying that f is differentiable and $df|_p$ is the composition of a rotation and a dilation. That is

$$\begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix} = \begin{bmatrix} r\cos(\theta) & r\sin(\theta) \\ -r\sin(\theta) & r\cos(\theta) \end{bmatrix}, \qquad r \in \boldsymbol{R}, \quad \theta \in [0, 2\pi).$$

Equating terms, we get

$$u_x = v_y, \qquad u_v = -v_x.$$

These are called the *Cauchy–Riemann equations*. Thus, if f is complex analytic, then its first partials vary continuously and satisfy the Cauchy–Riemann equations.

The converse is also true: f is complex analytic provided that df exists, is continuous, and satisfies the Cauchy–Riemann equations.

13.2 Cauchy's Theorem

Suppose γ is a smooth oriented arc in $\boldsymbol{C}$ and f is a complex valued function defined in a neighborhood of γ. We define a complex line integral along γ as follows. Letting $g : [a, b] \to \gamma$ be a smooth parametrization of γ that respects the orientation of γ, we define

$$\int_\gamma f \ dz = \int_a^b f(g(t)) \frac{dg}{dt} \ dt.$$

The same argument as in §8.6 shows that the answer only depends on γ and not the parametrization. Also, were we to switch the orientation, the value of the line integral would switch signs.

Exercise 4. Let λ be a counterclockwise oriented circle centered at 0, and let $f(z) = 1/z$. Prove that $\int_\lambda f \ dz = 2\pi i$.

If we have a finite union $\gamma = \{\gamma_j\}$ of smooth oriented arcs, we define

$$\int_\gamma f \ dz = \sum_j \int_{\gamma_j} f.$$

In particular, we want to consider the case when γ is a *circular polygon*. A circular polygon is an embedded loop made by concatenating line segments and arcs of circles; see Figure 13.1 for an example.

Theorem 13.1 (Cauchy) *Let γ be a circular polygon. Suppose that f is complex analytic in a neighborhood of the domain bounded by γ. Then $\int_\gamma f \ dz = 0$.*

Proof: Let $f = u + iv$. Letting dx and dy be the usual line elements, we can write

$$\int_{\partial D} f \ dz = \int_{\partial D} (u + iv)(dx + idy) = \int_{\partial D} (udx - vdy) + i \int_D (vdx + udy).$$

By Green's theorem, the integral on the right-hand side equals

$$\int_D (u_y + v_x)dxdy + i\int_D (u_x - v_y)dxdy.$$

Both pieces vanish, due to the Cauchy–Riemann equations. ♠

Remark: In §8.7 we proved Green's Theorem for polygons. The case of circular polygons follows from the polygon case and a straightforward limiting argument. Alternatively, most books on multivariable calculus have a proof of Green's Theorem in great generality; see, e.g., [SPI]. Cauchy's Theorem holds in the same generality that Green's Theorem holds, but the version we state is sufficient for all the applications we give.

13.3 The Cauchy Integral Formula

Here is the beautiful *Cauchy Integral Formula.*

Theorem 13.2 (Cauchy Integral Formula) *Let γ be a circular polygon, oriented counterclockwise around the domain D that it bounds. Let $a \in D-\gamma$. Suppose that f is complex analytic in a neighborhood U of D. Then*

$$f(a) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z-a}dz. \tag{45}$$

Proof: We translate the whole picture and consider without loss of generality the case when $a = 0$. The function $g(z) = f(z)/z$ is complex analytic in $U - \{0\}$. Let β be the circular polygon shown in Figure 13.1.

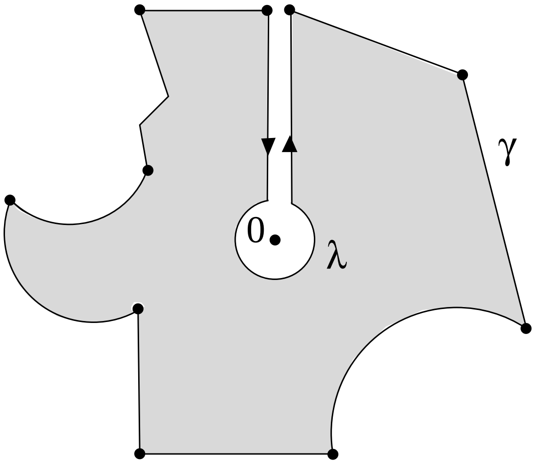

Figure 13.1.

We have

$$\int_\beta g\ dz = 0 \tag{46}$$

by Cauchy's Theorem. We allow the two oppositely oriented vertical segments in β to approach each other. In the limit, the contributions from the two vertical segments cancel out, and equation (46) yields

$$\int_\gamma g(z) = \int_\lambda g(z). \tag{47}$$

Here λ is a counterclockwise circle centered at 0. Define

$$I = \Big|\int_\gamma g(z)dz - 2\pi i f(0)\Big|. \tag{48}$$

We want to show that $I = 0$. Combining Exercise 4 and Equation 47, we have

$$I = \Big|\int_\gamma g(z)dz - f(0)\int_\lambda \frac{dz}{z}\Big| = \Big|\int_\lambda \frac{f(z)}{z}dz - \int_\lambda \frac{f(0)}{z}dz\Big|. \tag{49}$$

Now we have a bound on I that is expressed entirely in terms of λ. Rearranging the terms of the last integral, we have

$$I = \Big|\int_\lambda \frac{f(z) - f(0)}{z}dz\Big| \leq \text{length}(\lambda) \times 2|f'(0)|. \tag{50}$$

The last inequality holds once λ is sufficiently small. Letting λ shrink to a point, we see that $I = 0$, as desied. ♠

13.4 Differentiability

Here we use the Cauchy Integral Formula to prove some results about the differentiability of complex analytic functions. Our first result is not so important in itself, but it illustrates how one uses the Cauchy Integral Formula to get a formula for the derivative of a complex analytic function.

Theorem 13.3 *Suppose that f is a complex valued and continuously differentiable function defined in an open set U. If f satisfies the Cauchy Integral Formula with respect to every circle in U, then f is complex analytic in U.*

Proof: Let $a \in U$ and let $\gamma \subset U$ be a circle surrounding a. Using the Cauchy Integral Formula, we compute

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h} = \lim_{h\to 0}\frac{1}{2\pi i h}\Big(\int_\gamma \frac{f(z)}{z-a-h}dz - \int_\gamma \frac{f(z)}{z-a}dz\Big) =$$

$$\lim_{h\to 0}\frac{1}{2\pi i}\int_\gamma \frac{f(z)}{(z-a)(z-a-h)}dz == \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{(z-a)^2}dz. \qquad (51)$$

This tells us that f has a complex derivative at a and also gives a formula for it. ♠

Theorem 13.4 *Suppose that f is a complex analytic function defined in an open set U. Then f' is also complex analytic in U.*

Proof: Note that f' exists just by virtue of the fact that f is complex analytic. Since f is complex analytic in U, Theorem 13.3 holds for f. Equation (51) gives us a formula for f'. We compute

$$\lim_{h\to 0}\frac{f'(a+h)-f'(a)}{h} = \lim_{h\to 0}\frac{1}{2\pi i h}\Big(\int_\gamma \frac{f(z)}{(z-a-h)^2}dz - \int_\gamma \frac{f(z)}{(z-a)^2}dz\Big)$$

$$= \frac{2}{2\pi i}\int_\gamma \frac{f(z)}{(z-a)^3}dz. \qquad (52)$$

Here γ is some circle that surrounds a. Hence f' has a complex derivative throughout U and equation (52) gives a formula for it. In light of equation (52), the function f'' is continuous. Hence f' is complex analytic in U. ♠

An immediate corollary is that complex analytic functions are infinitely differentiable. The calculation in equation (52), when done inductively, yields the following formula for the nth derivative of a complex analytic function f.

$$f^{(n)}(a) = \frac{n!}{2\pi i}\int_\gamma \frac{f(z)}{(z-a)^{n+1}}dz. \qquad (53)$$

13.5 The Maximum Principle

Let f be a complex analytic function in a connected open set U. Here we will show that f cannot take on its maximum value at a point in U unless f is constant. We will assume that f takes on a maximum at some point $a \in U$, and we will derive a contradiction. If f has an interior maximum, we can compose f with translations and dilations and arrange the following properties.

- $|f(0)| = 1$.
- U contains the unit disk.
- $|F(z)| \leq 1$ for all $|z| = 1$.
- $|F(z)| < 1$ for some z such that $|z| = 1$.

Let γ be the unit circle. By the Cauchy Integral Formula we have

$$1 = |f(0)| = \frac{1}{2\pi}\Big|\int_\gamma \frac{f(z)}{z}\Big| \leq^* \frac{1}{2\pi}\int_\gamma |f(z)|dz < 1.$$

This is a contradiction. The starred inequality is essentially the triangle inequality. For later purposes we work out some consequences of the Maximum Principle.

Lemma 13.5 *Suppose that $f(z)/z^n$ is well defined at 0 and complex analytic in a neighborhood of the unit disk. Then $f(z) \leq M|z|^n$, where M is the maximum value of $|f(z)|$ on the unit circle.*

Proof: From the Maximum Principle we get the result that

$$|f(z)|/|z^n| \leq M.$$

Hence $|f(z)| \leq M|z|^n$. ♠

Lemma 13.6 *Suppose, for all n, that the function $f(z)/z^n$ is well defined at 0 and complex analytic in a neighborhood of the unit disk. Then f is identically 0 on the unit disk.*

Proof: From the preceding result, we have $|f(z)| \leq M|z|^n$. If $|z| < 1$, then

$$\lim_{n\to\infty} M|z|^n = 0.$$

Hence $|f(z)| = 0$ if $|z| < 1$. By continuity, $|f(z)| = 0$ if $|z| \leq 1$. ♠

13.6 Removable Singularities

Here we will prove the following result:

Theorem 13.7 *Let U be an open set that contains a point b. Suppose that f is complex analytic and bounded on $U - \{b\}$. Then $f(b)$ can be (uniquely) defined so that f is complex analytic in U.*

Proof: Let γ and β and λ be the loops used to prove the Cauchy Integral Formula. So, λ is a small loop surrounding b and γ is a big loop surrounding b. Let $|\lambda|$ denote the radius of λ. Let D be the open domain bounded by γ. We define $g : D \to \boldsymbol{C}$ by the integral

$$g(a) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z-a}dz.$$

The same calculation as in the proof of Theorem 13.4 shows that g is complex analytic on all of D. We will show that $f(a) = g(a)$ for all $a \in D - \{b\}$. Once we know this, we set $f(b) = g(b)$ and we are done.

Now suppose that $a \neq b$. Since $f(z)$ is bounded in a neighborhood of b we have

$$\lim_{|\lambda|\to 0}\int_\lambda \frac{f(z)}{z-a}dz = 0.$$

But, by the Cauchy Integral Formula,

$$f(a) = \frac{1}{2\pi i}\int_\beta \frac{f(z)}{z-a}dz$$

no matter which choice of λ we make. Therefore

$$f(a) = \lim_{|\lambda|\to 0}\frac{1}{2\pi i}\int_\beta \frac{f(z)}{z-a}dz = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z-a}dz = g(a).$$

So $f(a) = g(a)$ for all $a \in D - \{b\}$. ♠

Lemma 13.8 *Let D denote the unit disk. Suppose that f is complex analytic in a neighborhood of D and $|f(z)|/z^n$ is bounded on $D - \{0\}$. Then f is identically 0 on D.*

Proof: The function $f(z)/z^n$ is complex analytic in the unit disk by the above result. Lemma 13.6 now says that f is identically 0. ♠

13.7 Power Series

We say that a sequence $\{a_n\}$ of complex numbers satisfies the *unit convergence condition* (or UCC) if

$$\lim_{n\to\infty} a_n\rho^n = 0, \qquad \forall\rho\in[0,1). \tag{54}$$

The UCC implies that the terms in the sequence $\{|a_n|\rho^n\}$ decay exponentially fast for any $\rho < 1$. To see this, we choose any $\rho^* \in (\rho, 1)$ and note that

$$|a_n|\rho^n = |a_n|(\rho^*)^n \times \Big(\frac{\rho}{\rho^*}\Big)^n < \Big(\frac{\rho}{\rho^*}\Big)^n$$

for n sufficiently large.

Exercise 5. Suppose that $\{a_n\}$ satisfies the UCC. Let $k > 0$ be any integer and let C be any constant. Prove that the sequence $\{Cn^k a_n\}$ also satisfies the UCC.

Now we will discuss the convergence of power series to complex analytic functions, as well as the term-by-term differentiation of these series. Let $\{a_n\}$ be a sequence satisfying the UCC. First, we define a “finite series”, which is just a polynomial.

$$f_n(z) = \sum_{k=0}^{n} a_k z^k. \tag{55}$$

Lemma 13.9 *The sequence $\{f_n(z)\}$ is a Cauchy sequence of complex numbers for all $|z| < 1$.*

Proof: If $a, b > N$ and N is sufficiently large, then

$$|f_a(z) - f_b(z)| = |\sum_{n=a}^{b} a_n z^n| \le \sum_{n=a}^{b} |a_n||z|^n \le \sum_{N}^{\infty} \delta^n = \frac{\delta^n}{1-\delta}.$$

Here we have chosen some $\rho^* > |z|$ and taken $\delta = |z|/\rho^*$. This calculation establishes what we want. ♠

Lemma 13.9 says that the limit

$$f(z) = \sum a_n z^n = \lim_{n\to\infty} f_n(z) \tag{56}$$

exists provided that $|z| < 1$. Here is our main result about this infinite series.

Theorem 13.10 *$f(z)$ is complex analytic in the open unit disk and $f'(z)$ is obtained by differentiating the series term-by-term.*

Proof: Let $g_N = f - f_N$. Then

$$\frac{f(z+h)-f(z)}{h} = \frac{f_N(z+h)-f_N(z)}{h} + \frac{g_N(z+h)-g_N(z)}{h}.$$

From Exercise 1 above $f_N(z)$ is complex analytic. Also, the sequence $\{na_n\}$ satisfies the UCC by Exercise 4. Hence, $\lim_{N\to\infty} f'_N(z)$ exists at every point in the unit disk. Moreover, this limit is just obtained by differentiating the series for $f(z)$ term by term. To prove our result we just have to show that

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

This is the same as showing that

$$\lim_{N\to\infty}\lim_{h\to 0}\frac{g_N(z+h)-g_N(z)}{h} = 0.$$

On the individual terms we have the bound

$$|\frac{a_n(z+h)^n - a_n z^n}{h}| = |a_n||\frac{(z+h)^n - z^n}{h}| \leq^* n|a_n||z+h|^{n-1}.$$

The starred inequality comes from the fact that the map $\phi(z) = z^n$ expands distances in $\boldsymbol{C}$ by at most $n\delta^{n-1}$ as long as $|z| \leq \delta$.

As long as h is fairly small, we can choose some $\delta < 1$ and restrict our attention to the case $|z+h| < \delta < 1$. Given the above estimate, we get

$$|\frac{g_N(z+h)-g_N(z)}{h}| \leq \sum_{n=N}^{\infty} n|a_n|\delta^{n-1} = \sum_{n=N}^{\infty} n\delta|a_n|\delta^n = R_N.$$

(We are just calling the last expression R_N for convenience.) But the sequence $\{n\delta|a_n|\}$ satisfies the UCC by Exercise 5. Hence, the terms comprising R_N decay exponentially. Hence, $\lim_{N\to\infty} R_N = 0$. But the inequality above holds for any h with $|z+h| < \delta$. Hence

$$\lim_{N\to\infty}\lim_{h\to 0}|\frac{g_N(z+h)-g_N(z)}{h}| \leq \lim_{N\to\infty} R_N = 0.$$

This is what we wanted to prove. ♠

The above result, applied iteratively, shows that the kth complex derivative $f^{(k)}(z)$ is complex analytic in the open unit disk and is obtained by differentiating the series for $f(z)$ term-by-term k times.

Our discussion, which focused on the unit disk, generalizes in a straightforward way. Say that the sequence $\{b_n\}$ satisfies the *R-convergence criterion* if the sequence $\{a_n R^n\}$ satisfies the UCC. In this case the series $\sum b_n(z-z_0)^n$ is complex analytic in the open disk of radius R about z_0 and the same result as above applies.

13.8 Taylor Series

The basic result we want to prove is that a complex analytic function equals its Taylor series. We begin with a technical lemma.

Lemma 13.11 *Suppose that f is complex analytic in a neighborhood of the unit disk. Then the sequence*

$$\{f^{(n)}(0)/n!\}$$

is bounded and hence satisfies the UCC.

Proof: It follows immediately from equation (53) that $|f^{(n)}| \leq Mn!$, where M is the maximum value attained by f on the closed unit disk. ♠

Lemma 13.11 says that the Taylor series for f about 0 defines a power series which converges and is complex analytic in a neighborhood of the unit disk. The next result says that f coincides with its Taylor series in the unit disk.

Theorem 13.12 *Suppose that f is complex analytic in a neighborhood of the unit disk. Then f equals its Taylor series on the unit disk.*

Proof: Since the Taylor series $\widetilde{f}$ of f is defined and complex analytic on the unit disk, we can consider the difference function $f-\widetilde{f}$. This complex analytic function has zero Taylor series. Thus, it suffices to prove the following special case. If the Taylor series of f vanishes identically at 0, then f is zero on the whole unit disk.

If g is any function with $g(0) = 0$, we have

$$|g(z)| \leq \int_0^1 |g'(tz)|dt. \tag{57}$$

Here $g'(tz)$ is the complex derivative of the function $z \to g(tz)$. Equation (57) is best seen geometrically. The idea is that $|g'(tz)|$ measures the speed of the curve $t \to g(tz)$ which connects 0 to $g(z)$.

Let Δ be the closed unit disk. Fix n for the moment. Since $f^n(0) = 0$ we can choose $\delta > 0$ so that $|f^{(n)}(z)| < 1$ for all $|z| < \delta$. Applying equation (57) to $g = f^{(n-1)}$, we get

$$|f^{(n-1)}(z)| \leq |z|, \qquad \forall |z| \leq \delta. \tag{58}$$

Applying equation (57) to $g = f^{(n-2)}$ and using the bound in equation (58), we get

$$f^{(n-2)} \leq |z|^2/2, \qquad \forall |z| \leq \delta. \tag{59}$$

Continuing in this way, we get

$$|f(z)| \leq |z|^n/n!, \qquad \forall |z| \leq \delta. \tag{60}$$

In particular, $|f(z)|/|z|^n$ is bounded on $D_n - \{0\}$, where D_n is the disk of radius δ. Note that δ depends on n, but this does not bother us. By compactness, $|f(z)|/|z|^n$ is bounded on $\Delta - D_n$. Hence $|f(z)|/|z^n|$ is bounded on $\Delta - \{0\}$. Since this holds for all n, Lemma 13.8 says that f is identically 0 on the unit disk. ♠

Exercise 6. Define the *exponential function*

$$E(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

Prove that the series defining $E(z)$ converges on all of $\boldsymbol{C}$. Prove also that $E'(z) = E(z)$ and that $E(z_1 + z_2) = E(z_1)E(z_2)$. For this last part, you can do it by manipulating the series directly and applying the binomial theorem. The restriction of E to $\boldsymbol{R}$ coincides with the familiar exponential function.

Exercise 7. Define the two functions

$$C(z) = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} \cdots, \qquad S(z) = z - \frac{z^z}{z!} + \frac{z^5}{5!} - \frac{z^7}{7!} \cdots.$$

Show that these series converge for all $z \in \boldsymbol{C}$ and that $C(x) = \cos(x)$ and $S(x) = \sin(x)$ for all $x \in \boldsymbol{R}$. Verify that $E(z) = C(z) + iS(z)$.

Exercise 8. Let us define $\cos(x)$ and $\sin(x)$ such that the map

$$\gamma_0(x) = (\cos(x), \sin(x))$$

is the unit speed counterclockwise parametrization of the unit circle such that $\gamma(0) = (1, 0)$. Prove that $C(x) = \cos(x)$ and $S(x) = \sin(x)$ for all $x \in \boldsymbol{R}$. (*Hint*: Consider the map $\gamma_1(x) = (C(x), S(x))$. Check that

$$\frac{d}{dx}\Big(C^2(x) + S^2(x)\Big) = 0$$

using term-by-term differentiation. From here it is not too hard to show that γ_0 and γ_1 are the same parametrization of the unit circle.)

Exercise 9. Our main result in this section is definitely false for smooth functions that are not complex analytic. Consider the function

$$f(t) = \exp(-1/t^2), \qquad t > 0.$$

When $t \leq 0$ we define $f(t) = 0$. Prove that f is smooth and has a trivial Taylor series about 0. This shows that smooth functions need not equal their Taylor series.

14 Disk and Plane Rigidity

In this chapter, we apply some of the complex analysis developed in the previous chapter, notably the Maximum Principle and Theorem 13.7, to certain holomorphic maps of the disk and plane. The types of results we prove show that certain weak-seeming conditions placed on a complex analytic function actually place very strong restrictions on the function. These kinds of rigidity results provide a link between complex analysis and geoemtry.

As an application of the results, we will prove that stereographic projection maps circles on S^2 to circles in $\boldsymbol{C} \cup \infty$. While not the most elementary possible proof, our proof does give an application of the complex analysis we have been developing. For a geometric proof of the main result, see [HCV].

14.1 Disk Rigidity

We first prove Theorem 1.1, mentioned in Chapter 1.

Theorem 14.1 *Let f be biholomorphism from the unit disk to itself. Then f is a Möbius transformation.*

Proof: If $f(0) \neq 0$, then we can find a linear fractional automorphism H of Δ such that $f \circ H(0) = 0$. Thus, it suffices to consider the case when $f(0) = 0$. Since $f'(0)$ exists, the function $g(z) = f(z)/z$ is bounded in Δ. Hence, this function is complex analytic. Below we will show that $|f(z)| \leq |z|$ for all $z \in \Delta$, and the same argument, applied to f^{-1}, shows that $|f^{-1}(z)| \leq |z|$ for all $z \in \Delta$. These two inequalities show that $|f(z)| = |z|$ on Δ. But then $g(\Delta)$ is contained in the unit circle, a 1-dimensional curve. This is impossible unless g is a constant map. Hence there is a constant C such that $f(z) = Cz$. Hence f is a linear fractional transformation.

It remains to show that $|f(z)| \leq |z|$ for $z \in \Delta$. This is the same as showing that $|g(z)| \leq 1$ for all $z \in \Delta$. Let C_r be the circle of radius $r < 1$ about 0. Then $|g(z)| \leq 1/r$ on C_r. Hence $|g(z)| \leq 1/r$ if $|z| < r$ by the Maximum Principle. Letting $r \to 1$, we see that $|g(z)| \leq 1$ on Δ. This is what we wanted to prove. ♠

Exercise 1. Prove the same result for a biholomorphism from the upper halfplane to itself.

The next result shows the distinguished role played by the hyperbolic metric on the open unit disk, from the point of view of complex analysis.

Lemma 14.2 *Let Δ be the unit disk, equipped with the hyperbolic metric from §10.7. Let $f : \Delta \to \Delta$ be a complex analytic map, not necessarily a biholomorphism. Then f does not expand distances in the hyperbolic metric.*

Proof: We would like to see, at each point $p \in \Delta$, that the differential df maps vectors having hyperbolic length 1 to vectors having hyperbolic length at most 1. Call this the *no-stretch property*. It suffices to prove that the no-stretch property holds for each $p \in \Delta$. We can find Möbius transformations T_1 and T_2 such that $T_1(0) = p$ and $T_2(f(p)) = 0$, respectively. The map $g = T_2 \circ f \circ T_1$ satisfies $g(0) = 0$. Since T_1 and T_2 are hyperbolic isometries, g has the no-stretch property at 0 if and only if f has the no-stretch property at p. Since $g(0) = 0$, we just need to show that $|g'(0)| \leq 1$ to establish the no-stretch property for g at 0. The same argument as in the previous proof establishes this fact. ♠

Lemma 14.2 is more flexible than it first appears. Any disk W_0 in the plane has its own hyperbolic metric, so that a similarity carrying Δ to W_0 is a hyperbolic isometry. This principle should help you with the next exercise.

Exercise 2. Suppose that U is some open set in the plane and $w \in U$ is some point. Suppose also that $G : U \to \Delta$ is a holomorphic map. Prove that there is some disk $W \subset U$, centered at w, whose size does not depend on G, such that the hyperbolic distance from $G(w)$ to $G(w')$ is less than 1 for all $w' \in W$.

14.2 Liouville's Theorem

Here is Liouville's Theorem.

Theorem 14.3 (Liouville) *Suppose that f is a bounded holomorphic function on $\boldsymbol{C}$. Then f is constant.*

Proof: Equation (51) says that

$$f'(a) = \int_\gamma \frac{f(z)}{(z-a)^2} dz. \tag{61}$$

Taking γ to be a large circle of radius r about 0, we see that the right hand side of the above equation is at most C/r for some constant C. Letting $r \to \infty$, we see that $f'(a) = 0$. Since a is artitrary, f is constant. ♠

Exercise 3 (Challenge). A function $f : \boldsymbol{C} \to \boldsymbol{R}$ is called *harmonic* if it has the following property. For any disk D, the value of f at the center of D equals the average value of f on D. Prove that a bounded harmonic function is constant. This result is equivalent to Liouville's Theorem. (*Sketch*: You want to show that $f(a) = f(b)$ for all $a, b \in \boldsymbol{C}$. Consider the difference $C_r = A_r - B_r$, where A_r is the average of f on the disk of radius r about a and B_r is the average of f on the disk of radius r about b. Show that $\lim_{r\to\infty} C_r = 0$, by analyzing the intersection $A_r - B_r$ and observing that there is a lot of cancellation in the computation of $A_r - B_r$ when r is large.)

Exercise 4. Give an alternate proof of Liouville's Theorem by showing that the function $g(z) = f(z)/z$ is holomorphic in the whole plane and then applying the Maximum Principle.

Exercise 5. Use Liouville's Theorem to give another proof of the Fundamental Theorem of Algebra. (*Hint*: Let $P(z)$ be a complex polynomial supposedly with no roots. Consider $f(z) = 1/P(z)$.)

Exercise 6. Suppose that $g : \boldsymbol{C} \to \boldsymbol{C}$ is a continuous map with the following properties:

- $g(0) = 0$.
- g is holomorphic on $\boldsymbol{C} - \{0\}$.
- $|g(z)| < C|z|$ for $|z|$ sufficiently large.

Prove that $g(z) = Az$ for some constant A. Hint: First show that g is holomorphic on all of $\boldsymbol{C}$, then show the same thing for $h(z) = g(z)/z$.

Exercise 7 (Challenge). Suppose that $f : \boldsymbol{C} \to \boldsymbol{C}$ is a holomorphic function such that $|f(z)| < |z|^n$ for some n and all z with $|z|$ sufficiently large. Prove that f is a polynomial.

Lemma 14.4 *Suppose f is a homeomorphism of $\boldsymbol{C}$ that is complex analytic except at finitly many points. Then $f(z) = Az + B$ for some constants A and B.*

Proof: Combining Lemma 2.2 and Theorem 13.7, we see that f is complex analytic on all of $\boldsymbol{C}$. The function $f'(z)$ cannot identically vanish. So, we can compose f with translations and then assume that $f(0) = 0$ and $f'(0) > 0$. But then there is some $C > 0$ such that $|f(z)| > C|z|$ provided that $|z|$ is sufficiently small. Now consider the function

$$g(z) = \frac{1}{f(1/z)}. \tag{62}$$

Note that g satisfies the conditions of Exercise 4. Hence $g(z) = Az$. But then $f(z) = z/A$. Remembering that this new version of f is a translation of the original, we see that the original version of f has the form $Az + B$. ♠

14.3 Stereographic Projection Revisited

Let $\phi : S^2 \to \boldsymbol{C} \cup \infty$ be stereographic projection.

Lemma 14.5 *The differential $d\phi$ is a similarity on the tangent plane T_x at $x \in S^2 - \{(0,0,1)\}$.*

Proof: One can prove this result by a direct calculation, but we will give a geometric proof. Our proof refers to Figure 14.1. We think of $\boldsymbol{C}$ as the xy-plane. Let $T = T_x$ and let T' be the plane through x parallel to $\boldsymbol{C}$. Let L be the line joining $(0,0,1)$ to x. Figure 14.1 shows the intersection of all these objects with the plane Π containing $(0,0,0)$ and $(0,0,1)$ and x. The X Theorem from §8.3 implies that the lines $T \cap \Pi$ and $T' \cap \Pi$ make the same angle with $L = L \cap \Pi$. Hence, reflection in the plane $P = L^{\perp}$ carries T isometrically to T'.

The differential $d\phi$ has the following description: First reflect T to T' through P, then radially project T' to $\boldsymbol{C}$ through p. Thus $d\phi$ is the composition of an isometry and a similiarity, which is just another similarity. ♠

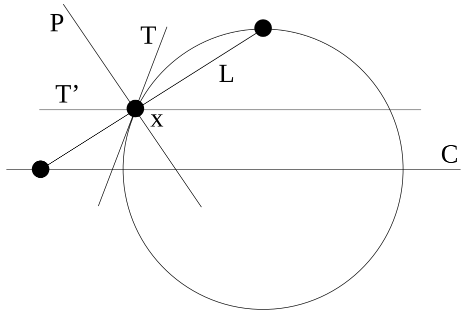

Figure 14.1. The Differential of Stereographic Projection

Exercise 8. Prove Lemma 14.5 by a direct calculation, using equation (32).

Lemma 14.6 *Suppose I is an isometry of S^2. Then $I' = \phi \circ I \circ \phi^{-1}$ is a linear fractional transformation.*

Proof: Here is where complex analysis comes in. We can find a linear fractional transformation T such that $J = T \circ I'$ fixes ∞. It suffices to show that J is a linear fractional transformation. The map J is smooth except at perhaps a finite list of points. (The points we are not certain about are various images and preimages of ∞.) Moreover, by Lemma 14.5, the differential dJ is a similarity at all but finitely many points. Hence J is a homeomorphism of $\boldsymbol{C}$ that is holomorphic except at finitely many points. By Lemma 14.4, the map J is linear, and hence a linear fractional transformation. But then I' is a linear fractional transformation. ♠

Now we come to the main application of the results in this section. Again, this result has a direct geometric proof, but we want to show how one can get the result from complex analysis.

Lemma 14.7 *Stereographic projection maps circles on S^2 to generalized circles in $\boldsymbol{C} \cup \infty$.*

Proof: Let C be a circle on S^2. Let I be an isometry of S^2 such that $I(C)$ contains $(0,0,1)$. As we remarked in §9.5, the curve $L = \phi(I(C))$ is a straight line (union ∞). Thus, $\phi(I(C))$ is a generalized circle. But

$$\phi(L) = I'(\phi(C)), \qquad I' = \phi \circ I \circ \phi^{-1}.$$

By Lemma 14.4, the map I' is a linear fractional transformation. Therefore, so is $J = (I')^{-1}$. But $\phi(C) = J(L)$, where J is a linear fractional transformation and L is a generalized circle. Since linear fractional transformations map generalized circles to generalized circles, as we saw in Chapter 10, we see that $J(L) = \phi(C)$ is also a generalized circle. ♠

Exercise 9. Generalize the definition of stereographic projection so that it works in all dimension and prove that generalized stereographic projection maps spheres to spheres. You should be able to deduce this from the 2-dimensional case and symmetry.

15 The Schwarz–Christoffel Transformation

In this chapter we will study some examples of Schwarz–Christoffel transformations. These maps turn out to give biholomorphisms between the upper half-plane and the interiors of polygons. For ease of exposition, we will restrict our attention to the case when the sides of the polygon are parallel to the coordinate axes. We call such polygons *rectilinear polygons*; see Figure 15.1.

