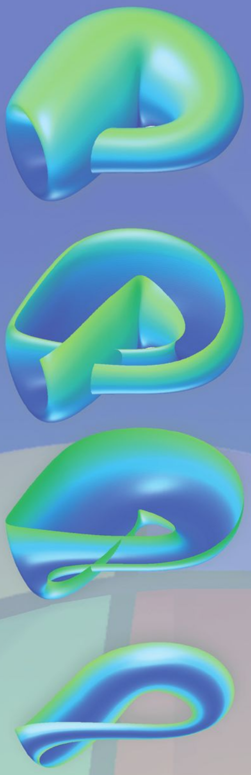
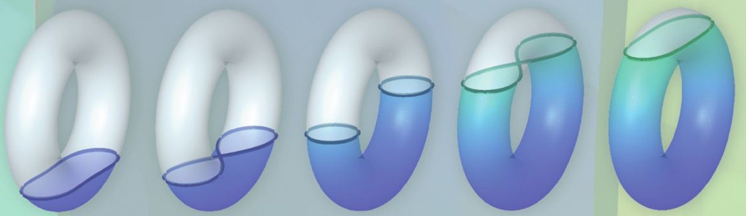
STUDENT MATHEMATICAL LIBRARY
Volume 46

Lectures on Surfaces

(Almost) Everything You Wanted to Know about Them

Anatole Katok
Vaughn Climenhaga





American Mathematical Society
Mathematics Advanced Study Semesters

Lectures on Surfaces

(Almost) Everything You Wanted to Know about Them

STUDENT MATHEMATICAL LIBRARY
Volume 46

Lectures on Surfaces

(Almost) Everything You Wanted to Know about Them

Anatole Katok
Vaughn Climenhaga

American Mathematical Society
Mathematics Advanced Study Semesters

Topographic data for illustration of Earth in Figure 1.12 is taken from the National Oceanic and Atmospheric Administration's ETOPO2v2 data set, available at `www.ngdc.noaa.gov`.

2000 *Mathematics Subject Classification.* Primary 51–01, 53–01, 57N05; Secondary 53A05, 57R05.

For additional information and updates on this book, visit
www.ams.org/bookpages/stml-46

Library of Congress Cataloging-in-Publication Data

Katok, A. B.
Lectures on surfaces : (almost) everything you wanted to know about them / Anatole Katok, Vaughn Climenhaga.
p. cm. — (Student mathematical library ; v. 46)
Includes bibliographical references and index.
ISBN 978-0-8218-4679-7 (alk. paper)
1. Surfaces. I. Climenhaga, Vaughn, 1982– II. Title.

QA571.K34 2008
516—dc22 2008029299

∞ The paper used in this book is acid-free and falls within the guidelines
established to ensure permanence and durability.
Visit the AMS home page at `http://www.ams.org/`

10 9 8 7 6 5 4 3 2 1 13 12 11 10 09 08

Contents

Foreword: MASS and REU at Penn State University

This book is part of a collection published jointly by the American Mathematical Society and the MASS (Mathematics Advanced Study Semesters) program as a part of the Student Mathematical Library series. The books in the collection are based on lecture notes for advanced undergraduate topics courses taught at the MASS and/or Penn State summer REU (Research Experiences for Undergraduates). Each book presents a self-contained exposition of a non-standard mathematical topic, often related to current research areas, accessible to undergraduate students familiar with an equivalent of two years of standard college mathematics and suitable as a text for an upper division undergraduate course.

Started in 1996, MASS is a semester-long program for advanced undergraduate students from across the USA. The program's curriculum amounts to sixteen credit hours. It includes three core courses from the general areas of algebra/number theory, geometry/topology and analysis/dynamical systems, custom designed every year; an interdisciplinary seminar; and a special colloquium. In addition, every participant completes three research projects, one for each core course. The participants are fully immersed into mathematics, and this, as well as intensive interaction among the students, usually leads to a dramatic increase in their mathematical enthusiasm and achievement. The program is unique for its kind in the United States.

The summer mathematical REU program is formally independent of MASS, but there is a significant interaction between the two: about half of the REU participants stay for the MASS semester in the fall. This makes it possible to offer research projects that require more than seven weeks (the length of the REU program) for completion. The summer program includes the MASS Fest, a two to three day conference at the end of the REU at which the participants present their research and that also serves as a MASS alumni reunion. A nonstandard feature of the Penn State REU is that, along with research projects, the participants are taught one or two intense topics courses.

Detailed information about the MASS and REU programs at Penn State can be found on the website `www.math.psu.edu/mass`.

Preface

This book is a result of the MASS course in geometry in the fall semester of 2007. MASS core courses are traditionally labeled as analysis, algebra, and geometry, but the understanding of each area is broad, e.g. number theory and combinatorics are allowed as algebra courses, topology is considered as a part of geometry, and dynamical systems as a part of analysis. No less importantly, an interaction of ideas and concepts from different areas of mathematics is highly valued.

The topic came to me as very natural under these conditions. Surfaces are among the most common and easily visualized mathematical objects, and their study brings into focus fundamental ideas, concepts, and methods from geometry proper, topology, complex analysis, Morse theory, group theory, and suchlike. At the same time, many of those notions appear in a technically simplified and more graphic form than in their general "natural" settings. So, here was an opportunity to acquaint a group of bright and motivated undergraduates with a wealth of concepts and ideas, many of which would be difficult for them to absorb if presented in a traditional fashion. This is the central idea of the course and the book reflects it closely.

The first, primarily expository, chapter introduces many (but not all) principal actors, such as the round sphere, flat torus, Möbius strip, Klein bottle, elliptic plane, and so on, as well as various methods of describing surfaces, beginning with the traditional representation by equations in three-dimensional space, proceeding to parametric representation, and introducing the less intuitive, but central for our purposes, representation as factor spaces. It also includes a preliminary discussion of the metric geometry of surfaces. Subsequent chapters introduce fundamental mathematical structures: topology, combinatorial (piecewise-linear) structure, smooth structure, Riemannian metric, and complex structure in the specific context of surfaces. The assumed background is the standard calculus sequence, some linear algebra, and rudiments of ODE and real analysis. All notions are introduced and discussed, and virtually all results proved, based on this background.

The focal point of the book is the Euler characteristic, which appears in many different guises and ties together concepts from combinatorics, algebraic topology, Morse theory, ODE, and Riemannian geometry. The repeated appearance of the Euler characteristic provides both a unifying theme and a powerful illustration of the notion of an invariant in all those theories.

A further idea of both the motivations and the material presented in the book may be found in the Table of Contents, which is quite detailed.

My plan for teaching the course was somewhat bold and ambitious, and could have easily miscarried had I not been blessed with a teaching assistant who became the book's co-author. I decided to use no text either for my own preparations or as a prop for students. Instead, I decided to present the material the way I understand it, with not only descriptions and examples, but also proofs, coming directly from my head. A mitigating factor was that, although sufficiently broadly educated, I am not a professional topologist or geometer. Hence, the stuff I had ready in my head or could easily reconstruct should not have been too obscure or overly challenging.

So, this is how the book came about. I prepared each lecture (usually without or with minimal written notes), and my TA, the third year Ph.D. student Vaughn Climenhaga, took notes and within 24 hours, usually less, prepared a very faithful and occasionally even somewhat embellished version typed in TeX. I usually did some very light editing before posting each installation for the students. Thus, the students had the text growing in front of their eyes in real time.

By the end of the Fall semester the notes were complete: additional work involved further editing and, in a few cases, completing and expanding proofs; a slight reordering of material to make each chapter consist of complete lectures; and in a couple of cases, merging two lectures into one, if in class a considerable repetition appeared. But otherwise the book fully retained the structure of the original one-semester course, and its expansion is due to the addition of a large number of pictures, a number of exercises (some were originally given in separate homework sets, others added later), and some "prose", i.e. discussions and informal explanations. All results presented in the book appeared in the course, and, as I said before, only in a few cases did proofs need to be polished or completed.

Aside from creating the original notes, my co-author Vaughn Climenhaga participated on equal terms in the editorial process, and, very importantly, he produced practically all of the pictures, including dozens of beautiful 3-dimensional images for which, in many cases, even the concept was solely his. Without him, I am absolutely sure that I would not have been able to turn my course into a book in anything approaching the present timeframe, and even if I did at all, the quality of the final product would have been considerably lower.

Anatole Katok

Chapter 1

Various Ways of Representing Surfaces and Basic Examples

Lecture 1

a. First examples. For many people, one of the most basic images of a surface is the surface of the Earth. Although it looks flat to the naked eye (at least in the absence of any striking geographic features), we learn early in our lives that it is in fact round, and that its shape is very well approximated by a sphere. Geometrically, the sphere is defined as the locus of points at a fixed distance, called the *radius*, from a given point, the centre. Using Cartesian coordinates and putting the origin at the centre, we derive the familiar equation

$$x^2 + y^2 + z^2 = R^2, \tag{1.1}$$

where R is the radius; the sphere is the set of all points in $\mathbb{R}^3$ whose coordinates (x, y, z) satisfy this equation.

Many other familiar shapes can also be defined geometrically and represented as the set of solutions of a single equation, as in (1.1). For example, the (round) cylinder is the locus of points at a fixed distance from a given straight line. If the line is taken to be the z-axis and the

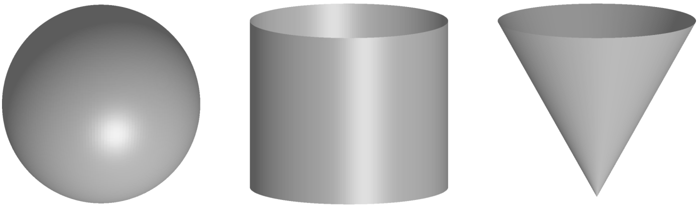


Figure 1.1. Three familiar surfaces.

distance is equal to R, the equation for the cylinder is

$$x^2 + y^2 = R^2. \tag{1.2}$$

Another surface familiar from elementary geometry (and also from ice-cream parlours) is the cone, which is obtained by rotating a straight line around another line which intersects it. If the axis of rotation is again the z-axis and the initial line lies in the xz-plane, with the equation $x = az$, then the cone is given by the equation

$$x^2 + y^2 = a^2 z^2. \tag{1.3}$$

Exercise 1.1. If we construct a surface of revolution using parallel lines instead of intersecting lines (as we did with the cone), we obtain a cylinder. There is a third possibility; the lines may be *skew*, that is, neither intersecting nor parallel. Describe the surface obtained in this case, and derive its equation.

We feel immediately that the three objects expressed by equations (1.1), (1.2), and (1.3), which are shown in Figure 1.1, are very different in a variety of robust ways. For example, the sphere is bounded—in fact, compact—while the cylinder and cone are not (contrary to what the picture might suggest). The sphere and cylinder are smooth everywhere, while the cone has a special point, the intersection of the two lines in the construction, which is the origin in (1.3).

These differences are qualitative, and would not be changed if we deformed each surface by a small amount—this reflects the fact that the three surfaces in question have different *topologies*. Such a deformation would, however, change the quantitative properties of a surface, which constitute its *geometry*. For example, stretching or

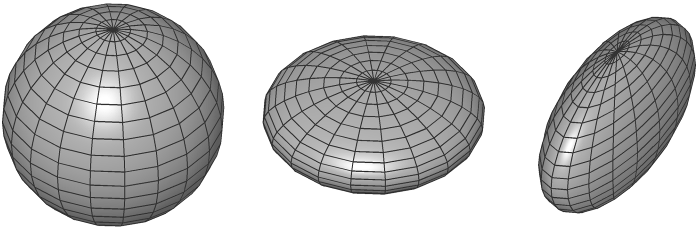


Figure 1.2. Three ellipsoids.

squeezing the sphere along the three coordinate axes produces an ellipsoid given by the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1, \tag{1.4}$$

where a, b, and c are parameters which depend on the degree of stretching or squeezing. Of the three surfaces above, the overall shape and crude properties of an ellipsoid (its topology) are most similar to that of a sphere, and are quite different from that of a cylinder or a cone; its geometry, however, displays many differences from the geometry of a sphere.[1] For example, the sphere has many symmetries (that is, rigid motions of the space which leave the sphere as a whole in place), while a triaxial ellipsoid (one for which all three numbers a, b, and c in (1.4) are different, such as the third shape shown in Figure 1.2) has only a few.

Exercise 1.2. Find all the symmetries for

(1) a triaxial ellipsoid;

(2) an ellipsoid of revolution for which $a = b \neq c$ (such as the second ellipsoid in Figure 1.2).

Consider separately the symmetries which can be effected by a continuous motion of the space and those which cannot, such as reflections with respect to planes.

[1] For the time being, we rely on intuitive ideas of what constitutes a general shape. For a reader steeped in mathematical rigor, we refer to notions of homeomorphism and diffeomorphism, which will be introduced later in Lectures 4 and 17, respectively, and say that two surfaces have similar shapes if they are homeomorphic, or diffeomorphic in the case of smooth surfaces.

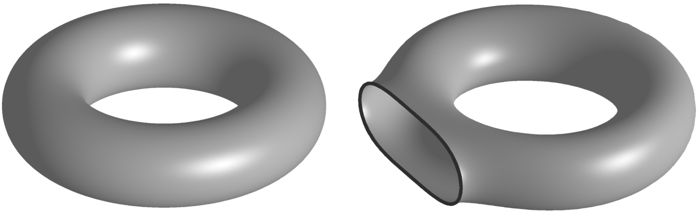


Figure 1.3. A torus and a handle.

Another familiar example of a surface is a torus—just as the sphere is the surface of an idealised ball, the torus is the surface of an idealised doughnut (or perhaps a bagel, depending on what sort of diet one is on). Like our first three examples, it is a surface of revolution, and may be obtained by rotating a circle around a line which lies in the plane of the circle, but does not intersect it. We will derive a nice equation (1.5) for the torus in the next lecture.

We can obtain new surfaces with qualitatively distinctive shapes by the procedure called "attaching a handle". A handle can be thought of as a torus with a hole (or if you like, an inner tube with a small patch cut out), as shown in Figure 1.3—this is attached to a hole cut in a given surface. Applying this procedure to a sphere produces a surface in the general shape of a torus. If we continue to attach more handles, we obtain something reminiscent of a pretzel with an increasing number of holes or, alternatively, a chain of tori linked to each other—Figure 1.4 shows a sphere with two handles. Like all the surfaces we have dealt with so far, these surfaces can also be represented by equations with a certain amount of effort (see Exercise 1.6).

b. Equations vs. other methods. We have obtained several different surfaces as the set of points whose coordinates (x, y, z) satisfy one equation or another. It is natural to ask what sort of equations will always yield nice, recognisable surfaces. Will any old equation do? Or must we impose some restrictions? And conversely, can we represent every surface by an equation?

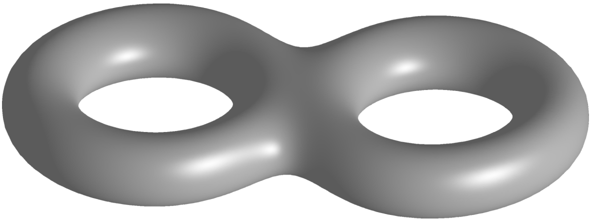


Figure 1.4. A sphere with two handles.

We begin by asking what sorts of equations are acceptable. By moving all the terms to the same side, any equation in x, y, and z can be written in the form $F(x, y, z) = 0$. If we hope to get a smooth surface, we must demand that the function F be at least differentiable—any of the equations (1.1), (1.2), (1.3), and (1.4) can be written in this form with a quadratic polynomial as the function F. But why are the sphere, the cylinder, and the ellipsoid all smooth, while the cone has a special point? The difference is clearly seen in the geometric description of the surfaces, since the line we use to define the cone passes through the axis of rotation, but it is not so easy to see what feature of the equations is responsible. How does this point of non-smoothness turn up in the equations?

The answer is that the origin is a *critical point* of the function $x^2+y^2-a^2z^2$ and lies on the surface defined by (1.3), while the other functions, $x^2 + y^2 + z^2 - R^2$, $x^2 + y^2 - R^2$, and $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} - 1$, have no critical points at the zero level. Thus, if we want to define a smooth surface in $\mathbb{R}^3$ by an equation of the form $F(x, y, z) = 0$, the function F should have no critical points at the zero level.

Turning to the other half of the relationship between surfaces and equations, we find that not every geometric object which common sense would call a surface can be represented as the solution set of an equation. One difficulty is caused by boundaries—notice that the cylinder defined in (1.2) is unbounded, and extends infinitely far in both the positive and negative z-directions. Suppose we want to consider a finite cylinder, which may be obtained by rotating an interval around a parallel line, or by rolling up a rectangular sheet of paper

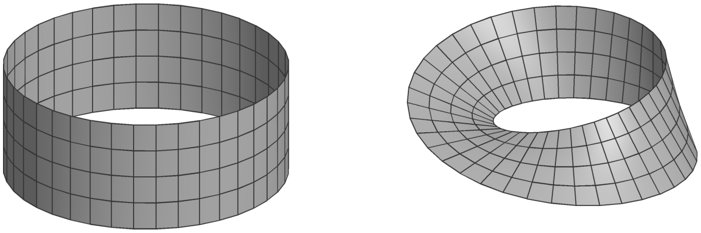


Figure 1.5. Two ways of gluing ends together.

and gluing together two opposite edges. How are we to represent such a surface by an equation?

One possibility is to add an auxiliary inequality—for example, one particular bounded cylinder is given as the solution set of

$$x^2 + y^2 = R^2, \qquad z^2 \leq 1.$$

This method solves the problem in some cases, but not all. Consider the second description of a cylinder given above, in which we take a band of paper and glue together the two ends—now look at what happens if we twist the band halfway around before gluing the ends together! The result is the famous *Möbius band* (or *Möbius strip*), shown in Figure 1.5. Its most surprising property is that it only has one side: an insect which crawls once around the band will find itself at the same place, but on the opposite side of the surface.

Now any surface which is given by an equation $F(x, y, z) = 0$ (with or without inequalities) and which does not contain any critical points must have two sides—the function F is positive on one side and negative on the other. It follows that the Möbius strip cannot be represented as the solution set of a 'nice' equation in the sense discussed above.

A related counterintuitive property of the Möbius strip has to do with closed curves. In the plane, any closed curve divides the plane into two regions[2]—on the Möbius strip, though, we can draw closed curves which have no "inside" or "outside". Consider the curve which divides the strip in half, so to speak, running halfway between the free

[2] This if the *Jordan Curve Theorem*, which we will state and prove rigorously in Lectures 34 and 35. It is not as easy as one might first think!

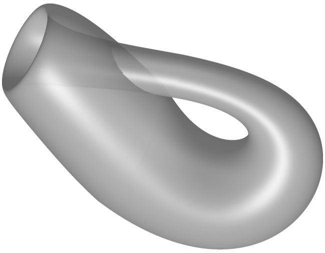


Figure 1.6. Immersing a Klein bottle in $\mathbb{R}^3$.

edges. If we take a pair of scissors and cut along this curve, we will be left with a single connected surface, rather than two disconnected pieces, which is what would happen if we performed the same operation on the cylinder, for example. This fact is intimately connected to the observation that if we place a clock at some point on this curve and move it once around the strip, when it returns it will be running counterclockwise!

The existence of the Möbius strip is the first indication that representing surfaces by equations is not sufficient. In the next lecture we will discuss an alternative way of representing it in an analytical fashion. Notice, however, that the Möbius strip, along with all our other examples, still lives comfortably in three-dimensional Euclidean space. Our next example challenges the assumption that all interesting surfaces can be realised this way.

If we want to glue together two opposite sides of a rectangle, we can either glue them with no twist, which produces a cylinder, or with a half-twist, which produces a Möbius strip.[3] A similar dichotomy arises if we decide to glue together the two ends of a cylinder. If we do this in the conventional way, we produce a torus—however, this is only one of two possible alignments for the pair of circles which are to be attached. The second possibility involves 'flipping' one of the ends around somehow, and results not in a torus, but in a *Klein bottle.* The closest we can come to visualising this in three dimensions is to have one end approach the other end not from outside the cylinder,

[3] A second half-twist will produce something which turns out to be homeomorphic to a cylinder, but with a different embedding in $\mathbb{R}^3$.

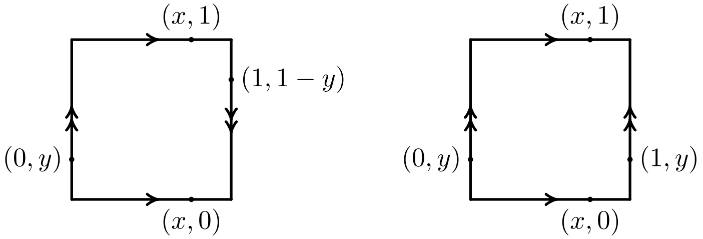


Figure 1.7. Planar models of a Klein bottle and a torus.

as with the torus, but from *inside*—to accomplish this, we must pass the end through the wall of the cylinder, creating a sort of twisted bottle (hence the name), as shown in Figure 1.6.

c. Planar models. Unlike the earlier examples, the Klein bottle cannot be embedded in $\mathbb{R}^3$, and so it is more difficult to represent properly. Abstractly, however, the procedure we followed to create it is not hard to describe, and this idea introduces a totally different way of looking at surfaces. We begin by taking the unit square for our rectangle:

$$X = \{ (x, y) \in \mathbb{R}^2 \mid 0 \leq x \leq 1,\ 0 \leq y \leq 1 \}.$$

We may then 'glue' together two opposite edges by declaring that for each value of x between 0 and 1, the pair of points $(x, 0)$ and $(x, 1)$ are now the same point. This gives an abstract representation of the cylinder—to obtain a Klein bottle, we must 'glue' together the two remaining edges with a flip.[4] We do this by considering each pair of points $(0, y)$ and $(1, 1 - y)$ as a single point—notice that all four corners are now identified. One easily checks that a piece of this object near every point looks like a piece of ordinary plane, so this seems to be a legitimate surface.[5]

Now we can look at the procedure just described and contemplate what happens when we identify both pairs of sides of the square in the conventional way: $(x, 0)$ with $(x, 1)$ and $(0, y)$ with $(1, y)$. We

[4] These edges are now "circles", in the topological sense at least, since $(0, 0)$ and $(0, 1)$ are the same point, and similarly for $(1, 0)$ and $(1, 1)$.

[5] Of course, we have not defined rigorously what we mean by a 'legitimate surface'. A two-dimensional smooth manifold (see Lecture 16) certainly qualifies.

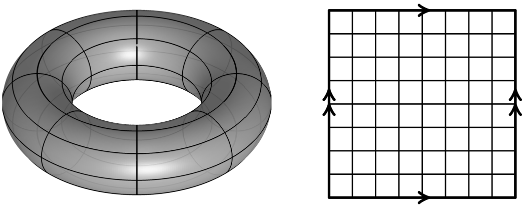


Figure 1.8. Meridians and parallels on two tori with different geometries.

obtain a surface resembling a torus as far as its global properties are concerned. For example, vertical and horizontal segments become closed curves which are identified with "parallels" and "meridians" of the torus of revolution—this will become clear in the next lecture when we introduce parametric representations of surfaces. However, the geometry of our surface, the *flat torus*, is different from that of the torus of revolution. For example, all vertical and all horizontal "circles" in the flat torus have the same length, while in the torus of revolution the meridians have the same length but the parallels do not (Figure 1.8). This is a consequence of the fact that although the cylinder in $\mathbb{R}^3$ has the same intrinsic geometry as the sheet of paper with only one pair of sides identified (that is, the paper is not stretched), it cannot be bent in $\mathbb{R}^3$ without a distortion. So far, our notion of intrinsic geometry is intuitive, but soon we will make it more precise.

Let us try to exhaust the possibilities of surface-building from a rectangular piece of paper. The only remaining way of identifying pairs of opposite sides is to identify both pairs of sides using a flip, so that we identify $(x, 0)$ with $(1 - x, 1)$ and $(0, y)$ with $(1, 1 - y)$. We will now turn our attention to this construction.

Exercise 1.3. Describe the surface obtained from the square by identifying points on pairs of adjacent sides, i.e. $(0, t)$ with $(1 - t, 1)$ and $(1, t)$ with $(1-t, 0)$. Pay attention both to the shape and to geometry.

d. Projective plane and flat torus as factor spaces. To get a more symmetric picture for the last construction, we may inflate the square to a disc into which the square is inscribed, project the boundary of the square radially to the circumference of the disc, and observe

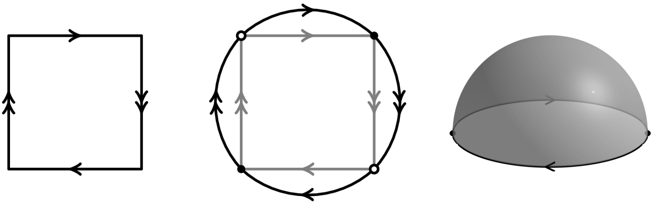


Figure 1.9. Various models for the real projective plane.

that the identified pairs become antipodal points on the boundary circle. Thus our object becomes the disc with pairs of opposite points on the boundary identified, as in Figure 1.9. To make this even more symmetric, inflate the disc to a hemisphere, keeping the boundary as the equator. Now we can add the other hemisphere and observe that each point of our object is represented by a pair of opposite points on the sphere.

Instead of taking pairs of antipodal points as the points of our surface, we may observe that any such pair determines a unique line in $\mathbb{R}^3$ passing through the centre of the sphere, and vice versa. Thus we may also think of our surface as the set of all lines through a particular point—the surface so obtained is called the *projective plane*, denoted $\mathbb{R}P^2$. An obvious advantage of the sphere representation over gluing is that it highlights the uniformity of the surface; all points look the same.

Inspired by the last construction, we may try to look at the flat torus differently. First recall that the circle can be represented either by an interval, say $[0, 1]$, with endpoints identified, or as the set of equivalence classes of real numbers modulo one, i.e. the set of all fractional parts of real numbers. If we simply think of all numbers with the same fractional part as the same element of the circle we come to the representation $S^1 = \mathbb{R}/\mathbb{Z}$—note that here every point on the circle is represented in the same way, in contrast to the interval with endpoints identified, where the choice of representation led to a false distinction between endpoints and non-endpoints. This choice of representation is a matter of fixing a *fundamental domain*; that is, a subset of $\mathbb{R}$ which contains exactly one element of each equivalence class, except along its boundary, where it may contain two or more. In this case, we may take any unit interval as our fundamental domain.

A similar observation may be made with two variables, where we observe that the (flat) torus $\mathbb{T}^2$ can be identified with the set of pairs of fractional parts of real numbers:

$$\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2,$$

where $\mathbb{Z}^2$ is the lattice of vectors with integer coordinates. These equivalence classes are represented by points in the unit square (the fundamental domain), once pairs of boundary points whose difference is an integer have been identified.

We may make one further step into abstraction; instead of vectors with integer coordinates, think about translations by those vectors. Then each equivalence class in $\mathbb{R}^2/\mathbb{Z}^2$ becomes an orbit of the group of such translations acting on $\mathbb{R}^2$, and our factor space (or *quotient space*) naturally becomes the space of orbits.

The same approach may be taken with the projective plane—notice that the flip on the sphere is a transformation which generates a group of two elements, since its square is the identity. The orbit of a point under the action of this group consists of the point itself, together with its antipode—identifying each such pair of points yields the projective plane, which can thus be thought of as the space of orbits of this two-element group acting on the sphere.

Exercise 1.4. Represent the cylinder, the infinite Möbius strip, and the Klein bottle as orbit spaces for some groups acting on the Euclidean plane $\mathbb{R}^2$. The infinite Möbius strip is the infinite rectangle $[0,1] \times \mathbb{R}$ with each pair of points $(0,y)$ and $(1,-y)$ identified.

Lecture 2

a. Equations for surfaces and local coordinates. Consider the problem of writing an equation for the torus; that is, finding a function $F\colon \mathbb{R}^3 \to \mathbb{R}$ such that the torus is the solution set $\{(x,y,z) \in \mathbb{R}^3 \mid F(x,y,z) = 0\}$. Because the torus is a surface of revolution, we begin with the equation for a circle in the xz-plane with radius 1 and centre at $(2,0)$:

$$S^1 = \{\, (x,z) \in \mathbb{R}^2 \mid (x-2)^2 + z^2 = 1 \,\}.$$

To obtain the surface of revolution, we replace x with the distance from the z-axis by making the substitution $x \mapsto \sqrt{x^2+y^2}$, and obtain

$$\mathbb{T}^2 = \left\{ (x,y,z) \in \mathbb{R}^3 \mid (\sqrt{x^2+y^2}-2)^2 + z^2 - 1 = 0 \right\}.$$

At first glance, then, setting $F(x,y,z) = (\sqrt{x^2+y^2}-2)^2 + z^2 - 1$ gives our desired solution. However, this suffers from the defect that F is not differentiable along the z-axis; we can overcome this fairly easily with a little algebra. Expanding the equation, isolating the square root, and squaring both sides, we obtain

$$\begin{aligned} x^2 + y^2 + 4 - 4\sqrt{x^2+y^2} + z^2 - 1 = 0,& \\ x^2+y^2+z^2+3 &= 4\sqrt{x^2+y^2}, \\ (x^2+y^2+z^2+3)^2 &= 16(x^2+y^2), \end{aligned}$$

and hence consider the function F defined by

$$F(x,y,z) = (x^2+y^2+z^2+3)^2 - 16(x^2+y^2). \tag{1.5}$$

It is easy to check that the new choice of F from (1.5) does not introduce any extraneous points to the solution set, and now F is differentiable on all of $\mathbb{R}^3$.

Exercise 1.5. Prove that a sphere with $m \geq 2$ handles cannot be represented as a surface of revolution.

Due to the result in Exercise 1.5, this argument cannot be applied directly to find an equation whose set of solutions look like a sphere with $m \geq 2$ handles, but we can reverse engineer the result to find a general method. Instead of beginning with a vertical plane, we consider the intersection of the torus and the horizontal xy-plane, which is given by two concentric circles. $F(x,y,0)$ is negative between the circles, hence $F(x,y,z) = F(x,y,0) + z^2 = 0$ has two solutions for those values of x and y, leading to the torus shape. By beginning with three or more circles (no longer concentric) we may use this idea to represent a sphere with any number of handles.

Exercise 1.6. Represent a sphere with two handles as the set of solutions of the equation $F(x,y,z) = 0$, where F is a differentiable function, and none of its critical points satisfy this equation.

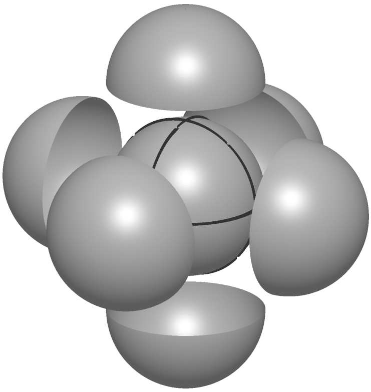


Figure 1.10. The sphere as a union of graphs.

What good is all this? What benefit do we gain from representing the torus, or any other surface, by an equation? Of course, it allows us to plug the equation into a computer and look at pretty pictures of our surface, but what we are really after is *coordinates* on our surface. After all, the surface is a two-dimensional affair, and so we should be able to describe its points using just two variables, but the equations we obtain are written in three variables.

To address this, we first backtrack a bit and discuss graphs of functions. Recall that given a function $f\colon \mathbb{R}^2 \to \mathbb{R}$, the graph of f is

$$\text{graph } f = \{\, (x,y,z) \in \mathbb{R}^3 \mid z = f(x,y) \,\}.$$

If f is 'nice', its graph is a 'nice' surface sitting in $\mathbb{R}^3$. Of course, most surfaces cannot be represented globally as the graph of such a function; the sphere, for instance, has two points on the z-axis, and hence we require at least two functions to describe it in this manner.

In fact, more than two functions are required if we adopt this approach. The unit sphere is given as the solution set of $x^2+y^2+z^2 = 1$, so we can write it as the union of the graphs of f_1 and f_2, where

$$f_1(x,y) = \sqrt{1-x^2-y^2},$$
$$f_2(x,y) = -\sqrt{1-x^2-y^2}.$$

The graph of f_1 is the northern hemisphere, and the graph of f_2 is the southern. However, we run into problems at the equator $z = 0$; for reasons which will be made apparent when we give the precise definition of a manifold (topological or differentiable), it is important that the domain on which we define each graph be *open*. In this particular case, this means we cannot include the equator in either the northern or the southern hemisphere, and must cover those points with other graphs. By using graphs with x or y as the dependent variable, we can cover the 'eastern' and 'western' hemispheres, as it were, but find that we require six graphs to deal with the entire sphere, as shown in Figure 1.10.

This approach has wide validity. Recall that $(x, y, z) \in \mathbb{R}^3$ is a *critical point* of a smooth function $F\colon \mathbb{R}^3 \to \mathbb{R}$ if the gradient of F vanishes at (x, y, z), and that a point is called *regular* if it is not critical. If S is the zero set of such a function, then at any regular point in S we can apply the Implicit Function Theorem and obtain a neighbourhood of the point which is the graph of some function; in essence, we are projecting patches of our surface to the various coordinate planes in $\mathbb{R}^3$. If our surface contains only regular points, this allows us to describe the entire surface in terms of these local coordinates.

As indicated in the first lecture, if the gradient vanishes at a point, the set of solutions may not look like a nice surface. A trivial example is the sphere of radius zero, $x^2 + y^2 + z^2 = 0$; a more interesting example is the cone $x^2 + y^2 - z^2 = 0$ near the origin.

b. Other ways of introducing local coordinates. From the geometric point of view, the choice of planes involved in representing a surface as the union of graphs of functions is somewhat arbitrary and unnatural; for example, the orthogonal projection of the northern hemisphere of S^2 to the xy-plane represents points in the 'arctic' quite well, but distorts things rather badly near the equator, where the derivative of the function blows up. If we are interested in angles, distances, and other geometric qualities of the surface, a more natural choice is to project to the tangent plane at each point; this will lead us eventually to the notion of a *Riemannian manifold*. If the previous approach represented an effort to draw a 'world map' of

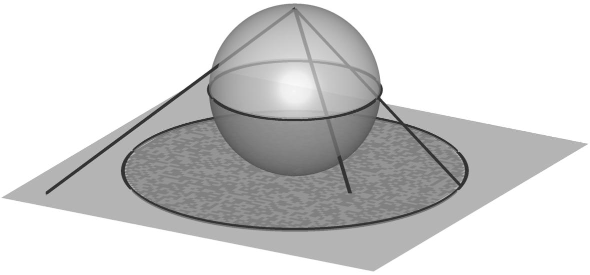


Figure 1.11. Stereographic projection from the sphere to the plane.

as much of the surface as possible, without regard to distortions near the edges, this approach represents publishing an atlas, with many smaller maps, each zoomed in on a small neighbourhood of each point in order to minimise distortions.

Orthogonal projections, whether to coordinate planes or tangent planes, form only a subset of the class of local coordinates on surfaces; there are many other members of this class besides. In the case of a sphere, one well-known example of local coordinates is stereographic projection (Figure 1.11), which gives a diffeomorphism[6] from the sphere minus a point to the plane.

Another example is given by the use of the familiar system of longitude and latitude to locate points on the surface of the earth; these resemble polar coordinates, mapping the sphere minus a point onto the open disc (Figure 1.12). The north pole is the centre of the disc, while the (deleted) south pole is its boundary; lines of longitude (meridians) become radii of the disc, while lines of latitude (parallels) become concentric circles around the origin.

However, if we want to measure distances on the sphere using any of these local coordinates, we cannot simply use the usual Euclidean distance in the disc or the plane—for example, the polar coordinates mentioned in the last example preserve distances along lines of longitude (radii), but distort distances along lines of latitude (circles centred at the origin). This is especially true near the boundary of the

[6] That is, a bijective differentiable map with differentiable inverse. See Lecture 17 for more details.

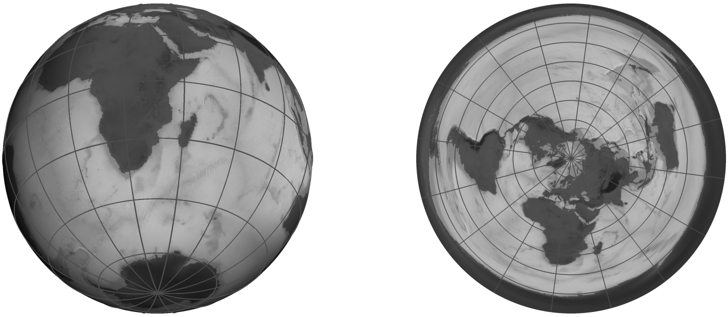


Figure 1.12. From the sphere to a disc via geographic coordinates.

disc, where the actual distance between points is much less than the Euclidean distance (since every point on the boundary is identified)—notice how much Antarctica is stretched out in Figure 1.12. This gives us our first example of a *Riemannian metric* (which for the time being we may simply think of as a notion of distance) on $\mathbb{D}^2$, apart from the usual Euclidean one.

Exercise 1.7. Stereographic projections from the north and south poles introduce two coordinate systems on the sphere minus the poles. Find the coordinate transformation from one of those systems to the other—that is, if a point on the sphere has coordinates (x, y) in the coordinate system projected from the north pole and (x', y') in the projection from the south, find (x', y') as a function of (x, y).

c. Parametric representations. While the idea of putting local coordinates on a surface will turn out to be more useful in general, we will occasionally have reason to deal with parametric representations. There are two important distinctions between these two methods of introducing coordinates on a surface.

First, local coordinates involve a map from the surface to a plane domain, while a parametric representation is a map from a plane domain to the surface. Formally, then, these two constructions are mutual inverses.

The second distinction is that a local coordinate system usually does not attempt to cover the entire surface by a single coordinate system, but rather uses several patches to accomplish the task. A parametric representation, on the other hand, usually involves a map from a plane domain to a surface which is onto, or at at least nearly so, as in the inverse to the stereographic projection. One should also keep in mind that, while the notion of an atlas of local coordinate systems has a precise meaning which we will describe in Chapter 3, the notion of parametric representation is somewhat vague.

Exercise 1.8. Write a parametric representation of the torus of revolution (1.5) using the 'latitude' (position of a plane section) and 'longitude' (the angular coordinate along a plane section) as parameters. Use this representation to construct a bijection between the flat torus from Lecture 1(d) and the torus of revolution.

d. Metrics on surfaces. As our discussion of local coordinates suggested, we must address the question of how the distance between two points on a surface is to be measured. In the case of the Euclidean plane, we have a formula, obtained directly from the Pythagorean theorem. For points on the sphere of radius R we also have a formula: the distance between two points is simply the angle they make with the centre of the sphere, multiplied by R. Properties of this distance, such as the triangle inequality, can be deduced via elementary geometry, or by representing the points as vectors in $\mathbb{R}^3$ and using properties of the inner product.

These explicit formulae are serendipitous consequences of the extremely symmetric shapes of the plane and the sphere. What is the correct notion of distance on an arbitrary surface? Recalling that in the plane at least, the shortest path between two points is a straight line, and it is precisely along this line that the distance given by the Pythagorean theorem is measured, we may suggest that the distance between two points should naturally be defined as the length of the shortest path connecting them.

In general, since we do not yet know whether such a shortest path always exists, the proper definition of distance is as the infimum of the set of lengths of paths connecting the two points. Of course, this requires that we have a definition for the length of a path on the surface. We can find the length of a path in $\mathbb{R}^3$ by approximating it with piecewise linear paths and then using the notion of distance in $\mathbb{R}^3$, which we already know. If our surface is not embedded in Euclidean space, however, we must replace this with an infinitesimal notion of distance, the Riemannian metric alluded to above. We will give a precise definition and discuss examples and properties of such metrics later in this course.

Lecture 3

a. More about the Möbius strip and projective plane. Let us go back to the Möbius strip. The most common way of introducing it is as a sheet of paper (or belt, carpet, etc.) whose ends have been attached after giving one of them a half-twist. In order to represent this surface parametrically, it is useful to consider the factor space construction, which was discussed in the first lecture for the Klein bottle and the flat torus, and which is even simpler in the case of the Möbius strip.

Begin with a rectangle R. We are going to identify each point on the left-hand vertical boundary of R with a point on the right-hand boundary; if we identify each point with the point directly opposite to it (on the same horizontal line), we obtain a cylinder. To obtain the Möbius strip, we identify the lower left corner with the upper right corner and then move inwards; in this fashion, if $R = [0,1] \times [0,1]$, the point $(0,t)$ is identified with the point $(1,1-t)$ for $0 \leq t \leq 1$.

To embed this in $\mathbb{R}^3$, we can effect the half-twist by a continuous uniform rotation of an interval (the vertical lines in the model) whose centre moves around a closed curve (say a circle), and which remains perpendicular to that circle. Using the x-coordinate in the model as the angular coordinate along the circle, and the y-coordinate as the distance along the interval, one can write a parametric representation of a Möbius strip in $\mathbb{R}^3$ (see Figure 1.5).

Exercise 1.9. Write explicit expressions for the parametric representation of a Möbius strip embedded into $\mathbb{R}^3$ without self-intersections described above.

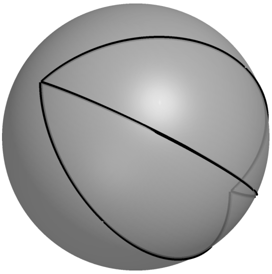


Figure 1.13. Multiple geodesics between antipodal points.

The projective plane with distance inherited from the sphere[7] is called the *elliptic plane*—it will be one of the star exhibits of this course. We can motivate its definition by considering the sphere as a geometric object, on which the notion of a line in Euclidean space is to be replaced by the concept of a *geodesic*; one key property of the former is that it is the shortest path between two points, and so informally at least, geodesics are simply curves which have this property. On the sphere, we will see that the geodesics are great circles, and so we may attempt to formulate various geometric propositions in this setting. However, this turns out to have some undesirable features from the point of view of conventional geometry; for example, every pair of geodesics intersects in *two* (diametrically opposite) points, not just one. Further, any two diametrically opposite points on the sphere can be joined by infinitely many geodesics (Figure 1.13), in stark contrast to the "two points determine a unique line" rule of Euclidean geometry.

Both of these difficulties are related to pairs of diametrically opposite points; the solution turns out to be to identify such points with each other. Identifying each point on the sphere with its antipode yields a quotient space, which is the projective plane described at the end of the first lecture. Alternatively, we can consider the flip map $I\colon (x,y,z) \mapsto (-x,-y,-z)$, which is an isometry of the sphere without fixed points. Declaring all members of a particular orbit of I to

[7] This simply means that the distance between two points in the projective plane is taken to be the minimum of pairwise distances between points in the sphere representing those points.

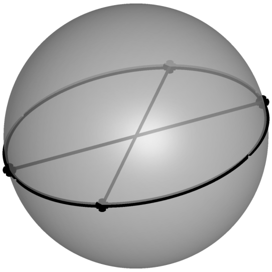


Figure 1.14. Determining distances in $\mathbb{R}P^2$ via central angles.

be the same point, we obtain the quotient space S^2/I, which is again the projective plane, or the elliptic plane when we are interested in the geometry.

In the elliptic plane, there is no such notion as the sign of an angle; we cannot consistently determine which angles are positive and which are negative. All the other geometric notions carry over, however; the distance between two points can still be found as the magnitude of the (acute) central angle they make (Figure 1.14), and the notions of angle between geodesics and length of geodesics are still well defined.

Exercise 1.10. Write at least five propositions from Euclidean geometry which are true in the elliptic plane and at least three propositions which are true in Euclidean geometry and are not true in the elliptic plane. Each proposition must include statements about configurations of lines and/or isometries, and no two should be trivial reformulations of each other.

b. A first glance at geodesics. Informally, as mentioned above, a *geodesic* is a curve of shortest length between two points; more precisely, it is a curve γ with the property that given any two points $\gamma(a)$ and $\gamma(b)$ whose parameter values a and b are sufficiently close together, any other curve from one point to the other will have length at least as great as the portion of γ between the two. Later in the course (Lecture 25), we will consider the question of whether such a curve always exists between two points, and whether it is unique.

The two most basic examples are the Euclidean spaces $\mathbb{R}^n$, where geodesics are straight lines, and the round sphere S^2, where geodesics

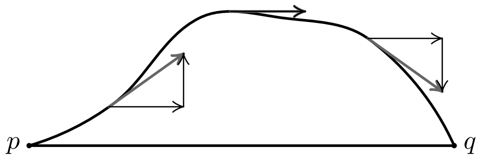


Figure 1.15. Decomposing tangent vectors to show that a straight line is the shortest smooth curve between two points.

are great circles. While the first fact is an article of faith in elementary geometry, it requires a proof using a certain amount of calculus. We will sketch the proof, but for a reader not familiar with calculations involving arbitrary curves, we recommend carrying out the argument in detail as an exercise.

Consider an arbitrary parametrised curve with endpoints p and q, and project it to the straight line pq. As a parametrised curve, the projection is no longer than the original curve—in fact, it is strictly shorter if the original curve does not lie entirely on the line.

If the curve is smooth, this follows from the formula for the length of the curve as the integral of the length of its tangent vector, which decomposes into two components, one parallel to the line pq, and one perpendicular (Figure 1.15). For an arbitrary curve, one can use an approximation by a polygonal curve—in either case, having established that the length of the original curve is greater than or equal to the length of the projected curve, one uses integration to show that the length of the projected curve is greater than or equal to the length of the interval pq, with equality if and only if the parameter is monotone (so that the curve is a reparametrised interval).

A very similar argument can be carried out on the sphere, using geographic coordinates around the point p and projection along parallels to the meridian (great circle) passing through p and q. In fact, once it is understood just what is needed for this argument, it can be adapted in many cases to find geodesics.

It is sometimes the case that one can find geodesics on other surfaces by reducing the question to a known situation. For example, the following exercise can be solved by reducing the question to the case of the Euclidean plane.

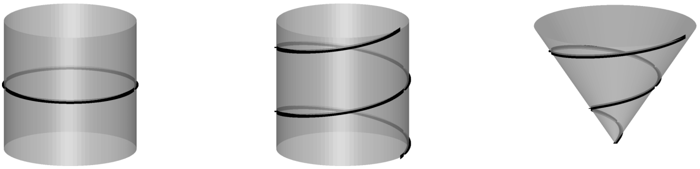


Figure 1.16. Three curves in $\mathbb{R}^3$.

Exercise 1.11. Find all geodesics on the round cylinder

$$\{\,(x,y,z)\in\mathbb{R}^3 \mid x^2+y^2=1\,\}$$

and the upper half of the round cone

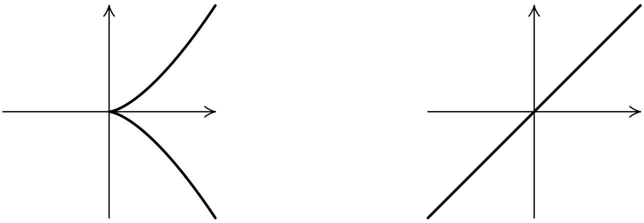
$$\{\,(x,y,z)\in\mathbb{R}^3 \mid x^2+y^2-z^2=0,\ z\geq 0\,\}.$$

c. Parametric representations of curves. We often write a curve in $\mathbb{R}^2$ as the solution of a particular equation; the unit circle, for example, is the set of points satisfying $x^2+y^2=1$. This implicit representation becomes more difficult in higher dimensions; in general, each equation we require the coordinates to satisfy will remove a degree of freedom (assuming independence) and hence a dimension, so to determine a curve in $\mathbb{R}^3$ we require not one, but two equations. Geometrically, we are obtaining a curve as the intersection of two surfaces, each specified by one of the equations. For example, the unit circle lying in the xy-plane is the solution set of

$$\begin{aligned} x^2+y^2&=1,\\ z&=0. \end{aligned}$$

which is the intersection of this plane with a cylinder of unit radius. This is a simple example for which these equations and the visualisation of the surfaces pose no real difficulty; there are many examples which are more difficult to deal with in this manner, but which can be easily written down using a *parametric representation.* That is, we define the curve in question as the set of all points given by

$$(x,y,z)=(f_1(t),f_2(t),f_3(t)),$$



$(x, y) = (t^2, t^3)$ $(x, y) = (t^3, t^3)$

Figure 1.17. Two curves with a vanishing tangent vector at $t = 0$.

where t lies in the interval $[a, b]$, whose endpoints a and b may be $\pm\infty$. In this representation, the circle discussed above would be written as

$$(x, y, z) = (\cos t, \sin t, 0)$$

with $0 \leq t \leq 2\pi$. If we replace the equation $z = 0$ with $z = t$, we obtain not a circle, but a helix; it takes a little more imagination to picture this as the intersection of two surfaces. We could also multiply the expressions for x and y by t to describe a spiral on the cone, whose implicit representation is again not immediate.

Exercise 1.12. Find two equations whose common solution set is the helix.

If we expect our curve to be smooth, we must impose certain conditions on the coordinate functions f_i. The first condition is that each f_i be continuously differentiable; this will guarantee the existence of a continuously varying tangent vector at every point along the curve. However, if we do not impose the further requirement that this tangent vector be non-vanishing, that is, that $(f_1')^2 + (f_2')^2 + (f_3')^2 \neq 0$ holds everywhere on the curve, then the curve may still fail to be smooth.

As a simple but important example of what may happen when this condition is violated, consider the curve $(x, y) = (t^2, t^3)$. The tangent vector $(2t, 3t^2)$ vanishes at $t = 0$, which appears as a *cusp* at the origin in Figure 1.17. So in this case, even though f_1 and f_2 are perfectly smooth functions, the curve itself is not smooth.

The non-vanishing condition is sufficient, but not necessary, to have a smooth curve; to see the latter, consider the curve $x = t^3$, $y = t^3$. The tangent vector vanishes when $t = 0$, but the curve itself is just the line $x = y$, which is as smooth as we could possibly ask for. In this case we could reparametrise the curve to obtain a parametric representation in which the tangent vector is everywhere non-vanishing.

d. Difficulties with representation by embedding. Parametric representations of curves (and surfaces as well), along with representations as level sets of functions (the implicit representations we saw before) all embed the curve or surface into an ambient Euclidean space, which so far has usually been $\mathbb{R}^3$. Our subsequent dealings have sometimes relied on properties of this ambient space; for example, the usual definition of the length of a curve relies on a broken line approach, in which the curve is approximated by a piecewise linear 'curve', whose length we can compute using the usual notion of Euclidean distance.

What happens, though, if our surface does not live in $\mathbb{R}^3$? We already touched upon this problem in Lecture 1(b), and now return to it in more depth, as $\mathbb{R}^3$ is not the proper setting for several of the surfaces we have seen so far. For example, $\mathbb{R}P^2$ cannot be embedded in $\mathbb{R}^3$, so if we are to compute the length of curves in the elliptic plane, we must either embed it in $\mathbb{R}^4$ or some higher-dimensional space, or else come up with a new definition of length, an issue to which we shall return in Lecture 23.

Our discussion of factor spaces in Lecture 1 was motivated by the example of the Klein bottle, which was defined as a factor space of the square, or rectangle, where the left and right edges are identified with direction reversed (as with the Möbius strip), but in addition, the top and bottom edges are identified (without reversing direction). We mentioned then that the Klein bottle cannot be embedded into $\mathbb{R}^3$, and that the closest one can come is to imagine rolling the square into a cylinder, then attaching the ends of the cylinder after passing one end through the wall of the cylinder into the interior.

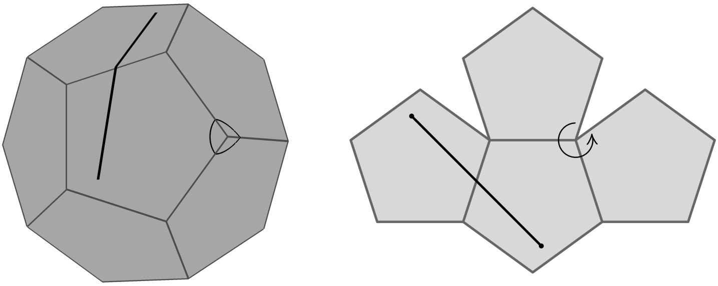


Figure 1.18. Life on a dodecahedron.

Of course, this results in the surface intersecting itself in a circle; in order to avoid this self-intersection, we could add a dimension and embed the surface in $\mathbb{R}^4$. Given the extra dimension to work with, we could begin with the immersion described above and perform the four-dimensional analogue of taking a string which is lying in a figure eight on a table, and lifting part of it off the surface of the table in order to avoid having it touch itself. No such manoeuvre is possible for the Klein bottle in three dimensions, but the immersion of the Klein bottle into $\mathbb{R}^3$ is still a popular shape, and some enterprising craftsman has been selling both 'Klein bottles' and beer mugs in the shape of Klein bottles at the yearly meetings of the American Mathematical Society. We had two such glass models of Klein bottles in the class, which were bought there: one is a conventional inverted bottle very similar to the image in Figure 1.6; the other is a "Klein beer mug", very close to a usual one in its outside shape and usable as a drinking vessel.

Even when an embedding exists, it is possible for the choice of embedding to obscure certain geometric properties of an object. Consider the surface of a dodecahedron (or any solid, for that matter). From the point of view of the embedding in $\mathbb{R}^3$, there are three kinds of points on the surface; a given point can lie either at a vertex, along an edge, or on a face. Being three-dimensional creatures, we see these as three distinct classes of points.

Now imagine that we are two-dimensional creatures living on the surface of the dodecahedron. We can tell whether or not we are at a vertex; at a vertex, the angles add up to less than 2π, whereas everywhere else, they add up to exactly 2π. However, we cannot tell whether or not we are at an edge; this has to do with the fact that given two points on adjacent faces, the way to find the shortest path between them is to unfold the two faces and place them flat on the plane (at which stage points on an edge look just like points on a face), draw a straight line between the two points in question, and then fold the surface back up (Figure 1.18). As far as our two-dimensional selves are concerned, points on an edge and points on a face are indistinguishable, since the unfolding process does not change any distances along the surface.

It is also possible that a surface which can be embedded in $\mathbb{R}^3$ will lose some of its nicer properties in the process. For example, the usual embedding of the torus destroys the symmetry between meridians and parallels; all of the meridians are the same length, but the length of the parallels varies. We can retain this symmetry by embedding in $\mathbb{R}^4$, the so-called *flat torus*. Parametrically, this is given by

$$\begin{aligned} x &= r\cos t, & y &= r\sin t, \\ z &= r\cos s, & w &= r\sin s, \end{aligned}$$

where $s, t \in [0, 2\pi]$. As we already mentioned, we can also obtain the flat torus as a factor space, using the same method as in the definition of the projective plane or Klein bottle. Beginning with a rectangle, we identify opposite sides (with no reversal of direction); alternately, we can consider the family of isometries of $\mathbb{R}^2$ given by $T_{m,n}\colon (x, y) \mapsto (x + m, y + n)$, where $m, n \in \mathbb{Z}$, and mod out by orbits. This construction of $\mathbb{T}^2$ as $\mathbb{R}^2/\mathbb{Z}^2$ is exactly analogous to the construction of the circle S^1 as $\mathbb{R}/\mathbb{Z}$.

We have seen that surfaces can be considered from different viewpoints: sometimes we treat them as geometric objects, with intrinsically defined distances, angles, and areas, while other times we treat them as ‘stretchable’ objects which can be bent and deformed, but not torn or broken. In mathematical language, this corresponds to considering different structures on surfaces, and this is the central theme of this course, which we will take up in earnest in the next lecture.

Before doing so, we would like to fix a linguistic ambiguity; for example, what should the word 'sphere' mean? How will we indicate whether we are treating a particular surface as a geometric object, or as a topological one (that is, one which may be deformed without changing the nature of the surface)? Our convention will be as follows: an indefinite article in front of the name, as in 'a sphere', 'a torus', or 'a projective plane', will mean that we consider the object in the topological sense, up to a homeomorphism. The use of an adjective or the definite article will generally signify a smaller class of objects, as in 'a sphere given by an equation'. Then 'a round sphere' would mean any sphere which has 'spherical geometry', that is, which is isometric to the actual sphere in Euclidean space. Similarly, 'a flat torus' signifies any torus with locally Euclidean geometry, while 'the flat torus' or 'the torus' will indicate the unit square with opposite sides identified, endowed with the appropriate geometry inherited from $\mathbb{R}^2$; sometimes we will call this object 'the standard flat torus'. 'The elliptic plane' indicates the factor space of the unit sphere in which antipodal points are identified, with geometry inherited from the sphere, and so on for various other examples which will arise.

Exercise 1.13. Write parametric representations for a projective plane in each of the following:

(1) $\mathbb{R}^3$ (with self-intersections).

(2) $\mathbb{R}^4$ (without self-intersections).

e. Regularity conditions for parametrically defined surfaces. A parametrisation of a surface in $\mathbb{R}^3$ is given by a region $U \subset \mathbb{R}^2$ with coordinates $(t, s) \in U$ and a set of three maps f_1, f_2, f_3; the surface is then the image of $F = (f_1, f_2, f_3)$, the set of all points $(x, y, z) = (f_1(t, s), f_2(t, s), f_3(t, s))$.