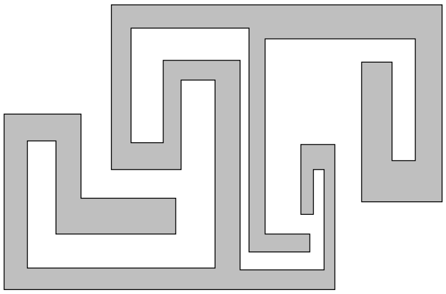

Figure 15.1. A rectilinear polygon

One remarkable thing about the Schwarz–Christoffel transformations is that there is, in a sense, an explicit formula for them. The book [DRT] has a great deal of information about these maps, including a discussion of their history.

15.1 The Basic Construction

Suppose that $x_1 < x_2 < \cdots < x_n \in \boldsymbol{R}$ and $e_1, \cdots, e_n$ are numbers such that $e_j = \pm 1/2$ for all j and $e_1 + \cdots + e_n = -2$.

Let $U \subset \boldsymbol{C}$ denote the upper half-plane. Let $U^* \subset \boldsymbol{C}$ denote the region obtained by deleting the closed downward pointing rays which start at $x_1, \ldots, x_n$. We are mainly interested in U, but the larger region U^* is convenient for technical purposes.

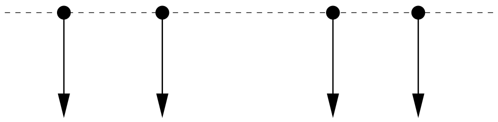

Figure 15.2. The region U^*

Consider the function

$$f(z) = (z - x_1)^{e_1} \cdots (z - x_n)^{e_n}. \tag{63}$$

If we try to define this function in all of $\boldsymbol{C}$ we run into trouble because we cannot consistently define f all the way around a loop which circles around x_j. Since U^* has no loops like this, f is defined and complex analytic in all of U^*.

We define a function $F : U^* \to \boldsymbol{C}$ as follows. First we set $F(i) = 0$. Next, for any $z \in U^*$, we let γ be a piecewise smooth path connecting 0 to z, and we set

$$F(z) = \int_\gamma f(z) dz. \tag{64}$$

Equation 64 is well defined by Theorem 13.1. It follows almost immediately from the Fundamental Theorem of Calculus that F is holomorphic in U^* and $F'(z) = f(z)$. In particular, $F'(z)$ never vanishes in U^*. Here is our main result about F.

Theorem 15.1 *F is well defined and continuous on $\boldsymbol{R} \cup \infty$. The image $F(\boldsymbol{R} \cup \infty)$ is a closed polygonal loop whose sides are alternately parallel to the real and imaginary axes. If $F(\boldsymbol{R} \cup \infty)$ is an embedded polygon, then F is a biholomorphism from U to the polygonal domain bounded by $F(\boldsymbol{R} \cup \infty)$.*

15.2 The Inverse Function Theorem

As a prelude to proving Theorem 15.1. We prove a special case of the Inverse Function Theorem. For the general case, see, e.g., [SPI].

Theorem 15.2 *Let f be a holomorphic map defined in a neighborhood of $z \in \boldsymbol{C}$. Suppose that $f'(z) \neq 0$. Then the restriction of f to a neighborhood of z has an inverse, and f^{-1} is also holomorphic.*

Proof: We can translate and scale so that $z = 0$ and $f(0) = 0$ and $f'(0) = 1$. Let D_r be the disk of radius r about 0. For r small, we have $|f'(z) - 1| < 1/100$ for all $z \in D_r$. Let $z_1 \neq z_2$ be two points in D_r. Let L be the straight line joining these points. Given our bounds on $f'(z)$ along L, we see that the curve $f(L)$ nearly has the same length as L and points almost in the same direction as L at all points. Hence $f(z_1) \neq f(z_2)$. Hence f is injective on D_r for r small.

The same argument shows that $f(\partial D_r)$ is a closed loop that is at least (say) $r/2$ from 0 and winds once around 0. Let Δ_r denote the set of points w such that $f(\partial D_r)$ winds once around w. Note that Δ_r is an open neighborhood of 0. Suppose there is some $w \in \Delta_r - f(D_r)$. Consider the 1-paramater family of loops $\gamma_t = f(t\partial D_r)$. For t close to 0, the loop γ_1 winds 0 times around w. On the other hand, γ_1 winds once around w. In order for the winding number to change in this way, γ_t must contain w for some t. But then $w \in f(D_r)$. Hence $f : D_r \to \Delta_r$ is a surjection.

Now we know that $f : D_r \to \Delta_r$ is a bijection. So, $f^{-1} : \Delta_r \to D_r$ exists. Our injectivity proof also shows that f^{-1} is continuous: f cannot map far away points close together. One way to see that f is differentiable at 0 is that the dilated maps $g_n(z) = nf(z/n)$ converge to a similarity as $n \to \infty$. But the dilated inverse of f is the inverse of the dilation of f. Hence, the dilations of f^{-1} also converge to a similarity. This shows that f^{-1} is differentiable at 0. The chain rule now shows that $(f^{-1})'(0) = 1/f'(0)$. The same argument works at any other point z in the interior of D_n. This shows that f^{-1} has a continuously varying complex derivative in the interior of Δ_n. Hence, f^{-1} is holomorphic in Δ_n. ♠

15.3 Proof of Theorem 15.1

We already know that F is defined on $\boldsymbol{R} - \{x_1, \ldots, x_n\}$, and F is pretty obviously continuous where defined.

Exercise 1. Prove that F is defined and continuous on $\boldsymbol{R} \cup \infty$. (*Hint*:

use the same definition at these points as for the other points. The finiteness of integrals such as

$$\int_0^1 \frac{1}{x^{1/2}} dx; \qquad \int_1^\infty x^{-2}\ dx$$

is what makes the definition work.)

Now we want to analyze the image $F(\boldsymbol{R} \cup \infty)$. The points $x_1, \ldots, x_n$ divide $\boldsymbol{R}$ into the $n+1$ intervals $I_0, \ldots, I_n$. Actually, $I_0 = (-\infty, x_1)$ and $I_n = (x_n, \infty)$ are rays. Let $J_k = F(I_k)$.

When we square f, we get

$$f^2(z) = (z - x_1)^{\pm 1} \cdots (Z - x_n)^{\pm 1}.$$

From this we see that f^2 is positive on I_0, negative on I_1, positive on I_2, and so on. So, f is real on I_0, pure imaginary on I_1, real on I_2, pure imaginary on I_3, and so on. But $F'(z) = f(z)$, and the argument of $F'(z)$ tells us how F rotates points in a neighborhood of z. Hence J_0 is a horizontal segment, J_1 is a vertical segment, J_2 is a horizontal segment, and so on. Since F is continuous on $\boldsymbol{R} \cup \infty$, we see that these segments all piece together to give the kind of path described in Theorem 15.1.

We orient $\boldsymbol{R}$ from $-\infty$ to $+\infty$. If you walk along $\boldsymbol{R}$, then U lies to your left. Being complex analytic, the map F is orientation preserving. This means that, as you walk around $F(\boldsymbol{R})$, the image $F(U)$ (at least locally) lies to your left.

Exercise 2. Show that $F(\boldsymbol{R})$ turns left at x_j if $e_j = -1/2$ and right if $e_j = -1/2$. Geometrically, $f(U)$ looks like one quadrant in a neighborhood of $f(x_j)$ if $e_j = -1/2$ and three quadrants if $e_j = 1/2$.

Given Exercise 2, and the fact that $e_1 + \cdots + e_n = -2$, the polygonal path $F(\boldsymbol{R})$ turns once around counterclockwise (the equivalent of 4 left turns.) Hence $F(I_0)$ and $F(I_n)$ travel in the same direction and fit together seamlessly.

Now we suppose that $F(\boldsymbol{R} \cup \infty)$ is an embedded polygon. Let R be the region bounded by $F(\boldsymbol{R} \cup \infty)$.

Lemma 15.3 *$F(U) \subset R$.*

Proof: Let $\overline{U} = U \cup \boldsymbol{R} \cup \infty$. The set $\overline{U}$ is a compact subset of the Riemann sphere $S^2 = \boldsymbol{C} \cup \infty$. Lemma 2.2 tells us that $F(\overline{U})$ is a bounded subset of $\boldsymbol{C}$. Since $\overline{U}$ is a compact subset of S^2 and F is continuous, $F(\overline{U})$ is compact.

If $F(U)$ is not a subset of R, we can find a point $p \in \overline{U}$ such that $F(p)$ lies in the boundary of $F(\overline{U})$ but not in ∂R. Note that p must lie in U because $F(\overline{U} - U) = \partial R$. By the Inverse Function Theorem, F maps a neighborhood of p onto a neighborhood of $F(p)$. But then $F(p)$ could not lie in the boundary of $F(\overline{U})$. This contradiction shows that $F(U) \subset R$. ♠

Exercise 3. Use essentially the same argument that we gave in §5.3, in connection with the Fundamental Theorem of Algebra, to show that $F(U) = R$.

Lemma 15.4 *F is one-to-one on U.*

Proof: Let $B \subset U$ denote the set of points z such that $F(z) = F(z')$ for some other $z' \in U$. Consider the extreme case when $B = U$. Choose some $z \in U^* \cap \boldsymbol{R}$, and let $\{z_n\}$ be a sequence of points in U converging to z. Let $\{z'_n\}$ be a sequence of points in U such that $F(z_n) = F(z'_n)$.

By Theorem 15.2, the map F is one-to-one in a neighborhood of z, so there is some minimum distance between z and z'_n. Passing to a subsequence, we can assume that z'_n converges to some point $z' \in \boldsymbol{R} \cup \infty$. From the minimum distance property, $z \neq z'$. By continuity $F(z) = F(z')$. But F is one-to-one on $\boldsymbol{R} \cup \infty$.

Now we know that $B \neq U$. We will show that B is both open and closed. Since U is connected, the only possibility is that $B = \emptyset$. Essentially, that same argument we just gave to show that $B \neq U$ shows that B is closed in U. We just have to show that B is open.

Suppose that $z \in B$ and $F(z') = F(z)$. By Theorem 15.2, F maps neighborhoods of z and z' onto neighborhoods of $F(z) = F(z')$. Hence B contains a neighborhood of z. Hence B is open. ♠

Lemma 15.4 and Exercise 3 combine to show that $F : U \to R$ is a complex analytic bijection. Theorem 15.2 now shows that F^{-1} is complex analytic. Hence F is a biholomorphism. This completes the proof of Theorem 15.1.

15.4 The Range of Possibilities

Theorem 15.1 explains how we can get *some* rectilinear polygons as images of the upper half-plane under a Schwarz–Christoffel transformation. It turns out that, up to scaling, we can get all of them this way. The idea is to show that we can vary the inputs of the construction so as to produce every possibility. Here is the main result.

Theorem 15.5 *Up to scaling, every rectilinear polygon is the image of the upper half-plane under a Schwarz–Christoffel transformation.*

The proof of Theorem 15.5 is a bit hard going, but I included it because I like the result and also because I will use Theorem 15.5 in the next chapter to prove the Riemann Mapping Theorem. Once we know the Riemann Mapping Theorem, we can say right away that every open solid polygon is the image of the upper half-plane under a biholomorphsm. However, without knowing Theorem 15.5, it seems difficult to prove, just from the Riemann Mapping Theorem, that every biholomorphism from the upper half-plane to a rectilinear polygon is given (up to composition with Möbius transformations) by a Schwarz–Christoffel transformation.

One unfortunate thing about our proof of Theorem 15.5 is that it is not completely self-contained. It relies on a basic result in topology known as Invariance of Domain. The Invariance of Domain result has always struck me as obviously true, but the proof is fairly difficult.

15.5 Invariance of Domain

The following result is Theorem 2B.3 in [HAT].

Theorem 15.6 (Invariance of Domain) *Suppose that $U \subset \boldsymbol{R}^n$ is an open set, and $\Phi : U \to \boldsymbol{R}^n$ is a continous and one-to-one map. Then $\Phi(U)$ is open in $\boldsymbol{R}^n$.*

We are mainly interested in a certain corollary of the Invariance of Domain result. Suppose that X and Y are spaces, both homeomorphic to open subsets of $\boldsymbol{R}^n$. A map $\Phi : X \to Y$ is *proper* if it has the following property. If $K \subset Y$ is compact, then $\Phi^{-1}(K) \subset X$ is compact.

Lemma 15.7 *Let X and Y be spaces, both homeomorphic to open subsets of $\boldsymbol{R}^n$. Suppose also that X is nonempty and Y is connected. If $\Phi : X \to Y$ is a one-to-one, continuous, and proper map, then $\Phi(X) = Y$.*

Proof: We suppose that this result is false and derive a contradiction. By Invariance of Domain, $\Phi(X)$ is an open subset of Y. Moreover, $\Phi(X)$ is nonempty. Since Y is connected, Y itself is the only subset of Y that is simultaneously open, closed, and nonempty. We conclude that $\Phi(X)$ is not closed. Hence, we can find a point

$$q \in \overline{\Phi(X)} - \Phi(X).$$

Given the location of q, we can find a sequence $\{p_k\} \in \Phi(X)$ such that $p_k \to q$.

We can choose $\{p_k\}$ so that it lies in a compact subset of Y. Since Φ is proper, there is a sequence $\{p'_k\}$, contained in a compact subset of X such that $\Phi(p'_k) = p_k$. Since $\{p'_k\}$ lies in a compact subset of X, this sequence has a convergent subsequence. Passing to this subsequence, we let $q' = \lim p_k \in X$. Since Φ is continuous, $\Phi(q') = q$. This contradicts the fact that $q \not\in \Phi(X)$. ♠

Given Lemma 15.7, the rest of our proof of Theorem 15.5 is self-contained.

15.6 The Existence Proof

Say that a *marked loop* is a counterclockwise oriented rectilinear loop with a preferred edge. We fix some length n sequence Σ of “lefts” and “rights”, with a total of 4 more “lefts” than ”rights”, and we let Y'_Σ denote the space of all marked polygons that have this sequence of turns as we trace around it counterclockwise, starting with the preferred edge. Let $Y_\Sigma \subset Y'_\Sigma$ denote the subset of embedded ones. Using the side lengths of the polygons, we consider Y'_Σ and Y_Σ as subsets of $\boldsymbol{R}^n$. This makes these sets into metric spaces.

Exercise 4. Let Σ be a sequence of length n, as above. Prove that Y_Σ and Y'_Σ are both homeomorphic to open subset of $\boldsymbol{R}^{n-2}$.

Exercise 5 (Challenge). Prove that Y_Σ is connected. (*Hint*: The result is certainly true for the sequence $\Sigma = LLLL$. Here Y_Σ is just the space of rectangles. In general, do induction on the length of Σ. Show that a rectilinear polygon always has a “spot” where you can continuously shrink one of the edges to a point without destroying the embedding property; see Figure 15.3.)

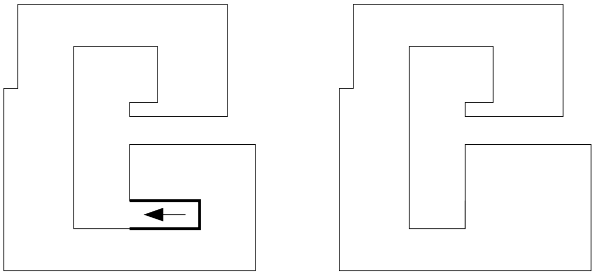

Figure 15.3. Shrinking an edge

Let Σ_1 and Σ_2 be two sequences. We write $\Sigma_1 \to \Sigma_2$ if Σ_2 is obtained from Σ_1 by the insertion of LR or RL somewhere in Σ_1. For any sequence Σ_2 except $LLLL$, there is some sequence Σ_1 such that $\Sigma_1 \to \Sigma_2$. Say that the sequence Σ_1 is *good* if some polygon in Y_Σ is the image $F(\boldsymbol{R} \cup \infty)$ for a Schwarz–Christoffel transform F.

Lemma 15.8 *All sequences are good.*

Proof: The sequence $LLLL$ is certainly good. We will prove the following statement. If $\Sigma_1 \to \Sigma_2$ and Σ_1 is good, then so is Σ_2. This lemma then follows from induction.

Let P be a polygon in Y_1 that is realized as the image $F(\boldsymbol{R} \cup \infty)$ for some Schwarz–Christoffel transformation F. Let $x_1, ..., x_n$ be the special points corresponding to F. The exponents $e_1, ..., e_n$ are chosen so as to match the sequence of lefts and rights in Σ_1.

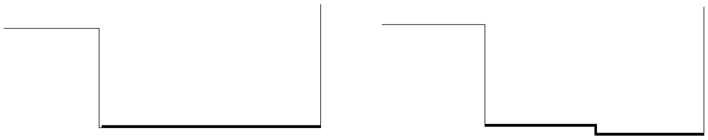

Figure 15.4. A zig-zag.

Let's say that Σ_2 is obtained from Σ_1 by inserting LR after the kth slot. Then, between x_k and x_{k+1}, we insert two new points x'_1 and x'_2. We place these points extremely close together, and right near the middle of the interval bounded by x_k and x_{k+1}. We chose additional exponents $e'_1 = -1/2$ and $e'_2 = 1/2$. Let F' be the new Schwarz–Christoffel transform based on the points $x_1, \ldots, x_k, x'_1, x'_2, x_{k+1}, \ldots, x_n$ and the corresponding exponents. When x'_1 and x'_2 are very close, the images $F(\boldsymbol{R} \cup \infty)$ and $F'(\boldsymbol{R} \cup \infty)$ are almost identical, except that the single edge $F(I_k)$ is replaced by a zig-zag, as shown in Figure 15.4. If this perturbation is small, the polygon in question is embedded. ♠

We fix a sequence of exponents $e_1, \ldots, e_n$, as above. These exponents determine the corresponding sequence Σ of lefts and rights. The input to our construction is a positive constant c and points $x_1 = -1$ and $0 = x_2 < x_3 < \cdots < x_n = 1$. Let X' be the set of possible inputs. X' is homeomorphic to $\boldsymbol{R}^{n-2}$.

Given some particular input $p \in X'$, the output is a polygonal loop $\Phi(p) = F(\boldsymbol{R} \cup \infty)$. Here F is the Schwarz–Christoffel transformation from equation (64), rescaled by c. We scale by c at the end for technical purposes. By construction $\Phi(p)$ is a point in the space of $Y' = Y'_\Sigma$. Thus, we have a map $\Phi : X' \to Y'$. The map Φ is pretty obviously continuous, given the formula for the Schwarz–Christoffel transformation.

Lemma 15.9 $\Phi : X' \to Y'$ *is one-to-one.*

Proof: Suppose that F_1 and F_2 are two Schwarz–Christoffel transformations such that $F_1(\boldsymbol{R} \cup \infty)$ and $F_2(\boldsymbol{R} \cup \infty)$ trace out the same polygon. We mean also that $F_1(x_i) = F_2(x_i)$ for all i. Define $G = F_1^{-1}F_2 : U \to U$. This is a biholomorphism that fixes -1, 0, and 1. By Exercise 6 below, G is the identity. ♠

Exercise 6. Let $G : U \to U$ be a biholomorphism from the upper half-plane to itself. Suppose that G fixes the points -1, 0, and 1. Prove that G is the identity map.

Recall that $Y \subset Y'$ is the subset of embedded polygons. We let $X = \Phi^{-1}(Y)$. Since all sequences are good, X is nonempty. Since Φ is continuous, X is open. To finish the proof of Theorem 15.5, we just have to show that $\Phi : X \to Y$ is proper. Then we can apply Lemma 15.7 and conclude that $\Phi(X) = Y$, as desired.

Let K be a compact subset of Y. We want to show that $\Phi^{-1}(K)$ is a compact subset of X. This is the same as showing that $\Phi^{-1}(K)$ is a compact subset of X'. We can put this another way. Suppose $\{p_k\}$ is a sequence of inputs that exits every compact subset of X'. We want to prove that $\Phi(p_k)$ exits every compact subset of Y.

Say that a *special interval* relative to the index k is an interval bounded by consecutive points $x_{k,j}$ and $x_{k,j+1}$. If $\{p_k\}$ exits every compact subset of X', then at least one of 3 things happens on a subsequence. Either $c_k \to \infty$ or $c_k \to 0$ or $\{c_k\}$ is bounded. In the last case, the length of the shortest special interval tends to 0 with k.

Suppose that $c_k \to \infty$. Now matter how we choose the input, all the points in the interval $[-2/3, -1/3] \subset I_0$ are at least $1/3$ units away from all the special points. Looking at the formula for F in equation (64), we see that $F(I_0)$ has length at least $(1/3)^{n+1}$. But then one of the sides of the kth output has length at least $c_k(1/3)^{n+1}$, a number that tends to ∞ with k. Hence $\Phi(p_k)$ exits every compact subset of Y.

Suppose that $c_k \to 0$. No matter what the input, we can find 3 points $y_1, y_2, y_3 \in \boldsymbol{R}$, all in distinct special intervals, that are all at least $1/2n$ from any of the endpoints of the special intervals. Looking at the formula for F in equation (64), we see that there is some constant C, independent of inputs, such that $|F(y_i)| < C$ for $i = 1, 2, 3$. But then the polygon corresponding to $\Phi(p_k)$ has three sides which come within Cc_k of the origin. But Cc_k tends to 0. This shows that points $\Phi(p_k)$ exit every compact set of Y.

The following result finishes our proof of Theorem 15.5.

Lemma 15.10 *If $\{c_k\}$ is bounded and the length of some special interval tends to* 0*, then $\Phi(p_k)$ exits every compact subset of Y.*

Proof: We will suppose that $\Phi(p_k)$ lies in a compact subset of Y and derive a contradiction. After a bounded amount of scaling, we can assume that $c_k = 1$ for all k. Let F_k be the Schwarz–Christoffel transformation associated to p_k. Let $P_k = F_k(\boldsymbol{R} \cup \infty)$. By compactness, there is some $D > 0$ such that the sides of P_k have length at most $1/D$ and the distance between any two distinct vertices of P_k is at least D. Here D is independent of k.

Passing to a subsequence, we can assume that $x_{k1}, \ldots, x_{kn}$ converges to points $x_{\infty,1}, \ldots, x_{\infty,m}$. Here $m < n$ because some points have coalesced. We have associated exponents $e_{\infty,1}, \ldots, e_{\infty,m}$, where $e_{\infty,k}$ is the sum of the exponents of the points that coalesce to $x_{\infty,k}$.

Exercise 7. Prove that $e_{\infty,k} \geq -1/2$. (*Hint*: Use the fact that the sides of P_k have length at most $1/D$, independent of k.)

Because the integrands for F_k converge at each point of U^*, the sequence $\{F_k\}$ of maps converges to a map $F_\infty : U^* \to \boldsymbol{C}$, defined exactly as in equation (64). Because $e_{\infty,k} \geq -1/2$ and $\sum e_{\infty,k} = -2$, the map F_∞ extends to be continuous on $\boldsymbol{R} \cup \infty$.

Choose an index m such that 2 or more points $x_{k,j}$ converge to $x_{\infty,m}$. Consider the special intervals A_∞ and B_∞ on either side of $x_{\infty,m}$. There are special intervals A_k and B_k such that $A_k \to A_\infty$ and $B_k \to B_\infty$. By our choice of index m, the intervals A_k and B_k are *not* consecutive.

Exercise 8. Prove the following result. There is some K such that $k > K$ implies that $F_k(A_k)$ contains all but $D/3$ of $F_\infty(A_\infty)$. (*Hint*: use the finiteness of of all the integrals involved and the convergence of the integrands. The same result holds for B in place of A.)

By Exercise 8, some endpoint of $F_k(A_k)$ is within $2D/3$ of some endpoint of $F_k(B_k)$. This contradicts the existence of D. ♠

16 Riemann Surfaces and Uniformization

The purpose of this chapter is to define the notion of a Riemann surface. A Riemann surface is essentially a surface that is built out of pieces of $\boldsymbol{C}$ glued together with complex analytic maps. Once we know about Riemann surfaces, we can speak about complex analytic maps between them. We will prove some basic results about such maps, relying on the material from the previous 3 chapters.

Following the discussion of Riemann surfaces, we will prove the Riemann Mapping Theorem. For another proof of this result, one that does not rely on Theorem 15.5, see [AHL].

The Riemann Mapping Theorem is a special case of the Poincaré Uniformization Theorem, a result we will state without proof. A proof can be found in [BE2]. After stating the Uniformization Theorem, we will deduce some consequences from it.

16.1 Riemann Surfaces

Let S be a surface. Recall that a smooth structure on S is a maximal collection of coordinate charts which have the property that the overlap functions are all smooth. A Riemann surface is defined in a similar way, with the word *complex analytic* replacing the word *smooth*. That is, a Riemann surface structure on a surface is a maximal collection of coordinate charts such that the overlap functions are all smooth. Here are some examples:

Open Subsets of $\boldsymbol{C}$. Any open subset of $\boldsymbol{C}$ is a Riemann surface. We can take the coordinate chart maps to be the identity.

The Riemann Sphere. We can think of S^2 as $\boldsymbol{C} \cup \infty$. Then $U_1 = \boldsymbol{C}$ is a neighborhood of $\{0\}$ and $U_2 = \boldsymbol{C} \cup \infty - \{0\}$ is a neighborhood of ∞. The identity map is a homeomorphism from U_1 to $\boldsymbol{C}$ and the map $f(z) = 1/z$ is a homeomorphism from U_2 to $\boldsymbol{C}$. The overlap $U_1 \cap U_2$ is $\boldsymbol{C} - \{0\}$ and the overlap function is just $f(z) = 1/z$, a complex analytic function. We already have a collection of (two) coordinate charts which cover S^2, and we can complete this collection to a maximal collection. This makes S^2 into a Riemann surface. This surface is known as the *Riemann sphere*.

Flat Tori. Let P be a parallelogram. If we glue the opposite sides of P together by translations, then we produce a closed surface. We can find a covering of S by coordinate charts whose overlap functions are translations, i.e., maps of the form $z \to z + C$ for various choices of the constant C. Such maps are complex analytic, and so we can make these flat tori into Riemann surfaces in a natural way.

Hyperbolic Surfaces. Recall that a hyperbolic structure on a surface is a maximal collection of coordinate charts into $\boldsymbol{H}^2$ such that the overlap functions are all restrictions of hyperbolic isometries. If we only use orientation preserving hyperbolic isometries, then these maps are all linear fractional transformations. Linear fractional transformations are complex analytic, and so a hyperbolic structure on a surface is always a Riemann surface structure.

Exercise 1. In §12.6 we discussed the notion of a Riemannian covering space. We can similarly define a *Riemann surface covering*. This would be a covering map between Riemann surfaces that is complex analytic. Given a covering map $E : \widetilde{S} \to S$, prove that $\widetilde{S}$ can be made into a Riemann surface such that E is a Riemann surface covering.

Exercise 2. Let $E(z)$ be the exponential function, as defined in Exercise 7 of §13.8. Prove that E is a covering map from $\boldsymbol{C}$ to $\boldsymbol{C} - \{0\}$. (*Hint*: Use the identities in Exercises 7-9 of §13.8 to get a handle on the geometry of E.)

Exercise 3 (Challenge). Let X be the space obtained by gluing together two copies of the solid unit square, along all sides (see Figure 16.1). Give X the structure of a Riemann surface (by finding local charts) so that there is a biholomorphic map between X and the Riemann sphere. (*Hint*: For the coordinate charts, the only tricky part is thinking about what to do at the vertices and edges. Think about the Christoffel transform between the square and the upper half plane.)

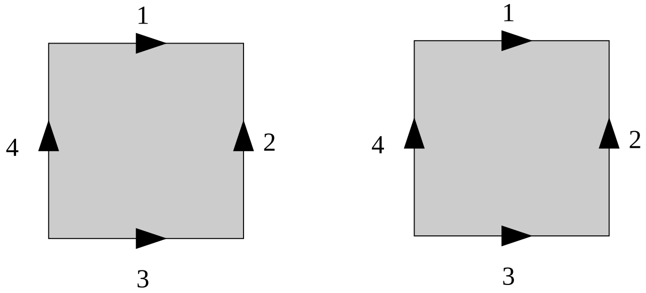

Figure 16.1. Gluing 2 squares

16.2 Maps Between Riemann Surfaces

Suppose S_1 and S_2 are two Riemann surfaces. A map $f : S_1 \to S_2$ is *complex analytic* in a neighborhood of $p_1 \in S_1$ if there are neighborhoods U_1 of p_1 and U_2 of $p_2 = f(p_1)$, together with coordinate charts $f_j : U_j \to \boldsymbol{C}$ such that the map $f_2 \circ f \circ f_1^{-1}$ is complex analytic. f is complex analytic on S_1 if f is complex analytic in a sufficiently small neighborhood of every point. We can use some of the machinery from Chapter 13 to prove nontrivial results about maps between Riemann surfaces. This chapter contains a sampler of these results.

Theorem 16.1 *There is no nontrivial complex analytic map from a compact Riemann surface into* $\boldsymbol{C}$.

Proof: Suppose $f : S \to \boldsymbol{C}$ is complex analytic. Since S is compact f achieves its maximum at some point $p \in S$. Let U be a coordinate chart about p and let $g : U \to C$ be a coordinate chart. Then $h = f \circ g^{-1}$ is a complex analytic map from the open set $g(U)$ into $\boldsymbol{C}$. Moreover, h takes its maximum value at an interior point of $g(U)$. But a nonconstant complex analytic map cannot have an interior maximum, according to the Maximum Principle from §13.5. ♠

On the other hand, there are plenty of complex analytic maps from the Riemann sphere to itself. For instance, any rational function $R(z) = \frac{P(z)}{Q(z)}$ is a complex analytic map from the Riemann sphere to itself. Here P and Q are polynomials. The set $R^{-1}(\infty)$ is contained in the set of zeros of Q.

Theorem 16.2 *There is no nonconstant complex analytic map from* $\boldsymbol{C}$ *into a hyperbolic surface.*

Proof: Let $f : \boldsymbol{C} \to S$ be a complex analytic map from $\boldsymbol{C}$ to S. Let $E : \boldsymbol{H}^2 \to S$ be the universal covering map. Using the lifting property for maps we can find a lifting $\widetilde{f} : \boldsymbol{C} \to \boldsymbol{H}^2$ such that $E \circ \widetilde{f} = f$. (We produce $\widetilde{f}$ by partitioning $\boldsymbol{C}$ into an infinite grid of squares, and applying the lifting theorem one square at a time.) By construction $\widetilde{f}$ is complex analytic. The point is that on small neighborhoods E^{-1} is defined and complex analytic; and $\widetilde{f} = f \circ E^{-1}$ on these small neighborhoods. However, we can take $\boldsymbol{H}^2$ as the open unit disk. So, $\widetilde{f}$ is a bounded complex analytic function on $\boldsymbol{C}$. However, all such maps are constant. Since $\widetilde{f}$ is constant, so is f. ♠

Any complex analytic homeomorphism from $\boldsymbol{C}$ to $\boldsymbol{C}$ is a linear map, by Corollary 14.4. Our proof of the next result uses this fact.

Theorem 16.3 *Suppose that* S *is a Riemann surface which has a non-Abelian fundamental group. Then there is no complex analytic covering map of the form* $E : \boldsymbol{C} \to S$.

Proof: Suppose that $E : \boldsymbol{C} \to S$ exists. Let G be the fundamental group of S. Then G acts on $\boldsymbol{C}$ as the deck group. Each element $g \in G$ acts as a complex analytic homeomorphism of $\boldsymbol{C}$. Hence g is a complex linear map. Being an element of the deck group, g acts without any fixed points. Therefore, g must be a translation. In short G is a group of translations. But any two translations commute and hence G is Abelian. This contradiction shows that E does not exist. ♠

16.3 The Riemann Mapping Theorem

Let Δ denote the open unit disk. Say that a *Jordan domain* is any set of the form $h(\Delta)$, where $h : \boldsymbol{C} \to \boldsymbol{C}$ is a homeomorphism.

Theorem 16.4 (Riemann Mapping Theorem) *Let D be any Jordan domain. There exists a biholomorphism from Δ to D.*

Riemann gave an intuitive description of the Riemann map. Imagine that the domain D is a uniformly conducting material, and that an electric potential of 1 is maintained at some interior point $x \in D$ and a potential of 0 is maintained on the boundary of D. The equipotential lines form loops around x, and the electricity flowing from x out to ∂D flows along lines perpendicular to the equipotential loops. The equipotential loops and the flow lines form a kind of wavy coordinate system on D. The Riemann map, if it is normalized to send 0 to x, sends the ordinary polar coordinate system on Δ to the wavy one.

We will give a proof of the Riemann Mapping Theorem that is based on Theorem 15.5.

Exercise 4. Prove the following statement. For any $\epsilon > 0$, there is an embedded rectilinear polygon P such that every point of ∂D is within ϵ of P, and vice versa.

We scale the picture so that $\Delta \subset D$. For each positive integer n, choose a rectilinear polygon that is within $1/n$ of ∂D in the sense of Exercise 4. Let D_n be the region bounded by this polygon. The polygon itself is ∂D_n.

Since Δ and the upper half plane are biholomorphically equivalent, Theorem 15.5 says that there is a biholomorphism $F_n : \Delta \to D_n$. Composing F_n with a Möbius transformation of Δ, we arrange that $F_n(0) = 0$ for all n. The rest of the proof amounts to showing that the sequence $\{F_n\}$ converges to the desired map.

Exercise 5. Let $r < 1$ and let $\Delta(r)$ denote the disk of radius r centered at the origin. Prove that there is some constant R, depending on r but not on n, such that $|F_n'(z)| < R$ for all $z \in \Delta(r)$. (*Hint*: Apply equation (61), using a circular loop $\gamma \subset \Delta$ that bounds $\Delta(r')$ for some $r' > 1$.)

Since D is bounded, we can pass to a subsequence so that $\{F_n(z)\}$ converges on a countable dense subset of points $z \in \Delta$. But then, Exercise 5 guarantees that $\{F_n(z)\}$ converges uniformly on each disk $\Delta(r)$. That is, for any $\epsilon > 0$, there is some N such that $n > N$ implies that $|F_m(z) - F_n(z)| < \epsilon$ for all $m, n > N$.

Let $F = \lim F_n$. We have a converging sequence of maps, all of which satisfy the Cauchy Integral Formula for all loops in Δ. Hence, F satisfies the Cauchy integral as well. Hence F is holomorphic. The main thing we want to rule out is that F is the constant map. The next lemma does this for us.

Lemma 16.5 $|F'(0)| \geq 1$.

Proof: Let $G_n = F_n^{-1}$. Recall that $F_n(0) = 0$ and $\Delta \subset D$. Hence $G_n(\Delta) \subset \Delta$ and $G_n(0) = 0$. By Lemma 14.2, we have the inequality $|G_n'(0)| \leq 1$. ♠

Exercise 6. Imitate the proof of Lemma 16.5 to show that $F'(z) > 0$ for all $z \in \Delta$.

Lemma 16.6 *F is one-to-one.*

Proof: Suppose that $F(z_1) = F(z_2)$. Then, by Theorem 15.2, there are disjoint open sets U_1 and U_2 such that $F(U_1) = F(U_2)$. But then $F_n(U_1)$ and $F_n(U_2)$ overlap for large n. This contradicts that F_n is one-to-one. ♠

Since $F'(z)$ never vanishes, Theorem 15.2 shows that F^{-1} is holomorphic. Now we know that F is a biholomorphism from Δ onto $F(\Delta)$. Certainly $F(\Delta) \subset D$. To finish the proof, we just have to show that $F(\Delta) = D$.

Choose $w \in D$. Let $z_n = F_n^{-1}(w)$. Call w *good* if the sequence $\{z_n\}$ remains within a compact subset of Δ. Otherwise call w *bad*. If w is good, there is at least one accumulation point $z \in \Delta$ of $\{z_n\}$. Since $F_n(z_n) = w$ and we have a uniform bound on $|F_n'|$ in a neighborhood of z, we have $F(z) = w$. We just have to show that every point in D is good.

Lemma 16.7 *w is contained in the interior of a disk $W \subset D$ with the following property. For all $w' \in W$, the hyperbolic distance between $F_n(w)$ and $F_n(w')$ is less than 1, independent of n.*

Proof: Apply Exercise 2 from Chapter 14 to the map $G = F_n^{-1}$ and some open set $U \subset D$ such that $w \in U$ and $U \subset F_n(\Delta)$ for all n. ♠

Note that w is good if and only if there is some K such that $\{z_n\}$ stays within K hyperbolic units of 0. It therefore follows from Lemma 16.7 and the triangle inequality that the set of good points is open. If $\{z_n\}$ stays with K hyperbolic units of 0, then $\{z'_n\}$ stays within $K+1$ units of 0. Here we have set $z'_n = F_n^{-1}(w')$. Similarly, it follows from Lemma 16.7 and the triangle inequality that the set of bad points is open. Finally, 0 is good. So, the set of good points is open, closed, and nonempty. Hence every point in D is good. Hence $f(\Delta) = D$.

16.4 The Uniformization Theorem

Here is the Poincaré Uniformization Theorem.

Theorem 16.8 (Poincaré Uniformization) *Suppose that A is a simply connected Riemann surface. Then one of three things is true:*

- *A is compact, and there is a biholomorphism between A and the Riemann sphere.*
- *A is noncompact and there is a biholomorphism between A and $\boldsymbol{C}$.*
- *A is noncompact and there is a biholomorphism between A and the open unit disk.*

Note that the Poincaré Uniformization Theorem contains the Riemann Mapping Theorem as a special case, when A is a Jordan domain. The main difference between the two results is that, in the Uniformization Theorem, A is not assumed to be a subset of $\boldsymbol{C}$.

16.5 The Small Picard Theorem

For the rest of the chapter, we deduce some nice consequences of the Uniformization Theorem.

Lemma 16.9 *There is a complex analytic covering map from the open unit disk to $\boldsymbol{C} - \{0,1\}$, the twice punctured plane.*

Proof: The universal cover X of $\boldsymbol{C} - \{0,1\}$ is a simply connected Riemann surface. Let $E : X \to \boldsymbol{C} - \{0,1\}$ be the covering map. If X is compact, then $E(X)$ is also compact, since the image of a compact set under a continuous map is compact. But $E(X) = \boldsymbol{C} - \{0,1\}$, which is noncompact. So, X is noncompact. If there is a biholomorphism between X and $\boldsymbol{C}$, then we have a complex analytic cover $\boldsymbol{C} \to \boldsymbol{C} - \{0,1\}$. However, the fundamental group of $\boldsymbol{C} - \{0,1\}$ is non-Abelian. This is a contradiction. We have only one alternative left in the Uniformization Theorem, and so there is a biholomorphism h between X and the open unit disk. But then $E \circ h^{-1}$ is the desired complex analytic covering map between the open unit disk and $\boldsymbol{C} - \{0,1\}$. ♠

Remark. In the concrete setting just discussed, it is possible to prove Lemma 16.9 directly, without appealing to the Uniformization Theorem. This is done in [AHL].

Lemma 16.9 is the main ingredient in the proof of the the following result, which is known as the *Small Picard Theorem*:

Theorem 16.10 *Let $f : \boldsymbol{C} \to \boldsymbol{C}$ be a nonconstant analytic map. Then either f is onto or f omits exactly one value.*

Proof: We will suppose that f omits at least two values and show that f is constant. We can scale f so that two of the omitted values are 0 and 1. Then $f : \boldsymbol{C} \to \boldsymbol{C} - \{0,1\}$. We have our holomorphic covering from the open unit disk Δ to $\boldsymbol{C} - \{0,1\}$. But then we can find a lift $\widetilde{f} : \boldsymbol{C} \to \Delta$. This map is a bounded complex analytic function, and hence constant. Hence f is constant as well. ♠

16.6 Implications for Compact Surfaces

The Uniformization Theorem is stated above in terms of simply connected Riemann surfaces, but it has nice implications for general surfaces. Here is a the main consequence for compact surfaces.

Theorem 16.11 *Let S be a compact and oriented Riemann surface.*

- *If S is homeomorphic to a sphere, then there is a biholomorphism between S and the Riemann sphere.*
- *If S is homeomorphic to the torus, then there is a biholomorphism between S and a flat torus.*
- *If S is a Riemann surface of negative Euler characteristic, then there is a biholomorphism between S and some hyperbolic surface.*

Proof: The sphere case is immediate from the Uniformization Theorem.

Suppose that S is not homeomorphic to a torus. Let $\widetilde{S}$ be the universal cover of S. Note that $\widetilde{S}$ is a simply connected Riemann surface. According to the Uniformization Theorem, there is either a biholomorphism between $\widetilde{S}$ and $\boldsymbol{C}$, or a biholomorphism between $\widetilde{S}$ and Δ, the open unit disk. In the former case, we would have a complex analytic covering map $\boldsymbol{C} \to S$. But S has non-Abelian fundamental group, so Theorem 16.3 rules out this possibility. Therefore, we have a complex analytic covering $\Delta \to S$ where Δ is the unit disk. Let G be the fundamental group of S. Then G acts on Δ as the deck group. Each element $g \in G$ is a biholomorphism of Δ. In Chapter 13 we proved that such maps are hyperbolic isometries. Hence G acts on Δ as a group of hyperbolic isometries. S is precisely the quotient of the hyperbolic plane by the orbit equivalence relation: Two points are equivalent iff there is some element of G which maps one to the other. Small neighborhoods of points in Δ contain unique members of equivalence classes, and so these little disks map injectively into S. The inverse maps give local coordinate charts into Δ, such that the overlap functions are restrictions of hyperbolic isometries. In short, S inherits its hyperbolic structure from Δ.

Suppose that S is homeomorphic to a torus. If there is a holomorphic covering $\Delta \to S$ then the same argument as just given shows that S is a hyperbolic surface and the fundamental group $\boldsymbol{Z}^2$ acts on Δ by hyperbolic isometries. This is only possible if all the elements of $\boldsymbol{Z}^2$ fix a common point on the unit circle. Such maps have the following property: For any $\epsilon > 0$ there is some point $x \in \Delta$ which is moved less than ϵ (as measured in the hyperbolic metric). But then S would have closed and homotopically nontrivial loop of length less than ϵ. This contradicts the fact that all sufficiently short loops on S are homotopically trivial. The contradiction shows that there is no holomorphic cover from Δ to S. Only one alternative for the Uniformization Theorem holds and so there is a holomorphic cover $\boldsymbol{C} \to S$. But now the deck transformations are all Euclidean translations and S inherits a Euclidean structure from $\boldsymbol{C}$ just as in the previous case. ♠

The above theorem is true in much more generality. For instance, suppose that $C \subset \boldsymbol{C}$ is a finite set of $N > 2$ points. Then there is a biholomorphism between $\boldsymbol{C} - C$ and a hyperbolic surface. The same result holds if C is a countably infinite set of points, or the middle-third Cantor set. It is hard to picture the universal cover of the complement of the middle-third Cantor set, but the Uniformization Theorem says that it is just the hyperbolic plane in disguise!

17 Flat Cone Surfaces

In this chapter we revisit the idea of gluing polygons together to form a surface. In a sense, we return to the question taking the most naive point of view possible. We keep the Euclidean geometry of the component pieces and see what we get when they are glued together. This point of view leads to the definition of a *flat cone surface*.

After we define flat cone surfaces, we will prove a fundamental result about them, the combinatorial Gauss–Bonnet Theorem. The combinatorial Gauss–Bonnet Theorem is am analogue the Gauss–Bonnet Theorem from differential geometry; compare Theorem 12.4.

Following the proof of the combinatorial Gauss–Bonnet Theorem, we give an application of flat cone surfaces to the study of polygonal billiards. This is a theme that will take up both this chapter and the next. All the material about billiards can be found, in much greater detail in [MAT].

17.1 Sectors and Euclidean Cones

A *sector* in $\boldsymbol{R}^2$ is the closure of one of the 2 components of $\boldsymbol{R}^2 - \rho_1 - \rho_2$, where ρ_1 and ρ_2 are two distinct rays emanating from the origin. For example, the nonnegative quadrant is a sector. The *angle* of the sector is defined as the angle between ρ_1 and ρ_2 as measured from inside the sector. For instance, the angle of the nonnegative quadrant is $\pi/2$.

Two sectors in $\boldsymbol{R}^2$ can be glued together isometrically along one of their edges. A *Euclidean cone* is a space obtained by gluing together, in a cyclic pattern, a finite number of sectors. The *angle* of the Euclidean cone is the sum of the angles of the sectors. The *cone point* is the equivalence class of the origin(s) under the gluing. The cone point is the only point which potentially does not have a neighborhood locally isometric to $\boldsymbol{R}^2$.

Note that two isometric Euclidean cones might have different descriptions. For instance, $\boldsymbol{R}^2$ can be broken into 4 quadrants or 8 sectors of angle $\pi/4$.

Exercise 1. Prove that two Euclidean cones are isometric if and only if they have the same angle.

Exercise 2. Define the unit circle in a Euclidean cone to be the set of points which are 1 unit away from the cone point. On the cone of angle 4π

find the shortest path between every pair of points on the unit circle. This problem breaks down into finitely many cases, depending on where the points are located.

Exercise 3. Let C be a Euclidean cone, with cone point x. Say that a vector field on $C - x$ is locally constant if an isometry carrying any open set of $C - x$ into $\boldsymbol{R}^2$ carries the vector field to a constant vector field. Prove that $C - x$ has a parallel vector field in a neighborhood of x if and only if the cone angle of C is a multiple of 2π. (*Hint*: Unroll C into the plane and watch the vector field as you go once around the cone point.)

17.2 Euclidean Cone Surfaces

We defined in §3.2 what it means for a surface to be oriented—it does not contain any Möbius bands. For ease of exposition, we only consider oriented surfaces.

Say that a compact oriented surface Σ is a *Euclidean cone surface* if it has the following two properties:

- Every point $p \in \Sigma$ has a neighborhood which is isometric to a neighborhood of the cone point in a Euclidean cone of angle $\theta(p)$.
- We have $\theta(p) = 2\pi$ for all but finitely many points.

The points p, where $\theta(p) \neq 2\pi$, are called the *cone points*. The quantity

$$\delta(p) = 2\pi - \theta(p)$$

is called the *angle deficit*. So, there are only finitely many points with nonzero angle deficit, and these deficits could be positive or negative.

Here are two examples:

- Let P be a convex polyhedron in $\boldsymbol{R}^3$. Then ∂P is a Euclidean cone surface. The metric on ∂P is the intrinsic one: the distance between two points is the length of the shortest curve which remains on ∂P and joins the points.
- Let $P_1, \ldots, P_n$ be a finite union of polygons. Suppose that these polygons can be glued together, isometrically along their edges, so that the result is a surface. Then the surface in question is a Euclidean cone surface if it is given its intrinsic metric, i.e., the shortest path metric.

Amazingly, every example of type 2 is also an example of type 1 provided that the underlying surface is a sphere and all the angle deficits are positive. This result is known as the Alexandrov Theorem. (To make this strictly true we have to allow for the possibility that P is contained in a plane in $\boldsymbol{R}^3$.) One interesting open problem is to determine the combinatorics of the convex polyhedron you get, based on the intrinsic geometry of the cone surface.

17.3 The Gauss–Bonnet Theorem

Here is combinatorial version of the Gauss–Bonnet Theorem:

Theorem 17.1 *If S is a compact cone surface, then*

$$\sum_p \delta(p) = 4\pi\chi(S).$$

Here the sum is taken over all angle deficits.

Proof: A *Euclidean triangle* on a Euclidean cone surface S is a region isometric to (you guessed it) a Euclidean triangle. For instance, on the boundary of a tetrahedron, there are 4 obvious maximal Euclidean triangles. Two triangles on a cone surface *intersect* normally if they are either disjoint or share a vertex or share an edge. A *triangulation* of S is a decomposition of S into finitely many triangles, such that each pair of triangles intersects normally.

Exercise 4. Prove that every Euclidean cone surface has a triangulation.

Choose a triangulation of S. Let $T_1, \ldots, T_F$ be the list of triangles in the triangulation. Each T_i has associated to it three angles a_i, b_i, c_i, with $a_i + b_i + c_i = \pi$. The cone points are all at vertices of the triangles, and so

$$\sum_p \delta(p) = 2\pi V - (\sum_{i=1}^F a_i + \sum_{i=1}^F b_i + \sum_{i=1}^F c_i).$$

In other words, we add up all the angles and see how the total sum differs from the expected $2\pi V$. Given that $a_i + b_i + c_i = \pi$, we have

$$\sum_p \delta(p) = 2\pi V - \pi F = 2\pi(V - F/2) =^* 2\pi(V + F - E) = 2\pi\chi(S).$$

The starred equality has the following explanation. Each triangle contributes $3/2$ edges to the total number of edges. That is, $E = 3F/2 = F + F/2$. Hence $-F/2 = F - E$. ♠

For comparison, we mention that the differential geometric version of the Gauss–Bonnet Theorem says that the total curvature of a surface S is $2\pi\chi(S)$, where χ is the Euler characteristic of S; see §3.4. One can view the combinatorial Gauss–Bonnet Theorem as the limit of the differential geometric version, in which all the curvature is concentrated at finitely many points. At the same time, one can view the differential geometric version as a limit of the combinatorial version, in which the curvature gradually diffuses out, over larger and larger finite sets of points so that it becomes continuously distributed.

17.4 Translation Surfaces

A Euclidean cone surface is a *translation surface* if all the cone angles are integer multiples of 2π. For instance, the octagon surface discussed extensively in Chapter 1 is a translation surface when the octagon is interpreted as a regular Euclidean octagon.

Theorem 17.2 *Let S be a flat cone surface, and let C be a finite list of points in S. Then $S - C$ admits a parallel vector field if and only if S is a translation surface.*

Proof: Suppose first that such a vector field exists. Let $x_1, \ldots, x_n$ be the points of C. Let $U_1, \ldots, U_n$ be disk neighborhoods of $x_1, \ldots, x_n$, respectively. Since $U_k - x_k$ admits a parallel vector field, the cone angle at x_k is an integer multiple of 2π. This is Exercise 3 above.

Now we prove the converse. Choose some basepoint $x \in \Sigma - C$. Let $v(x)$ be some unit vector tangent to x. Our goal is to define a unit vector $v(y)$ for each point $y \in \Sigma - C$. Here is the construction. Let γ be any smooth curve which joins x to y and stays in $\Sigma - C$. Say that a vector field along γ is *parallel* if, in the local coordinates, the vectors are all translates of each other. Since every point of γ has a neighborhood which is isometric to a disk in $\boldsymbol{R}^2$, there is a unique parallel vector field along γ which agrees with $v(x)$ at x. We define $v(y)$ to be the vector of this parallel vector field at y. If this is really well defined, then in small neighborhoods, our vector field consists entirely of parallel vectors.

To finish our proof, we need to see that this definition is independent of the path γ. If γ_1 and γ_2 are paths connecting x to y, and are homotopic relative their endpoints, then we can produce a finite sequence of paths $\gamma_1 = \beta_1, \ldots, \beta_n = \gamma_2$ such that β_i and β_{i+1} agree except in a region which is contained in a single Euclidean disk. (You get the β curves just by doing the homotopy a little bit at a time.) Within the Euclidean disk, you can see that the vector field along β_i must be parallel to the vector field along β_{i+1}, because both vector fields just consist of a bunch of parallel vectors, and the two vector fields agree at some point in the disk. Since this is true for all i, the two methods for defining $v(y)$ agree.

The fundamental group $\pi_1(\Sigma - C)$ is generated by loops which travel from x into a small neighborhood of one of the cone points, wind around the cone point, and then come back. If γ_1 and γ_2 are arbitrary paths joining x to y, then γ_1 is homotopic relative to the endpoints to $\delta_1 * \cdots * \delta_k * \gamma_2$, where each δ_i is one of the special loops just mentioned. Each loop δ_i starts and ends at x. We just have to see that the parallel vector field along δ_i agrees with $v(x)$ at both ends. Everything boils down to what happens in a neighborhood of the cone point.

By Exercise 3, we can define a parallel vector field in the neighborhood of each point of C. Call these vector fields the "background vector fields". The parallel vector fields along our looks have constant length and make constant angles with the relevant background vector field. So, the parallel vector field along one of our loops comes exactly back to itself when the loop is done. ♠

Recall that a *gluing diagram* for a surface is a list of finitely many polygons, together with a recipe for gluing together the sides of the polygon in pairs.

Lemma 17.3 *Suppose that S is a flat cone surface obtained from a gluing diagram in which the two sides in each glued pair are parallel. Then S is a translation surface.*

Proof: Once we show that S is orientable, we will know that S is a cone surface. On each polygon, we consider the standard pair of vector fields V_1 and V_2. Here V_j consists of vectors parallel to the basis vector e_j. Given the nature of the gluing maps, the vector fields piece together across the edges to give parallel vector fields V_1 and V_2 defined on the complement of finitely many points.

We first show that S is orientable. If S is not orientable, then S contains a Möbius band M. By shrinking M if necessary, we can arrange that M lies entirely in the region where both V_1 and V_2 are defined. But then we can define a continuous pair of linearly independent vector fields on a Möbius band. This is easily seen to be impossible. Hence S is oriented.

It now follows from Lemma 17.2 that S is a translation surface. ♠

In light of Lemma 17.3, the surface obtained by gluing (with translations) the opposite sides of a regular $2n$-gon is a translation surface.

Translation Principle. Whenever we consider gluing diagrams for translation surfaces, in which more than one polygon is involved, we always think of the polygons in the plane as being pairwise disjoint. How the polygons sit in the plane is really not so important, in the following sense. Suppose that $P_1, \ldots, P_n$ are the polygons involved in a gluing diagram for some surface. Suppose that $Q_1, \ldots, Q_n$ are new polygons, such that Q_k is a translation of P_k for all k, and the pattern of gluing for the Q's is the same as the pattern of gluing for the Ps. Then the two resulting surfaces are canonically isometric. The canonical isometry is obtained by piecing together the translations that carry each P_k to Q_k. We mention this rather obvious principle because it guarantees that certain constructions, which seem based on arbitrary choices, are actually well defined independent of these choices.

17.5 Billiards and Translation Surfaces

Let P be a Euclidean polygon. A *billiard path* in P is the motion taken by an infinitesimal frictionless ball as it rolls around inside P, bouncing off the walls according to the laws of inelastic collisions: the angle of incidence equals the angle of reflection; see Figure 17.1 below. We make a convention that a path stops if it lands precisely at a vertex. (The infinitesimal ball falls into the infinitesimal pocket.)

The billiard path is *periodic* if it eventually repeats itself. Geometrically, a periodic billiard path corresponds to a polygonal path Q with the following properties:

- $Q \subset P$ (that is, the solid planar region).

- The vertices of Q are contained in the interiors of the edges of P.
- Q obeys the angle of incidence rule discussed above.

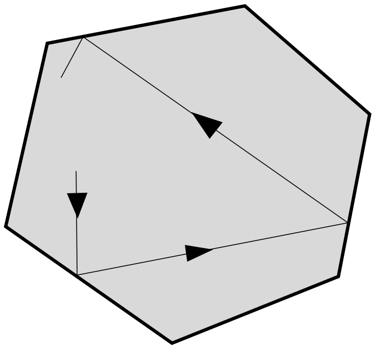

Figure 17.1. Polygonal billiards

Exercise 5. Find (with proof) all the examples of periodic billiard paths in a square which do not have self-intersections. So, the path Q has to be embedded.

The polygon P is called *rational* if all its angles are rational multiples of π. For instance, the equilateral triangle is a rational polygon.

In this section I will explain how to associate a translation surface to a rational polygon. This is a classical construction, attributed by some people to A. Katok and A.N. Zemylakov. The geometry of the translation surface encodes many of the features of billiards in the polygon.

For each edge e of P there is a reflection R_e in the line through the origin parallel to e. Like all reflections, R_e has order 2. That is, $R_e \circ R_e$ is the identity map. Let G be the group generated by the elements $R_1, \ldots, R_n$. Here R_j stands for R_{e_j} and $e_1, \ldots, e_n$ is the complete list of edges. If e_i and e_j are parallel, then $R_i = R_j$. If P is a rational polygon then there is some N such that e_j is parallel to some Nth root of unity. But then G is a group of order at most $2N$. In particular, G is a finite group.

For each $g \in G$, we define a polygon

$$P_g = g(P_g) + V_g. \qquad (65)$$

Here V_g is a vector included so that all the polygons $\{P_g |\ g \in G\}$ are disjoint. Thanks to the Translation Principle, the surface we will produce is independent of the choices of the translation vectors.

To form a gluing diagram, we declare that every two edges of the form

$$e_1 = g(e) + V_g, \qquad e_2 = gr(e) + V_{gr}, \qquad r = R_e. \tag{66}$$

are glued together by a translation. Here e is an arbitrary edge of P. Since $gr(e) = g(e)$, the edges e_1 and e_2 are parallel. Hence, it makes sense to glue them by a translation. Note also that $(gr)r = g$. So, our instructions tell us to glue e_1 to e_2 if and only if they tell us to glue e_2 to e_1. Let S be the space obtained from the gluing diagram. Since the edges are glued in pairs, S is a surface. By Lemma 17.3, S is a translation surface.

Here we work out the example where P is an isosceles triangle with small angles $2\pi/8$. In this case, the group G has order 16 and our surface will be made from 16 isometric copies of P.

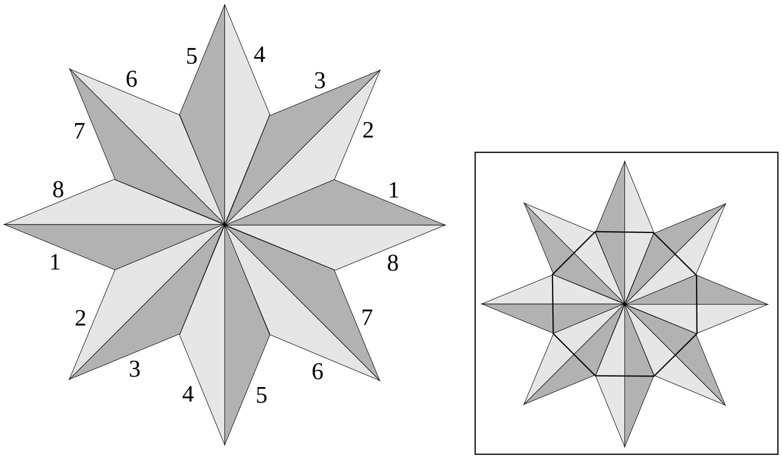

Figure 17.2. Gluing diagram for a translation surface

Figure 17.2 shows the resulting gluing diagram. We have chosen the translations so that all the long sides have already been glued together. Also, we have colored the triangles alternately light and dark so as to better show the pattern. The numbers around the outside of the figure indicate the gluing pattern for the short edges.

The gluing pattern in Figure 17.2 has an alternate description. Take two regular Euclidean octagons and glue each side of one to the opposite side of the other. The smaller inset picture in Figure 17.2 shows one of the two octagons. The other octagon is splayed open, and made by gluing together the pieces that are outside the octagon shown.

A path $\gamma \in \widehat{P}$ is called *straight* if every point $p \in \gamma$ has a neighborhood U with the following property. any isometry between U and a subset of $\boldsymbol{R}^2$ maps $\gamma \cap U$ to a straight line segment. (For concreteness we can always take U to be a little Euclidean ball centered at p.) There is an obvious map $\pi : X \to P$. We just forget the group element involved. This forgetting respects the way we have done the gluing and so π is a well-defined continuous map from $\widehat{P}$ to P. The map π is somewhat like a covering map, except that it is not locally a homeomorphism around points on the edges or vertices.

Lemma 17.4 *Suppose $\widehat{\gamma}$ is a straight path on $\widehat{P}$ which does not go through any vertices of $\widehat{P}$. Then $\gamma = \pi(\widehat{\gamma})$ is a billiard path on P.*

Proof: By construction γ is a polygonal path whose only vertices are contained in the interiors of edges of P. We just have to check the perfect K condition at each vertex. You can see why this works by building a physical model: Take a piece of paper and make a crease in it by folding it in half (and then unfolding it.) Now draw a straight line on the paper which crosses the crease. This straight line corresponds to a piece of $\widehat{\gamma}$ which crosses an edge. When you fold the paper in half you see the straight line turn back at the crease and form a perfect K. This folded path corresponds to γ. ♠

The converse is also true:

Lemma 17.5 *Suppose that γ is a billiard path on P. Then there is a straight path $\widehat{\gamma}$ on $\widehat{P}$ such that $\pi(\widehat{\gamma}) = \gamma$.*

Proof: We use the fact that the map π is almost a covering map. Think of γ as a parametrized path $\gamma : \boldsymbol{R} \to P$, with $\gamma(0)$ contained in the interior of P. We define $\widehat{\gamma}(0)$ to be the corresponding interior point of (P, g), where $g \in G$ is any initial element of G we like. We can define $\widehat{\gamma}(t)$ until the first value $t_1 > 0$ such that $\gamma(t_1)$ lies on an edge, say e_1, of P. But then we can define $\widehat{\gamma}$ in a neighborhood of t_1 in such a way that $\widehat{\gamma}(t_1 - s) \in (P, g)$ and $\widehat{\gamma}(t + s) \in (P, e_1 g)$ for $s > 0$ small. If you think about the folding construction described in the previous lemma, you will see that the straight path $\gamma(t_1 - \epsilon, t_1 + \epsilon)$ projects to $\widehat{\gamma}(t_1 - \epsilon, t_1 + \epsilon)$. Here ϵ is some small value which depends on the location of $\gamma(t_1)$. We can define $\widehat{\gamma}$ for $t > t_1$ until we reach the next time t_2 such that $\gamma(t_2)$ lies in an edge of P. Then we repeat the above construction for parameter values in a neighborhood of t_2. And so on.

This process continues indefinitely, and defines $\widehat{\gamma}$ for all $t \geq 0$. Now we go in the other direction and define $\widehat{\gamma}$ for all $t < 0$. ♠

Note that $\widehat{\gamma}$ is a closed loop in $\widehat{P}$ if and only if γ is a periodic billiard path. Thus, the closed straight loops in $\widehat{P}$ correspond, via π, to periodic billiard paths in P.

Exercise 6. Suppose that P is the regular 7-gon. What is the Euler characteristic of $\widehat{P}$? As a much harder problem, can you find a formula for the Euler characteristic of $\widehat{P}$ as a function of the angles of P?

Exercise 7. The same construction can be made when P has some irrational angles. What do you get if P is a right triangle with the two small angles irrational multiples of π?

17.6 Special Maps on a Translation Surface

We would like to understand how straight lines move around on the polygon P. Recall that $\widehat{P}$ is the translation surface made by suitably gluing together finitely many copies of P. We are going to prove a dynamical result about the action of certain maps on $\widehat{P}$. The result we prove holds in much greater generality, but to keep the discussion self-contained, we are going to consider only one special map on $\widehat{P}$.

Choose some direction on $\widehat{P}$. Given $x \in \widehat{P}$, let $f(x)$ be the point you get by starting at x and moving for one unit in the given direction. The map f is defined except at those points x whose corresponding path hits a cone point. This means that f is defined except on a set contained in a finite union of line segments. When f is defined at x, a sufficiently small disk about x just moves forward in exactly the same way that x does. This means that f is an isometry when restricted to sufficiently small disks. On a large disk, which cuts across the set where f is not defined, f is a piecewise isometry; it has the effect of splitting the disk into several pieces and mapping each piece isometrically to some place on $\widehat{P}$.