As with parametric representations of curves, we need a regularity condition to ensure that our surface is in fact smooth, without cusps or singularities. We once again require that the functions f_i be continuously differentiable, but now it is insufficient to simply require that the matrix of derivatives Df be non-zero. Rather, we require that it have maximal rank; the matrix is given by

$$Df = \begin{pmatrix} \partial_s f_1 & \partial_t f_1 \\ \partial_s f_2 & \partial_t f_2 \\ \partial_s f_3 & \partial_t f_3 \end{pmatrix}$$

and so our requirement is that the two tangent vectors to the surface, given by the columns of Df, be linearly independent. Under this condition, the Implicit Function Theorem guarantees that the parametric representation is locally bijective and that its inverse is differentiable.

Parametric representations may of course have singularities. A good example is the representation of the sphere given by the inverse map to the geographic coordinates, which maps an open disc regularly onto the sphere with a point removed, and collapses the boundary of the disc into this single point.

Lecture 4

a. Remarks on metric spaces and topology. Geometry in its most immediate form deals with measuring distances.[8] For this reason, *metric spaces* are fundamental objects in the study of geometry. In the geometric context, the distance function itself is the object of interest; this stands in contrast to the situation in analysis, where metric spaces are still fundamental (as spaces of functions, for example), but where the metric is introduced primarily in order to have a notion of convergence, and so the *topology* induced by the metric is the primary object of interest, while the metric itself stands somewhat in the background.

A metric space is a set X, together with a metric, or distance function, $d\colon X \times X \to \mathbb{R}_0^+$, which satisfies the following axioms for all values of the arguments:

(1) Positivity: $d(x, y) \geq 0$, with equality iff $x = y$.

(2) Symmetry: $d(x, y) = d(y, x)$.

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

[8] The reader should be aware, however, that in modern mathematical terminology, the word ‘geometry’ may appear with adjectives like ‘affine’ or ‘projective’. Those branches of geometry study structures which do not involve distances directly.

The last of these is generally the most interesting, and is sometimes useful in the following equivalent form:

$$d(x,y) \geq |d(x,z) - d(y,z)|.$$

Once we have defined a metric on a space X, we immediately have a topology on X induced by that metric. The *ball* in X with centre x and radius r is given by

$$B(x,r) = \{\, y \in X \mid d(x,y) < r \,\}.$$

Then a set $A \subset X$ is said to be *open* if for every $x \in A$, there exists $r > 0$ such that $B(x,r) \subset A$, and A is *closed* if its complement $X \setminus A$ is open. We now have two equivalent notions of convergence: in the metric sense, $x_n \to x$ if $d(x_n, x) \to 0$, while the topological definition requires that for every open set U containing x, there exist some N such that for every $n > N$, we have $x_n \in U$. It is not hard to see that these are equivalent.

Similarly for the definition of continuity; we say that a function $f\colon X \to Y$ is *continuous* if $x_n \to x$ implies $f(x_n) \to f(x)$. The equivalent definition in more topological language is that continuity requires $f^{-1}(U) \subset X$ to be open whenever $U \subset Y$ is open. We say that f is a *homeomorphism* if it is a bijection and if both f and f^{-1} are continuous.

Exercise 1.14. Show that the two sets of definitions (metric and topological) in the previous two paragraphs are equivalent.

Within mathematics, there are two broad categories of concepts and definitions with which we are concerned. In the first instance, we seek to fully describe and understand a particular kind of structure. We make a particular definition or construction, and then seek to either show that there is only one object (up to some appropriate notion of isomorphism) which fits our definition, or to give some sort of classification which exhausts all the possibilities. Examples of this approach include Euclidean space, which is unique once we specify dimension, or Jordan normal form, which is unique for a given matrix up to a permutation of the basis vectors, as well as finite simple groups, or semisimple Lie algebras, for which we can (eventually) obtain a complete classification.

No such uniqueness or classification result is possible with metric spaces and topological spaces in general; these definitions are examples of the second category of mathematical objects, and are generalities rather than specifics. In and of themselves, they are far too general to allow any sort of complete classification or universal understanding, but they have enough properties to allow us to eliminate much of the tedious case by case analysis, which would otherwise be necessary when proving facts about the objects in which we are really interested. The general notion of a group, or of a Banach space, also falls into this category of generalities.

Before moving on, there are three definitions of which we ought to remind ourselves. First, recall that a metric space is *complete* if every Cauchy sequence converges. This is not a purely topological property, since we need a metric in order to define Cauchy sequences; to illustrate this fact, notice that the open interval $(0,1)$ and the real line $\mathbb{R}$ are homeomorphic, but that the former is not complete, while the latter is.

Secondly, we say that a metric space (or subset thereof) is *compact* if every sequence has a convergent subsequence. In the context of general topological spaces, this property is known as sequential compactness, and the definition of compactness is given as the requirement that every open cover have a finite subcover; for our purposes, since we will be dealing with metric spaces, the two definitions are equivalent. There is also a notion of *precompactness*, which requires every sequence to have a *Cauchy* subsequence.

The knowledge that X is compact allows us to draw a number of conclusions; the most commonly used one is that every continuous function $f\colon X \to \mathbb{R}$ is bounded, and in fact achieves its maximum and minimum. In particular, the product space $X \times X$ is compact, and so the distance function is bounded.

Finally, we say that X is *connected* if it cannot be written as the union of non-empty disjoint open sets; that is, if $X = A \cup B$, with A and B open and $A \cap B = \emptyset$, implies either $A = X$ or $B = X$. There is also a notion of *path connectedness*, which requires for any two points $x, y \in X$ the existence of a continuous function $f\colon [0,1] \to X$ such that $f(0) = x$ and $f(1) = y$. As is the case with the two forms of compactness above, these are not equivalent for arbitrary topological spaces (or even for arbitrary metric spaces—the usual counterexample is the union of the graph of $\sin(1/x)$ with the vertical axis), but will be equivalent on the class of spaces with which we are concerned.

b. Homeomorphisms and isometries. In the topological context, the natural notion of equivalence between two spaces is that of homeomorphism, which we defined above as a continuous bijection with continuous inverse. Two topological spaces are *homeomorphic* if there exists a homeomorphism between them. Any property common to all homeomorphic spaces is called a *topological invariant*; this naturally includes any property defined in purely topological terms, such as connectedness, path-connectedness, and compactness.

Some invariants require a little more work; for example, we would like to believe that dimension is a topological invariant, and this is in fact true,[9] but proving that $\mathbb{R}^m$ and $\mathbb{R}^n$ are not homeomorphic for $m \neq n$ requires non-trivial tools.

A considerable part of this course deals with topological invariants of compact surfaces, and in particular, the task of classifying such surfaces up to a homeomorphism. We will almost succeed in solving this problem completely; the only assumption we will have to make is that the surfaces in question admit one of several natural additional structures. In fact this assumption turns out to be true for any surface, but we do not prove this in this course.

The natural equivalence relation in the geometric setting is isometry; a map $f\colon X \to Y$ between metric spaces is *isometric* if

$$d_Y(f(x_1), f(x_2)) = d_X(x_1, x_2)$$

for every $x_1, x_2 \in X$. If in addition f is a bijection, we say that f is an *isometry.* We are particularly interested in the set of isometries from X to itself,

$$\mathrm{Isom}(X, d) = \{\, f\colon X \to X \mid f \text{ is an isometry} \,\},$$

which we can think of as the symmetries of X. In general, the more symmetric X is, the larger this set.

[9] At least for the usual definition of dimension; we mention an alternate definition in the next section.

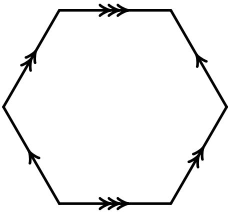


Figure 1.19. A planar model on a hexagon.

In fact, Isom(X, d) is not just a set; it has a natural binary operation given by composition, under which is becomes a group. This is an example of a very natural and general kind of group which is often of interest; all the bijections are from some fixed set to itself, with composition as the group operation. On a finite set, this gives the symmetric group S_n, the group of permutations. On an infinite set, the group of all bijections becomes somewhat unwieldy, and it is more natural to consider the subgroup of bijections which preserve a particular structure, in this case the metric structure of the space. Another common example of this is the general linear group $GL(n, \mathbb{R})$, which is the group of all bijections from $\mathbb{R}^n$ to itself preserving the linear structure of the space.

In the next lecture, we will discuss the isometry groups of Euclidean space and of the sphere.

Exercise 1.15. Consider a regular hexagon with pairs of opposite sides identified by the corresponding translations, as in Figure 1.19.

(1) Prove that it is a torus.

(2) Prove that locally, it is isometric to Euclidean plane.

(3) Prove that it is not isometric to the standard flat torus.

c. Other notions of dimension. As mentioned above, we usually think of dimension as a topological invariant. However, for general compact metric spaces there is another notion of dimension which is a metric invariant, rather than a topological one. The main idea is to capture the rate at which volume (or some other kind of measure) scales with the metric; for example, a cube in $\mathbb{R}^n$ with side length r has volume r^n, and the exponent n is the dimension of the space.

In general, given a compact metric space X, for any $\varepsilon > 0$, let $N(\varepsilon)$ be the minimum number of ε-balls required to cover X; that is, the minimum number of points $x_1, \dots, x_{N(\varepsilon)}$ in X such that every point in X lies within ε of some x_i. This may be thought of as measuring the average 'volume' of an ε-ball, in some sense; the *upper box dimension* of X is defined to be

$$\bar{d}_{\text{box}}(X) = \limsup_{\varepsilon \to 0} \frac{\log N(\varepsilon)}{\log 1/\varepsilon}.$$

We take the upper limit because the limit itself may not exist. The *lower box dimension* is defined similarly, taking the lower limit instead. These notions of dimension do not behave quite so nicely as we would like in all situations; for example, the set of rational numbers, which is countable, has upper and lower box dimension equal to one.

There is a more effective notion of *Hausdorff dimension*, which eliminates the need to distinguish between upper and lower limits, and which is equal to zero for any countable set; because its definition requires an understanding of measure theory, we will not discuss it here. For 'good' sets all three definitions coincide, and are central to the study of fractal geometry; however, they are not topological invariants, so our claim in the last section must be understood to apply only to a strictly topological notion of dimension.

d. Geodesics. When we are interested in a metric space as a geometric object, rather than as something in analysis or topology, it is of particular interest to examine those triples (x, y, z) for which the triangle inequality becomes degenerate, that is, for which $d(x, z) = d(x, y) + d(y, z)$.

For example, if our space X is just the Euclidean plane $\mathbb{R}^2$ with distance function given by Pythagoras' formula,

$$d((x_1, x_2), (y_1, y_2)) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2},$$

then the triangle inequality is a consequence of the Cauchy-Schwarz inequality, and we have equality in the one iff we have equality in the other; this occurs iff y lies in the line segment $[x, z]$, so that the three points x, y, z are in fact collinear.

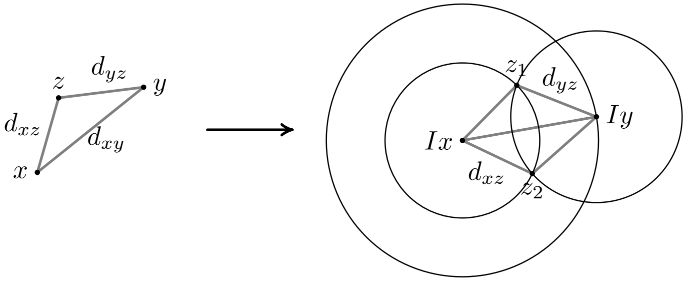


Figure 1.20. Images of three points determine an isometry.

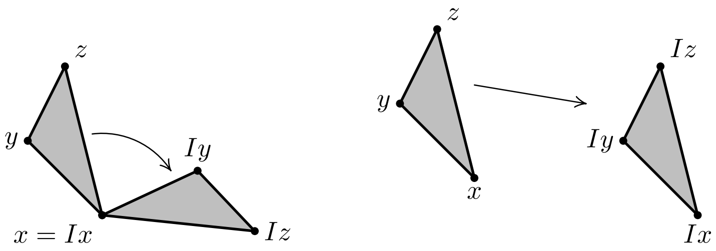
A similar observation holds on the sphere, where the triangle inequality becomes degenerate for the triple (x, y, z) iff y lies along the shorter arc of the great circle connecting x and z. So in both these cases, degeneracy occurs when the points lie along a geodesic; this suggests that in general, a characteristic property of a geodesic is the relation $d(x, z) = d(x, y) + d(y, z)$ whenever y lies between two points x and z which are sufficiently close along the curve.

Lecture 5

a. Isometries of the Euclidean plane. There are three ways to describe and study isometries of the Euclidean plane: synthetic; as affine maps in two real dimensions; and as affine maps in one complex dimension. The last two methods are closely related. We begin with observations using the traditional synthetic approach.

If we fix three non-collinear points in $\mathbb{R}^2$ and want to describe the location of a fourth, it is enough to know its distance from each of the first three. This may readily be seen from the fact that three circles whose centres are not collinear intersect in at most one point.

As a consequence of this, an isometry of $\mathbb{R}^2$ is completely determined by its action on three non-collinear points. In fact, if we have an isometry $I\colon \mathbb{R}^2 \to \mathbb{R}^2$, and three such points x, y, z, as in Figure 1.20, the choice of Ix constrains Iy to lie on the circle with centre Ix and radius $d(x, y)$, and once we have chosen Iy, there are only two possibilities for Iz; one (z_1) corresponds to the case where I preserves orientation, the other (z_2) corresponds to the case where



Rotation—one fixed point

Translation—no fixed points

Figure 1.21. Orientation preserving isometries.

orientation is reversed. So for two pairs of distinct points a, b and a', b' such that the distances between a and b and between a' and b' coincide, there are exactly two isometries which map a to a' and b to b'; one of these will be orientation preserving, the other orientation reversing.

Passing to algebraic descriptions, notice that any isometry I must carry lines to lines, since as we saw last time, three points in the plane are collinear iff the triangle inequality becomes degenerate. Thus it is an *affine map*—that is, a composition of a linear map and a translation—so it may be written as $I\colon x \mapsto Ax + b$, where $b \in \mathbb{R}^2$ and A is a 2×2 matrix. In fact, A must be orthogonal, which means that we can write things in terms of the complex plane $\mathbb{C}$ and get (in the orientation preserving case) $I\colon z \mapsto az + b$, where $a, b \in \mathbb{C}$ and $|a| = 1$. In the orientation reversing case, we have $I\colon z \mapsto a\bar{z} + b$.

Using the preceding discussion, we can now classify any isometry of the Euclidean plane as belonging to one of four types, depending on whether it preserves or reverses orientation, and whether or not it has a fixed point.

Case 1: An orientation preserving isometry which possesses a fixed point is a *rotation.* Let x be the fixed point, $Ix = x$. Fix another point y; both y and Iy lie on a circle of radius $d(x, y)$ around x. The rotation about x which takes y to Iy satisfies these criteria, which are enough to uniquely determine I given that it preserves orientation; hence I is exactly this rotation.

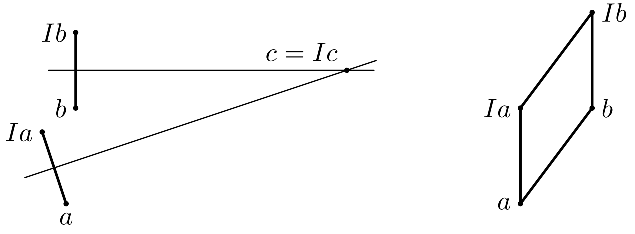


Figure 1.22. An orientation preserving isometry with no fixed points is a translation.

Rotations are entirely determined by the centre of rotation and the angle of rotation, and so are specified by three parameters.

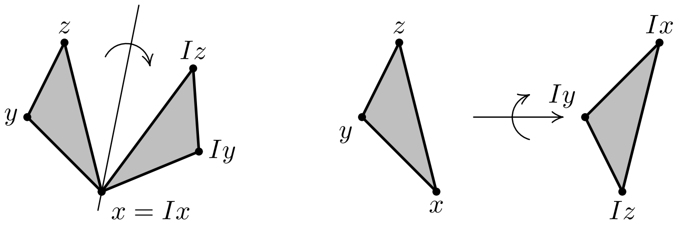
Case 2: An orientation preserving isometry I with no fixed points is a *translation*. The easiest way to see that is to use the complex algebraic description. Writing $Iz = az + b$ with $|a| = 1$, we observe that if $a \neq 1$, we can solve $az + b = z$ to find a fixed point for I. Since no such point exists, we have $a = 1$, hence $I\colon z \mapsto z + b$ is a translation.

One can also make a purely synthetic argument for this case; we show that the intervals $[a, Ia]$ and $[b, Ib]$ must be parallel and of equal length for every a, b. Indeed, if they fail to be parallel for some a, b, then their perpendicular bisectors intersect in some point c, as shown in Figure 1.22. Since $[a, Ia, c]$ and $[b, Ib, c]$ are isosceles triangles, we have $d(a, c) = d(Ia, c)$ and $d(b, c) = d(Ib, c)$, hence $Ic = c$ since I preserves orientations.

But I has no fixed point, and so $[a, Ia]$ and $[b, Ib]$ must be parallel; since I is an isometry, $d(Ia, Ib) = d(a, b)$, and hence the quadrilateral $[a, Ia, Ib, b]$ is a parallelogram. It follows that the intervals $[a, Ia]$ are all parallel and of equal length, and so I is a translation.

We only require two parameters to specify a translation; since the space of translations is two-dimensional, almost every orientation preserving isometry is a rotation, and hence has a fixed point.

Case 3: An orientation reversing isometry which possesses a fixed point is a *reflection*. Say $Ix = x$, and fix $y \neq x$. Let ℓ be the line bisecting the angle formed by the points y, x, Iy. Using the same



Reflection—a line of fixed points Glide reflection—no fixed points

Figure 1.23. Orientation reversing isometries.

approach as in case 1, the reflection through ℓ takes x to Ix and y to Iy; since it reverses orientation, I is exactly this reflection.

It takes two parameters to specify a line, and hence a reflection, so the space of reflections is two-dimensional.

Case 4: An orientation reversing isometry with no fixed point is a *glide reflection.* Let T be the unique translation that takes x to Ix. Then $I = R \circ T$ where $R = I \circ T^{-1}$ is an orientation reversing isometry which fixes Ix. By the above, R must be a reflection through some line ℓ. Decompose T as $T_1 \circ T_2$, where T_1 is a translation by a vector perpendicular to ℓ, and T_2 is a translation by a vector parallel to ℓ. Then $I = R \circ T_1 \circ T_2$, and $R \circ T_1$ is a reflection through a line parallel to ℓ, hence I is the composition of a translation T_2 and a reflection $R \circ T_1$ which commute; that is, a glide reflection.

A glide reflection is specified by three parameters; hence the space of glide reflections is three-dimensional, so almost every orientation reversing isometry is a glide reflection, and hence has no fixed point.

The group $\operatorname{Isom}(\mathbb{R}^2)$ is a topological group with two components; one component comprises the orientation preserving isometries, the other the orientation reversing isometries. From the above discussions of how many parameters are needed to specify an isometry, we see that the group is three-dimensional; in fact, it has a nice embedding into the group $GL(3, \mathbb{R})$ of invertible 3×3 matrices:

$$\operatorname{Isom}(\mathbb{R}^2) = \left\{ \begin{pmatrix} O(2) & \mathbb{R}^2 \\ 0 & 1 \end{pmatrix} : \begin{pmatrix} \mathbb{R}^2 \\ 1 \end{pmatrix} \to \begin{pmatrix} \mathbb{R}^2 \\ 1 \end{pmatrix} \right\}.$$

Here $O(2)$ is the group of real-valued orthogonal 2×2 matrices, and the plane upon which $\mathrm{Isom}(\mathbb{R}^2)$ acts is the horizontal plane $z = 1$ in $\mathbb{R}^3$.

Exercise 1.16. Prove that every isometry of the Euclidean plane can be represented as a product of at most three reflections.

Exercise 1.17. Consider all possible configurations of two and three lines in the plane: two lines may be either parallel or intersecting; for three lines there are a few more options. Identify the product of reflections in those lines for each case as one of four types of isometries.

Exercise 1.18. Consider an orientation reversing isometry in the complex form $z \mapsto a\bar{z} + b$. Find a condition on $a, b \in \mathbb{C}$ which will determine if it is a reflection or a glide reflection, and identify the axis in both cases.

b. Isometries of the sphere and the elliptic plane. By counting dimensions in the isometry group of the Euclidean plane, we argued that almost every orientation preserving isometry has a fixed point, while almost every orientation reversing isometry has no fixed point. In the next lecture, we will see that the picture for the sphere is somewhat similar—now *any* orientation preserving isometry has a fixed point, and most orientation reversing ones have none. For the elliptic plane, however, it will turn out to be dramatically different: *any isometry has a fixed point*, and can in fact be interpreted as a rotation!

Many of the arguments in the previous section carry over to the sphere; the same techniques of taking intersections of circles, etc. still apply. The classification of isometries on the sphere is somewhat simpler, since every orientation preserving isometry has a fixed point, while every orientation reversing isometry (other than reflection in a great circle) has a point of period two, which becomes a fixed point when we pass to the elliptic plane.

We will be able to show that every orientation preserving isometry of the sphere comes from a rotation of $\mathbb{R}^3$, and that the product of two rotations is itself a rotation. This is slightly different from the case with $\mathrm{Isom}(\mathbb{R}^2)$, where the product could either be a rotation, or if the two angles of rotation summed to zero (or a multiple of 2π), a translation. We will, in fact, be able to obtain $\mathrm{Isom}(S^2)$ as a group of 3×3 matrices in a much more natural way than we did for $\mathrm{Isom}(\mathbb{R}^2)$ above, since any isometry of S^2 extends to a linear orthogonal map of $\mathbb{R}^3$, and so we will be able to use linear algebra directly.

Lecture 6

a. Classification of isometries of the sphere and the elliptic plane. There are two approaches we can take to investigating isometries of the sphere S^2; we saw this dichotomy begin to appear when we examined $\mathrm{Isom}(\mathbb{R}^2)$. The first is the *synthetic* approach, which treats the problem using the tools of solid geometry; this is the approach used by the Greek geometers of late antiquity in developing spherical geometry for use in astronomy.

The second approach, which we will follow below, uses methods of linear algebra; translating the question about geometry to a question about matrices puts a wide range of techniques at our disposal, which will prove enlightening, and rather more useful now than it was in the case of the plane, when the relevant matrices were only 2×2.

The first important result is that there is a natural bijection (which is in fact a group isomorphism) between $\mathrm{Isom}(S^2)$ and $O(3)$, the group of real orthogonal 3×3 matrices. The latter is defined by

$$O(3) = \{\, A \in M_3(\mathbb{R}) \mid A^T A = I \,\}.$$

That is, $O(3)$ comprises those matrices for which the transpose and the inverse coincide. This has a nice geometric interpretation; we can think of the columns of a 3×3 matrix as vectors in $\mathbb{R}^3$, so that $A = (a_1|a_2|a_3)$, where $a_i \in \mathbb{R}^3$. (In fact, a_i is the image of the i^{th} basis vector e_i under the action of A.) Then A lies in $O(3)$ iff $\{a_1, a_2, a_3\}$ forms an orthonormal basis for $\mathbb{R}^3$, that is, if $\langle a_i, a_j \rangle = \delta_{ij}$, where $\langle \cdot, \cdot \rangle$ denotes the inner product, and δ_{ij} is the Kronecker delta, which takes the value 1 if $i = j$, and 0 otherwise. The same criterion applies if we consider the rows of A rather than the columns.

Since $\det(A^T) = \det(A)$, any matrix $A \in O(3)$ has determinant ± 1; the sign of the determinant indicates whether the map preserves or reverses orientation. The group of real orthogonal matrices with determinant equal to positive one is the *special orthogonal group* $SO(3)$.

In order to see that the members of $O(3)$ are in fact the isometries of S^2, we could take the synthetic approach and look at the images of three points not all lying on the same geodesic, as we did with $\mathrm{Isom}(\mathbb{R}^2)$; in particular, we can take the standard basis vectors e_1, e_2, e_3.

An alternate approach is to extend the isometry to $\mathbb{R}^3$ by homogeneity. That is, given an isometry $I\colon S^2 \to S^2$, we can define a linear map $A\colon \mathbb{R}^3 \to \mathbb{R}^3$ by

$$Ax = \|x\| \cdot I\left(\frac{x}{\|x\|}\right).$$

It follows that A preserves lengths in $\mathbb{R}^3$, and in fact, this is sufficient to show that it preserves angles as well. This can be seen using a technique called *polarisation*, which allows us to express the inner product in terms of the norm, and hence show the general result that preservation of norm implies preservation of inner product:

$$\begin{aligned}\|x+y\|^2 &= \langle x+y, x+y\rangle \\ &= \langle x,x\rangle + 2\langle x,y\rangle + \langle y,y\rangle \\ &= \|x\|^2 + \|y\|^2 + 2\langle x,y\rangle, \\ \langle x,y\rangle &= \frac{1}{2}(\|x+y\|^2 - \|x\|^2 - \|y\|^2).\end{aligned}$$

This is a useful trick to remember, and it allows us to show that a symmetric bilinear form is determined by its diagonal part. In our particular case, it shows that the matrix A we obtained is in fact in $O(3)$, since it preserves both lengths and angles.

The matrix $A \in O(3)$ has three eigenvalues, some of which may be complex. Because A is orthogonal, we have $|\lambda| = 1$ for each eigenvalue λ; further, because the determinant is the product of the eigenvalues, we have $\lambda_1\lambda_2\lambda_3 = \pm 1$. The entries of the matrix A are real, hence the coefficients of the characteristic polynomial are as well; this implies that if λ is an eigenvalue, so is its complex conjugate $\bar{\lambda}$.

There are two cases to consider. Suppose $\det(A) = 1$. Then the eigenvalues are $\lambda, \bar{\lambda}$, and 1, where $\lambda = e^{i\alpha}$ lies on the unit circle in the complex plane. Let x be the eigenvector corresponding to the eigenvalue 1, and note that A acts on the plane orthogonal to x by rotation by α; hence A is a rotation by α around the axis through x.

The second case, $\det(A) = -1$, can be dealt with by noting that A can be written as a composition of $-I$ (reflection through the origin) with a matrix with positive determinant, which must be a rotation, by the above discussion. Upon passing to the elliptic plane $\mathbb{R}P^2$, the reflection $-I$ becomes the identity, so that *every* isometry of $\mathbb{R}P^2$ is a rotation.

This result, that every isometry of the sphere is either a rotation or the composition of a rotation and a reflection through the origin, shows that every isometry has either a fixed point or a point of period two, which becomes a fixed point upon passing to the quotient space $\mathbb{R}P^2$.

As a concrete example of how all isometries become rotations in $\mathbb{R}P^2$, consider the map A given by reflection through the xy-plane, $A(x, y, z) = (x, y, -z)$. Let R be rotation by π about the z-axis, given by $R(x, y, z) = (-x, -y, z)$. Then $A = R \circ (-I)$, so that as maps on $\mathbb{R}P^2$, A and R coincide. Further, any point $(x, y, 0)$ on the equator of the sphere is fixed by this map, so that R fixes not only one point in $\mathbb{R}P^2$, but many.

Exercise 1.19. Let x and y be two points in the elliptic plane.

(1) Prove that there are at most two shortest curves connecting x and y.

(2) Find a necessary and sufficient condition for uniqueness of the shortest curve connecting x and y.

b. Area of a spherical triangle. In the Euclidean plane, the most symmetric formula for determining the area of a triangle is Heron's formula

$$A = \sqrt{s(s-a)(s-b)(s-c)},$$

where a, b, c are the lengths of the sides, and $s = \frac{1}{2}(a + b + c)$ is the semiperimeter of the triangle. There are other, less symmetric, formulae available to us if we know the lengths of two sides and the measure of the angle between them, or two angles and a side; if all

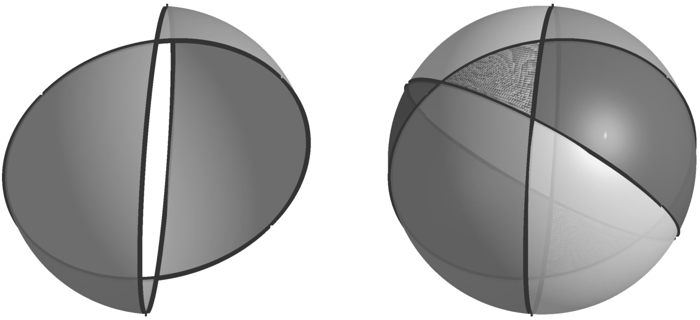


Figure 1.24. Determining the area of a spherical triangle.

we have are the angles, however, we cannot determine the area, since the triangle could be scaled up or down, preserving the angles while changing the area.

This is not the case on the surface of the sphere; given a spherical triangle, that is, the region on the sphere enclosed by three geodesics (great circles), we can find the area of the triangle via a wonderfully elegant formula in terms of the angles, as follows.

Consider the ‘wedge’ lying between two lines of longitude on the surface of a sphere, with an angle α between them. The area of this wedge is proportional to α, and since the surface area of the sphere with radius R is $4\pi R^2$, it follows that the area of the wedge is $\frac{\alpha}{2\pi}4\pi R^2 = 2\alpha R^2$. If we take this together with its mirror image (upon reflection through the origin), which lies on the other side of the sphere, runs between the same poles, and has the same area, then the area of the ‘double wedge’ shown in Figure 1.24 is $4\alpha R^2$.

Now consider a spherical triangle with angles α, β, and γ. Put the vertex with angle α at the north pole, and consider the double wedge lying between the two great circles which form the angle α. Paint this double wedge red; as we saw above, it has area $4\alpha R^2$.

Repeat this process with the angle β, painting the new double wedge yellow, and with γ, painting that double wedge blue. Now every point on the sphere has been painted exactly one colour (or, as in Figure 1.24, one particular shade of gray), with the exception of the points lying inside our triangle, and the points diametrically opposite them, which have been painted all three colours. (We neglect the boundaries of the wedges, since they have area zero.) Hence if we add up the areas of the double wedges, we obtain

$$\begin{aligned}\sum \text{areas of wedges} &= \text{blue area} + \text{yellow area} + \text{red area} \\ &= (\text{area of sphere}) + 4 \times (\text{area of triangle}),\end{aligned}$$

which allows us to write an equation for the area A of the triangle:

$$4(\alpha + \beta + \gamma)R^2 = 4\pi R^2 + 4A.$$

Solving, we see that

$$A = R^2(\alpha + \beta + \gamma - \pi). \tag{1.6}$$

Thus the area of the triangle is directly proportional to its *angular excess*; this result has no analogue in planar geometry, due to the flatness of the Euclidean plane. As we will see later on in the course, it does have an analogue in the hyperbolic plane, where the angles of a triangle add up to less than π, and the area is proportional to the *angular defect.*

Exercise 1.20. Express the area of a geodesic polygon on the sphere in terms of its angles.

Lecture 7

a. Spaces with lots of isometries. In our discussion of the isometries of $\mathbb{R}^2$, S^2, and $\mathbb{R}P^2$, we have observed a number of differences between the various spaces, as well as a number of similarities. One of the most important similarities is the high degree of symmetry each of these spaces possesses, as evidenced by the size of their isometry groups.

We can make this a little more concrete by observing that the isometry group acts *transitively* on each of these spaces; given any two points a and b in the plane, on the sphere, or in the projective plane, there is an isometry I of the space such that $Ia = b$.

In fact, we can make the stronger observation that the group acts transitively on the set of unit tangent vectors. That is to say, if v is a unit tangent vector at a, which can be thought of as indicating a particular direction along the surface from the point a, and w is a unit tangent vector at b, then not only can we find an isometry that carries a to b, but we can find one that carries v to w.

Another example of a surface with this property is the hyperbolic plane, which will appear in Chapter 4, and has the remarkable property that its isometry group allows not one but three natural representations as a matrix group (or a factor of such a group by its two-element centre).

In fact, these four examples are the only surfaces for which isometries act transitively on unit tangent vectors. There are of course a number of higher-dimensional spaces with this property: Euclidean spaces, spheres, and projective spaces, which are all analogues of their two-dimensional counterparts, immediately come to mind, and there are many more besides.

As an example of a space for which this property fails, consider the flat torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$. The property holds locally, in the neighbourhood of a point, but does not hold on the entire space. While $\text{Isom}(\mathbb{T}^2)$ acts transitively on points, it does not act transitively on tangent vectors; some directions lie along geodesics which are closed curves, while other directions do not. Another example is given by the cylinder, and examples of a different nature will appear later when we consider the hyperbolic plane and its factors.

What sorts of isometries does $\mathbb{T}^2$ have? We may consider translations $z \mapsto z + z_0$; rotations of $\mathbb{R}^2$, however, will not generally lead to isometries of $\mathbb{T}^2$, since they will usually fail to preserve the lattice $\mathbb{Z}^2$. The rotation by $\pi/2$ about the origin is permissible, as are the flips around the x- and y-axes, and around the line $x = y$.

In general, $\mathbb{Z}^2$ must be mapped to itself or a translation of itself, and so the isometry group is generated by the group of translations, along with the symmetry group of the lattice. The latter group is simply D_4, the dihedral group on four letters, which arises as the symmetry group of the square.

Exercise 1.21. Describe all the isometries of

(1) the ‘hexagonal’ torus of Exercise 1.15;

(2) the flat Möbius strip;

(3) the flat Klein bottle, i.e. the square with appropriately identified pairs of opposite sides.

Consider a more general class of examples, which generalise the construction of the flat torus as $\mathbb{R}^2/\mathbb{Z}^2$. Let L be a *lattice* in $\mathbb{R}^2$—that is, a set of vectors of the form $\{\, mu + nv \mid m, n \in \mathbb{Z} \,\}$, where u and v are two fixed linearly independent vectors. We can identify the factor space $\mathbb{R}^2/L$ with the parallelogram

$$\{\, su + tv \mid 0 \leq s, t \leq 1 \,\}$$

with pairs of opposite sides identified by translations.

Exercise 1.22. Show that the following statements hold.

(1) The factor space $\mathbb{R}^2/L$ is homeomorphic to a torus;
(2) $\mathbb{R}^2/L$ has a natural metric which is locally isometric to $\mathbb{R}^2$;
(3) The isometry group acts transitively on $\mathbb{R}^2/L$.

The 'crystallographic restriction' property established in the following exercise aids in the classification of isometries of these tori.

Exercise 1.23. Show that any non-trivial isometry of $\mathbb{R}^2/L$ with a fixed point has period 2, 3, 4, or 6.

b. Symmetric spaces. The discussion of spaces with lots of isometries is related to the notion of a *symmetric space*, which we will now examine more closely. In what follows, we assume certain properties of geodesics which will be formally described (but not proved) later in this course. In particular, we assume that there is a unique geodesic passing through a given point in a given direction, and that there is a unique shortest geodesic connecting any two sufficiently close points. Of course, all of this assumes the metric on our surface is given in a nice way, as has been the case with all examples considered so far.[10]

Given a point x on a surface X, we define the *geodesic flip* through x, denoted by I_x, as follows. For each geodesic γ passing through x, each point y lying on γ is sent to the point on γ which is the same

[10] These notions of direction and 'nice' metrics, which are rather vague at the moment, will be made more precise when we discuss smooth manifolds and Riemannian metrics in Chapters 3 and 4.

distance along the geodesic from x as y is, but in the other direction. It is immediate that this map preserves lengths along geodesics through x; it may happen, however, that the distances *between* these geodesics vary, in which case the map would not be isometric.

If the map is indeed isometric on some neighbourhood of x, and if this property holds for the geodesic flip I_x through any point $x \in X$, then we say that X is *locally symmetric.* The classification of such spaces (in any dimension) is one of the triumphs of Lie theory. Notice that the geodesic flip may not be extendable to a globally defined isometry, so the isometry group of a locally symmetric space may be (and sometimes is) quite small. Although we have not yet encountered any such examples, later on (Lecture 31) we will construct the hyperbolic octagon, whose isometry group can be shown to be finite, even though the space is locally symmetric.

Given two nearby points x, y, we can take the point z lying at the midpoint of the geodesic segment connecting them. Then $I_z x = y$. If X is connected (and hence path-connected) then any two points can be connected by a finite chain of neighbourhoods where these local isometries are defined. This implies that for any two points in a locally symmetric space, there exists an isometry between small enough neighbourhoods of those points. In other words, locally such a space looks the same near every point.

If for any point $x \in X$ the geodesic flip I_x can be defined not just locally, but globally (that is, extended to the entire surface X), and if it is in fact an isometry of X, then we say X is *globally symmetric.* In this case, the group of isometries $\mathrm{Isom}(X)$ acts transitively on all of X.

In the previous lecture we discussed a related, but stronger, notion, in which we require $\mathrm{Isom}(X)$ to act transitively not only on points in X, but also on unit tangent vectors. If this holds, then in particular, given any $x \in X$, there is an isometry of X taking some tangent vector at x to its opposite; this isometry must then be the geodesic flip, and so X is globally symmetric. It is *not* the case, however, that every globally symmetric space has this property of transitive action on tangent vectors; the flat torus is one example.

Examples of symmetric spaces are given by $\mathbb{R}^n$, S^n, and $\mathbb{R}P^n$, as well as by their direct products, about which we will say more momentarily. First, notice that the flat torus is symmetric, being the direct product of two symmetric spaces S^1. However, the embedding of the torus into $\mathbb{R}^3$ produces a space which is *not* symmetric, since the isometry group does not act transitively on the points of the surface. In fact, the isometry group of the embedded torus of revolution (the bagel) in $\mathbb{R}^3$ is a finite extension of a one-dimensional group of rotations, while the isometry group of the flat torus is, as we saw last time, a finite extension of a two-dimensional group of translations. Hence the two surfaces are homeomorphic but not isometric.

The flat torus $\mathbb{R}^2/\mathbb{Z}^2$ has no isometric embedding into $\mathbb{R}^3$, but it is isometric to the embedded torus in $\mathbb{R}^4$ given as the zero set of the two equations

$$\begin{aligned} x_1^2 + x_2^2 &= 1, \\ x_3^2 + x_4^2 &= 1. \end{aligned}$$

c. Remarks concerning direct products. Given any two sets X and Y, we can define their *direct product*, sometimes called the *Cartesian product*, as the set of all ordered pairs (x, y):

$$X \times Y = \{\, (x, y) \mid x \in X,\ y \in Y \,\}.$$

It is very often the case that if X and Y carry an extra structure, such as that of a group, a topological space, or a metric space, then this structure can be carried over to the direct product in a natural way. For example, the direct product of two groups is a group under pointwise multiplication, and the direct product of two topological spaces is a topological space in the product topology.

If X and Y carry metrics d_X and d_Y, then we can put a metric on $X \times Y$ in the same manner as we put a metric on $\mathbb{R}^2$, by defining

$$d((x, y), (x', y')) = \sqrt{d_X(x, x')^2 + d_Y(y, y')^2}.$$

If there are geodesics on X and Y, we can define geodesics on $X \times Y$, and hence can define the geodesic flip, which can be shown to satisfy the formula

$$I_{(x,y)}(x', y') = (I_x(x'), I_y(y')).$$

In the case $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$, this corresponds to the fact that the composition of a flip about a vertical line with a flip about a horizontal line is equivalent to rotation by π around the intersection of the two lines.

With the geodesic flip defined, we can then ask whether the product space $X \times Y$ is symmetric, and it turns out that if X and Y are both symmetric spaces, so is their direct product $X \times Y$. In this manner we can obtain many higher-dimensional examples, and so if we were to attempt to classify such spaces, we would want to focus on those which are irreducible in that they cannot be decomposed as the direct product of two lower-dimensional spaces, since the other examples will be built from these.

The direct product provides a common means by which we decompose objects of interest into simpler examples in order to gain a complete understanding. We find many examples of this in linear algebra, in which context the phrase *direct sum* is also sometimes used. Any finite-dimensional vector space can be written as the direct product of n copies of $\mathbb{R}$; this is just the statement that any finite-dimensional vector space has a basis. A more sophisticated application of this process is the decomposition of a linear transformation in terms of its action upon its eigenspaces, so that a symmetric matrix can be written as the direct product of one-dimensional transformations, while for a general matrix, we have the Jordan normal form.

This process is also used in the classification of finitely generated abelian groups, where we decompose the group of interest into a direct sum of copies of $\mathbb{Z}$ and cyclic groups whose order is a power of a prime, so that no further decomposition is possible. Thus the natural counterpart to the study of how a particular sort of mathematical structure can be decomposed is the study of what instances of that structure are, in some appropriate sense, *irreducible.*

Chapter 2

Combinatorial Structure and Topological Classification of Surfaces

Lecture 8

a. Topology and combinatorial structure on surfaces. Let X be a topological space. To avoid pathological cases, assume that X is *metrisable*, i.e. it is possible to place a metric d on X which induces the given topology. Note that there are many choices of metric which will be equivalent from the topological point of view. Once we have chosen a distance function, we can define balls of fixed radius around points, open and closed sets, convergence, closure, boundary, interior, and so on just as we do in real analysis.

Such an X is a *manifold* if for every point $x \in X$, there exists some open neighbourhood U_x containing x which is homeomorphic to $\mathbb{R}^n$; i.e. there exists a homeomorphism $\phi_x \colon U_x \to \mathbb{R}^n$. Thus a manifold is a topological space which locally looks like Euclidean space.

Exercise 2.1.

(1) Show that every connected manifold is path-connected.

(2) Construct an example of a compact connected metric space which is not path-connected.

We would like to say that the dimension n of the Euclidean space in question is also the dimension of the manifold, and is the same for every point x; two issues arise. The first is that if the space X is not connected, n may vary across the different components; this is easily avoided by assuming in addition that X is connected.

The second is more subtle. The proof that n is the same for every ϕ_x ought to go something like this: "Given two homeomorphisms $\phi_x\colon U_x \to \mathbb{R}^m$ and $\phi_y\colon U_y \to \mathbb{R}^n$, we can find a path from x to y in M. Then we can find points $x_1, \dots, x_k$ along the path such that $x_1 = x$, $x_k = y$, $\phi_{x_i}\colon U_{x_i} \to \mathbb{R}^{n_i}$ is a homeomorphism, and $U_{x_i} \cap U_{x_{i+1}} \neq \emptyset$ for every i. Thus that intersection is homeomorphic to open sets in both $\mathbb{R}^{n_i}$ and $\mathbb{R}^{n_{i+1}}$, and so $n_i = n_{i+1}$, because..."

Because what? This is where our intuition claims something stronger than our knowledge (at least for the moment). The above proof can be used to establish that $\mathbb{R}^m$ and $\mathbb{R}^n$ are homeomorphic, and we want to say that this can only happen if $m = n$. This is, in fact, true, but the general proof is somewhat more slippery than we might at first think.

It is relatively straightforward to show that $\mathbb{R}$ and $\mathbb{R}^2$ are not homeomorphic, although it should be noted that the Peano curve gives an example of a continuous map from $\mathbb{R}$ onto $\mathbb{R}^2$. This cannot be made into a homeomorphism, however; indeed, if $f\colon \mathbb{R} \to \mathbb{R}^2$ is a homeomorphism, then $f\colon \mathbb{R} \setminus \{0\} \to \mathbb{R}^2 \setminus \{f(0)\}$ is also a homeomorphism, but the latter space is connected and the former is not. Since connectedness is a topological property (we can define it entirely in terms of open and closed sets, without reference to a metric or any other structure), it is preserved by homeomorphisms, and hence we have a contradiction, showing that $\mathbb{R}$ is not homeomorphic to $\mathbb{R}^2$.

This argument actually shows that $\mathbb{R}$ is not homeomorphic to *any* $\mathbb{R}^n$ for $n \geq 2$, and naturally suggests a similar approach to showing, for example, that $\mathbb{R}^2$ is not homeomorphic to $\mathbb{R}^3$. Removing a line from $\mathbb{R}^2$ disconnects it, while removing a line from $\mathbb{R}^3$ leaves it connected. However, we cannot say in general what form the image of the line we remove from $\mathbb{R}^2$ will have in order to show that $\mathbb{R}^3$ remains connected. If we start with a line in $\mathbb{R}^3$ and take its preimage in $\mathbb{R}^2$, we have a continuous non-self-intersecting curve in the plane;

that such a curve separates $\mathbb{R}^2$ into two connected components is the content of the famous Jordan Curve Theorem, one of the cornerstones of two-dimensional topology. We will prove this theorem in Lectures 34 and 35.

The preceding discussion illustrates one of the difficulties inherent to topology; the notion of continuity is not a particularly nice one to work with all of the time, since continuous functions can be quite unpleasant, and the field is home to many pathological counterexamples. If, however, we restrict ourselves to *differentiable* objects, then things become much easier, and we have a whole array of local tools at our disposal, using the fact that the idea of *direction* is now made meaningful by the presence of tangent vectors. So far we have defined the notion of a *topological manifold*; by adding more structure, we can work with *differentiable* manifolds, in which context the equivalence relation of homeomorphism is replaced with that of *diffeomorphism.* This will be one of the central topics later in this course, beginning in Chapter 3.

For the time being, let us return to the continuous case. Having made these definitions, we can now give a proper definition of a surface; a surface is simply a two-dimensional manifold. One of our primary goals will be the classification of compact surfaces up to homeomorphism. Thus, we will need a reliable way of determining whether two surfaces are homeomorphic.

If two surfaces are in fact homeomorphic, we can demonstrate this by simply exhibiting a homeomorphism from one to the other. To show that they are *not* homeomorphic, however, often requires a little more ingenuity. For example, why is the torus not homeomorphic to the sphere? Intuitively it is clear that one cannot be deformed into the other, but a rigorous proof is harder to come by. One method is to follow our sketch of the proof that $\mathbb{R}^2$ is not homeomorphic to $\mathbb{R}^n$ for any $n \geq 3$; a 'nice' curve on the surface of the sphere disconnects it, which is not the case for every curve on the torus. So we can consider a curve which fails to disconnect the torus and claim that its image disconnects the sphere; this is again the Jordan Curve Theorem.

There are other proofs of this result as well, but they all require an alternative set of tools with which to approach the problem. For example, once we have the definition of a fundamental group and develop the basic theory of covering spaces, it becomes immediate that the sphere and the torus are not homeomorphic, since they have different fundamental groups. This illustrates a common approach to such problems, that of finding an *invariant*. If we can exhibit some property of a surface which is invariant under homeomorphisms, then two surfaces for which that property differs cannot be homeomorphic; in this case, the property is the fundamental group, or the property of being simply connected.

One approach to classifying surfaces is to restrict ourselves to the differentiable case, where everything is smooth, and then see what we can learn about the continuous case from that analysis. This echoes, for example, the approximation of continuous functions by polynomials in numerical analysis.

Another approach, which we will examine more closely in the next lecture, is to decompose our surface into a combination of simple pieces and take a combinatorial approach. For example, we could study surfaces which can be built up as the union of triangles glued along the edges. The strategy then is to first classify all surfaces which can be obtained this way (or, equivalently, all surfaces which allow such a combinatorial structure), and then to show that *every* surface can be so obtained.

We will occupy ourselves primarily with the first part, which is fun, includes combinatorics and algebra, and provides a good set of tools dealing with various examples and questions. The second part would drag us into hard general topology, starting from the Jordan Curve Theorem (which we will prove at the end of the course), and involves subtle approximation constructions which we will not discuss in this course. Fortunately, once this is established it can be taken for granted.

Exercise 2.2. Consider the space obtained from the torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ by identifying every point x with $-x$. Prove that it is homeomorphic to the sphere.

b. Triangulation. Because non-compact surfaces can be extremely complicated, we will fix our attention for the next while on compact surfaces, with the goal of describing all possible compact surfaces up to homeomorphism. That is, we will construct a list of representative examples, which will provide a classification in the sense that

(1) Every compact surface is homeomorphic to a surface from the list.

(2) No two surfaces from the list are homeomorphic to each other.

Along the way, we will describe convenient sets of invariants characterising the surfaces up to a homeomorphism.

So far, we have defined a surface as a two-dimensional manifold. We will begin by considering surfaces with an additional combinatorial structure, a *triangulation*, which avoids local complexity by building the surface up from simple pieces.

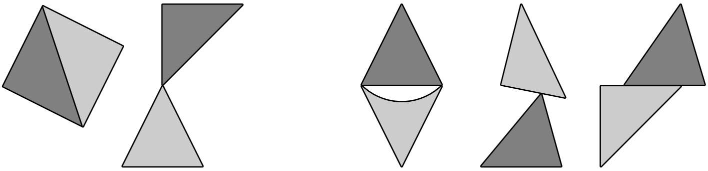
Definition 2.1. The *standard n-simplex*, denoted σ^n, is the subset of $\mathbb{R}^{n+1}$ given by

$$\sigma^n = \left\{ (x_0, \dots, x_n) \in \mathbb{R}^{n+1} \;\middle|\; x_i \geq 0 \;\forall i,\; \sum_{i=0}^{n} x_i = 1 \right\}.$$

We also use σ^n to denote any homeomorphic image of the standard n-simplex along with the *barycentric coordinates* $(x_0, \dots, x_n)$, and refer to such an image as an n-simplex.

We will only use the low-dimensional simplices σ^0, σ^1, and σ^2. The 0-simplex is simply a point, while the 1-simplex is an interval with a coordinate; that is, if A and B are the endpoints of the interval, then any point in the interval can be written as $tA + (1-t)B$, where $t \in [0,1]$, or more symmetrically as $tA + sB$, where $t, s \geq 0$ and $t + s = 1$.

The 2-simplex is a triangle; if the vertices are A, B, and C, then any point in the triangle can be written as $t_1 A + t_2 B + t_3 C$, where $t_i \geq 0$ and $t_1 + t_2 + t_3 = 1$. Some motivation for the term *barycentric coordinates* is given by the fact that if a point mass measuring t_i is placed at each vertex, then $t_1 A + t_2 B + t_3 C$ gives the location of the centre of mass of the triangle.



Properly attached Improperly attached

Figure 2.1. Attaching 2-simplices.

The boundary of an n-simplex σ^n is a union of $n+1$ different $(n-1)$-simplices, and the barycentric coordinates on these simplices come in a natural way from the coordinates on σ^n. For example, in the 2-simplex $\{t_1A + t_2B + t_3C\}$, the part of the boundary opposite C is the 1-simplex $\{t_1A + t_2B\}$.

A simplex also carries an orientation corresponding to the ordering of the vertices; this orientation is preserved by even permutations of the vertices, and reversed by odd ones. Hence there are two different orientations of a 2-simplex; one corresponds to the orderings (going clockwise, for instance) ABC, BCA, and CAB, while the other corresponds to CBA, BAC, and ACB.

The method by which we will build a surface out of simplices is called *triangulation*. We will say that two simplices are *properly attached* if their intersection is a simplex whose barycentric coordinates are given by the restriction of the coordinates on the two intersecting simplices.

Figure 2.1 gives examples of properly and improperly attached simplices. Informally, a collection of properly attached simplices is a *simplicial complex*, and a triangulation is a simplicial complex which is also a manifold. We can make this precise as follows:

Definition 2.2. A *triangulation* of a surface S is a collection $\mathcal{T}$ of 2-simplices, $\mathcal{T} = \{\sigma_i^2\}_{i=1}^n$, such that the following hold:

(1) $S = \bigcup_{i=1}^n \sigma_i^2$.

(2) For every $i \neq j$, the intersection $\sigma_i^2 \cap \sigma_j^2$ is either a 1-simplex σ_{ij}^1, a 0-simplex σ_{ij}^0, or the empty set $\emptyset$.

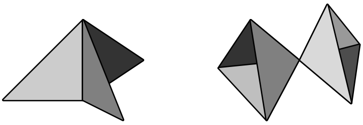


Figure 2.2. Properly attached simplices in forbidden configurations.

(3) Every 1-simplex is in the boundary of exactly two of the σ_i^2; that is, $\sigma_{ij}^1 = \sigma_{k\ell}^1$ iff $(i,j) = (k,\ell)$.

(4) Every 0-simplex is in the boundary of several σ_i^2 which may be arranged in a cyclic order; that is, given σ^0, the set of σ_i^2 which contain σ^0 can be put in a list $\sigma_{i_1}^2, \dots, \sigma_{i_k}^2$ in such a way that $\sigma_{i_j}^2 \cap \sigma_{i_{j+1}}^2$ is a 1-simplex for each $1 \leq j \leq k$ (where $\sigma_{i_{k+1}}^2 = \sigma_{i_1}^2$).

Properties (1) and (2) are fundamental to the concept of a *simplicial complex*, while properties (3) and (4) ensure that $\mathcal{T}$ is in fact a surface. In property (3), we could replace the words "exactly two" with "at most two"; this would allow for the possibility of a surface with a boundary.

The final two properties forbid the sorts of (three-dimensional) configurations seen in Figure 2.2, which serves to ensure that the triangulation is locally homeomorphic to $\mathbb{R}^2$; the details of this are left as an exercise for the reader.

Exercise 2.3. Show that if we define S by the union in property 1, and the simplices in $\mathcal{T}$ satisfy the other properties listed, then S is in fact a surface.

We now turn our attention to a particular surface, the sphere, and investigate possible triangulations. In particular, what is the triangulation of S^2 which uses the smallest possible number of simplices?

Any convex polyhedron is homeomorphic to the sphere, and we may subdivide any face which is not already a triangle to obtain a triangulation of S^2. Note that we will often refer to 2-simplices as *faces*, 1-simplices as *edges*, and 0-simplices as *vertices*. Among all polyhedra, the tetrahedron has the smallest number of faces, and in fact, this is the best we can do. Given any triangulation of S^2 (or any surface for that matter), fix a 2-simplex, or face; since each edge must belong to exactly two faces, and since any two faces intersect in at most one edge, there must be at least three distinct faces besides the one we chose, for a total of at least four, as in the tetrahedron.

As an example of a fairly natural construction which is *not* a triangulation, consider the torus. We obtain the torus as the quotient space of the square by an equivalence relation on its edges, and there is a very natural triangulation of the square into two 2-simplices by drawing a diagonal. This is not a triangulation of the torus, however, since the two triangles intersect along all three edges after passing to the quotient space.

One final example of a triangulation is provided by the icosahedron, which has 20 faces; passing to the quotient space obtained by identifying opposite faces, we have a triangulation of the projective plane using 10 simplices (it must be checked that all properties of a triangulation still hold after taking the quotient).

In general, triangulations provide an excellent theoretical tool for use in proofs, but are not the ideal technique for constructions or computations regarding particular surfaces; we will eventually discuss other methods more suited to those tasks.

c. Euler characteristic.

Definition 2.3. Given a triangulation $\mathcal{T}$, let F be the number of 2-simplices σ^2 (faces), E the number of 1-simplices σ^1 (edges), and V the number of 0-simplices σ^0 (vertices). Then the *Euler characteristic* of the triangulation is given by

$$\chi(\mathcal{T}) = F - E + V. \tag{2.1}$$

For the five regular polyhedra, we have the following table—here V', E', and F' represent the number of vertices, edges, and faces of the triangulations of the cube and dodecahedron obtained by partitioning each square face into two triangles, and each pentagonal face into three.

	V	E	F	V'	E'	F'	χ
tetrahedron	4	6	4				2
cube	8	12	6	8	18	12	2
octahedron	6	12	8				2
dodecahedron	20	30	12	20	54	36	2
icosahedron	12	30	20				2

Two features of this table are worthy of note. First note that for the cube and for the dodecahedron, we obtain $\chi = 2$ whether we calculate with V, E, F or V', E', F'; the act of subdividing each face does not change the Euler characteristic. Secondly, each of these polyhedra has the same Euler characteristic. This last turns out to be a consequence of the fact that they are all homeomorphic to the sphere S^2, and leads us to a quite general theorem.

Theorem 2.4. *Given a surface S, any two triangulations $\mathcal{T}_1$ and $\mathcal{T}_2$ of S have the same Euler characteristic.*

Proof. The proof will proceed in four steps.

(1) Define *barycentric subdivisions*, which will allow us to refine a triangulation $\mathcal{T}$.

(2) Show that χ is preserved under barycentric subdivisions, so that refining a triangulation does not change its Euler characteristic.

(3) Define a process of *coarsening*, and show that it also preserves χ.

(4) Given any two triangulations $\mathcal{T}_1$ and $\mathcal{T}_2$, refine $\mathcal{T}_1$ until we can use its vertices and edges to approximate the vertices and edges of $\mathcal{T}_2$, then coarsen this refinement into a true approximation of $\mathcal{T}_2$ itself.

This will allow us to speak of $\chi(S)$ rather than $\chi(\mathcal{T})$, and to compare properties of surfaces via properties of their triangulations.

Barycentric subdivision. Given a face σ^2 of $\mathcal{T}$, draw three lines, each originating at a vertex, passing through the point $(1/3, 1/3, 1/3)$, and ending at the midpoint of the opposite side. This partitions σ^2 into six smaller triangles (Figure 2.3), which inherit their barycentric coordinates from an appropriate scaling of the coordinates on σ^2.

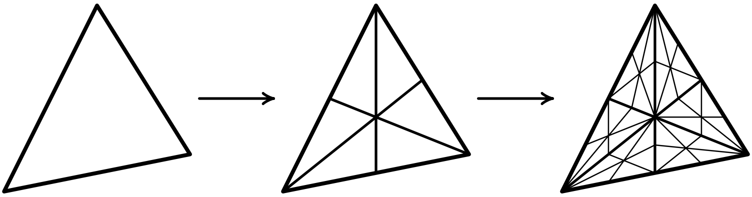


Figure 2.3. Two successive barycentric subdivisions.