Given a set $S \subset \widehat{P}$, we define

$$\text{area}(S) = \sum_{i=1}^{n} \text{area}(S \cap P_i) \tag{67}$$

Here $P_1, \ldots, P_n$ are the polygons out of which $\widehat{P}$ is made. This definition assumes that you know how to compute area inside the Euclidean plane. For the kinds of complicated sets which arise in dynamical systems, one actually needs some measure theory to give a rigorous definition. In our setting here, we are just going to be computing the areas of sets which are obtained by cutting a disk into finitely many pieces along straight line segments.

Exercise 8. Let Δ be a disk. Let $\Delta_n \subset \Delta$ denote the set of points $p \in \Delta$ such that f^k is well-defined on p for all $k = 1, \ldots, n$. Note that $\Delta - \Delta_n$ is contained in a finite union of line segments. Define $f^n(\Delta) = f(\Delta_n)$. Prove that $f^n(\Delta)$ and Δ have the same area. (*Hint*: $f^n(\Delta)$ is is obtained by translating small pieces of Δ_n isometrically to various parts of $\widehat{P}$. These pieces do not overlap because f^{-1} exists and has the same properties as f.)

Theorem 17.6 *Let $p \in \widehat{P}$ be any point on which f and all its iterates are defined, and let $\epsilon > 0$ be arbitrary. Then there is some $q \in \widehat{P}$ and some n such that $d(p,q) < \epsilon$ and $d(p, f^n(q)) < \epsilon$.*

Proof: Let Δ be the disk of radius ϵ about p. Let $D_0 = \Delta$ and let $D_n = F^n(\Delta)$. By Exercise 8, the sets $D_0, D_1, D_2, \ldots$ all have the same area. Since $\widehat{P}$ has finite area, these sets cannot all be disjoint from each other. Hence there are two sets D_a and D_b, which intersect at some point x_a. We take $a < b$. But then D_{a-1} and D_{b-1} intersect at $x_{a-1} = f^{-1}(x_a)$. Continuing in this way, we see that D_0 and D_{b-a} intersect at some point x_0. By construction x_0 lies within ϵ of p and $f^{b-a}(x_0)$ also lies within ϵ of p. ♠

Theorem 17.6 works more generally when all we know is that f is an area preserving map of $\widehat{P}$ that is defined except on a set of zero area (or, more technically zero measure). The proof is essentially the same, but one has to deal more carefully with the concept of area.

Even for area preserving maps, Theorem 17.6 is a toy version of a much stronger and more general result known as the *Poincaré Recurrence Theorem*.

17.7 Existence of Periodic Billiard Paths

It is a theorem of Howie Masur that every rational polygon has a periodic billiard path. In fact, Masur gives bounds on the number of such billiard paths of length at most L. He proves that there are at least $L^2/C - C$ of them, and at most $CL + C$ of them, for some constant C which depends on the polygon. In some cases, it is possible to get sharper results. For instance:

Exercise 9. Prove that there is a constant C such that

$$\lim_{L \to \infty} N(L)/L^2 = C,$$

where $N(L)$ is the number of periodic billiard paths of length less than L on the unit square. What is C?

In this section, I will sketch an elementary proof, due to Boshernitsyn, that every rational polygon has at least one periodic billiard path. You will see that the proof actually gives the existence of many periodic billiard paths, but no bounds like the ones mentioned above.

We choose a direction perpendicular to one of the sides of P and let $f : \widehat{P} \to \widehat{P}$ be the function considered in the previous section. Let $p \in \widehat{P}$ be some point. We think of p as the lift of $\gamma(0)$, where $\gamma : \boldsymbol{R} \to P$ is a billiard path which, at time 0, is travelling perpendicular to a side of P. That is, γ is travelling parallel to V at time 0. By Theorem 17.6, there is some q very close to p and some n so that $q = \widehat{\beta}(0)$ and $f^n(q) = \widehat{\beta}(n)$ are very close together, and $\widehat{\beta}(0)$ is very close to $\widehat{\gamma}(0)$. Here $\widehat{\beta}(0)$ is a straight path in $\widehat{P}$ which goes through q at time 0.

If $\widehat{\beta}(0)$ and $\widehat{\gamma}(0)$ are sufficiently close, then these two points are on the same polygon of $\widehat{P}$. Hence β and γ are travelling in the same direction at time 0. Likewise β is travelling in the same direction at times 0 and n. In short, β travels perpendicular to a side of P at time 0 and also at some much later time n. This means that β hits the same side of P twice, and both times at right angles. But then β is periodic. Each time it hits P perpendicularly, β just reverses itself and retraces its path. Figure 17.3 shows an example of such a path.

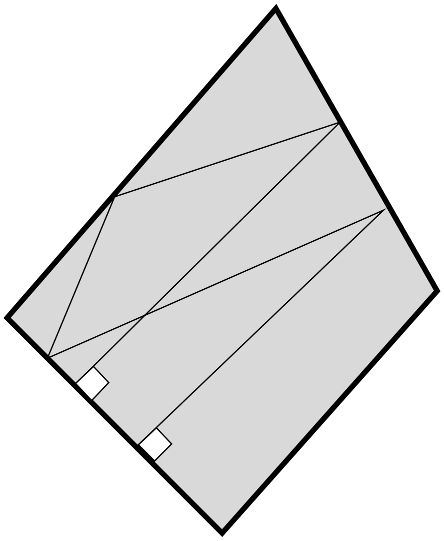

Figure 17.3. A periodic billiard path

18 Translation Surfaces and the Veech Group

In the previous chapter we explained a construction which starts with a rational polygon and produces a translation surface. The straight line flow on the polygon controls the nature of billiards in the polygon. In this chapter we will study the group of affine automorphisms of a translation surface. This group is known as the *Veech group*.

It turns out that the Veech group can be interpreted as a group of symmetries of the hyperbolic plane. So, starting with polygonal billiards, we get back to hyperbolic geometry. We will work out a nontrivial example of a Veech group at the end of the chapter. I learned this particular example from Pat Hooper, and the presentation I give is pretty close to the way he explained it to me.

A lot of the material in this chapter can be found in various surveys of rational billiards; see, e.g., [MAT].

18.1 Affine Automorphisms

Recall that an *affine map* of $\boldsymbol{R}^2$ is a map of the form $x \to Ax + B$, where A is a 2×2 invertible and orientation-preserving matrix and B is another vector. If $B = 0$, then the map is linear. Note that the set of affine maps of $\boldsymbol{R}^2$ forms a group under composition.

Suppose that Σ is a translation surface. An affine automorphism of Σ is a homeomorphism $\phi : \Sigma \to \Sigma$ such that the following hold:

- ϕ permutes the nontrivial cone points of Σ.
- Every ordinary point of Σ has a neighborhood in which ϕ is an affine map.

The second condition needs a bit more explanation. Let $p \in \Sigma$ be an ordinary point. This is to say that there is a small disk Δ_p about p and an isometry I_p from Δ_p to a small disk in $\boldsymbol{R}^2$. The same goes for the point $q = \phi(p)$. The map $I_q \circ \phi \circ I_p^{-1}$ is defined on the open set $U = I_p(\Delta_p) \subset \boldsymbol{R}^2$ and maps it to another open set $I_q(\Delta_q) \subset \boldsymbol{R}^2$. The second condition says that this map is the restriction of an affine map to U.

We denote the set of all affine automorphisms of Σ as $A(\Sigma)$. It is easy to see that the composition of two affine automorphisms of Σ is again an affine automorphism. Likewise, the inverse of an affine automorphism of Σ is an affine automorphism of Σ. In short, $A(\Sigma)$ is a group.

Exercise 1. Let A be a 2×2 matrix with integer entries and determinant 1. Let B any vector, and let Σ be the square torus. You can think of Σ as $(\boldsymbol{R}/\boldsymbol{Z})^2$. Let ϕ be the map $\phi([x]) = [Ax + B]$. Prove that ϕ is an affine automorphism of Σ. Thus, the square torus has a huge affine automorphism group.

Exercise 2. Give an example of a translation surface which has no non-trivial affine automorphisms.

Exercise 3 (Challenge). The affine automorphisms group of the square torus is uncountable since it contains any translation. However, prove that the affine automorphism group of a surface with at least one cone point is countable. (*Hint*: It suffices to consider the subgroup G that preserves all the cone points. Try to show that this subgroup is discrete, in the sense that any element of G sufficiently close to the identity must actually be the identity. Draw many segments connecting all the cone points, and consider the action of an element near the identity on these many line segments.)

18.2 The Diffential Representation

Let $SL_2(\boldsymbol{R})$ denote the group of 2×2 matrices having real entries and determinant 1. Given a group A, a *representation* of A into $SL_2(\boldsymbol{R})$ is a homomorphism $\rho : A \to SL_2(\boldsymbol{R})$. Here is one explanation for this terminology: The elements of A might be somehow abstract, but a representation is a way of, well, representing these elements concretely as matrices. A representation doesn't have to be one-to-one or onto, but of course representations with these additional properties are especially nice.

Here we explain a canonical representation $\rho : A(\Sigma) \to SL_2(\boldsymbol{R})$. The basic property of Σ we use is that there are canonical identifications between any pair of tangent planes $T_p(\Sigma)$ and $T_q(\Sigma)$, defined as follows: By Theorem 17.2, there exists a parallel vector field on $\Sigma - C$, where C is the set of cone points. Given $p, q \in S - C$, we can find an isometry I from a neighborhood of p to a neighborhood of q such that $I(p) = q$. If we insist that I preserves both the orientation and the parallel field, then I is unique. Moreover, I is independent of the choice of parallel field. The differential dI isometrically maps $T_p(\Sigma)$ to $T_q(\Sigma)$. We set $\phi_{pq} = dI$. So, in short

$$\phi_{pq} : T_p(\Sigma) \to T_q(\Sigma) \tag{68}$$

is a canonical isometry. One immediate consequence of our definition is that

$$\phi_{pr} = \phi_{qr} \circ \phi_{pq}, \qquad \phi_{qp} = \phi_{pq}^{-1}. \tag{69}$$

Now, given an element $f \in A(\Sigma)$ we choose and ordinary point $p \in \Sigma$, and let $q = f(p)$. Let df_p be the differential of f at p. This means that df_p is a linear map from $T_p(\Sigma)$ to $T_q(\Sigma)$. Note that the composition

$$M(f,p) = \phi_{qp} \circ df_p$$

is a linear isomorphism from $T_p(\Sigma)$ to itself. Using the isometry I_p, we can identify $T_p(\Sigma)$ with, say, the tangent plane to $\boldsymbol{R}^2$ at the origin. We let $\rho(f)$ be the linear transformation of $\boldsymbol{R}^2$ which corresponds to $M(f,p)$ under the identification.

We claim that $\rho(f)$ is independent of the choice of point p. To see this, we note that the map $\rho(f)$ has the following alternate description. Using the coordinate charts I_p and I_q discussed above, the map $\rho(f)$ is just the linear part of

$$dI_q \circ df_p \circ dI_p^{-1}.$$

The linear part of an affine map does not depend on the point. Hence $\rho(f)$ has the same definition independent of which point we use inside our local coordinate chart. But the surface is connected, so $\rho(f)$ does not depend on the choice of point at all.

The determinant of $\rho(f)$ measures the factor by which f increases area in a neighborhood of any point. Since the whole surface has finite area and $\rho(f)$ is an automorphism, $\rho(f)$ must have determinant 1. Hence we can interpret $\rho(f)$ as an element of $SL_2(\boldsymbol{R})$. The map $f \to \rho(f)$ is a homomorphism because of the chain rule: The linear differential of a composition of maps is just the composition of the linear differential of the invididual maps. And composition of linear maps is the same thing as matrix multiplication in $SL_2(\boldsymbol{R})$.

We have now constructed the representation $\rho : A(\Sigma) \to SL_2(\boldsymbol{R})$. We let $V(\Sigma) = \rho(A(\Sigma))$. The matrix group $V(\Sigma)$ is sometimes called the *Veech group*. Below we will work out the Veech group associated to the "double octagon" example discussed toward the end of §17.5. Before we get to examples, however, we need to develop a bit more of the theory.

18.3 Hyperbolic Group Actions

Recall that $\boldsymbol{H}^2$ is the hyperbolic plane. We work in the upper half plane model. Every element of $SL_2(\boldsymbol{R})$ acts on $\boldsymbol{H}^2$ isometrically, as a linear fractional transformation; see §10.3. In particular, the Veech group V acts on $\boldsymbol{H}^2$. The *orbit* of a point $x \in \boldsymbol{H}^2$ is defined to be the set

$$\{g(x) |\ g \in V\}.$$

We define an equivalence relation on points in $\boldsymbol{H}^2$ by saying that two points are equivalent iff they lie in the same orbit.

V is said to act *properly discontinuously* on $\boldsymbol{H}^2$ if, for every metric ball $B \subset \boldsymbol{H}^2$, the set

$$\{g \in V |\ g(B) \cap B \neq \emptyset\}$$

is a finite set. In other words, all but finitely elements of V have such a drastic action on $\boldsymbol{H}^2$ that they move the ball B completely off itself.

Exercise 4. Let $SL_2(\boldsymbol{Z})$ be the group of 2×2 integer matrices having determinant 1. Prove that $SL_2(\boldsymbol{Z})$ acts properly discontinuously on $\boldsymbol{H}^2$.

Before we establish the main result in this section, we give one more definition. Two groups $G_1, G_2 \in SL_2(\boldsymbol{R})$ are *conjugate* if there is some $g \in SL_2(\boldsymbol{R})$ such that $G_2 = gG_1g^{-1}$.

Exercise 5. Suppose that G_1 and G_2 are conjugate. Prove that G_1 acts properly discontinuously on $\boldsymbol{H}^2$ if and only if G_2 does.

Theorem 18.1 *If V is the Veech group of a surface, then V acts properly discontinuously on $\boldsymbol{H}^2$.*

We will sketch the proof of Theorem 18.1 in the next section.

Whether or not V acts properly discontinuously, we can form the quotient $\boldsymbol{H}^2/V$ as follows. We define two points $x, y \in \boldsymbol{H}^2$ to be equivalent if there is some $g \in V$ such that $g(x) = y$. Then $\boldsymbol{H}^2/V$ is defined to be the set of equivalence classes of points. In the case where V acts properly discontinuously, the quotient is particularly nice:

Theorem 18.2 *If V acts properly discontinuously on $\boldsymbol{H}^2$, then we can remove a countable discrete set of points T from $\boldsymbol{H}^2$ such that the quotient $(\boldsymbol{H}^2 - T)/V$ is a hyperbolic surface.*

Proof: Before we start we note that all the elements of V act in an orientation-preserving way, so that there are no reflections in V. (For the orientation-reversing case, the statement of the result is slightly different.)

Let T be the set of points $x \in \boldsymbol{H}^2$ such that $g(x) = x$ for some nontrivial $g \in V$. The set T must be discrete in the sense that there is some $\epsilon > 0$ such that any ball of radius ϵ contains at most one point of T. Otherwise we could find some ball B which contained infinitely many points of T, and we would contradict the proper discontinuity. Note that T is invariant under V: If $x \in T$ is fixed by g, then $y = h(x)$ is fixed by hgh^{-1}. Thus, the quotient $(\boldsymbol{H}^2 - T)/V$ makes sense. Every $x \in \boldsymbol{H}^2 - T$ has a neighborhood Δ_x such that $g(\Delta_x) \cap \Delta_x = \emptyset$ for any nontrivial g. To see this, let d_g denote the hyperbolic distance between $g(x)$ and x. Since $x \not\in T$, the number d_g is positive. The proper discontinuity prevents there being a sequence $\{g_i\}$ with $\{d_{g_i}\}$ converging to 0. Hence there is some positive lower bound to d_g, which is what we need.

Now we know that each $x \in \boldsymbol{H}^2 - T$ has a little neighborhood which is moved completely off itself by all of G (except the identity). This little neighborhood therefore maps injectively into the quotient $(\boldsymbol{H}^2 - T)$ and serves as a coordinate chart about x. ♠

Note that the quotient $\boldsymbol{H}^2/V$ still makes sense, and actually it is obtained from $(\boldsymbol{H}^2 - T)/V$ just by adding finitely many points. We define the *covolume* of V to be the volume of $(\boldsymbol{H}^2 - T)/V$. The group V is said to be a *lattice* if V has finite covolume. Σ is said to be a *Veech surface* if V is a lattice. For instance, $SL_2(\boldsymbol{Z})$ is a lattice.

18.4 Proof of Theorem 18.1

We first take care of a trivial case of Theorem 18.1.

Exercise 6. Suppose that Σ is a translation surface with no cone points. Prove that Σ is isometric to a flat torus.

Exercise 7. Prove Theorem 18.1 in the case when the surface has no cone points.

From now on, we consider the case when Σ has at least one cone point. In this case, Σ is homeomorphic to a surface having negative Euler characteristic. Let C be the set of cone points of Σ. We call a map $\gamma : [0,1] \to \Sigma$ a *saddle connection* if the follwing hold.

- $\gamma(t) \in C$ if and only if $t = 0, 1$.
- The restriction of γ to $(0,1)$ is locally a straight line.

Exercise 8. Prove that Σ has a pair of non-parallel saddle connections that intersect at a point of $\Sigma - C$.

Lemma 18.3 *Let f be an affine automorphism of Σ. Let γ_1 and γ_2 be a pair of saddle connections, as in Exercise 8. Suppose f preserves the endpoints of γ_1 and γ_2, and $f(\gamma_j) = \gamma_j$ for $j = 1, 2$. Then f is in the kernel of the differential representation ρ.*

Proof: The restriction of an affine map to a straight line is just a dilation. Hence, the restriction of f to γ_j is just a dilation. Since $f(\gamma_j) = \gamma_j$, the dilation factor must be one: the total length is preserved. So f is the identity on γ_j.

Let p be an intersection point of γ_1 and γ_2. We know that $f(p) = p$. Since γ_1 and γ_2 are nonparallel, we see that df_p fixes two independent directions at p. Hence df_p is the identity. But then $\rho(f)$ is the identity. ♠

We suppose that there is some ball B and an infinite collection $\{g_i\} \in V$ such that $g_i(B) \cap B \neq \emptyset$. It is a general principle of compactness that there must be elements of our set which are arbitrarily close to each other. Hence, we can find an infinite list of distinct elements of V whose action on $\boldsymbol{H}^2$ converges to the action of the identity element.

What this means in terms of Σ is that we can find an infinite sequence $\{f_j\}$ of affine automorphisms such that $\rho(f_i)$ is not the identity but $\rho(f_i)$ converges to the identity as $i \to \infty$. All these elements permute the set of cone points somehow. So, by taking suitable powers of our elements, we can assume that each f_i fixes each cone point of Σ.

Let γ_1 and γ_2 be the saddle connections from Exercise 8. The segment $f_k(\gamma_1)$ is another saddle connection that connects the same two cone points as does γ_1. For k large, $f_k(\gamma_1)$ and γ_1 nearly point in the same direction and nearly have the same length. If they do not point in exactly the same direction, they cannot connect the same two endpoints. The two paths start out at the same cone point but then slowly diverge, so that one of them misses the cone point at the other end. Figure 18.1 shows what we mean.

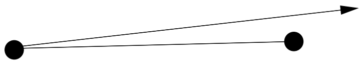

Figure 18.1. Nearly parallel paths

This means that $f_k(\gamma_1)$ and γ_1 point in exactly the same direction for k large. But then $f_k(\gamma_1) = \gamma_1$. The same argument shows that $f_k(\gamma_2) = \gamma_2$ for k large. But then, by the previous result, $\rho(f_k)$ is the identity for large k. This contradiction finishes the proof.

18.5 Triangle Groups

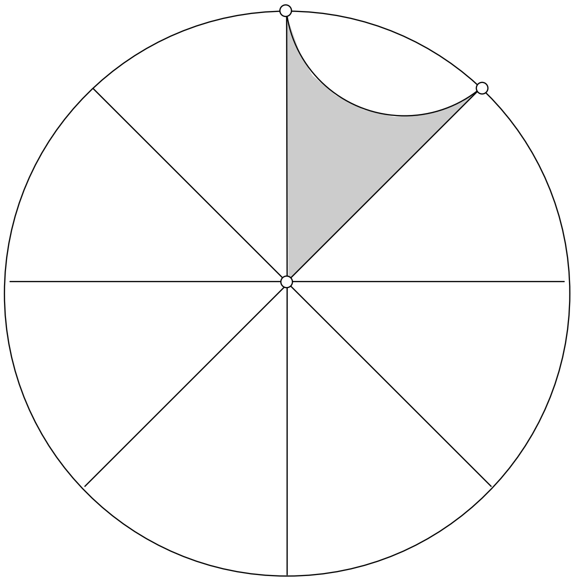

Figure 18.2. The hyperbolic triangle of interest

Recall that a geodesic hyperbolic triangle is a triangle in $\boldsymbol{H}^2$ whose sides are either geodesic segments, geodesic rays, or geodesics. The case of interest to us is the geodesic triangle with 2 ideal vertices and one other vertex having interior angle $2\pi/8$. Figure 18.2 shows a picture of the triangle we mean, drawn in the disk model. This triangle is known as the $(8,\infty,\infty)$ triangle.

Lemma 18.4 *Let γ be any geodesic in $\boldsymbol{H}^2$. Then there is an order 2 hyperbolic isometry which fixes γ.*

Proof: Thinking of $\boldsymbol{H}^2$ as the upper half-plane, the map $z \to -\overline{z}$ fixes the imaginary axis, which is a geodesic. We have already seen that any two geodesics are isometric to each other. If g is an isometry taking the geodesic γ_1 to the geodesic γ_2 and I is an order 2 isometry fixing γ_1, then gIg^{-1} is the desired order 2 isometry fixing γ_2. Thus, we can start with the one reflection desribed above and construct all the others by conjugation. ♠

The order 2 hyperbolic isometry fixing γ is called a *hyperbolic reflection* in γ. Given any geodesic triangle Δ, we can form the group $G(\Delta) \subset SL_2(\boldsymbol{R})$ as follows. We let I_1, I_2, I_3 be hyperbolic reflections fixing the 3 sides of Δ and then we let $G(\Delta)$ be the group generated by words of even length in I_1, I_2, I_3. For instance, I_1I_2 and $I_1I_2I_1I_3$ all belong to G but $I_1I_2I_3$ does not. All the elements in G are orientation preserving and it turns out that we can find matrices in $SL_2(\boldsymbol{R})$ for the elements I_1I_2, I_2I_3, and I_3I_1. This is enough to show that G actually comes from a subgroup of $SL_2(\boldsymbol{R})$.

18.6 Linear and Hyperbolic Reflections

As preparation for the Veech group example we will work out, we discuss how to convert between certain linear maps as they act on $\boldsymbol{R}^2$ and the corresponding linear fractional actions on $\boldsymbol{H}^2$.

Say that a *linear reflection* is a linear transformation $T : \boldsymbol{R}^2 \to \boldsymbol{R}^2$ such that $T(v) = v$ and $T(w) = -w$ for some basis $\{v, w\}$ of $\boldsymbol{R}^2$. The corresponding linear fractional transformation acting on $\boldsymbol{H}^2$ is a hyperbolic reflection. This can be seen by considering the special case when $v = (1, 0)$ and $w = (0, 1)$: all other cases are conjugate to this one.

The map T is determined by the pair (v, w), but more than one basis determines T. The basis (C_1v, C_2w) also determines T, where C_1 and C_2 are any 2 nonzero constants. For this reason, it is really the pair (L_1, L_2) that determines T, where L_1 is the line through v and L_2 is the line through w. The map T fixes L_1 pointwise and reverse L_2.

The map $-T$ fixes L_2 pointwise and reverses L_1. For this reason, the unordered pair $\{L_1, L_2\}$ determines the pair of maps $\{T, -T\}$. The map $\pm T$ corresponds to a hyperbolic reflection, and each hyperbolic reflection corresponds to a pair $\pm T$ of maps. In short, each hyperbolic reflection is determined by an unordered pair $\{L_1, L_2\}$ of lines through the origin. We call such a pair of lines a *cross*.

Let us first consider the case when L_1 and L_2 are perpendicular. In this case, we call $\{L_1, L_2\}$ a *plus*, because the two lines look like a + symbol, up to rotation. If we work in the disk model Δ of the hyperbolic plane, we can normalize so that the hyperbolic reflections corresponding to pluses all fix some geodesic through the origin in $\boldsymbol{C}$. Figure 18.3 shows two examples. On the left-hand side of Figure 18.3 we show two pluses, one drawn thickly and one drawn thinly. On the right hand side of Figure 18.3, we show the geodesics in Δ fixed by the corresponding hyperbolic reflections.

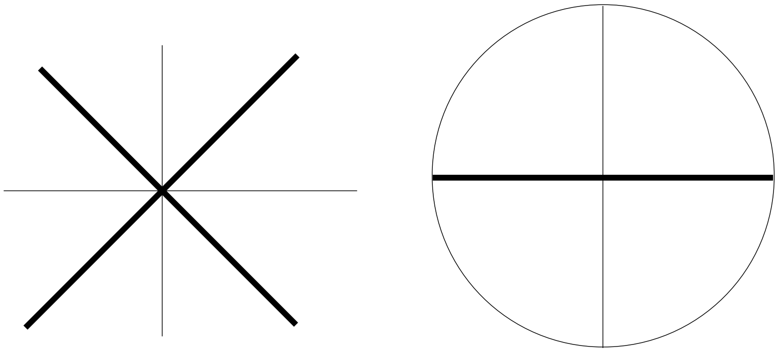

Figure 18.3. Euclidean and hyperbolic reflections

Exercise 9. Let θ be the smallest angle between the lines of one plus and the lines of another. Prove that the corresponding geodesics in Δ meet at an angle of 2θ.

In light of Exercise 9, we can draw 3 crosses whose corresponding geodesics in Δ are three sides of the $(8, \infty, \infty)$ triangle shown in Figure 18.2. Two of the crosses are pluses and one is not. The crosses are drawn thickly, and the thin lines are present for reference. The thin lines are evenly spaced in the radial sense.

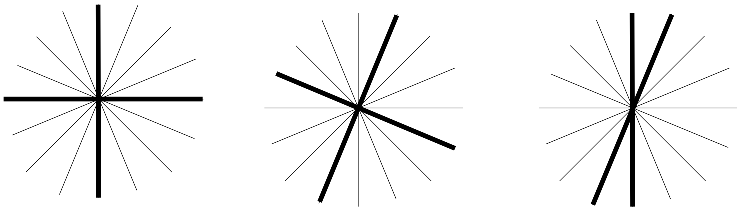

Figure 18.4. Three special crosses

Here is why this works. Let $\pm T_1$, $\pm T_2$, and $\pm T_3$ be the (pairs of) hyperbolic reflections corresponding to each of the three crosses. Let R_j be the hyperbolic reflection corresponding to $\pm T_j$. By construction R_1 and R_2 each fix one of the "Euclideanly straight" sides of the triangle in Figure 18.2.

We claim that R_3 fixes the third side of the triangle in Figure 18.2. The central point is that the third cross shares a line with each of the first two crosses. If the signs are appropriately chosen, the element T_1T_3 is a parabolic element that fixes the vertical line through the origin. From this we see that R_1R_3 is a parabolic element fixing the top vertex of our triangle. This is only possible if R_3 fixes this same point. A similar argument shows that R_3 fixes the other ideal vertex of our triangle. Hence, as claimed, R_3 fixes the edge joining these vertices.

We have gone through all this trouble because we want to recognize the $(8,\infty,\infty)$ triangle group as a subgroup of the group of all affine automorphisms of a certain translation surface. We will work this out in the next section.

18.7 Behold, The Double Octagon!

We will compute the Veech group of the translation surface associated to the Euclidean isosceles triangle having small angle $2\pi/8$. As we saw in §17.5, this surface is obtained from a gluing diagram involving two regular Euclidean octagons. Each side of one octagon is glued to the opposite side of the other. Let Σ be this surface.

Theorem 18.5 *$V(\Sigma)$ is the even subgroup of the $(8,\infty,\infty)$ reflection triangle group.*

The $(8,\infty,\infty)$ triangle group is the group generated by the three hyperbolic reflections R_1, R_2, R_3 considered in the previous section. The *even subgroup* consists of elements made from composing an even number of these elements. The even subgroup has index 2 in the whole group. The point is that every element of the reflection triangle group is either odd or even.

We will sketch a proof of Theorem 18.5. To make things work well, we define an *anti-affine automorphism* to be a homeomorphism of Σ which is locally anti-affine, meaning that the map locally has the form $x \to L(x) + C$, where L is an orientation-reversing linear map and C is some constant vector. The linear reflections considered in the previous section are of this form.

Let $\widehat{A}(\Sigma)$ be the group of these maps, and let $\widehat{V} = \rho(\widehat{A})$, where ρ is the differential representation as above. We will show that $\widehat{V}$ coincides with the group $\widehat{G}$ generated by the reflections in the sides of the $(8, \infty, \infty)$ triangle. The odd elements of $\widehat{A}$ are orientation reversing and the even elements are orientation preserving. So, the Veech group corresponds to the images of the even elements.

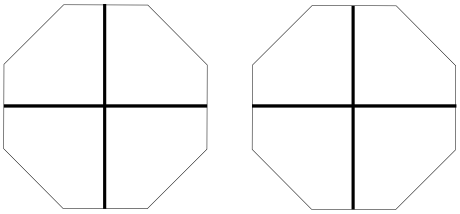

Figure 18.5. The first cross

Figure 18.5 shows the octagons involved in the gluing diagram for Σ. Again, each side of the left octagon is glued to the opposite side of the right octagon by a translation. Simultaneous reflection in the vertical sides of Σ induces an element T_1 of $\widehat{A}$. The differential of this map, evaluated at the center of the first octagon, fixes the vertical line through the center and reverses the horizontal line. The element $\pm dT_1$ therefore corresponds to the first plus in Figure 18.4. Hence $\rho(\pm T_1) = R_1$. Figure 18.6 does for R_2 with Figure 18.5 does for R_1. Here we take T_2 to be simultaneous reflection in the diagonals of positive slope.

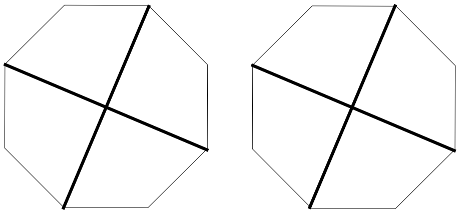

Figure 18.6. The second cross

So far we have used fairly trivial symmetries of our surface. Now we have to do something nontrivial to see the anti-affine automorphism that corresponds to the third cross. Figure 18.7 shows the cross $\{L_1, L_2\}$ we are aiming for, drawn on one of the octagons. The auxilliary line L_3 will be explained momentarily.

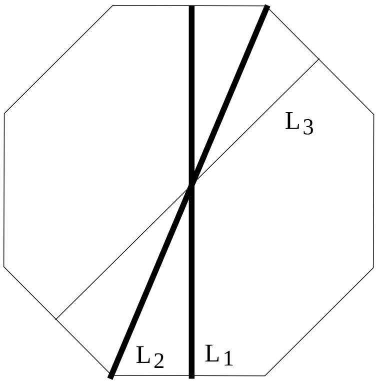

Figure 18.7. The third cross

We will produce an automorphism $g : \Sigma \to \Sigma$ such that g fixes L_2 pointwise and $g(L_1) = L_3$ in a length-preserving and height-reversing way. That is, g maps the top vertex of L_1 to the bottom vertex of L_3 and vice versa. At the same time, the map T_2 fixes L_2 pointwise and maps L_3 to L_1 in a length-preserving way and height-preserving way. But then the composition $T_3 = T_2 \circ g$ fixes L_2 pointwise and reverses L_1. By construction, the maps $\pm T_3$ correspond to our third cross. We set $R_3 = \rho(\pm T_3)$, and we have the desired map.

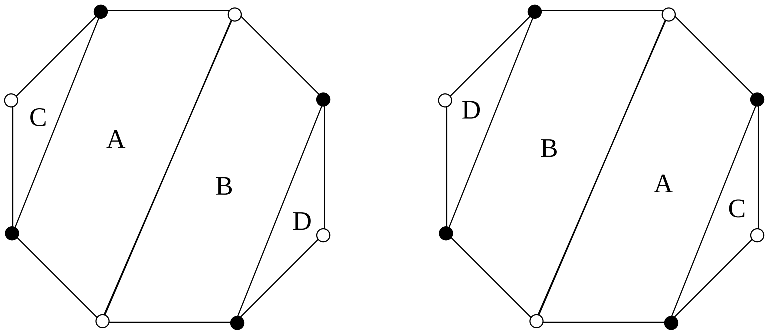

Figure 18.8. Cylinder decomposition

Now we turn our attention to the construction of the map g. Figure 18.8 shows a decomposition of Σ into 4 cylinders, labelled A, B, C, D. Remember, each side of the left cylinder is glued to the opposite side of the right cylinder. Thus, for instance, the two A pieces on the left and right glue together to make the A cylinder. The A and B cylinders are isometric to each other and the C and D cylinders are isometric to each other. Here is the miracle that makes everything work.

Exercise 10. Prove that the A and C cylinders are similar to each other. Hence, all 4 cylinders are similar to each other.

For starters, we have g do the same thing on each octagon. Figure 18.9 shows how g acts on one of the octagons. g maps the points labelled x to the points labelled y, in the manner suggested by the arrows. These points are at the midpoints of the relevant edges.

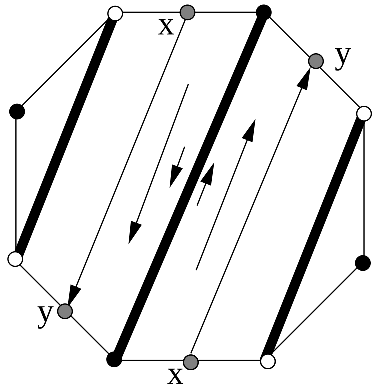

Figure 18.9. Action of the automorphism

Assume for the moment that there really is a locally affine automorphism of Σ that has this action. That is, assume that g really exists. By construction g fixes L_1 pointwise and g maps L_1 to L_3 in a length-preserving and height-reversing way. The point is that L_1 connects the two x points and L_3 connects the two y points as shown in Figure 18.9.

It only remains to show that g actually exists. First of all, we define g in a neighborhood of the "centerline" L_2. We start extending g outward until it is defined on the A cylinder. The lines connecting the x points to the y points glue together to form the central loops of the A and B cylinders. By construction g shifts these central loops half way around. Hence g extends to be the identity on the boundary of the A cylinder. Even though g is the identity on the boundary of the A cylinder, g is not the identity on A: it is what is called a *Dehn twist*. The same discussion works for the B cylinder.

Now we consider the C cylinder. So far, g is defined on one boundary component of the C cylinder, and g is the identity on this boundary component. Because the A and C cylinders are similar, g extends to all of the C cylinder in such a way as to be the identity on both boundary components. The action of g on the C cylinder is the same, up to scaling, as the action of g on the A cylinder. A similar thing works for the D cylinder. So, all in all, g is a Dehn twist of each of the 4 cylinders, and the 4 separate maps fit together seamlessly because g is the identity on every boundary component of every cylinder. This establishes the existence of g.

Now we know that $\widehat{V}(\Sigma)$ contains the $(8,\infty,\infty)$ reflection triangle group. Hence, the Veech group $V(\Sigma)$ contains the even subgroup of the $(8,\infty,\infty)$ reflection triangle group. To finish our proof, we will show that $\widehat{V}(\Sigma)$ is precisely the reflection triangle group. Let Y denote the $(8,\infty,\infty)$ triangle. Let $\widehat{G}$ be the group generated by hyperbolic reflections in the sides of Y.

Exercise 11 (Challenge). Suppose that Γ is a group acting properly discontinuously on $\boldsymbol{H}^2$ and $\widehat{G} \subset \Gamma$. Prove that either $\Gamma = \widehat{G}$ or else Γ is the group generated by the reflections in the sides of the geodesic triangle obtained by bisecting the Y in half.

If $\widehat{V}$ does not equal $\widehat{G}$, then Σ has an extra isometric symmetry which fixes the centers of the octagons. (This corresponds to the extra element, reflection in the bisector of Y.) But the octagons do not have any line of symmetry between the two drawn in our figures above. Hence, this extra symmetry does not exist. Hence $\widehat{V}(\Sigma) = \widehat{G}$. This is what we wanted to prove.

Exercise 12 (Challenge). Do all the same things as above for the translation surface associated to the isosceles triangle having small angles π/n for $n = 4, 6, 8, \ldots$.

19 Continued Fractions

The purpose of this chapter is to describe continued fractions and their connection to hyperbolic geometry. One motivating factor for including a chapter on continued fractions (besides their obvious beauty) is that it gives us a nice way to introduce the modular group. The modular group makes its appearance several times in subsequent chapters. See the book [DAV] for an excellent treatment.

19.1 The Gauss Map

Given any $x \in (0,1)$ we define

$$\gamma(x) = (1/x) - \text{floor}(1/x). \tag{70}$$

Here, floor(y) is the greatest integer less than or equal to y. The Gauss map has a nice geometric interpretation, as shown in Figure 19.1. We start with a $1 \times x$ rectangle, and remove as many $x \times x$ squares as we can. Then we take the left over (shaded) rectangle and turn it 90 degrees. The resulting rectangle is proportional to a $1 \times \gamma(x)$ rectangle. Starting with $x_0 = x$, we can form the sequence $x_0, x_1, x_2, \ldots$ where $x_{k+1} = \gamma(x_k)$. This sequence is defined until we reach an index k for which $x_k = 0$. Once $x_k = 0$, the point x_{k+1} is not defined.

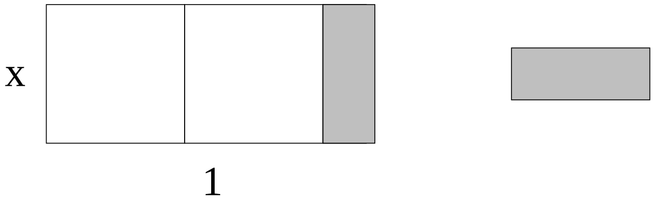

Figure 19.1. Cutting down a rectangle

Exercise 1. Prove that the sequence $\{x_k\}$ terminates at a finite index if and only if x_0 is rational.

Consider the rational case. We have a sequence $x_0, \ldots, x_n$, where $x_n = 0$. We define

$$a_{k+1} = \text{floor}(1/x_k); \qquad k = 0, \ldots, n-1. \tag{71}$$

The numbers a_k also have a geometric interpretation. Referring to Figure 19.1, where $x = x_k$, the number a_{k+1} tells us the number of squares we can remove before we are left with the shaded rectangle. In Figure 19.1, $a_{k+1} = 2$. Figure 19.2 shows a more extended example. Starting with $x_0 = 7/24$, we have the following.

- $a_1 = \text{floor}(24/7) = 3$.
- $x_1 = 24/7 - 3 = 3/7$.
- $a_2 = \text{floor}(7/3) = 2$.
- $x_2 = (7/3) - 2 = 1/3$.
- $a_3 = \text{floor}(3) = 3$.
- $x_3 = 0$.

In Figure 19.2 we can read off the sequence $(a_1, a_2, a_3) = (3, 2, 3)$ by looking at the number of squares of each size in the picture. The overall rectangle is $1 \times x_0$.

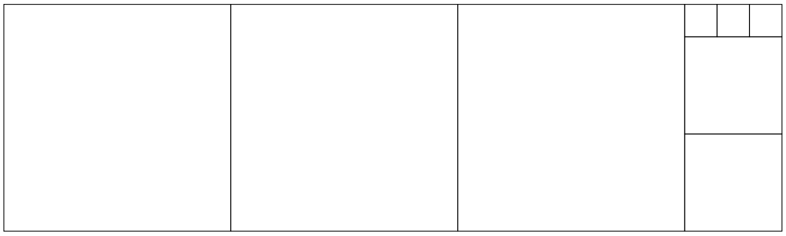

Figure 19.2. A 7/24 by 1 rectangle cut into squares

19.2 Continued Fractions

Again, sticking to the rational case, we can get an expression for x_0 in terms of $a_1, \ldots, a_n$. In general, we have

$$x_{k+1} = \frac{1}{x_k} - a_{k+1},$$

which leads to

$$x_k = \frac{1}{a_{k+1} + x_{k+1}}. \tag{72}$$

But then we can say that

$$x_0 = \frac{1}{a_1 + x_1} = \frac{1}{a_1 + \cfrac{1}{a_2 + x_2}} = \frac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + x_3}}} \cdots. \tag{73}$$

We introduce the notation

$$\alpha_1 = \frac{1}{a_1}, \qquad \alpha_2 = \frac{1}{a_1 + \cfrac{1}{a_2}}, \qquad \alpha_3 = \frac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3}}} \cdots. \tag{74}$$

In making these definitions, we are chopping off the x_k in each expression in equation (73). The value of α_k depends on k, but $x_0 = \alpha_n$ because $x_n = 0$.

Cosidering the example from the previous section, we have

$$\alpha_1 = \frac{1}{3}, \qquad \alpha_2 = \frac{1}{3 + \cfrac{1}{2}} = \frac{2}{7}, \qquad \alpha_3 = x_0 = \frac{7}{24}.$$

We say that two rational numbers p_1/q_1 and p_2/q_2 are *Farey related* if

$$\det \begin{bmatrix} p_1 & p_2 \\ q_1 & q_2 \end{bmatrix} = p_1 q_2 - p_2 q_1 = \pm 1. \tag{75}$$

In this case, we write $p_1/q_2 \sim p_2/q_2$. For instance $1/3 \sim 2/7$ and $2/7 \sim 7/24$. This is no accident.

Exercise 2. Starting with any rational $x_0 \in (0,1)$ we get a sequence $\{\alpha_k\}$ as above. Prove that $\alpha_k \sim \alpha_{k+1}$ for all k.

Exerxise 3. Consider the sequence of differences $\beta_k = \alpha_{k+1} - \alpha_k$. Prove that the signs of β_k alternate. Thus, the sequence $\alpha_1, \alpha_2, \alpha_3, \ldots$ alternately over-approximates and under-approximates $x_0 = \alpha_n$.

Exercise 4. Prove that the denominator of α_{k+1} is greater than the denominator of α_k for all k. In particular, the α-sequence does not repeat. With a little bit of extra effort, you can show that the sequence of denominators grows at least exponentially.

19.3 The Farey Graph

Now we will switch gears and discuss an object in hyperbolic geometry. Let $\boldsymbol{H}^2$ denote the upper half-plane model of the hyperbolic plane. We form a geodesic graph $\mathcal{G}$ in $\boldsymbol{H}^2$ as follows. The vertices of the graph are the rational points in $\boldsymbol{R} \cup \infty$, the ideal boundary of $\boldsymbol{H}^2$. The point ∞ counts as rational, and is considered to be the fraction $1/0$. The edges of the graph are geodesics joining Farey related rationals. For instance, the vertices

$$0 = \frac{0}{1}, \qquad 1 = \frac{1}{1}, \qquad \infty = \frac{0}{1}$$

are the vertices of an ideal triangle T_0 whose boundary lies in $\mathcal{G}$.

Let $\Gamma = SL_2(\boldsymbol{Z})$ denote the group of integer 2×2 matrices acting on $\boldsymbol{H}^2$ by linear fractional transformations. As usual, Γ also acts on $\boldsymbol{R} \cup \infty$. The group Γ is known as the *modular group*.

Technical Remark. Before we launch into a discussion about Γ, there is one technical point we need to clear up. The matrices A and $-A$ give rise to the same linear fractional transformation, so sometimes people introduce the notation $PSL_2(\boldsymbol{Z})$ to denote the quotient group $SL_2(\boldsymbol{Z})/\pm$, in which each element is an equivalence class consisting of $\{A, -A\}$. This irritating distinction really plays no role in our discussions, but you should keep in mind that a *matrix* is really not quite the same thing as a linear fractional transformation, due to the redundancy just mentioned. Nonetheless matrices represent linear transformations.

Exercise 4. Let $g \in \Gamma$ be some element. Suppose $r_1 \sim r_2$. Prove that $g(r_1) \sim g(r_2)$. In particular, g is a symmetry of $\mathcal{G}$.

Now we know that Γ acts as a group of symmetries of $\mathcal{G}$. We can say more. Suppose e is an edge of $\mathcal{G}$, connecting p_1/q_1 to p_2/q_2. The matrix

$$\begin{bmatrix} p_1 & p_2 \\ q_1 & q_2 \end{bmatrix}^{-1}$$

carries e to the edge connecting $0 = 0/1$ to $\infty = 1/0$. We call this latter edge *our favorite*. In other words, we can find a symmetry of $\mathcal{G}$ that carries any edge we like to our favorite edge. Since Γ is a group, we can find an element of Γ carrying any one edge e_1 of $\mathcal{G}$ to any other edge e_2. We just compose the element that carries e_1 to our favorite edge with the inverse of the element that carries e_2 to our favorite edge. In short Γ acts transitively on the edges of $\mathcal{G}$.

Exercise 5. Prove that no two edges of $\mathcal{G}$ cross each other.

We have exhibited an ideal triangle T_0 whose boundary lies in $\mathcal{G}$. Our favorite edge is an edge of this triangle. It is also an edge of the ideal triangle T_1 with vertices

$$\frac{0}{1}, \qquad \frac{1}{0}, \qquad \frac{-1}{1}.$$

The boundary of this triangle lies in $\mathcal{G}$ as well. Thus, our favorite edge is flanked by two ideal triangles whose boundaries lie in $\mathcal{G}$. But then, by symmetry, this holds for every edge of $\mathcal{G}$. Starting out from T_0 and moving outward in a tree-like manner, we recognize that $\mathcal{G}$ is the set of edges of a triangulation of $\boldsymbol{H}^2$ by ideal triangles. Figure 19.3 shows a finite portion of $\mathcal{G}$. The vertical line on the left is our favorite line. The vertical line on the right connects 1 to ∞.

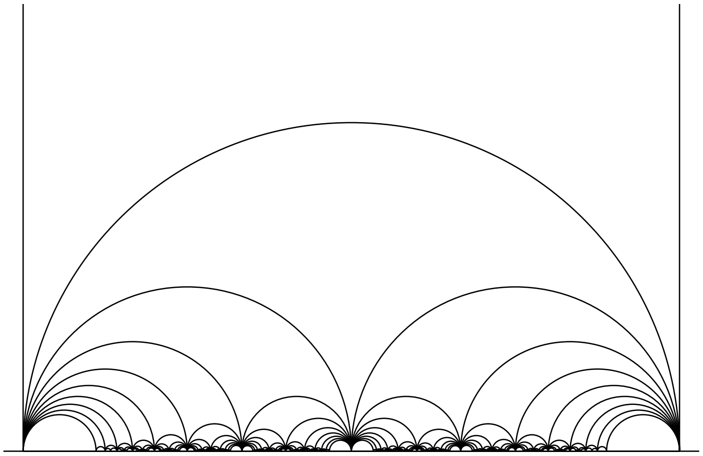

Figure 19.3. A portion of the Farey graph

19.4 Structure of the Modular Group

The Farey graph gives a good way to understand the structure of the modular group. Since it is built out of ideal triangles, the Farey graph has 3-fold symmetry built into it. The matrix

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

represents the order 3 linear fractional transformation $A \in \Gamma$ that permutes the vertices of the ideal triangle T_0 discussed above. More precisely, A has the action

$$0 \to 1 \to \infty \to 0.$$

The Farey graph also has 2-fold symmetry. The matrix

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

represents the order 2 linear fractional transformation $B \in \Gamma$ that has the action

$$0 \to \infty, \qquad 1 \to -1.$$

The element B is a rotation about the "midpoint" of the edge of the Farey graph that joins 0 to ∞. Put another way, B swaps the triangle $T_0 = (0, 1, \infty)$ with the adjacent triangle $T' = (0, -1, \infty)$.

Exercise 6. Prove that the element BAB rotates the triangle T' with vertices $(0, -1, \infty)$.

Exercise 7. Some finite string of letters, just using A and B, is called *reduced* if the strings AAA and BB do not occur in it. We make this definition because the hyperbolic isometries A^3 and B^2 are the identity. Prove that any element of the modular group has the form $w(A, B)$ where w is a reduced word. (*Hint*: Show, just using A's and B's, that you can move your favorite edge to any other edge in two ways. This is exactly what the modular group can do.)

Exercise 8. Let $w(A, B)$ be some nontrivial reduced work. Prove that $w(A, B)$ is a nontrivial element of the modular group.

19.5 Continued Fractions and the Farey Graph

Let's go back to continued fractions and see how they fit in with the Farey graph. Let $x_0 \in (0, 1)$ be a rational number. We have the sequence of approximations $\alpha_1, \ldots, \alpha_n = x_0$ as in equation (74). It is convenient to also define

$$\alpha_{-1} = \infty, \qquad \alpha_{-0} = 0; \tag{76}$$

If we consider the larger sequence $\alpha_{-1}, \ldots, \alpha_n$, the statements of Exercises 2 and 3 remain true. In particular, we have a path $P(x_0)$ in the Farey graph that connects ∞ to x_0, obtained by connecting ∞ to 0 to α_1, etc. The example given above does not produce such a nice picture, so we will give some other examples.

Let $x_0 = 5/8$. This gives us

$$a_1 = \cdots = a_5 = 1$$

and

$$\alpha_1 = 1, \qquad \alpha_2 = \frac{1}{2}, \qquad \alpha_3 = \frac{2}{3}, \qquad \alpha_4 = \frac{3}{5}, \qquad \alpha_5 = x_0 = \frac{5}{8}.$$

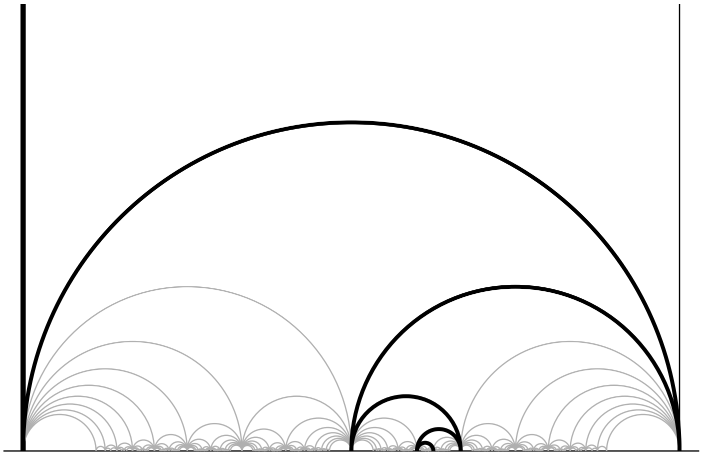

Figure 19.4: The Farey path associated to 5/8

Taking $x_0 = 5/7$ gives

$$a_1 = 1, \qquad a_2 = 2, \qquad a_3 = 2.$$

and

$$\alpha_1 = 1, \qquad \alpha_2 = \frac{2}{3}, \qquad \alpha_3 = \frac{5}{7}.$$

There are three things we would like to point out about these pictures. First, they make a zig-zag pattern. This always happens, thanks to Exercises 3 and 4 above. Exercise 3 says that the path cannot backtrack on itself, and then Exercise 4 forces the back-and-forth behavior.

Second, we can read off the numbers $a_1, \ldots, a_n$ by looking at the "amount of turning" the path makes at each vertex. In Figure 19.4, our path turns "one click" at α_0, then "two clicks" at α_1, then "two clicks" at α_2. This corresponds to the sequence $(1, 2, 2)$. Similarly, the path in Figure 19.3 turns "one click" at each vertex, and this corresponds to the sequence $(1, 1, 1, 1, 1)$.

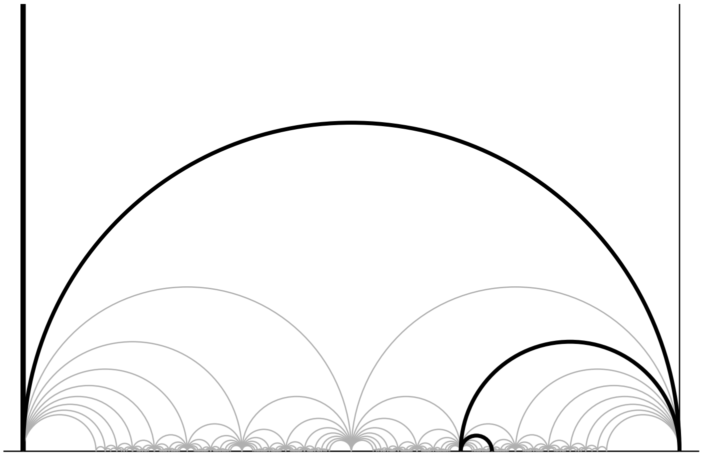

Figure 19.4. The Farey path associated to 5/7

Exercise 9. Prove that the observation about the turns holds for any rational $x_0 \in (0, 1)$.

Third, the diameter of the kth arc in our path is less than $1/k(k-1)$. This is a terrible estimate, but it will serve our purposes below. To understand this estimate, note that the kth arc connects $\alpha_{k-1} = p_{k-1}/q_{k-1}$ to $\alpha_k = p_k/q_k$, and $\alpha_{k-1} \sim \alpha_k$. The diameter of the kth arc is

$$|\alpha_{k-1} - \alpha_k| = \left|\frac{p_{k-1}}{q_{k-1}} - \frac{p_k}{q_k}\right| =^* \frac{1}{q_{k-1}q_k} \leq \frac{1}{k(k-1)}.$$

The starred equation comes from the fact that α_{k-1} and α_k are Farey related. The last inequality comes from Exercise 4. As we mentioned in Exercise 4, the denominators of the α-sequence grow at least exponentially. So, actually, the arcs in our path shrink exponentially fast.

19.6 The Irrational Case

So far, we have concentrated on the case when x_0 is rational. If x_0 is irrational, then we produce an infinite sequence $\{\alpha_k\}$ of rational numbers that approximate x. From what we have said above, we have

$$x \in [\alpha_k, \alpha_{k+1}] \quad \text{or} \quad x \in [\alpha_{k+1}, \alpha_k] \tag{77}$$

for each index k, with the choice depending on the parity of k, and also

$$\lim_{k\to\infty} |\alpha_k - \alpha_{k+1}| = 0. \tag{78}$$

Therefore,

$$x_0 = \lim_{k\to\infty} \alpha_k. \tag{79}$$

The corresponding infinite path in the Farey graph starts at ∞ and zig-zags downward forever, limiting on x.

The nicest possible example is probably

$$x_0 = \frac{\sqrt{5}-1}{2} = 1/\phi,$$

where ϕ is the golden ratio. In this case, $a_k = 1$ for all k and α_k is always ratio of two consecutive Fibonacci numbers. The path in this case starts out as in Figure 19.3 and continues the pattern forever. Taking some liberties with the notation, we can write

$$\frac{1}{\phi} = \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cdots}}}$$

Since $\phi = 1 + (1/\phi)$ we can equally well write

$$\phi = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cdots}}} \tag{80}$$

The $\{a_k\}$ sequence is known as the *continued fraction expansion* of x_0. In case $x_0 > 1$, we pad the sequence with floor(x_0). So, $1/\phi$ has continued fraction expansion $1, 1, 1, \ldots$ and ϕ has continued fraction expansion $1, 1, 1, 1 \ldots$

Exercise 10. Find the continued fraction expansion of $\sqrt{k}$ for $k = 2, 3, 5, 7$.

The subject of continued fractions is a vast one. Here are a few basic facts:

- An irrational number $x_0 \in (0, 1)$ is the root of an integer quadratic equation $ax^2 + bx + c = 0$ if and only if it has a continued fraction expansion that is eventually periodic. [DAV] has a proof.

- The famous number e has continued fraction expansion

$$2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10 \cdots$$

- The continued fraction expansion of π is not known.

In spite of having a huge literature, the subject of continued fractions abounds with unsolved problems. For instance, it is unknown whether the $\{a_k\}$ sequence for the cube root of 2 is unbounded. In fact, this is unknown for any root of an integer polynomial equation that is neither quadratic irrational nor rational.

20 Teichmüller Space and Moduli Space

The purpose of this chapter is to introduce the notions of the Teichmüller space and the moduli space of a closed surface. I will also discuss the mapping class group, which is the group of symmetries of Teichmüller space. The theory of these objects is vast and deep. My purpose is just to introduce the basic objects in an intuitive way. I learned most of the material here (at least in the negative Euler characteristic case) from [THU]. The book [RAT] has a careful treatment from a similar point of view. There are a number of more advanced works devoted entirely to the topics introduced here. For instance, see [GAR] and [FMA].

First we will deal with the case of tori, and then we will deal with negative Euler characteristic case.

20.1 Parallelograms

Say that a *marked parallelogram* is a parallelogram P with a distinguished vertex v, a distinguished first side e_1, and a distinguished second side e_2. The sides e_1 and e_2 should meet at v, as in Figure 20.1. We say that two marked parallelograms P_1 and P_2 are *equivalent* if there is an orientation-preserving similarity, i.e., a translation followed by a dilation followed by a rotation, that maps P_1 to P_2 and preserves all the markings.

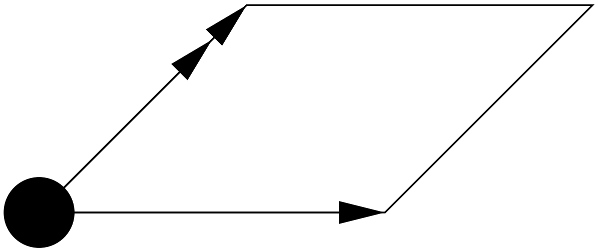

Figure 20.1. A marked parallelogram

We think of P as a subset of $\boldsymbol{C}$. If we have a marked parallelogram, we can translate it so that $v = 0$ and e_1 points from 0 to 1. Then e_2 points from 0 to some $z \in \boldsymbol{C} - \boldsymbol{R}$. We only consider "half" of the possibilities, the case when $z \in \boldsymbol{H}^2$, considered as the upper half plane of $\boldsymbol{C}$.

Exercise 1. Prove that $z(P_1) = z(P_2)$ if and only if P_1 and P_2 are equivalent.

We can also reverse the process. Given $z \in \boldsymbol{H}^2$, we can form a marked parallelogram P such that $z(P) = z$. We simply choose the parallelogram with vertices $(0, 1, z, 1+z)$ and mark it in the obvious way. In short, we can say that there is a natural bijection between the set $\mathcal{T}$ of marked parallelograms and $\boldsymbol{H}^2$, the upper half plane. We can even take this one step further. $\boldsymbol{H}^2$ has its usual hyperbolic metric, and we can transfer this metric onto $\mathcal{T}$. That is, the distance between P_1 and P_2 is defined to be the hyperbolic distance between $z(P_1)$ and $z(P_2)$.

20.2 Flat Tori

In §6.3 we discussed the surface one gets by gluing the opposite sides of a parallelogram. Such a surface is known as a *flat torus*. One thing we did not discuss in §6.3 is the effect of changing the shape of the underlying parallelogram, i.e., the topic of the previous section. When we consider the idea of doing the construction in §6.3 for *all possible parallelograms* we are led to the notions of Teichmüller space and moduli space. So, the idea is to unite the discussion in §6.3 with the discussion in the previous section.

Say that a *flat torus* is a surface T that is locally Euclidean and also homeomorphic to a torus. Recall that the universal cover of T is $\boldsymbol{R}^2$ and the fundamental group of T is $\boldsymbol{Z}^2$.

Definition 20.1. Say that a *marked flat torus* is a flat torus, together with a distinguished pair of elements $\gamma_1, \gamma_2 \in \pi_1(T)$ which generate $\pi_1(T)$. We say that two marked tori T_1 and T_2 are *equivalent* if there is an orientation-preserving similarity that carries T_1 to T_2 and induces a map on the fundamental groups that carries the one distinguished generating set to the other.

Given a marked flat torus T, we can produce a marked parallelogram, as follows: We think of $\pi_1(T, v)$ as the deck group, acting on $\boldsymbol{R}^2$ by translations. We then consider the parallelogram P with vertices

$$0, \qquad \gamma_1(0), \qquad \gamma_2(0), \qquad \gamma_1(0) + \gamma_2(0). \tag{81}$$

The distinguished vertex is 0, and the kth distinguished edge points from 0 to $\gamma_k(0)$. We insist that the marking of P is *positively oriented* So, again, we weed out redundancy by only considering half the possibilities. We can also reverse the process. If we start with a marked parallelogram P, as in Figure

20.1, we can glue the opposite sides of P. The glued-up sides are loops which represent γ_1 and γ_2.

Exercise 2. Prove that two marked tori are equivalent if and only if the corresponding marked parallelograms are equivalent.

Now we will redo the same construction from another point of view. Let Σ denote our favorite flat torus, say, the one obtained by identifying the opposite sides of a square.

Definition 20.2. A *marked flat torus* is a triple (Σ, T, ϕ), where T is a flat torus and $\phi : \Sigma \to T$ is an orientation-preserving homeomorphism. We say that two triples (Σ, T_1, ϕ_1) and (Σ, T_2, ϕ_1) are *equivalent* if there is an orientation-preserving similarity $f : T_1 \to T_2$ such that $f \circ \phi_1$ and ϕ_2 are homotopic maps.

We can convert between the one notion of marked torus and the other. To go from Definition 20.2 to Definition 20.1, we first choose a distinguished pair of elements of $\pi_1(\Sigma)$: We let γ_1 denote the loop that is the glued-up horizontal edge, and we let γ_2 be the loop that is the glued-up vertical edge. Then $\phi^*(\gamma_1)$ and $\phi^*(\gamma_2)$ are the distinguished elements of $\pi_1(T)$. To go from Definition 20.1 to Definition 20.2, we recall that our original notion of a marked torus gives us a description of the torus as a glued-up parallelogram P. We just map the unit square to P by an affine map in such a way that the gluings are respected, and then we interpret this map as a map from Σ to our torus. This produces a triple (Σ, T, ϕ), where ϕ is not just a homeomorphism but actually locally affine.

Exercise 3. Prove that our conversion between the two notions of marked tori respects the equivalence relations. That is, the two notions are really the same notion.

Since all the notions we have discussed are the same, we think of $\mathcal{T}$ as the space of marked tori in the second sense. That is, we work with equivalence classes of triples (Σ, T, ϕ). We have a canonical identification of $\mathcal{T}$ with $\boldsymbol{H}^2$, the hyperbolic plane. With this interpretation, $\mathcal{T}$ is known as the *Teichmüller space* of (marked) flat structures on the torus.

20.3 The Modular Group Again

Now we are going to bring the modular group into the picture. We discussed this group in §19.3 and §19.4. First of all, we can interpret our favorite flat torus Σ as the quotient $\boldsymbol{R}^2/\boldsymbol{Z}^2$. This observation is important in what we do next.

Let $\mathcal{G} = SL_2(\boldsymbol{Z})$ denote the group of integer 2×2 matrices of determinant 1, the modular group. Any $g \in \mathcal{G}$ acts on $\boldsymbol{R}^2$ as an orientation preserving linear transformation, and $g(\boldsymbol{Z}^2) = \boldsymbol{Z}^2$. This means that g induces an orientation-preserving homeomorphism $g : \Sigma \to \Sigma$. We give this homeomorphism the same name as the linear transformation which induces it.

Given a triple (Σ, T, ϕ), we define the new triple $(\Sigma, T, \phi \circ g^{-1})$. That is, we keep the same surface T, but we change $\phi : \Sigma \to T$ to the map given by the composition $\Sigma \to \Sigma \to T$, with the first arrow given by g^{-1}. We use g^{-1} in place of the more obvious choice of g for technical reasons that we will explain momentarily.