Notice also that subdivisions of edges inherit coordinates in a consistent manner from both of the faces which are being subdivided; this is an advantage that barycentric subdivision enjoys over other possible methods of subdividing the 2-simplices.

Invariance of χ. Given a triangulation $\mathcal{T}$, let $\mathcal{T}'$ denote its barycentric subdivision. Each face is divided into six parts, so $F' = 6F$. Similarly, each edge is divided into two new edges, and each face has six new edges drawn in its interior, so $E' = 2E + 6F$. Finally, one new vertex is drawn on each edge, and one more in the centre of each face, so $V' = V + E + F$. Putting this all together, we obtain

$$\begin{aligned}\chi(\mathcal{T}') &= V' - E' + F' \\ &= (V + E + F) - (2E + 6F) + 6F \\ &= V - E + F \\ &= \chi(\mathcal{T}),\end{aligned}$$

and so the Euler characteristic is preserved by barycentric subdivision.

Lecture 9

a. Continuation of the proof of Theorem 2.4. In order to complete the proof that two triangulations of the same surface have the same Euler characteristic, we first prove two lemmas. For our purposes here, a polygon is a region of the plane bounded by a closed broken line.

Lemma 2.5. *Any polygon can be triangulated.*

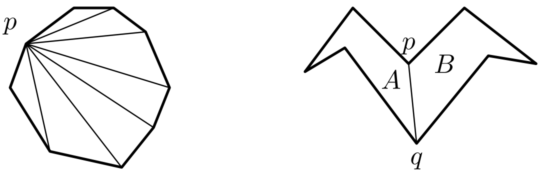


Figure 2.4. Triangulating polygons.

Proof. In the convex case, we can triangulate an n-gon P by fixing a vertex p, and then drawing $n-3$ diagonals from p, one to each vertex which is distinct from and not adjacent to p.

If the polygon P is non-convex, we proceed by induction on the number of sides. Fix a vertex p at which the angle is greater than π (if no such vertex exists, we are back in the convex case); it must be the case that some other vertex q is visible from p, in the sense that the line segment $[p, q]$ lies inside the polygon. Note that unlike the convex case, it may no longer happen that q can be taken to be adjacent to a neighbour of p (Figure 2.4).

Now each of A and B has fewer sides than P, and hence can by triangulated by the inductive hypothesis. This gives a triangulation of our original polygon, and proves the lemma. □

Definition 2.6. Two triangulations $\mathcal{T}_1$ and $\mathcal{T}_2$ are *affinely equivalent* if there exists a bijection $f\colon \mathcal{T}_1 \to \mathcal{T}_2$ which preserves the simplicial structure (that is, the image of a 2-simplex in $\mathcal{T}_1$ is a 2-simplex in $\mathcal{T}_2$, and so on) and whose restriction to any 2-simplex is an affine map.

Lemma 2.7. *Any triangulated polygon is affinely equivalent to a convex triangulated polygon.*

Proof. Once again, we proceed by induction. For $n = 3$, triangles are convex, so there is nothing to prove. For $n \geq 4$, decompose P into the union of an $(n-1)$-gon P' and a triangle T which is attached to P' along an edge e. By the inductive hypothesis, P' is affinely equivalent to a convex triangulated polygon $f(P')$, and to show that P is as well, we must attach an affine image of T along $f(e)$ in such a way that the polygon remains convex.

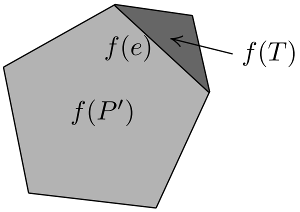


Figure 2.5. Obtaining a convex triangulation.

Any two triangles are affinely equivalent, and so we can make two angles of $f(T)$ as small as we like. In particular, the two angles at either end of $f(e)$ are each less than π, and so we can make two angles of $f(T)$ small enough that gluing $f(T)$ along $f(e)$ does not increase either of these angles beyond π, and the polygon remains convex. □

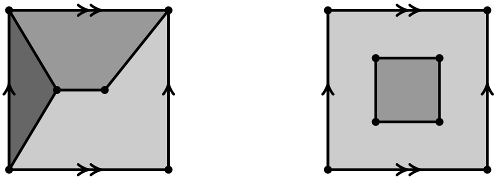
We now return to the proof that $\chi(\mathcal{T}_1) = \chi(\mathcal{T}_2)$. We begin by defining an analogue of triangulation using polygons with any number of sides.

Definition 2.8. A *map*[1] of a surface S is a partition of S into polygons such that the intersection of any two polygons is a union of some number of edges and/or vertices from each of the two (a polygon may also have a non-trivial intersection of this kind with itself). A *coarsening* of a triangulation $\mathcal{T}$ of S is a map of S in which each polygon is the union of 2-simplices from $\mathcal{T}$.

This definition allows certain configurations which were forbidden when using triangulations, as illustrated in Figure 2.6(a). Certain other configurations are still forbidden, however. For example, the requirement that the boundary of each polygon be a single closed curve forbids 'nestings' of the sort shown in Figure 2.6(b).

Remark. As we are doing topology at the moment, rather than geometry, it is natural to include the 1-gon (the disc with a marked 'vertex' on its boundary) and the 2-gon (the disc with the boundary divided into two 'edges') among the polygons. Notice, however, that by adding extra 'unnecessary' vertices which divide some edges, one

[1] In the sense of a 'geographic map', rather than in the usual mathematical sense of a 'mapping' or 'function'.



(a) Permissible (b) Forbidden

Figure 2.6. Permissible and forbidden configurations for maps of the torus.

can always assume that any polygon within a map has at least three sides.

It is a straightforward matter to verify that the Euler characteristic is preserved by coarsening, since joining together two faces eliminates both an edge and a face, and hence preserves χ. Note furthermore that when we compute $\chi(\mathcal{M})$ for a map $\mathcal{M}$, we may, if we like, disregard vertices with only two edges, and count edges separated by such vertices as a single edge, since by doing so we eliminate both a vertex and an edge, and so preserve χ. This will be the convention we follow in the remainder of the proof.

The final step of the proof requires approximating $\mathcal{T}_2$ with a coarsening of a refinement of $\mathcal{T}_1$. That is, if we denote by $\mathcal{T}_1^n$ the refinement of $\mathcal{T}_1$ obtained by performing n consecutive barycentric subdivisions, then we want to find a map $\mathcal{M}$ which is simultaneously

(1) a coarsening of $\mathcal{T}_1^n$;

(2) an approximation of $\mathcal{T}_2$, in a sense which will soon be made precise.

The latter requirement will, in particular, imply that $V(\mathcal{M}) = V(\mathcal{T}_2)$, and similarly for E and F. Thus we will have $\chi(\mathcal{T}_2) = \chi(\mathcal{M}) = \chi(\mathcal{T}_1)$.

To make precise the notion of 'approximation', observe that a triangulation gives us not only a combinatorial structure on a surface, but a metric one as well. Using the Pythagorean distance formula in terms of barycentric coordinates, we can define the distance between any two points on the same face—for convenience, we will scale distances so that edges of 2-simplices have unit length. Once the distance

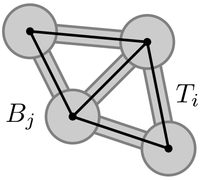


Figure 2.7. A neighbourhood of $\mathcal{T}_2$.

is defined on each face, we may define the length of a piecewise linear path on the surface, and so distances between points on different faces can be defined as the infimum of lengths of all such paths connecting the points.

Consider then the metric induced on S by the triangulation $\mathcal{T}_2$, which we hope to approximate. Let B_i be the ball of radius $1/3$ around the i^{th} vertex, and T_i the 'tube' of radius $1/10$ around the i^{th} edge, as indicated in Figure 2.7; then $B_i \cap B_j = \emptyset$ for $i \neq j$, and similarly for T_i.

The plan now is to consider a refinement $\mathcal{T}_1^n$, where n is very large. Then the edges of $\mathcal{T}_1^n$ form a sort of mesh, as shown in Figure 2.8. For sufficiently large n, the diameter of $\mathcal{T}_1^n$ (in the metric induced by $\mathcal{T}_2$) will be small enough that the mesh contains a path through each tube T_i from the ball B_j at one end to the ball B_k at the other. We will also be able to choose vertices in the mesh within each B_i and join them to these paths in such a way as to obtain a map $\mathcal{M}$ which is a coarsening of $\mathcal{T}_1^n$ and which has one vertex within each B_i, one edge for each tube T_i, and one face for each face of $\mathcal{T}_2$. It will then follow that the Euler characteristic is the same for $\mathcal{M}$ and $\mathcal{T}_2$.

Given an edge e of $\mathcal{T}_2$ running from vertex v_1 to vertex v_2, let B_1 and B_2 denote the ε_1-balls around v_1 and v_2, respectively, and let T denote the tube around e between B_1 and B_2. Note carefully that although we are accustomed to thinking of e as a straight line, there is no compatibility condition between the affine structures of the two triangulations $\mathcal{T}_1$ and $\mathcal{T}_2$—hence in Figures 2.8, 2.9, and 2.10, which are drawn with reference to the coordinates on $\mathcal{T}_1$, the edge e appears as a somewhat arbitrary continuous curve $\gamma\colon [0,1] \to S$.

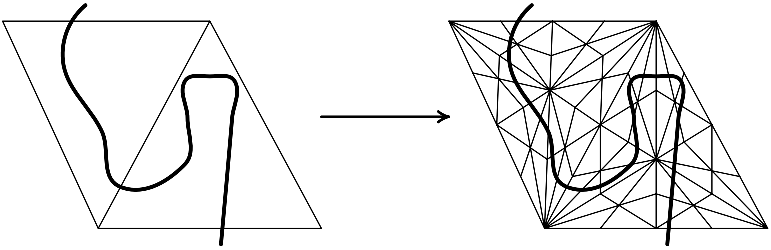


Figure 2.8. Refining $\mathcal{T}_1$ to approximate an edge of $\mathcal{T}_2$.

As the parameter t along the curve increases from 0 to 1, let x_0 be the last point of e which lies in the closure of B_1. Let x_1 be the first point of e after x_0 which intersects an edge of $\mathcal{T}_1^n$. Thereafter, let x_{k+1} be the first point of e after x_k which lies along an edge of $\mathcal{T}_1^n$ which does not contain x_k, and terminate the sequence with the first point x_N which lies in the closure of B_2.

Now the sequence $x_0, x_1, \dots, x_N$ determines a sequence of edges in the mesh $\mathcal{T}_1^n$. If x_k does not lie on a vertex, then it determines a unique edge e_k; if x_k does coincide with a vertex of the mesh, then choose an edge e_k which has x_k as one endpoint and an endpoint of e_{k-1} as the other.

This gives a sequence of edges $e_1, \dots, e_N$, each of which shares an endpoint with each of its neighbours. As shown in Figure 2.9, this may not be a true path, due to the presence of configurations which may be thought of as 'fans' and 'loops'. If we can eliminate these, we will have our desired path.

This elimination may be accomplished by beginning at an endpoint y_1 of e_1 and following not e_1, but the last e_k (greatest value of k) to have y_1 as an endpoint. This takes us to a vertex y_2, which must be an endpoint of e_{k+1} since the latter shares an endpoint with e_k, and y_1 is never to be visited again.

Again we follow the last e_ℓ to have y_1 as an endpoint, and iterate this procedure, eventually ending at y_M, an endpoint of e_N. We can follow an edge from y_M to a point $y_{M+1} \in B_2$, and similarly can find $y_0 \in B_1$. Let $\tilde{e}$ denote the broken line path from y_0 to y_{M+1}. Provided n was taken large enough, $\tilde{e}$ lies entirely inside the tube T.

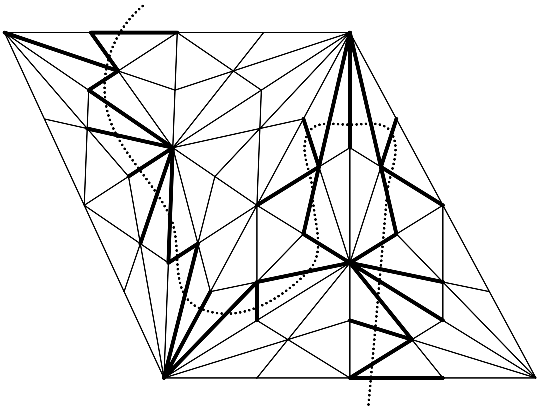


Figure 2.9. Determining a sequence of edges in the mesh.

We carry out this procedure along every edge e of $\mathcal{T}_2$, and then turn our attention to the balls B_i. Given a ball B_i, let m be the number of edges coming into B_i, and denote the corresponding endpoints of the broken paths $\tilde{e}$ by $z_1, \ldots, z_m$. Choose any vertex v of $\mathcal{T}_1^n$ lying inside B_i, and connect v to z_1 by a path along edges of the mesh $\mathcal{T}_1^n$. Then connect it to z_2 by a path which does not intersect the first, and continue until it is connected to every z_j.

After repeating this in every B_i, we have a map $\mathcal{M}$ of S which contains

(1) One vertex in each B_i, and hence the same number of vertices as $\mathcal{T}_2$, because the B_i are disjoint.

(2) One edge corresponding to each edge of $\mathcal{T}_2$, because the T_i are disjoint, and we constructed the edges $\tilde{e}$ so as not to intersect themselves or each other.

(3) One polygonal region corresponding to each 2-simplex of $\mathcal{T}_2$, because of the non-intersecting nature of the edges.

Hence V, E, and F all agree on $\mathcal{M}$ and $\mathcal{T}$; further, $\mathcal{M}$ is a coarsening of $\mathcal{T}_1^n$, and hence we have

$$\chi(\mathcal{T}_2) = \chi(\mathcal{M}) = \chi(\mathcal{T}_1^n) = \chi(\mathcal{T}_1). \qquad \square$$

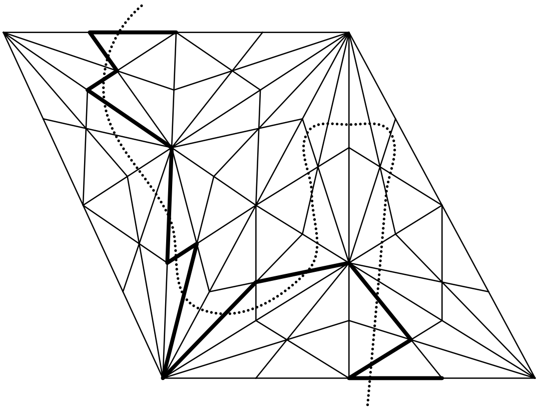
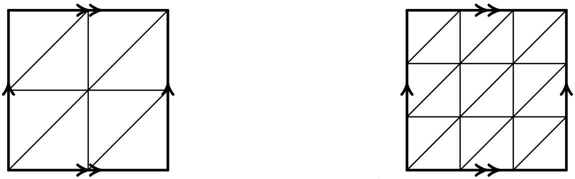


Figure 2.10. A true path along the mesh.

The sort of technical drudgery involved in the above proof is common in *point set topology.* In *algebraic topology*, one is often able to bypass considerations of this type by considering a coarser equivalence relation than homeomorphism, namely that of *homotopy equivalence*, which makes no distinction between, for example, the unit disc and a single point, or between an annulus and a circle.

This allows us to avoid certain convoluted constructions such as the one we have just been through, but has drawbacks of its own. While the fundamental invariants studied in algebraic topology, particularly homotopy and homology groups, are indeed invariant under homotopy equivalence, other important topological invariants such as dimension are not, and so there are topological results which cannot be achieved using this method. For example, the question of how many simple closed curves can be removed from a surface before it is disconnected is related to the Euler characteristic, but the proof requires an argument closer to the one we have just given, rather than one involving homotopy.

b. Calculation of Euler characteristic. We have already seen that the Euler characteristic of any regular polyhedron is 2, and with the above result on independence from choice of triangulation, we can now state unequivocally that $\chi(S^2) = 2$.



Not a triangulation

A true triangulation

Figure 2.11. Attempts at triangulating the torus.

Consider the triangulation of the torus in Figure 2.11. We must be careful how we count vertices and edges because of the identifications made between opposite sides. The four corners are all the same vertex, and the eight remaining vertices along the edge are identified in four pairs. Adding the four vertices in the interior, we have $1 + 4 + 4 = 9$ vertices. Similarly, the 12 outside edges come in six pairs, and we add 21 interior edges for a total of $E = 27$. Finally, there are 18 faces, so $\chi = V - E + F = 9 - 27 + 18 = 0$.

Exercise 2.4. Prove that the minimal number of vertices in a triangulation of the torus is seven.

For the projective plane $\mathbb{R}P^2$, we could be careful and choose a particular symmetric triangulation of S^2 which remains a triangulation after identifying antipodal points, such as the icosahedron, or we could be a little more careless and simply consider a very fine symmetric triangulation $\mathcal{T}$ which is guaranteed to remain a triangulation when we pass to its projection $\tilde{\mathcal{T}}$ in $\mathbb{R}P^2$. Then we have $2\tilde{F} = F$, $2\tilde{E} = E$, and $2\tilde{V} = V$, so it follows that $\chi(\mathbb{R}P^2) = 1$.

Exercise 2.5. Find a triangulation of the projective plane which uses the fewest possible simplices.

This argument works quite generally whenever we have a covering map from one space to another. In particular, since the map $(x, y) \mapsto (2x, 2y)$ is a covering map from the flat torus to itself, we have $4\chi(\mathbb{T}^2) = \chi(\mathbb{T}^2)$, which provides an alternative proof that $\chi(\mathbb{T}^2) = 0$.

Exercise 2.6. Calculate the Euler characteristic of the Klein bottle.

Lecture 10

a. From triangulations to maps. Let M_1 and M_2 be two surfaces equipped with triangulations $\mathcal{T}_1$ and $\mathcal{T}_2$. If $f\colon M_1 \to M_2$ is a homeomorphism, then $f(\mathcal{T}_1)$ is a triangulation of M_2, hence $\chi(\mathcal{T}_1) = \chi(\mathcal{T}_2)$; it follows that $\chi(M_1) = \chi(M_2)$, so the Euler characteristic is a topological invariant of a compact triangulable surface.

We have left unanswered (and up until now, unasked) the question of whether any compact surface (compact two-dimensional manifold) admits a triangulation. This is in fact the case, but we will not present the proof of this result in this course, as it requires not only considerable combinatorial ingenuity, but also techniques of point set topology at a higher level than used in the proof of Theorem 2.4.

Rather, we shall turn our attention from triangulations to maps, which we introduced briefly in the proof of Theorem 2.4. We will see, in particular, that the proof given there really establishes the more general result that the Euler characteristic is an invariant not just of triangulations, but of maps. First, though, a few comments about maps are in order.

The most obvious distinction between maps and triangulations is the list of permissible shapes; maps may comprise polygons with any number of sides, while triangulations are restricted to triangles. However, there is another, more subtle distinction. A triangulation comes equipped with barycentric coordinates on each triangle, so when we attach two triangles, it is obvious how the gluing along each edge is to be carried out. This is not the case for a map; the polygons lack a native affine structure, and so in particular there is no canonical way to attach along edges. With this in mind, let us make more precise the definition of a map, which so far we have thought of as a union of "properly attached" polygons.

We begin with the standard n-gons S_n, which are modeled by the regular n-gons lying in the complex plane $\mathbb{C}$ with vertices at the n^{th} roots of unity $\exp(2\pi ik/n)$, $1 \le k \le n$. As was mentioned before, we allow the case $n = 2$; S_2 is modeled by the unit circle $\{\, z \in \mathbb{C} \mid |z| = 1 \,\}$ with two vertices at ± 1 and two edges, one the top half of the circle, the other the bottom half. We also allow the case $n = 1$; the

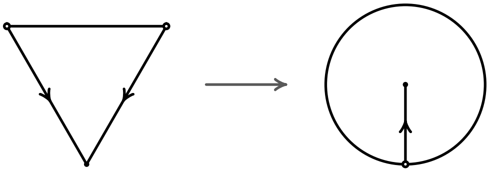


Figure 2.12. A map with a 'spike'.

model for S_1 has the entire unit circle as its single edge, and $z = 1$ as its single vertex.

Now a *generalised polygon* P on a surface M is simply the image of some S_n under a continuous function $f\colon S_n \to M$ satisfying certain conditions:

(1) The restriction of f to the interior of S_n is a homeomorphism onto its image.

(2) Given any edge e of S_n, the restriction of f to e is a homeomorphism onto its image.

The images under f of edges of S_n are themselves referred to as edges, and similarly for vertices. This allows us to make the following formal definition:

Definition 2.9. Given a surface M, a *map* on M is a decomposition of M as a union of generalised polygons (not disjoint), $M = \bigcup_{i=1}^{n} P_i$, along with the associated functions $f_i\colon S_{n_i} \to P_i$, satisfying:

(1) Given $i \neq j$, the intersection $P_i \cap P_j$ is a union of edges of P_i and P_j.

(2) Any point $x \in M$ which is not a vertex has at most two preimages; in particular, it lies in at most two of the P_i.

The latter condition ensures that each edge is identified with at most (in fact, exactly) one other. With the precise definition in hand, we can now state the following:

Theorem 2.10. *Let M be a compact surface which admits a triangulation $\mathcal{T}$, and let $\mathcal{M}$ be any map on M. Then $\chi(\mathcal{M}) = \chi(\mathcal{T})$, and hence any two such maps have the same Euler characteristic.*

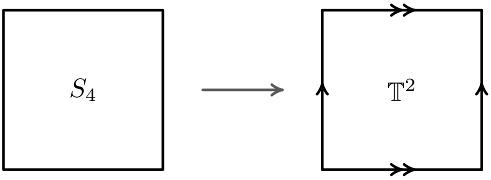


Figure 2.13. A map of the torus using only a single face.

Proof. Proceed exactly as in the proof of Theorem 2.4, taking $\mathcal{T}_1 = \mathcal{T}$ and replacing $\mathcal{T}_2$ with $\mathcal{M}$; note that we may just as easily approximate the map $\mathcal{M}$ with the mesh $\mathcal{T}_1^n$ as the triangulation $\mathcal{T}_2$. □

The definition of a map allows for very general configurations; for example, we can have 'spikes', as in Figure 2.12, which is the image of S_3 under a function identifying two adjacent sides in the direction indicated. We can also represent the torus $\mathbb{T}^2$ as a map with a single face, which is just the familiar planar model shown in Figure 2.13, the image of S_4 under a function identifying opposite edges—the usual parametric embedding of the flat torus in $\mathbb{R}^4$ gives a concrete realisation of this map.

This last example illustrates the greater utility provided by maps for purposes of computation and classification. As indicated in the previous lecture, triangulations are not terribly effective for these two purposes, despite being powerful theoretical tools. We will see very shortly that maps do not suffer from this shortcoming.

Theorem 2.11. *Any surface M which admits a map must necessarily admit a map with a single face.*

Proof. The proof is by induction on the number of faces, and the only difficulty is a slight technical one. Given two generalised polygons P_i and P_j which share an edge, we would like to erase that edge and combine the two polygons into one; P_i is an image of S_{n_i}, and P_j of S_{n_j}, so we would like to obtain $P_i \cup P_j$ as an image of $S_{n_i+n_j-2}$ under some function f_{ij}. However, because there is no affine structure on the polygons, and hence no *a priori* agreement in any meaningful sense between f_i and f_j along the edge we wish to remove, we must explicitly construct f_{ij}.

By Lemma 2.5, we can triangulate both S_{n_i} and S_{n_j}, and these triangulations carry over to triangulations of P_i and P_j. Taking the union of these gives a triangulation of $P_i \cup P_j$, which we can coarsen by removing all edges and vertices in the interiors of P_i and P_j, as well as all those lying along the edge we wish to remove.

In this way we obtain a single face in place of the two which were there before, decreasing the number of faces in the map by one. The result follows by induction. □

This leads us to the following result which will prove very valuable in our classification of surfaces:

Corollary 2.1. *Every compact triangulable surface is homeomorphic to a polygon with pairs of sides identified (which must therefore have an even number of sides).*

Remark. The process of investigating higher-dimensional manifolds via the analogue of triangulation, known as *simplicial decomposition*, is in general much more difficult. In three dimensions, for example, it is not obvious what requirement should be placed on the set of 3-simplices intersecting at a common vertex in order that the neighbourhood of that vertex be homeomorphic to $\mathbb{R}^3$, whereas in two dimensions the requirement was simply that the 2-simplices be arranged cyclically.

More critically, there are examples of higher-dimensional topological manifolds which admit no simplicial decomposition. The existence of such manifolds, which defies straightforward intuition, is among the most striking results of topology.

We also note that all our considerations can also be carried out for surfaces with a boundary; that is, two-dimensional manifolds where we allow two different types of points. Interior points have neighbourhoods homeomorphic to $\mathbb{R}^2$, while boundary points have neighbourhoods homeomorphic to $\mathbb{R}^2_+ = \{ (x, y) \in \mathbb{R}^2 \mid y \geq 0 \}$. Such surfaces may be usefully thought of as compact surfaces without boundary which have had holes removed.

b. Examples. We now know from Corollary 2.1 that we can classify compact triangulable surfaces by examining the quotient spaces of

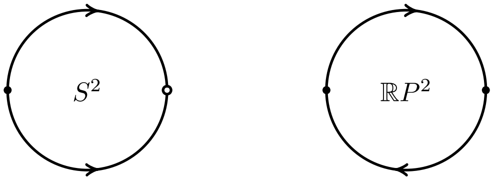


Figure 2.14. The two possible models on S_2.

various polygons upon identifying various pairs of sides. Let us begin our investigation of these *planar models* with the possibilities for the 2-gon.

S_2 is just the unit disc in $\mathbb{C}$ with ± 1 singled out as the vertices. An identification of the two edges is accomplished by a homeomorphism from one to the other along which we will 'glue' the edges. The reader may verify that perturbing the homeomorphism slightly does not change the resulting quotient space; all that matters is the *direction* of the homeomorphism. That is, if we move from left to right along the top edge, does the corresponding point on the bottom edge move from left to right or from right to left?

The two cases are shown in Figure 2.14. In the first case, we have the quotient space of the disc $D^2 \subset \mathbb{C}$ by the equivalence relation $z \sim \bar{z}$ for $|z| = 1$, which is the sphere S^2. We note that the model has two vertices, one face, and one edge (since the top and bottom edges are identified), so $\chi = V - E + F = 2 - 1 + 1 = 2$, as expected for the sphere.

In the second case, the equivalence relation is given by $z \sim -z$, and we obtain the projective plane $\mathbb{R}P^2$. This may be seen from the fact that $\mathbb{R}P^2$ is the northern hemisphere of S^2 with antipodal equatorial points identified; upon orthogonal projection to the equatorial plane we obtain the disc with antipodal boundary points identified, which is the picture in Figure 2.14. Note that the vertices ± 1 are identified, so the model has $V = E = F = 1$ and hence $\chi = 1$, which agrees with our original calculation for $\chi(\mathbb{R}P^2)$.

We now pass to the case where P is a 4-gon, or square. We must first decide which pairs or edges will be identified; we can either identify opposite sides or two sets of adjacent sides. For each pair,

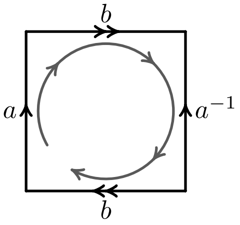


Figure 2.15. Notation for models on S_4.

there are two possible orientations, which we may think of as forward and backward, so given a choice of how to pair up the edges, there are three possibilities; both pairs forward, both backward, or one each way.

We can make this more precise with some notation, which is illustrated in Figure 2.15. Let us assign each edge a letter, and use each letter exactly twice. Two edges with the same letter are to be identified, and the direction is determined by whether the letter appears as, for example, a or a^{-1}. If we draw arrows on the sides indicating the direction of identification, and then make a circuit clockwise around the square beginning in the lower left, we write each side as x if we are moving in the direction of the arrows, and x^{-1} if we are moving opposite the direction of the arrows.

Thus we would write Figure 2.15, which is a model of the Klein bottle, as $aba^{-1}b$, because starting at the lower left, we encounter first a side labeled a with an arrow pointing clockwise, then a side b with an arrow pointing clockwise. We then encounter a with an arrow pointing *counterclockwise*, so we write a^{-1}, and finally a second side b with an arrow pointing clockwise, so we write b.

In this way we can write the six possible identifications, up to rotations and relabelings, as

$$\begin{array}{lll} aabb, & aa^{-1}bb, & aa^{-1}bb^{-1}, \\ abab, & aba^{-1}b, & aba^{-1}b^{-1}. \end{array}$$

For example, the labeling $aba^{-1}b^{-1}$ is quickly seen to be our usual model of the torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$, and Figure 2.15 shows the labeling $aba^{-1}b$ to be the Klein bottle. But what are the other four?

If we look at $aa^{-1}bb^{-1}$, for example, we can compute $\chi = 2$, and so we might conjecture that it represents the sphere, since so far all of the surfaces we have computed χ for have had distinct Euler characteristics. However, while S^2 and $\mathbb{R}P^2$ are in fact the only surfaces with $\chi = 2$ and 1, respectively, it is no longer true for $\chi \leq 0$ that the Euler characteristic uniquely determines the surface.

Exercise 2.7. Prove by direct construction that

- $aabb$ is another representation for the Klein bottle;
- $aa^{-1}bb$ is the projective plane;
- $aa^{-1}bb^{-1}$ is indeed the sphere;
- $abab$ is the projective plane.

Exercise 2.8. Consider all surfaces which can be obtained by identifying pairs of sides in a hexagon. Divide them into groups of mutually homeomorphic surfaces, and prove that surfaces from different groups are not homeomorphic.

In turns out that we will need another invariant to construct our list of surfaces; this will lead us to the notion of *orientability.* With this tool in hand, we will be able to construct a complete list, and to identify any planar model with a surface on our list via the method of cutting and pasting.

Lecture 11

a. Euler characteristic of planar models. So far we have seen planar models for four different surfaces; two of these used the 2-gon and two the 4-gon. We can list these in terms of the identifications made between various sides as we complete a circuit around the boundary, as explained previously:

edge identifications	surface	Euler characteristic
aa^{-1}	sphere	2
aa	projective plane	1
$aba^{-1}b^{-1}$	torus	0
$abab^{-1}$ or $aabb$	Klein bottle	0

To compute the Euler characteristic χ of a planar model on a $2m$-gon, we may observe that $F = 1$ and $E = m$ after passing to the quotient space, so the only variable is the number of vertices after all identifications have been made. If we write q_i for the number of edges attached to the i^{th} vertex, then we have the relation $2E = \sum_{i=1}^{V} q_i$.

A vertex attached to a single edge constitutes a 'spike' which may be removed without changing the topology of the surface. In general, this allows us to obtain a planar model on a $2(m-1)$-gon, and so we may assume that $q_i \geq 2$ for every i. We can go further and note that a vertex with $q_i = 2$ is in some sense superfluous, and can be removed, combining the two adjacent edges into one, to again obtain a planar model on a $2(m-1)$-gon. Thus for any planar model without spikes or unnecessary vertices, we have $2E \geq 3V$, and it follows that

$$\chi = V - E + F \leq \frac{2}{3}E - E + 1 = 1 - \frac{m}{3}.$$

Upon making the further observations that χ is an integer and that $V \geq 1$, we have convenient bounds on the Euler characteristic in terms of the number of sides of the planar model:

$$2 - m \leq \chi \leq 1 - \left\lceil \frac{m}{3} \right\rceil .$$

Note that these only apply if the model is simplified, in the sense discussed above. The astute reader will observe that the bounds we have obtained forbid positive values of χ, and hence cannot apply to our models of the sphere and the projective plane. This is because in the model of the sphere as a 2-gon with edges identified, both vertices are spikes, but we cannot remove them to make a simpler model without eliminating every edge of the 2-gon. Similarly, for the projective plane, the single edge has both ends at the same vertex, so the vertex has degree two, but cannot be removed without eliminating every vertex of the 2-gon.

We now return to the question of planar models on the 4-gon. At least two vertices must be identified, so $1 \leq V \leq 3$, hence χ must be one of 0, 1, or 2. We have seen surfaces with each of these values already, and it turns out that these are the only options.

b. Attaching handles. Given a surface M, we can 'attach a handle' by cutting two holes in the surface, taking a cylinder C, and gluing

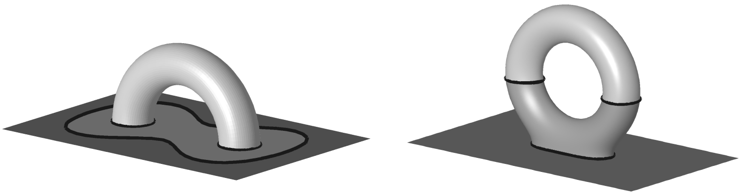


Figure 2.16. Attaching a handle—two equivalent constructions.

one end of C to each hole (Figure 2.16). For example, if we begin with a sphere and attach a handle in this manner, we obtain a surface homeomorphic to a torus.

Consider a neighbourhood of the two holes to which the cylinder is attached; this is homeomorphic to a disc with two holes, the so-called 'pair of pants' surface (also shown in Figure 3.14). Gluing one end of C to each hole, we obtain a torus with a hole, and so attaching a handle in the manner described above is equivalent to cutting a single hole and gluing our torus with a hole along its boundary; this is the procedure mentioned at the very beginning of this course, back in Lecture 1.

So far this is rather vague and imprecise; what does 'cutting a hole' mean, anyway? We want to say that we remove a homeomorphic image of a disc and glue along its boundary; will we obtain the same object no matter which disc we remove? Just how standard is a hole?

If we consider attaching a handle to a sphere, we could appeal to the Jordan Curve Theorem, which states that any homeomorphic image of a circle on the sphere separates it into two disjoint regions, each homeomorphic to a disc. We then remove one of these discs, and glue the torus with a hole along the boundary circle.

Alternately, we can return to our combinatorial approach, and examine methods for cutting holes in our planar models. The usual model of the torus is the square with opposite edges identified; where is the best place to cut the hole? As shown in Figure 2.17, we cut the hole in a corner, so that the torus with a hole has a planar model on a pentagon.

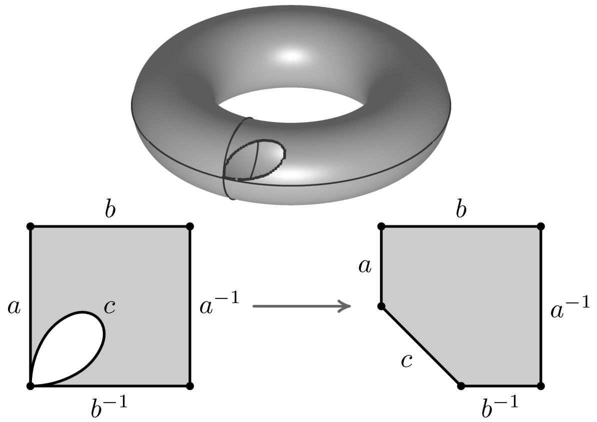


Figure 2.17. Cutting a hole in a torus.

Now if we begin with a planar model on any $2m$-gon and cut a hole in this manner, we can attach the torus with a hole as shown in Figure 2.18 to obtain a planar model on a $2(m+2)$-gon. Since all five vertices of the torus with a hole are identified, we do not add any new vertices by doing this, and we still have $F = 1$; thus the net result of this process is to increase the number of edges by two, and hence to decrease the Euler characteristic by two.

As we have seen, the sphere with one handle is a torus, which has a planar model on the 4-gon. Using the above process, we may attach a second handle and obtain a planar model on the 8-gon (Figure 2.19);

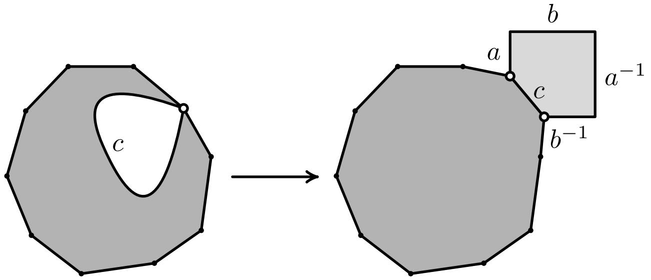


Figure 2.18. Attaching a handle to a planar model.

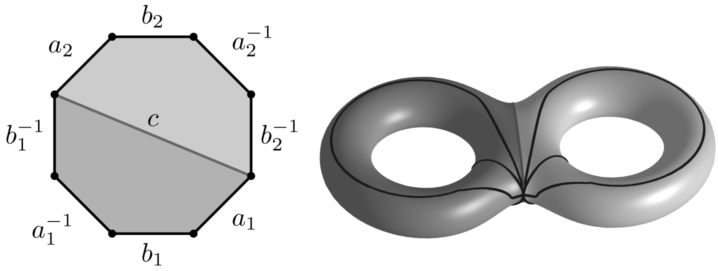


Figure 2.19. A sphere with two handles.

in general, after attaching m handles, we have a planar model on the $4m$-gon, with identifications given as in the table below:

m	identifications	V	E	F	χ
1	$aba^{-1}b^{-1}$	1	2	1	0
2	$a_1b_1a_1^{-1}b_1^{-1}a_2b_2a_2^{-1}b_2^{-1}$	1	4	1	-2
m	$a_1b_1a_1^{-1}b_1^{-1}\cdots a_mb_ma_m^{-1}b_m^{-1}$	1	$2m$	1	$2-2m$

We will see eventually that this list is exhaustive; any compact orientable surface which admits a triangulation is homeomorphic to the sphere with m handles, for some $m \in \mathbb{N}_0$. First, though, we must discuss the notion of *orientability.*

Exercise 2.9. Prove that for every $m \geq 1$, both the regular $4m$-gon and the regular $4m+2$-gon with pairs of opposite sides identified by translations are homeomorphic to the sphere with m handles.

c. Orientability. What does it mean for a surface to be orientable? The usual first example of a non-orientable surface is the Möbius strip; it is often said that the strip "only has one side", which distinguishes it from orientable surfaces such as the sphere and the torus. Another way of saying this is that if we place a clock on this surface and move it once around the strip, returning to its original position, it will have reversed directions and be running counterclockwise.

However, we are dealing with surfaces as topological objects, and the notion of direction along a surface, which we need to apply the above method in its simplest incarnation, properly belongs to the

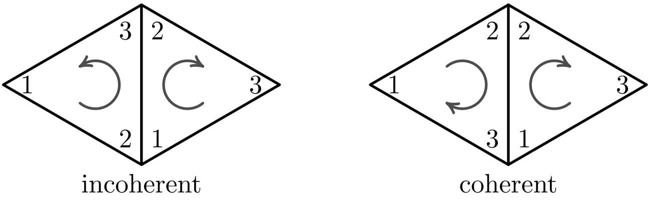


Figure 2.20. Coherent and incoherent orientations of 2-simplices.

study of differentiable manifolds, rather than topological ones. Orientability is in fact a topological invariant, and so we proceed as we did for the Euler characteristic, by first considering surfaces with triangulations.

An orientation of a triangle is simply an ordering of its vertices; this is preserved by even permutations of the vertices, and reversed by odd permutations. Thus we label the vertices 1, 2, and 3, and think of traversing the boundary of the triangle in the direction given by $1 \to 2 \to 3 \to 1$. We say that two adjacent 2-simplices are oriented *coherently* if they induce *opposite* orderings (or orientations) on the edge in which they intersect, as illustrated in Figure 2.20.

Definition 2.12. A triangulation $\mathcal{T}$ of a surface M is *orientable* if its 2-simplices admit a coherent collection of orientations.

Exercise 2.10. Show that no triangulation of the Möbius strip is orientable.

Exercise 2.11. Prove that for a surface with a triangulation, the following two definitions of orientability are equivalent:

(1) All triangles can be oriented in a coherent way.

(2) One can chose a positive direction of rotation at every point which changes continuously.

Theorem 2.13. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two triangulations of a surface M. Then $\mathcal{T}_1$ is orientable if and only if $\mathcal{T}_2$ is orientable.*

Proof. As in the proof of Theorem 2.4, we refine, coarsen, and approximate. Notice first that orientability is inherited by barycentric

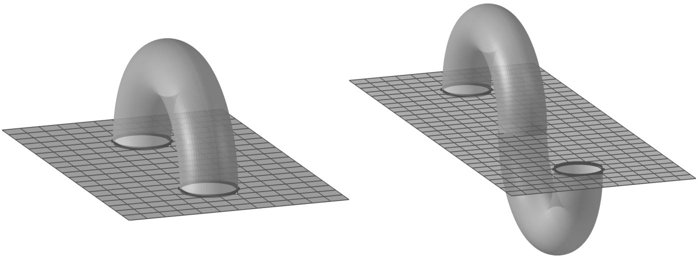


Figure 2.21. Two handles, one inverted, one not.

subdivision. Furthermore, we can extend the definition of orientability to maps, and one sees immediately that orientability is preserved under coarsening. Finally, a slight perturbation of an orientable triangulation is also orientable, and so we may approximate $\mathcal{T}_2$ with a coarsening of a refinement of $\mathcal{T}_1$ to obtain the result. □

d. Inverted handles and Möbius caps. Looking at the usual immersion of the Klein bottle into $\mathbb{R}^3$, we may see that it is homeomorphic to a sphere with two holes which has had a cylinder attached, but in which the attachment has been made in different directions along the two circles.

This is the notion of an *inverted handle*; after removing two holes from the surface M, take a patch of the surface which contains both and which can be given an orientation. Then take orientations of the two circles which are *not* coherent with respect to this orientation, and attach the ends of the cylinder according to these (Figure 2.21).

We have seen that attaching a single inverted handle to a sphere results in a Klein bottle. What surface do we obtain if we attach a second inverted handle?

We postpone the answer, and first consider the result of attaching an inverted handle to the projective plane. Because the projective plane is not orientable, we may slide one of the holes all the way around the surface and return it to its original position, reversing its orientation in the process. Thus attaching an inverted handle gives

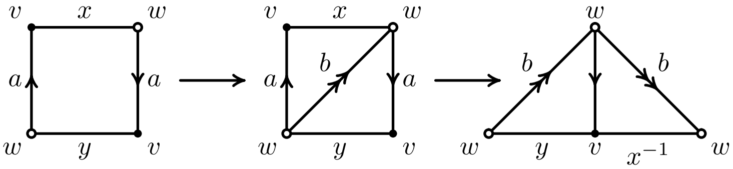


Figure 2.22. Planar models of the Möbius strip.

the same surface as attaching a regular handle in the case when the original surface is not orientable.

One final remark is in order. Just as with regular handles, the process of attaching an inverted handle decreases the Euler characteristic by two. Since no connected surface has $\chi > 2$, we cannot obtain the projective plane by attaching an inverted handle to anything. Rather, we may obtain it by attaching a *Möbius cap* to the sphere.

This attachment, also known as a *cross cap*, is carried out by removing a disc from the surface, and then identifying opposite points on its boundary. Alternately, we may think of gluing its boundary circle to the boundary circle of a Möbius strip; we will discuss this in more detail next time. For the time being, we merely note that attaching a Möbius cap to the sphere results in the projective plane, and ask the reader to consider what surface results if we attach a Möbius cap to the projective plane.

Lecture 12

a. Non-orientable surfaces and Möbius caps. As indicated in Figure 2.22, a planar model of the Möbius strip M is given by 4-gon with identifications $axay$. The edges x and y are not identified with anything else, and so remain as points on the boundary of the Möbius strip. The two vertices labeled v are identified with each other, as are the two vertices labeled w. Thus the boundary of M is given by following x from v to w, and y from w to v, and we see that the boundary of the Möbius strip is simply a single circle, as the third part of Figure 2.22 makes clear.

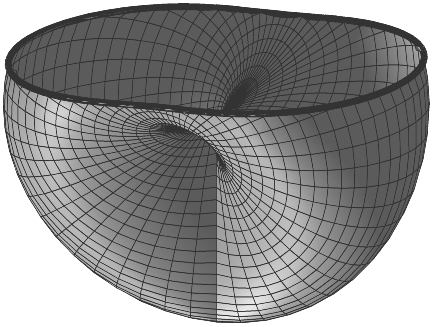


Figure 2.23. A cross cap; the Möbius strip immersed in $\mathbb{R}^3$ with self-intersection.

If we immerse the Möbius strip in $\mathbb{R}^3$ as shown in Figure 2.23, then the boundary circle $x^{-1}y$ is shown at the top of the figure, and the edge a runs along the lower portion of the surface. Beginning with any surface, we can cut a hole in the surface and attach a Möbius strip along the circle forming its boundary; this is the action of adding a Möbius cap, or cross cap.

This construction immediately makes the surface non-orientable, since any 'clock' can be brought to the Möbius strip and moved once around it, reversing its direction. We see from this that while orientability is a property of the entire surface, non-orientability is in some sense a local property; not in the sense that it can be defined in terms of neighbourhoods of points, but in the sense that if a portion of the surface is non-orientable, then the entire surface is non-orientable, no matter what constructions we may make elsewhere.

b. Calculation of Euler characteristic. We will follow a very general procedure of constructing surfaces by making various attachments. Suppose we are given a surface and then cut out a number of holes; then we are left with a surface whose boundary is a disjoint union of homeomorphic images of the circle S^1. Then there are three ways we can fill each hole by attaching standard surfaces to each image of S^1:

(1) Attach a cap, that is, a homeomorphic image of a disc, to a hole.

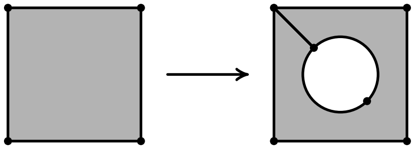


Figure 2.24. Cutting a hole in a surface.

(2) Attach a Möbius cap, a homeomorphic image of the Möbius strip, to a hole.

(3) Attach a handle to two holes, or, equivalently, attach a torus with a hole to a single hole.

Note that we make no mention of inverted handles, which we discussed in the previous lecture. The reason for this will shortly be made clear; first we ask what effect each of these attachments has on the Euler characteristic χ. Let us see what each does to a map of the surface.

As shown in Figure 2.24, cutting a single hole has the effect of adding two vertices, three edges, and leaving F constant,[2] so it decreases χ by 1. If we fill the hole with a cap, we add a face and leave E and V unchanged, so χ is returned to its original value. Hence the overall effect of cutting a hole and attaching a cap is to preserve the Euler characteristic (indeed, removing a hole and attaching a cap produces a surface which is homeomorphic to the original surface).

Similarly, attaching a Möbius cap adds a face; in addition, however, it adds an edge (the edge a from Figure 2.22), so that χ remains the same as it was for the surface with the hole. Hence the overall effect of cutting a hole and attaching a Möbius cap is to decrease the Euler characteristic by 1.

If we cut two holes to attach a handle, we decrease χ by 2. Attaching the handle itself adds two faces and two edges, leaving χ unchanged. Hence the overall effect of cutting two holes and attaching a handle is to decrease the Euler characteristic by 2, as we saw before.

[2] Of course, there are various other ways we could create a map for the new surface; we could, for example, take a minimalist approach and simply add a single edge with both ends attached to a preexisting vertex, as in Figure 2.17. Any choice we make will decrease χ by 1.

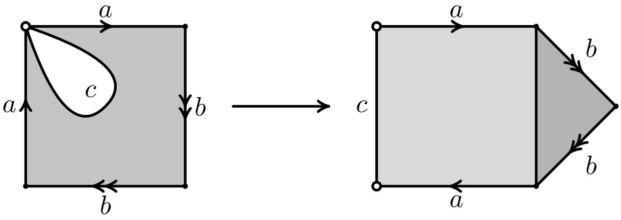


Figure 2.25. An inverted handle is two Möbius caps.

Once we establish that every surface can be obtained from the sphere by these constructions, these considerations illustrate why every orientable surface has even Euler characteristic.

Now what about adding an inverted handle? Why has it been left off our list? It turns out that attaching an inverted handle is equivalent to attaching two Möbius caps. Indeed, just as attaching a handle to two holes is equivalent to attaching a torus with a hole to a single hole, we can think of attaching an inverted handle as attaching a Klein bottle with a hole. A planar model of the Klein bottle on the 4-gon is given by the identifications *aabb*, and Figure 2.22 suggests a proof that each of *aa* and *bb* is equivalent to attaching a Möbius cap. Then Figure 2.25 shows that attaching an inverted handle is equivalent to attaching two Möbius caps.

The reader is encouraged to work through the details of these constructions independently; the concepts involved are not difficult, but care must be taken in counting vertices, edges, and faces to compute the Euler characteristic.

c. Covering non-orientable surfaces.

Definition 2.14. A (finite) *covering space* of a surface S is a connected surface $\tilde{S}$ together with a map $f\colon \tilde{S} \to S$ such that the following conditions hold:

(1) There exists $n \in \mathbb{N}$ such that given any point $x \in S$, the preimage $f^{-1}(x) \subset \tilde{S}$ consists of n distinct points.

(2) For every point $x \in S$, there exists a neighbourhood U_x of x such that $f^{-1}(U_x) = \bigcup_{i=1}^{n} V_i$, where each $V_i \subset \tilde{S}$ is open,

$V_i \cap V_j = \emptyset$ for $i \neq j$, and the restriction of f to V_i is a homeomorphism between V_i and U_x.

Then we say that $\tilde{S}$ is an *n-fold covering space*, or sometimes an *n-fold cover.*

Remark. One may also consider *infinite covering spaces*, where the pre-image of a neighbourhood of any point consists of a countable collection of homeomorphic images of the neighbourhood. The standard projection of the real line onto the circle is the simplest example of such a covering; projection of the plane onto the torus is another. These examples show that covering spaces and factor spaces are in some sense dual constructions to each other.

We have already seen one important example of a covering space; S^2 is a double cover of the projective plane $\mathbb{R}P^2$. This is an immediate consequence of the definition we gave for $\mathbb{R}P^2$, with the quotient map providing the covering map f.

Another example is given by the Möbius strip, which has the cylinder as a double cover. Consider the infinite strip $X = \mathbb{R} \times [-1, 1]$ with the translation $\tau\colon (x, y) \mapsto (x + 2, y)$. Then we obtain the cylinder as the quotient space of X by the action of τ; that is, we identify each point with all of its images under iterates of τ. A square root of τ is given by $\sigma\colon (x, y) \mapsto (x + 1, -y)$, and the quotient space of X by the action of σ is the Möbius strip. The covering map arises naturally as the canonical projection

$$\begin{aligned} f\colon \qquad\qquad & X/\sigma^2 \to X/\sigma, \\ \{\, (x + 2n, y) \mid n \in \mathbb{Z} \,\} & \mapsto \{\, (x + n, (-1)^n y) \mid n \in \mathbb{Z} \,\}. \end{aligned}$$

A similar argument, whose details are left to the reader, shows that the torus is a double cover for the Klein bottle.

We repeat our observation from a previous lecture that the Euler characteristics of S and $\tilde{S}$ are related; in particular, if $\tilde{S}$ is an n-fold cover of S, we have $\chi(\tilde{S}) = n\chi(S)$.

These examples point us towards a general result concerning non-orientable surfaces. Specifically, we have the following:

Proposition 2.1. *Every non-orientable surface has an orientable double cover.*

Proof. We follow an approach which is of wide utility both in topology and in other fields of mathematics; we define the problem away. We would like to associate a fixed orientation to each point of S; since we cannot do this, we define $\tilde{S}$ as follows: a point on $\tilde{S}$ is just a point of S together with a particular orientation at that point.

Locally, this looks like taking the direct product $S \times \{\pm 1\}$, so that each point in S appears twice in $\tilde{S}$, once with a positive orientation and once with a negative one. Of course, this is not true globally, precisely since S is non-orientable, and so we cannot define positive and negative in a coherent sense over the whole surface.

So far this gives us a set of points $\tilde{S}$ along with a natural projection $f\colon \tilde{S} \to S$. In order for $\tilde{S}$ to be a surface, we must describe its topology. We may define a set $U \subset \tilde{S}$ to be open if its image $f(U)$ in S is an open set, and if in addition we may define a coherent orientation on $f(U)$ which agrees with the orientation associated with each point in U. This gives a basis for the topology on $\tilde{S}$, and it is now immediate that f is a covering map.

If our original surface S were orientable, this procedure would give us a disconnected space, the union of two disjoint copies of S. Because S is non-orientable, we may find a path $\gamma\colon [0,1] \to S$ such that $\gamma(0) = \gamma(1)$, and following γ reverses orientation. Hence given any two points $x, y \in \tilde{S}$, we can find paths η_1 from $f(x)$ to $\gamma(0)$ and η_2 from $\gamma(0)$ to $f(y)$; then one of $\eta_1 \circ \eta_2$ or $\eta_1 \circ \gamma \circ \eta_2$ must give a path from x to y, and it follows that $\tilde{S}$ is connected.

Finally, $\tilde{S}$ is orientable by the construction. □

Exercise 2.12. Find a necessary and sufficient condition on k and l so that there is a covering map from the sphere with k handles onto the sphere with l handles.

d. Classification of orientable surfaces.

Theorem 2.15. *If M is a compact, closed (without boundary), orientable surface which has a map, then there exists an integer $m \geq 0$ such that M is homeomorphic to the sphere with m handles.*

Proof. We begin by outlining the general strategy, and postpone a detailed proof until the next lecture.

By Theorem 2.11, our surface admits a map with a single face, so we can consider a model on a $2n$-gon. After making the appropriate cancellations, we may assume that our model has no spikes; that is, no two adjacent edges have the same label. This corresponds to cancelling inverses in the sequence of identifications, so we forbid appearances of aa^{-1}, bb^{-1}, etc. Next we demonstrate a technique to modify our map so that it has only a single vertex; because the Euler characteristic must be preserved, this will of necessity decrease the number of edges.

Because M is orientable, it may be shown that in the sequence of identifications, each side must appear once in each direction. That is, we cannot have $abab$, but must have $aba^{-1}b^{-1}$, and so on. Then by considering the pair of identified edges which have the fewest other edges between them, we may find two symbols a and b which appear in the order $aba^{-1}b^{-1}$; note that there may be other edges in between these appearances. However, by assuming that all vertices of the $2n$-gon are identified, we may show that the model is equivalent to a $2(n-2)$-gon with a handle attached, and then proceed by induction on the Euler characteristic.

Lecture 13

a. Proof of the classification theorem. Given a map on a closed compact orientable surface S, we follow the steps indicated last time to show that S is homeomorphic to the standard model for a sphere with m handles, that is, a $4m$-gon with identifications

$$a_1 b_1 a_1^{-1} b_1^{-1} \dots a_m b_m a_m^{-1} b_m^{-1}.$$

By Theorem 2.11, we may take the map on S to be a single polygonal face with pairs of edges identified. Because we may obtain our map as the coarsening of a triangulation, we have an affine structure along the edges of the face, which may be used in the identification process.

If S is the sphere with model aa^{-1}, then we are done; otherwise, any spikes aa^{-1} may be eliminated without changing the topology of

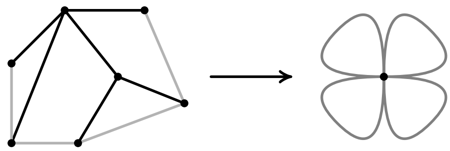


Figure 2.26. Collapsing a maximal tree.

the surface, and so we may assume that no symbol appears next to its inverse.

Before carrying out the inductive step, we make some definitions, and prove a lemma which allows us to assume that all vertices of the polygon are identified.

Definition 2.16. The *n-skeleton* of a triangulation (or in general, of a simplicial complex) is the union of all simplices of dimension $\leq n$.

In particular, the 1-skeleton is the 'frame' around which the triangulation is built; since it comprises 0-simplices (vertices) connected by 1-simplices (edges), it is in fact a graph. We can make the analogous definition for a map, and this is the concept we will now utilise.

Definition 2.17. A *tree* is a graph without cycles.

It is easy to show by induction that any (finite connected) graph admits a maximal tree, that is, a subgraph which is a tree and which is not properly contained in any other tree. This last condition is equivalent to the requirement that the subgraph contain every vertex of the graph. Unless the graph itself is a tree, maximal trees are never unique.

Figure 2.26 illustrates a common construction in algebraic topology; we consider a graph G, a maximal tree T (the darker edges in the picture), and identify T to a point, obtaining a quotient space G/T, which will have one vertex and n edges, and will be a 'bouquet' of circles all connected at a single point. G/T is *homotopic* to G, but not homeomorphic. Consequently, we will not use this construction directly, but will use it to motivate the proof of the following lemma.

Lemma 2.18. *Given a map $\mathcal{M}$ on a surface S, there exists a map $\tilde{\mathcal{M}}$ on S with the same number of faces as $\mathcal{M}$, but with only one vertex.*

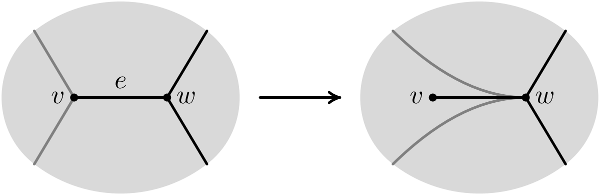


Figure 2.27. Turning a leaf into a spike.

Proof. Let G be the 1-skeleton of $\mathcal{M}$, and let T be a maximal tree of G. Consider a 'leaf' of T, that is, a vertex v which is connected to only one edge e (in T; its degree in G may be greater).

Let w be the vertex at the other end of e. As shown in Figure 2.27, take every other edge (besides e) which is attached to v, and attach it instead to w. This gives a new map on S in which the edge between v and w is a spike, and so may be eliminated.

We continue this procedure until T consists of just a single point; the resulting map is $\tilde{\mathcal{M}}$, and since the step of moving edges from v to w does not change the number of faces, we see that we are done. □

Further, because S is orientable, the direction of each identification is specified for us. Indeed, if any symbol a_0 appears twice (as $\ldots a_0 \ldots a_0 \ldots$, rather than $\ldots a_0 \ldots a_0^{-1} \ldots$), then we have the configuration seen in Figure 2.30. Moving a 'clock' once through H from a_0 to a_0 reverses its orientation, which cannot happen if S is orientable, and so we see that if a symbol a appears in the identifying sequence, so does its inverse a^{-1}.

Returning to our proof of the theorem, we now have a surface whose 1-skeleton is a bouquet of circles $a, b, c, \ldots$, which we draw as a polygonal map with certain identifications $a \sim a^{-1}$, etc. The next step is to find a handle.

Given any symbol a, we write the distance between a and a^{-1} as $\operatorname{dist}(a)$; here by 'distance' we mean the number of edges between a and a^{-1} as we proceed around the boundary of the polygon in either direction (whichever gives us the shorter distance). For example,

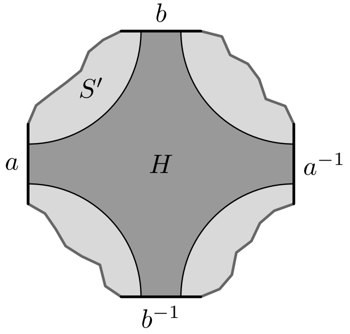


Figure 2.28. The configuration $aba^{-1}b^{-1}$.

in the sequence aa^{-1}, we have $\operatorname{dist}(a) = 0$, and in $abca^{-1}b^{-1}c^{-1}$, $\operatorname{dist}(a) = 2$.

Now choose a such that $\operatorname{dist}(a)$ is minimal; that is, $\operatorname{dist}(a) \leq \operatorname{dist}(b)$ for any other symbol b. Because we have eliminated spikes, we have $\operatorname{dist}(a) \geq 1$, so some symbol b lies between a and a^{-1}. Further, since $\operatorname{dist}(a)$ is minimal, b^{-1} cannot lie between a and a^{-1}, so our sequence must look something like $\ldots a \ldots b \ldots a^{-1} \ldots b^{-1} \ldots$.

This configuration is illustrated by Figure 2.28. The region H is homeomorphic to a torus with a hole, as shown in Figure 2.29.

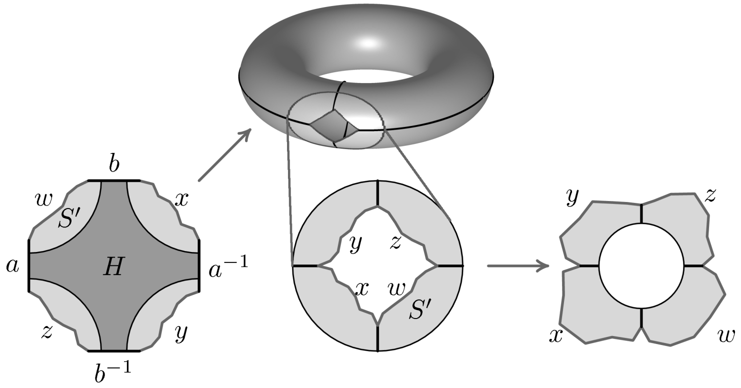


Figure 2.29. A visualisation of H and S'.

The lighter region, labeled S' and comprising four faces which are joined when a, a^{-1} and b, b^{-1} are identified, models a surface with a hole; Figure 2.29 illustrates the fact that upon filling the hole with a disc, we obtain a planar model which satisfies the conditions of our theorem, and which has four fewer edges than our original model. By induction, this is homeomorphic to the standard model of the sphere with m handles, for some value of m, and so reattaching the handle H shows that our surface S is homeomorphic to the standard model of the sphere with $m+1$ handles. □

Remark. In higher dimensions, a complete classification along these lines would be much more difficult to accomplish. Indeed, one of the great achievements in mathematics in recent years has been an essential completion of the classification of 3-manifolds, which was achieved by Perelman's proof of the Thurston geometrisation conjecture; in particular, this settled the famous Poincaré conjecture.

There are other models for the sphere with m handles besides the standard one; one of the most symmetric is given by the sequence of identifications

$$a_1 \dots a_{2m} a_1^{-1} \dots a_{2m}^{-1}$$

which is just the $4m$-gon with opposite sides identified. This and other models have the same topology as the standard model (recall Exercise 2.9), and so do not add any new surfaces to our list, but are sometimes useful for understanding various geometric structures which will appear later in this course. For example, in this model the edge identifications can be effected by parallel translations, and thus one obtains a Euclidean structure everywhere on the surface, with the exception of the vertices, which become 'super-conic' points at which the total angle is a multiple of 2π.

Example 2.19. For $m = 2$, all eight vertices of the regular octagon are identified, producing a sphere with two handles equipped with a Euclidean structure everywhere except at the single vertex, where the total angle is 6π. Later in this course we will see that this fact can be interpreted as a particular limit case of the celebrated Gauss-Bonnet Theorem.

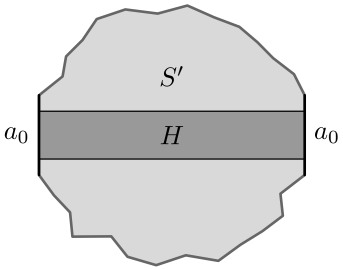


Figure 2.30. The configuration a_0a_0.

b. Non-orientable surfaces: Classification and models. Already we have seen that non-orientable surfaces may have several models of equal utility; while the projective plane is best represented as aa on a 2-gon, the Klein bottle may be thought of on a 4-gon both as $abab^{-1}$ or $aabb$. A similar situation continues to hold as we move to planar models with more sides; we are, however, able to use a similar process to the one above and obtain a complete classification.

As before, we may remove spikes and reduce to the case of a single vertex. If every pair of edges to be identified includes a symbol and its inverse, then the above proof applies, and the surface is orientable. Hence a non-orientable surface must include the configuration $\dots a_0 \dots a_0 \dots$, as shown in Figure 2.30. Following the same procedure as in the proof of the theorem, we may remove a Möbius cap from the surface and replace it with a disc to obtain a planar model, with fewer edges, of a surface S'.

If S' is orientable, we apply the theorem and have that our original surface S can be represented by the identifications

$$a_0a_0a_1b_1a_1^{-1}b_1^{-1}\dots a_mb_ma_m^{-1}b_m^{-1}.$$

If S' is not orientable, we use the same argument to remove another Möbius cap, and continue until we obtain either an orientable surface or the projective plane aa.

Recall that gluing a handle to a non-orientable surface is equivalent to gluing two Möbius caps. Hence we may write *any* non-orientable surface in terms of the identifications

$$a_1a_1a_2a_2\dots a_na_n,$$

which yield a sphere with n Möbius caps. This gives a canonical form for non-orientable surfaces, although there are others we could choose; for instance, we can use the above observations to write any non-orientable surface as either one or two Möbius caps attached to a sphere with handles.

Lecture 14

a. Chain complexes and Betti numbers. We now turn our attention to a concept which may at first appear quite unnatural, but which is in fact of great utility, and is central to much of modern mathematics; the idea of *homology.* As we will see, the initial definitions are purely algebraic, but the theory is central to modern topology, and also has broad applications in algebra and, somewhat surprisingly, also in analysis.

We begin with some linear algebra. Rather than considering a single linear transformation between two linear spaces, we consider a sequence of linear spaces with certain transformations between them. This is made precise as follows:

Definition 2.20. A *chain complex* $\mathcal{C}$ is a sequence of linear spaces C_k over some field (or more generally, modules over a ring) with linear maps $\partial_k \colon C_k \to C_{k-1}$, called *boundary operators*, which satisfy the identity $\partial_k \circ \partial_{k+1} = 0$.

Thus we have a picture reminiscent of an *exact sequence*:

$$0 \xrightarrow{\partial_{m+1}} C_m \xrightarrow{\partial_m} C_{m-1} \xrightarrow{\partial_{m-1}} \cdots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0.$$

The requirement that the composition of two consecutive boundary operators be trivial may be expressed setwise as $\operatorname{im}\partial_{k+1} \subset \ker\partial_k$; that is, the image of each boundary operator is a subspace of the kernel of the next. Exact sequences are characterised by the condition that this containment is in fact equality for every k. The *homology groups* associated with the chain complex $\mathcal{C}$ will, in some sense, measure how far it is from being exact.

Definition 2.21. Given a chain complex $\mathcal{C}$, the k^{th} *homology group* is the quotient

$$H_k(\mathcal{C}) = \ker \partial_k / \operatorname{im} \partial_{k+1}.$$

The elements of C_k are referred to as *chains*. For reasons which will become apparent when we discuss the application of chain complexes and homology to surfaces, we refer to elements of $\ker \partial_k$ as *cycles*, and elements of $\operatorname{im} \partial_k$ as *boundaries*. That is, cycles are chains which are taken to zero by the appropriate boundary operator, and boundaries are chains which may be obtained as the image of another chain under a boundary operator. Then the homology groups may be thought of as comprising cycles modulo boundaries.

The homology groups are quotients of the C_k, and hence carry the same structure. If the C_k are finite-dimensional vector spaces over $\mathbb{R}$ (or $\mathbb{C}$), then so are the homology groups. We refer to this as homology with coefficients in $\mathbb{R}$ (or $\mathbb{C}$) to indicate what structure the H_k possess. For such spaces, the only invariant is dimension; that is, two finite-dimensional vector spaces are isomorphic if and only if they have the same dimension. So we may describe the homology of $\mathcal{C}$ by the dimensions of the homology groups; these are the *Betti numbers* $\beta_k = \dim H_k(\mathcal{C})$.

Similarly, if the C_k are finitely generated abelian groups (finitely generated modules over $\mathbb{Z}$), then so are the homology groups, and we speak of $\mathbb{Z}$-homology, or homology with integer coefficients. In this case, we have the following fundamental result from algebra:

Proposition 2.2. *Any finitely generated abelian group G is isomorphic to $\mathbb{Z}^d \times F$, where F is finite, abelian, and may be written as the direct product of* primary cyclic groups *$\mathbb{Z}/p^k\mathbb{Z}$.*

This provides a decomposition of G into the *free part* $\mathbb{Z}^d$ and the *torsion part* F. We refer to d as the *rank of the free part*; the set $\{d, p_1^{k_1}, \dots, p_n^{k_n}\}$ uniquely determines the group G up to isomorphism, so the rank of the free part, together with the orders of the primary cyclic groups in the torsion part, provides us with a complete system of invariants.

In this case, we take the Betti number β_k to be the rank of the free part of $H_k(\mathcal{C})$; because this does not completely characterise $H_k(\mathcal{C})$, certain issues arise in the application of $\mathbb{Z}$-homology that do not arise when we use real or complex coefficients. This added intricacy is reflected in applications, where $\mathbb{Z}$-homology provides more information than homology with real or complex coefficients.

b. Homology of surfaces. With the exception of some suggestive terminology (cycles, boundaries, etc.), we have not yet drawn any connection between homological algebra and any geometrical concepts. In fact, we will find that the connections are rich and meaningful, and help to clarify the concepts just introduced by relating them to things we already know from our study of surfaces.

To this end, consider a surface with a triangulation $\mathcal{T}$, or more generally, any simplicial complex. We will define a chain complex $\mathcal{C}(\mathcal{T})$, examine the geometric interpretation of the spaces $C_k(\mathcal{T})$ and the boundary operators ∂_k, and find a striking relationship between the Euler characteristic $\chi(\mathcal{T})$ and the Betti numbers β_k. While the algebraic definition of $\mathcal{C}$ assigned no particular interpretation to the indices k, for our purposes here they are to be thought of as indicating the dimension of the objects from which C_k, H_k, β_k, etc., will be determined.

In what follows, we will use the $\mathbb{R}$-homology throughout; we could just as well use coefficients in $\mathbb{C}$, or in $\mathbb{Z}$, although in this latter case, certain technical issues arise, which were alluded to above, and we will postpone these for the time being.

The chain complex $\mathcal{C}$ is given by the sequence of spaces and operators

$$0 \xrightarrow{\partial_3} C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0.$$

We begin by describing the space $C_0(\mathcal{T})$; this is the set of linear combinations of the vertices of $\mathcal{T}$. This is on a purely formal level, and is not to be thought of as having any geometric meaning; perhaps the best visualisation is to put a single real number at each vertex,

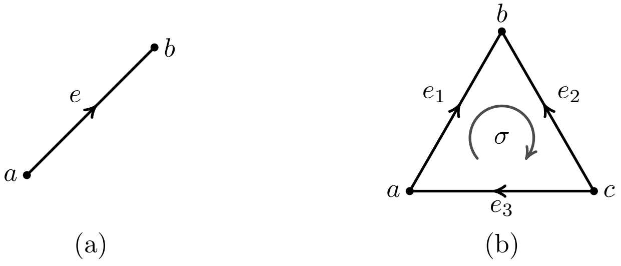


Figure 2.31. The boundary operator on (a) edges and (b) faces.

in which case each choice of real numbers corresponds to an element of $C_0(\mathcal{T})$.

Because the range of ∂_0 is trivial, the map itself must be trivial, so there is nothing to specify here.

What about $C_1(\mathcal{T})$? We begin by giving each edge of $\mathcal{T}$ an orientation; $C_1(\mathcal{T})$ is generated by these oriented edges, just as $C_0(\mathcal{T})$ was generated by the vertices. (In that case, we could not speak of the orientation of a single vertex, so the issue did not arise.) If we are doing $\mathbb{Z}$-homology, then for some edge e we can think of $n \cdot e$ as representing $|n|$ journeys along e in the direction specified when $n \geq 0$, and in the opposite direction when n is negative. If the coefficients are in $\mathbb{R}$ or $\mathbb{C}$, then it is probably best to think of the construction in a purely formal sense.

Our definition of ∂_1 will begin to demonstrate why the maps ∂_k are referred to as boundary operators. Given an oriented edge e, we must define $\partial_1(e)$ as a linear combination of vertices; once we have done this for each edge, ∂_1 will be defined on all of $C_1(\mathcal{T})$, since the oriented edges form a basis. Suppose our edge e runs from one vertex a to some other vertex b, as in Figure 2.31(a). Then we may define $\partial_1(e) = b - a$, so that ∂_1 of an edge is a linear combination of the boundaries of that edge.

Because ∂_0 is the zero map, the identity $\partial_0 \circ \partial_1 = 0$ is immediate, and needs no further verification.

Given our definitions of $C_0(\mathcal{T})$ and $C_1(\mathcal{T})$ as the linear spaces spanned by the oriented 0-simplices and 1-simplices, respectively, it is reasonable to expect that $C_2(\mathcal{T})$ ought to be spanned by the oriented 2-simplices, and this is indeed the definition we make. It is important to realise here that we do not impose any coherence requirement on these orientations; they are simply fixed arbitrarily for each face. A similar observation applies to the orientations of the 1-simplices.

Since ∂_3 has trivial domain, it must be a trivial map, so the only remaining piece of $\mathcal{C}$ to identify is the boundary operator $\partial_2 \colon C_2(\mathcal{T}) \to C_1(\mathcal{T})$. Analogously to the case with ∂_1, we will consider a 2-simplex σ and define $\partial_2(\sigma)$ as a linear combination of the 1-simplices e_i which form its boundary. The sign on each edge is determined by the relative orientations of σ and e_i; the edge is given a coefficient of $+1$ if the orientations agree, and -1 if they disagree. So for the 2-simplex shown in Figure 2.31(b), we have

$$\partial_2 \sigma = e_1 - e_2 + e_3.$$

Finally, we must verify that $\partial_1 \circ \partial_2 = 0$. (Again, ∂_3 is trivial, so $\partial_2 \circ \partial_3 = 0$ is immediate.) This is straightforward; in Figure 2.31(b), for example, we have

$$\partial_1 \partial_2 \sigma = (b - a) - (b - c) + (a - c) = 0.$$

These definitions can, of course, be continued for $k \geq 3$ in the case of manifolds or simplicial complexes of higher dimension. The primary difference is that the concept of orientation is no longer as straightforward to visualise, and must be defined in terms of even and odd permutations; this poses no additional technical difficulty, however.

c. A second interpretation of Euler characteristic. The geometric definition of $C_k(\mathcal{T})$ also lends some legitimacy to the use of the terms *chains*, *cycles*, and *boundaries* for elements of C_k, $\ker \partial_k$, and $\operatorname{im} \partial_k$, respectively. As a concrete example, consider an element of $C_1(\mathcal{T})$ such as $2e_1 + e_2 - e_3 + e_4$ as shown in Figure 2.32. The individual edges may be thought of as the links of a chain, which in general may lie in several pieces, as for instance in the chain $e_3 + e_4 \in C_1(\mathcal{T})$.

Due to the definition of the boundary operator ∂_1, a chain lies in the kernel of ∂_1 iff it ‘closes up’; neither of the examples just given lie in $\ker \partial_1$, although $e_1 + e_2 - e_3$ does. Similarly, the boundaries in

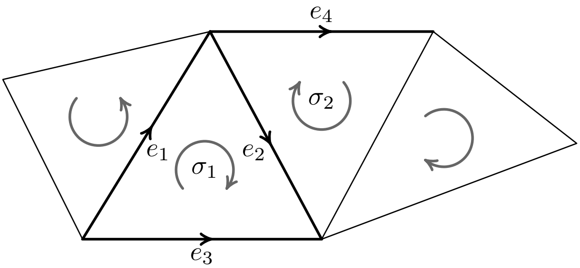


Figure 2.32. A chain of edges in $C_1(\mathcal{T})$.

C_1 are those chains which lie in the image of ∂_2, and these are seen to be the boundaries of a chain of 2-simplices. As always, orientation is important; $e_1 + e_2 - e_3$ is a boundary, but $e_1 + e_2 + e_3$ is not.

The condition that $\partial_k \circ \partial_{k+1} = 0$ implies that every boundary is a cycle; the question of which cycles are boundaries is precisely the issue at the heart of homology theory.

Let $B_k(\mathcal{T})$ be the dimension of the space of boundaries $\operatorname{im} \partial_{k+1}$, and $Z_k(\mathcal{T})$ the dimension of the space of cycles $\ker \partial_k$. Then the Betti number β_k, which is the dimension of the homology group $H_k(\mathcal{T})$, is given by $Z_k(\mathcal{T}) - B_k(\mathcal{T})$. We will now relate this to the Euler characteristic by determining the relationship between V, E, and F and the values of Z_k and B_k.

The above formula for the Betti numbers uses the following fundamental relation from linear algebra:

$$\text{dimension} = \text{rank} + \text{nullity}.$$

In our current context, this states that

$$\begin{aligned} \dim C_k &= \dim \operatorname{im} \partial_k + \dim \ker \partial_k \\ &= B_{k-1} + Z_k. \end{aligned}$$

Note that for $k = 0$, we just have $\dim C_0 = Z_0$ since ∂_0 is the zero map, and also that $B_2 = 0$ since ∂_3 is the trivial map. We now make the observation that $\dim C_0$ is just the number of vertices V, $\dim C_1$ is the number of edges E, and $\dim C_2$ is the number of faces F. Hence we have

$$\begin{aligned}\chi(\mathcal{T}) &= F - E + V \\ &= \dim C_2 - \dim C_1 + \dim C_0 \\ &= (B_1 + Z_2) - (B_0 + Z_1) + Z_0 \\ &= (Z_2 - B_2) - (Z_1 - B_1) + (Z_0 - B_0) \\ &= \beta_2 - \beta_1 + \beta_0.\end{aligned}$$

The Euler characteristic is the alternating sum of the Betti numbers! This provides an alternate definition of the Euler characteristic, which can easily be extended to higher dimensions for arbitrary simplicial complexes.

Remark. Since we know that the Euler characteristic is independent of a particular choice of triangulation, we obtain as a corollary that the alternating sum of the Betti numbers does not depend of the triangulation either. As we will soon see, the same applies to each Betti number separately.

Lecture 15

a. Interpretation of the Betti numbers. Although we now know that the Betti numbers tell us the Euler characteristic, we do not yet have a sense of what topological information they may carry on their own. This interpretation, however, turns out to be quite useful.

Proposition 2.3. *Any connected surface has $\beta_0 = 1$.*

Proof. Consider the edges of $\mathcal{T}$. Each has a boundary consisting of two vertices; if an edge e runs between vertices a and b, then $\partial_1(e) = b - a$, and so the sum of the coefficients of $\partial_1(e)$ is $1 + (-1) = 0$. It follows that the image of any linear combination of edges has coefficients which sum to zero; that is, every boundary in $C_1(\mathcal{T})$ has coefficients which sum to zero. It may be checked that this condition is sufficient; given a chain (of vertices) $\tilde{v} = \sum_{i=1}^{n} x_i v_i \in C_1(\mathcal{T})$ such that $\sum_{i=1}^{n} x_i = 0$, find a chain (of edges) $\tilde{e} \in C_2(\mathcal{T})$ which corresponds to a path from v_n to v_{n-1}. Then $\partial_1(x_n\tilde{e}) = x_n v_{n-1} - x_n v_n$, so $\tilde{v} + \partial_1(x_n\tilde{e})$ also has coefficients which sum to zero, but has only $n - 1$ non-zero coefficients. We may proceed by induction in this way to show that $\tilde{v} \in \operatorname{im} \partial_1$, so that $\tilde{v}$ is in fact a boundary. □

In general, β_0 is the number of connected components. We turn next to β_2, before returning to ponder the significance of β_1.

Proposition 2.4. *Any connected orientable surface has $\beta_2 = 1$; any non-orientable surface has $\beta_2 = 0$.*

Proof. Let $\tilde{\sigma} = \sum_{i=1}^n x_i \sigma_i \in C_2$ be a non-trivial chain (of faces), and consider under which circumstances we might have $\partial_2 \tilde{\sigma} = 0$. For each i, $\partial_2 x_i \sigma_i$ is a linear combination of three edges, each with coefficient $\pm x_i$. Since each edge e appears as a boundary of exactly two faces, say σ_i and σ_j, the coefficient of e in $\partial_2 \tilde{\sigma}$ will vanish iff $x_i \sigma_i$ and $x_j \sigma_j$ correspond to a coherent orientation of σ_i and σ_j. (Recall that in the definition of C_2, each face is assigned an arbitrary orientation, which is preserved by positive coefficients and reversed by negative ones.)

Now if $\tilde{\sigma} \in \ker \partial_2$, the orientations given to the faces by the coefficients are all coherent, and so the surface is orientable. Hence a non-orientable surface has $\beta_2 = 0$.

Conversely, an orientation on the surface gives rise to an element of the kernel, as just described. Because the surface is connected, an orientation on one face induces an orientation on all others, and so there is only one coherent orientation (up to sign); hence the kernel has only a single dimension, and $\beta_2 = 1$. □

Definition 2.22. An orientable surface is homeomorphic to a sphere with handles. The number of handles is the *genus* of the surface. For a non-orientable surface, which must be homeomorphic to a sphere with Möbius caps, the genus is the number of Möbius caps.

Consider a surface S of genus m. If S is orientable, we have $\chi = 2 - 2m$, since each handle reduces the Euler characteristic by 2, and also $\chi = \beta_0 - \beta_1 + \beta_2 = 2 - \beta_1$. For a non-orientable surface, each Möbius cap reduces χ by 1, and so $2 - m = \beta_0 - \beta_1 + \beta_2 = 1 - \beta_1$. We have proved the following:

Proposition 2.5. *For an orientable surface, β_1 is twice the genus. For a non-orientable surface, β_1 is the genus minus one.*

This development of the homology of a surface has so far depended on the particular triangulation $\mathcal{T}$. Indirectly, we have seen that the Betti numbers at least are independent of the choice of triangulation by giving them a topological interpretation; this is in some sense cheating, since it only works for surfaces and is not the most general proof.

In general, while the chain complex $\mathcal{C}$ depends on the triangulation $\mathcal{T}$, the homology sequence $\{H_k\}$ does not. The proof of this follows exactly the same lines as the proof of Theorem 2.4; we can define all the relevant concepts for maps as well as triangulations, and show that homology is preserved by barycentric subdivision, coarsening, and so on, as we did before. Notice that calculations with maps, which may have few vertices and faces, are much less cumbersome that those with triangulations.

Remark. Moving beyond surfaces, the notion of a *CW-complex* generalises the definition of a simplicial complex in a way which is similar to, but more complicated than, how maps generalise triangulations. These CW-complexes are fundamental objects in the study of modern topology, but lie beyond the scope of this course.

b. Torsion in the first homology and non-orientability. The above treatment has glossed over some subtle points that arise for non-orientable surfaces. On the projective plane, for instance, we have $\beta_1 = 0$, which suggests that every cycle of edges may be the boundary of a chain of faces. This is, however, not true; taking the sphere with antipodal points identified as our model, consider the path which runs halfway around the equator. This is a cycle and lies in the kernel of ∂_1, but is not in the image of ∂_2.

Here we see the difference between $\mathbb{R}$-homology and the richer $\mathbb{Z}$-homology; while the first $\mathbb{R}$-homology group of $\mathbb{R}P^2$ is trivial, the first $\mathbb{Z}$-homology group is $\mathbb{Z}/2\mathbb{Z}$, the group of two elements, whose presence is not noticed by the Betti number.

Exercise 2.13. Calculate the $\mathbb{Z}$-homology of the sphere with m Möbius caps.

c. Another derivation of interpretation of Betti numbers. By considering maps rather than triangulations, we can make the geometric interpretation of the Betti numbers β_0, β_1, and β_2 somewhat more transparent. As in the proof of Theorem 2.15, we may obtain any closed compact surface as a planar model on some $2n$-gon with all vertices identified. Then the chain complex

$$0 \longrightarrow C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$$

is given explicitly (using real coefficients) by

$$0 \longrightarrow \mathbb{R} \xrightarrow{\partial_2} \mathbb{R}^n \xrightarrow{\partial_1} \mathbb{R} \xrightarrow{\partial_0} 0$$

where $C_2 = \mathbb{R}$ is the space spanned by the single face, $C_1 = \mathbb{R}^n$ the space spanned by the n (pairs of) edges, and $C_0 = \mathbb{R}$ the space spanned by the single vertex. Because all vertices are identified, $\partial_1 = 0$; hence

$$H_0 = \ker \partial_0 / \operatorname{im} \partial_1 = \mathbb{R}/\{0\} = \mathbb{R}$$

and so $\beta_0 = 1$. Turning to the boundary operator ∂_2, we see that it takes the edges of the face and 'forgets' their order.[3] For example, if our model is an 8-gon σ with identifications $abc^{-1}dacb^{-1}d$, we have

$$\partial_2 \sigma = a + b - c + d + a + c - b + d = 2a + 2d.$$

If our surface is orientable, it is homeomorphic to the sphere with m handles, so $n = 2m$ and each symbol appears in pairs a, a^{-1}. Then $\ker \partial_2 = \mathbb{R} = C_2$ and we have

$$H_2 = \ker \partial_2 / \operatorname{im} \partial_3 = \mathbb{R}/\{0\} = \mathbb{R},$$

which gives $\beta_2 = 1$. In this case $\partial_2 = 0$ implies that

$$H_1 = \ker \partial_1 / \operatorname{im} \partial_2 = \mathbb{R}^{2m}/\{0\} = \mathbb{R}^{2m}$$

and so $\beta_1 = 2m$ is twice the genus.

If our surface is non-orientable, then $\ker \partial_2 = \{0\}$ and $\dim \operatorname{im} \partial_2 = 1$, so we have

$$\begin{aligned} H_2 &= \{0\}, \\ H_1 &= \mathbb{R}^n/\mathbb{R} = \mathbb{R}^{n-1}; \end{aligned}$$

[3] The reader with some knowledge of homotopy theory may see a relationship between this characterisation of ∂_2 and the fact that H_1 is the abelianisation of the fundamental group.

hence $\beta_2 = 0$, and $\beta_1 = n - 1$ is one less than the number of Möbius caps.

Of course, all of this relies on the fact that our development of homology theory for triangulations can also be carried out for maps. This extension is fairly straightforward, since any polygon has two possible orientations, which can in particular be derived from coherent orientations of triangles in any triangulation of the polygon. Then we may define the chain complex associated with a map as we did before, and proceed verbatim through the development of the theory.

Chapter 3

Differentiable Structure on Surfaces: Real and Complex

Lecture 16

a. Charts and atlases. Thus far we have considered primarily the topological properties of surfaces, and so the key concept has been that of a topological manifold, something locally homeomorphic to Euclidean space. The more combinatorial concepts of triangulations and maps have entered as auxiliary tools, giving the surface some extra structure which has proved useful in our classification programme, but also coming with two drawbacks. The first of these is technical; we have not established that these concepts are universally applicable, that is, that every surface admits a triangulation. The second drawback is more aesthetic, having to do with the fact that the extra combinatorial structure is not particularly natural; triangulations are effective theoretical tools, and maps have proved useful in performing computations and classifications, but neither is in any sense a natural generalisation of the definition of a topological manifold.

The purpose of the present chapter is to study an extra structure on manifolds which is quite natural; namely, that of a differentiable (or smooth) manifold. We begin by recalling the definition of a manifold in terms of coordinate charts, and then impose an added differentiability requirement on the transition maps from one patch (set of coordinates) to another.

Definition 3.1. A topological space S is a surface if it admits an atlas. An *atlas* $\mathcal{A}$ on S is a collection of open sets (*patches*) U_α together with maps (*charts*) $\phi_\alpha \colon U_\alpha \to \mathbb{R}^2$ such that

(1) The charts cover S; that is, $\bigcup_\alpha U_\alpha = S$.

(2) ϕ_α is a homeomorphism for every α.

Given two charts ϕ_α, ϕ_β, the *transition map* between them is

$$\phi_\beta \circ \phi_\alpha^{-1} \colon \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta).$$

We say that an atlas $\mathcal{A}$ is *differentiable* (or *smooth*) if every transition map is differentiable and has non-vanishing Jacobian determinant. Equivalently, each transition map is to be differentiable with differentiable inverse.

Note that the collection $\mathcal{A}$ may be infinite, or even uncountable. If we write $\phi_\beta \circ \phi_\alpha^{-1}(x,y) = (f(x,y), g(x,y))$, then the requirement that the Jacobian determinant is non-vanishing may be rewritten as

$$\det \begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{pmatrix} \neq 0$$

for every $(x,y) \in \mathbb{R}^2$. Given that the transition map is a bijection, this is equivalent to the condition that the inverse be differentiable.

How differentiable is a differentiable surface? The usual meaning of the word ‘smooth’, and the one which we will primarily use, is $\mathcal{C}^\infty$; that is, infinitely often differentiable. We could also consider $\mathcal{C}^r$-surfaces, for which the transition maps are only required to have continuous derivatives up to order r.[1]

The definition above can be generalised by replacing $\mathbb{R}^2$ with $\mathbb{R}^n$, in which case we replace the word ‘surface’ with the word ‘manifold’,

[1] The reader should be aware that depending on the context and the author, the word ‘smooth’ may only imply the existence of finitely many derivatives. In general, if no further specifics are given, the most common meaning of the word is “as smooth as need be to guarantee whatever results I’m about to claim.”

and speak about an n-dimensional *differentiable manifold* (or sometimes *smooth manifold*).

Definition 3.2. Two smooth atlases $\mathcal{A}$ and $\mathcal{B}$ on a surface S are *compatible* if the union is a smooth atlas.

In general, a single topological manifold may admit several different mutually incompatible smooth structures. For example, $\mathbb{R}$ is a one-dimensional smooth manifold with atlas $\mathcal{A}$ given by a single map, the identity $\mathrm{Id}\colon \mathbb{R} \to \mathbb{R}$. We may consider an atlas $\mathcal{B} = \{\phi\}$ which is also given by a single piecewise linear map

$$\phi(x) = \begin{cases} x & \text{if } x \leq 0, \\ 2x & \text{if } x \geq 0. \end{cases}$$

Because $\mathcal{B}$ comprises only a single chart, the only transition map is the identity map $\phi \circ \phi^{-1}$, hence the atlas is smooth. However, because $\phi \circ \mathrm{Id}^{-1} = \phi$ is not smooth, $\mathcal{A}$ and $\mathcal{B}$ are not compatible.

Although incompatible, these differentiable structures on the line, along with similarly obtained structures on other manifolds, are equivalent in a natural sense; namely, there exists a *homeomorphism* which takes one structure into the other. It turns out that in dimensions one (trivially) and two (via triangulation), *all* differentiable structures on a given manifold are equivalent in this sense. We will discuss this in more detail later; for now we content ourselves with observing that this fails in higher dimensions, where the situation becomes more bizarre. The 7-dimensional sphere, for example, admits 28 mutually non-equivalent differentiable structures.

A brief comment about the local invertibility condition is in order. In one dimension, the requirement that the Jacobian determinant be non-vanishing reduces to the condition that $f'(x) \neq 0$. Given a map $f\colon \mathbb{R} \to \mathbb{R}$ with this property, it follows that f^{-1} exists and is continuously differentiable. Notice that the inverse may exist if f' vanishes, but will not be $\mathcal{C}^1$; the standard example is $f\colon x \mapsto x^3$, for which the derivative of f^{-1} has a singularity at 0.

Moving up a dimension, the two-dimensional version of the Inverse Function Theorem states that if $f, g\colon \mathbb{R}^2 \to \mathbb{R}$ are $\mathcal{C}^1$, and $F = (f, g)\colon \mathbb{R}^2 \to \mathbb{R}^2$ has non-vanishing Jacobian determinant, then

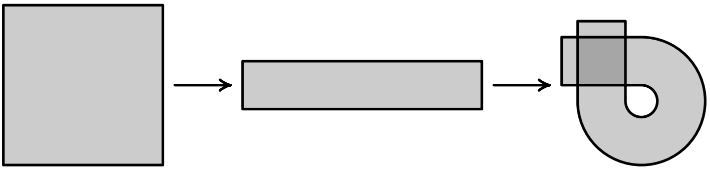


Figure 3.1. A regular map with no global inverse.

for any $(u_0, v_0) = F(x_0, y_0) = (f(x_0, y_0), g(x_0, y_0))$, there exists some neighbourhood U of (u_0, v_0) and a continuously differentiable map $\Phi = (\phi, \psi)\colon U \to \mathbb{R}^2$ such that $\Phi \circ F(x, y) = (x, y)$ for every $(x, y) \in \Phi(U)$, and that in addition,

$$\begin{pmatrix} \frac{\partial \phi}{\partial u} & \frac{\partial \phi}{\partial v} \\ \frac{\partial \psi}{\partial u} & \frac{\partial \psi}{\partial v} \end{pmatrix} \Bigg|_{(u_0, v_0)} = \begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{pmatrix} \Bigg|_{(x_0, y_0)}^{-1}.$$

An addendum to this theorem is that if F is in fact $\mathcal{C}^k$, then so is its inverse, so that regularity is passed to the inverse function.

Once local existence of the inverse has been established, the formula for the Jacobian follows from differentiating the equation $\Phi \circ F(x, y) = (x, y)$. It is important to recognise, however, that in the multi-dimensional case the theorem only guarantees *local* existence of an inverse. Figure 3.1 shows an example of a continuously differentiable map from the unit square to itself which has non-vanishing Jacobian determinant but which is not globally invertible. Thus in the definition of a smooth manifold, the existence of the global inverse of a transition map comes not from the Inverse Function Theorem, but from the bijective nature of the charts ϕ_α.

Finally, we note that none of this discussion has made any reference to metric properties of the surface. These will become important when we discuss Riemannian manifolds, but play no explicit role in the theory of differentiable manifolds.

b. First examples of atlases. We have seen a definition of smooth charts and atlases on a compact surface; the definition works equally well for non-compact surfaces, and it is natural to consider such cases since individual charts are already non-compact.

Example 3.3. The open disc $D^2 = \{ (x,y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1 \}$ is a non-compact surface on which we can place a smooth atlas with a single chart. Using polar coordinates, we may write the chart as

$$\begin{aligned} \phi\colon \quad D^2 &\to \mathbb{R}^2, \\ (r,\theta) &\mapsto (\rho(r),\theta), \end{aligned}$$

where $\rho\colon [0,1) \to [0,\infty)$ is to be a smooth function chosen so as to make ϕ smooth at the origin. We can play it safe and define ρ piecewise, setting it to be the identity map on $[0,\varepsilon)$ and then choosing a smooth extension which goes to infinity as $r \to 1$.

It is slightly trickier to write a single explicit formula. We offer one, inspired by elementary considerations from complex analysis. Consider the map from D^2 to the upper half-plane given by the complex equation

$$F(z) = \frac{1 - iz}{z - i}. \tag{3.1}$$

In real coordinates, this becomes

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

Composing this with the map $(x,y) \mapsto (x, y - \frac{1}{y})$, which maps the upper half-plane to the entire plane, we obtain the desired formula:

$$\Phi(x,y) = \left(\frac{2x}{x^2 + (y-1)^2}, \frac{1 - x^2 - y^2}{x^2 + (y-1)^2} - \frac{x^2 + (y-1)^2}{1 - x^2 - y^2} \right). \tag{3.2}$$

This example shows that if we so desire, we may use the open disc as the local model for a surface, rather than the entire plane; we will do this most of the time from now on.

Example 3.4. Another useful method is to use open rectangles as patches. In this case as well, a single chart is sufficient, since the map

$$(x,y) \mapsto \left(\tan\left(\frac{\pi a x}{b - a} + \frac{\pi(a+b)}{2(b-a)} \right), \tan\left(\frac{\pi c x}{d - c} + \frac{\pi(c+d)}{2(d-c)} \right) \right) \tag{3.3}$$

maps the rectangle $(a,b) \times (c,d)$ onto the whole plane.

Example 3.5. Any open subset $U \subset \mathbb{R}^2$ is a non-compact surface on which we can place a smooth atlas with infinitely many charts. In

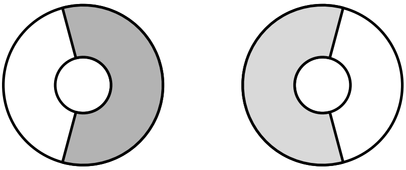


Figure 3.2. Two charts on an annulus.

particular, since U is open, for every point $x_0 \in U$ there exists $r > 0$ such that $B(x_0, r) \subset U$. Each open disc $B(x_0, r)$ is equivalent to the standard open unit disc via translation and homothety (isotropic expansion or contraction), whose composition forms the affine map

$$\phi_{x_0,r} \colon x \mapsto \frac{x - x_0}{r}$$

Since the previous example allows us to use the standard open disc as our model, these charts ϕ will form a smooth atlas provided the transition maps are smooth. But these transition maps are just the composition of two affine maps, and hence are affine maps themselves, so the result follows.

These examples show that we may use *any* open subset of $\mathbb{R}^2$, and not just the open disc or a rectangle, as our model for patches of a surface. Of course, to transform an atlas modeled on different open sets into an atlas modeled on the disc or the whole plane, we may have to increase the number of charts.

The last example used infinitely many charts (one for each point) to cover an open set; it is illuminating to consider what the minimum number of charts we can use is for various sets.

Example 3.6. By considering the annulus, we see immediately that it is not always possible to cover the set with a single chart. In this case, two charts are sufficient, as shown in Figure 3.2, and polar coordinates give a homeomorphism from each region to a rectangle in the (r, θ)-plane.

The same result holds for a cylinder, which is homeomorphic to the annulus. Indeed, the plane with any number of (round) holes can be covered with two charts, as shown in Figure 3.3, but not with one.

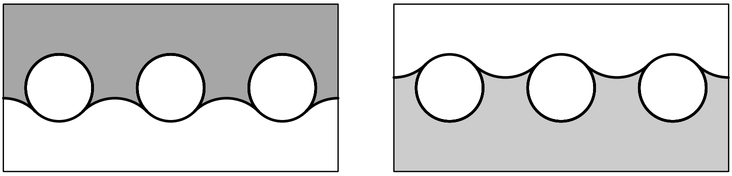


Figure 3.3. Two charts suffice to cover $\mathbb{R}^2$ with holes.

Lecture 17

a. Differentiable manifolds. We now have in our hands the definition of a differentiable manifold. While this definition is rather more involved than the definition of a topological manifold, it is in many ways a better object to work with. The key property of the latter was that at the local level, it has the topological structure of Euclidean space; by requiring that the differentiable structure be carried over as well, we place local coordinates on the manifold, which enable us to use the whole arsenal of tools from multivariable calculus.

It is worth noting that a particular set of local coordinates has no intrinsic meaning; the same smooth structure may be described by many different sets of local coordinates around a point. This fact has two important consequences in our treatment of smooth manifolds.

The first consequence is that we must always be concerned with how things behave with respect to allowable changes of coordinates; it is important to understand what happens on the regions where charts overlap when we work in the various sets of local coordinates which are available to us.

The second consequence is that we will eventually be motivated to establish coordinate-free notation for the objects with which we are concerned, and to give definitions which make no reference (or as little reference as possible) to a particular system of local coordinates. This will allow us to avoid the technical drudgery of working through coordinate changes at every turn.

We recall the definition of a smooth chart on a surface S; an open set $U \subset S$, together with a homeomorphism $\phi\colon U \to D^2$. The local coordinates on U are given by $\phi^{-1}: D^2 \to U$, as shown in Figure 3.4,

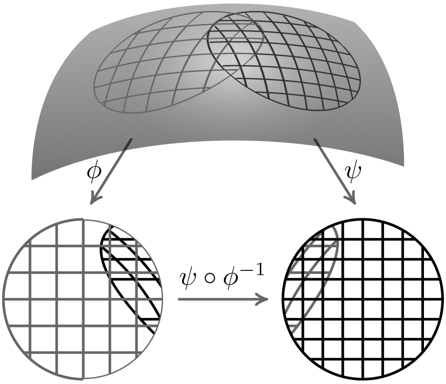


Figure 3.4. Two charts and their transition map.

and the condition that a collection of charts forms a smooth atlas is given by the requirement that the transition maps $\phi \circ \psi^{-1}$ be smooth and satisfy the conditions of the Inverse Function Theorem. That is, we require the Jacobian matrix to be invertible at each point, from which the theorem allows us to conclude the existence of a local inverse.

b. Diffeomorphisms. A major theme in modern mathematics is the investigation of various sorts of structures. To wit, we begin with a set X and proceed to list certain axioms or properties which are to be satisfied by the elements of X; in this way we may place on X the structure of a group, a metric space, a vector space, etc.

Having made this intrinsic definition, we must then confront the question of just what it means for two such objects to be indistinguishable from this intrinsic point of view. In order to answer this question, we must establish a particular equivalence relation on the class of all objects endowed with the structure we defined. These equivalence relations are fundamental to the study of these objects; some familiar examples are shown in the table.

Structure	Equivalence relation
sets	bijection
groups	isomorphism
linear transformations	conjugacy
metric spaces	isometry
topological spaces	homeomorphism
smooth manifolds	diffeomorphism

If we restrict to a subclass of examples of a particular structure, we use the same equivalence relation. So, for example, the proper equivalence relation on the class of finitely generated abelian groups is still isomorphism, just as it is for groups in general, and the equivalence relation for topological manifolds is still homeomorphism, since they form a subclass of the class of topological spaces.

The eventual goal, when it is possible, is to understand a particular sort of structure by obtaining a complete classification. That is, we explicitly construct a list of examples of the structure with the property that every other example of the structure is equivalent to something on our list. For example, this is accomplished by Jordan normal form in the case of (finite-dimensional) linear transformations.

It often happens that we must restrict to a subclass, as discussed in Lecture 4, in order to have any hope of a complete classification. For example, it is sheer folly to attempt a complete classification of all groups, but classification theorems have been obtained for finitely generated abelian groups, and even for finite simple groups. Similarly, topological spaces resist a general classification, but we have seen a classification of the subclass of two-dimensional topological manifolds; we will now consider the specific case of differentiable manifolds.

Definition 3.7. Given two smooth surfaces S and S' with atlases $\mathcal{A}$ and $\mathcal{A}'$, respectively, and a homeomorphism $f\colon S \to S'$, any chart $\phi\colon U \to D^2$ in $\mathcal{A}$ can be carried to a chart $\phi \circ f^{-1}\colon f(U) \to D^2$ on S'. If we do this for all charts in $\mathcal{A}$, we obtain a smooth atlas $\tilde{\mathcal{A}}$ on S'; we say that f is a *diffeomorphism* if $\tilde{\mathcal{A}}$ and $\mathcal{A}'$ are compatible.

This is a rather formal definition. To make it more intuitive, we may observe that $f\colon S \to S'$ is a diffeomorphism if it is a bijection whose representation $\phi \circ f \circ \psi^{-1}$ in any pair of local coordinates is a smooth function with invertible matrix of derivatives.

Exercise 3.1. Prove that the standard flat torus is diffeomorphic to the standard torus of revolution.

We are now faced with the problem of classifying smooth surfaces up to diffeomorphism. It is natural to ask whether all surfaces admit a differentiable structure, and if so, whether this structure is unique. The proof that the answer to both questions is yes will come later, via triangulations and maps. For the time being, we investigate the relationship between triangulations and smooth structures.

Proposition 3.6. *Given a triangulation $\mathcal{T}$ of a surface S, there exists a smooth atlas $\mathcal{A}$ on S, and vice versa.*

Idea of proof. To define a smooth atlas, we must first exhibit a chart at each point of the surface, and then show that the transition maps are smooth. There are three kinds of points on S; those lying in the interior of a 2-simplex, those lying on an edge, and those lying at a vertex. Since the affine coordinates on each 2-simplex provide a homeomorphism between its interior and the interior of a triangle, we have a natural chart on each interior point.

Edge points are also relatively straightforward to deal with; because the barycentric coordinates on neighbouring 2-simplices must agree on their edge of intersection, there is a natural homeomorphism between the interior of their union and the interior of a quadrilateral (the union of two triangles), which gives a chart at each point on the edge.

Vertices are another matter—we want to follow the same argument that we can simply take the union of the neighbouring 2-simplices and obtain a chart from the homeomorphism, but the naïve approach fails. The reason for this is that the angles around our vertex may not add up to 2π; recall our discussion in Lecture 3 of an ant or some other two-dimensional creature wandering around the surface of a dodecahedron. Points on an edge are indistinguishable from points on a face, but vertices are different, precisely for the reason that the sum of the angles may not be 2π (recall Figure 1.18).

Having understood the problem, it is no great challenge to address it properly, and we will do so in the next lecture.

The proof in the opposite direction, the construction of a triangulation given a smooth structure, requires certain tools in order to reconcile triangulations within different coordinate charts. We choose to do it using a Riemannian metric and taking short geodesic arcs as edges of the triangles; this will be discussed in Lecture 31. □

At this point it is natural to give an interpretation of orientability in terms of smooth structure. There are two possible orientations on $\mathbb{R}^2$, corresponding to declaring either clockwise or counterclockwise as the positive direction of rotation.[2] This orientation is naturally inherited by a chart on a surface; orientability of a surface corresponds to the possibility of choosing one of the two orientations on each chart so that they match on the intersections.

Exercise 3.2. Prove that if for a particular atlas on a surface (compact or not) the transition maps between any two charts have positive determinant at any point of intersection, then the surface is orientable.

c. More examples of charts and atlases. We can interpret some of the examples from the previous lecture as constructions of diffeomorphisms between various manifolds. Specifically, we proved that the disc and rectangles are diffeomorphic to the whole plane $\mathbb{R}^2$, with formulae (3.2) and (3.3) providing the corresponding diffeomorphisms.

More generally, we can say that a (two-dimensional) differentiable manifold which admits an atlas consisting of a single chart is diffeomorphic to $\mathbb{R}^2$. Motivated by this observation, we may look for more examples of this sort.

Example 3.8. Any bounded convex region U in the plane can be covered with a single chart. Simply fix a point $p \in U$; then the idea is to stretch or shrink each line segment from p to the boundary of U so that they are all the same length, and we obtain a copy of D^2. If

[2] The definition of orientability in higher dimensions is not quite so straightforward, and requires consideration of even and odd permutations.

we do this in the obvious linear way, we will obtain a map which is not differentiable at p, so we must construct it piecewise, as suggested last time for the diffeomorphism between D^2 and $\mathbb{R}^2$; near p the map is taken to be the identity, so that all the stretching and shrinking happens away from a neighbourhood of p.

This argument does not use the full strength of the convexity of U, but rather relies only on the fact that each ray from p intersects the boundary precisely once; a region U satisfying this condition for some $p \in U$ is said to be *star-shaped* or *convex from a point.*

Passing from open subsets of the plane—where there is a natural smooth structure provided by any covering with discs—to other more general surfaces, we face the problem of defining a natural smooth structure.

Consider, for example, the surface S defined by an equation of the form $F(x, y, z) = 0$, where F is a smooth function with no critical points at the zero level. Then at every point of S, at least one of the partial derivatives does not vanish, and hence by the Implicit Function Theorem, the corresponding coordinate can be expressed as a differentiable function of the other two. This gives a local chart in a small neighbourhood of the point. Compatibility—that is, smoothness of the transition maps between two charts—is obvious if the charts are obtained using the same coordinate but has to be checked if different coordinates are used. This will be done carefully in the next lecture. In the meantime, let us consider the specific example of the round sphere.

Example 3.9. Following the above recipe, one can try to project the sphere to each of the coordinate planes; consider first the horizontal xy-plane. Each hemisphere projects bijectively onto the unit disc and is thus covered by a chart, as was shown in Figure 1.10. The equator is not covered by either of these charts (recall that patches are open sets), but it is covered by the four charts arising from the projections to the four remaining vertical coordinate planes.

Now we must check compatibility of the charts by confirming that the transition map between any two charts is differentiable. Without loss of generality, consider the intersection of the two charts on the hemispheres defined by $y > 0$ and $z > 0$. In the first chart, x and z may serve as coordinates; in the second, we use x and y. The coordinate charts are given by the projections $\phi\colon (x, y, z) \mapsto (x, z)$ and $\psi\colon (x, y, z) \mapsto (x, y)$, and their inverses are

$$\phi^{-1}(x, z) = \left(x, \sqrt{1 - x^2 - z^2}, z\right),$$
$$\psi^{-1}(x, y) = \left(x, y, \sqrt{1 - x^2 - y^2}\right).$$

These lead to the transition maps

$$\psi \circ \phi^{-1}(x, z) = \left(x, \sqrt{1 - x^2 - z^2}\right),$$
$$\phi \circ \psi^{-1}(x, y) = \left(x, \sqrt{1 - x^2 - y^2}\right),$$

and we see that

$$\frac{\partial y}{\partial z} = \frac{-2z}{\sqrt{1 - x^2 - z^2}} < 0,$$
$$\frac{\partial z}{\partial y} = \frac{-2y}{\sqrt{1 - x^2 - y^2}} < 0.$$

Since obviously $\partial x / \partial x = 1$, these derivatives are precisely the corresponding Jacobian determinants, and so the transition maps $(x, z) \mapsto (x, y)$ and $(x, y) \mapsto (x, z)$ from one coordinate system to the other both satisfy the compatibility condition. A similar analysis goes through for transition maps involving the other four charts in our atlas.

As we have already seen, the atlas in this example is not the only atlas we might put on the sphere. There are other, more economical, atlases available, which turn out to generate the same differentiable structure.

As the sphere is compact, and hence not homeomorphic to the plane, it cannot be covered with a single chart; however, it can be covered with just two, via stereographic projection from two antipodal points. Notice that stereographic projection from a point p maps the $S^2 \setminus \{p\}$ onto the plane, thus providing a chart on the sphere *with a single point removed*, as was shown in Figure 1.11.

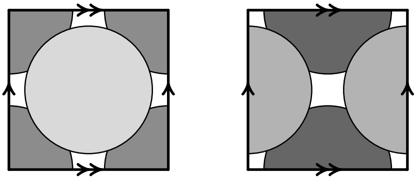


Figure 3.5. An atlas on $\mathbb{T}^2$ with four charts.

Exercise 3.3. Prove that the stereographic projections from two antipodal points are compatible with each other and with the hemispheric charts described above.

Exercise 3.4. Let F be a differentiable convex function on $\mathbb{R}^3$; i.e.

$$F(t\vec{x} + (1-t)\vec{y}) \leq tF(\vec{x}) + (1-t)F(\vec{y})$$

for every $\vec{x}$, $\vec{y} \in \mathbb{R}^3$ and $t \in [0,1]$. Suppose that c is a regular value of F and that the surface $F = c$ is non-empty and compact. Show that this surface is diffeomorphic to the standard sphere given by $x^2 + y^2 + z^2 = 1$.

Example 3.10. The standard flat torus illustrates another way to introduce a natural differentiable structure, which is somewhat less visual, but which does not require even elementary calculations of the kind we performed for the sphere. Namely, one simply *projects* the standard smooth structure from the plane $\mathbb{R}^2$ to the torus $\mathbb{R}^2/\mathbb{Z}^2$. To do that, notice that every disc of radius less than 1/2 is mapped injectively to the torus by the natural projection, and thus defines a chart. As in the case of open subsets of the plane, the transition maps between local coordinates coming from different discs are given by affine maps, hence the charts are compatible.

One can immediately address the question of the minimal number of charts required. Four is obviously sufficient (Figure 3.5)—the torus is covered by the discs of radius 2/5 centered in the center of the square, at the midpoints of two non-identified sides, and at the vertex. If instead of discs one uses certain other domains in the plane which allow a single chart, one can reduce this number to three. Two charts are not sufficient to cover the torus, but the proof is far from straightforward.

Lecture 18

Until the last few examples of the preceding lecture, we had dealt primarily with smooth manifolds in the abstract. Those examples illustrated some possible techniques for defining smooth structures on particular sorts of manifolds; we will now examine systematic methods for this process. Since the definition of a smooth surface is given in terms of charts from the surface to the plane, the first idea is of course to inherit a smooth structure directly from the plane by defining the charts explicitly. We will also see examples in which a surface inherits its smooth structure from another surface with which we are already familiar.

a. Embedded surfaces. First, consider embedded surfaces in $\mathbb{R}^3$. That is, let $F\colon \mathbb{R}^3 \to \mathbb{R}$ be a smooth function, let $S = \{\,(x,y,z) \in \mathbb{R}^3 \mid F(x,y,z) = 0\,\}$ be its zero set, and assume that 0 is a regular value for F. There are two basic methods of associating a smooth atlas with this surface.

Coordinate projections. Each $(x,y,z) \in S$ is a regular point, and hence the gradient $\nabla F(x,y,z) \neq 0$, so the Implicit Function Theorem gives us coordinate charts around each point via projection to one of the three coordinate planes in $\mathbb{R}^3$, as we have already seen. The Implicit Function Theorem also guarantees smoothness of the transition maps, and it follows that these projections define a smooth atlas on S.

Tangent plane projections. We may also take a more symmetric and geometrically natural approach and project not to the coordinate planes, but to the tangent planes at each point. Aside from an intrinsic aesthetic appeal, this method has an advantage of distorting geometry of the surface in the minimal possible way, since projection carries a neighbourhood of each point to the plane best fitted to the surface at that point. A disadvantage of this method is that it is more complicated computationally.

b. Gluing surfaces. A second method for obtaining a smooth structure is to take two or more surfaces on which we have a known smooth structure, cut a certain number of holes in each of them, and then

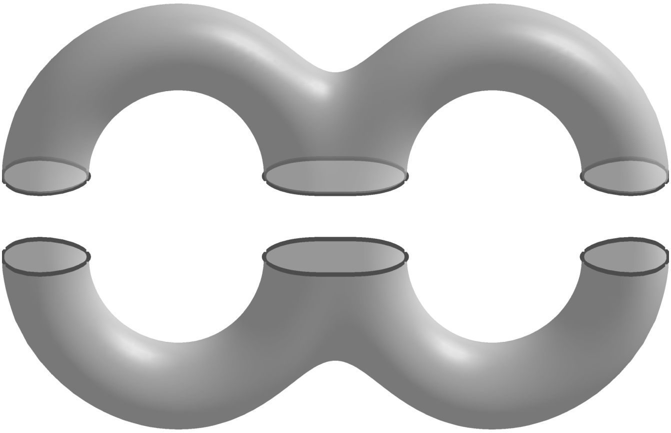


Figure 3.6. Gluing to get a sphere with two handles.

glue along these holes. This is illustrated in Figure 3.6, which shows two copies of a sphere with three holes being glued together to obtain a surface of genus two.

Topologically, a sphere with a hole is simply a disc, so a sphere with n holes is homeomorphic to a disc with $n-1$ holes (Figure 3.7). There is a natural smooth structure on the disc, coming from the plane, and so each of the two pieces in Figure 3.6 has a smooth structure. If we are careful in how we glue the two pieces together, and glue along a neighbourhood of the boundary (since the patches need to overlap), then it may be checked that these give rise to a smooth structure on the union, which in this case is a sphere with two handles.