Exercise 4. Prove that (Σ, T_1, ϕ_1) and (Σ, T_2, ϕ_2) are equivalent if and only if $(\Sigma, T_1, \phi_1 \circ g)$ and $(\Sigma, T_2, \phi_2 \circ g)$ are equivalent.

The group $\mathcal{G}$ *acts* on the space $\mathcal{T}$ in the sense that

$$g_1(g_2(x)) = (g_1 \circ g_2)(x), \tag{82}$$

for all $g_1, g_2 \in \mathcal{G}$ and all $x \in \mathcal{T}$. Here $g_1 \circ g_2$ means "first do g_2 and then do g_1". To see this, let x be a point represented by the triple (Σ, T, ϕ). We compute

$$\begin{aligned} g_1(g_2(x)) &= g_1(\Sigma, T, \phi \circ g_2^{-1})(\Sigma, T, \phi \circ g_2^{-1} \circ g_1^{-1}) \\ &= (\Sigma, T, \phi \circ (g_1 \circ g_2)^{-1}) = (g_1 \circ g_2)(x). \end{aligned}$$

From this calculation, you can see why we used the inverse: it makes the compositions come out the right way.

We have an explicit identification of $\mathcal{T}$ with $\boldsymbol{H}^2$, and we can see how a particular matrix

$$g = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{83}$$

acts on $\mathcal{T}$ in these coordinates. Let $x = (\Sigma, T, \phi)$ as above. We put T in the best possible position, so that T is obtained by gluing the opposite sides of the parallelogram $(0,1,z,1+z)$, and so that ϕ is induced by the linear transformation carrying $(1,0)$ to $(1,0)$ and $(0,1)$ to (x,y). Here $z=x+iy$. When we lift ϕ to the universal covers of Σ and T, respectively, we get the same linear transformation. In other words, the linear transformation

$$\widehat{\phi}=\begin{bmatrix}1 & x\\ 0 & y\end{bmatrix} \tag{84}$$

induces the homeomorphism ϕ. The linear transformation

$$\widehat{\phi}\circ g^{-1}=\begin{bmatrix}d-cx & -b+ax\\ -cy & ay\end{bmatrix} \tag{85}$$

induces the homeomorphism $\phi\circ g^{-1}$.

To figure out the point $g(x)$ we just have to compute the shape of the marked parallelogram $g(T_0)$, where T_0 is the unit square whose distinguished point is the origin, whose first distinguished edge is $1\equiv(1,0)$, and whose second distinguished edge is $i\equiv(0,1)$. Here we are listing both the coordinates in $\boldsymbol{C}$ and in $\boldsymbol{R}^2$. We compute that $\widehat{\phi}$ maps the first and second edges, respectively, to $cz-d$ and $az-b$. So, if $x\in\mathcal{T}$ corresponds to $z\in\boldsymbol{H}^2$, then $g(z)$ corresponds to

$$\frac{az-b}{cz-d}. \tag{86}$$

Except for the minus signs, this is the usual linear fractional action of g on $\boldsymbol{H}^2$.

20.4 Moduli Space

The quotient $\mathcal{M}=\mathcal{T}/\mathcal{G}$ is known as *moduli space*. To interpret this quotient in terms of tori, we (temporarily) let $\mathcal{M}'$ denote the set of equivalence classes of flat tori. Here we say that two flat tori are equivalent if there is an orientation-preserving similarity carrying one to the other. We are going to construct a natural bijection between $\mathcal{M}$ and $\mathcal{M}'$. Once we have this bijection, we can forget about $\mathcal{M}'$ and simply realize that $\mathcal{M}$ is the space of equivalence classes of flat tori.

There is an obvious map from $\mathcal{T}$ to $\mathcal{M}'$. Given a triple (Σ,T,ϕ), we simply consider the torus T alone. It is a tautology that this map respects the notions of equivalence on both $\mathcal{T}$ and $\mathcal{M}'$. The action of $\mathcal{G}$ on $\mathcal{T}$ has no effect on the underlying torus—only the map is changed—so actually we get a map from $\mathcal{M}$ to $\mathcal{M}'$.

At the same time, there is a map from $\mathcal{M}'$ to $\mathcal{M}$. Given a flat torus T, we arbitrarily choose a homeomorphism $\phi : \Sigma \to T$, and we consider the image of the triple $(\Sigma, T, \phi) \in \mathcal{M}$. To see that this map is independent of choices, consider the two triples (Σ, T, ϕ_1) and (Σ, T, ϕ_2). The map $\phi_2^{-1} \circ \phi_1$ is an orientation-preserving homeomorphism of Σ. We claim that this map is homotopic to some linear homeomorphism of Σ, given by $g \in \mathcal{G}$. Assuming this claim for the moment, we see that $(\Sigma, T, \phi_2) = g(\Sigma, T, \phi_1)$ for some $g \in \mathcal{G}$. That is, both choices lead to the same point in $\mathcal{M}$.

Assuming our claim, we see that there is a natural bijection between $\mathcal{M}$ and $\mathcal{M}'$, so we may identify $\mathcal{M}$ as the space of equivalence classes of flat tori. The space $\mathcal{M}$ is known as *moduli space*.

It is worth pointing out that we just considered $\mathcal{M}'$ as a set, but we might have put a metric on $\mathcal{M}'$ in some way. Any reasonable choice would make our bijection between $\mathcal{M}$ and $\mathcal{M}'$ a homeomorphism. We will refrain from doing this here, because below we will actually do it for surfaces of negative Euler characteristic.

Our only piece of unfinished business is to show that $\gamma = \phi_2^{-1} \circ \phi_1$ is homotopic to the action of some $g \in SL_2(\boldsymbol{Z})$. Note that γ acts on the fundamental group $\pi_1(\Sigma)$. Since γ is an orientation-preserving homeomorphism, γ has the same action on $\pi_1(\Sigma)$ as does some $g \in SL_2(\boldsymbol{Z})$. So, replacing γ by $\gamma \circ g^{-1}$, we can assume that γ acts as the identity on $\pi_1(\Sigma)$. Our task now is to show that γ is homotopic to the identity map.

A formal proof of this fact is a bit tedious, but we will sketch the idea. Let e_1 and e_2 be the usual horizontal and vertical loops on Σ. Since $\gamma(e_1)$ is homotopic to e_1, we first adjust γ so that it is the identity on e_1. Next, we adjust γ so that it is the identity on $e_1 \cup e_2$. But now we can cut Σ open and interpret γ as a continuous map from the unit square to itself which is the identity on the boundary. The following exercise finishes the proof.

Exercise 5. Prove that a continuous map from the unit square to itself, which is the identity on the boundary, is homotopic to the identity map.

There is an important lesson to take away from this section. The space $\mathcal{M}$ has a more direct and simple definition than the space $\mathcal{T}$. However, it was useful to define $\mathcal{T}$ first and then realize $\mathcal{M}$ as a quotient of $\mathcal{T}$. We hope that this lesson motivates the definition of the Teichmüller space of surfaces having a fixed negative Euler characteristic.

20.5 Teichmüller Space

We would like to make all the same constructions that we made above in the negative Euler characteristic case, but there is one fine point we want to iron out. Above, we considered flat tori and similarities between them, and below we will consider hyperbolic surfaces and isometries between. Given any flat torus, we can always rescale the metric so that it has unit area. If we only work with unit area tori, then the natural maps between them are (orientation-preserving) isometries. The point here is that an area-preserving and orientation preserving similarity is an isometry. So, we might have redone the whole theory above using unit area tori and isometries. This point of view is more natural in the negative Euler characteristic case, because two hyperbolic surfaces with the same topology always have the same area; see Theorem 12.4.

Now we are ready to go. We will fix a number $g \geq 2$, the genus of the surfaces we consider. Recall that the genus g of a surface S satisfies the equation

$$\chi(S) = 2 - 2g. \tag{87}$$

Here χ is the Euler characteristic, as discussed in §3.4. Thus, a torus has genus 1, and the octagon surface has genus 2. In general, a genus g surface is a "g-holed torus" that is locally isometric to $\boldsymbol{H}^2$, the hyperbolic plane. We are going to build $\mathcal{T}_g$, the Teichmüller space of genus g hyperbolic surfaces.

We first fix our favorite surface of genus g, and call it Σ. Unlike in the torus case, a "favorite" does not immediately jump out. My personal favorite is the one obtained by gluing together the opposite sides of a regular hyperbolic $4g$-gon. In any case, we look at triples of the form (Σ, M, ϕ), where M is a hyperbolic surface of genus g and $\phi : \Sigma \to M$ is a homeomorphism. We say that two triples (Σ, M_1, ϕ_1) and (Σ, M_2, ϕ_2) are *equivalent* if there is an isometry $f : M_1 \to M_2$ such that $f \circ \phi_1$ and ϕ_2 are homotopic maps.

When we worked out the case of the torus, we had a natural way of putting coordinates on the space $\mathcal{T}$. The point is that $\mathcal{T}$ is really just $\boldsymbol{H}^2$ in disguise. This time, it is not so obvious what to do. So, we will first make $\mathcal{T}_g$ into a metric space. The idea behind the next definition is to make sense of surfaces being nearby to each other without quite being the same. We say that a homeomorphism $f : M_1 \to M_2$ is a $(1+\epsilon)$-isometry if the inequality

$$1 - \epsilon \leq \frac{d_2(x_2, y_2)}{d_1(x_1, y_2)} \leq 1 + \epsilon \tag{88}$$

holds for all quadruples x_1, y_1, x_2, y_2 with $x_1, y_1 \in M_1$ and $x_2 = f(x_1)$ and $y_2 = f(y_1)$. The functions d_1 and d_2 are the metrics on M_1 and M_2, respectively.

Define the distance between 2 triples (Σ, M_1, ϕ_1) and (Σ, M_2, ϕ_2) to be the infimal ϵ with the property that there is a $(1+\epsilon)$ isometry $f : M_1 \to M_2$ such that $f \circ \phi_1$ and ϕ_2 are homotopic maps.

Exercise 6. Prove that the equivalence relation we have defined respects the distance we have defined. Hence, the distance between two equivalence classes makes sense. This is how we make $\mathcal{T}_g$ into a metric space.

In the case of the torus, there was a perfectly canonical metric on $\mathcal{T}$, namely the hyperbolic metric. In the higher genus case, the metric we have defined is pretty good but not perfectly canonical. There are a number of canonical metrics on $\mathcal{T}_g$. The two most commonly used are the Teichmüller metric and the Weil–Petersson metric. One vexing thing is that these two common metrics are pretty different from each other. So, while there are several nice ways to view $\mathcal{T}_g$, there does not seem to be one best way. What is best depends upon the context.

20.6 The Mapping Class Group

When we worked with our favorite flat torus Σ, the one based on the unit square, we saw that $SL_2(\boldsymbol{Z})$ arose naturally as the group of locally linear and orientation-preserving homeomorphisms of Σ. For a hyperbolic surface, it is not immediately obvious that there is a similarly natural group of homeomorphisms. However, it turns out that there is such a group.

Above, we sketched a proof that any orientation-preserving homeomorphism of the flat square torus Σ is homotopic to the action of an element of $SL_2(\boldsymbol{Z})$. In fact, we can equally well say that $SL_2(\boldsymbol{Z})$ is the quotient of the group of orientation-preserving similarities of Σ, modulo homotopy. That is, two such homeomorphisms are equivalent, and considered the same, if they are homotopic. This is a definition that carries over immediately to the higher genus case.

We fix some initial hyperbolic surface Σ of genus g. The *mapping class group* is defined as the group of equivalence classes of homeomorphisms of Σ, where two homeomorphisms are equivalent if they are homotopic. The group is often denoted $\mathcal{MCG}_g$. It is a kind of generalization of $SL_2(\boldsymbol{Z})$. The definition of the mapping class group depends only mildly on the choice of Σ. Any other choice would lead to an isomorphic group.

Exercise 7. The mapping class group is certainly well defined as a set. Prove that it is well defined as a group, that is, the group law respects the equivalence classes.

People have focused quite a bit of attention on the mapping class group in recent years. There are many open problems about this group. One of the most well-known open problems is as follows. For each genus g, does there exist some $n = n_g$ and a one-to-one homomorphism $\phi : \mathcal{MCG}_g \to GL_n(\boldsymbol{C})$? Here $GL_n(\boldsymbol{C})$ is the group of complex valued $n \times n$ matrices having nonzero determinant. Or, more briefly, *is the mapping class group linear?*. Subgroups of $GL_n(\boldsymbol{C})$ are called linear.

The group $\mathcal{MCG}_g$ acts on $\mathcal{T}_g$. The homeomorphism $g : \Sigma \to \Sigma$ acts on the triple (M, Σ, ϕ) by sending it to the triple $(M, \Sigma, \phi \circ g^{-1})$. From the way we have defined the mapping class group, this action respects the equivalence relation used to define $\mathcal{T}_g$.

Exercise 8. Recall that we defined a metric on $\mathcal{T}_g$ above. Prove that each element of $\mathcal{MCG}_g$ acts as an isometry on $\mathcal{T}_g$.

Now that we have defined $\mathcal{MCG}_g$ and $\mathcal{T}_g$, we define $\mathcal{M}_g$ to be the quotient space.

$$\mathcal{M}_g = \mathcal{T}_g / \mathcal{MCG}_g. \tag{89}$$

The space $\mathcal{M}_g$ is known as the *moduli space* of genus g hyperbolic surfaces.

As in the torus case, we could take the alternate route and define $\mathcal{M}_g$ as the set of hyperbolic surfaces of genus g equipped with a metric like the one defined above for $\mathcal{T}_g$. That is, the distance between two hyperbolic surfaces is the infimal ϵ such that there is a $(1+\epsilon)$ isometry between them.

21 Topology of Teichmüller Space

We defined the Teichmüller space $\mathcal{T}_g$ in the previous chapter. In the case of the torus, the Teichmüller space is just a copy of $\boldsymbol{H}^2$; in particular, it is homeomorphic to $\boldsymbol{R}^2$. Here will sketch a well-known proof, mainly through a series of exercises, that $\mathcal{T}_g$ is homeomorphic to $\boldsymbol{R}^{6g-6}$.

This beautiful result sets up a picture that is in some ways very similar to the one for the torus. We have the mapping class group $\mathcal{MCG}_g$ acting isometrically on a space that is homeomorphic to $\boldsymbol{R}^{6g-g}$ (but having a funny metric on it). All the complexity in the topology of the moduli space $\mathcal{M}_g$ comes from the operation of taking the quotient of a topologically trivial space by the action of a complicated group. This is exactly what happens for the torus, except that the space $\mathcal{M}$ and the group $SL_2(\boldsymbol{Z})$ are not so complicated.

21.1 Pairs of Pants

A *pair of pants* is a hyperbolic surface-with-boundary that is obtained by taking two identical copies of a right-angled hexagon and gluing 3 of the sides. Figures 21.1 and 21.2 show this. These kinds of gluings were considered in detail in Chapter 12.

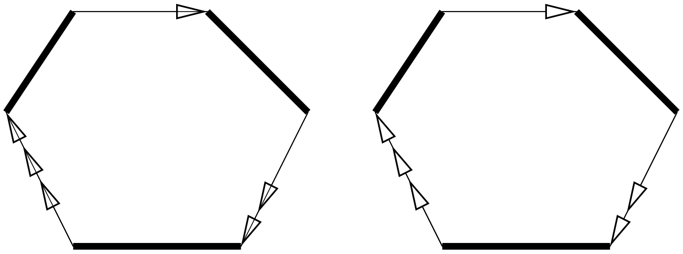

Figure 21.1. Hexagon gluing

Exercise 1. Let l_1, l_2, l_3 be three positive numbers. Prove that there is a right-angled hexagon whose "odd" sides have lengths l_1, l_2, l_3. Prove that this hexagon is unique up to an isometry.

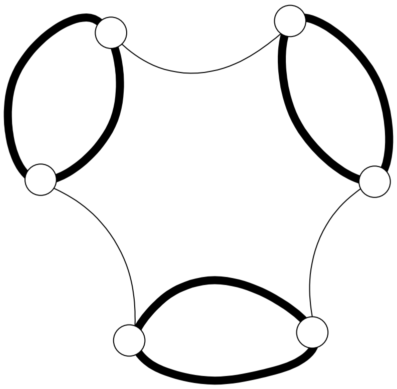

Figure 21.2. A pair of pants

A pair of pants is homeomorphic to a sphere with 3 holes. The boundaries are *totally geodesic* in the sense that every point on the boundary has a neighborhood that is isometric to a half-disk in $\boldsymbol{H}^2$. By *half-disk*, we mean the portion of a disk that lies to one side of a diameter.

Exercise 2. Suppose that M is a surface with boundary whose interior is locally isometric to $\boldsymbol{H}^2$ and whose boundary is totally geodesic. Suppose also that M is homeomorphic to a sphere with 3 holes. Prove that M is a pair of pants, in the sense that M can be built by gluing together 2 hexagons in the manner described above. (*Hint*: Consider the 3 geodesics arcs obtained by connecting the 3 boundary components, in pairwise fashion, by shortest curves. Cut M open along these curves and watch M fall apart into two right-angled hexagons. Use Exercise 1 to show that these hexagons are copies of each other.)

Exercise 3. Let l_1, l_2, l_3 be three positive real numbers. Prove that there exists a unique pair of pants, up to isometry, whose boundaries have lengths l_1, l_2, l_3, respectively.

To avoid confusion, we will shorten *pair of pants* to *pant*.

21.2 Pants Decompositions

We say that a *pants decomposition* of a hyperbolic surface is a realization of that surface as a finite union of pants, glued together along their boundaries.

In this section we will prove that every hyperbolic surface has a pants decomposition. Actually, we will prove that any hyperbolic surface has many such.

Suppose that M is a hyperbolic surface and γ is a closed loop on M. Recall that $\boldsymbol{H}^2$ is the universal cover of M and that $M = \boldsymbol{H}^2/\Gamma$, where Γ is the deck group. Let $\widetilde{\gamma}$ denote a lift of γ to $\boldsymbol{H}^2$. Since γ is a closed loop, there must be some nontrivial element $g \in \Gamma$ such that $g(\gamma) = \gamma$. According to the classification of isometries of the hyperbolic plane, g is either elliptic, hyperbolic, or parabolic. Since M is a compact surface, there is some $\epsilon > 0$ such that every ϵ ball on M is embedded. This means that g moves every point of $\boldsymbol{H}^2$ by at least ϵ. But this means that g is hyperbolic; see §10.9. In particular, g has two fixed points on ∂H^2.

From this picture, we see that $\widehat{\gamma}$ has two accumulation points on ∂H^2, namely the fixed points of g. There is a unique geodesic $\widetilde{\beta}$ connecting the two fixed points of g. This geodesic is the axis of g. The quotient $\beta = \widetilde{\beta}/g$ is called the *geodesic representative* of γ. Intuitively, if we think of γ as a rubber band that has been perhaps stretched out of its natural position, then β represents the curve assumed when the rubber band snaps back into position. The following exercise justifies this point of view.

Exercise 4. Prove that γ and β are homotopic.

The curve γ is called *simple* if γ has no self-intersections. If γ is simple, then no two lifts of γ to the universal cover intersect. If β the geodesic representative is not simple, then two lifts $\widetilde{\beta}_1$ and $\widetilde{\beta}_2$ in $\boldsymbol{H}^2$ cross each other. But then the endpoints of $\widetilde{\beta}_1$ separate the endpoints of $\widetilde{\beta}_2$ on ∂H^2. But then we can find corresponding lifts $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ of γ whose (ideal) endpoints have the same property. This forces these lifts to cross, which means that γ does have a self-intersection. Figure 21.3 shows the situation.

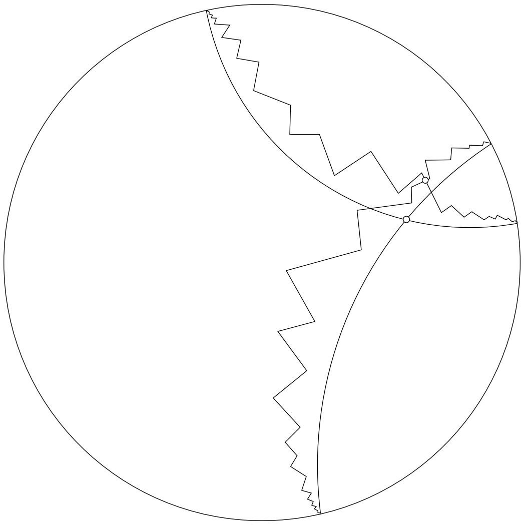

Figure 21.3. Intersecting curves

So, if γ is simple, then so is β. A similar argument proves the following generalization: if $\{\gamma_i\}$ is a finite list of pairwise disjoint simple loops, then so is $\{\beta_i\}$, the list of geodesic representatives.

The process of replacing a simple loop by its geodesic representative is a magical one. You might imagine that the geodesic replacements could somehow crash into themselves or to each other, even though the original curves do not. But, as we just explained, this is not the case. So, if we want to find a pants decomposition of a hyperbolic surface M, all we have to do is find simple loops that divide M into 3-holed spheres. Then we replace all these loops with their geodesic representatives, and we have our pants decomposition.

Now we can give the intuitive idea behind the main result of this chapter. We start by choosing our favorite pants decomposition of our favorite surface Σ. Let $S \subset \Sigma$ denote the set of curves of this decomposition. By definition $\Sigma - S$ is a union of 3-holed spheres. A point of $\mathcal{T}_g$ is an equivalence class of triple (Σ, M, ϕ). We define a pants decomposition on M by taking the geodesic representatives of the curves in the set $\phi(S) \subset M$.

Exercise 5. Prove that there will be exactly $2g - 2$ pants in the decomposition, with $3g - 3$ boundary curves.

We get $3g - 3$ real numbers by considering the lengths of the boundary curves in the pants decomposition. Each curve is contained in 2 pants. We get an additional $3g-3$ numbers by considering how two pants are glued together along their common boundary. These other $3g-3$ numbers are usually called the *twist parameters*. Our construction respects the equivalence relation on triples and gives a well-defined map. We from $\mathcal{T}_g$ to $\boldsymbol{R}^{6g-6}$. will see that this map is a homeomorphism.

21.3 Special Maps and Triples

In this section we prepare for our main construction. First, we choose our favorite kind of pant. This is the pant obtained by gluing together two identical regular hexagons. Each pant, including our favorite, has two special points on each boundary component. These are the points which are images of hexagon vertices; see Figure 21.2 above.

We (re)choose our favorite surface Σ so that it made from gluing together $3g-3$ of our favorite kind of pants. We insist that the pants are glued in such a way that the marked points are matched up. This still doesn't determine the surface exactly, but we just pick one from amongst the various possibilities.

Let $P(l_1, l_2, l_3)$ denote the pant having boundary lengths l_1, l_2, l_3. We choose some reasonable homeomorphism from our favorite pant P_0 to $P(l_1, l_2, l_3)$. The best way to to this is to choose a homeomorphism from the regular right-angled hexagon to a hexagon with side lengths $l_k/2$ that sends vertices to vertices, and then to double this map, so to speak.

More generally, given a 6-tuple $(l_1, l_2, l_3, \theta_1, \theta_2, \theta_3)$, we choose a map from P_0 to $P(l_1, l_2, l_3)$ which agrees with our original map above, but which makes a twist of $\theta_k/2$ (say) clockwise in a small neighborhood of the kth boundary component. This new map is obtained from the original one by giving a kind of twist. The new map is very similar to the Dehn twists we discussed in §18.7. We call such a map *special* and denote it by $\mu(l_1, l_2, l_3, \theta_1, \theta_2, \theta_3)$.

Given a triple (Σ, M, ϕ), we get a pants decomposition of M, as described above. We call the triple *special* if the following hold:

- The restriction of ϕ to each pant of Σ is one of our special maps.

- If the restriction of ϕ to some pant P is the map

$$\mu(l_1, l_2, l_3, \theta_1, \theta_2, \theta_3),$$

then the restriction of ϕ to the pant P' that meets P along the kth boundary of P is $\mu(\dots l_k \dots, \dots, \theta_k, \dots)$. The other 4 coordinates can be different.

Here we sketch the proof that any triple is equivalent to a special triple. We warn the reader that a formal construction is filled with many details that we are not including. We hope that this sketch is sufficient for the interested reader to give a careful proof. [RAT] has all the details.

Lemma 21.1 *Any triple is equivalent to a special triple.*

Proof (sketch). Starting with the triple (Σ, M, ϕ_1), we make a homotopy between ϕ_1 and a homeomorphism $\phi_2 : \Sigma \to M$ that maps the set S of geodesics to the set of geodesics on the natural pants decomposition of M. Next, we make a homotopy between ϕ_2 and a map $\phi_3 : \Sigma \to M$ that agrees with a special map outside a small neighborhood of S, the only place where we do not have control over ϕ_2. This map agrees with the special map on the outside of $3g - 3$ small annuli.

Each annulus on Σ is divided in half. The two halves of an annulus are subsets of the two different pants that glue together along a common boundary component. The center curve of the annulus is the common boundary component. We have a foliation of each annulus by circles that are, in a sense, parallel to the center curve. We have a similar picture for the corresponding annulus on M. We just adjust ϕ_3 so that it twists these circles "at a constant rate", evenly dividing the total twist between the two halves, so to speak. The final map ϕ_4 is homotopic to the original one, and is special on each pant. Thus (Σ, M, ϕ_4) is equivalent to the original triple (Σ, M, ϕ_1). ♠

21.4 The End of the Proof

Now we construct our map from $\mathcal{T}_g$ to $\boldsymbol{R}^{6g-6}$. We start with a triple (Σ, M, ϕ) representing a point in $\mathcal{T}_g$. By Lemma 21.1, it suffices to consider a special triple. However, for special triples, we can assign a pair (l, θ) to each geodesic in the set $S \subset \Sigma$ of pants boundaries. This gives us the map from $\mathcal{T}_g$ to $\boldsymbol{R}^{6g-6}$. Call this map Φ.

The map Φ is surjective, essentially thanks to Exercise 3. We can build pants with any boundary lengths we like, and then we can glue them together with as much twisting as we like. The map Φ is injective because the coordinates on S give us complete instructions for how to assemble the surface M and the map $\phi : \Sigma \to M$. Hence Φ is a bijection between $\mathcal{T}_g$ and $\boldsymbol{R}^{6g-6}$.

The map Φ^{-1} is continuous. If we have special maps corresponding to nearly identical parameters, the corresponding pants are nearly isometric to each other, and the twisting is nearly the same. This allows us to build a map between the two surfaces that is nearly an isometry and in the correct homotopy class.

Showing that Φ is continuous is the most tedious part of the argument. Here we explain the proof. Let (M, Σ_j, ϕ_j) for $j = 1, 2$ be two very nearby special triples. Suppose that $f : M_1 \to M_2$ is a $(1 + \epsilon)$-isometry.

If ϵ is small, the map f carries each pant on M_1 to a 3-holed sphere on M_2 whose boundaries are very nearly geodesics. The process of replacing each near geodesic by an actual geodesic shortens the curve. Hence, the pants boundaries on M_2 are at most $(1 + \epsilon)$ times as long as their counterparts on M_1 and vice versa. Hence, the length parameters $\{l_k\}$ labelling each curve in S are about the same for each triple.

Exercise 6. Let P_1 be a pant on M_1, and let P_2 be the corresponding pant on M_2. Prove that $f(\partial P_1)$ is contained in an ϵ'-neighborhood of P_2, where $\epsilon' \to 0$ with ϵ. (*Hint*: Lift the picture to the universal cover, and show that a curve that nearly minimizes length in $\boldsymbol{H}^2$ must be close to an actual geodesic.)

Exercise 7. Each pant on M_1 decomposes into two identical right-angled hexagons. Let H be such a hexagon. Prove that $f(H)$ is within ϵ' of one of the corresponding hexagons on M_2. Here $\epsilon' \to 0$ as $\epsilon \to 0$. (*Hint*: By Exercise 5, f maps the pant boundaries to curves that are very close to their geodesic representatives. This takes care of half the sides of H. Next, f maps each other side s of H to a curve that nearly realizes the distance between two components of a pant on M_2. Prove that this forces $f(s)$ to be near the true minimizer.)

Each pant boundary β on M_1 has 4 special points. Two of these points come from one pant incident to β and the other two come from the other pant incident to β. We call these collections of points *special quads*.

Exercise 8. Let Q denote one of the special quads. Prove $f(Q)$ is within ϵ'

of the corresponding special quad on M_2. Here $\epsilon' \to 0$ as $\epsilon \to 0$.

For M_1, the parameter $\theta/(2\pi)$ on each curve of S can be deduced from the distances between the points of the relevant special quads. We now conclude from Exercise 8 that the θ parameters for M_1 are close mod 2π to the corresponding θ parameters for M_2.

Finally, if the θ parameter for M_1 differs by nearly an integer from the corresponding θ parameter for M_2, then we can find an γ on M_1, crossing over the image of the offending boundary component, such that $f(\gamma)$ twists more times around and is considerably longer, as shown in Figure 21.4. Figure 21.4 shows the case when $\theta = 0$ for M_1 but $\theta = 1$ for M_2.

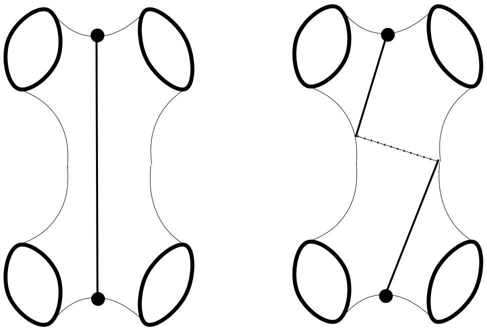

Figure 21.4. Integer twist

All in all, we have shown (modulo some details) that the map Φ is a homeomorphism from $\mathcal{T}_g$ to $\boldsymbol{R}^{6g-6}$.

22 The Banach–Tarski Theorem

The purpose of this chapter is to prove the Banach–Tarski Theorem. My account is somewhat similar to the one given in [WAG]. At first, the Banach–Tarski Theorem does not look too much like a result about surfaces, but in some sense it is a result about rotations of the sphere. The proof I give also brings in the modular group in an essential way.

22.1 The Result

We say that $A, B \subset \boldsymbol{R}^3$ are *equivalent* if there are finite partitions into disjoint pieces,

$$A = A_1 \cup \cdots \cup A_n, \qquad B = B_1 \cup \cdots \cup B_n,$$

and isometries $I_1, \ldots, I_n$ such that $I_j(A_j) = B_j$ for all j. In this case, we write $A \sim B$. When $A \sim B$ it means, informally, that one can cut A into n pieces, like a puzzle, and reassemble those pieces into B. The implied map $A \to B$ is, by definition, a *piecewise isometric map*.

Exercise 1. Prove that $\sim$ is an equivalence relation.

The Banach–Tarski Theorem requires the Axiom of Choice. See [DEV] for a discussion of this axiom. Here is the precise version that is needed.

Real Axiom of Choice (RAC). Let $\{X_\alpha\}$ be a disjoint union of subsets of $\boldsymbol{R}^3$. Then there exists a set $S \subset \bigcup X_\alpha$ such that S contains exactly one element of X_α for each α.

The RAC might seem obvious, or at least harmless. What the Banach–Tarski Theorem shows is that a highly counterintuitive result comes as a consequence of assuming that the RAC is true.

Say that A is a *good set* if A is bounded and A contains a ball.

Theorem 22.1 (Banach–Tarski) *Assume the RAC If A and B are arbitrary good sets, then $A \sim B$.*

In light of the fact that $\sim$ is an equivalence relation, it suffices to prove the Banach–Tarski Theorem in the case that A is a ball of radius 1.

What makes this theorem amazing is that A could be a tiny ball and B could be an enormous ball. At first you might think that this result contradicts such physical properties as "conservation of mass". The usual reply to this objection is that the pieces needed to make the puzzle are so complicated that they do not have mass. Another reply is that we are not talking about physical objects that are made out of atoms.