Since we saw in the previous lecture that a disc with any number of (circular) holes can be covered with two charts, this construction shows that a sphere with any number of handles admits an atlas with four charts; by the classification theorem, we can now put a smooth structure on any orientable surface which admits a triangulation.

c. Quotient spaces. A third construction is available any time we are considering a quotient space and already have a smooth structure on the covering surface. This was the case in the example at the

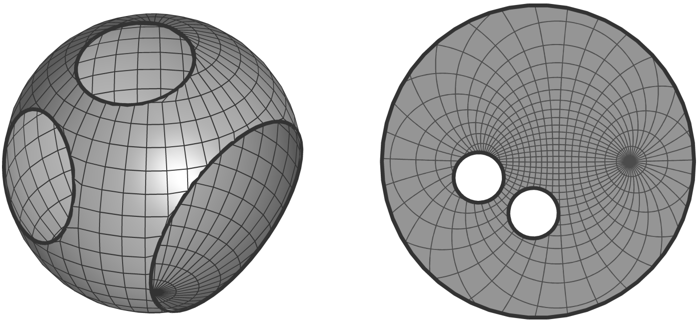


Figure 3.7. A sphere with three holes.

end of the previous lecture, where the torus was to inherit its smooth structure from the plane, its cover.

Suppose $\pi\colon \tilde{S} \to S$ is a quotient map. We would like to define charts on S as images of charts on $\tilde{S}$; unfortunately, this fails in general, because if we begin with 'too large' a patch U on $\tilde{S}$, the map $\pi\colon U \to \pi(U)$ may not be injective. However, because open subsets of patches can also be used as patches, we can guarantee injectivity by only considering images of charts whose patches are 'small enough' in precisely that sense.

For the example of the flat torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$, the covering space was $\mathbb{R}^2$, and the quotient map π was the map taking each point to its equivalence class under integer translations. Then the condition that a patch on $\mathbb{R}^2$ be 'small enough' is the requirement that it contain no more than one element from each equivalence class; for example, a disc of radius $> \sqrt{2}$ is mapped to the whole torus by π, and is too large, while any disc of radius $< 1/2$ works just fine.

Applying this approach to the sphere with six charts given by projecting each hemisphere to the appropriate coordinate plane, we obtain a smooth atlas on the projective plane which consists of three charts, since each pair of opposite hemispheres need only be covered once in the quotient space. This same technique lets us put a smooth atlas on any non-orientable surface admitting a triangulation, since any such surface is has an orientable double cover, which admits a smooth structure as discussed above.

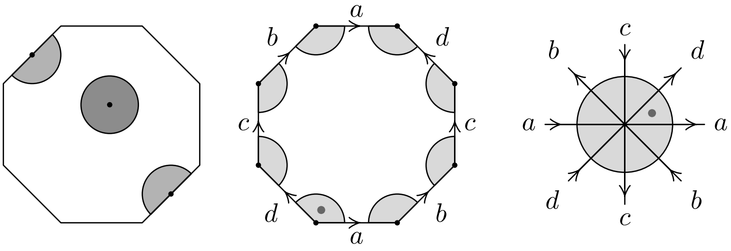


Figure 3.8. Patches for a planar model on the octagon.

We will show later that *any* compact surface admits a smooth atlas with just three charts. It is much more difficult to prove that this is nearly always optimal, in that except in the case of the sphere (which can be covered with just two charts via stereographic projection), no smooth atlas with two charts exists.

d. Removing singularities. Suppose we wish to put a smooth structure on a sphere with two handles via the standard planar model on an octagon, using the method just described for quotient spaces. At points in the interior of the octagon, and along the edges, there is no trouble; as shown in Figure 3.8, we may simply take as a patch containing the point a small disc which does not contain any vertices of the octagon, and use as our chart the standard affine map between our disc and the standard one.

Around the single vertex v (recall that all eight vertices of the octagon are identified), things do not work out quite so neatly. A small disc around v has eight components in the planar model, each of which is a pie piece subtending an angle of $3\pi/4$. Combinatorially, they are to be ordered cyclically to give a neighbourhood homeomorphic to a disc, but this cannot be carried out naïvely since their angles sum to 6π, rather than 2π. Thus it is not immediately obvious what the homeomorphism between the disc around v and the standard disc ought to be.

There are several ways of resolving this difficulty, which is representative of a whole class of situations where a natural structure possesses isolated singularities. One particularly elegant solution comes from complex analysis.

The map $z \mapsto z^3$ is a smooth map from the unit disc in the complex plane to itself under which each point (besides the origin) has exactly three preimages. We may choose a particular branch of the inverse function $z \mapsto z^{1/3}$ so that the wedge

$$\left\{ z \in \mathbb{C} \;\middle|\; |z| < 1,\ \arg z \in \left[0, \frac{3\pi}{4}\right] \right\}$$

is mapped to the wedge

$$\left\{ z \in \mathbb{C} \;\middle|\; |z| < 1,\ \arg z \in \left[0, \frac{\pi}{4}\right] \right\}.$$

If we 'squeeze' each of the eight wedges in this way (after a suitable rescaling to obtain a radius of 1), their union is precisely the unit disc (after appropriate rotations), as shown in Figure 3.8. Notice that the transition maps between this chart and the charts around nearby points in the interior are in fact smooth. This can be easily seen since the interior and edge charts are obtained from the standard Euclidean coordinates by affine transformations, so the transition functions to and from the vertex chart are given by compositions of affine transformations with the coordinate expression of the function $z \mapsto z^3$ and its inverse away from the origin, which satisfy both differentiability and Jacobian invertibility conditions.

A version of this method allows us to introduce a smooth structure on a surface with a triangulation; by adjusting angles of triangles near a vertex, their sum may be made equal to 2π in a way compatible with the natural smooth structure inside the triangles and around the edges.

Lecture 19

a. Riemann surfaces: Definition and first examples. The idea of using one complex variable rather than two real ones for our coordinate charts has very rich results, and the method of the previous example bears more fruit than one might at first expect. We must begin our (brief) foray into the subject of complex manifolds with a definition from basic complex analysis.

Definition 3.11. Given an open domain $U \subset \mathbb{C}$, a map $f\colon U \to \mathbb{C}$ is *holomorphic* if the derivative

$$f'(z) = \lim_{h \to 0} \frac{1}{h}(f(z+h) - f(z))$$

exists for every $z \in U$.

For computational purposes, existence of f' is often checked via the *Cauchy-Riemann equations.* Geometrically, the requirement is that f preserve (signed) angles between smooth curves as a map from $\mathbb{R}^2$ to $\mathbb{R}^2$; such a map is called *conformal.*

In striking contrast to the real case, existence of a single complex derivative is enough to guarantee that f is smooth, and even analytic; not only must f have infinitely many continuous derivatives, but there is a neighbourhood around each point $z \in U$ on which the power series expansion of f converges absolutely to f. This equivalence of holomorphicity and analyticity for complex functions is one of the most fundamental theorems in complex analysis.

A consequence of this is that the class of holomorphic functions $\mathbb{C} \to \mathbb{C}$ is in some sense smaller and more rigid than the class of differentiable, or even smooth, functions $\mathbb{R}^2 \to \mathbb{R}^2$. Given two smooth functions f and g on separated domains $U, V \subset \mathbb{R}^2$, we can 'glue' them together to obtain a smooth function $h\colon W \to \mathbb{R}^2$, where $W \supset U \cup V$, with $h|_U = f$, $h|_V = g$. The principle of analytic continuation prevents a similar procedure from being possible in the complex plane.

Definition 3.12. A *complex manifold* is a topological space equipped with a holomorphic atlas; that is, each point has a neighbourhood homeomorphic to the open disc in $\mathbb{C}$, such that the transition maps between charts are holomorphic. A one-dimensional complex manifold is called a *Riemann surface.*

Note that a Riemann surface has one *complex* dimension, and hence two *real* dimensions, so it is in fact a surface in the sense that we have been discussing.

There is a natural complex structure on $\mathbb{R}^2$ making it into the Riemann surface $\mathbb{C}$. Notice that since the model for a chart is a unit disc and not $\mathbb{C}$ itself (a distinction which will have a significance here that it did not in the real case), we need to produce an atlas. We can of course cover $\mathbb{C}$ by infinitely many discs and observe that the transition maps are translations, which are obviously holomorphic.

More interestingly, one can produce a finite atlas by noticing that there is an invertible holomorphic function which maps any half-plane onto the unit disc. For the upper half-plane, this function is given by the inverse F^{-1} of equation (3.1), and for an arbitrary half-plane H by the composition of F^{-1} with a complex affine map taking H into the upper half-plane. Thus one can take two half-planes $\operatorname{Im} z > 0$ and $\operatorname{Im} z < 1$ as charts.

Any connected open domain in $\mathbb{C}$ inherits a complex structure from $\mathbb{C}$ itself, and hence is a Riemann surface. Together with the above observation, this shows, as in the real case, that we can use any simply connected open subset of $\mathbb{C}$, including $\mathbb{C}$ itself, as the model for our atlas, rather than restricting ourselves to the unit disc.

Another set of examples are the flat tori $\mathbb{C}/L$ introduced in Exercise 1.22, since transition maps are given by translations, which in complex notation have the form $z \mapsto z + c$ and are obviously differentiable as complex functions. We consider these surfaces in more detail later.

The next example is the sphere S^2, which can be made into the *Riemann sphere* by equipping it with a complex structure as follows:

As a topological space, the sphere is the one-point compactification of the plane. Setwise, we write this as $S^2 = \mathbb{C} \cup \{\infty\}$; that is, we obtain the Riemann sphere by adding a point at infinity to the complex plane. Then we have an atlas consisting of two charts; the first is given by the identity map $\mathbb{C} \to \mathbb{C}$, and the second is given by the reciprocal map

$$\begin{aligned} S^2 \setminus \{0\} &\to \mathbb{C}, \\ z &\mapsto \frac{1}{z}. \end{aligned}$$

These are very closely related to stereographic projection; topologically speaking, the atlases are equivalent, and a comparison of the formulae is left as an exercise.

The Riemann sphere is arguably the most important example of a Riemann surface, even more so than $\mathbb{C}$ itself. As justification for this claim, consider a fractional linear transformation of the form

$$f\colon z \mapsto \frac{az+b}{cz+d}$$

where $a, b, c, d \in \mathbb{C}$. As a map from $\mathbb{C}$ to itself, f has a pole at $z = -d/c$, where the transformation is undefined, and the range of the transformation is not the entire complex plane, but rather $\mathbb{C} - \{a/c\}$. However, if we consider f as a transformation of the Riemann sphere S^2, then it is in fact a bijection, with $f(-d/c) = \infty$ and $f(\infty) = a/c$.

A similar consideration applies to any rational function

$$f\colon z \mapsto \frac{P(z)}{Q(z)}$$

where P, Q are polynomials in z. By including the point at infinity in our space, we allow the map to be well-defined everywhere, although for non-linear polynomials it will no longer be one-to-one.

So the sphere is a Riemann surface; what about the other compact surfaces? Consider the construction in the last lecture of a smooth structure on the surface of genus two via its planar model on the octagon with pairs of opposite sides identified. In light of our subsequent definition of complex structure, we see that this is in fact what we constructed, and so the surface of genus two can be made into a Riemann surface. An immediate generalisation of this procedure to polygons with more sides shows that any compact orientable surface, whatever its genus, can be made into a Riemann surface.

This leaves the non-orientable surfaces to be considered. We know that every non-orientable surface has an orientable double cover, which admits a complex structure by the above discussion. It is natural to ask whether this structure is inherited by the non-orientable quotient space; for example, is $\mathbb{R}P^2$ a Riemann surface, inheriting a complex structure from the Riemann sphere? In fact, it is not; because holomorphic maps preserve *signed* angles (which we will see later in this lecture), it can be shown that any surface admitting a holomorphic structure is in fact orientable, which prohibits the existence of such a structure on the projective plane, or any other non-orientable surface. The essential obstacle is the fact that complex conjugation, $z \mapsto \bar{z}$, is not a holomorphic map, because while it preserves the magnitude of angles, it does not preserve their sign.

b. Holomorphic equivalence of Riemann surfaces. Just as the notion of diffeomorphism provided an equivalence relation for real differentiable surfaces, there is a notion of *holomorphic equivalence* for complex (Riemann) surfaces. As we will soon see, rigidity of holomorphic functions implies that complex manifolds which are diffeomorphic as real differentiable manifolds may not be holomorphically equivalent. In particular, the unit disc is not holomorphically equivalent to $\mathbb{C}$.

The existence of Riemann surfaces which are not holomorphically equivalent despite being diffeomorphic means that holomorphic equivalence is stronger than diffeomorphism. Thus, although a given surface admits only one smooth structure up to diffeomorphism, it often admits many different non-holomorphic complex structures.

As a first example of this phenomenon, we demonstrate that although the disc D^2 and the plane $\mathbb{C}$ are equivalent as smooth manifolds, they are *not* equivalent as complex manifolds. We demonstrated that they are diffeomorphic by showing that D^2 is diffeomorphic to the upper half-plane, which is in turn diffeomorphic to $\mathbb{C}$. The first of these diffeomorphisms can in fact be chosen to be a holomorphic equivalence, as in equation (3.1), so that D^2 is equivalent to the upper half-plane as a Riemann surface. However, the diffeomorphism from the upper half-plane to $\mathbb{C}$ is not a holomorphic equivalence, and in fact, no such equivalence exists.[3]

The fact that D^2 and $\mathbb{C}$ are not holomorphically equivalent is a consequence of *Liouville's theorem*, which states that any function on the entire complex plane $\mathbb{C}$ which is both holomorphic and bounded must in fact be constant; this is in turn a consequence of Cauchy's integral formula, one of the fundamental results in complex analysis. Once this theorem is known, we can simply observe that any polynomial $p(z)$ is holomorphic and bounded on D^2, so that if $\psi\colon \mathbb{C} \to D^2$ were a holomorphic equivalence, then $\psi \circ p$ would be a bounded holomorphic function on $\mathbb{C}$, contradicting the theorem.

[3] The reader with some knowledge of hyperbolic geometry may wish to consider the ramifications of this paragraph in that context.

Using various elementary functions, one can construct holomorphic equivalences between the unit disc and a variety of plane domains.

Exercise 3.5. Construct a holomorphic equivalence between the unit disc $D^2 = \{\, z \in \mathbb{C} \mid |z| < 1 \,\}$ and

(1) the strip $\{\, z \in \mathbb{C} \mid 0 < \operatorname{Re} z < 1 \,\}$;

(2) the half-strip $\{\, z \in \mathbb{C} \mid 0 < \operatorname{Re} z < 1,\ 0 < \operatorname{Im} z \,\}$;

(3) the unit square $\{\, z \in \mathbb{C} \mid 0 < \operatorname{Re} z < 1,\ 0 < \operatorname{Im} z < 1 \,\}$;

(4) the exterior of an ellipse, together with the point at infinity.

In fact, the holomorphic non-equivalence between the complex plane and the disc (and hence any of the domains listed above) makes the whole plane rather an exception. The celebrated *Riemann Mapping Theorem* asserts that any *proper* subset of $\mathbb{C}$ homeomorphic to the unit disc is, in fact, holomorphically equivalent to it.

A further example of diffeomorphic Riemann surfaces which are not holomorphically equivalent is given by complex tori. As in Exercise 1.22, consider *any* lattice in the complex plane given by

$$L = \{\, mu + nv \mid m, n \in \mathbb{Z} \,\}$$

where $u, v \in \mathbb{C} = \mathbb{R}^2$ are linearly independent over $\mathbb{R}$. Then $\mathbb{C}/L$ is a Riemann surface which is diffeomorphic to the standard flat torus $\mathbb{T}^2 = \mathbb{C}/\mathbb{Z}^2$, but which may carry a different complex structure.

Proposition 3.7. *Given two lattices $L_1, L_2 \subset \mathbb{C}$, the tori $\mathbb{C}/L_1$ and $\mathbb{C}/L_2$ are equivalent as Riemann surfaces if and only if there is a* linear *holomorphic function F such that $FL_1 = L_2$.*

Proof. We will use Liouville's theorem once again.

Let $f\colon \mathbb{C}/L_1 \to \mathbb{C}/L_2$ be a holomorphic equivalence. Extending to the covering space $\mathbb{C}$, we obtain a holomorphic function F on $\mathbb{C}$ which maps each translation of L_1 onto a translation of L_2. Take two generators u and v of L_1. Then $F(z+u) - F(z) \in L_2$ and $F(z+v) - F(z) \in L_2$ for every $z \in \mathbb{C}$, hence both differences are constant because they vary continuously within the discrete set L_2. Differentiating, we obtain $F'(z) = F'(z+u) = F'(z+v)$, and so F'

is holomorphic and doubly periodic, hence bounded. By Liouville's theorem, it is a constant, and hence F is a linear function. □

There are other interesting examples of holomorphic equivalence. For example, since holomorphic maps act transitively on the Riemann sphere, the Riemann sphere with *any* point removed is holomorphically equivalent to $\mathbb{C}$.

The exponential function has period $2\pi i$ and misses the values 0 and ∞. Moreover, it takes every other value exactly once in the domain $0 \leq \operatorname{Im} z < 2\pi$. Thus it gives a holomorphic equivalence between the cylinder $\mathbb{C}/2\pi i\mathbb{Z}$ and the Riemann sphere with the two points 0 and ∞ removed.

Exercise 3.6. Prove that the Riemann sphere with any two points removed is holomorphically equivalent to the cylinder $\mathbb{C}/u\mathbb{Z}$ for any $u \in \mathbb{C}$.

Exercise 3.7. Prove that the Riemann spheres with any three points removed are all holomorphically equivalent.

c. Conformal property of holomorphic functions and invariance of angles on Riemann surfaces. It is instructive to consider the geometric significance of all this; what geometric structure is preserved by complex equivalence that is not preserved by smooth equivalence?

We began our discussion of surfaces with purely topological considerations; at the local level, a surface looks like $\mathbb{R}^2$, so we can define *coordinates* on the surface. By adding a smooth structure and requiring that the transition maps $\phi \circ \psi^{-1}$ be not only continuous, but differentiable, we gave meaning to the notion of *direction* on the surface; we will soon examine this in more detail when we consider tangent spaces. Most recently, we have added a complex structure, which demands that the transition maps be holomorphic; geometrically, they must preserve signed angles, so we can now speak about *angles* on the surface without reference to a particular coordinate chart.

Let us make this more explicit. Given two smooth curves γ and η on our surface which intersect in a point p, we may take a chart ϕ on

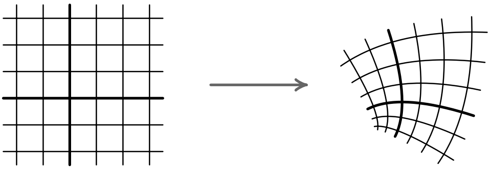


Figure 3.9. Conformal maps preserve angles.

a neighbourhood of p. Then $\phi(\gamma)$ and $\phi(\eta)$ are smooth curves in the complex plane which intersect at 0. We may take the tangent lines to these curves, measure the (signed) angle between them, and then declare this to be the angle between γ and η at p. Had we taken some other chart ψ, we would have measured the angle between the two curves $\psi(\gamma)$ and $\psi(\eta)$ in $\mathbb{C}$; however, because these are the images of the curves $\phi(\gamma)$ and $\phi(\eta)$ under the transition map $\psi \circ \phi^{-1}$, which preserves signed angles because it is holomorphic, we would obtain the same measurement. Thus our definition is independent of the particular choice of coordinate chart.

The fact that holomorphic functions preserve angles is a standard one from complex analysis, and is not difficult to see using some basic ideas from calculus. In the context of functions of one real variable, the usual linear approximation to $f\colon \mathbb{R} \to \mathbb{R}$ at a point x_0 is given by the map

$$x \mapsto f(x_0) + f'(x_0)(x - x_0)$$

and has a graph which is simply the tangent line at $(x_0, f(x_0))$ to the graph of f. In higher dimensions, the derivative $f'(x_0)$ is replaced by the Jacobian matrix; in the case of a map $\phi\colon \mathbb{C} \to \mathbb{C}$ in one complex variable (for example, the transition map between two charts), we have the complex derivative $\phi(z_0)$ (which may be thought of as a 2×2 real matrix in a standard way). Because ϕ is analytic, we may use the power series expansion around z_0 on some small neighbourhood:

$$\phi(z) = \phi(z_0) + \phi'(z_0)(z - z_0) + \text{(higher order terms)}.$$

Given two curves through z_0 meeting at an angle θ, we want to confirm that their images under ϕ also meet at the angle θ. The constant term $\phi(z_0)$ merely gives the point of intersection, and does not affect the angle. The higher order terms also have no effect on the angle, since

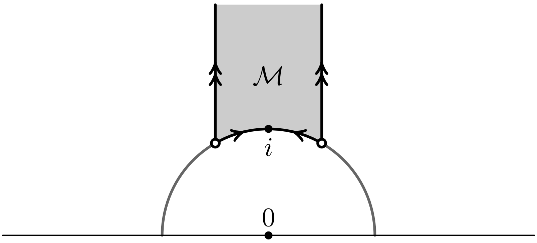


Figure 3.10. Fundamental domain for the modular surface.

they do not affect the tangent lines to the images of the curves at $\phi(z_0)$.

Hence we need only examine the effect of multiplication by the complex number $\phi'(z_0)$. The geometric effect of multiplication by a complex number is homothety (expansion or contraction by the modulus of the number) followed by rotation (by the number's argument); both of these preserve the angle between two lines, and hence holomorphic maps preserve angles.

d. Complex tori and the modular surface. We will now build upon the classification of the complex tori given by Proposition 3.7.

Exercise 3.8. Let L_1 and L_2 be two lattices in $\mathbb{C}$. Show that the tori $\mathbb{C}/L_1$ and $\mathbb{C}/L_2$ are holomorphically equivalent if and only if the angle from a shortest vector to a shortest non-collinear vector in each of the lattices is the same and the ratios of the lengths of those vectors are equal.

This exercise shows that there is a two-parameter family of different complex tori; one parameter is the angle between the generating vectors, and the other is the ratio of their lengths. The next exercise examines a standardised way of choosing the generators for the lattice.

Exercise 3.9. Show that one can choose the shortest vector u in a lattice to be 1, and the second shortest v to be in the region

$$\mathcal{M} = \{ z \in \mathbb{C} \mid |z| \geq 1, |\operatorname{Re} z| \leq 1/2 \} \tag{3.4}$$

shown in Figure 3.10. In addition, show that this requirement determines v uniquely if it lies in the interior of $\mathcal{M}$, and that if it lies on the boundary, then it is determined up to the identifications $-\frac{1}{2} + it \sim \frac{1}{2} + it$ on the vertical boundary and $z \sim -1/z$ on the circular part.

Notice that since $z \mapsto z + 1$ and $z \mapsto -1/z$ are holomorphic functions, the domain $\mathcal{M}$ with the given identifications possesses a natural complex structure, with the exception of the two 'conic' points i and $\frac{\pm 1 + \sqrt{3}i}{2}$, where the total angle after making identifications collapses to π and $2\pi/3$, respectively. This can be relieved by introducing the coordinates $w = (z - i)^2$ and $w = (z - \frac{1+\sqrt{3}i}{2})^3$ near those points. However, it turns out to be more useful to keep the conic points and consider $\mathcal{M}$ as a complex surface with two conical singularities somewhat similar to the standard cone (1.3). It is called the *modular surface*, and plays an extraordinarily important role in number theory and the theory of group representations.

We have thus encountered a very interesting phenomenon: the collection of classes of Riemann surfaces on the torus (up to holomorphic equivalence) is itself naturally endowed with the structure of a Riemann surface! The presence of a complex structure on this collection of equivalence classes, called *Teichmüller space*, is a simple, albeit highly non-trivial, manifestation of a general phenomenon seen throughout different areas of mathematics, wherein the set of invariants of a structure of a certain kind itself possesses a similar structure.

Lecture 20

a. Differentiable functions on real surfaces. In various aspects of the study of surfaces, an important role is played by the class of 'nice' functions on a given surface. For complex (Riemann) surfaces, the natural class to consider is the set of compex-valued holomorphic functions, while for real smooth surfaces, one considers differentiable real-valued functions. There is an important difference here; in the complex case, we deal with functions of one (complex) variable, and so the dimensions of the domain and the range are same, while in the real case, we consider functions of two (real) variables. In the complex case, then, the level set of a given value z, that is, the set of points on the surface at which f takes the value z, is generally a discrete set of points, while in the real case, the level set is usually a smooth curve. In particular, this allows the possibility of 'building up' a real smooth surface by considering the level sets of a sufficiently nice function; this procedure, which we will do later on in this lecture, is one of the basic constructions of Morse theory.

Definition 3.13. Given a function $f: S \to \mathbb{R}$ on a smooth surface, we say that f is *differentiable* if its coordinate representation $f \circ \phi^{-1}: \mathbb{R}^2 \to \mathbb{R}$ is differentiable for every chart $\phi: U \to \mathbb{R}^2$.

We first note that if f is differentiable in one coordinate chart on a neighbourhood, then it is differentiable in any other chart on that same neighbourhood. Indeed, if we have two charts $\phi: U \to D^2$ and $\psi: V \to D^2$, the coordinate representation of f using ϕ is given by

$$f_U = f \circ \phi^{-1}: D^2 \to \mathbb{R}$$

and the representation using ψ is

$$\begin{aligned} f_V &= f \circ \psi^{-1} \\ &= (f \circ \phi^{-1}) \circ (\phi \circ \psi^{-1}) \\ &= f_U \circ (\phi \circ \psi^{-1}). \end{aligned}$$

The transition map $\phi \circ \psi^{-1}$ is smooth and has smooth inverse, so f_V is differentiable on $\psi(U \cap V)$ iff f_U is differentiable on $\phi(U \cap V)$.

Definition 3.14. Given a chart $\phi: U \to D^2$ and a function $f: S \to \mathbb{R}$, the point $p \in U$ is a *critical point* for f if the gradient $\nabla(f \circ \phi^{-1})$ vanishes at p. If the gradient is non-zero at p, we say that p is a *regular point.*

Differentiating the above formula relating f_V and f_U, we have

$$\nabla f_V = D(\phi \circ \psi^{-1}) \nabla f_U$$

where $D(\phi \circ \psi^{-1})$ is the Jacobian of the transition map. By the axioms of a smooth manifold, this has non-zero determinant and hence is invertible, so $\nabla f_V = 0$ if and only if $\nabla f_U = 0$. We have proved the following lemma.

Lemma 3.15. *The critical points of a differentiable function are independent of the particular choice of coordinate chart.*

We now show that away from its critical points, any function can be made to assume a standard form by choosing an appropriate coordinate chart.

Lemma 3.16. *Given a differentiable function $f\colon S \to \mathbb{R}$ and a regular point $p \in S$, there exists a chart $\phi\colon U \to D^2$ around p in which $f_U(x, y) = f(\phi^{-1}(x, y)) = f(p) + x$.*

Proof. Take any coordinates (u, v) around p; because p is not a critical point, we may assume without loss of generality that $\frac{\partial f}{\partial u} \neq 0$. (Here we are abusing notation by using f to stand for both the function $S \to \mathbb{R}$ and its coordinate representation $D^2 \to \mathbb{R}$.)

Then by the Implicit Function Theorem, we may write v as a function of f and u, and hence we can use these as our coordinates. □

The next exercise establishes a similar result in the complex case.

Exercise 3.10. Given a holomorphic function f on a Riemann surface and a point p such that $f'(p) \neq 0$ for some choice of local coordinate, show that one can find a holomorphic chart ϕ around p such that $f(w) = f(p) + \phi(w)$.

So much for the regular points. But what happens at the critical points? We cannot hope for a single standard sort of chart around critical points in the same manner as we just obtained for regular points, because critical points of f have various properties which must remain invariant under changes of coordinates. For example, some critical points are isolated, while others are not. For the time being, we consider only isolated critical points; that is, points $p \in S$ such that for some neighbourhood U, p is the only critical point contained in U.

Even so, there are various possibilities. We typically use critical points as a tool to optimise the value of f; we may find that a particular critical point is a local maximum, a local minimum, or neither, and this classification is independent of our choice of coordinates. In the one-dimensional case, we classified critical points by looking at the second derivative; in two dimensions, the object of interest is the *Hessian matrix*

$$D^2 f(p) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2}(p) & \frac{\partial^2 f}{\partial x \partial y}(p) \\ \frac{\partial^2 f}{\partial y \partial x}(p) & \frac{\partial^2 f}{\partial y^2}(p) \end{pmatrix}.$$

Note that the form of this matrix will only be meaningful if p is a critical point, since otherwise the Hessian vanishes in the coordinate system specified by the above lemma.

Recall from linear algebra that given a symmetric 2×2 matrix

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

such as the one above, we can either use A to define a linear transformation $\mathbb{R}^2 \to \mathbb{R}^2$ by

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto A \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ bx + cy \end{pmatrix} \tag{3.5}$$

or to define a quadratic form $\mathbb{R}^2 \to \mathbb{R}$ by

$$\begin{aligned} \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x \\ y \end{pmatrix}^T A \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\ &= ax^2 + 2bxy + cy^2. \end{aligned}$$

For the Hessian, it is the latter meaning which is relevant here, rather than the more familiar use as a linear transformation. For a linear transformation, the matrix A transforms under a change of coordinates to the matrix $C^{-1}AC$, where C is the matrix specifying the new coordinates; for a quadratic form, A becomes instead $C^T AC$.

It is a basic property of the determinant that $\det C^T = \det C$, and so $\det(C^T AC) = \det(C)^2 \det A$. Thus the sign of the determinant is preserved by changes of coordinates. Assuming the matrix $D^2 f(p)$ is non-degenerate, we have three possibilities:

(1) $\det D^2 f(p) > 0$ and $D^2 f(p)$ is positive definite. Then p is a local minimum for f.

(2) $\det D^2 f(p) > 0$ and $D^2 f(p)$ is negative definite. Then p is a local maximum for f.

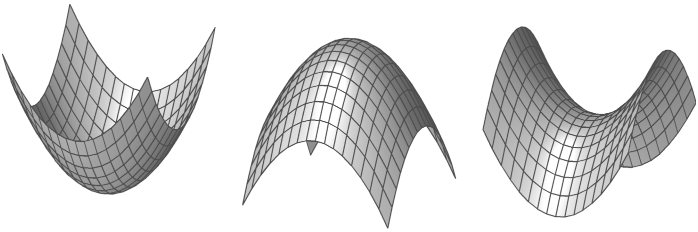


Figure 3.11. Three non-degenerate critical points.

(3) $\det D^2 f(p) < 0$. Then p is a saddle; neither a minimum nor a maximum.

We can now make a linear change of coordinates which brings the quadratic part of the function to a particularly simple form, so that the graph is as shown in Figure 3.11. In all cases the remainder term will be $o(x^2 + y^2)$.

(1) In the first case, there exists a local coordinate system in which $f(x, y) = f(0, 0) + x^2 + y^2 + o(x^2 + y^2)$.

(2) In the second case, there exists a local coordinate system in which $f(x, y) = f(0, 0) - (x^2 + y^2) + o(x^2 + y^2)$.

(3) In the third case, there exists a local coordinate system in which $f(x, y) = f(0, 0) + x^2 - y^2 + o(x^2 + y^2)$.

Exercise 3.11. Prove that any critical point p with $\det D^2 f(p) \neq 0$ is isolated from other critical points.

In fact, the consideration of the behavior of a function near a non-degenerate critical point is made more convenient by a useful technical result called the *Morse lemma*, which states that under an appropriate choice of local coordinates, the error term in the above representation can be eliminated. We present the proof in the most interesting case, that of a saddle, as a series of exercises.

Exercise 3.12. Let p be a non-degenerate saddle point of the function f. Show that locally, the level set $\{ (x, y) \mid f(x, y) = f(p) \}$ is a union of two smooth curves which are tangent at the origin to the diagonals $y = x$ and $y = -x$.

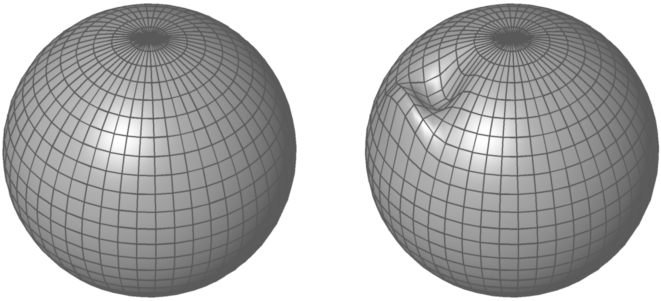


Figure 3.12. Two spheres with different height functions.

Exercise 3.13. Under the same assumption, show that there exist local coordinates (x', y') such that locally, the set $\{ (x, y) \mid f(x, y) = f(p) \}$ is a union of the diagonals $y' = x'$ and $y' = -x'$ themselves.

Exercise 3.14. Show that there exists a smooth map in a neighbourhood of p which is the identity on the diagonals $y' = x'$ and $y' = -x'$, and which maps the curves $f = c$ to hyperbolas $x'y' = c$ for every constant c.

For the other two cases, we will need only a weaker statement which parallels that of Exercise 3.12.

Exercise 3.15. Let p be a non-degenerate minimum of the function f. Show that there exists $\varepsilon > 0$ such that for any c with $f(p) < c < f(p) + \varepsilon$, the level set $f(x, y) = c$ is locally a smooth curve which intersects every ray in the (x, y) coordinates at a single point, and which is transversal (not tangential) to those rays.

b. Morse functions. Given a compact surface S and a smooth function $f \colon S \to \mathbb{R}$, basic topological arguments imply that f achieves its maximum and minimum on S; since the gradient of f in any coordinate representation vanishes at each of these, f must have at least two critical points.

We can easily construct an example where f has no other critical points aside from these two; consider the sphere $S^2 = \{ (x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1 \}$ and the height function $f \colon (x, y, z) \mapsto z$. Then

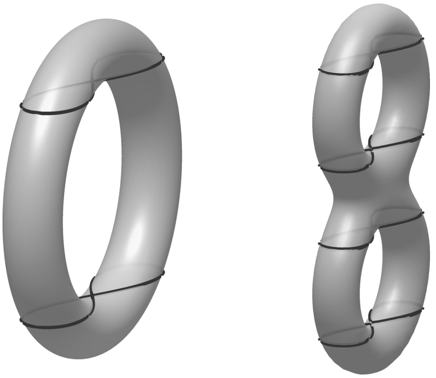


Figure 3.13. Defining a Morse function on a sphere with one or two handles.

f has a maximum at the north pole $(0, 0, 1)$, a minimum at the south pole $(0, 0, -1)$, and no other critical points.

If we perturb the sphere slightly, as shown in Figure 3.12, we will introduce a new pair of local extrema; one local maximum and one local minimum. Along with these we will create two saddle points, so that all in all the perturbed sphere has six critical points; two maxima, two saddles, and two minima.

Another interesting example is given by the standard torus of revolution standing sideways as shown in Figure 3.13, again with the height function $f\colon (x, y, z) \mapsto z$. Now f has one maximum and one minimum, along with two saddles at $(0, 0, \pm 1)$. A similar procedure yields a smooth function on the sphere with m handles having one maximum, one minimum, and $2m$ saddles; the case $m = 2$ is shown, with critical levels drawn for the four saddle points.

Definition 3.17. Let S be a smooth surface and $f\colon S \to \mathbb{R}$ a smooth function. f is called a *Morse function* if every critical point p of f is *non-degenerate*; i.e. the Hessian matrix $D^2 f(p)$ is invertible.

It follows from the definition that every critical point of a Morse function is either a maximum, or a minimum, or a saddle.

Exercise 3.16. Represent the second surface shown in Figure 3.13 (or one homeomorphic to it) as a regular level set of a smooth function, and prove that the height function is indeed a Morse function with one minimum, one maximum, and four saddles.

We will find that looking at the level sets of a Morse function $f\colon S \to \mathbb{R}$ and how they change from one level to another reveals a great deal of information about the surface S. In fact, we can describe a procedure to reconstruct S (up to diffeomorphism) from knowledge of just the critical points of f.

First suppose that for a particular $c \in \mathbb{R}$ the level set $f^{-1}(c) \subset S$ has no critical points (that is, c is a regular value). Then by the same argument used to establish that the level set $F^{-1}(c)$ is a surface (2-dimensional manifold) whenever c is a regular value of $F\colon \mathbb{R}^3 \to \mathbb{R}$, we can deduce from the Implicit Function Theorem and the Inverse Function Theorem that $f^{-1}(c)$ is a 1-dimensional submanifold of S. Since every compact 1-dimensional manifold is a disjoint union of circles, it follows that $f^{-1}(c)$ has this form.

Now what happens if c is a critical value? Let $p \in f^{-1}(c)$ be a critical point; then by the Morse lemma we may choose local coordinates around p such that f takes a standard form.[4] There are three possibilities:

(1) p is a local minimum, $f = c + x^2 + y^2$. Then for c' slightly smaller than c, the level set $f^{-1}(c')$ does not contain any points near p. For $c' = c$, it contains just one point, p, and for c' slightly greater than c, $x^2 + y^2 = c' - c$ defines a circle. Thus as we increase the value of c' through c, a circle is born around the critical point p.

(2) p is a local maximum, $f = c - (x^2 + y^2)$. The reverse of the above process occurs; the circle which exists for $c' < c$ shrinks to a point at $c' = c$ and then vanishes for $c' > c$. As we increase the value of c' through c, a circle dies around p.

[4] The use of the Morse lemma in our considerations is convenient, but not essential. In the minimum and maximum cases, we only need Exercise 3.15, while Exercise 3.12 suffices for the case of a saddle.

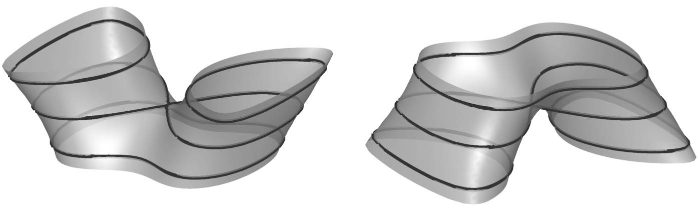


Figure 3.14. Level sets $f^{-1}(c')$ passing through a saddle point.

(3) p is a saddle, $f = c + x^2 - y^2$. For $c' < c$, the (local) level set is a hyperbola opening left and right; for $c' = c$ it is two lines intersecting at p, and for $c' > c$ it is a hyperbola opening up and down. At the global level, we know that between critical points, the level sets are unions of circles, so there are two possibilities, as illustrated in Figure 3.14; as we pass through c, two circles may join and become one, or one circle may split and become two.

With these in mind, we may reconstruct S by increasing c through the range of f; this is the central idea of *Morse theory*, which has very powerful applications in a more general setting than we will consider here. Although the process is much more complicated in higher dimensions, the techniques developed from this theory are involved in the proof of the generalisation of the famous Poincaré conjecture for manifolds of dimension ≥ 5, one of the landmark achievements of mathematics in the third quarter of the twentieth century.[5] The very rough outline of the method is to start from a Morse function on a given manifold which satisfies the assumptions of the Poincaré conjecture—i.e. has certain invariants identical to those of a sphere—and modify it to decrease the number of critical points until only one maximum and one minimum remain.

c. The third incarnation of Euler characteristic. At a more down-to-earth level, we will now show how to use Morse functions to describe a third incarnation of the Euler characteristic χ for surfaces.

[5] This brought a Fields Medal to Stephen Smale in 1966; the solution of the conjecture in the two remaining dimensions—first in dimension four, and then in the original three—resulted in two more Fields Medals later.

If we count the various sorts of critical points on the surfaces we have examined so far (using the height function as our Morse function each time), we have the following:

Surface	maxima	saddles	minima	χ
sphere	1	0	1	2
(perturbed) sphere	2	2	2	2
torus	1	2	1	0
sphere with m handles	1	$2m$	1	$2-2m$

Note that in each case, the Euler characteristic χ is equal to the alternating sum of the three columns; in fact, this is true in general.

Theorem 3.18. *For any Morse function $f\colon S \to \mathbb{R}$, the Euler characteristic is related to the number of critical points by the formula*

$$\chi = (\#\text{ of maxima}) - (\#\text{ of saddles}) + (\#\text{ of minima}). \tag{3.6}$$

Before proving the theorem, we describe the general method and examine what happens in the case of the torus. We proceed by examining the *sublevel sets*

$$S_c = f^{-1}((-\infty, c]) = \{\, x \in S \mid f(x) \leq c \,\}.$$

Let m and M be the minimum and maximum values, respectively, assumed by f on S. Then for $c < m$, we have $S_c = \emptyset$, and for $c \geq M$, $S_c = S$. The real story is what happens in between m and M...

The next observation to make is that nothing interesting happens at non-critical levels. This is the content of the following lemma, which intuitively looks quite plausible, although a rigorous proof requires certain tools which we will not develop until later (see Lecture 36(b)).

Lemma 3.19. *Given a Morse function $f\colon S \to \mathbb{R}$ and $a, b \in \mathbb{R}$ such that every $c \in (a, b)$ is a regular value ($f^{-1}(c)$ contains no critical points), then S_c and $S_{c'}$ are diffeomorphic for every $c, c' \in (a, b)$.*

Thus for the torus shown in Figures 3.13 and 3.15, with inner radius 1 and outer radius 2, all the action happens at $f(x) = \pm 1, \pm 3$. In between those points, the boundary of S_c is the level set $f^{-1}(c)$, which we know to be a disjoint union of circles. The four critical

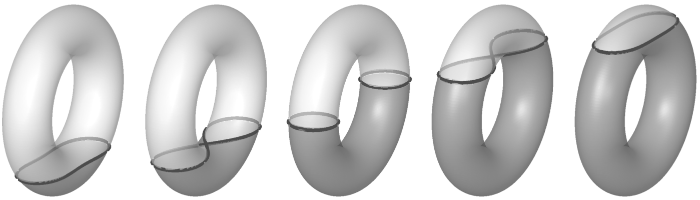


Figure 3.15. Sublevel sets on the vertical torus.

points run through the four possibilities enumerated in our earlier discussion:

(1) At $c = -3$, a circle is born, so the empty set is replaced by a disc.

(2) At $c = -1$, one circle splits into two, so the disc is replaced by a cylinder.

(3) At $c = 1$, the two circles rejoin and become one, so the cylinder is replaced by a torus with a hole.

(4) At $c = 3$, the circle dies, so the hole is filled with a cap, and we obtain the entire torus.

Proof of Theorem 3.18. It follows from Lemma 3.19 that between critical levels, the changes in S_c are only quantitative, not qualitative, and have no effect on the Euler characteristic; in order to prove the theorem, therefore, it suffices to examine the change in χ as we pass through each of the various sorts of critical points. To accomplish this, we first extend the definition of χ to allow non-connected manifolds; this will allow examples with $\chi > 2$, which is impossible in the connected case.

Now there are three cases to examine. If $f^{-1}(c)$ contains a local minimum of f, then passing through c corresponds to adding a new disc, as we saw, and hence increases χ by one. Similarly, passing through a local maximum corresponds to filling in a hole with a disc, which involves adding a face and leaving the number of edges and faces unchanged, and so also increases χ by one.

It remains only to show that passing through a saddle point decreases χ by one. Figure 3.16 shows the sublevel sets $S_{c'}$ (viewed from

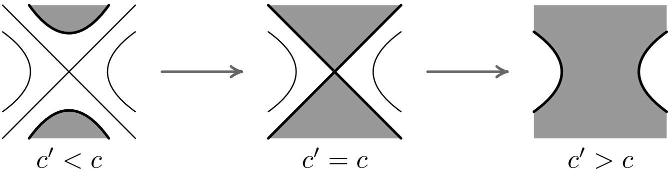


Figure 3.16. Sublevel sets near a saddle.

above) for values of c' near the critical value c. Upon passing through the saddle, the number of edges and vertices remains the same, but two faces which previously were separate are joined into one. Hence the alternating sum $\chi = V - E + F$ is decreased by one. □

If we carry out this construction a bit more carefully, we can actually obtain a complete classification of smooth surfaces using Morse functions as our tool; this was in fact the inspiration for the proof we gave of the classification theorem for compact orientable surfaces (Theorem 2.15), and is a 'baby version' of the arguments used in higher dimension, like those on which the afore-mentioned proof of the Poincaré conjecture in dimensions five and above is based.

Exercise 3.17. Consider the function $f(x,y) = \sin(4\pi x)\cos(6\pi y)$ on the standard flat torus $\mathbb{R}^2/\mathbb{Z}^2$.

(1) Prove that it is a Morse function, and calculate the number of minima, saddles, and maxima.
(2) Describe the evolution of the sublevel sets $f^{-1}((-\infty, c))$ as c varies from the lowest minimum value to the highest maximum value.

Lecture 21

a. Functions with degenerate critical points. Having successfully used the ideas of Morse theory to reconstruct the surface S and run across our old friend, the Euler characteristic, we would now like to extend the same ideas and techniques to the case where our function $f\colon S \to \mathbb{R}$ may fail to be Morse by having degenerate critical points.

We begin by noting that in the non-degenerate Morse case, we obtained the Euler characteristic by giving each critical point a 'weight' of either +1 (for a maximum or a minimum) or −1 (for a saddle), and then summing over all critical points. In order to extend our calculations to include degenerate critical points (for which the Hessian matrix $D^2 f$ has zero determinant), we must similarly define the *Morse index* for these points. The goal will be to define for each critical point p, degenerate or not, the Morse index $\operatorname{ind}_f(p)$ in such a way that the following formula holds:

$$\chi = \sum_{\nabla f(p)=0} \operatorname{ind}_f(p). \tag{3.7}$$

It is instructive to begin by considering degenerate critical points in one dimension. Given a smooth function $f\colon \mathbb{R} \to \mathbb{R}$, non-degenerate critical points of f will be either minima or maxima, near which f will behave like $x \mapsto \pm x^2$. An example of a degenerate critical point is given by $f\colon x \mapsto x^3$, which has $f'(0) = f''(0) = 0$. 0 is a critical point since f' vanishes, and it is degenerate since the Hessian, which in this case is just the 1×1 matrix $[f'']$, has zero determinant.

What happens to the critical point at 0 if we perturb the function f slightly? For concreteness, let ε be small (either positive or negative) and let $f_\varepsilon(x) = x^3 + \varepsilon x$, so that f_0 is our original function f. Then $f_\varepsilon' = 3x^2 + \varepsilon$; for $\varepsilon > 0$, we have $f_\varepsilon'(x) > 0$ everywhere, and hence f has no critical points. For $\varepsilon < 0$, f_ε has two critical points at $\pm\sqrt{-\varepsilon/3}$; one of these is a local maximum and the other is a minimum.

We note that the above analysis goes through no matter how small the perturbation is; the degenerate critical point either vanishes or splits into two non-degenerate critical points. This is in sharp contrast to the case where the critical point is already non-degenerate; because the condition $\det(D^2 f) \neq 0$ is an *open* condition, sufficiently small perturbations will not effect any qualitative changes. Near a degenerate critical point, though, perturbing f will result in some sort of *bifurcation*, as we have just seen.

We now examine a two-dimensional example of this sort of behaviour. The function $f(x, y) = xy(x + y)$ has the level set $f^{-1}(0)$

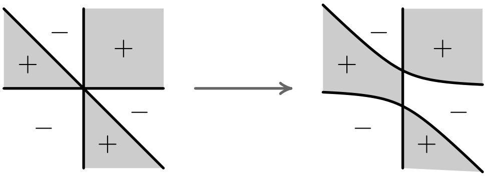


Figure 3.17. Perturbing f near a degenerate critical point.

shown in the first graph in Figure 3.17, and takes the signs indicated. Differentiating, we have

$$\begin{aligned}
Df &= (2xy + y^2, x^2 + 2xy), \\
D^2 f &= \begin{pmatrix} 2y & 2x+2y \\ 2x+2y & 2x \end{pmatrix}, \\
\det(D^2 f) &= 4(xy - (x+y)^2) \\
&= -4(x^2 + xy + y^2).
\end{aligned}$$

Thus the critical point $(0,0)$ is degenerate;[6] we can perturb f by adding εx, and obtain

$$\begin{aligned}
f(x,y) &= xy(x+y) + \varepsilon x, \\
Df &= (2xy + y^2 + \varepsilon, x^2 + 2xy).
\end{aligned}$$

Because the perturbation was linear, $D^2 f_\varepsilon = D^2 f$. To find the critical points, we observe that $Df = 0$ implies $x = 0$ or $x = -2y$; in the former case, we have $y^2 + \varepsilon = 0$, and in the latter, we have $-3y^2 + \varepsilon = 0$. Hence the fixed points are given by

parameter	fixed point(s)
$\varepsilon < 0$	$(0, \pm\sqrt{-\varepsilon})$
$\varepsilon = 0$	$(0,0)$
$\varepsilon > 0$	$(\mp 2\sqrt{\frac{\varepsilon}{3}}, \pm\sqrt{\frac{\varepsilon}{3}})$

The second graph in Figure 3.17 shows the situation for $\varepsilon < 0$, where the single degenerate critical point has bifurcated into two non-degenerate critical points, both saddles since $\det(D^2 f_\varepsilon) = 4\varepsilon < 0$. A precise visualisation of the case $\varepsilon > 0$ is left for the reader.

[6] A degenerate critical point of this nature is sometimes known as a 'monkey saddle' because of the extra depression in the graph, which could accomodate a simian tail.

Now if we think of our function f as a height function on S as we did for the sphere and the vertical torus, then small perturbations of f in the neighbourhood of a critical point correspond to 'warping' S slightly, which ought to have no effect on the Euler characteristic. Then the above example leads us to expect that the complicated saddle exhibited by $f(x, y) = xy(x + y)$ ought to be counted as two regular saddles, and so the Morse index of this particular degenerate critical point ought to be -2. In fact, an argument analogous to the one given last time shows that passing through a saddle of this form corresponds to joining *three* faces into a single face, while leaving the number of edges and vertices constant, and hence decreases χ by 2.

How are we to make this general, though? Our approach so far has been relatively *ad hoc*; we will now develop in a more systematic manner a theory which will allow us to assign an index to each isolated critical point and hence find the Euler characteristic of a surface in terms of *any* smooth function with isolated critical points, whether degenerate or not.

The first step will be to define the degree of a map from the circle to itself. We will then consider vector fields on a surface, in particular the points where they vanish, and use this notion of degree to define the index of the vector field at such a point. Finally, we will observe that associated to every smooth function $f\colon S \to \mathbb{R}$ is a natural vector field given by the gradient of f, and hence define the index of a critical point p as the index of the gradient vector field around p.

As we do all this, it ought to be remembered that while the details of the construction depend upon a particular choice of coordinates on the surface, the final result, the value of the index, will be independent of our choice of chart.

Exercise 3.18. Consider the function $f(x, y) = xy(x + y)(x - y)$ in a neighbourhood of the origin. Construct a perturbation of this function which has only non-degenerate critical points, and use this perturbation to calculate the change in Euler characteristics of the sublevel sets effected by the passage through such a critical point.

Exercise 3.19. Consider the function $f(x, y) = xy(x + y) + r(x, y)$ where r is a function which vanishes at the origin, together with all its partial derivatives of orders one, two, and three.

(1) Prove that in a neighbourhood of the origin, the level set $f(x, y) = 0$ is a union of three smooth curves tangent at the origin to the x-axis, the y-axis, and the diagonal $y = -x$.

(2) Prove that the Euler characteristic of the sublevel set $S_c = f^{-1}((-\infty, c])$ for this function decreases by two as c passes through the zero level from negative to positive, and argue that the same result holds on a compact surface with a function which has the local form given above, and has no other critical points with the same value.

b. Degree of a circle map. Given a map $f\colon S^1 \to S^1$ from the circle to itself, we can think of the circle as being wrapped around itself a number of times by f; we will call this number the *degree* of the map. This is made precise as follows.

Recall that S^1 can be given as the quotient space $\mathbb{R}/\mathbb{Z}$, or the unit interval $[0, 1]$ with ends identified. Then we can think of f as a function not on the circle, but on the real line. That is, we can define a function $F\colon \mathbb{R} \to \mathbb{R}$ (called the *lift* of f) such that

$$f(x + \mathbb{Z}) = F(x) + \mathbb{Z}.$$

(Recall that points in the quotient space $\mathbb{R}/\mathbb{Z}$ are equivalence classes $x + \mathbb{Z} = \{\dots, x - 1, x, x + 1, \dots\}$.) First choose any $F(0) \in f(0 + \mathbb{Z})$; once $F(0)$ is fixed, the requirement that F be continuous determines $F(x)$ for every $x \in \mathbb{R}$.

Passing once around the circle brings us back to where we began; this corresponds to increasing x by 1, and when we return to the starting point, we must have the same value of f, hence $F(1) \in F(0) + \mathbb{Z}$, so $F(1) - F(0) \in \mathbb{Z}$. Notice that any given continuous circle map f has infinitely many different lifts, and any two lifts differ by an integer constant.

Definition 3.20. Given $f\colon S^1 \to S^1$ and $F\colon \mathbb{R} \to \mathbb{R}$ defined as above, the integer $F(1) - F(0)$ is the *degree* of the circle map f, and is denoted by $\deg f$.

The first half of Figure 3.18 shows a circle map f (actually, a graph of the lift F) with degree 2.

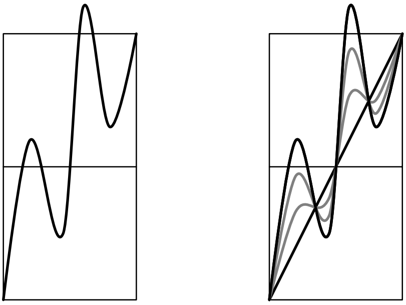


Figure 3.18. A circle map is homotopic to a linear map.

For our purposes, the most important property of the degree is that it is continuous in the uniform $\mathcal{C}^0$ topology; in other words, changing the image of the function f by an amount $< \varepsilon$ at each point of S^1 will only change the degree of f by an amount $< \varepsilon$. Since the degree takes integer values, this implies that it must in fact remain the same; we say that it is *locally constant.*

In particular, if $\{f_t\}_{t\in[0,1]}$ is a continuous family of maps, the degree of f_t is constant with respect to t. If two functions $f_0, f_1 \colon S^1 \to S^1$ can be connected by such a continuous family, we say that they are *homotopic*; what we have just shown is that degree is a *homotopy invariant.*

Further, as shown in the second half of Figure 3.18, we can construct a linear homotopy from any circle map with degree n to the linear map $E_n \colon x \mapsto nx$ via the family of functions

$$F_t(x) = (1-t)F(x) + tnx$$

with $F_0 = F$ and $F_1 \colon x \mapsto kx$. This shows that two circle maps with the same degree are homotopic to each other, and so degree completely classifies circle maps up to homotopy, with representatives of each homotopy class being given by the standard linear maps E_n for $n \in \mathbb{Z}$. The map E_n wraps the circle around itself n times, and in additive notation ($S^1 = \mathbb{R}/\mathbb{Z}$) is written as

$$E_n(x) = nx \mod 1.$$

In multiplicative notation ($S^1 = \{ z \in \mathbb{C} \mid |z| = 1 \}$), this becomes

$$E_n(z) = z^n.$$

We may think of a circle map f as recording the progress of a runner around a track. At time $t = 0$, he is at the start line, and then proceeds around the track with direction and speed determined by f'.[7] At time $t = 1$, he crosses the finish line (which is in the same place on the track as the start line)—the degree of the map f is just the number of laps he has completed in between.

In these terms, our current method of measuring degree corresponds to the point of view of the runner; he keeps track of the distance he has run, counting counterclockwise as positive and clockwise as negative, and after completing the race tells us how far he has gone. An equally valid point of view is that of a spectator sitting in the stands somewhere along the track, counting the number of times the runner goes by. If the runner passes the spectator going counterclockwise, the count increases by one; if the runner passes in a clockwise direction, the count decreases by one. Then at the end of the race, the spectator will also have an accurate count of the number of laps the runner has completed.

In terms of the function f, this point of view amounts to choosing a point $y \in S^1$, looking at the set of preimages $f^{-1}(y)$, assigning each the value ± 1 based on the sign of f' at that point, and then summing over these values; the sum will be the degree of f.

Exercise 3.20. Prove that the degree of the composition $f \circ g$ of two circle maps f and g is equal to the product of the degrees of f and g.

Exercise 3.21. Prove that a continuous circle map of degree d has at least $|d - 1|$ fixed points.

It is not immediately obvious what the higher-dimensional generalisation of all this ought to be. The two n-dimensional analogues of the circle S^1 are the n-torus, which is the direct product of n circles, and the n-sphere S^n. We might attempt to generalise the definition of degree to maps of either of these manifolds.

[7] Note, though, that the definition of degree goes through even if f is only continuous, and not differentiable.

There are some mathematical contexts, such as Fourier analysis, in which the torus is the natural generalisation of the circle, and the definition of degree extends quite naturally in this direction. For a map $f\colon \mathbb{T}^n \to \mathbb{T}^n$, we could essentially repeat the above discussion, noting that the n-torus is the quotient space $\mathbb{R}^n/\mathbb{Z}^n$, lifting the map f to $F\colon \mathbb{R}^n \to \mathbb{R}^n$, and obtaining a vector in $\mathbb{Z}^n$ as the degree.

It turns out, though, that for our purposes here, the n-sphere is the relevant manifold.[8] Because $\mathbb{R}^n$ is not a covering space of S^n, the definition above does not generalise in the naïve way, and it is not at first apparent how we ought to count the number of times that the sphere wraps around itself under the action of f.

Despite this, we can in fact generalise the concept of degree to higher dimensions, and this is a fundamental definition in algebraic and differential topology. However, it requires us to define and work with the homology groups of the sphere, and all in all would get us into deeper waters than we are prepared for at the moment.

A somewhat more manageable approach works for smooth maps. In this case we can adopt the second point of view, which in the one-dimensional case was that of the spectator watching the runner go by, and involved using the notion of 'positive' and 'negative' regular preimages of a given point. This works in higher dimensions as well, and we may once again define the degree as the number of positive preimages minus the number of negative ones, just as for the circle.