22.2 The Schroeder–Bernstein Theorem

The Schroeder–Bernstein Theorem says the following. If there are injective maps from A into B and from B into A, then there is a bijection between A and B. This result works for any sets and any maps. (Any book on set theory has this result, but you can extract the general proof from the proof of the next result.)

In case A and B are subsets of $\boldsymbol{R}^3$ and the injections are piecewise isometric maps, then the bijection manufactured by the proof is also piecewise isometric. To formalize this situation, we write $A \prec B$ if $A \sim B'$ for some subset $B' \subset B$. This is another way of saying that there is a piecewise isometric injection from A into B.

Lemma 22.2 *If $A \prec B$ and $B \prec A$, then $A \sim B$.*

Proof: We have injective and piecewise isometric maps $f : A \to B$ and $g : B \to A$. Say that an n-chain is a sequence of the form $x_n \to \cdots \to x_0 \in A$, where

- $x_k \in A$ if $k > 0$ is even. In this case $f(x_k) = x_{k-1}$
- $x_k \in B$ if k is odd. In this case $g(x_k) = x_{k-1}$.

For each $a \in A$, let $n(a)$ denote the length of the longest n-chain that ends in $a = x_0$. It might be that $n(a) = \infty$. Let $A_n = \{a \in A |\ n(a) = n\}$. Swapping the roles of A and B, define B_n similarly.

Now observe the following:

- $f(A_k) = B_{k+1}$ for $k = 0, 2, 4, \ldots$.
- $g^{-1}(A_k) = B_{k-1}$ for $k = 1, 3, 5$.
- $f(A_\infty) = B_\infty$.

The restriction of f to

$$A' = A_0 \cup A_2 \cup \cdots \cup A_\infty$$

is an injective piecewise isometry and the restriction of g^{-1} to

$$A'' = A - A' = A_1 \cup A_3 \cup A_5 ...$$

is also an injective piecewise isometry. (Note that A'' does not include A_∞.) Define $h(a) = f(a)$ if $a \in A'$ and $h(a) = g^{-1}(a)$ if $a \in A''$. By construction

$$f(A') \cap g^{-1}(A'') = \emptyset.$$

Hence h is an injection. Also,

$$B = f(A') \cup g^{-1}(A'').$$

Hence h is a surjection. Hence h is a bijection. By construction h is a piecewise isometric map. ♠

22.3 The Doubling Theorem

We write $B \succ A$ if there is a partition $B = B_1 \cup \ldots \cup B_n$ and isometries $I_1, \ldots, I_n$ such that $A \subset \bigcup I_j(B_j)$. In other words, we can break B into finitely many pieces and use these pieces to cover A. The sets $I_1(B_1), \ldots, I_n(B_n)$ need not be disjoint from each other.

Here is a result that sounds simpler (but not really much less surprising) than the Banach–Tarski Theorem.

Theorem 22.3 (Doubling) *Assume the RAC Then there exist* 3 *disjoint unit balls A, B_1, B_2 such that $A \succ B$, where $B = B_1 \cup B_2$.*

Now we will reduce the Banach–Tarski Theorem to the Doubling Theorem.

Lemma 22.4 *If $B \succ A$, then $A \prec B$.*

Proof: Assume $B \succ A$. Define

- $A_1 = A \cap I_1(B_1)$.
- $A_2 = A \cap I_2(B_2) - A_1$.
- $A_3 = A \cap I_3(B_3) - A_1 - A_2$, etc.

Then $A = A_1 \cup \cdots \cup A_n$ is a partition. Let $B'_j = I_j^{-1} A_j$ and let $B' = \bigcup B'_j$. Then $B'_1 \cup \cdots \cup B'_n$ is a partition of B'. By construction $A \sim B' \subset B$. Hence $A \prec B$. ♠

Let B_r denote the unit ball of radius r centered at the origin.

We will assume the RAC and the Doubling Theorem. We claim that $B_r \sim B_s$ for any $r, s > 0$. By scaling, we can assume that $1 = r < s$. Clearly, $B_1 \prec B_s$. In light of the lemmas in the previous section, it suffices to prove that $B_1 \succ B_s$. There is some n such that B_s can be covered by 2^n translates of B_1. Iterating the Doubling Theorem n times, we see that B_1 is equivalent to 2^n disjoint translates of B_1. But then $B_1 \succ B_s$. This proves what we want. Now we know that $B_r \sim B_s$ for all $r, s > 0$.

We have already mentioned that it suffices to prove the Banach–Tarski Theorem when $B = B_1$, the unit ball. But $B_r \subset A \subset B_s$ for some pair of balls B_r and B_s. Since $B_r \sim B_s$ and $A \subset B_s$, we have $B_r \succ A$. This implies that $A \prec B_r$. But $B_r \prec A$. Hence $A \sim B_r$. But $B_r \sim B_1$. Hence $A \sim B_1$. This finishes the reduction of the Banach–Tarski Theorem to the Doubling Theorem.

22.4 Depleted Balls

We are left to prove the Doubling Theorem. The Doubling Theorem is about as simple as can be, but unfortunately some technical complications arise when we try to prove the Doubling Theorem directly. The way around these complications is to prove a related result instead.

Say that a *depleted ball* is a set of the form $B - X$, where B is a unit ball and X is a countable union of lines through the center of B.

Exercise 2. Prove that any unit ball can be covered by 3 isometric images of any depleted ball.

Theorem 22.5 (Depleted Ball) *Assume the RAC Then there exists a depleted ball Σ and a partition $\Sigma = \Sigma_1 \cup \Sigma_2 \cup \Sigma_3$ such that the following hold:*

- Σ_i *and* Σ_j *are isometric for all pairs* i, j.
- $\Sigma_3 \succ \Sigma_1 \cup \Sigma_2$.

Lemma 22.6 *Assume the RAC Then there are* 9 *disjoint depleted balls* A, $B_1, \ldots, B_8$ *such that* $A \sim B$ *where* $B = B_1 \cup \cdots \cup B_8$.

Proof: Iterating the conclusion of the Depleted Ball Theorem, we see that $\Sigma_1 \succ Y$, where Y is any finite union of isometric copies of Σ_1. Our lemma follows almost immediate from this. ♠

Exercise 3. Deduce the Doubling Theorem from the last lemma.

To finish the proof of the Banach–Tarski Theorem, we just have to prove the Depleted Ball Theorem.

22.5 The Depleted Ball Theorem

Proving the Depleted Ball Theorem is the most interesting part of the proof of the Banach–Tarski Theorem. The rest is really just "window dressing". This is the part of the proof that brings in the modular group.

Consider the countable group

$$\Gamma = \langle A, B | A^3 = B^2 = \text{identity} \rangle.$$

In other words, Γ is the group of all words in A and B subject to the relations that A^3 and B^2 are the identity word.

Exercise 4. Prove that Γ is isomorphic, as a group, to the modular group discussed in §19.3 and §19.4. (*Hint*: Put together Exercises 7 and 8 in §19.4.)

We have a partition $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$, where

- Γ_1 consists of those words starting with A.
- Γ_2 consists of those words starting with A^2.
- Γ_3 consists of the empty word and also those words starting with B.

We have the following structure:

$$A\Gamma_k = \Gamma_{k+1}, \qquad \Gamma_1 \cup \Gamma_2 \subset B\Gamma_3.$$

Indices are taken mod 3 for the first equation. These two algebraic facts are quite close to the conclusion of the Depleted Ball Theorem. The trick is to convert the algebra into geometry. Let B denote the unit ball in $\boldsymbol{R}^3$, and let $SO(3)$ denote the group of rotations of B. Below we will prove the following technical lemma.

Lemma 22.7 *There exists an injective homomorphism $\rho : \Gamma \to SO(3)$.*

We choose our injective homomorphism, and we identify A and B with their images under ρ. So, A is an order 3 rotation of B and B is an order 2 rotation of B. In general, we identify elements of Γ with their images under ρ.

A nontrivial element of $SO(3)$ is a rotation about some line through the origin. We say that a line in $\boldsymbol{R}^3$ is *bad* if it is the line fixed by some element of Γ. Since Γ is a countable group, there are only countably many bad lines. Let X denote the union of these bad lines, and let $\Sigma = B - X$. Then Σ is a depleted ball. Moreover, the group Γ acts *freely* on Σ in the following sense. If $\gamma(p) = p$ for some $\gamma \in \Gamma$ and some $p \in \Sigma$, then γ is the identity element.

We have an equivalence relation on Σ. We write $p_1 \sim p_2$ if and only if $p_1 = \gamma(p_2)$ for some $\gamma \in \Gamma$. The fact that Γ is a group implies easily that $\sim$ is an equivalence relation. This gives us an uncountable partition

$$\Sigma = \bigcup \Sigma_\alpha$$

into the equivalence classes. By the RAC, we can find a new set $S \subset \Sigma$ such that S has one member in common with each S_α.

Lemma 22.8 *Let $\gamma_1, \gamma_2 \in \Gamma$ be distinct elements. Then $\gamma_1(S) \cap \gamma_2(S) = \emptyset$.*

Proof: We argue by contradiction. Suppose that $p \in \gamma_1(S) \cap \gamma_2(S)$. We have $\gamma_j^{-1}(p) \in S$ for $j = 1, 2$. But $\gamma_1^{-1}(p)$ and $\gamma_2^{-1}(p)$ are in the same Γ orbit. Since S intersects each Γ orbit exactly once, we have $\gamma_1^{-1}(p) = \gamma_2^{-1}(p)$. But then $\gamma_2\gamma_1^{-1}(p) = p$. Since Γ acts freely on Σ, we have $\gamma_2\gamma_1^{-1} =$ identity. Hence $\gamma_1 = \gamma_2$. This is a contradiction. ♠

Lemma 22.9

$$\Sigma = \bigcup_{\gamma \in \Gamma} \gamma(S).$$

Proof: Choose $p \in \Sigma$. By construction, there is some $q \in S$ such that $p \sim q$. This means that $p = \gamma(q)$ for some $\gamma \in \Gamma$. Hence $p \in \gamma(S)$. ♠

Now define

$$\Sigma_k = \Gamma_k(S) := \bigcup_{\gamma \in \Gamma_k} (S).$$

The previous two results show that $\Sigma = \Sigma_1 \cup \Sigma_2 \cup \Sigma_3$ is a partition of Σ. At the same time

$$A(\Sigma_k) = \Sigma_{k+1}, \qquad B(\Sigma_3) = B\Gamma_3(S) \supset (\Gamma_1 \cup \Gamma_2)(S) = \Sigma_1 \cup \Sigma_2.$$

The first part of this equation shows that Σ_i and Σ_j are isometric for all i, j. The second part shows that $\Sigma_1 \cup \Sigma_2$ is isometric to a subset of Σ_3. Hence $\Sigma_3 \succ (\Sigma_1 \cup \Sigma_2)$. This proves the Depleted Ball Theorem.

22.6 The Injective Homomorphism

The last piece of unfinished business is to produce an injective homomorphism $\rho : \Gamma \to SO(3)$. Let $\phi : S^2 \to \boldsymbol{C} \cup \infty$ be stereographic projection, as in §9.5 and §14.3. We say that two points $z, w \in \boldsymbol{C} \cup \infty$ are *partner points* if

$$w = -1/\overline{z}. \tag{90}$$

In particular, 0 and ∞ are partner points.

Exercise 5. Prove that ϕ maps antipodal points on S^2 to partner points.

Exercise 6. Let T_1 and T_2 be two linear fractional transformations, both of which fix two distinct points $z, w \in \boldsymbol{C}$. Suppose also that the differentials dT_1 and dT_2 are the same map at z. Prove that $T_1 = T_2$.

Lemma 22.10 *Suppose that α is an order 3 linear fractional transformation that fixes two partner points z and $-1/\overline{z}$ in $\boldsymbol{C}$. Then the map $\phi^{-1} \circ \alpha \circ \phi$ is an isometric rotation of S^2.*

Proof: Let α be as in Lemma 22.10. We know by Exercise 5 that the map $\alpha' = \phi^{-1} \circ \alpha \circ \phi$ fixes two antipodal points and has order 3. We can find an isometry I' of S^2 that has order 3 and fixes these same two points. Let $I = \phi \circ I' \circ \phi^{-1}$. By Lemma 14.6, the map I is a linear fractional transformation. Note that I and α fix the same two points and dI and $d\alpha$ have the same action at either point. Hence $I = \alpha$, by Exercise 6. Hence $I' = \alpha'$, as desired. This completes the proof. ♠

Let $SL_2(\boldsymbol{C})$ denote the group of 2×2 matrices, with complex entries, having determinant 1. As in Chapter 10, these matrices represent linear fractional transformations.

Exercise 7. Let $z, w \in \boldsymbol{C} - \{0\}$ be distinct points. Prove that there exists an element $T_{z,w}$ of $SL_2(\boldsymbol{C})$ that represents a linear fractional transformation that carries 0 to z and ∞ to w.

The matrix

$$\beta = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}$$

represents the linear fractional transformation that has order 2 rotation about 0 and ∞. The matrix

$$\alpha_{z,w} = T_{z,w} \circ \begin{bmatrix} \omega & 0 \\ 0 & \omega^5 \end{bmatrix} \circ T_{z,w}^{-1}, \qquad \omega = \exp(\pi i/3),$$

represents an order 3 linear fractional transformation that fixes z and w. The entries of $\alpha_{z,w}$ are polynomials in z and w.

Once we pick $z \in \boldsymbol{C} - \{0\}$, we define a homomorphism $\rho_z : \Gamma \to SL_2(\boldsymbol{C})$ by the rule

$$\rho_z(A) = \alpha_{z,w}, \qquad w = 1/\overline{z}, \qquad \rho_z(B) = \beta.$$

Note that $\rho_z(\beta)$ doesn't depend on z. Now we define $\rho : \Gamma \to SO(3)$ by the rule

$$\rho = \phi^{-1} \circ \rho_z \circ \phi,$$

where ϕ is stereographic projection. Lemma 22.10 guarantees that $\rho(A)$ is an isometric rotation, and this is obvious for $\rho(B)$. Note that ρ is injective if and only if ρ_z is injective. So, at this point, we can forget about ρ entirely and just worry about choosing z so that ρ_z is injective. This is a problem entirely about linear fractional transformations.

Given any $\gamma \in \Gamma$, let $S(\gamma) \subset \boldsymbol{C}$ denote those $z \in \boldsymbol{C} - \{0\}$ such that $\rho_z(\gamma)$ is not the identity matrix. Setting $z = x + iy$, we see that the coordinates of $w = 1/z$ are rational functions of x and y. Therefore, the entries of $\rho_z(\gamma)$ are rational functions of x and y. Any rational function of x and y either vanishes identically or vanishes on a nowhere dense set. In particular, $S(\gamma)$ is either empty or else an open dense set.

Exercise 8. The Baire Category Theorem (for the plane) says that the intersection of a countable collection of open dense subsets of $\boldsymbol{C}$ is nonempty. Prove this result.

Suppose for the moment that $S(\gamma)$ is nonempty for all nontrivial $\gamma \in \Gamma$. Then, by the Baire Category Theorem, the intersection

$$\bigcap_{\gamma} S(\gamma)$$

is nonempty. Choosing any z in this intersection leads to an injective ρ_z. So, to finish our proof, we just have to show that $S(\gamma)$ is always nonempty.

We will fix γ and show that $S(\gamma)$ is nonempty. Recall that the element $\alpha_{z,w}$ is defined for all pair of distinct $z, w \in \boldsymbol{C}$. Accordingly, we can define the homomorphism $\rho_{z,w}$ by sending A to $\alpha_{z,w}$ and B to β. This lets us speak about the matrix entries of $\rho_{z,w}(\gamma)$. These are polynomials on the two complex variables z and w. Let F_{ij} be one of these polynomials.

Lemma 22.11 *F_{ij} is nontrivial for some i, j.*

Proof: Here is the crucial observation. We can choose (z, w) so that the image $\rho(\Gamma)$ is conjugate to the hyperbolic modular group discussed in §19.3 and §20.3. For this choice of (z, w), the matrix $\rho_{z,w}(\gamma)$ is not the identity. The point is that the corresponding element in the modular group does something nontrivial to the hyperbolic plane. Hence, the matrix coefficients of this matrix, as functions of z and w, cannot be constant. ♠

We let $F = F_{ij}$ for the indices guaranteed by the previous result. Let

$$R_\Delta = \{(z, -1/\overline{z}) |\ z \in \boldsymbol{C} - \{0\}\}.$$

We really only care about the restriction of F to R_Δ, because the other points in $\boldsymbol{C}^2$ do not correspond to homomorphisms from Γ into $SO(3)$.

Intuitively, what makes the next lemma work is that $\boldsymbol{R}_\Delta$ is a "big" subset of $\boldsymbol{C}^2$. An algebraic geometer would say that $\boldsymbol{R}_\Delta$ is *Zariski dense*, and that would be the end of the proof, but we will work out what we need from scratch. For the interested reader, most books on algebraic geometry will have a discussion of Zariski Density. See, for instance, [KEN].

Lemma 22.12 *F is nonconstant on R_Δ.*

Proof: This is a general result about polynomials in $\boldsymbol{C}^2$ and does not depend on the specific structure of F. We will suppose that F is constant in $\boldsymbol{R}_\Delta$ and derive a contradiction. Consider the following rational map on $\boldsymbol{C}^2$:

$$\theta(z_1, z_2) = \Big(z_1 + 1/z_2, i(z_1 - 1/z_2)\Big).$$

By construction $\theta(R_\Delta)$ is open in $\boldsymbol{R}^2$. The function $\theta \circ F \circ \theta^{-1}$ is a rational function on $\boldsymbol{C}^2$ that is constant on an open subset of $\boldsymbol{R}^2$. (A rational function is the ratio of two polynomials.) This forces $\theta \circ F \circ \theta^{-1}$ to be globally constant. But then F is globally constant as well. This contradiction completes the proof. ♠

Since F is not constant on R_Δ, the matrix $\rho_z(\gamma)$ cannot be constant on all of R_Δ. Hence $S(\gamma)$ is nonempty. This completes the last piece of unfinished business. There is an injective homomorphism $\rho : \Gamma \to SO(3)$.

23 Dehn's Dissection Theorem

We saw in §8.5 that any two polygons of the same area are dissection equivalent to each other. The purpose of this chapter is to prove Dehn's Dissection Theorem, which shows that the analogous result in 3 dimensions is false.

23.1 The Result

A *polyhedron* is a solid body whose boundary is a finite union of polygons, called *faces*. We require that any two faces are either disjoint or share a common edge or share a common vertex. Finally, we require that any edge common to two faces is not common to any other face.

A *dissection* of a polyhedron P is a description of P as a finite union of smaller polyhedra,

$$P = P_1 \cup \cdots \cup P_n, \tag{91}$$

such that the smaller polyhedra have pairwise disjoint interiors. Note that there is not an additional assumption, say, that the smaller polyhedra meet face to face.

Two polyhedra P and Q are *scissors congruent* if there are dissections $P = P_1 \cup \cdots \cup P_n$ and $Q = Q_1 \cup \cdots \cup Q_n$ such that each P_k is isometric to Q_k. Sometimes, one requires that all the isometries are orientation-preserving, but in fact and two shapes that are scissors congruent *via* general isometries are also scissors congruent *via* orientation preserving isometries. (This little fact isn't something that is important for our purposes.)

In 1900, David Hilbert posed 23 problems, now known as the *Hilbert Problems*. Hilbert's Third Problem asks if every two polyhedra of the same volume are scissors congruent to each other. (Hilbert conjectured that the answer was no.) The Hilbert Problems have inspired a huge amount of mathematics since 1900, but the third one was solved in 1901, by Max Dehn. In 1901, Dehn proved the following result.

Theorem 23.1 *The cube and the regular tetrahedron (of the same volume) are not scissors congruent.*

Exercise 1. Say that a *prism* is a polyhedron with 5 faces, two of which are parallel. So, a prism has a triangular cross-section. Prove that any two prisms of the same volume are scissors congruent. (*Hint*: After some effort you can boil this down to the the polygon dissection theorem.)

23.2 Dihedral Angles

The *dihedral angle* is an angle we attach to an edge of a polyhedron. To define this angle, we rotate the polyhedron so that the edge in question is vertical, and then we look directly down on the polyhedron. The two faces containing our edge appear as line segments, and the dihedral angle is the angle between these line segments. We will measure dihedral angles in such a way that a right angle has measure 1/4. All the dihedral angles of a cube are 1/4.

All edges of a regular tetrahedron have the same dihedral angle. We are going to prove that this common angle is irrational. Geometrically, this is the same as saying that one cannot fit finitely many tetrahedra precisely around an edge, even if these tetrahedra are permitted to wrap around more than once before closing back up.

We will place our tetrahedron in space so that one edge is vertical. Rather than work in $\boldsymbol{R}^3$, it is useful to work in $\boldsymbol{C}\times\boldsymbol{R}$, where $\boldsymbol{C}$ is the complex plane. This is a nice way to distinguish the vertical direction. Consider the complex number

$$\omega = \frac{1}{3} + \frac{2\sqrt{2}}{3}i. \qquad (92)$$

Note that $|\omega| = 1$. Let T_0 be the tetrahedron with vertices

$$(1,0), \qquad (\omega,0), \qquad \Big(0,\frac{1}{\sqrt{3}}\Big), \qquad \Big(0,\frac{-1}{\sqrt{3}}\Big).$$

Exercise 2. Check that T_0 is a regular tetrahedron.

Consider the new tetrahedron T_n, with vertices

$$(\omega^n,0), \qquad (\omega^{n+1},0), \qquad \Big(0,\frac{1}{\sqrt{3}}\Big), \qquad \Big(0,\frac{-1}{\sqrt{3}}\Big).$$

The tetrahedra $T_0, T_1, T_2, \ldots$ are just rotated copies of T_0. We are rotating about the vertical axis. Notice that T_{n+1} and T_n share a face for every n. To say that the dihedral angle is irrational is the same as saying that the list $T_0, T_1, T_2, \ldots$ is infinite. This is the same as saying that there is no n such that $\omega^n = 1$.

In the next section, we will rule out the possibility that $\omega^n = 1$ for any positive integer n. This means that $T_0, T_1, T_2 \ldots$ really is an infinite list. Hence, the common dihedral angle associated to the edges of a regular tetrahedron is irrational.

23.3 Irrationality Proof

The point of this section is to prove the following result: The complex number

$$\omega = \frac{1}{3} + \frac{2\sqrt{2}}{3}i \tag{93}$$

does not satisfy the equation $\omega^n = 1$ for any positive integer n.

Exercise 3. Check that $\omega^n \neq 1$ for $n = 1, 2, 3, 4, 5, 6$. Also check that $\omega^2 = (2/3)\omega - 1$.

In light of Exercise 2, we just have to check the case $n \geq 7$. Let $G(\omega)$ be the set of numbers of the form $a + b\omega$, where a and b are integers. This set is discrete: every disk intersects only finitely many elements of $G(\omega)$. The point here is that ω is not real. So, considered as vectors in the plane, 1 and ω are linearly independent over the reals.

Let $n \geq 7$ be the smallest value such that (supposedly) $\omega^n = 1$. Let $\boldsymbol{Z}[\omega]$ denote the set of numbers of the form

$$a_1\omega + a_2\omega^2 + \cdots + a_n\omega^n \tag{94}$$

where $a_1, \ldots, a_n$ are integers. $\boldsymbol{Z}[\omega]$ has the nice property that

$$(\omega^a - \omega^b)^c \in \boldsymbol{Z}[\omega] \tag{95}$$

for any positive integers a, b, c. This comes from the fact that $\omega^n = 1$. There are at least 7 powers of ω crowded on the unit circle, so at least 2 of them must be closer than 1 unit apart. But that means we can find integers a and b such that $|z| < 1$, where $z = \omega^a - \omega^b$. The numbers $z, z^2, z^3 \ldots$ all belong to $\boldsymbol{Z}[\omega]$, and these numbers are distinct because $|z^{n+1}| = |z||z^n| < |z^n|$. So, $\boldsymbol{Z}[\omega]$ is not discrete.

Using Exercise 3, we get

$$\omega^3 = \omega \times \omega^2 = \omega \times ((2/3)\omega - 1) = (2/3)\omega^2 - \omega = (5/9)\omega - (2/3),$$

and similarly for higher powers of ω. In general,

$$3^n(a_1\omega + \cdots + a_n\omega^n) = \text{integer} + \text{integer} \times \omega. \tag{96}$$

for any choice of integers $a_1, \ldots, a_n$. But then $\boldsymbol{Z}[\omega]$ is contained in a scaled-down copy of $G(\omega)$ and hence is discrete. But $\boldsymbol{Z}[\omega]$ is not discrete, and we have a contradiction.

23.4 Rational Vector Spaces

Let $\mathcal{R} = \{r_1, \ldots, r_n\}$ be a finite list of real numbers. Let V be the set of all numbers of the form

$$a_0 + a_1 r_1 + \cdots + a_N r_N, \qquad a_0, a_1, \cdots, a_N \in \boldsymbol{Q}.$$

V is a finite dimensional $\boldsymbol{Q}$-vector space.

We declare two elements $v_1, v_2 \in V$ to be *equivalent* if $v_1 - v_2 \in \boldsymbol{Q}$. In this case we write $v_1 \sim v_2$. Let $[v]$ denote the set of all elements of V that are equivalent to v. Let W denote the set of equivalence classes of V. The two operations are given by

$$[v] + [w] = [v + w], \qquad r[v] = [rv].$$

The 0-element is given by $[0]$.

Exercise 4. Prove that these definitions make sense, and turn W into another finite-dimensional $\boldsymbol{Q}$-vector space.

Let $v_1, \ldots, v_m$ be a basis for V, and let $w_1, \ldots, w_n$ be a basis for W. The *tensor product* $V \otimes W$ is the $\boldsymbol{Q}$-vector space of formal linear combinations

$$\sum_{i,j} a_{ij}(v_i \otimes w_j), \qquad a_{ij} \in \boldsymbol{Q} \tag{97}$$

Here $v_i \otimes w_j$ is just a formal symbol, but in a compatible way the symbol $\otimes$ defines a bilinear map from $V \times W$ into $V \otimes W$:

$$\Big(\sum a_i v_i\Big) \otimes \Big(\sum b_j w_j\Big) = \sum a_i b_j (v_i \otimes w_j). \tag{98}$$

The $m \times n$ elements $\{1(v_i \otimes w_j)\}$ serve as as a basis for $V \otimes W$.

Here is a basic property of $V \otimes W$. If $v \in V$ is nonzero and $w \in W$ is nonzero, then $v \otimes w$ is nonzero. One sees this simply by writing v and w out in a basis and considering equation (98). At least one product $a_i b_j$ will be nonzero. In particular

$$6 \otimes \delta \neq 0, \tag{99}$$

where δ is the dijedral angle of the regular tetrahedron, and $\mathcal{R}$ is chosen so as to contain δ.

23.5 Dehn's Invariant

Let $\mathcal{R} = \{r_1, \ldots, r_N\}$ be a finite list of real numbers, and let V and W be the two examples of vector spaces given in Examples 1 and 2 above. Once again, V is the set of all numbers of the form

$$a_0 + a_1 r_1 + \cdots + a_n r_N, \qquad a_0, \ldots, a_N \in \boldsymbol{Q},$$

and W is the set of equivalence classes in V.

Suppose that X is a polyhedron. Let $\lambda_1, \ldots, \lambda_k$ denote the side lengths of all the edges of X. Let $\theta_1, \ldots, \theta_k$ be the dihedral angles, listed in the same order. We say that X is *adapted* to $\mathcal{R}$ if

$$\lambda_1, \ldots, \lambda_k, \theta_1, \ldots, \theta_k \in \mathcal{R}. \tag{100}$$

If X is adapted to $\mathcal{R}$, we define the *Dehn invariant* as

$$\langle X \rangle = \sum_{i=1}^{k} (\lambda_i \otimes [\theta_i]) \in V \otimes W. \tag{101}$$

The operation $\otimes$ is as in equation (97), and the addition makes sense because $V \otimes W$ is a vector space.

Suppose now that P and Q are a cube and a regular tetrahedron having the same volume. Assume $\mathcal{R}$ is chosen large enough so that P and Q are both adapted to $\mathcal{R}$. Let λ_P and λ_Q denote the side lengths of P and Q, respectively. Let δ_P and δ_Q denote the respective dihedral angles. We have $[\delta_P] = [1/4] = [0]$, because $1/4$ is rational. On the other hand, we have already seen that δ_Q is irrational. Hence $[\delta_Q] \neq [0]$. This gives us

$$\langle P \rangle = 12\lambda_P \otimes [\delta_P] = [0], \qquad \langle Q \rangle = 6\lambda_Q \otimes [\delta_Q] \neq [0]. \tag{102}$$

In particular,

$$\langle P \rangle \neq \langle Q \rangle. \tag{103}$$

To prove Dehn's Theorem, our strategy is to show that the Dehn invariant is the same for two polyhedra that are scissors congruent. The result in the next section is the key step in this argument.

23.6 Clean Dissections

Say that a *clean dissection* of a polyhedron X is a dissection $X = X_1 \cup \cdots \cup X_N$, where each pair of polyhedra are either disjoint or share precisely a lower-dimensional face. Let $\mathcal{R}$ be as above.

Lemma 23.2 *Suppose that $X = X_1 \cup \cdots \cup X_N$ is a clean dissection and all polyhedra are adapted to $\mathcal{R}$. Then $\langle X \rangle = \langle X_1 \rangle + \cdots + \langle X_N \rangle$.*

Proof: We will let Y stand for a typical polyhedron on our list. Say that a *flag* is a pair (e, Y), where e is an edge of Y. Then

$$\langle X_1 \rangle + \cdots + \langle X_N \rangle = S = \sum_{f \in F} \lambda(f) \otimes \theta'(f).$$

Here F is the set of all flags and $\lambda(f)$ and $\theta'(f)$ are the length and dihedral angle associated to the flag f.

We classify the flag (e, Y) as one of three types:

- **Type-1.** e does not lie on the boundary of P.
- **Type-2.** e lies in the boundary of P, but not in an edge.
- **Type-3.** e lies in an edge of P.

We can write $S = S_1 + S_2 + S_3$, where S_j is the sum over flags of Type j.

Call two flags (e, Y) and (e', Y') *strongly equivalent* iff $e = e'$. Given a Type-1 edge e, let $\theta_1, \ldots, \theta_m$ denote the dijedral angles associated to the flags involving e. From the clean dissection property, these polyhedra fit exactly around e, so that (with our special units) $\theta_1 + \cdots + \theta_m = 1$. Hence

$$\sum \lambda(e) \otimes [\theta_j] = \lambda(e) \otimes \sum [\theta_j] = \lambda(e) \otimes [1] = 0.$$

Summing over all Type-1 equivalence classes, we find that $S_1 = 0$. A similar argument shows that $S_2 = 0$. In this case $\theta_1 + \cdots + \theta_k = 1/2$.

Now we show that $S_3 = \langle X \rangle$. Define a *weak equivalence class* as follows. (e, P) and (e', P') are weakly equivalent iff e and e' lie in the same edge of X. The set of weak equivalence classes is bijective with the set of edges of X. Let e be some edge of X, with length and dihedral angle λ and θ. Let $e_1, \ldots, e_m$ be the different edges that appear in weak equivalence class named by e. With the obvious notation $\lambda = \lambda_1 + \cdots + \lambda_k$. Let $\theta_{j1}, \ldots, \theta_{jm_j}$ denote the dihedral angles associated to the strong equivalence class involving e_j. We have $\theta_{j1} + \cdots \theta_{jm_j} = \theta$. Summing over the weak equivalence class, we get

$$\sum_{jk} \lambda_j \otimes [\theta_{jk}] = \sum_j \lambda_j \otimes [\theta] = \lambda(e) \otimes [\theta(e)].$$

Summing over all weak equivalence classes, we get $S_3 = \langle X \rangle$, as desired. ♠

23.7 The Proof

Let P be a cube, and let Q be a tetrahedron. We will suppose that we have a scissors congruence between $P = P_1 \cup \cdots \cup P_n$ and $Q = Q_1 \cup \cdots \cup Q_n$.