One must, however, justify this procedure by showing that for any smooth map $f\colon S^n \to S^n$, there are indeed points in S^n whose preimages are all regular (the matrix of partial derivatives in local coordinates has non-zero determinant), and then show that the degree so defined is the same for all regular values. To appreciate the subtlety of this procedure, the reader is encouraged to work out the details for the one-dimensional case.

[8] This is because we will eventually use the degree of a circle map to define the index of a vector field, and then use the index of the gradient vector field to define the Morse index of a degenerate critical point. Defining the index of a vector field at a critical point involves obtaining a map on the boundary of a small neighbourhood of that point, which topologically is a sphere rather than a torus.

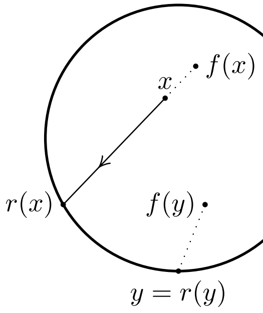


Figure 3.19. Proof of Brouwer's theorem.

c. Brouwer's fixed point theorem. Before using the notion of degree to define the index of a zero for a vector field, we will present a remarkable application of degree theory which is of a purely topological nature.

Theorem 3.21 (Brouwer's fixed point theorem in dimension two)**.** *Let X be any space homeomorphic to the closed disc D^2. Then any continuous map of X into itself has a fixed point.*

Proof. We consider the standard closed disc

$$D^2 = \{ (x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1 \}$$

and argue by contradiction. Suppose $f \colon D^2 \to D^2$ is a continuous map without fixed points. For $p \in D^2$ consider the open half-line (ray) beginning at $f(p)$ and passing through the point p. This half-line intersects the unit circle S^1, which is the boundary of the disc D^2, at a single point which we will denote by $r(p)$. Notice that $r(p) = p$ for $p \in \partial D^2 = S^1$, and that the map $r \colon D^2 \to \partial D^2$ thus defined is continuous (because f has no fixed points).

Now for each $t \in (0, 1]$, consider the circle map $r_t \colon S^1 \to S^1$ given by restricting r to the circle of radius t around the origin. As we have already seen, $r_1 = \mathrm{Id}$ and hence has degree one. As $t \to 0$, the map r_t converges to a constant map, and hence for small enough t, we must have $\deg r_t = 0$. Since the degree of r_t depends continuously on t, this is impossible, and we have our contradiction. □

Brouwer's theorem holds for discs in any dimension. In dimension one, the statement is a trivial corollary of the Intermediate Value Theorem. In higher dimensions, the scheme of the proof remains the same, but it uses the more complicated notion of degree for a sphere map.

Lecture 22

a. Zeroes of a vector field and their indices. We end this chapter by applying the notion of degree to a continuous vector field. A precise definition of the phrase *vector field*, and the accompanying notions of tangent vectors, tangent spaces, etc., will come later in the lecture. For the time being, we consider a particular set of local coordinates which uses as its patch the open set $U \subset \mathbb{R}^2$. Then a vector field assigns a vector to each point of U, and so we denote it by

$$\begin{aligned} X\colon \qquad U &\to \mathbb{R}^2, \\ (x,y) &\mapsto (u,v). \end{aligned}$$

In this way, X specifies a direction and magnitude at each point $(x,y) \in U$, and we say that the vector field is continuous, smooth, etc. if the map X is continuous, smooth, etc.

The idea is to look at the rotation of the vector field around a point where it vanishes. First we note that around a point where the vector field is non-zero, it is nearly constant on a small neighbourhood (in fact, it can be made exactly constant by an appropriate choice of coordinates), and hence points in a particular direction, without any rotation. Around a point where the vector field (u,v) vanishes, however, the situation is different.

Let $p = (x_0, y_0)$ be an isolated zero of X, and consider a small circle going counterclockwise around p, parametrised by $\gamma\colon S^1 \to U$. To each point (x,y) on the circle the vector field assigns a vector (u,v), and by normalising (u,v), we obtain a unit vector, which is just a point on the unit circle. In this way the vector field near p defines a circle map $\phi_\gamma\colon S^1 \to S^1$ by

$$\phi_\gamma\colon t \mapsto \frac{(u(\gamma(t)), v(\gamma(t)))}{\|(u(\gamma(t)), v(\gamma(t)))\|}. \tag{3.8}$$

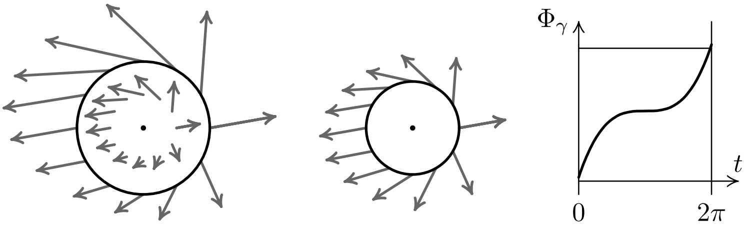


Figure 3.20. A critical point of a vector field with degree one.

Figure 3.20 shows an example of a vector field, its normalisation along the curve, and the lift Φ_γ of the circle map.

Definition 3.22. The *index* of a critical point $p = (x_0, y_0)$ of a vector field is the degree of the circle map ϕ_γ in (3.8). We denote this value by $\operatorname{ind}_p X$.

It is vital to our definition that γ does not go through any critical points of X; that is, the vector field must be non-vanishing along the curve (otherwise we would not be able to normalise). Further, X should not vanish at any other point in the region bounded by γ other than (x_0, y_0), or the value we derive for the index will not be accurate; we will see why this is so in Lecture 36, when we consider indices for curves enclosing more than one point.

Thus we see that we must take some care in our choice of γ; on the other hand, nothing in this definition actually uses the fact that we took the image of γ to be a circle of a particular radius. By continuously deforming γ into a circle of a different radius, or any other simple closed curve around p, we vary the induced circle map ϕ_γ, and hence the index $\operatorname{ind}_p X$, continuously, provided all the curves through which we deform satisfy the two conditions of the previous paragraph. Since the index is an integer, it remains constant, and hence is the same for any such curve. This also shows that the index is invariant under a change of coordinates, since such a change merely takes the circle to some other valid curve.

Is there any condition which characterises a 'valid curve', beyond the above requirements that X be non-vanishing on the curve and its interior, except at p? If we draw various curves which can be reached via permissible deformations, we see that they all ‘go around p exactly once counterclockwise’. It turns out that a formalisation of this notion is the proper condition, and that any curve satisfying it can be permissibly deformed into a small circle around p.

This leads us to the definition of the index of a curve with respect to a point. To this end, let γ be a closed curve in the plane (which may be self-intersecting), and let p be any point not on the curve. Then we may define a circle map ϕ by

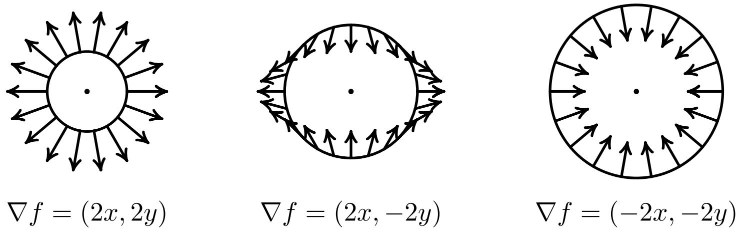
$$\phi\colon t \mapsto \frac{\gamma(t) - p}{\|\gamma(t) - p\|}.$$

Definition 3.23. Given γ, p, ϕ as above, the *index* of γ with respect to p, also called the *winding number*, is the degree of ϕ, and is denoted by $\operatorname{ind}_p \gamma$.

In light of this definition, the rather vague statement that ‘γ goes around p once’ ought to be replaced by the requirement that $\operatorname{ind}_p \gamma = 1$. Note that the index may depend on our choice of coordinates; consider the equator of the sphere, and a point not on the equator. Then in one of the charts of stereographic projection, the index will be 0, while in the other, it will be ±1. For this reason, we should speak about the index of a curve *in the plane*, rather than on a surface, unless we have fixed a coordinate chart.

This notion of index is central to complex analysis, where it comes into play in the statement (and proof) of the residue theorem, which generalises Cauchy’s integral formula. It also plays a somewhat surprising role in the proof of the Fundamental Theorem of Algebra, which we will investigate in Lecture 33, and we will use it in our proof of the Jordan Curve Theorem in Lectures 34 and 35.

For the time being, the salient fact is that we can now apply this theory to the gradient vector field ∇f in order to define the index of any isolated critical point of f. Of course, the theory works for *any* vector field, whether or not it arises as the gradient of a smooth function; this will eventually lead us to discover yet another incarnation of the Euler characteristic in Lecture 36.



$\nabla f = (2x, 2y)$ $\nabla f = (2x, -2y)$ $\nabla f = (-2x, -2y)$

Figure 3.21. Three gradient vector fields around a critical point.

b. Calculation of index. We now turn to the case which served as a motivation for introducing the index for a vector field. Take a nondegenerate critical point of a function f, introduce local coordinates in which the function is quadratic, and consider the gradient vector field $\nabla f = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y})$.

Then ∇f is equal to one of $(2x, 2y)$, $(2x, -2y)$, or $(-2x, -2y)$, as in Figure 3.21, depending on whether the critical point is a minimum, a saddle, or a maximum. As a reference curve γ, consider a small circle parametrised by t, which we take to be the argument (angle) divided by 2π. For the three cases shown, the maps ϕ_γ are

$$\phi_\gamma(t) = t, \quad \phi_\gamma(t) = -t, \quad \text{and} \quad \phi_\gamma(t) = t + 1/2.$$

The degree in the first and last cases is equal to 1, and in the second to -1, which is in complete agreement with the derivation of the index formula (3.7) from the counting formula (3.6) for critical points.

More generally, if the vector field is transversal (not tangent) to a curve γ at every point, as in the first and third cases above, then the index of the critical point is equal to one. This is obvious for a circle (or for any star-shaped curve), since in this case the map ϕ_γ is homotopic to the identity.

Notice that in our three standard cases, the gradient vector fields turned out to be linear. One can consider more general linear vector fields of the form $(ax + by, cx + dy)$. Let $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$—then the origin is an isolated zero if and only if $\det A \neq 0$.

By a linear coordinate change, the matrix A, and hence the vector field as well, can be brought either to diagonal form (if the eigenvalues are real), or to the form λR (if the eigenvalues are complex), where λ

is a positive scalar and R is the matrix of a rotation. Notice that in all cases except that of one positive and one negative eigenvalue, the vector field is transversal to a circle γ around the origin, and hence the index is equal to one. In the remaining case, the diagonal vector field can be deformed to $(x, -y)$, and hence the map ϕ_γ is homotopic to $t \mapsto -t$. Thus the index is equal to -1, and we have completely categorised the linear case.

Notice furthermore that a non-linear vector field near a critical point can be split into the sum of its linear part and an error term, which is of the size $o(|x|+|y|)$. If the linear part is non-degenerate (its matrix is invertible), then on a small enough circle, the vector field is homotopic to the linear part. Thus we have proved the following statement.

Proposition 3.8. *Let p be an isolated zero of a vector field X, and let D_pX be the linear part of X at p. If D_pX has matrix A in a local coordinate system, then the index of X at p is equal to the sign of the determinant of A.*

On the other hand, if the linear part is degenerate, it is of no use in studying the non-linear vector field. The extreme case appears when the linear part vanishes completely, as for the gradients of the functions $xy(x + y)$ and $xy(x + y)(x - y)$, which were discussed in Lecture 21(a). For these examples, one can explicitly construct the maps ϕ_γ for a small circle γ, and prove by brute force that the index is equal to 2 for the first, and 3 for the second, as expected. Later, in Lecture 36, we will be able to see this by perturbing a degenerate critical point into several non-degenerate ones, and using additivity of the index for a curve which winds around several critical points.

Finally, we have to address the fact that our considerations so far have depended on a choice of coordinates. There are two aspects to this. In the first place, because our definition of vector fields was given in terms of local coordinates, we need to see what happens when the coordinate system changes. We will take up this issue momentarily, and our conclusion will be entirely satisfactory—vector fields can be defined independently of local coordinates, and in particular, coordinate changes carry zeroes into zeroes, nice curves surrounding isolated zeroes into other such curves, and preserve the index.

The second issue has to do with calculating the index of an isolated critical point for a function. For that purpose we used the gradient vector field $\nabla f = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y})$, which depends on the choice of coordinates. Fortunately, the transformation of the gradient vector field under a coordinate change is not too drastic, and in particular, it does not affect the index. To see that, notice that in any coordinate system, the gradient points in the direction in which the function increases most quickly. The tangent to the level curve of the function divides the tangent space into two half-planes, and the fact that the function increases in the direction of the gradient specifies in which half-plane the gradient vector lies. Because the level curve and its tangent are naturally defined independently of a coordinate system, this half-plane is too, and so the independence of the index of a gradient vector field from a particular choice of coordinates follows from the following fact.

Proposition 3.9. *Let X and Y be vector fields with the same isolated fixed point p, such that in a small neighbourhood of p, the directions of X and Y are never opposite. In other words, the angle between X and Y is strictly between $-\pi$ and π. Then*

$$\operatorname{ind}_p X = \operatorname{ind}_p Y.$$

Proof. Consider the continuous family of vector fields given by $X_t = (1-t)X + tY$ for $0 \leq t \leq 1$. Our condition implies that p is an isolated fixed point for all X_t, and so for a small circle γ the map ϕ_γ may be constructed for X_t. Then $\operatorname{ind}_p X_t$ depends continuously on t, and is thus constant. □

c. Tangent vectors, tangent spaces, and the tangent bundle. We have been using the terminology of vector fields and tangent vectors rather carelessly up to this point, and so it is time to make these notions more precise.

Given a smooth surface S embedded in $\mathbb{R}^3$, we have a clear geometric definition of the tangent plane to S at a point $p \in S$. We would like to generalise this definition to an arbitrary smooth surface (or indeed, a smooth manifold of any dimension) without reference to a particular embedding in Euclidean space. To do this, we will need to give a definition of tangent vectors and tangent spaces in terms of

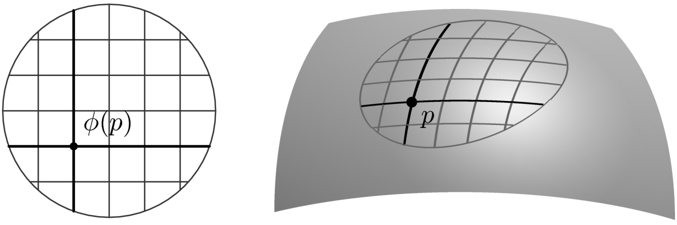


Figure 3.22. Defining the tangent space using coordinates.

the various coordinate patches and charts which make up a smooth atlas. First we will define tangent vectors at a point p, then we will define the tangent space T_pS as the linear space comprising all such vectors; finally, the tangent bundle TS will be the disjoint union of all the tangent spaces.

We begin by considering a single chart $\phi\colon U \to \mathbb{R}^2$; to each point $p \in S$, we want to somehow associate a two-dimensional linear space (since our surface is two-dimensional). We will also require that this space behave well under coordinate changes, in that such changes must preserve its linear structure. There are two ways of accomplishing this, neither of which is entirely satisfactory from a visual point of view. Consequently, the reader is advised to approach the following as being, to some degree at least, a purely formal construction, the geometric meaning of which will become apparent in time.

The first idea is to look at the two coordinate axes in $\mathbb{R}^2$ and to consider their preimages in S, which are smooth curves intersecting at p. We expect a smooth curve to have a tangent vector at each point, so we may write $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ for the tangent vectors to the preimages of the x-axis and the y-axis, respectively.

As the notation suggests, this has an interpretation in terms of directional derivatives; for the time being, we treat these as formal symbols, and call any linear combination of them a tangent vector to S at p. Then the tangent plane at p is

$$T_pS = \left\{ t\frac{\partial}{\partial x} + s\frac{\partial}{\partial y} \;\middle|\; (t, s) \in \mathbb{R} \right\}.$$

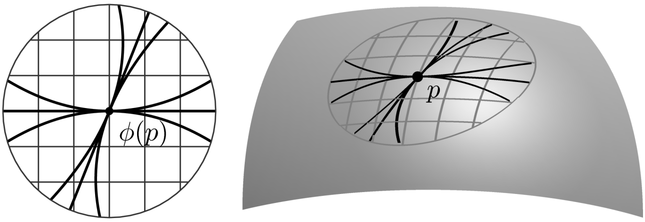


Figure 3.23. A coordinate-free definition of the tangent space.

What happens to this definition under a change of coordinates $F\colon (x,y) \mapsto (u,v)$? We will see in the next lecture that as one might expect, we transform the tangent vectors according to the rule

$$\frac{\partial}{\partial u} = \frac{\partial x}{\partial u}\frac{\partial}{\partial x} + \frac{\partial y}{\partial u}\frac{\partial}{\partial y}$$

which will allow us to write the change of coordinates in the tangent space as a linear map in terms of the Jacobian of F.

The second possible idea to follow in constructing the tangent spaces, which we mention only briefly here, is to consider equivalence classes of curves through p. Given two smooth curves γ and η which pass through p at time 0, we say that γ and η are *tangent* at p if

$$\|\phi(\gamma(t)) - \phi(\eta(t))\| = o(t),$$

that is, if

$$\lim_{t\to 0} \frac{\|\phi(\gamma(t)) - \phi(\eta(t))\|}{t} = 0,$$

where $\|\cdot\|$ is the norm in the coordinate space $\mathbb{R}^2$. Then the tangent space T_pS is given as the set of equivalence classes of smooth curves up to tangency; this definition is in some sense coordinate-free, because the equivalence classes will be the same for any choice of coordinates and any choice of norm.

In the next chapter, we will use the idea of tangent spaces to define a notion of *Riemannian metric* on a surface. From the geometric point of view, this is the most important structure carried by a surface, and so we will devote all of Chapter 4 to Riemannian metrics and geometry.

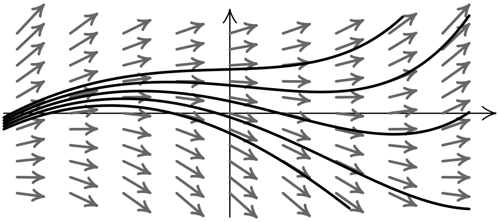


Figure 3.24. Some integral curves for a vector field.

Our business with other aspects of smooth structure is far from over, however. For example, we only briefly mentioned some important results such as the expression of the Euler characteristic as the sum of the indices of zeroes for a vector field (the fourth incarnation of Euler characteristic). Further, we have not developed our understanding of vector fields themselves far enough. There are two main directions in which this theory is developed. The integration of vector fields to produce *flows* (one-parameter groups of diffeomorphisms of surfaces), along with the study of such flows, is the beginning of the qualitative theory of ordinary differential equations (ODEs); we will not devote much time to this topic, but will spend some time in the final chapter pursuing a second direction, examining the structure of vector fields near non-degenerate zeroes. In that discussion, Riemannian metrics will provide a useful auxiliary tool, letting us associate a vector field to certain functions and maps in Lecture 36. They will also prove useful in dealing with topological questions, when we prove the existence of tubular neighbourhoods in Lemma 5.6.

Exercise 3.22. An *integral curve* of a vector field X is a smooth curve such that X is tangent to the curve at any point of the curve, as shown in Figure 3.24.

Construct a smooth vector field on the torus without zeroes and without closed integral curves.

Exercise 3.23. Construct an example of a vector field on the sphere with a single zero, and calculate the index of this zero.

Exercise 3.24. Prove that the index of an isolated zero of a smooth vector field can take any integer value.

Chapter 4

Riemannian Metrics and Geometry of Surfaces

Lecture 23

a. Definition of a Riemannian metric. The definition of the tangent space given in the previous lecture formalises the idea of being able to discuss *directions* on a manifold. In order to formulate and address problems of a geometric nature, we must also have a notion of *distance.* To this end, we will now give the definition of a *Riemannian metric*, one of the core ideas in modern geometry.

Consider a surface S embedded in $\mathbb{R}^3$. We have a natural metric (notion of distance) in $\mathbb{R}^3$ given by Pythagoras' formula, which is to be inherited by S in some fashion; that is, we want to define distances on S in terms of the ambient metric in $\mathbb{R}^3$. Given two points $x, x' \in S$, the most obvious way to do this is to declare their distance to be equal to the Euclidean distance in $\mathbb{R}^3$:

$$d(x,x') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + (x_3 - x_3')^2}.$$

This idea, however natural, is not the correct one. If we think of x and x' as two cities on the surface of the earth, what we are really interested in is not the length of the shortest tunnel *through* the earth from one to the other, which is what this formula gives us, but the distance we must travel *along the surface* to get from one to the other.

Hence the proper definition of $d(x,x')$ is as the length of the shortest path $\gamma\colon [0,1] \to S$ with $\gamma(0) = x$, $\gamma(1) = x'$. For a surface in $\mathbb{R}^3$, we can determine this length via the arclength integral

$$\ell(\gamma) = \int_0^1 \|\gamma'(t)\|\, dt = \int_0^1 \sqrt{\langle \gamma'(t), \gamma'(t)\rangle}\, dt.$$

For a general surface defined without reference to a particular embedding, we need a way of defining the length of the tangent vector $\gamma'(t)$, and this is what a Riemannian metric will give us.

Recall that for a point p on a smooth surface S, we denote the tangent space at p by T_pS. For an embedded surface in $\mathbb{R}^3$, we usually picture the tangent plane as also lying in $\mathbb{R}^3$ and being somehow attached to the surface at p. The problem with this picture is that this plane may intersect the surface at other points as well, and will certainly intersect other tangent planes, even though we want to think of the tangent bundle as being the *disjoint* union of the tangent spaces.

This is easier to visualise if we consider a one-dimensional manifold, the circle. Then the tangent space at each point is simply a line, and if we attach disjoint lines to each point on a circle, we obtain a cylinder, a non-compact two-dimensional manifold, as the tangent bundle of S^1. The tangent bundle of a surface will be a non-compact four-dimensional manifold, which is locally (but not necessarily globally) the direct product of the surface and $\mathbb{R}^2$.

Given an atlas $\mathcal{A}$ on S, we obtain an atlas on the tangent bundle TS with charts given by

$$\begin{aligned} \phi \times \mathrm{Id}\colon \qquad & U \times \mathbb{R}^2 \to \mathbb{R}^4 = \mathbb{R}^2 \times \mathbb{R}^2, \\ & (p, u\partial_x + v\partial_y) \mapsto (x,y,u,v), \end{aligned}$$

where $\phi\colon p \mapsto (x,y)$ is a chart on U, and we use the notation $\partial_x = \frac{\partial}{\partial x}$, $\partial_y = \frac{\partial}{\partial y}$ for the basis vectors in the tangent space T_pS.

To give the definition of a Riemannian metric, we must first recall the definition of an *inner product* on the vector space T_pS.

Definition 4.1. An *inner product* on T_pS is a function

$$\begin{aligned} \langle \cdot, \cdot \rangle_p \colon T_pS \times T_pS &\to \mathbb{R}, \\ (u,v) &\mapsto \langle u, v\rangle_p \end{aligned}$$

with the following properties:

(1) Symmetry: $\langle u, v\rangle_p = \langle v, u\rangle_p$ for all $u, v \in T_pS$.

(2) Bilinearity—that is, linearity in each argument:

$$\begin{aligned}\langle \lambda u_1 + u_2, v\rangle_p &= \lambda\langle u_1, v\rangle_p + \langle u_2, v\rangle_p,\\ \langle u, \lambda v_1 + v_2\rangle_p &= \lambda\langle u, v_1\rangle_p + \langle u, v_2\rangle_p\end{aligned}$$

for all $u, v, u_i, v_i \in T_pS$, $\lambda \in \mathbb{R}$.

(3) Positive definiteness: $\langle u, u\rangle_p \geq 0$, with equality iff $u = 0$.

Such a function is called a *positive definite symmetric bilinear form.* Note that given symmetry, bilinearity follows from linearity in the first variable.

Definition 4.2. A *Riemannian metric* on a surface S is a family of inner products on the tangent spaces T_pS which depend smoothly on the point p.

This definition can of course be made for a manifold in any dimension, not just a surface; a manifold equipped with a Riemannian metric is known as a *Riemannian manifold.*

What does 'smooth' mean in this context? If we write the Riemannian metric in terms of local coordinates, a tangent vector u may be written in terms of its coordinate representation with respect to the standard basis $\{\partial_x, \partial_y\}$ as $u = u_1\partial_x + u_2\partial_y$. If instead of thinking of u_1 and u_2 as fixed real numbers, we allow them to be smooth functions of the coordinates x and y, we obtain a *smooth vector field* $u(x, y) = u_1(x, y)\partial_x + u_2(x, y)\partial_y$ which comprises one tangent vector in each tangent space T_pS, where p varies over the patch U. Now given two such vector fields u and v, we may write the inner products of $u(p)$ and $v(p)$ at any point $p = \phi^{-1}(x, y) \in U$ in terms of the Riemannian metric, using the assumption of bilinearity and symmetry:

$$\begin{aligned}\langle u, v\rangle_p &= \langle u_1\partial_x + u_2\partial_y, v_1\partial_x + v_2\partial_y\rangle_p\\ &= u_1v_1\langle\partial_x, \partial_x\rangle_p + (u_1v_2 + u_2v_1)\langle\partial_x, \partial_y\rangle_p + u_2v_2\langle\partial_y, \partial_y\rangle_p\\ &= \begin{pmatrix} u_1 & u_2\end{pmatrix}\begin{pmatrix} a(x,y) & b(x,y)\\ b(x,y) & c(x,y)\end{pmatrix}\begin{pmatrix} v_1\\ v_2\end{pmatrix},\end{aligned}$$

where $a, b, c\colon \mathbb{R}^2 \to \mathbb{R}$ are given by

$$\begin{aligned} a(x,y) &= \langle \partial_x, \partial_x \rangle_p, \\ b(x,y) &= \langle \partial_x, \partial_y \rangle_p = \langle \partial_y, \partial_x \rangle_p, \\ c(x,y) &= \langle \partial_y, \partial_y \rangle_p. \end{aligned}$$

Then the condition that the metric be smooth may be given by requiring a, b, c to be smooth functions; equivalently, given any two smooth vector fields u, v, we require the map $p \mapsto \langle u, v \rangle_p$ to be smooth.

What does it mean in terms of the above discussion to require that the metric be positive definite? Clearly $\langle \partial_x, \partial_x \rangle_p > 0$, and similarly for ∂_y, so we have $a, c > 0$. This is necessary but not sufficient; it turns out that in addition, $\left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$ must have positive determinant.

Note that this is an open condition; given a matrix A with positive determinant and positive diagonal terms, a small perturbation of A will still have positive determinant and positive diagonal terms, and so small perturbations of our metric will still be positive definite.

One can also check that the matrix A which defines the metric transforms under a change of coordinates to the matrix $C^T AC$, where C is the Jacobian matrix of the transition map. To see this, let (x, y) and (x', y') be two coordinate systems on a neighbourhood of S, with the transition map given by $\phi\colon (x, y) \mapsto (x', y')$. To determine the change of coordinates on the tangent space, we suppose that $(u, v) = u\partial_x + v\partial_y$ is mapped to $(u', v') = u'\partial_{x'} + v'\partial_{y'}$. Then

$$\begin{aligned} u'\partial_{x'} + v'\partial_{y'} &= u\partial_x + v\partial_y \\ &= u\left(\frac{\partial x'}{\partial x}\partial_{x'} + \frac{\partial y'}{\partial x}\partial_{y'}\right) + v\left(\frac{\partial x'}{\partial y}\partial_{x'} + \frac{\partial y'}{\partial y}\partial_{y'}\right) \\ &= \left(\frac{\partial x'}{\partial x}u + \frac{\partial x'}{\partial y}v\right)\partial_{x'} + \left(\frac{\partial y'}{\partial x}u + \frac{\partial y'}{\partial y}v\right)\partial_{y'}, \end{aligned}$$

which we may write more succinctly as

$$\begin{pmatrix} u' \\ v' \end{pmatrix} = D\phi \begin{pmatrix} u \\ v \end{pmatrix},$$

where $D\phi$ is the Jacobian of the transition map ϕ. Hence since for two vector fields $u = (u_1, u_2)$ and $v = (v_1, v_2)$ we have $\langle u, v \rangle_p = u^T Av$, the change of coordinates which gives $u' = (D\phi)u$ and $v' = (D\phi)v$

leads to

$$\begin{aligned}\langle u', v'\rangle_p &= \langle (D\phi)u, (D\phi)v\rangle_p \\ &= u^T(D\phi)^T A(D\phi)v\end{aligned}$$

which is the change of coordinates formula mentioned above.

Exercise 4.1. Express the Riemannian metric on the round sphere in the following coordinate systems:

(1) geographic coordinates;

(2) polar coordinates in the planes of stereographic projections;

(3) polar coordinates in the planes of coordinate projections.

Exercise 4.2. Express the Riemannian metric on the torus of revolution (1.5) in the 'geographic' coordinates (θ, ϕ), where θ is the angle between a plane section passing through the z-axis and the xz-plane, and ϕ is the angular coordinate on a plane section.

Exercise 4.3. Consider the regular octagon with pairs of opposite sides identified, with the smooth structure defined in Lecture 19(d) (see Figure 3.8).

Define a smooth Riemannian metric on this surface in such a way that *angles* between tangent vectors at any point other than the vertex are equal to the Euclidean angles.

b. Partitions of unity. The above definition of a Riemannian metric relies on a choice of local coordinates at each point, and so in order to define a Riemannian metric on the entire surface, we must define it locally on each patch. However, the formula just derived for the change of coordinates must be satisfied where the patches overlap, so we cannot simply choose an arbitrary positive definite symmetric matrix varying smoothly from point to point within each patch. In particular, we cannot obtain a Riemannian structure on a smooth surface by simply defining the metric by the identity matrix within each patch, since the change of coordinates formula will probably fail on the intersections of the different patches.

To overcome this difficulty, we require a tool for passing from the local setting to the global. The tool we will use is a *partition of unity*, which has wide applicability in topology and geometry anytime we want to 'patch together' a collection of objects which have a linear structure and are locally defined.

By 'linear structure', we mean that the objects of interest form a vector space. For example, given two smooth functions f_1 and f_2 on a surface, and any numbers λ_1, λ_2, the linear combination $\lambda_1 f_1 + \lambda_2 f_2$ is also a smooth function, and so the set of smooth functions has a linear structure; a similar observation holds for smooth vector fields.

In the case of Riemannian metrics, it is not hard to verify that the sum of two positive definite symmetric matrices A_1 and A_2 will itself be positive definite and symmetric; however, multiplying A_1 or A_2 by a negative constant will not result in a positive definite matrix, so we must restrict ourselves to multiplication by non-negative values of λ_1 and λ_2. We say that the set of positive definite symmetric matrices, and hence the set of Riemannian metrics, forms a *cone*; as it turns out, this will be sufficient to allow us to apply the partition of unity.

Definition 4.3. Let $\{U_1, \dots, U_N\}$ be a finite cover of S by coordinate patches. A *smooth partition of unity* is a collection $\{\rho_1, \dots, \rho_N\}$ of smooth functions $S \to \mathbb{R}$ which satisfy the following conditions:

(1) $\operatorname{supp}(\rho_i) = \overline{\{\, x \mid \rho_i(x) \neq 0 \,\}} \subset U_i$;

(2) $\rho_i \geq 0$;

(3) $\sum_{i=1}^N \rho_i \equiv 1$.

We will defer until the next lecture a proof that any finite cover of S by coordinate patches admits a smooth partition of unity, and content ourselves for the time being with briefly mentioning the use of this new object.

Suppose we have a collection of functions, or vector fields, or Riemannian metrics, which are only defined locally; that is, for each patch U_i we have a function (or vector field, etc.) A_i which is defined on U_i but nowhere else. Then we can construct a globally defined function (or whatever) A by using the partition of unity, as suggested in Figure 4.1:

$$A = \sum_{i=1}^N \rho_i A_i.$$

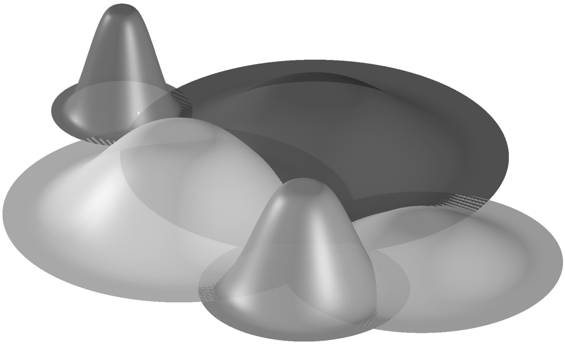


Figure 4.1. Using a partition of unity to build a global object.

The careful reader will protest that A, which is meant to be defined on all of S, is being written as a sum of things which are not so defined. This is where the properties of the partition of unity $\{\rho_i\}$ are vital; because ρ_i vanishes where A_i is not defined, we may simply ignore those terms, and take our sum over only those terms which are defined and not equal to zero. Furthermore, because each ρ_i is smooth, any regularity properties of the locally defined A_i are passed to A itself.

This method of gluing together locally defined objects which have no *a priori* relation to each other is often the only way of defining 'good' global objects, and has wide applicability.

Lecture 24

a. Existence of partitions of unity. We now formally state and prove the theorem on the existence of smooth partitions of unity to which we alluded in the previous lecture.

Theorem 4.4. *Let $\mathcal{U} = \{(U_i, \phi_i)\}_{i=1}^N$ be a finite smooth atlas on a compact surface S, where $U_i \subset S$ are open patches and*

$$\phi_i \colon U_i \to D^2 = \{\, (x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1 \,\}$$

are coordinate charts. Then there exists a smooth partition of unity subordinate to $\mathcal{U}$, that is, smooth functions $\rho_i \colon S \to \mathbb{R}$ such that

(1) $\operatorname{supp}(\rho_i) \subset U_i$;

(2) $\rho_i \geq 0$;

(3) $\sum_{i=1}^{N} \rho_i \equiv 1$.

Proof. We begin with a more general lemma, which applies to any compact topological manifold and does not rely on the smooth structure of our surface. The key idea is that because we are dealing with an *open* cover, we can shrink the patches U_i by some small amount and still cover the entire surface; with this lemma in hand, we will proceed to construct smooth functions ρ_i which have the closures of these shrunken patches as their supports.

Lemma 4.5. *Given a finite smooth atlas $\mathcal{U}$ as above, there exists $\varepsilon > 0$ such that the sets*

$$U_i^\varepsilon = \phi_i^{-1}(D_{1-\varepsilon}^2)$$

still cover S, where $D_r^2 = \{\, (x,y) \in \mathbb{R}^2 \mid x^2 + y^2 < r^2 \,\}$ is the disc of radius r.

Proof of the lemma. We proceed by contradiction; if no such ε exists, then for every $\varepsilon > 0$ we have

$$\bigcup_{i=1}^{N} U_i^\varepsilon \subsetneq S$$

and hence there exists a sequence of points $x_n \in S$ such that $x_n \notin U_i^\varepsilon$ for any $1 \leq i \leq N$. By compactness, $(x_n)_{n=1}^\infty$ has a convergent subsequence; without loss of generality, we may assume that the entire sequence converges to some point $x \in S$.

Now $x \in U_i$ for some i, so write $\phi_i(x) = (t,s) \in D^2$. Then $t^2 + s^2 < 1$ so there exists $\delta > 0$ such that $t^2 + s^2 < (1-\delta)^2$. Hence since $x_n \to x$, there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have $x_n \in U_i$, $\phi(x_n) = (t_n, s_n)$, and $t_n^2 + s_n^2 < (1-\delta/2)^2$. Thus $x_n \in U_i^{\delta/2}$, contradicting our original assumption. □

As mentioned above, this proof makes no reference to the smooth structure of S, and works for a compact manifold of arbitrary dimension by replacing (t,s) with $(t_1, \dots, t_k)$. In the case of a non-compact manifold and an infinite cover, the lemma is not true as stated, but a

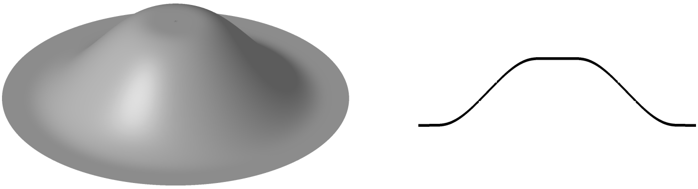


Figure 4.2. Our desired bump function and its radial profile.

similar result is still true and may be used to establish our theorem; we restrict ourselves here to the compact case, however.

Given $\varepsilon > 0$ as in the lemma, we now want to construct smooth functions $\rho_i \colon S \to \mathbb{R}$ such that $\rho_i > 0$ on U_i^ε and $\rho_i = 0$ on $U_i \setminus U_i^\varepsilon$, implying $\operatorname{supp}(\rho_i) = \overline{U_i^\varepsilon} \subset U_i$. The construction of such *bump functions* begins by considering the one-dimensional case.

What we would like in the one-dimensional case is a smooth function $F_\varepsilon \colon \mathbb{R} \to \mathbb{R}$ whose graph is as shown in Figure 4.2. Assuming all the derivatives of F_ε vanish at 0, we can then define ρ_i radially using F_ε. The first task, then, is to construct such an F_ε.

A smooth function which vanishes on one side of a point a must necessarily have all derivatives equal to zero at a; hence we begin by recalling the standard example (Figure 4.3(a)) of a smooth function for which all derivatives vanish at 0, but which is not identically zero on any neighbourhood of the origin. Define $f \colon \mathbb{R} \to \mathbb{R}$ piecewise by

$$f(x) = \begin{cases} 0 & x \leq 0, \\ e^{-1/x^2} & x > 0. \end{cases}$$

Exercise 4.4. Using the fact that the exponential function grows faster than any polynomial, show that $f^{(n)}(0) = 0$ for all $n \geq 0$.

Now to obtain a smooth function with compact support, we fix $a, b \in \mathbb{R}$ and consider the function

$$f_{a,b}(x) = f(x - a) \cdot f(b - x)$$

whose graph is shown in Figure 4.3(b). Since $\operatorname{supp}(f_{a,b}) = [a, b]$, we would like to simply define our bump function by

$$\rho_i(\phi_i^{-1}(x, y)) = f_{-1+\varepsilon, 1-\varepsilon}(\sqrt{x^2 + y^2}).$$

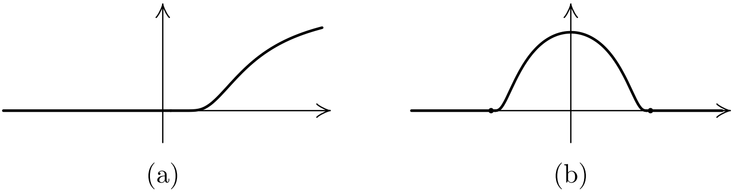


Figure 4.3. Building a partition of unity.

However, this function will not be smooth at $(x, y) = (0, 0)$, since not all derivatives of $f_{a,b}$ vanish at $\frac{a+b}{2}$. To remedy this situation, we define a smooth function $g_{a,b}$ which is constant on $(-\infty, a]$ and vanishes on $[b, \infty)$ by integrating $f_{a,b}$:

$$g_{a,b}(x) = \int_x^\infty f_{a,b}(t)\,dt$$

Note that we could take as our upper bound of integration any real number larger than b.

Now we can once more define a candidate bump function $\tilde{\rho}_i$ by

$$\tilde{\rho}_i(p) = \begin{cases} g_{\varepsilon,1-\varepsilon}(\sqrt{x^2+y^2}), & p = \phi_i^{-1}(x,y) \in U_i, \\ 0, & p \notin U_i. \end{cases}$$

By the construction of $g_{a,b}$, it is immediate that $\tilde{\rho}_i$ is smooth, non-negative, and has support $U_i^\varepsilon \subset U_i$; the only thing left to obtain a partition of unity is the requirement that the functions sum to 1 at each point. This is easily accomplished with a simple normalisation procedure; by the lemma, the patches U_i^ε cover S, and hence $\sum_{i=1}^N \tilde{\rho}_i(x) > 0$ for every $x \in S$. By defining

$$\rho_i(x) = \frac{\tilde{\rho}_i(x)}{\sum_{i=1}^N \tilde{\rho}_i(x)}$$

we have the desired smooth partition of unity. □

This construction relies heavily on the dramatic difference between smoothness and analyticity for real functions; in the complex case, where the two are equivalent, no such argument would have been possible. In essence, we are using the pathological nature of smooth real functions for our own ends.

b. Global properties from local and infinitesimal. In the previous lecture, we described how to use a partition of unity (which we now know exists) to construct a Riemannian metric on any compact smooth surface S. This gives a useful example of producing a global object from components which are only defined *locally*.

Riemannian metrics are an outstanding example of how an *infinitesimally* defined object leads to global or, more appropriately, *macroscopic*, considerations. In the first approximation, Riemannian geometry is modeled on Euclidean geometry; this can be likened to approximating a differentiable function near a point by a linear one with the same value at the point and the slope equal to the derivative at the point.

Certain properties of the function such as convexity, or, geometrically speaking, the curvature of the function's graph, are lost in such an approximation since they depend on higher derivatives. If the second derivative does not vanish, then quadratic approximation recovers at least the convexity properties of the function and the curvature of its graph.

For a Riemannian metric, the linear approximation corresponds to an approximation by a metric with constant coefficients in a given coordinate system; it obviously misses important geometric properties, such as the radius of the sphere in the case of a spherical metric, which is closely related to the curvature. The recipe, then, is clear: we must recover these properties by considering the change in the Riemannian metric from point to point, which we will do by taking the first (and, if necessary, higher) derivatives of the coefficients into account. We will come back to this in a systematic way later in Lecture 32.

But there is also another aspect to the relationship between global and infinitesimal properties. Let us look at the basic calculus example again.

In order to find the minima of a differentiable function, which is a global property, we examine how the function should behave near such a point and deduce that the point must be critical, i.e. all partial derivatives must vanish. Then, in order to determine whether a critical point is a minimum, a maximum or neither, we apply the second derivative (Hessian) test. Finally, having determined all local minima, we simply compare the values of the function at those points to determine the global minimum.

A similar method works in finding curves which play the role of straight lines in Riemannian geometry, the *geodesics*. The distance between two points $a, b \in S$ is defined as the minimum of lengths of paths connecting a and b; the question of finding a shortest path is, on the face of it, a rather difficult global question, requiring us to somehow consider all possible paths from a to b. By using a local approach, we will be able to identify the analogues of the critical points in the previous problem; that is, the paths which cannot be made shorter by a small perturbation.

This *variational approach* leads eventually to the second-order *Euler-Lagrange differential equation* for a geodesic parametrised by arc length, and will allow us to restrict our search to a much smaller class of paths. We will see that the solution is uniquely defined by the initial condition and initial "velocity", i.e. the tangent vector of length one at the initial point. A counterpart of the second derivative test can be used to show that for any two sufficiently close points the solution indeed has minimal length.

Unlike the case of the Euclidean plane, the situation becomes more complicated if the endpoints are far away or if a geodesic comes back or close to the initial point. The latter is inevitable on compact surfaces, even if the geometry locally looks Euclidean, as it does for the flat torus.

c. Lengths, angles, and areas. By recalling some facts from Euclidean geometry, we observe that the choice of a Riemannian metric allows us to define lengths, angles, and areas in the tangent space to a surface S at a point p.

First note that given a tangent vector $u \in T_pS$, we can define the length (or *norm*) of u by the formula $\|u\|_p = \sqrt{\langle u, u \rangle_p}$.

Now consider the triangle shown in Figure 4.4. The law of cosines states that

$$\|u + v\|_p^2 = \|u\|_p^2 + \|v\|_p^2 + 2\|u\|_p\|v\|_p \cos \alpha.$$

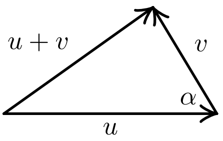


Figure 4.4. Calculating the angle between two vectors.

Combining this with the above formula for the length, we have

$$\langle u+v, u+v\rangle_p = \langle u,u\rangle_p + \langle v,v\rangle_p + 2\|u\|_p\|v\|_p \cos\alpha,$$

and expanding the left side using the properties of the inner product yields

$$\langle u+v, u+v\rangle_p = \langle u,u\rangle_p + \langle v,v\rangle_p + 2\langle u,v\rangle_p;$$

whence we have

$$\alpha = \arccos \frac{\langle u,v\rangle_p}{\|u\|_p\|v\|_p}$$

and so a Riemannian metric allows us to define angles between tangent vectors.

Finally, once lengths and angles are defined, we also have a notion of area. For example, the parallelogram spanned by the vectors $u, v \in T_pS$ has area $\|u\|_p\|v\|_p \sin\alpha$, where α is the angle between u and v.

These are all infinitesimal notions, being defined in the tangent space. We can in fact obtain global counterparts to all of these, which are defined on the surface itself.

The case of the angle is the easiest since it requires only differentiation and no integration. Namely, given two smooth curves γ, $\eta\colon (-\varepsilon,\varepsilon) \to S$ with $\gamma(0) = \eta(0) = p$, the tangent vectors $\gamma'(0)$ and $\eta'(0)$ both lie in T_pS, and so the angle between the two curves is defined as the angle between their tangent vectors at the point of intersection p.

The length of a smooth curve $\gamma\colon [a,b] \to S$ is given by the formula

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

It must be checked that this length is independent of a particular parametrisation of γ; that is, given a smooth monotone increasing function $s\colon [c,d] \to [a,b]$, the curve $\tilde{\gamma}$ defined by

$$\tilde{\gamma}(s) = \gamma(s(t))$$

should have the same length as γ. This follows immediately from the change of variable formula from the calculus of one variable.

Finally, we can define the area for a domain $D \subset S$ bounded by piecewise smooth curves. One can cut such a domain into finitely many pieces such that every piece lies inside a single coordinate patch. Thus we will begin by considering such domains; let (x,y) be local coordinates and $\rho(x,y)$ be the area of the parallelogram spanned by the coordinate vector fields $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$. Then the area of a domain D is defined as

$$a(D) = \int_D \rho(x,y)\,dx\,dy.$$

The change of variables formula from the calculus of two variables shows that this definition is independent of coordinate changes, and hence the area of a large domain can be defined as the sum of the areas of its pieces which lie inside single coordinate patches.

Exercise 4.5. Prove that if we only know the lengths of all tangent vectors for a particular Riemannian metric, we can find the angles between them in terms of those lengths, and thus recover the metric completely.

Lecture 25

a. Geometry via a Riemannian metric. The concept of a Riemannian manifold, introduced in the previous two lectures, lies at the heart of modern geometry. Indeed, when we use the word 'geometry' nowadays, what is usually meant is the study of Riemannian manifolds; this covers both Euclidean and non-Euclidean cases, including hyperbolic geometry and the geometry of projective space.

Three of the main ingredients of two-dimensional geometry are length, angle, and area. Previously, we saw how to define the infinitesimal versions of these on the tangent space, and went through the process of obtaining the macroscopic versions by a process of integration (in the case of length and area) or differentiation (in the case of angle).

Comment on notation: It is common to see a Riemannian metric defined on a patch by the equation

$$ds^2 = a(x,y)\,dx^2 + 2b(x,y)\,dx\,dy + c(x,y)\,dy^2, \tag{4.1}$$

which specifies the magnitude of an infinitesimal displacement in terms of its coordinates. This corresponds to our definition of the inner product on each tangent space as being given by the matrix

$$\begin{pmatrix} a(x,y) & b(x,y) \\ b(x,y) & c(x,y) \end{pmatrix}.$$

In the case of a Euclidean metric, when this matrix becomes the identity, we have the familiar formula

$$ds^2 = dx^2 + dy^2.$$

In our discussion of complex manifolds and Riemann surfaces we encountered the notion of a conformal map, which preserves angles but not necessarily distances. A related concept for a Riemannian metric is the idea of a *conformal change*, which replaces the metric given by ds^2 with another given by $\rho(x,y)^2\,ds^2$, where $\rho(x,y)$ is a non-vanishing smooth function. That is, the length of each tangent vector in the tangent bundle is scaled by a factor which depends only on the base point (x,y), and not on the particular vector itself.

This operation gives us a useful tool in classifying Riemannian metrics on surfaces, in that via a conformal change, we can put every metric on a compact surface into some canonical form. It turns out, for instance, that every Riemannian metric on the sphere is *conformally equivalent* to the usual round metric obtained by embedding the unit sphere in $\mathbb{R}^3$. Similarly, any metric on the torus is conformally equivalent to some flat metric; it should be pointed out, however, that the various flat metrics, which may be obtained by using different parallelograms (or rectangles) as our planar model for the torus, are *not* conformally equivalent. These facts, and their even more non-trivial generalisations for surfaces of higher genus, rely on advanced (albeit by now standard) results from complex analysis called *regularisation theorems.*

Exercise 4.6. Consider a Riemannian metric given in terms of local coordinates (4.1). Interpret the following conditions in terms of the coefficients of the metric:

(1) The coordinate curves $x =$ constant and $y =$ constant are orthogonal.

(2) The coordinate curves $x =$ constant and $y =$ constant form the angle $\pi/4$ at each point.

(3) The area determined by the metric coincides with the usual area $dx\,dy$.

b. Differential equations. We briefly recall some notions from the theory of ordinary differential equations. Given a system of n first order ODEs

$$\dot{x}_i = f_i(x, t)$$

for $x \in U \subset \mathbb{R}^n$, we are in general unable to find an explicit closed form solution $x(t)$. However, provided the functions f_i are 'nice enough'—for example, if they are continuously differentiable—it is possible to prove that for every set of initial conditions there exists a unique solution $x(t)$ on some interval $t \in (0, t_0)$.

Such existence and uniqueness results are central to the study of ordinary differential equations, and their counterparts also appear in the study of partial differential equations. We will rely on this sort of result when we investigate geodesic curves on a surface; in particular, the existence of geodesics will hinge on the existence of solutions to the Euler-Lagrange equations, which can be brought to the form

$$\ddot{x}_i = g_i(x, \dot{x})$$

for some functions g_i.

This system of n second-order ODEs reduces to a system of $2n$ first-order ODEs using the standard trick of setting $v = \dot{x}$, which gives

$$\begin{aligned} \dot{x}_i &= v_i(t), \\ \dot{v}_i &= g_i(x, v, t), \end{aligned}$$

allowing us to apply the existence and uniqueness theorem mentioned above.

c. Geodesics. We now have definitions of length, angle, and area on a surface endowed with a Riemannian metric; we have not yet dealt, however, with the analogue of a basic geometric object, a straight line. To this end, we fix two points a and b on the surface and look for the shortest curve between them (though *a priori* we have no guarantee that such a curve exists and is unique).

Since the length of a curve is defined via a parametrisation of the curve, we are dealing with a real-valued function (the length) whose domain is the set of all parametrised curves $\gamma\colon [0,s] \to S$ with $\gamma(0) = a$, $\gamma(s) = b$. This is an extremely large set, being a sort of infinite-dimensional manifold; in this context, the function assigning a length to each parametrised curve is referred to as a *functional*, which we can write as

$$\ell\colon \gamma \mapsto \int_0^s \|\gamma'(t)\|_{\gamma(t)}\, dt.$$

Now we are looking for the curve (or curves) γ which minimise this functional; we would like to use a sort of derivative to identify critical points which will be the candidates for minima. This is hampered by the fact that if γ is such a minimum, then any reparametrisation of γ is also a minimum, since length is independent of parametrisation. This will mean that the 'critical curves' for the length functional are not isolated in the set of parametrised curves, which is problematic.

The way around this problem is to choose a preferred parametrisation for each curve; specifically, we focus on the parametrisation by *arc length*, for which

$$\int_0^t \|\gamma'(\tau)\|\, d\tau = t$$

for every $t \geq 0$. This is obviously equivalent to the condition $\|\gamma'(t)\| = 1$ for all t, for which reason this is sometimes referred to as the *unit speed* parametrisation.

We single out these parametrisations not by restricting our space of curves to arc length parametrisations, but by considering a slightly different functional, the *action* α, defined by

$$\alpha(\gamma) = \int_0^s \|\gamma'(t)\|^2\, dt.$$

In the case $\|\gamma'\| \equiv 1$, we have $\alpha(\gamma) = \ell(\gamma)$; a variant of the Cauchy-Schwarz inequality shows that for any other parametrisation, $\alpha(\gamma) > \ell(\gamma)$, and so the minima of α are precisely the minima of ℓ which are parametrised by arc length.

Some justification for the inequality may be given by considering the problem of minimising $x_1^2 + \cdots + x_n^2$ subject to the restrictions $x_i \geq 0$, $x_1 + \cdots + x_n = 1$. This is equivalent to finding the point on the unit simplex closest to the origin; the unique minimum occurs when $x_i = 1/n$ for every i.

By using tools from the calculus of variations, one may obtain a criterion for a curve γ to be a critical point of the action functional α. We will not carry out the details, but rather will state without proof the *Euler-Lagrange equation*, which applies in a more general setting than just the problem of minimising the action functional.

Proposition 4.10 (Euler-Lagrange equation). *If $\gamma\colon [a,b] \to \mathbb{R}^n$ minimises the functional $\int L(x,\dot{x})\,dt$, then the partial derivatives of the cost function L are related by the equation*

$$\frac{\partial L}{\partial x} = \frac{d}{dt}\frac{\partial L}{\partial \dot{x}} \tag{4.2}$$

at every point along $\gamma([a,b])$.

Applying this criterion to the action functional, for which the cost function is $\|\dot{x}\|^2$, we obtain a second order ODE, the solution of which is a geodesic. Assuming the Riemannian metric is $\mathcal{C}^2$, the existence and uniqueness theorem discussed above applies, and we have the following important result:

Proposition 4.11. *Given a $\mathcal{C}^2$ Riemannian metric on a smooth surface, there exists $\varepsilon > 0$ such that for every $v \in T_pS$ with $\|v\| = 1$, there exists a unique curve $\gamma_v\colon [0,\infty) \to S$ satisfying*

1. $\gamma_v'(0) = v$;
2. $\|\gamma_v'\| \equiv 1$;
3. *If $|t_1 - t_2| \leq \varepsilon$ then $\gamma\colon [t_1,t_2] \to S$ is the unique shortest curve between $\gamma(t_1)$ and $\gamma(t_2)$.*

The final property is the key property of geodesics, and establishes that for points which are close enough *along the geodesic*, it does in fact minimise length. In a similar vein, the following result can also be shown by using the previous proposition along with the Implicit Function Theorem, as well as the fact (which we did not state yet) that γ_v depends smoothly on v.

Proposition 4.12. *Under the conditions above, there exists $\varepsilon > 0$ such that if $p, q \in S$ lie a distance $< \varepsilon$ apart, then there exists a unique shortest curve $\gamma_{p,q}$ from p to q in the arc length parametrisation. Further, if $v = \gamma'_{p,q}(0)$, then $\gamma_{p,q} = \gamma_v$.*

Both these propositions deal with small scales; if we go farther away along a geodesic, various sorts of behaviour are possible. In the Euclidean plane, nothing changes; two points determine a unique straight line, no matter what the distance between them is. On the sphere, however, we recall that the geodesics are great circles, and so all the geodesics γ_v converge at the point antipodal to p (as in Figure 1.13). This is an instance of the problem of *conjugate points.*

On a flat torus, the situation is different yet again. Any two points on the flat torus can be connected by infinitely many geodesics, but they will be of different lengths, unlike on the sphere, where all great circles have the same length.

Although we have not derived explicit differential equations for the geodesics, and did not provide any other general recipe for finding them, there are certain cases where Propositions 4.11 and 4.12 allow us to identify specific geodesics. The key idea here is the symmetry; the following exercises demonstrate how this idea is used in several situations.

Exercise 4.7. Prove that great circles are geodesics on the round sphere, and that there are no other geodesics. Use only Proposition 4.12, and do not carry out any calculations.

Exercise 4.8. Let $F(x, y, z)$ be a differentiable function which is even in z, i.e. $F(x, y, z) = F(x, y, -z)$, and for which 0 is not a critical value. Prove that every connected component of the curve $\{ (x, y, z) \mid F(x, y, z) = 0, z = 0 \}$ is a geodesic on the surface $F = 0$.

Exercise 4.9. Prove that every ellipsoid has at least three closed geodesics without self-intersections.

Exercise 4.10. Let S be a surface of revolution, and P a plane passing through the axis of rotation. Show that the connected components $S \cap P$ are geodesics in S.

Exercise 4.11. Assume a smooth surface S in $\mathbb{R}^3$ is symmetric with respect to a rotation R by π around an axis which intersects S at an isolated point p. Show that R acts as the geodesic flip near p, i.e. it keeps any geodesic passing through p invariant, and reflects it through p, preserving the length parameter.

Lecture 26

a. First glance at curvature. We now turn our attention to what is perhaps the most important isometric invariant of a surface endowed with a Riemannian metric, the *curvature*. Specifically, we shall be interested in what is referred to as the *Gaussian curvature*; there are other sorts of curvature as well, but we shall not dwell on them, and so any mention of curvature in what follows refers to Gaussian curvature, unless otherwise stated.

Two of the standard examples to keep in mind during our discussion of curvature are the Euclidean plane and the round sphere. As one might expect, the plane has zero curvature, while the sphere has a curvature which varies according to its radius; a sphere with small radius will have a large curvature, and conversely. In fact, we will see that the curvature of a sphere with radius R is $1/R^2$; some motivation for the fact that the curvature varies as the inverse square of the radius, and not some other power, may be given by the observation that under this definition, the *total curvature* of the sphere, obtained by integrating the curvature at each point with respect to the area generated by the metric over the entire surface, is in fact independent of the radius, since the surface area grows as $4\pi R^2$.

These two examples exhibit zero curvature and positive curvature, respectively, at every point; is there a corresponding surface exhibiting constant negative curvature? There is indeed such an example, the hyperbolic plane, which cannot be isometrically embedded into $\mathbb{R}^3$—we will turn our attention to this example following some preliminary remarks concerning the definition of curvature without reference to an embedding. After a thorough study of the invaluable example of the hyperbolic plane, we will then flesh out these preliminary remarks with a more systematic discussion of curvature.

Traditionally, differential geometry has considered various aspects of curvature which arise from the particular embedding of a surface into $\mathbb{R}^3$. In this approach, curvature is first studied in terms of the *extrinsic* properties of a surface; that is, with reference to the ambient space $\mathbb{R}^3$ and the particular choice of embedding. With a fair amount of work, one comes eventually to Gauss' Theorema Egregium (see e.g. Section 20.1 in Coxeter's *Introduction to Geometry*), which gives a characterisation of one of the several curvature characteristics of the embedded surface in purely *intrinsic* terms; that is, using only the properties of the Riemannian metric on the surface.

The difference between the extrinsic and intrinsic points of view is made apparent when we compare the idea of curvature for curves and for surfaces. Given a curve γ in $\mathbb{R}^2$, the curvature is given by the speed of rotation of the unit tangent vector. That is, if we consider the arc length parametrisation and let $\theta(s)$ denote the angle between $\gamma'(s)$ and the positive x-axis, then the curvature κ at a point $\gamma(s)$ is given by

$$\kappa(\gamma(s)) = \frac{d\theta}{ds}.$$

Similarly to our earlier claim regarding the sphere, we find that a circle of radius R has a constant curvature of $1/R$.

Two observations must now be made, however. The first is that the curvature is a property not only of the curve, but also of its orientation; if we parametrise the curve in the opposite direction, the curvature will change sign. The second, which will illustrate the difference between curves and surfaces with regard to curvature, is that the curvature is completely dependent upon the extrinsic properties of the curve; after all, any small neighbourhood of a smooth curve is isometric to an interval on a straight line, and so the intrinsic properties of the curve have nothing whatsoever to do with the curvature.

In marked contrast to this, we will see that intrinsic properties are sufficient to determine curvature of a surface. For the moment, let us address the question of what properties the curvature ought to have. Just what sort of beast are we after here?

The curvature is to be a real-valued function $\kappa\colon S \to \mathbb{R}$ which is invariant under isometries; that is, which is intrinsically determined. Furthermore, it is to have the property that if we scale the metric by a constant λ, then we scale the curvature by $1/\lambda^2$; that is, if $\tilde{S}$ is the surface S with Riemannian metric given by

$$d\tilde{s}^2 = \lambda^2 ds^2,$$

then the curvature $\tilde{\kappa}\colon \tilde{S} \to \mathbb{R}$ is given by

$$\tilde{\kappa} = \frac{1}{\lambda^2}\kappa.$$

What do these properties tell us about the curvature of a sphere? Given any two points p, q on the sphere, we can find a rotation which takes p to q; because rotations are isometries, the fact that κ is to be invariant under isometries implies that $\kappa(p) = \kappa(q)$, and hence the sphere has constant curvature. Further, scaling the metric by λ is equivalent to scaling the radius R by λ, and hence the constant value of the curvature is proportional to $1/R^2$ by the second property above. Thus the two properties above are sufficient to determine the curvature up to this constant of proportionality, which is chosen so that the unit sphere has curvature $\kappa = 1$.

A similar series of observations holds for the Euclidean plane; as for the sphere, any point p can be carried to any other point q by an isometry (in particular, translation by $q - p$), and so the isometry group acts transitively; hence the curvature must be constant. Furthermore, because scaling the metric results in a copy of the plane which is isometric to the original, this constant must be zero, so $\kappa = 0$ for the Euclidean plane.

To characterise curvature in purely geometric terms, without reference to an embedding, consider a small circle around a point p on a surface S. The definition of a circle as the set of all points at a fixed distance r from p still makes perfect sense; however, the usual formulae for circumference and area will need to be modified with a small error term. On a sphere, for instance, a circle around the north pole with radius r will have a shorter circumference than a circle in

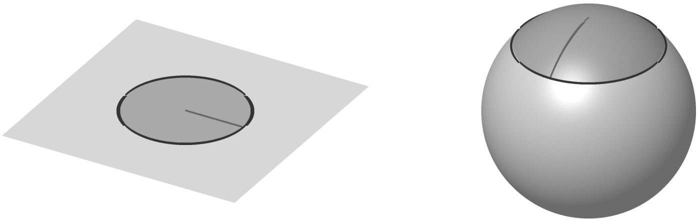


Figure 4.5. Relating curvature to the circumference of a circle.

the plane with radius r (Figure 4.5). We will see that

$$\text{circumference} = 2\pi r - cr^3 + o(r^3)$$

where c is a constant related to the curvature. Upon integration, we will obtain an expression for the area of the disc as

$$\text{area} = \pi r^2 - \frac{c}{4} r^4 + o(r^4).$$

b. The hyperbolic plane: two conformal models.

b.1. *The upper half-plane model.* In order to exhibit a surface with constant *negative* curvature, we pull a proverbial rabbit from our sleeve, or hat, or some other piece of proverbial clothing, and give without motivation the definition of the upper half-plane model of hyperbolic geometry due to Henri Poincaré, arguably the greatest mathematician since Gauss and Riemann. Our surface will be H^2, defined as

$$H^2 = \{\, (x,y) \in \mathbb{R}^2 \mid y > 0 \,\} = \{\, z \in \mathbb{C} \mid \operatorname{Im} z > 0 \,\},$$

where it is useful to keep in mind the formulation in terms of complex numbers in order to describe the isometry group of H^2.

The metric on H^2 is given by a conformal change of the standard metric:

$$ds^2 = \frac{dx^2 + dy^2}{y^2}. \tag{4.3}$$

The fact that the denominator vanishes when $y = 0$ gives some justification for the fact that we consider only the upper half-plane, and not the entire plane. From (4.3) it is apparent that Euclidean lengths are increased when y is small, and decreased when y is large; Figure 4.6

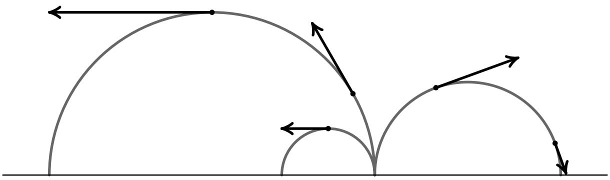


Figure 4.6. Unit tangent vectors in the hyperbolic plane.

shows some unit tangent vectors. All of these have unit length in the hyperbolic metric, and so their Euclidean lengths vary as y varies.

In order to show that H^2 has constant curvature, we will show that isometries act transitively. To see this, it will suffice to exhibit two particular classes of isometries.

(1) *Translations.* Given a real number t, the translation by t which takes z to $z + t$ (or in real coordinates, (x, y) to $(x + t, y)$) is an isometry since the metric does not depend on the horizontal coordinate x.

(2) *Homotheties.* For any $\lambda > 0$, the map $z \mapsto \lambda z$ turns out to be an isometry; this is most easily seen by writing the metric as

$$ds = \frac{(dx^2 + dy^2)^{\frac{1}{2}}}{y}$$

from which it is clear that multiplying both x and y by λ does not change ds.

Since any composition of these two types of isometries is itself an isometry, the isometry group acts transitively on H^2; given $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$, we can first scale z_1 by y_2/y_1 so that the imaginary parts are the same, and then translate by the difference in the real parts. It follows that H^2 has constant curvature.

Acting transitively on the surface itself is not the whole story, however; in the case of the sphere and the Euclidean plane, the isometry group acts transitively not only on the surface, but also on the unit tangent bundle.

By way of explaining this last statement, recall the general fact that given any smooth map $f\colon S \to S$, the Jacobian Df_p at a point p

defines a linear transformation between the tangent spaces T_pS and $T_{f(p)}S$, so that the pair (f, Df) acts on the tangent bundle as

$$\begin{aligned}(f, Df)\colon \quad TS &\to TS,\\ (p, v) &\mapsto (f(p), Df_p v).\end{aligned}$$

Now f is an isometry iff Df acts isometrically on each tangent space; in particular, it must preserve the norm. Thus we restrict our attention to tangent vectors of norm one, which form the *unit tangent bundle*; for each isometry f acting on S, the pair (f, Df) acts isometrically on the unit tangent bundle of S.

For both S^2 and $\mathbb{R}^2$, this action is transitive; given any two points $p, q \in S$ and unit tangent vectors $v \in T_pS$, $w \in T_qS$, there exists an isometry $f\colon S \to S$ such that

$$\begin{aligned}f(p) &= q,\\ Df_p(v) &= w.\end{aligned}$$

To see that a similar property holds for H^2, we must consider all the isometries and not just those generated by the two classes mentioned so far. For example, we have not yet considered the orientation reversing isometry $(x, y) \mapsto (-x, y)$.

We will prove later (Proposition 4.14) that every orientation preserving isometry of H^2 has the form

$$f\colon z \mapsto \frac{az + b}{cz + d},$$

where $a, b, c, d \in \mathbb{R}$. This condition guarantees that f fixes the real line, which must hold for any isometry of H^2. We also require that $ad - bc \neq 0$, since otherwise the image of f is a single point; in fact, we must have $ad - bc > 0$; otherwise f swaps the upper and lower half-planes.

As given, f appears to depend on four real parameters, while considerations similar to those in the analysis of the isometry groups of S^2 and $\mathbb{R}^2$ suggest that three parameters ought to be sufficient. Indeed, scaling all four coefficients by a factor $\lambda > 0$ leaves the transformation f unchanged, but scales the quantity $ad - bc$ by λ^2; hence we may require in addition that $ad - bc = 1$, and now we see that f belongs to a three-parameter group.

The condition $ad - bc = 1$ is obviously reminiscent of the condition $\det A = 1$ for a 2×2 matrix $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. In fact, if given such a matrix A we denote the transformation given above by f_A, then a little algebra verifies that

$$f_{AB} = f_A \circ f_B$$

and so the isometry group of H^2 is isomorphic to $SL(2, \mathbb{R})$, the group of 2×2 real matrices with unit determinant, modulo the provision that $f_I = f_{-I} = \mathrm{Id}$, and so we must take the quotient of $SL(2, \mathbb{R})$ by its centre $\{\pm I\}$. This quotient is denoted $PSL(2, \mathbb{R})$, and hence we will have

$$\mathrm{Isom}(H^2) = PSL(2, \mathbb{R}) = SL(2, \mathbb{R}) / \pm I$$

once we show that f_A is an isometry for every $A \in SL(2, \mathbb{R})$, and that every isometry is of this form. One way to prove the first statement (the second will be Proposition 4.14) is to show that every such f_A can be decomposed as a product of isometries which have one of the following three forms:

$$\begin{aligned} z &\mapsto z + t, \\ z &\mapsto \lambda z, \\ z &\mapsto -\frac{1}{z}, \end{aligned}$$

where $t \in \mathbb{R}$ and $\lambda \in \mathbb{R}^+$ define one-parameter families of isometries. This is equivalent to showing that $SL(2, \mathbb{R})$ is generated by the matrices

$$\left\{ \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \,\middle|\, t \in \mathbb{R} \right\} \bigcup \left\{ \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \,\middle|\, \lambda \in \mathbb{R}^+ \right\} \bigcup \left\{ \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \right\}.$$

We have seen already that the first two transformations preserve the metric (4.3). To see that $z \mapsto \tilde{z} = -1/z$ is an isometry, one must suffer through a small amount of algebra and use the fact that for $z = x + iy$ we have

$$\tilde{z} = -\frac{1}{z} = \frac{-1}{x + iy} = -\frac{x - iy}{x^2 + y^2} = \frac{-x + iy}{x^2 + y^2},$$

which allows us to compute

$$d\tilde{x} = \frac{(x^2 - y^2)\, dx - 2xy\, dy}{(x^2 + y^2)^2}$$

along with a similar formula for $d\tilde{y}$; together, these let us deduce that $d\tilde{s} = ds$, showing that the map is an isometry.

Later we will give other proofs that any fractional linear transformation with real coefficients and non-vanishing determinant is a hyperbolic isometry.

b.2. *The disc model.* Remember that at least one motivation for considering the hyperbolic plane was to provide an ideal model of a surface of negative curvature.[1] In attempting to define curvature via excess or defect in the length of a small circle or area of a small disc, and to calculate it explicitly for the hyperbolic plane, we will find that our life is made easier by the introduction of a different model, which is also due to Poincaré. This is given by an open unit disc, for which the boundary of the disc plays the same role as was played by the real line with respect to H^2 (the so-called *ideal boundary*). The metric is given by

$$ds^2 = \frac{4(dx^2 + dy^2)}{(1 - x^2 - y^2)^2}, \tag{4.4}$$

and we may see that this model is the image of H^2 under a conformal transformation, for example

$$z \mapsto \frac{iz + 1}{z + i}. \tag{4.5}$$

An advantage of this model is that rotation around the origin is an isometry, and so hyperbolic circles around the origin are simply Euclidean circles in the plane with the same centre—of course, the hyperbolic radius is different from the Euclidean radius. This rotation is exactly the one type of isometry which does not have a convenient 'natural' representation in the upper half-plane model; thus it is useful to switch back and forth between the two models depending on the type of symmetry for which a particular problem calls.

b.3. *Embedded surfaces.* It is natural to ask whether one can realise the hyperbolic plane as a surface in $\mathbb{R}^3$. This turns out not to be possible for the whole plane (although the proof is not simple); however, there are surfaces in $\mathbb{R}^3$ whose intrinsic geometry is *locally* isometric

[1] There are of course plenty of other reasons—it is sufficient to recall that the geometry of the hyperbolic plane is the original non-Euclidean geometry where all the standard axioms except for the fifth postulate hold.

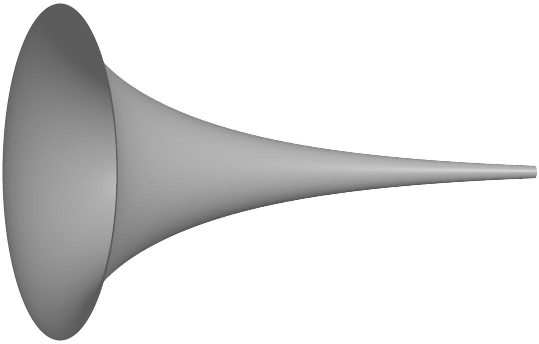


Figure 4.7. A pseudosphere.

to that of hyperbolic plane, in the same manner as the cylinder, for example, is locally Euclidean, despite not being globally isometric to $\mathbb{R}^2$.

The classic example of such a surface is the *pseudosphere* (Figure 4.7), the surface of revolution around the x-axis of the curve in the xz-plane called a *tractrix*, which is given parametrically by

$$(x, z)(t) = \left(t - \frac{\sinh t}{\cosh t}, \frac{1}{\cosh t}\right)$$

where $t \geq 0$. In order to see that the pseudosphere is locally isometric to the hyperbolic plane, one introduces coordinates on the pseudosphere in which the Riemannian metric induced from $\mathbb{R}^3$ has the same form as in the upper half-plane model of the hyperbolic plane.

c. Geodesics and distances on H^2. On an arbitrary surface with a Riemannian metric, the process of defining an explicit distance function and describing the geodesics can be quite tortuous. For the two spaces of constant curvature that we have already encountered, the solution turns out to be quite simple; on the Euclidean plane, geodesics are straight lines and the distance between two points is given by Pythagoras' formula, while on the sphere, geodesics are great circles and the distance between two points is proportional to the central angle they subtend.

One might expect, then, that the situation on H^2 exhibits a similar simplicity, and this will in fact turn out to be the case. Let us first consider two points $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ with equal real parts $x_1 = x_2 = x$ and $y_2 > y_1$. Then it is fairly straightforward to see that the shortest path between z_1 and z_2 is a vertical line. For this curve we have

$$\ell(\gamma) = \int_{y_1}^{y_2} \left\| \frac{d}{dt}(x+it) \right\|_{x+it} dt = \int_{y_1}^{y_2} \frac{1}{t}\, dt = \log y_2 - \log y_1 \tag{4.6}$$

and the length of any other curve will be greater than this value due to the contribution of the horizontal components of the tangent vectors—we will present this argument in more detail in the next lecture. It follows that vertical lines are geodesics in H^2.

Isometries preserve geodesics, and hence the image of a vertical line under any of the isometries discussed above is also a geodesic. Horizontal translation and scaling by a constant will map a vertical line to another vertical line, but the map $z \mapsto -1/z$ behaves differently. This map is the composition of reflection about the imaginary axis with the map $z \mapsto -1/\bar{z}$, and the latter is simply inversion in the unit circle. We encountered this map in Exercise 1.7 as the map

$$(x, y) \mapsto \left(\frac{x}{x^2+y^2}, \frac{y}{x^2+y^2} \right)$$

which arises as the transition map between stereographic projections from the north and south poles. It may be checked that this map takes lines to circles and circles to lines (with the exception of lines through the origin, which are mapped into themselves, and circles centred at the origin, which are taken into other circles centred at the origin); in particular, vertical lines are mapped to circles whose centres lie on the x-axis, and hence half-circles in H^2 with centres on the real axis are also geodesics.[2]

Because the three classes of isometries just mentioned generate the isometry group of H^2, which acts transitively on the tangent bundle, these are all the geodesics.

[2] In the next lecture we will prove that any fractional linear transformation $z \mapsto \frac{az+b}{cz+d}$, where a, b, c, d are arbitrary *complex* numbers such that $ad - bc \neq 0$, maps lines and circles into lines and circles.

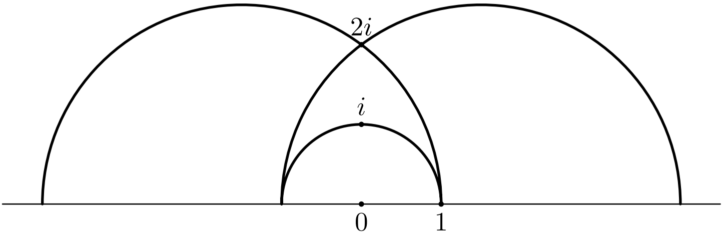


Figure 4.8. Failure of the parallel postulate in H^2.

With this characterisation of geodesics in hand, we can immediately see that Euclid's parallel postulate fails in the hyperbolic plane; given the upper half of the unit circle, which is a geodesic, and the point $2i$, which is a point not on that geodesic, there are many geodesics passing through $2i$ which do not intersect the upper half of the unit circle, as shown in Figure 4.8.

We now come to the question of giving a formula for the distance between two points $z_1, z_2 \in H^2$. Distance must be an isometric invariant, and must also be additive along geodesics. We may construct a geodesic connecting z_1 and z_2 by drawing the perpendicular bisector of the line segment between them and taking the intersection of this bisector with the real line. The circle centred at this point of intersection which passes through z_1 and z_2 will be the geodesic we seek.

As shown in Figure 4.9, let w_1 and w_2 be the points at which this circle intersects the real line; we will prove later (Lemma 4.7) that the *cross-ratio*

$$(z_1, z_2; w_1, w_2) = \frac{z_1 - w_1}{z_2 - w_1} \div \frac{z_1 - w_2}{z_2 - w_2} \tag{4.7}$$

is preserved by all isometries of H^2. It turns out to be multiplicative along geodesics, not additive; if we place a third point z_3 between z_1 and z_2 along the circle as in Figure 4.9, we will have

$$\left| \frac{z_1 - w_1}{z_2 - w_1} \div \frac{z_1 - w_2}{z_2 - w_2} \right| = \left| \frac{z_1 - w_1}{z_3 - w_1} \div \frac{z_1 - w_2}{z_3 - w_2} \right| \times \left| \frac{z_3 - w_1}{z_2 - w_1} \div \frac{z_3 - w_2}{z_2 - w_2} \right|.$$

Hence to obtain a true distance function which is additive along geodesics, we must take the logarithm of the cross-ratio. Notice from

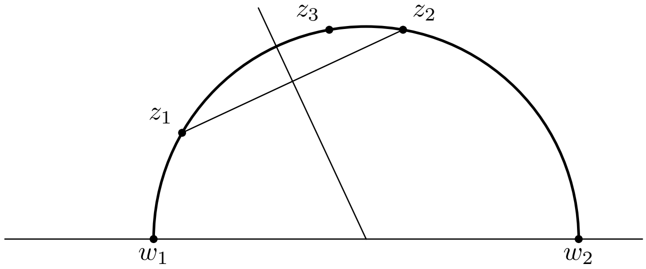


Figure 4.9. Using cross-ratio to define distance.

equation (4.6) that

$$d(iy_1, iy_2) = \log |(iy_1, iy_2; 0, \infty)|.$$

Since every pair of points can be mapped by an isometry to a pair of points on the imaginary axis, invariance of the cross-ratio implies that

$$d(z_1, z_2) = \log \left| \frac{z_1 - w_1}{z_2 - w_1} \right| - \log \left| \frac{z_1 - w_2}{z_2 - w_2} \right|. \tag{4.8}$$

Exercise 4.12. Prove the following formula for the hyperbolic distance between two points z_1 and z_2 in the upper half-plane:

$$d(z_1, z_2) = \log \frac{|z_1 - \bar{z}_2| + |z_1 - z_2|}{|z_1 - \bar{z}_2| - |z_1 - z_2|}.$$

Lecture 27

a. Detailed discussion of geodesics and isometries in the upper half-plane model. One of our key examples throughout this course has been the flat torus, a surface whose name indicates that it is a surface of constant zero curvature, and which has Euler characteristic zero. We have also seen that the sphere, which has positive Euler characteristic, has constant positive curvature.

From our considerations of the hyperbolic plane, which we will continue in this lecture, we will eventually see that a sphere with m handles, $m \geq 2$, which is a surface of negative Euler characteristic, can be endowed with a metric under which it has constant negative curvature.

These examples suggest that there might be some connection between curvature and Euler characteristic; this is the content of the *Gauss-Bonnet Theorem*, which we will come to later on.

For the time being, we postpone further discussion of curvature until we have examined the hyperbolic plane in greater detail. Recall the Poincaré upper half-plane model:

$$H^2 = \{\, (x, y) \in \mathbb{R}^2 \mid y > 0\} = \{\, z \in \mathbb{C} \mid \operatorname{Im} z > 0 \,\}.$$

The hyperbolic metric on the upper half-plane is given by a conformal change of the Euclidean metric:

$$ds^2 = \frac{dx^2 + dy^2}{y^2}.$$

Visually, this means that to obtain hyperbolic distances from Euclidean ones, we stretch the plane near the real axis, where $y = \operatorname{Im} z$ is small, and shrink it far away from the real axis, where y is large. Thus if we take a vertical strip which has constant Euclidean width, such as

$$X = \{\, (x, y) \in H^2 \mid 0 \le x \le 1 \,\},$$

and glue the left and right edges together, we will obtain a sort of funnel, or trumpet, in the hyperbolic metric, which is very narrow at large values of y, and flares out hyperbolically as y goes to 0. Part of this construction (at the narrow end of the funnel) is realised on the surface of the pseudosphere mentioned in Lecture 26(b.3).

Now we will present a detailed derivation of the distance formula (4.8), beginning with the special case (4.6). So we take two points $z_1 = x + iy_1$ and $z_2 = x + iy_2$ which lie on the same vertical half-line, where $y_1 < y_2$. The curve $\gamma\colon [y_1, y_2] \to H^2$ given by

$$\gamma(t) = x + it$$

has length given by

$$\ell(\gamma) = \int_{y_1}^{y_2} \|\gamma'(t)\|\, dt = \int_{y_1}^{y_2} \frac{1}{t}\, dt = \log y_2 - \log y_1.$$

To see that this is in fact minimal, let $\eta\colon [a, b] \to H^2$ be any smooth curve with $\eta(a) = z_1$, $\eta(b) = z_2$, and write $\eta(t) = x(t) + iy(t)$. Then we have

$$\begin{aligned}\ell(\eta) &= \int_a^b \|\eta'(t)\|\,dt = \int_a^b \frac{\sqrt{x'(t)^2+y'(t)^2}}{y(t)}\,dt \\ &\geq \int_a^b \frac{|y'(t)|}{y(t)}\,dt \geq \int_a^b \frac{d}{dt}\log y(t)\,dt = \log y_2 - \log y_1\end{aligned}$$

with equality iff $x'(t) \equiv 0$ and $y'(t) > 0$. Hence vertical lines are geodesics in H^2.

To determine what the rest of the geodesics in H^2 look like, we will examine the images of vertical lines under isometries. First we give another proof (independently of any decomposition of the transformation into a product of simple ones) that *fractional linear transformations*

$$f\colon z \mapsto \frac{az+b}{cz+d},$$

where $a, b, c, d \in \mathbb{R}$ are such that $ad - bc = 1$, are indeed isometries of the hyperbolic plane. If we attempt to write f in terms of the real and imaginary parts of z, we quickly discover why the use of complex numbers to represent H^2 is so convenient:

$$\begin{aligned}f(x,y) &= f(x+iy) \\ &= \frac{ax+iay+b}{cx+icy+d} \\ &= \frac{ax+b+iay}{cx+d+icy}\cdot\frac{ax+b-iay}{cx+d-icy} \\ &= \frac{(ax+b)(cx+d)+acy^2+i(acxy+ady-acxy-bcy)}{(cx+d)^2+(cy)^2} \\ &= F(x,y) + \frac{iy}{(cx+d)^2+c^2y^2}.\end{aligned}$$

The exact form of the real part $F(x,y)$ is unimportant for our purposes here, since ds is independent of the value of x. It is important, however, to note that the denominator of the imaginary part is given by

$$(cx+d)^2 + c^2y^2 = |cx+d+icy|^2 = |cz+d|^2,$$

and hence if we write $f(x,y) = (\tilde{x}, \tilde{y})$, we have

$$\tilde{y} = \frac{y}{|cz+d|^2}.$$

How are we to show that this is an isometry? One conceivable plan of attack would be to compute the distance formula on H^2 and then show directly that the distance between $f(z_1)$ and $f(z_2)$ is the same as the distance between z_1 and z_2 for any two points $z_1, z_2 \in H^2$. This, however, requires computation of an explicit distance formula, which is in fact our ultimate goal. To avoid a vicious circle, we take the infinitesimal point of view and examine the action of f on tangent vectors. That is, we recall that given a map $f\colon \mathbb{R}^2 \to \mathbb{R}^2$, the Jacobian derivative Df is a linear map from $\mathbb{R}^2$ to $\mathbb{R}^2$ which takes tangent vectors at (x, y) to tangent vectors at $f(x, y)$. If f is in addition a holomorphic map from $\mathbb{C}$ to (shining) $\mathbb{C}$, then this map $Df_{(x,y)}$ will act on $\mathbb{R}^2$ ($\mathbb{C}$) as multiplication by a complex number $f'(z)$. Geometrically, this means that Df is the composition of a homothety (by the modulus of $f'(z)$) and a rotation (by the argument of $f'(z)$).

In the case of a fractional linear transformation given by the formula above, we have

$$\begin{aligned} f'(z) &= \frac{d}{dz}\frac{az+b}{cz+d} = \frac{a(cz+d)-c(az+b)}{(cz+d)^2} \\ &= \frac{ad-bc}{(cz+d)^2} = \frac{1}{(cz+d)^2} \end{aligned}$$

and hence, writing $f(x, y) = (\tilde{x}, \tilde{y})$ and recalling the form of $\tilde{y}$, we have

$$|f'(z)| = \frac{\tilde{y}}{y}.$$

Now f takes the point $z = x+iy \in H^2$ to the point $\tilde{z} = \tilde{x}+i\tilde{y}$, and Df_z takes the tangent vector $v \in T_zH^2$ to the vector $Df_z v \in T_{\tilde{z}}H^2$. Because Df_z is homothety composed with rotation, we have, *in the Euclidean norm on* $\mathbb{R}^2$,

$$\|Df(v)\|_{\text{Euc}} = |f'(z)| \cdot \|v\|_{\text{Euc}}.$$

The hyperbolic norm is just the Euclidean norm divided by the y-coordinate, and so we have

$$\|Df(v)\|_{\tilde{z}} = \frac{\|Df(v)\|_{\text{Euc}}}{\tilde{y}} = \frac{|f'(z)|}{\tilde{y}}\|v\|_{\text{Euc}} = \frac{1}{y}\|v\|_{\text{Euc}} = \|v\|_z.$$

This is the infinitesimal condition for f to be an isometry; with this fact in hand, it quickly follows that f preserves the length of any curve γ, and hence preserves geodesics and the distances between points.

b. The cross-ratio. The knowledge that fractional linear transformations are isometries allows us to find the rest of the geodesics in H^2; these are simply the images under isometries of the vertical half-lines discussed earlier. This in turn will give us the tools we need to compute the explicit formula (4.8) for the distance between two points $z_1, z_2 \in H^2$. To this end, we make the following definition (the following discussion is valid in $\mathbb{C}$ generally, not just H^2):

Definition 4.6. Given $z_1, z_2, z_3, z_4 \in \mathbb{C}$, the *cross-ratio* is the complex number

$$(z_1, z_2; z_3, z_4) = \frac{z_1 - z_3}{z_2 - z_3} \div \frac{z_1 - z_4}{z_2 - z_4}.$$

This generalises (4.7), where the last two points were taken on the real line. It turns out that *any* fractional linear transformation, whether or not the coefficients lie in $\mathbb{R}$, preserves the cross-ratio.

Lemma 4.7. *Given any $a, b, c, d \in \mathbb{C}$ with $ad - bc \neq 0$ and any $z_1, z_2, z_3, z_4 \in \mathbb{C}$, define w_1, w_2, w_3, w_4 by*

$$w_j = \frac{az_j + b}{cz_j + d}$$

for $1 \leq j \leq 4$. Then

$$(w_1, w_2; w_3, w_4) = (z_1, z_2; z_3, z_4).$$

Proof. Straightforward computation; substitute the expressions for w_i into the cross-ratio formula, clear denominators, and notice that constant and quadratic terms (in z_i) cancel out additively, while linear coefficients cancel multiplicatively, leaving the cross-ratio of the z_i as the result. □

As a simpler example of this general idea, one can notice that if we consider triples (z_1, z_2, z_3) of complex numbers, then the *simple ratio*

$$\frac{z_1 - z_3}{z_2 - z_3}$$

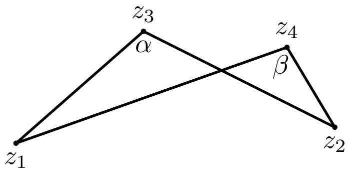


Figure 4.10. Interpreting the cross-ratio of four numbers.

is preserved by the linear map $z \mapsto az+b$ for any $a, b \in \mathbb{C}$. Indeed, the complex number $z_1 - z_3$ is represented by the vector pointing from z_3 to z_1, and similarly $z_2 - z_3$ is the vector from z_3 to z_2. Recall that the argument of the ratio of two complex numbers is given by the difference in their arguments; hence the argument of the above ratio is the angle made by the points z_1, z_3, z_2 taken in that order.

Furthermore, linear transformations are characterised by the fact that they preserve the simple ratio; this can easily be seen by fixing two points z_1 and z_2, and then expressing $f(z)$ in terms of z from the equality

$$\frac{z_1 - z}{z_2 - z} = \frac{f(z_1) - f(z)}{f(z_2) - f(z)}.$$

Later we will use the same argument to show that fractional linear transformations are characterised by the property of preserving the cross-ratio (Lemma 4.9).

As with the simple ratio, the cross-ratio can be interpreted geometrically. Let α be the angle made by z_1, z_3, z_2 in that order, and β the angle made by z_1, z_4, z_2, as in Figure 4.10. Then the argument of the cross-ratio is just $\alpha - \beta$. In particular, if $\alpha = \beta$, then the cross-ratio is a positive real number; this happens iff the points z_1, z_2, z_3, z_4 all lie on a circle with z_1 adjacent to z_2 and z_3 adjacent to z_4 as in the picture, or if they are collinear.

If $\alpha - \beta = \pi$, the four points still lie on a circle (or possibly a line), but now the order is changed; z_4 will have moved to a position between z_1 and z_2 on the circumference. The upshot of all of this is that the cross-ratio is a real number iff the four points lie on a circle or a line. Because fractional linear transformations preserve cross-ratios, we have proved the following theorem.

Theorem 4.8. *If γ is a line or a circle in $\mathbb{C}$ and $f\colon \mathbb{C} \to \mathbb{C}$ is a fractional linear transformation, then $f(\gamma)$ is also a line or a circle.*

There are other ways of proving this theorem, but they involve either a fair amount of algebra using the characterisations of lines and circles in terms of z and $\bar{z}$, or a synthetic argument which requires the decomposition of fractional linear transformations into maps of particular types.

It is worth noting that if we think of all this as happening on the Riemann sphere rather than on the complex plane, we can dispense with this business of 'lines and circles'. Recall that the Riemann sphere is the complex plane $\mathbb{C}$ together with a point at infinity; circles in the plane are circles on the sphere which do not pass through the point at infinity, and lines in the plane are circles on the sphere which do pass through the point at infinity. Fractional linear transformations also assume a nicer form, once we make the definitions

$$f(\infty) = \frac{a}{c}, \quad f\left(-\frac{d}{c}\right) = \infty.$$

Returning to the hyperbolic plane, we now make use of the fact that fractional linear transformations preserve angles (because they are conformal) and cross-ratios (as we saw above). In particular, the image of a vertical line under such a transformation f is either a vertical line, which we already know to be a geodesic, or a circle; because angles are preserved and because f preserves the real line (by virtue of having coefficients in $\mathbb{R}$), this circle must intersect $\mathbb{R}$ perpendicularly, and hence must have its centre on the real line.

This allows us to conclude our detailed derivation of the distance formula (4.8) by establishing that semicircles whose centre lies in $\mathbb{R}$ are also geodesics. Let f be a fractional linear transformation which maps the vertical half-line $\{ z \in H^2 \mid \operatorname{Re} z = 0 \}$ to the semicircle $\{ z \in H^2 \mid |z - a_0| = r \}$. Given two points z_1, z_2 lying on the semicircle, we have $z_1 = f(iy_1)$ and $z_2 = f(iy_2)$; hence $d(z_1, z_2) = d(iy_1, iy_2)$ since f is an isometry.

Furthermore, supposing without loss of generality that $y_1 > y_2$, we see that $f(0)$ and $f(\infty)$ are the two points where the circle intersects $\mathbb{R}$. Denote these by w_1 and w_2, respectively; then w_1 lies closer to z_1, and w_2 lies closer to z_2. Since f preserves cross-ratios, we have

$$\begin{aligned}(z_1, z_2; w_1, w_2) &= (iy_1, iy_2; 0, \infty)\\ &= \frac{iy_1 - 0}{iy_2 - 0} \div \frac{iy_1 - \infty}{iy_2 - \infty} = \frac{y_1}{y_2}\end{aligned}$$

and recalling that $d(iy_1, iy_2) = \log y_1 - \log y_2 = \log(y_1/y_2)$, the fact that f is an isometry implies

$$d(z_1, z_2) = \log(z_1, z_2; w_1, w_2).$$

If we remove the assumption that $y_1 > y_2$, we must take the absolute value of this quantity.

In order to show that this analysis is complete, we must show that there are no other geodesics in H^2 other than those described here. This will follow once we know that any two points $z_1, z_2 \in H^2$ either lie on a vertical half-line or on a semicircle whose centre is in $\mathbb{R}$, and that any such half-line or semicircle can be obtained as the image of the imaginary axis under a fractional linear transformation.

The former assertion is straightforward, as described in the previous lecture (Figure 4.9). To see the latter, note that horizontal translation $z \mapsto z + t$ and homothety $z \mapsto \lambda z$ are both fractional linear transformations, and that using these, we can obtain any vertical half-line from any other, and any semicircle centred in $\mathbb{R}$ from any other. Thus we need only obtain a circle from a line, and this is accomplished by considering the image of the vertical line $\operatorname{Re} z = 1$ under the fractional linear transformation $z \mapsto -1/z$, which will be a circle of radius $1/2$ centred at $-1/2$.

Exercise 4.13. Prove that fractional linear transformations of the form

$$z \mapsto \frac{az + \bar{c}}{cz + \bar{a}},$$

where $a, c \in \mathbb{C}$ satisfy $a\bar{a} - c\bar{c} = 1$, represent isometries of the hyperbolic plane in the disc model.

c. Circles in the hyperbolic plane. Theorem 4.8 raises a natural question: what is the intrinsic meaning of the curves in the hyperbolic plane which are represented in the models by lines, rays, intervals, circles, or arcs of circles?

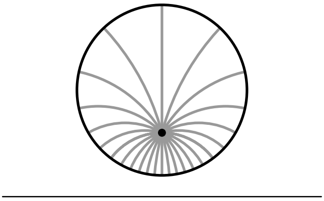


Figure 4.11. Hyperbolic centre and radii of a circle in H^2.

As we have seen, some of these are geodesics; in fact, a necessary and sufficient condition for that is that the curve (or its extension) cross the real line (in the half-plane model) or the unit disc (in the disc model) at a right angle. But what are the rest?

We have seen one example: in the disc model, the circles centred at the origin represent circles in the hyperbolic metric. Hence any image of such a circle under a fractional linear hyperbolic isometry, which must be a (Euclidean) circle by Theorem 4.8, also represents a hyperbolic circle. Now using the inverse of the transformation (4.5), these circles are mapped to circles in the upper half-plane, which thus also represent hyperbolic circles. In the upper half-plane, any circle can be mapped into any other circle by a linear transformation with real coefficients, and so we conclude that any circle inside the upper half-plane represents a hyperbolic circle. Applying (4.5), we reach the same conclusion for the disc model.

Finally, we need to show that *any* hyperbolic circle is represented this way. Let γ be a hyperbolic circle in the disc model and p be its centre in the hyperbolic metric. There is a hyperbolic isometry, represented by a fractional linear transformation, which maps p into the origin, and γ into a hyperbolic circle centred at the origin, which is represented by a Euclidean circle. Hence γ is a Euclidean circle as well. This carries over to the upper half-plane model, and so we have proved the following fact:

Proposition 4.13. *In both the upper half-plane and the disc models, circles in hyperbolic metric are represented by Euclidean circles; conversely, every Euclidean circle which lies inside the half-plane or the disc represents a hyperbolic circle.*

What about Euclidean circles which do not lie inside the upper half-plane or the disc, but which intersect the ideal boundary? They are not closed curves in H^2, and so cannot be circles; if they do not meet the ideal boundary at a right angle, they are not geodesics. So what are they? We will address this question in Lecture 29, where we make a more detailed study of the isometries of the hyperbolic plane.

Exercise 4.14. Calculate the hyperbolic radius and the hyperbolic centre of the circle in H^2 given by the equation

$$\|z - 2i - 1\|^2_{\text{Euc}} = 9/4.$$

Lecture 28

a. Three approaches to hyperbolic geometry. As we continue to plan our assault on the mountain of hyperbolic geometry, there are three main approaches that we might take: the synthetic, the analytic, and the algebraic.

a.1. The first of these, the *synthetic* approach, proceeds along the same lines as the classical Euclidean geometry which is (or used to be, at any rate) taught as part of any high school education. One approaches the subject axiomatically, formulating several postulates and then deriving theorems from these basic assumptions. From this point of view, the only difference between the standard Euclidean geometry one learns in school and the hyperbolic non-Euclidean geometry we are investigating here is the failure of Euclid's fifth postulate, the parallel postulate, in our present case.

This postulate can be stated in many forms; the most common formulation is the statement that given a line and a point not on that line, there exists exactly one line through the point which never intersects the original line. One could also state that the measures of the angles of any triangle sum to π radians, or that there exist triangles with equal angles which are not isometric, and there are many other equivalent formulations.

In hyperbolic geometry, this postulate is no longer valid; however, any theorem of Euclidean geometry which does not rely on this postulate still holds. The common body of such results is known as *absolute* or *neutral geometry*, and the historical approach from the time of Euclid until the work of Lobachevsky and Bolyai in the nineteenth century was to attempt to prove that the parallel postulate in fact follows from the others. The synthetic approach, then, uses the result that if the parallel postulate can be added to the axioms of absolute geometry without fear of contradiction, then its negation can as well, and proceeds axiomatically assuming that negation.

a.2. The second approach is the *analytic* one, which we have made some use of thus far; one derives and then makes use of formulae for lengths, angles, and areas. This approach has the advantage of being the most general of the three, in that it can be applied to any surface, whereas both the synthetic and the algebraic approaches have limited applicability beyond the highly symmetric examples of the Euclidean and hyperbolic (and, to a certain extent, elliptic) planes. Hyperbolic trigonometry can be associated with this approach too.

a.3. For the time being, however, we will make use of the symmetry possessed by the hyperbolic plane, which allows us to take the third option, the *algebraic* approach. In this approach, we study the isometry group of H^2 and use properties of isometries to understand various aspects of the surface itself, a process in which linear algebra becomes an powerful and invaluable tool.

b. Characterisation of isometries. First, then, we must obtain a complete description of the isometries of H^2. We saw in the previous lecture that fractional linear transformations of the form

$$z \mapsto \frac{az+b}{cz+d}$$

are orientation preserving isometries of H^2 in the upper half-plane model for any $a, b, c, d \in \mathbb{R}$ with $ad - bc = 1$. But what about orientation reversing isometries? Since the composition of two orientation reversing isometries is an orientation preserving isometry, once we have understood the orientation preserving isometries it will suffice to exhibit a single orientation reversing isometry. Such an isometry is given by the map

$$z \mapsto -\bar{z}$$

which is reflection in the imaginary axis. By composing this with fractional linear transformations of the above form, we obtain a family of orientation reversing isometries of the form

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

where again, $a, b, c, d \in \mathbb{R}$ are such that $ad - bc = 1$. By changing the sign on a and c, we can write each of these isometries as

$$z \mapsto \frac{a\bar{z}+b}{c\bar{z}+d} \tag{4.9}$$

where $ad - bc = -1$.

Now we claim that these are in fact all of the isometries of H^2. The following argument for the hyperbolic plane can in fact be made to work in much greater generality, and says that for any surface the isometry group is not 'too big'.

We will show that any isometry I is uniquely determined by the images of three points which do not lie on the same geodesic (recall Figure 1.20). Given that $I(A) = \tilde{A}$ and $I(B) = \tilde{B}$, let γ be the unique geodesic connecting A and B, and $\tilde{\gamma}$ the unique geodesic connecting $\tilde{A}$ and $\tilde{B}$. Then because $I(\gamma)$ is also a geodesic connecting $\tilde{A}$ and $\tilde{B}$, we must have $I(x) \in \tilde{\gamma}$ for every $x \in \gamma$. Furthermore, the distance along γ from x to A must be the same as the distance along $\tilde{\gamma}$ from $I(x)$ to $\tilde{A}$, and similarly for B. This requirement uniquely determines the point $I(x)$.

This demonstrates that the action of I on two points of a geodesic is sufficient to determine it uniquely on the entire geodesic. It follows that I is uniquely determined on the three geodesics connecting A, B, and C by its action on those three points; thus we know the action of I on a geodesic triangle. But now given any point $y \in S$, we may draw a geodesic through y which passes through two points of that triangle; it follows that the action of I on those two points, which we know, determines $I(y)$.

Thus we have established uniqueness, but not existence, of an isometry taking A, B, and C to $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$. Indeed, given two sets of three points, it is not in general true that some isometry carries one set to the other. As a minimal requirement, we see that the pairwise distances between the points must be the same; we must have $d(A, B) = d(\tilde{A}, \tilde{B})$ and so on. If our surface is symmetric enough, this condition will be sufficient, as is the case for the Euclidean plane and the round sphere; we will soon see that this is also the case for H^2. First, we prove a fundamental lemma concerning fractional linear transformations in general.

Lemma 4.9. *Let* (z_1, z_2, z_3) *and* (w_1, w_2, w_3) *be two triples of distinct points in the extended complex plane* $\mathbb{C} \cup \{\infty\}$ *(the Riemann sphere). Then there exist unique coefficients* $a, b, c, d \in \mathbb{C}$ *such that the fractional linear transformation*

$$f \colon z \mapsto \frac{az + b}{cz + d}$$

satisfies $f(z_j) = w_j$ *for* $j = 1, 2, 3$*. Furthermore, any map from* $\mathbb{C} \cup \{\infty\}$ *to itself which preserves the cross-ratio is a fractional linear transformation.*

Proof. Recall that fractional linear transformations preserve cross-ratios, and hence if for some $z \in \mathbb{C}$ the f we are looking for has $f(z) = w$, we must have

$$(z_1, z_2; z_3, z) = (w_1, w_2; w_3, w). \tag{4.10}$$

Using the expression for the cross-ratio, we have

$$\frac{(z_1 - z_3)(z_2 - z)}{(z_2 - z_3)(z_1 - z)} = \frac{(w_1 - w_3)(w_2 - w)}{(w_2 - w_3)(w_1 - w)},$$

and solving this equation for w in terms of z will give the desired fractional linear transformation:

$$w = \frac{w_1(z_1 - z_3)(w_2 - w_3)(z_2 - z) - w_2(z_2 - z_3)(w_1 - w_3)(z_1 - z)}{(z_1 - z_3)(w_2 - w_3)(z_2 - z) - (z_2 - z_3)(w_1 - w_3)(z_1 - z)}. \tag{4.11}$$

Since (4.10) implies (4.11) we also get the second statement. □

Proposition 4.14. *Given points* $z_1, z_2, z_3, w_1, w_2, w_3 \in H^2$ *satisfying* $d(z_j, z_k) = d(w_j, w_k)$ *for each pair of indices* (j, k)*, there exists a unique isometry taking* z_k *to* w_k*. If the geodesic triangles* z_1, z_2, z_3 *and* w_1, w_2, w_3 *have the same orientation, this isometry is orientation preserving and is represented by a fractional linear transformation; otherwise it is orientation reversing and has the form* (4.9)*.*

Remark. The first part of this proposition states that given two triangles in H^2 whose corresponding sides are of equal length, there

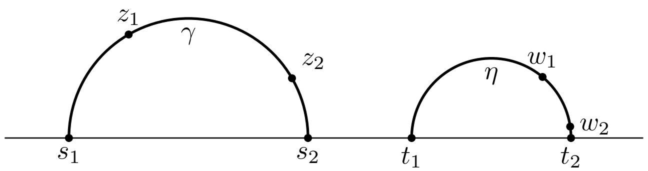


Figure 4.12. The images of two points determine a unique fractional linear transformation.

exists an isometry of H^2 taking one triangle to the other. This statement is true in Euclidean geometry as well, and in fact holds as a result in absolute geometry. As such, it could be proven in a purely synthetic manner; while such an approach does in fact succeed, we will take another path and use our knowledge of fractional linear transformations.

Notice that, while Lemma 4.9 gives us a fractional linear transformation which is a candidate to be an isometry, this candidate is the desired isometry only if the orientations of the triangles z_1, z_2, z_3 and w_1, w_2, w_3 coincide.

We first prove that the group of fractional linear transformations with real coefficients acts transitively on *pairs* of points (z_1, z_2), where the distance $d(z_1, z_2)$ is fixed. We then use the fact that a third point z_3 has only two possible images under an isometry, and that the choice of one of these as w_3 determines whether the isometry preserves or reverses orientation.

Proposition 4.15. *Given points $z_1, z_2, w_1, w_2 \in H^2$ with $d(z_1, z_2) = d(w_1, w_2)$, there exists a unique fractional linear transformation f satisfying $f(z_j) = w_j$ for $j = 1, 2$. This transformation f has real coefficients and hence is an isometry of H^2.*

Proof. Let γ be the geodesic connecting z_1 and z_2, and η the geodesic connecting w_1 and w_2. Let s_1 and s_2 be the two points where γ intersects $\mathbb{R}$, with s_1 nearer to z_1 and s_2 nearer to z_2, and define t_1 and t_2 similarly on η, as shown in Figure 4.12.

By Lemma 4.9, there exists a unique fractional linear transformation f with complex coefficients such that $f(s_1) = t_1$, $f(z_1) = w_1$, and $f(z_2) = w_2$. In order to complete the proof, we must show that f in fact preserves the real line, and hence has real coefficients.

Recalling our distance formula for H^2 in terms of the cross-ratio, the condition that $d(z_1, z_2) = d(w_1, w_2)$ can be rewritten as

$$(z_1, z_2; s_1, s_2) = (w_1, w_2; t_1, t_2).$$

From the proof of Lemma 4.9, this was exactly the formula that we solved for t_2 to find $f(s_2)$; it follows that $f(s_2) = t_2$. Since f is a conformal map which takes lines and circles to lines and circles, and since $\mathbb{R}$ intersects γ orthogonally at s_1 and s_2, the image of $\mathbb{R}$ is a line or circle which intersects η orthogonally at t_1 and t_2, and hence is in fact $\mathbb{R}$.

Now $f(\mathbb{R}) = \mathbb{R}$, so f has real coefficients and is in fact an isometry of H^2. □

In order to obtain Proposition 4.14, we need only extend the result of this proposition to take into account the position of the third point, which determines whether the isometry preserves or reverses orientation. To this end, note that the condition $d(w_1, w_3) = d(z_1, z_3)$ implies that w_3 lies on a circle of radius $d(z_1, z_3)$ centred at w_1; similarly, it also lies on a circle of radius $d(z_2, z_3)$ centred at w_3.

Assuming z_1, z_2, z_3 do not all lie on the same geodesic, there are exactly two points which lie on both circles, each an equal distance from the geodesic connecting z_1 and z_2. One of these will necessarily be the image of z_3 under the fractional linear transformation f found above; the other one is $(r \circ f)(z_3)$ where r denotes reflection in the geodesic η.

To better describe r, pick any point $z \in H^2$ and consider the geodesic ζ which passes through z and meets η orthogonally. Denote by $d(z, \eta)$ the distance from z to the point of intersection; then the reflection $r(z)$ is the point on ζ a distance $d(z, \eta)$ beyond this point. Alternatively, we may recall that the map $R\colon z \mapsto -\bar{z}$ is reflection in the imaginary axis, which is an orientation reversing isometry. There exists a unique fractional linear transformation g taking η to the imaginary axis; then r is simply the conjugation $g^{-1} \circ R \circ g$.

Exercise 4.15. Prove that the group of orientation preserving isometries of H^2 in the unit disc model is the group of all fractional linear transformations of the form

$$z \mapsto \frac{az + \bar{c}}{cz + \bar{a}}$$

where $a, c \in \mathbb{C}$ satisfy $a\bar{a} - c\bar{c} = 1$.

Lecture 29

a. Classification of isometries. Now we turn to the task of classifying these isometries and understanding what they look like geometrically.

a.1. *Fixed points in the extended plane.* For the time being we restrict ourselves to orientation preserving isometries. We begin by considering the fractional linear transformation f as a map on all of $\mathbb{C}$ (or, more precisely, on the Riemann sphere $\mathbb{C} \cup \{\infty\}$) and look for fixed points, given by

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

Clearing the denominator and simplifying gives the quadratic equation

$$cz^2 + (d - a)z - b = 0$$

whose roots are

$$\begin{aligned} z &= \frac{1}{2c}\left(a - d \pm \sqrt{(a-d)^2 + 4bc}\right) \\ &= \frac{1}{2c}\left(a - d \pm \sqrt{(a+d)^2 - 4(ad - bc)}\right) \\ &= \frac{1}{2c}\left(a - d \pm \sqrt{(a+d)^2 - 4}\right). \end{aligned}$$

Note that the quantity $a + d$ is just the trace of the matrix of coefficients $X = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, which we already know has unit determinant. Let λ and μ be the eigenvalues of X; then $\lambda\mu = \det X = 1$, so $\mu = 1/\lambda$, and we have

$$a + d = \operatorname{Tr} X = \lambda + \mu = \lambda + \frac{1}{\lambda}.$$

There are three possibilities to consider regarding the nature of the fixed point or points $z = f(z)$:

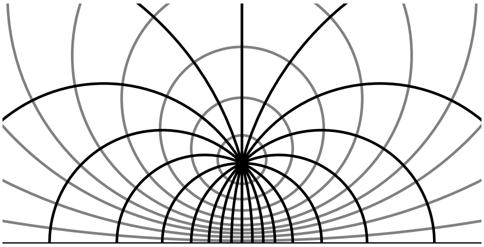


Figure 4.13. Geodesics passing through i and hyperbolic circles centred at i.

- **(E):** $|a+d| < 2$, corresponding to $\lambda = e^{i\alpha}$ for some $\alpha \in \mathbb{R}$. In this case there are two fixed points z and $\bar{z}$, with $\operatorname{Im} z > 0$ and hence $z \in H^2$.
- **(P):** $|a+d| = 2$, corresponding to $\lambda = 1$ (since X and $-X$ give the same transformation). In this case there is exactly one fixed point $z \in \mathbb{R}$.
- **(H):** $|a+d| > 2$, corresponding to $\mu < 1 < \lambda$. In this case, there are two fixed points $z_1, z_2 \in \mathbb{R}$.

a.2. *Elliptic isometries.* Let us examine each of these in turn, beginning with **(E)**, where f fixes a unique point $z \in H^2$. Consider a geodesic γ passing through z. Then $f(\gamma)$ will also be a geodesic passing through z; let α be the angle it makes with γ at z. Then because f preserves angles, it must take any geodesic η passing through z to the unique geodesic which passes through z and makes an angle of α with η. Thus f is analogous to what we term rotation in the Euclidean context; since f preserves lengths, we can determine its action on any point in H^2 based solely on knowledge of the angle of rotation α. As our choice of notation suggests, this angle turns out to be equal to the argument of the eigenvalue λ.

As an example of a map of this form, consider

$$f\colon z \mapsto \frac{(\cos\alpha)z + \sin\alpha}{(-\sin\alpha)z + \cos\alpha}$$

which is rotation by α around the point i; the geodesics passing through i are the dark curves in Figure 4.13. The lighter curves are the circles whose (hyperbolic) centre lies at i; each of these curves intersects all of the geodesics orthogonally, and is left invariant by f.

This map does not seem terribly symmetric when viewed as a transformation of the upper half-plane; however, if we look at f in the unit disc model, we see that i is taken to the origin, and f corresponds to the rotation by α around the origin in the usual sense. Thus we associate with a rotation (as well as with the family of all rotations around a given point p) two families of curves:

(1) The *pencil* of all geodesics passing through p; each element of this family maps to another, and rotations around p act transitively on this family.

(2) The family of circles around p which are orthogonal to the geodesics from the first family. Each circle is invariant under rotations, and rotations around p act transitively on each circle.

We will discover similar pictures for the remaining two cases.

a.3. *Parabolic isometries.* Case **(P)** can be considered as a limiting case of the previous situation where the fixed point p goes to infinity. Let $t \in \mathbb{R} \cup \{\infty\}$ be the unique fixed point in the Riemann sphere, which lies on the ideal boundary. As with the family of rotations around p, we can consider the family of all orientation preserving isometries preserving t; notice that as in that case, this family is a *one-parameter group* whose members we will denote by $p_s^{(t)}$, where $s \in \mathbb{R}$. As above, one can see two invariant families of curves:

(1) The pencil of all geodesics passing through t (dark curves in Figure 4.14)—each element of this family maps to another, and the group $\{p_s^{(t)}\}$ acts transitively on this family.

(2) The family of *limit circles*, more commonly called *horocycles* (light curves in Figure 4.14), which are orthogonal to the geodesics from the first family. They are represented by circles tangent to $\mathbb{R}$ at t, or by horizontal lines if $t = \infty$. Each horocycle is invariant under $p_s^{(t)}$, and the group acts transitively on each horocycle.

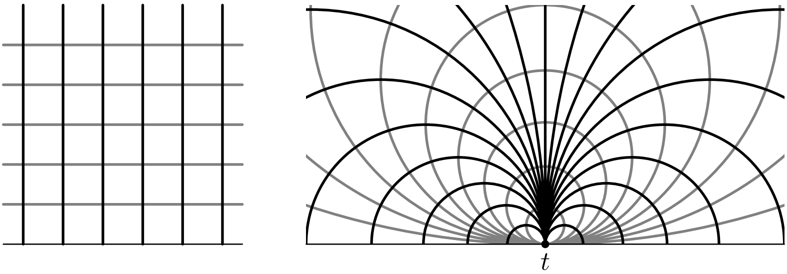


Figure 4.14. Parallel geodesics and horocycles for parabolic isometries.

A useful (but visually somewhat misleading) example is given by the case $t = \infty$ with

$$p_s^{(\infty)} z = z + s.$$

We will see later in the lecture that for the parabolic case, the 'angle' s does not have properties similar to the rotation angle α. In particular, it is not an invariant of the isometry.