We first produce new dissections of P and Q that are clean. Here is the construction. Let $\Pi_1, \ldots, \Pi_k$ denote the union of all the planes obtained by extending the faces of any polyhedron in the above dissection of P. Say that a *chunk* is the closure of a component of $\boldsymbol{R} - \bigcup \Pi_j$. Then we have clean dissections

$$P_i = P_{i1} \cup \cdots \cup P_{in_i} \tag{104}$$

of each P_i into chunks, and also the clean dissection

$$P = \bigcup P_{ij} \tag{105}$$

of P into chunks. We make all the same definitions for Q. The dissections in equation (105) for P and Q might not define a scissors congruence, but we don't care.

Let $\mathcal{R}$ denote the finite list of lengths and dihedral angles that arise in any of the polyhedra appearing in our constructions involving P and Q. Let $V \otimes W$ be the vector space defined as in the previous sections, relative to $\mathcal{R}$. Computing the Dehn invariants in $V \otimes W$, we have

$$\langle P \rangle = \sum \langle P_{ij} \rangle = \sum \langle P_i \rangle = \sum \langle Q_i \rangle = \sum \langle Q_{ij} \rangle = \langle Q \rangle. \tag{106}$$

The first equality is obtained by applying Lemma 23.2 to the dissection in equation (105). The second equality is obtained by applying Lemma 23.2 to each dissection in equation (104) and adding the results. The middle equality comes from the obvious isometric invariance of the Dehn invariant. The last two equalities have the same explanations as the first two. In short, $\langle P\rangle = \langle Q\rangle$. This contradicts our computation that $\langle P\rangle \neq \langle Q\rangle$. The only way out of the contradiction is that the cube and the tetrahedron are not scissors congruent.

Exercise 5. Consider all the unit area platonic solids. Which are scissors congruent to which?

24 The Cauchy Rigidity Theorem

The purpose of this chapter is to prove the Cauchy Rigidity Theorem for strictly convex polyhedra. One can find another proof in the book [AIZ] As the authors point out therein, Cauchy's original proof was flawed, and a correct proof from comes from a letter from I.J. Schoenberg to K. Zaremba.

The proof is half geometrical and half combinatorial. The geometrical half of the proof I give is very similar to what Aigner and Ziegler do, except that I spell out some of the intermediate steps in more detail. The combinatorial half can be done in many ways, and I give an argument based on the combinatorial Gauss–Bonnet Theorem; see §17.3.

24.1 The Main Result

A *polyhedron* is a solid body whose boundary is a finite union of polygons, called *faces*. A polyhedron P is called *strictly convex* if, for each face f of P, there is a half-space Π_f such $P \subset \Pi_f$ and $P \cap \partial\Pi_f = f$. The boundary $\partial\Pi_f$ is the plane extending f. The cube is a classic example of a strictly convex polyhedron.

Say that two polyhedra P and P' are *flexes* of each other if there is a homeomorphism from ∂P to $\partial P'$ which is an isometry when restricted to each face. In other words, there is a combinatorics-respecting bijection between the faces of P and the faces of P' such that corresponding faces are isometric to each other. Making the same definition for polygons, we observe that any pair of rhombuses, having unit side length, are flexes of each other. The Cauchy Rigidity Theorem rules out this behavior in 3 dimensions, at least for strictly convex polyhedra.

Theorem 24.1 (Cauchy) *Let P and P' be two strictly convex polyhedra. If P and P' are flexes of each other then P and P' are isometric.*

Exercise 1. Show by example that the Cauchy Rigidity Theorem is false when the convexity assumption is dropped.

Amazingly, Robert Connelly discovered examples of continuous families of polyhedra, in which every two are flexes of each other. In other words, Connelly's examples actually flex in a literal sense.

24.2 The Dual Graph

There is a nice graph that lies on the surface of P, called the *dual graph*. We place one new vertex per face of P, and join two vertices by an edge if and only if the corresponding faces share an edge.

There is a nice geometric way to picture the dual graph. Let S denote a set of points, one per interior face of P. Just to be definite, we choose the center of mass of each face of P. Then, let P^* denote the convex polygon whose vertex set is S. Formally, we can say that P^* is the *convex hull* of S, namely the intersection of all closed and convex subsets of $\boldsymbol{R}^3$ that contain S. The dual graph is exactly the union of edges and vertices of P^*.

Exercise 2. When P is a platonic solid, P^* is also a platonic solid. The construction pairs up the cube with the octahedron, the dodecahedron with the icosahedron, and the tetrahedron with itself, or, rather, a slightly smaller tetrahedron. Try to draw pictures of these cases.

To get perhaps the nicest picture of the dual graph, we surround P^* by a large sphere and then project the dual graph onto the surface of the sphere by a radial projection from some point in the interior of P^*. Finally, we identify this large sphere with S^2, the unit sphere. This gives us a graph Γ on S^2, all of whose edges are arcs of circles. Each component of $S^2 - \Gamma$ is a polygon whose boundary is made from circular arcs. The important thing for us is just that each component is homeomorphic to a disk.

There is a natural correspondence between the edges of the polygon and the edges of the dual graph. When we draw the dual graph directly on P, each edge of the dual graph crosses one edge of P, and vice versa. So, if we have some kind of labelling of the edges of P, we can transfer it in the obvious way to a labelling of the edges of the dual graph Γ, which we think about in its final incarnation, as a graph on S^2.

24.3 Outline of the Proof

Each edge e of P has a partner edge e' of P'. Let $\theta(e)$ be the dihedral angle of P at e, and let $\theta(e')$ be the corresponding dihedral angle of P' at e'. (Recall that the dihedral angle is the angle made by the planes incident to the edge.) We label the edge e by $(+)$, $(-)$, or (0) according as to whether the sign of $\theta(e) - \theta(e')$ is positive, negative, or zero.

We transfer our labelling to the dual graph, $\Gamma \subset S^2$. Each component C of $S^2 - \Gamma$ is bounded by a circuit in Γ. We get a cyclically ordered list $L(C)$ of members of $\{+, -, 0\}$ by reading the labels of this circle, say, in clockwise order.

We call $L = L(C)$ *bad* if, after we delete all the 0's from L, we have a nonempty list that changes from $+$ to $-$ at most once as we cycle through it. Otherwise, we call L *good*. For instance $(+0 + - - -00+)$ is a bad list, and $(+ + - - - + -)$ is a good list.

Below we will prove two results. The first is geometrical and the second is combinatorial.

Lemma 24.2 *For any component C of $S^2 - \Gamma$, the list $L(C)$ is good.*

Lemma 24.3 *Let Γ be a graph on S^2 such that each component of $S^2 - \Gamma$ is an embedded topological disk. Suppose that the edges of Γ are labelled nontrivially by elements of $\{+, -, 0\}$. Then there is at least one component C of $S^2 - \Gamma$ such that $L(C)$ is a bad list.*

Our two lemmas contradict each other unless the labelling of Γ is completely trivial. But then $\theta(e) = \theta(e')$ for all edges e of P. But this easily implies that P and P' are isometric.

Exercise 3. Build half an octahedron by taping together 4 cardboard equilateral triangles about a vertex. The portion of Γ corresponding to these faces is a quadrilateral. Physically flex the object and observe that the only possible nontrivial labelling is $(+ - +-)$ or, of course, $(- + -+)$. Compare this with Exercise 5 from Chapter 9.

Exercise 4. Without looking at the long-winded proof below, prove Lemma 24.3 for the cube.

Exercises 3 and 4 combine to prove Cauchy's Theorem for P and P', when P is a regular octahedron.

Exercise 5. Imitating Exercises 3 and 4, Give a proof of Cauchy's Theorem for the regular icosahedron.

24.4 Proof of Lemma 24.3

Let P be a polygon whose edges are labelled $(+)$ and $(-)$. We say that P has a *good labelling* if the list of labels around its edges is good; that is, there at least 2 sign changes from $(+)$ to $(-)$. A quadrilateral with a good labelling must be labelled $(+,-,+,-)$, up to cyclic ordering.

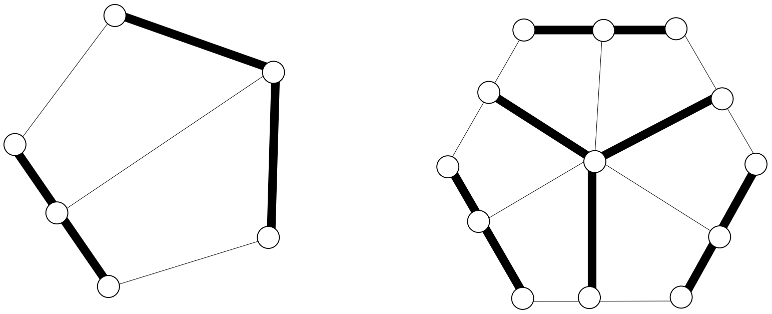

Figure 24.1. Adding edges and vertices

Figure 24.1 shows examples of how one can divide a polygon with a good labelling into quadrilaterals with good labellings. In the subdivision process, you are allowed to add both edges and vertices.

Exercise 6. Prove Lemma 24.4 below.

Lemma 24.4 *Suppose that P is a polygon with a good labelling. Let P_v denote the vertex set of P. We can partition P into alternately labelled quadrilaterals, extending the labelling on P, such that the following is true:*

- *Let w be a vertex of a quadrtilateral that lies in the interior of P. Then w is a vertex of at least 4 quadrilaterals.*
- *Let w be a vertex of a quadrilateral that lies in $\partial P - P_v$. Then w is a vertex of 2 quadrilaterals.*

First we prove Lemma 24.3 in the special case that all labels of Γ are nonzero. By Lemma 24.4, we can partition each component of $S^2 - \Gamma$ into alternately labelled quadrilaterals. These partitions fit together to partition S^2 itself into alternately labelled quadrilaterals. Now we make 3 observations.

- Each quadrilateral vertex in the interior of a component of $S^2 - \Gamma$ is a vertex of 4 quadrilaterals.

- Each vertex in the interior of an edge is a vertex of exactly 4 quadrilaterals, two coming from each side.

- Each vertex of Γ is a vertex of at least 3 quadrilaterals, by the valence condition. However, the edges emanating from a vertex must alternate in sign, given that our quadrilaterals are all alternately labelled. Hence, each vertex of Γ is a vertex of at least 4 quadrilaterals.

In short, every quadrilateral vertex is the vertex of at least 4 quadrilaterals.

Now we build a Euclidean cone surface based on our partition. We glue together unit squares using the combinatorial pattern given by our quadrilaterals. Call the resulting surface Σ. By construction, the cone angle of Σ is at least 2π at each quadrilateral vertex. The remaining points of Σ are locally Euclidean. Hence, the total combinatorial curvature of Σ is nonpositive. But Σ is homeomorphic to S^2. This contradicts the combinatorial Gauss–Bonnet Theorem.

Now consider the general case, where there are possibly edges labelled with a zero. Our proof goes by induction on the number Z of edges that have the zero label. We already treated the case when $Z = 0$. In general, suppose that e is an edge labelled 0. There are two cases.

In the first case, suppose that the closed edge E is embedded. We can form a new graph Γ_e in S^2 by collapsing e to a point and dragging all the edges of Γ incident to e to this new point; see Figure 24.3.

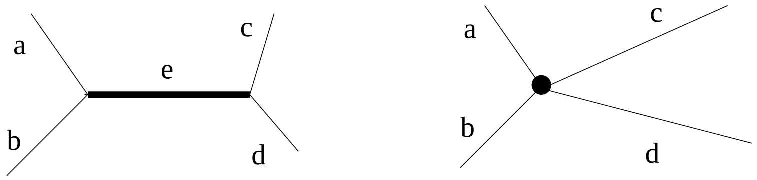

Figure 24.2. Collapsing an edge

Our operation only changes the two components of $S^2 - \Gamma$ that share e. These components remain topological disks: we are just shrinking one of their edges to a point. Moreover, since e is labelled 0, the lists associated to each of these two components remain good. In short Γ_e satisfies the same hypotheses as Γ but has one fewer edge labelled 0.

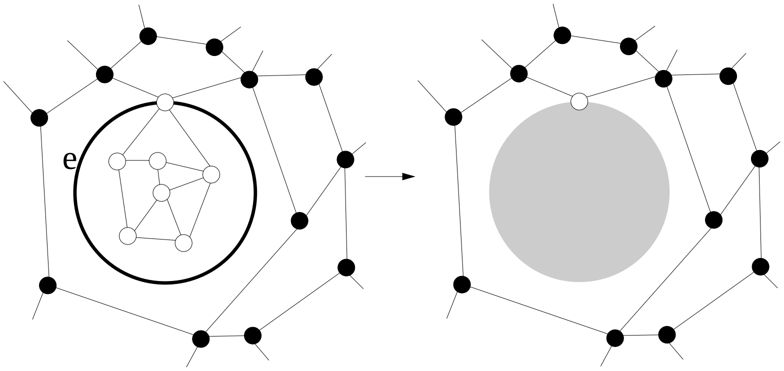

Figure 24.3. Replacing the inner part of the graph with a disk

In the other case, e is a loop in S^2. Note that e divides S^2 into two disks. At least one of these disks—say the outer one, as in Figure 24.3—contains some edges of Γ. Pick such a disk, and then erase the portion of Γ contained in the other disk – the inner disk in Figure 24.3. Finally, erase e. Figure 24.4 shows this operation. The result is a smaller graph that satisfies the hypotheses of Lemma 24.3 but the Z value has decreased by one.

24.5 Proof of Lemma 24.2

Say that a *spherical arm* is a connected polygonal arc contained in the boundary of a convex spherical polygon. Thus, a spherical arm is made from a finite union of arcs of great circles, meeting end to end. We insist that the two endpoints of the spherical arm are distinct, so that the spherical arm does not make a complete circuit around the spherical polygon. Given the notion of convexity discussed in Chapter 9, a spherical arm is necessarily contained in a hemisphere.

Suppose that $A(0)$ and $A(1)$ are spherical arms, each consisting of n geodesic segments. Let $A_1(k), \ldots, A_n(k)$ be the geodesic segments comprising $A(k)$, taken in order. Let $a_1(k), \ldots, a_n(k)$ be the vertices of $A(k)$. Finally, let $\theta_j(k)$ be the interior angle of $A(k)$ at $a_j(k)$. The choice of interior angle makes sense, thanks to convexity.

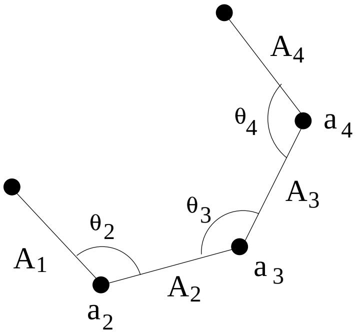

Figure 24.4. A spherical arm of length 4

Below we will prove Cauchy's Arm Lemma:

Lemma 24.5 (Cauchy's Arm Lemma) *Assume that $\theta_j(0) \leq \theta_j(1)$ for all j, with strict inequality for at least one index. Then*

$$d(a_0(0), a_n(0)) < d(a_0(1), a_n(1)).$$

Here d denotes spherical distance.

Before proving Cauchy's Arm Lemma, let's use it to prove Lemma 24.2. Let v be a vertex of a strictly convex polyhedron. Let Σ be a small sphere centered at v. The intersection $\partial P \cap \Sigma$ is a convex spherical polygon. Dilating the picture, we think of this polygon as existing on S^2, the unit sphere.

We can make this construction for partner vertices v and v' on P and P', respectively. This produces two convex spherical polygons C and C'. The lengths of the edges of C are the same as the lengths of the corresponding edges of C'. We label the vertices of C as in Cauchy's Arm Lemma from the previous chapter, depending on the comparison between the two internal angles at the vertices.

If our list of labels is not good, we can find a chord of C so that all the $(+)$ labels occur on one side and all the $(-)$ labels occur on the other. This is shown in Figure 24.5. Let p and q be the endpoints of this chord. Let C_1 denote one of the arcs of C connecting p to q, and let C_2 denote the other. Let C_1' and C_2' be the corresponding chords on C'.

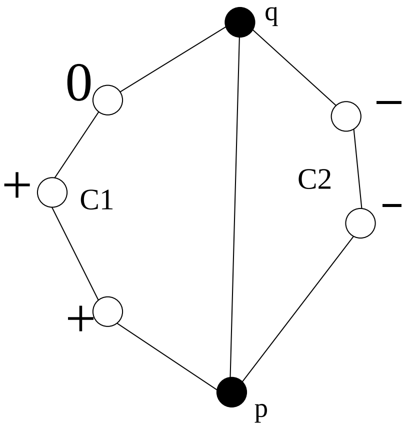

Figure 24.5. Dividing the polygon in half

Applying Cauchy's Arm Lemma to C_1 and C_1', we conclude that

$$\|p-q\| > \|p'-q'\|.$$

Applying the Arm Lemma to C_2 and C_2', we get the opposite inequality. This is a contradiction.

24.6 Euclidean Intuition Does Not Work

The proof of Cauchy's Arm Lemma is actually rather difficult, though the result seems obvious based on Euclidean intuition.

Cauchy's mistake was that he assumed a result from Euclidean geometry that is false in the spherical case. In this section, I'll highlight the difference between the Euclidean and spherical cases. My reason for doing this is to justify the difficulty it takes to actually prove Cauchy's Arm Lemma.

Consider a Euclidean version of the main construction. We say that a *Euclidean arm* is a connected arc of a convex Euclidean polygon. Suppose that $A(0)$ is a Euclidean arm. We can make a polygonal arc $A(t)$, for $t > 0$ by increasing the last angle of $A(0)$. Call this angle $\theta(t)$. We keep everything else fixed. One should picture a person flexing his finger; Figure 24.5 shows the situation.

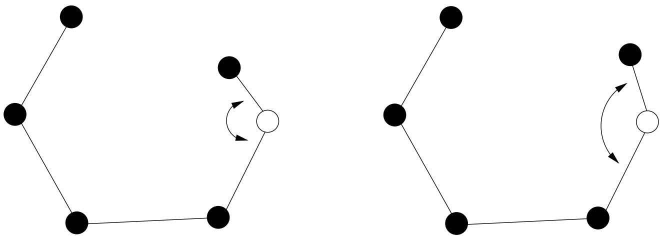

Figure 24.6. Flexing a Euclidean arm.

Exercise 7. Prove that $A(t)$ is a Euclidean arm provided that $\theta(t) < \pi$. That is, the process of opening up one of the angles cannot destroy convexity.

Exercise 8. Show, by example, that the conclusion of Exercise 7 is no longer true in the spherical case.

The natural approach to proving Cauchy's Arm Lemma is to simply open up one of the arms a bit at a time, showing that the distance between the endpoints keeps increasing. Unfortunately, in the spherical case, the object can cease to be an arm at some point because one can lose the convexity.

24.7 Proof of Cauchy's Arm Lemma

Let's analyze the problem of flexing a spherical arm. Let $B(t)$ be a spherical arm for all $0 \leq t < s$. Let $b_0, \ldots, b_{n-1}, b_n(t)$ be the points of $B(t)$. Only the last point changes. Let $B_1, \ldots, B_{n-1}, B_n(t)$ be the segments of $B(t)$. We suppose that the angle $\theta(t)$ at b_{n-1} increases as $t \to s$, but that $\theta(s) < \pi$.

Lemma 24.6 *Whether or not $B(s)$ is a spherical arm, $B(s)$ lies in some open hemisphere.*

Proof: Suppose not. Let $\widehat{B}_n(t)$ be the great circle extending $B_n(t)$, and let $H(t)$ denote the open hemisphere containing $B(t) - B_n(t)$ for t small. By Exercise 4 of Chapter 9, and continuity, $H(t)$ contains $B(t) - B_n(t)$ for all $t < s$. But $H(s)$ cannot contain $B(s) - B_n(s)$, because then we could move $B(s)$ by a tiny amount so that it lies in $H(s)$.

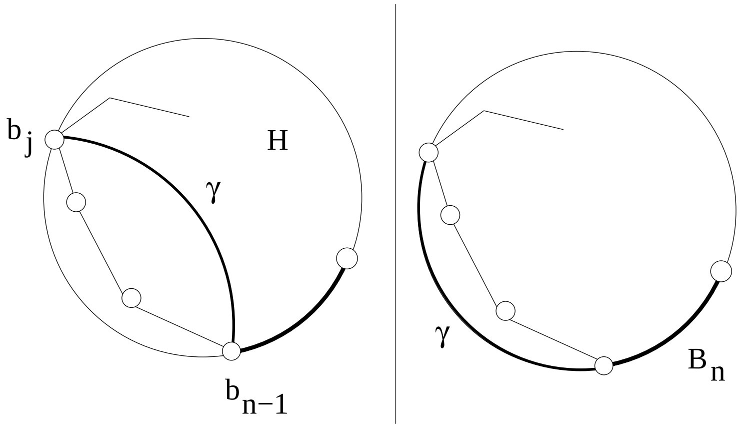

Figure 24.7. An arm and a great circle

The only possibility is that there is some vertex $b_j(s)$ contained in $\widehat{B}_n(s)$. Here $j \leq n-2$. Let γ be the geodesic joining b_{n-1} to b_j. The angle between γ and $B_n(t)$ is bounded away from both 0 and π. The endpoints of γ are not antipodal because they are vertices of a spherical arm. Therefore, γ is the unique geodesic connecting its endpoints. But, the condition $b_j \in \widehat{B}_n(s)$ forces $\gamma \subset \widehat{B}_n(s)$. This is a contradiction. ♠

We keep going with the same set-up as in the previous lemma.

Lemma 24.7 *If $B(s)$ is not a spherical arm, then b_0, b_1, and $b_n(s)$ lie on the same arc of a great semicircle, with b_1 between b_0 and $b_n(s)$.*

Proof: From the previous result, we know that $B(t)$ lies in some open hemisphere for all $t \leq s$. Since $B(s)$ is not a spherical arm, there are 3 points $\beta_0, \beta_1, \beta_2 \in B(s)$, not all on the same edge of $B(s)$ but all lying on the same geodesic segment β. These 3 points cannot lie in any spherical arm, so one of the points, say β_0, must lie in $B_n(s) - b_{n-1}$.

For the same reason as in the previous lemma, β does not lie in the great circle $\widehat{B}_n(s)$. At the same time, β cannot be transverse to $B(s)$ at any β_j. Otherwise, by stability, we would have a similar triple of points for all t sufficiently close to s.

Suppose β_0 is an interior point of $B_n(s)$. Since $\beta \not\in \widehat{B}_n(s)$, the segment β is transverse to $B(s)$ at β_0. This is a contradiction. Hence $\beta_0 = B_n(s)$.

The points β_1 and β_2 both lie on the spherical arm $A = B(s) - B_n(s)$. If β_1 and β_2 do not lie on the same edge of A, then β is transverse to A at both

β_1 and β_2. This is a contradiction. Hence β_1 and β_2 lie on the same edge of A. Since $A \subset B(s)$, we see that β_1 and β_2 lie on the same side, say, the jth side, of $B(s)$.

If $j > 1$ then we have the topological picture shown (for $j = 2$) in Figure 24.8. This picture is implied by the fact that $B(s) - B_n(s)$ is a spherical arm. But then there is a geodesic nearby β, through $b_n(s)$, which intersects $B(s)$ transversely at the two other intersection points. The same stability argument as above gives us a contradiction.

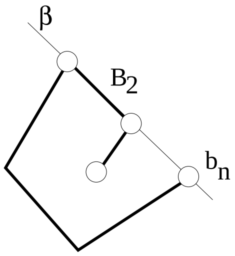

Figure 24.8. The limiting shape

Now we know that b_0 and b_1 and $b_n(s)$ lie on the same great circle. Since all these points lie in an open hemisphere, all these points lie in the interior of some great semicircle. Finally, observe that the geodesic connecting b_0 to b_2 crosses the geodesic connecting b_1 to $b_n(t)$ for all $t < s$. Taking a limit, as $t \to s$, establishes that b_1 lies between b_0 and $b_n(s)$. ♠

Finally, we prove Cauchy's Arm Lemma. The proof goes by induction on n. In the case where $n = 2$, the result follows from Exercise 5 of Chapter 9. Consider the special case when $\theta_j(0) = \theta_j(1)$ for some j. Then we can produce a new spherical arm B by replacing $A_j \cup A_{j+1}$ by the single geodesic segment connecting a_j to a_{j+1}. Here we are just cutting off a corner. The spherical arms $B(0)$ and $B(1)$ satisfy the same hypotheses as do $A(0)$ and $A(1)$, and our basic move has not changed the endpoints. Hence, by induction $d(a_0(0), a_n(0)) < d(a_0(1), a_n(1))$.

For $t \in [0, 1]$, let $\theta_{n-1}(t)$ be the angle that linearly interpolates between $\theta_{n-1}(0)$ and $\theta_{n-1}(1)$. Let $B(t)$ denote the polygonal curve that is the same as $A(0)$, except that we move A_n so that the angle between A_n and A_{n-1} is $\theta(t)$. Let $B_j(t)$ be the jth segment of $B(t)$, and let $b_j(t)$ be the jth vertex. We have set things up so that $B_j(t) = A_j(0)$ for $j = 1, \ldots, n-1$ and $b_j(t) = a_j(0)$ for $j = 1, \ldots, n-1$. Only the last segment moves.

Suppose that $B(1)$ is a spherical arm. Only one angle of $B(1)$ differs from $A(1)$. At the same time, the last angle of $B(1)$ is the same as the last angle of $A(1)$. Thus, we may apply the special case we have already considered, twice, to get the following chain of inequalities:

$$d(a_0(0), a_n(0)) < d(b_0(0), b_n(0)) < d(a_0(1), a_n(1)).$$

Now we come to the hard part of the proof. Suppose that $B(1)$ is not a spherical arm. By Exercise 8 above, this case really can happen. If $B(1)$ is not a spherical arm, then there is some s such that $B(t)$ is a spherical arm for all $t < s$, but $B(s)$ is not a spherical arm. By Lemma 24.7, the points $b_0(s), b_1(s), b_n(s)$ lie on the same great half-circle, with $b_0(s)$ between $b_1(s)$ and $b_n(s)$. Therefore

$$d(b_1(s), b_n(s)) = d(b_0(s), b_1(s)) + d(d_0(s), b_n(s)). \qquad (107)$$

We have

$$\begin{gathered} d(a_0(1), a_n(1)) \geq^1 \\ d(a_1(1), a_n(1)) - d(a_0(1), a_1(1)) \geq^2 \\ \lim_{t \to s} d(a_1(t), a_n(t)) - d(a_0(1), a_1(1)) =^3 \\ \lim_{t \to s} d(a_1(t), a_n(t)) - d(a_0(t), a_1(t)) =^4 \\ d(a_1(s), a_n(s)) - d(a_0(s), a_1(s))) =^5 d(a_0(s), a_n(s)). \end{gathered}$$

The first inequality is the triangle inequality. The second inequality is the induction step applied to the spherical arm obtained from $B(t)$ by chopping off the first segment $B_1(t)$. The third equality comes from the fact that $b_0(t)$ and $b_1(t)$ are independent of t. The fourth equality is continuity. The fifth equality is equation (107).

On the other hand, choosing any $u \in (0, s)$, we have

$$\begin{gathered} d(a_0(s), a_n(s)) \\ = \lim_{t \to s} d(a_0(t), a_n(t)) \geq^1 \\ d(a_0(u), a_n(u)) > \\ d(a_0(0), a_n(0). \end{gathered}$$

The first inequality comes from the special case (some angles equal) applied to $B(u)$ and $B(u)$. The last inequality comes from the special case appied to $B(u)$ and $B(0)$. Our last two equations combine to give the statement in Cauchy's Arm Lemma. This completes the proof.

References

[AHL] L. Ahlfors, *Complex Analysis*, McGraw-Hill, New York, 1952.

[AIZ] M. Aigner and G. Ziegler, *Proofs from The Book*, Springer-Verlag, 1998.

[BAL] T. Banchoff and S. Lovett, *Differential Geometry of Curves and Surfaces*, A. K. Peters, Ltd., Natick, MA, 2010.

[BE1] A. Beardon, *The Geometry of Discrete Groups*, Graduate Texts in Mathematics **91**, Springer-Verlag, New York, 1983.

[BE2] A. Beardon, *A Primer on Riemann Surfaces*, L.M.S. Lecture Note Series **78**, Cambridge University Press, Cambridge, 1984.

[BRO] M. Beck and S. Robins *Computing the Continuous Discretely*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 2007.

[CHE] C. Chevalley, *Theory of Lie Groups*, Princeton Mathematical Series **8**, Princeton University Press, Princeton, NJ, 1999.

[DAV] H. Davenport *The Higher Arithmetic (8th ed.)*, Canbridge University Press, Cambridge, 2008.

[DEV] K. Devlin *The Joy of Sets: Fundamentals of Contemporary Set Theory (2nd Ed.)*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1993.

[DOC] M. DoCarmo, *Riemannian Geometry*, Mathematics Theory and Applications, Birkhauser, Boston, 1992.

[DRT] A. Driscoll and L. N. Trefethen, *Schwarz-Christoffel Mapping*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, 2002.

[FMA] B. Farb and D. Margalit, *A Primer on Mapping Class Groups*, Princeton University Press, Princeton, (to appear).

[GAR] F. Gardiner, *Teichmüller Theory and Quadratic Differentials*, Pure and Applied Mathematics, John Wiley and Sons, New York, 1987.

[GPO] V. Guillemin and A. Pollack *Differential Topology*, Prentice Hall, Englewood Cliffs, NJ, 1974.

[HAT] A. Hatcher, *Algebraic Topology*, Cambridge University Press, Cambridge, 2002.

[HCV] D. Hilbert and A. Cohn-Vossen, *Geometry and the Imagination*, Chelsea Publishing Company, New York, 1952.

[HER] I. M. Herstein, *Topics in Algebra, 2nd Ed.*, John Wiley and Sons, Xerox College Publishing, Lexington, MA, 1975.

[KAT] S. Katok, *Fuchsian Groups*, Chicago Lectures in Mathematics, University of Chicago Press, Chicago, 1992.

[KEN] K. Kendig, *Elementary Algebraic Geometry*, Graduate Texts in Mathematics, Springer-Verlag, New York, 1977.

[KIN] L. C. Kinsey, *The Topology of Surfaces*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1993.

[MAT] H. Masur and S. Tabachnikov, *Rational Billiards and Flat Structures*, Handbook of Dynamical Systems Vol 1A, 1015-1089, North-Holland, Amsterdam, 2002.

[MUN] J. R. Munkries, *Topology*, Prentice Hall, Englewood Cliffs, NJ, 1975.

[RAT] J. Ratcliff, *Foundations of Hyperbolic Manifolds*, Graduate Texts in Mathematics, **149**, Springer-Verlag, New York, 1994.

[SPI] M. Spivak, *Calculus on Manifolds. A Modern Approach to Classical Theorems of Advanced Calculus*, W. A. Benjamin, New York–Amsterdam, 1965.

[TAP] K. Tapp, *Matrix Groups for Undergraduates*, Student Mathematical Library, **29**, American Math Society (2006)

[THU] W. P. Thurston, *The Geometry and Topology of* 3*-Manifolds*, Lecture Notes, Princeton University Press, Princeton, NJ, 1978.

[WAG] S. Wagon, *The Banach-Tarski Paradox*, Encyclopedia of Mathematics and Its Applications, Cambridge University Press, Cambridge, 1985.

[WAL] W. Wallace, *Question 269*, in Thomas Leyborne, Math. Repository III, London, 1814.