Exercise 4.16. Show that given two points $z_1, z_2 \in H^2$, there are exactly two different horocycles which pass through z_1 and z_2.

a.4. *Hyperbolic isometries.* Finally, consider the case **(H)**, in which we have two real fixed points $w_1 < w_2$. Since f takes geodesics to geodesics and fixes w_1 and w_2, the semicircle γ which intersects $\mathbb{R}$ at w_1 and w_2 is mapped to itself by f, and so f acts as translation along this curve by a fixed distance. The geodesic γ is the only geodesic invariant under the transformation; in a sense, it plays the same role as the centre of rotation in the elliptic case, a role for which there is no counterpart in the parabolic case.

To see what the action of f is on the rest of H^2, consider as above two invariant families of curves:

(1) The family of geodesics which intersect γ orthogonally (the dark curves in Figure 4.15). If η is a member of this family, then f will carry η to another member of the family; which member is determined by the effect of f on the point where η intersects γ.

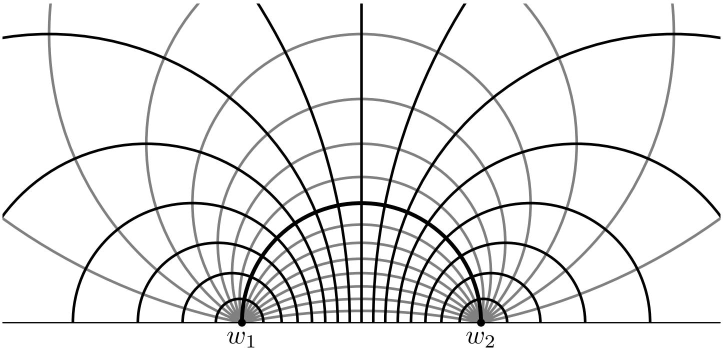


Figure 4.15. Orthogonal geodesics and equidistant curves for the geodesic connecting w_1 and w_2.

(2) The family of curves orthogonal to these geodesics (the light curves in Figure 4.15)—these are the *equidistant curves* (or *hypercircles*). Such a curve ζ is defined as the locus of points which lie a fixed distance from the geodesic γ; in Euclidean geometry this condition defines a geodesic, but this is no longer the case in the hyperbolic plane. Each equidistant curve ζ is carried into itself by the action of f.

A good example of maps f falling into the case **(H)** are the maps which fix 0 and ∞:

$$f\colon z \mapsto \lambda^2 z.$$

In this case the geodesic γ connecting the fixed points is the imaginary axis (the vertical line in Figure 4.16), the geodesics intersecting γ orthogonally are the (Euclidean) circles centred at the origin (the dark curves), and the equidistant curves are the (Euclidean) lines emanating from the origin (the lighter curves).

To be precise, given any geodesic γ in the hyperbolic plane, we define an r*-equidistant curve* as one of the two connected components of the locus of points at a distance r from γ.

Exercise 4.17. For any given $r > 0$, show that there are exactly two different r-equidistant curves (for some geodesics) which pass through two given points in the hyperbolic plane.

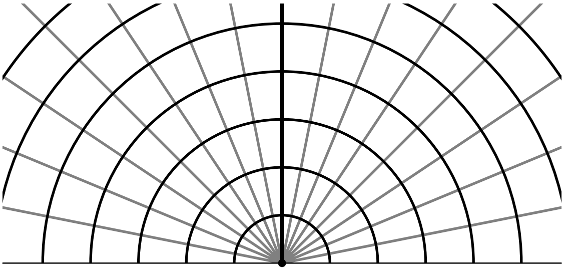


Figure 4.16. Orthogonal geodesics and equidistant curves for the imaginary axis.

Thus we have answered the question about the significance of (Euclidean) circles tangent to the real lines and arcs which intersect it. The former (along with horizontal lines) are horocycles, and the latter (along with rays intersecting the real line) are equidistant curves. Notice that all horocycles are isometric to each other (they can be viewed as circles of infinite radius), whereas for equidistant curves there is an isometry invariant, namely the angle between the curve and the real line. One can see that this angle uniquely determines the distance r between an equidistant curve and its geodesic, and vice versa. The correspondence between the two can be easily calculated in the particular case shown in Figure 4.16.

Exercise 4.18. The arc of the circle $|z-2i|^2 = 8$ in the upper half-plane represents an r-equidistant curve. Find r.

a.5. *Canonical form for elliptic, parabolic, and hyperbolic isometries.* The technique of understanding an isometry by showing that it is conjugate to a particular standard transformation has great utility in our classification of isometries of H^2. Recall that we have a one-to-one correspondence between 2×2 real matrices with unit determinant (up to a choice of sign) and fractional linear transformations preserving $\mathbb{R}$, which are the isometries of H^2 that preserve orientation:

$$PSL(2,\mathbb{R}) = SL(2,\mathbb{R})/\pm \mathrm{Id} \longleftrightarrow \mathrm{Isom}^+(H^2),$$

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \longleftrightarrow f_A \colon z \mapsto \frac{az+b}{cz+d}.$$

Composition of isometries corresponds to matrix multiplication:

$$f_A \circ f_B = f_{AB}.$$

We may easily verify that two maps f_A and f_B corresponding to conjugate matrices are themselves conjugate; that is, if $A = CBC^{-1}$ for some $C \in GL(2, \mathbb{R})$, we may assume without loss of generality that $C \in SL(2, \mathbb{R})$ by scaling C by its determinant. Then we have

$$f_A = f_C \circ f_B \circ f_C^{-1}.$$

It follows that f_A and f_B have the same geometric properties: fixed points, actions on geodesics, etc. Conjugation by f_C has the effect of changing coordinates by an isometry, and so the intrinsic geometric properties of an isometry are conjugacy invariants. For example, in the Euclidean plane, any two rotations by an angle α around different fixed points x and y are conjugated by the translation taking x to y, and any two translations by vectors of equal length are conjugated by any rotation by the angle between those vectors. Thus, in the Euclidean plane, the conjugacy invariants are the angle of rotation and the length of the translation.

In order to classify orientation preserving isometries of H^2, it suffices to understand certain canonical examples. We begin by recalling the following result from linear algebra:

Proposition 4.16. *Every matrix in $SL(2, \mathbb{R})$ is conjugate to one of the following (up to sign):*

(E): *An* elliptic *matrix of the form*

$$\begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}, \qquad \alpha \in \mathbb{R}.$$

(P): *The* parabolic *matrix*

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

(H): *A* hyperbolic *matrix of the form*

$$\begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}, \qquad t \in (0, \infty).$$

The three cases **(E)**, **(P)**, and **(H)** for the matrix A correspond to the three cases discussed above for the fractional linear transformation f_A. Recall that the isometries corresponding to the elliptic case **(E)** have one fixed point in H^2, those corresponding to the parabolic case **(P)** have one fixed point on the *ideal boundary* $\mathbb{R} \cup \{\infty\}$, and those corresponding to the hyperbolic case **(H)** have two fixed points on the ideal boundary.

The only invariants under conjugation are the parameters α (up to a sign) and t, which correspond to the angle of rotation and the distance of translation, respectively. Thus two orientation preserving isometries of H^2 are conjugate *in the full isometry group of* H^2 iff they fall into the same category **(E)**, **(P)**, or **(H)** and have the same value of the invariant α or t, if applicable.

Notice that if we consider only conjugacy by orientation preserving isometries, then α itself (rather than its absolute value) is an invariant in the elliptic case, and the two parabolic matrices $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 & -1 \\ 0 & 1 \end{smallmatrix}\right)$ are not conjugate. In contrast, the conjugacy classes in the hyperbolic case do not change.

Thus we see that there are both similarities and differences between the structure of the group of orientation preserving isometries in the Euclidean and hyperbolic planes. Among the similarities is the possible number of fixed points: one or none. Isometries with one point—rotations—look completely similar, but the set of isometries with no fixed points—which in the Euclidean case is just translations—is more complicated in the hyperbolic case, including both parabolic and hyperbolic isometries.

An important difference in the structure of the isometry groups comes from the following observation. Recall that a subgroup H of a group G is *normal* if for any $h \in H$ and $g \in G$ the conjugate $g^{-1}hg$ remains in H. It is not hard to show that in the group of isometries of the Euclidean plane, translations form a normal subgroup; the situation in the hyperbolic case is rather different.

Exercise 4.19. Prove that the group of isometries of the hyperbolic plane has no non-trivial normal subgroups, i.e. the only normal subgroups are the whole group and the trivial subgroup containing only the identity.

Another example of a difference between the two cases comes when we consider the decomposition of orientation preserving isometries into reflections—this is possible in both the Euclidean and the hyperbolic planes, and any orientation preserving isometry can be had as a product of two reflections. In the Euclidean plane, there are two possibilities—either the lines of reflection intersect, and the product is a rotation, or the lines are parallel, and the product is a translation. In the hyperbolic plane, there are three possibilities for the relationship of the lines (geodesics) of reflection: once again, they may intersect or be parallel (i.e. have a common point at infinity), but now a new option arises; they may also be ultraparallel (see Figure 4.17). We will discuss this in more detail shortly.

Exercise 4.20. Prove that the product of reflections in two geodesics in the hyperbolic plane is elliptic, parabolic, or hyperbolic, respectively, depending on whether the two axes of reflection intersect, are parallel, or are ultraparallel.

a.6. *Orientation reversing isometries.* Using representation (4.9) and following the same strategy, we try to look for fixed points of orientation reversing isometries. The fixed point equation takes the form

$$c|z|^2 + dz - a\bar{z} - b = 0.$$

Separating real and imaginary parts, we get two cases:

(1) $d + a = 0$. In this case, there is a whole geodesic of fixed points, and the transformation is a reflection in this geodesic, which geometrically is represented as inversion (if the geodesic is a semicircle) or the usual sort of reflection (if the geodesic is a vertical ray).

(2) $d + a \neq 0$. In this case, there are two fixed points on the (extended) real line, and the geodesic connecting these points is preserved, so the transformation is a glide reflection, and can be written as the composition of reflection in this geodesic and a hyperbolic isometry with this geodesic as its axis.

Thus the picture for orientation reversing isometries is somewhat more similar to the Euclidean case.

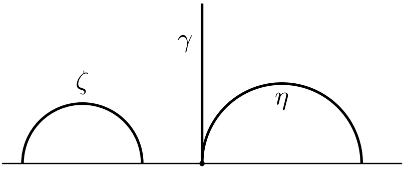


Figure 4.17. Parallels and ultraparallels.

b. Geometric interpretation of isometries. From the synthetic point of view, the fundamental difference between Euclidean and hyperbolic geometry is the failure of the parallel postulate in the latter case. To be more precise, suppose we have a geodesic (line) γ and a point p not lying on γ, and consider the set of all geodesics (lines) through p which do not intersect γ. In the Euclidean case, there is exactly one such geodesic, and we say that it is parallel to γ. In the hyperbolic case, not only are there many such geodesics, but they come in two different classes, as shown in Figure 4.17.

The curves γ, η, and ζ in Figure 4.17 are all geodesics, and neither η nor ζ intersects γ in H^2. However, η and γ both approach the same point on the ideal boundary, while ζ and γ do not exhibit any such asymptotic behaviour. We say that η and γ are *parallel*, while ζ and γ are *ultraparallel*.

Each point x on the ideal boundary corresponds to a family of parallel geodesics which are asymptotic to x, as shown in Figure 4.14. The parallel geodesics asymptotic to ∞ are simply the vertical lines, while the parallel geodesics asymptotic to some point $x \in \mathbb{R}$ form a sort of bouquet of curves.

A recurrent theme in our description of isometries has been the construction of orthogonal families of curves. Given the family of parallel geodesics asymptotic to x, one may consider the family of curves which are orthogonal to these geodesics at every point; such curves are called *horocycles*. As shown in Figure 4.14, the horocycles for the family of geodesics asymptotic to ∞ are horizontal lines, while the horocycles for the family of geodesics asymptotic to $x \in \mathbb{R}$ are Euclidean circles tangent to $\mathbb{R}$ at x.

The reason horocycles are sometimes called limit circles is illustrated by the following construction: fix a point $p \in H^2$ and a geodesic ray γ which starts at p. For each $r > 0$ consider the circle of radius r with centre on γ which passes through p; as $r \to \infty$, these circles converge to the horocycle orthogonal to γ.

What do we mean by this last statement? In what sense do the circles 'converge' to the horocycle? For any fixed value of r, the circle in the construction lies arbitrarily far from some points on the horocycle (those which are 'near' the ideal boundary), and so we certainly cannot expect any sort of uniform convergence in the hyperbolic metric. Rather, convergence in the hyperbolic plane must be understood as convergence of pieces of fixed, albeit arbitrarily large, length—that is, given $R > 0$, the arcs of length R lying on the circles in the above construction with p at their midpoint do in fact converge uniformly to a piece of the horocycle, and R may be taken as large as we wish.

The situation is slightly different in the model, where we do have genuine uniform convergence, as the complete (Euclidean) circles representing (hyperbolic) circles converge to the (Euclidean) circle representing the horocycle.

This distinction between the intrinsic and extrinsic viewpoints raises other questions; for example, the above distinction between parallel and ultraparallel geodesics relies on this particular model of H^2 and the fact that points at infinity are represented by real numbers, and so seems rooted in the extrinsic description of H^2. Can we distinguish between the two sorts of asymptotic behaviour intrinsically, without reference to the ideal boundary?

It turns out that we can; given two ultraparallel geodesics γ and η, the distance from γ to η grows without bound; that is, given any $C \in \mathbb{R}$, there exists a point $z \in \gamma$ such that no point of η is within a distance C of z. On the other hand, given two parallel geodesics, this distance remains bounded, and in fact goes to zero.

To see this, let γ be the imaginary axis; then the equidistant curves are Euclidean lines through the origin, as shown in Figure 4.18, and η is a Euclidean circle which is tangent to γ at the origin. The distance from γ to the equidistant curves is a function of the slope

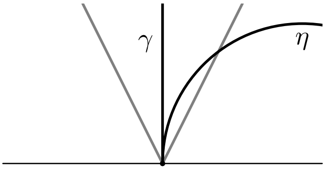


Figure 4.18. Distance between parallel geodesics.

of the lines; steeper slope corresponds to smaller distance, and the points in between the curves are just the points which lie within that distance of γ. But now for any slope of the lines, η will eventually lie between the two equidistant curves, since its slope becomes vertical as it approaches the ideal boundary, and hence the distance between γ and η goes to zero.

One can see the same result by considering a geodesic η which is parallel to γ not at 0, but at ∞; then η is simply a vertical Euclidean line, which obviously lies between the equidistant curves for large enough values of y.

To get an idea of how quickly the distance goes to 0 in Figure 4.18, recall that the hyperbolic distance between two nearby points is roughly the Euclidean distance divided by the height y, and that the Euclidean distance between a point on the circle η in Figure 4.18 and the imaginary axis is roughly y^2 for points near the origin; hence

$$\text{hyperbolic distance} \sim \frac{\text{Euclidean distance}}{y} \sim \frac{y^2}{y} = y \to 0.$$

With this understanding of circles, parallels, ultraparallels, and horocycles, we can now return to the task of giving geometric meaning to the various categories of isometries. In each case, we found two families of curves which intersect each other orthogonally; one of these will comprise geodesics which are carried to each other by the isometry, and the other family will comprise curves which are invariant under the isometry.

In the elliptic case **(E)**, the isometry f is to be thought of as rotation around the unique fixed point p by some angle α; the two families of curves are shown in Figure 4.13. Given $v \in T_pH^2$, denote by γ_v the unique geodesic passing through p with $\gamma'(p) = v$. Then we have

$$f\colon \{\gamma_v\}_{v \in T_pH^2} \to \{\gamma_v\}_{v \in T_pH^2},$$
$$\gamma_v \mapsto \gamma_w,$$

where $w \in T_pH^2$ is the image of v under rotation by α in the tangent space. Taking the family of curves orthogonal to the curves γ_v at each point of H^2, we have the one-parameter family of circles

$$\{\eta_r\}_{r \in (0,\infty)}$$

each of which is left invariant by f.

In the parabolic case **(P)**, the map f is just horizontal translation $z \mapsto z + 1$. Note that by conjugating this map with a homothety, and a reflection if necessary, we obtain horizontal translation by any distance, so any horizontal translation is conjugate to the canonical example. Given $t \in \mathbb{R}$, let γ_t be the vertical line $\operatorname{Re} z = t$; then the geodesics γ_t are all asymptotic to the fixed point ∞ of f, and we have

$$f\colon \{\gamma_t\}_{t \in \mathbb{R}} \to \{\gamma_t\}_{t \in \mathbb{R}},$$
$$\gamma_t \mapsto \gamma_{t+1}.$$

The invariant curves for f are the horocycles, which in this case are horizontal lines η_t, $t \in \mathbb{R}$. For a general parabolic map, the fixed point x may lie on $\mathbb{R}$ rather than at ∞; in this case, the geodesics and horocycles asymptotic to x are as shown in the second image in Figure 4.14. The invariant family of geodesics consists of geodesics parallel to each other.

Finally, in the hyperbolic case **(H)**, the standard form is $f_A(z) = \lambda^2 z$ for $\lambda = e^t$, and the map is simply a homothety from the origin. There is exactly one invariant geodesic, the imaginary axis, and the other invariant curves are the equidistant curves, which in this case are Euclidean lines through the origin. The curves orthogonal to these at each point are the geodesics γ_r ultraparallel to each other, shown in Figure 4.16, where γ_r is the unique geodesic passing through the point ir and intersecting the imaginary axis orthogonally. The map f_A acts on this family by taking γ_r to $\gamma_{\lambda^2 r}$.

In the general hyperbolic case, the two fixed points will lie on the real axis, and the situation is as shown in Figure 4.15. The invariant geodesic η_0 is the semicircle connecting the fixed points, and the equidistant curves are the other circles passing through those two points. The family of orthogonal curves comprises the geodesics intersecting η_0 orthogonally, as shown in the picture.

Lecture 30

a. Area of triangles in different geometries. In our earlier investigations of spherical and elliptic geometry (by the latter we mean the geometry of the projective plane with metric inherited from the sphere), we found that the area of a triangle was proportional to its *angular excess*, the amount by which the sum of its angles exceeds π. For a sphere of radius R, the constant of proportionality was $R^2 = 1/\kappa$, where κ is the curvature of the surface.

In Euclidean geometry, the existence of any such formula was precluded by the presence of similarity transformations, diffeomorphisms of $\mathbb{R}^2$ which expand or shrink the metric by a uniform constant.

In the hyperbolic plane, we find ourselves in a situation reminiscent of the spherical case. We will find that the area of a hyperbolic triangle is proportional to the angular *defect*, the amount by which the sum of its angles falls short of π, and that the constant of proportionality is again given by the reciprocal of the curvature.

We begin with a simple observation, which is that every hyperbolic triangle does in fact have angles whose sum is less than π (otherwise the above claim would imply that some triangles have area ≤ 0).

For that we use the open disc model of the hyperbolic plane, and note that given any triangle, we can use an isometry to position one of its vertices at the origin; thus two of the sides of the triangle will be (Euclidean) lines through the origin, as shown in Figure 4.19. Then because the third side, which is part of a Euclidean circle, is convex in the Euclidean sense, the sum of the angles is less than π.

This implies the remarkable 'fourth criterion of equality of triangles' above and beyond the three criteria which are common to both the Euclidean and hyperbolic planes.

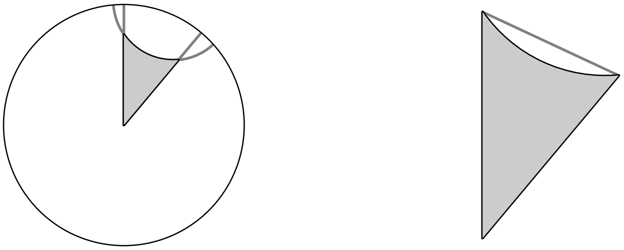


Figure 4.19. A hyperbolic triangle has angles whose sum is less than π.

Proposition 4.17. *Two geodesic triangles with pairwise equal angles are isometric.*

Proof. We will use the disc model. Without loss of generality, we may assume that both triangles have one vertex at the centre O and that two of their sides lie on the same radii. Thus we have triangles AOB and $A'OB'$ where the vertices A and A' lie on one radius, and B and B' on another. The angles OAB and $OA'B'$ are equal and so are the angles OBA and $OB'A'$.

Now there are two possibilities; either the arcs AB and $A'B'$ intersect, or they do not. Assume first that they intersect at some point C. Then the triangle ACA' has two angles which add to π, which is impossible. Hence without loss of generality, we may assume that arc AA' lies inside the triangle OBB'. Then the sum of the angles of the geodesic quadrangle $AA'BB'$ is equal to 2π, which is again an impossibility since it can be split into two geodesic triangles, at least one of which must therefore have angles whose sum is $\geq \pi$. This contradiction implies $A = A'$ and $B = B'$. □

b. Area and angular defect in hyperbolic geometry. Our proof of the area formula is due to Gauss, and follows the exposition in Coxeter's book *Introduction to Geometry* (Sections 16.4 and 16.5). It is essentially a synthetic proof, and as such does not give us a value for the constant of proportionality; to obtain that value, we must turn to analytic methods. We will also deviate slightly from the true synthetic approach by using drawings in the two models of H^2.

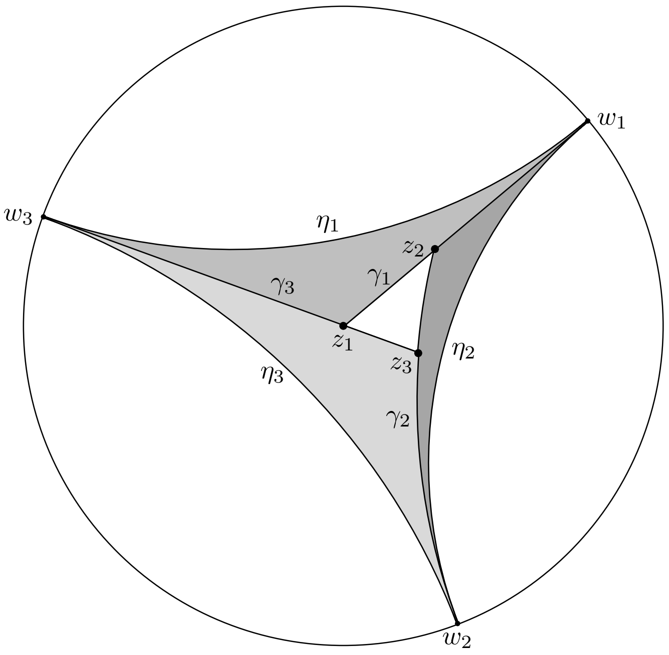


Figure 4.20. Computing the area of a hyperbolic triangle.

As with so many things, non-Euclidean geometry was first discovered and investigated by Gauss, who kept his results secret because he had no proof that his geometry was consistent. Eventually, the introduction of several models (of which the Poincaré half-plane and open disc models were not the earliest) showed that hyperbolic geometry is consistent, contingent upon the consistency of Euclidean geometry; a contradiction in the former would necessarily lead to a contradiction in the latter.

Theorem 4.10. *Given a hyperbolic triangle Δ with angles α, β, and γ, the area A of Δ is given by*

$$A = \frac{1}{-\kappa}(\pi - \alpha - \beta - \gamma), \tag{4.12}$$

where κ is the curvature, whose value is -1 for the standard upper half-plane and open disc models.

Proof. The proof of the analogous formula for the sphere involved partitioning it into segments and using an inclusion-exclusion formula. This relied on the fact that the area of the sphere is finite; in our present case, we must be more careful, as the hyperbolic plane has infinite area. However, we can recover a setting in which a similar proof works by considering *asymptotic triangles*, which turn out to have finite area.

The idea is as follows: let z_1, z_2, z_3 denote the vertices of the triangle, and without loss of generality, take z_1 to be the origin in the open disc model. As shown in Figure 4.20, draw the half-geodesic γ_1 which begins at z_1 and passes through z_2; similarly, draw the half-geodesics γ_2 and γ_3 beginning at z_2 and z_3, and passing through z_3 and z_1, respectively. Let w_j denote the point at infinity approached by γ_j as it nears the boundary of the disc.

Now draw three more geodesics, as shown in the picture; η_1 is to be asymptotic to w_3 and w_1, η_2 is to be asymptotic to w_1 and w_2, and η_3 is to be asymptotic to w_2 and w_3. Then the region T_0 bounded by η_1, η_2, and η_3 is a *triply asymptotic triangle*. If we write T_j for the *doubly asymptotic triangle* whose vertices are z_j, w_j, and w_{j-1}, we can decompose T_0 as the disjoint union

$$T_0 = T_1 \cup T_2 \cup T_3 \cup \Delta$$

and so the area $A(\Delta)$ may be found by computing the areas of the regions T_j, provided they are finite.

Since these regions are not bounded, it is not at first obvious why they should have finite area. We begin by making two observations concerning triply asymptotic triangles.

First, all triply asymptotic triangles are isometric. That is, given $w_1, w_2, w_3 \in \partial D^2$ and $\tilde{w}_1, \tilde{w}_2, \tilde{w}_3 \in \partial D^2$ with the same orientation, Lemma 4.9 guarantees the existence of a unique fractional linear transformation f taking w_j to $\tilde{w}_j$, which must then preserve ∂D^2 and map the interior to the interior, and hence is an isometry of H^2.

Secondly, a triply asymptotic triangle will have finite area iff each of its 'arms' does, where by an 'arm' we mean the section of the triangle which approaches infinity. How do we compute the area of such an arm? A prototypical example is the singly asymptotic triangle

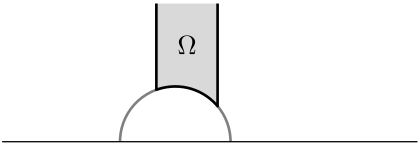


Figure 4.21. A singly asymptotic triangle.

shown in Figure 4.21, where we use the half-plane model and choose ∞ as the point on the ideal boundary, so two of the geodesics are vertical lines. The infinitesimal area element at each point is given by $\frac{1}{y^2}\,dx\,dy$ where dx and dy are Euclidean displacements, and so the area of the shaded region Ω is

$$A(\Omega) = \int_\Omega \frac{1}{y^2}\,dx\,dy,$$

which converges as $y \to \infty$, and hence Ω has finite area. It follows that the area of a triply asymptotic triangle is finite, and independent of our choice of triangle; denote this area by μ. Note that any hyperbolic triangle is contained in a triply asymptotic triangle, and so every hyperbolic triangle must have area less than μ.

In order to complete our calculations for A, we must find a formula for the areas of the doubly asymptotic triangles T_1, T_2, and T_3 (the shaded triangles in Figure 4.20). Note first that by using an isometry to place the non-infinite vertex of a doubly asymptotic triangle at the origin, we see that the area depends only on the angle at the vertex. Given an angle θ, let $f(\theta)$ denote the area of the doubly asymptotic triangle with angle $\pi - \theta$, so that if θ_j is the angle in the triangle at the vertex z_j, then $A(T_j) = f(\theta_j)$.

We may obtain a triply asymptotic triangle as the disjoint union of two doubly asymptotic triangles with angles $\pi - \alpha$, $\pi - \beta$ where $\alpha + \beta = \pi$, and hence

$$f(\alpha) + f(\beta) = \mu.$$

Similarly, we may obtain a triply asymptotic triangle as the disjoint union of three doubly asymptotic triangles with angles $\pi - \alpha$, $\pi - \beta$, and $\pi - \gamma$, where $(\pi-\alpha)+(\pi-\beta)+(\pi-\gamma) = 2\pi$ and hence $\alpha+\beta+\gamma = \pi$, so we have

$$f(\alpha)+f(\beta)+f(\gamma)=\mu$$

for such α, β, γ. We may rewrite the above two equations as

$$\begin{aligned} f(\alpha+\beta)+f(\pi-\alpha-\beta)&=\mu, \\ f(\alpha)+f(\beta)+f(\pi-\alpha-\beta)&=\mu, \end{aligned}$$

and comparing the two gives

$$f(\alpha+\beta)=f(\alpha)+f(\beta)$$

so that f is in fact a linear function. Further, the limit $\alpha \to \pi$ corresponds to a doubly asymptotic triangle whose non-zero angle shrinks and goes to zero, and so the triangle becomes triply asymptotic; hence $f(\pi)=\mu$, and we have

$$f(\theta)=\frac{\mu}{\pi}\theta.$$

It follows that

$$\begin{aligned} A(\Delta)=T_0-T_1-T_2-T_3&=\mu-\frac{\mu}{\pi}(\theta_1+\theta_2+\theta_3) \\ &=\frac{\mu}{\pi}(\pi-\theta_1-\theta_2-\theta_3), \end{aligned}$$

which proves our formula, with constant of proportionality $\frac{1}{-\kappa}=\frac{\mu}{\pi}$.

In order to calculate the coefficient of proportionality for the standard half-plane model consider the triply asymptotic triangle T in the upper half-plane bounded by the unit circle $|z|=1$ and the vertical lines $\operatorname{Re} z=1$ and $\operatorname{Re} z=-1$. The area of T is given by

$$\begin{aligned} \mu=\int_T \frac{1}{y^2}\,dx\,dy &= \int_{-1}^{1}\int_{\sqrt{1-x^2}}^{\infty}\frac{1}{y^2}\,dy\,dx \\ &= \int_{-1}^{1}\frac{1}{\sqrt{1-x^2}}\,dx=\int_{-\pi/2}^{\pi/2}d\theta=\pi \end{aligned}$$

using the substitution $x=\sin\theta$. This confirms the choice $\kappa=-1$ for the usual model. □

Note that the formula is valid not only for finite triangles, but also for asymptotic triangles, since taking a vertex to infinity is equivalent to taking the corresponding angle to zero.

The above proof that the area μ of a triply asymptotic triangle is finite relied on analytic methods, rather than purely synthetic ones.

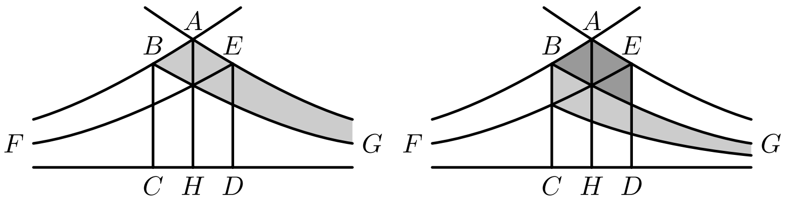


Figure 4.22. Decomposing an asymptotic triangle.

We sketch the purely synthetic proof given in Coxeter's book, which relies only on the fact that area is additive and that reflections are isometries. As before, it suffices to prove that the area of a singly asymptotic triangle is finite.

Consider such a triangle, given by the shaded region in Figure 4.22. Here we begin with the asymptotic triangle ABG and extend the geodesic AB to the point F at infinity. Then we draw the geodesic asymptotic to F and G and add the perpendicular AH, which bisects the angle at A. Note that all the curves in this picture represent geodesics—as this is a purely synthetic picture, it does not refer to either of the models, and in particular, does not include the ideal boundary. Reflecting BG in the line AH gives the geodesic EF; the geodesics BC, ED bisect the appropriate angles and meet the geodesic FG orthogonally.

The bulk of the proof is in the assertion that by repeated reflections first in ED and then in AH, the rest of the shaded region can be brought into the pentagon $ABCDE$. The first step is shown in Figure 4.22, and the details of the proof are left to the reader. Once it is established that ABG can be decomposed into triangles whose isometric images fill $ABCDE$ disjointly, it follows immediately that the area of ABG is finite, and the proof is complete.

Exercise 4.21. Find all the isometries which preserve a triply asymptotic triangle.

Exercise 4.22. Consider a line in the hyperbolic plane and a doubly asymptotic triangle for which this line is one of the sides. Assume the angle at the finite vertex is fixed, and find the locus of all finite vertices.

Lecture 31

a. Hyperbolic metrics on surfaces of higher genus. One model we considered for the flat torus was the real plane modulo the integer lattice. More formally, we took the quotient space $\mathbb{R}^2/\mathbb{Z}^2$, in which points on the torus corresponded to orbits in $\mathbb{R}^2$ of the subgroup $\Gamma \subset \operatorname{Isom}(\mathbb{R}^2)$ comprising integer translations. The discrete subgroup Γ is generated by the translations $(x, y) \mapsto (x + 1, y)$ and $(x, y) \mapsto (x, y+1)$, and the orbit of a point (x, y) in $\mathbb{R}^2$ under the action of Γ is simply the set containing all the images of (x, y) under compositions of these maps and their inverses.

Thus far we have not seen an analogous model for surfaces of higher genus; in the course of this lecture, we will exhibit such a model, but in the hyperbolic plane, rather than the Euclidean. To motivate this, consider an equivalent way of looking at the above model. Rather than taking points on the torus to be entire orbits of Γ, we may restrict our attention to a single *fundamental domain* which contains exactly one point from each orbit, with the exception of boundary points, which are identified somehow.

In the case of the torus, a fundamental domain is given by the unit square $[0, 1] \times [0, 1]$, and opposite edges are identified via the two translations mentioned above, which generate Γ. This is our familiar planar model for the torus, and we see that the images of the fundamental domain under Γ tile the Euclidean plane.

In the course of our topological classification of surfaces, we constructed such planar models for every compact surface, and it is natural to ask if the algebraic construction which works so well for the torus might also be carried out for these planar models. As a concrete example, consider the octagon with opposite sides identified via the four translations

$$\begin{aligned}
f_1 &\colon (x, y) \mapsto (x + 2, y),\\
f_2 &\colon (x, y) \mapsto (x + \sqrt{2}, y + \sqrt{2}),\\
f_3 &\colon (x, y) \mapsto (x, y + 2),\\
f_4 &\colon (x, y) \mapsto (x - \sqrt{2}, y + \sqrt{2}).
\end{aligned}$$

This is a planar model of a surface S with genus two, and so we might hope that if we consider the subgroup $\Gamma \subset \mathrm{Isom}(\mathbb{R}^2)$ generated by $\{f_1, f_2, f_3, f_4\}$ and take the quotient space $\mathbb{R}^2/\Gamma$, we would obtain that same surface. However, things do not work out so nicely; indeed, it is straightforward to verify that the orbit under Γ of each point $(x, y) \in \mathbb{R}^2$ is in fact dense in the plane.

We may gain some insight into the problem by realising that if this approach were to work, the images of the octagon under the isometries in Γ would tile the plane, as was the case for the unit square under integer translations. This is impossible, because the angles of the octagon do not add up correctly—indeed, if just three octagons were to meet at a common vertex, the sum of their angles would be $9\pi/4$, which is already greater than 2π.

Here we encounter the same difficulty we ran into when attempting to place a smooth structure on S. In order for the surface to inherit the geometry of the space tiled by its fundamental domain (this space, if simply connected, is known as its *universal cover*), the eight wedges which make up a neighbourhood of the vertex in the fundamental domain must all be put together into a disc surrounding that vertex; this requires that their angles sum to 2π, not 6π as is the case in the current planar model.

In the Euclidean plane, this is impossible; any octagon, regardless of shape and size, has angles which sum to 6π merely by virtue of being an octagon. We have seen, however, that things are different in the hyperbolic plane, where triangles, at least, have angles whose sum is less than that of their Euclidean counterparts. By decomposing a geodesic polygon in the hyperbolic plane into triangles, we see that a similar formula holds, and the area is proportional to the angular defect vis-à-vis the corresponding Euclidean polygon.

In particular, a geodesic octagon in the hyperbolic plane with area 4π will have angles whose sum is 2π. We will find that H^2 can in fact be tiled with such octagons, and that everything works out just as it did for $\mathbb{R}^2$ and the torus. In order to see this, we must find isometries which will identify the sides of the octagon; while we no longer have translations available in the Euclidean sense, we do have

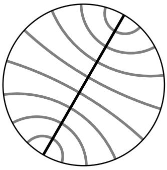


Figure 4.23. Geodesics for a hyperbolic translation.

isometries falling into the case **(H)** discussed last time, which may be thought of as hyperbolic translations.

Given such an isometry f, we have two fixed points at infinity and a unique geodesic γ connecting them. There exists $r > 0$ such that any point $p \in \gamma$ is taken by f to a point $f(p) \in \gamma$ with $d(p, f(p)) = r$. Indeed, given a geodesic γ and a distance r, there exists a unique isometry f with these properties (provided we specify in which direction along γ the points are to be moved).

f also preserves the equidistant curves of γ; we will be most interested, though, in the family of orthogonal geodesics which are pairwise ultraparallel and which are parametrised by their intersection with γ. If we choose coordinates on the open disc model in which γ is a Euclidean line through the origin, then we have the picture shown in Figure 4.23.

Returning to the question of finding a good model for the surface with genus two, consider four geodesics through the origin in H^2 which make angles of 0, $\pi/4$, $\pi/2$, and $3\pi/4$ with the horizontal. We may draw eight more geodesics, each orthogonal to one of the original four, such that each of the eight new geodesics has the same Euclidean radius.

For small values of this radius, these geodesics do not intersect, and are ultraparallel, as shown in the first panel of Figure 4.24. As the radius is increased, neighbouring geodesics eventually become parallel and meet at infinity, as shown in the second panel; at this point the angle between neighbouring geodesics is 0. As the radius is increased still further, as shown in the third panel, this angle increases as well, and the geodesics now intersect in H^2 itself to form an octagon.

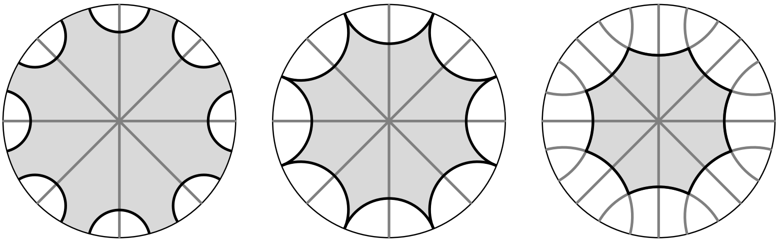


Figure 4.24. Various attempts at a hyperbolic octagon.

In the limit as the radius goes to 1, the octagon becomes more and more nearly Euclidean; correspondingly, its area goes to 0. The individual angles approach (but do not reach) $3\pi/4$, and so their sum approaches (but does not reach) 6π. By the Intermediate Value Theorem, there is some value of the Euclidean radius for which the sum of the angles of the octagon is exactly 2π; this is the octagon we want.

Recalling our discussion of hyperbolic translations, we see that the four geodesics passing through the origin, together with the distance given by the diameter of the octagon, are sufficient to specify four isometries f_1, f_2, f_3, and f_4.

Let Γ be the subgroup of $\text{Isom}(H^2)$ generated by $\{f_1, f_2, f_3, f_4\}$, and consider the quotient space H^2/Γ whose points are orbits of Γ—then as desired, we obtain the surface of genus two. The geodesic octagon found above is the fundamental domain, and its images under Γ tile H^2, just as the images of the unit square under integer translations tile $\mathbb{R}^2$. It may be checked that although the isometries f_j do not commute, they do satisfy the relation

$$f_1 \circ f_2 \circ f_3 \circ f_4 \circ f_1^{-1} \circ f_2^{-1} \circ f_3^{-1} \circ f_4^{-1} = \text{Id},$$

which is reminiscent of our earlier method of cataloguing edge identifications for planar models.

Thus we have succeeded in placing a locally hyperbolic metric on the surface of genus two, as follows: on the interior of the octagon, S obtains its metric directly from H^2; along the edges, we may obtain a patch by using one of the isometries f_j and again inherit the metric from H^2. Finally, at the vertex, where we ran into so much difficulty in defining a smooth structure, there is now no trouble, because the angle is $\pi/4$, and so under the appropriate isometries, the images of the eight wedges in the fundamental domain all come together to fill a neighbourhood of the vertex in H^2, and the metric is passed down without incident.

We may use a similar construction to place a locally hyperbolic metric on any compact orientable surface of genus $g \geq 2$. Beginning with $4g-4$ geodesics through the origin, we find a $(4g-4)$-gon in H^2 whose angles sum to 2π and which has opposite edges identified by hyperbolic translations. By using the fact that any non-orientable surface has an orientable double cover, we also have a locally hyperbolic metric on any compact surface with negative Euler characteristic.

Recall that we can obtain a topological torus by taking any parallelogram and identifying opposite edges by translation, but that these tori will in general have different metric structures. For example, the subgroups of $\text{Isom}(\mathbb{R}^2)$ defined by

$$\begin{aligned}\Gamma &= \langle\, (x,y) \mapsto (x+1,y), (x,y) \mapsto (x,y+1) \,\rangle, \\ \Gamma' &= \langle\, (x,y) \mapsto (x+1,y), (x,y) \mapsto (x+1,y+1) \,\rangle\end{aligned}$$

yield different flat metric structures on the tori $\mathbb{R}^2/\Gamma$ and $\mathbb{R}^2/\Gamma'$, although the two surfaces are identical topologically. Similarly, we may choose a different set of isometries $g_1, g_2, g_3, g_4 \in \text{Isom}(H^2)$ and take the quotient space of H^2 by the action of these isometries; provided the relation

$$g_1 \circ g_2 \circ g_3 \circ g_4 \circ g_1^{-1} \circ g_2^{-1} \circ g_3^{-1} \circ g_4^{-1} = \text{Id}$$

still holds, this quotient space will be a surface of genus two, but with a different hyperbolic metric. This observation is the precursor to what is known as Teichmüller theory.

b. Curvature, area, and Euler characteristic. Why is it that we were able to put a flat metric on the torus, which has $\chi = 0$, but not on surfaces of higher genus, for which $\chi < 0$? We have just seen that although we could not put a flat metric on these surfaces, we could give them a locally hyperbolic metric; might it be possible to do this for the torus as well?

In order to put a locally hyperbolic metric on the torus, we must find a planar model which lies in H^2. If we proceed as before, drawing two orthogonal geodesics passing through the origin and then varying the geodesics orthogonal to these, we obtain an asymptotic quadrilateral. Identifying opposite sides of this quadrilateral with the appropriate hyperbolic translations yields a surface S which is topologically equivalent to a punctured torus; that is, a torus with a point removed. The metric induced on the torus by H^2 has a singularity at this point.

So far this is exactly the picture we began with for surfaces of higher genus; for example, an asymptotic octagon with opposite sides identified corresponds to a surface of genus two with a single point removed and a singularity in the metric around this point. However, for those surfaces we were able to remove the singularity by bringing the geodesics bounding the planar model closer to the origin. This is of no use for the hyperbolic quadrilateral, because as long as the quadrilateral has positive area, the sum of its angles will be less than 2π, and so the singularity at the vertex persists.[3]

This method fails, then, to yield a locally hyperbolic metric on the torus. A deeper reason for this failure is given by the following theorem, which relates area, curvature, and Euler characteristic, and foreshadows the important *Gauss-Bonnet Theorem.* Armed with this theorem, we will be able to state categorically that it is impossible to place a locally hyperbolic metric on the torus, whether by the method attempted above, or by any other.

Theorem 4.11. *Let S be a compact surface with a locally hyperbolic metric (that is, a surface with a metric which is locally isometric to patches of H^2), and let $A(S)$ denote the total area of S. Then*

$$A(S) = -2\pi\chi(S).$$

In general, if S is a compact surface with constant curvature κ, then

$$\kappa A(S) = 2\pi\chi(S).$$

Proof. We use the angular defect formula for the area of a hyperbolic triangle, applied to a geodesic triangulation of S. The existence of

[3] What happens if we try this with a hexagon instead of a quadrilateral?

such a triangulation is easy to establish, and the details are technical rather than conceptual; simply choose a large number of points, draw geodesics connecting them to obtain a geodesic map, and then refine the map until a triangulation is obtained.

Using this triangulation, we have the usual formula for the Euler characteristic:

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

Furthermore, as for any triangulation, counting edges gives $3F = 2E$, and so $F = 2E - 2F$. Finally, for every triangle τ in the triangulation, the angular defect formula (4.12) tells us that

$$A(\tau) = \pi - \alpha - \beta - \gamma$$

where α, β, γ are the angles of the triangle. Summing over all τ yields

$$A(S) = \pi F - 2\pi V$$

since the angles around each vertex sum to 2π, and every angle is counted exactly once. The above information now yields the straightforward calculation

$$A(S) = \pi(F - 2V) = \pi(2E - 2F - 2V) = -2\pi\chi(S)$$

which establishes the first formula.

Note that if we write (4.12) in the form

$$\kappa A(\tau) = \alpha + \beta + \gamma - \pi, \tag{4.13}$$

then this proof goes through for the sphere (1.6) and the plane as well, which have $\kappa = 1, 0$. To take care of all possible values of κ, recall that if we scale the metric by a constant factor, the area scales as the square of that factor, and the curvature scales as the inverse of the area, so that the product $\kappa A(S)$ remains constant and equal to $2\pi\chi(S)$.

To gain the second statement in Theorem 4.11, it must be shown that every surface of constant curvature is locally isometric to either a sphere, the plane, or a hyperbolic plane,[4] depending on whether the curvature is positive, zero, or negative (note that the global structure may be quite different, though). The proof of this uses geodesic polar

[4] We say ‘a’ hyperbolic plane because a value other than -1 for the curvature demands that we scale the metric, as described above.

coordinates, which we will see in the next lecture, and the notion of a *Jacobi field*, which is beyond the scope of this course (for example, see Proposition 10.9 in John M. Lee's *Riemannian Manifolds: An Introduction to Curvature*).

Taking for granted this classification of such surfaces, the above argument establishes the claim for the three basic models, and hence for any surface of constant curvature. □

We originally defined the Euler characteristic in terms of triangulations, and then saw it turn up in homology via the Betti numbers, and in Morse theory via critical points of smooth functions. Theorem 4.11 illustrates yet another guise of the Euler characteristic, this time in terms of curvature and area:

$$\chi(S) = \frac{\kappa A(S)}{2\pi}.$$

This result can in fact be extended to surfaces whose curvature is not constant, as we will soon see when we study the Gauss-Bonnet Theorem. The idea will be to take a triangulation which is fine enough that curvature is nearly constant on each triangle, and then approximate the area of each triangle by using the angular defect/excess formula in its general form (4.13). By showing that this formula remains correct up to a higher order error term in the case of variable curvature, we will be able to replace the expression $\kappa A(S)$ with the integral of the curvature, obtaining the general expression

$$\chi(S) = \frac{1}{2\pi} \int_S \kappa(x)\, dA(x)$$

for the Euler characteristic.

Lecture 32

a. Geodesic polar coordinates. Up to this point, we have discussed curvature in certain specific settings without giving a general definition of curvature for an arbitrary surface with a Riemannian metric. In order to do this, we first recall the three types of surfaces of constant curvature that we have considered so far, and express the metric on each in *geodesic polar coordinates* around a particular point.

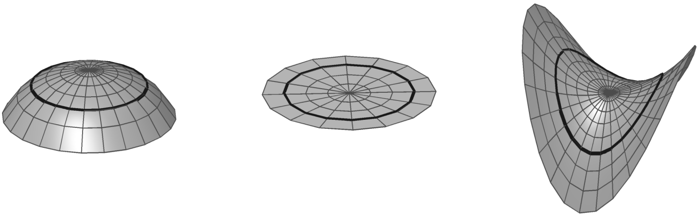


Figure 4.25. Geodesic polar coordinates on surfaces with positive, zero, and negative curvature.

To be more precise, we fix a point $p \in S$ and choose polar coordinates on a neighbourhood U of p such that a point $q \in U$ has coordinates (r, θ), where r is the distance from p to q along the unique geodesic of minimal length connecting the two points, and θ is the angle this geodesic makes with a fixed reference geodesic through p.

Let us conside our three standard symmetric examples, which correspond to the three cases shown in Figure 4.25 (although of course H^2 cannot actually be embedded in $\mathbb{R}^3$).

On the Euclidean plane with p taken to be the origin, these are just the usual polar coordinates (r, θ); the geodesics through p are straight lines through the origin, and the metric is given by

$$ds^2 = dr^2 + r^2\, d\theta^2. \tag{4.14}$$

On the sphere with radius R, we may take p to be the north pole. Then the geodesics through p are the meridians (lines of constant longitude); the point $q = (r, \theta)$ has longitude given by θ and latitude chosen so that its distance from the north pole along that line of longitude is r. One immediately sees that the metric in these coordinates is

$$ds^2 = dr^2 + R^2 \sin^2\left(\frac{r}{R}\right) d\theta^2. \tag{4.15}$$

Finally, on H^2 in the disc model with p as the origin, we see that the geodesics through p are straight lines through the origin, and a straightforward calculation shows that the metric (4.4) becomes

$$ds^2 = dr^2 + \sinh^2 r\, d\theta^2 \tag{4.16}$$

in geodesic polar coordinates.

In general, in the geodesic polar coordinates described above, the curves $\theta = \text{constant}$ are geodesics, while for small values of c the curves $r = c$ are circles centred at p, i.e. the loci of points whose distance from p is precisely c. The circles intersect the geodesics $\theta = \text{constant}$ orthogonally; if it were not so, varying θ along a circle would change r, a contradiction. This fact, together with the definition of r, implies that the metric in these coordinates has the form

$$ds^2 = dr^2 + (g(r,\theta))^2\, d\theta^2 \tag{4.17}$$

where $g\colon \mathbb{R}^2 \to \mathbb{R}$ is some smooth function which is positive for $r > 0$ (away from p) and vanishes at p.

Exercise 4.23. Let S be the surface of revolution around the z-axis of the curve $z = \phi(x)$ in the xz-coordinate plane, where ϕ is an even function. Express the function g which appears in the geodesic polar coordinates (4.17) around the point $(0, 0, \phi(0))$ in terms of the function ϕ.

b. Curvature as an error term in the circle length formula. Our description of curvature in terms of geodesic polar coordinates must come from the properties of the function g. We make the following definition, and then offer some geometric justification.

Definition 4.12. With g as above, the *curvature* of S at a point $q = (r, \theta)$ is

$$\kappa(q) = \kappa(r,\theta) = -\frac{g_{rr}}{g} = -\frac{1}{g}\frac{\partial^2 g}{\partial r^2}. \tag{4.18}$$

Notice that in the three symmetric cases (4.14), (4.15), and (4.16), one obtains $\kappa \equiv 0$, R, and -1, respectively.

Notice also that since g vanishes at $r = 0$, this expression does not initially define curvature at the point p, the centre of the geodesic polar coordinate system. We will show after Theorem 4.13 that the limit of the right hand part of the expression (4.18) as $r \to 0$ exists, and this can be taken as the curvature at that point.

As we will see, this definition makes the proof of the Gauss-Bonnet Theorem (which we will come to shortly) relatively straightforward. However, it lacks a clear geometric interpretation, and also has the weakness of being dependent (*a priori*, at least) on our particular choice of a coordinate system around p. What if we were to define our polar coordinates around some other point on S? Why should we expect to obtain the same value for κ at each point?

We address the first of these issues now, giving a coordinate-free interpretation of the curvature at p in terms of the circumference of small circles around p, assuming that the limit of $-g_{rr}/g$ as $r \to 0$ exists and is finite, which we will show in the next section. Fix $r > 0$ and let $C_p(r)$ be the circle of radius r around p. Abusing notation slightly, we write $\ell(r)$ for the circumference of this circle, and we see that

$$\ell(r) = \ell(C_p(r)) = \int_{C_p(r)} ds = \int_0^{2\pi} g(r,\theta)\, d\theta.$$

In what follows, we sweep issues of the smoothness of g under the rug; everything we say regarding the error estimates on g may be verified using results from ODE theory and the calculus of variations, but we will not get bogged down in the details here.

We fix a value of θ and take the Taylor expansion of g in r around 0; note that the constant term vanishes because $g(0) = 0$. To find the linear term, note that as r goes to zero, we approach the Euclidean case, and the circumference is $2\pi r$ plus some higher order terms. Thus we have

$$g_r|_{r=0} = 1.$$

The quadratic term requires a value of g_{rr} at $r = 0$; because $g(0) = 0$ and $\kappa(p) = -\lim_{r\to 0} g_{rr}/g$ is finite, we must have

$$g_{rr}|_{r=0} = 0$$

and so the quadratic term vanishes. Finally, since $g_{rr} = -\kappa g + o(r) = -\kappa r + o(r)$, we have

$$g_{rrr}|_{r=0} = -\kappa,$$

which allows us to write the Taylor expansion for g as

$$g(r,\theta) = r - \frac{\kappa}{6} r^3 + o(r^3).$$

It follows that the circumference is given by

$$\ell(r) = 2\pi r - \frac{\pi\kappa}{3} r^3 + o(r^3)$$

and we have the following formula for the curvature κ:

$$\kappa(p) = 3 \lim_{r \to 0} \frac{2\pi r - \ell(r)}{\pi r^3}. \tag{4.19}$$

This gives a nice geometric interpretation of curvature; however, because it only applies to the curvature at the origin of the coordinate system, we will need a different argument to show that $\kappa(q)$ is in fact independent of our choice of origin p, and that the limit used above does in fact exist.

Exercise 4.24. Given a surface with Riemannian metric, express the curvature at a point through the error term in the area of a disc centred at this point as the radius goes to zero.

Exercise 4.25. Let S be a surface in $\mathbb{R}^3$, fix a point $p \in S$, and suppose that the degree of tangency between S and its tangent plane at p is greater than one—that is, the distance between a point on the surface at a distance r from p and its projection to the tangent plane is $O(r^3)$. Prove that the curvature of S at p is equal to zero.

Exercise 4.26. As before, write $\ell(r)$ and $A(r)$ for the length and area of a circle and a disc of radius r in the hyperbolic plane. Find

(1) $\lim_{r\to\infty} \frac{\log \ell(r)}{r}$;

(2) $\lim_{r\to\infty} \frac{\log A(r)}{r}$.

The conclusion of the last exercise is rather surprising. It shows that both the length of a circle and the area of a disc grow *exponentially* with the radius, in contrast with the linear and quadratic growth seen in Euclidean geometry.

c. The Gauss-Bonnet Theorem. We are now in a position to state and prove the Gauss-Bonnet Theorem for a general geodesic triangle with variable curvature (Figure 4.26). In the following, we write dS for an infinitesimal area element.

Theorem 4.13. *Let A, B, and C be the vertices of a geodesic triangle Δ on a surface S, and let α, β, and γ be the angles at these vertices. Then the integral of the curvature of S over Δ is equal to the angular*

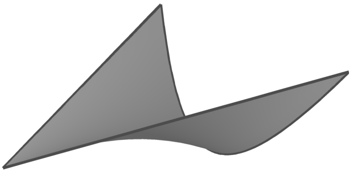


Figure 4.26. A geodesic triangle with variable curvature.

excess:

$$\int_\Delta \kappa\, dS = \alpha + \beta + \gamma - \pi. \tag{4.20}$$

Proof. Choosing geodesic polar coordinates centred at A, the integral in question is

$$\int_\Delta \kappa\, dS = \int_\Delta \kappa g(r,\theta)\, dr\, d\theta = -\int_\Delta g_{rr}(r,\theta)\, dr\, d\theta.$$

For $0 \le \theta \le \alpha$, let γ_θ be the geodesic through A which makes an angle of θ with the geodesic AB (as in Figure 4.27), and let $\rho(\theta)$ be the distance along γ_θ from A to the opposite side BC. Then the above integral may be rewritten as

$$\begin{aligned}-\int_0^\alpha \int_0^{\rho(\theta)} g_{rr}(r,\theta)\, dr\, d\theta &= -\int_0^\alpha g_r(r,\theta)\Big|_{r=0}^{r=\rho(\theta)} d\theta \\ &= \int_0^\alpha -g_r(\rho(\theta),\theta) + 1\, d\theta = \alpha - \int_0^\alpha g_r(\rho(\theta),\theta)\, d\theta.\end{aligned}$$

Lemma 4.14. *With γ_θ as above, let $\psi(\theta)$ be the angle of intersection of γ_θ and the geodesic BC. Then*

$$\frac{d\psi}{d\theta} = -g_r(\rho(\theta),\theta).$$

Proof. Parametrising the geodesic BC by arc length s, we see (Figure 4.27) that

$$\cos\psi = \frac{dr}{ds}, \qquad \sin\psi = g\frac{d\theta}{ds}.$$

Differentiating the first and then using the second yields

$$\frac{d^2r}{ds^2} = -\sin\psi\frac{d\psi}{ds} = -g\frac{d\theta}{ds}\frac{d\psi}{ds}. \tag{4.21}$$

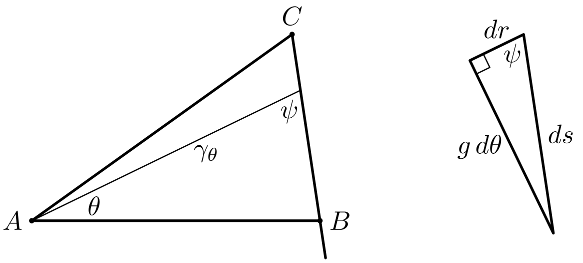


Figure 4.27. Computing $\frac{d\psi}{d\theta}$.

In order to complete the proof, we need to appeal to the equations for a geodesic, which we did not derive explicitly. However, for the particular case of geodesic polar coordinates, they are not so bad; recall that we defined geodesics as those curves which minimise the action functional

$$\int_a^b \frac{1}{2}\|\dot{\gamma}\|^2\,dt = \int_a^b \frac{1}{2}\dot{r}^2 + g(r,\theta)^2\dot{\theta}^2\,dt,$$

and so since BC is a geodesic, the Euler-Lagrange equations from Proposition 4.10 become

$$\begin{aligned} gg_r\dot{\theta}^2 &= \frac{d}{dt}\dot{r} = \ddot{r}, \\ gg_\theta\dot{\theta}^2 &= \frac{d}{dt}g^2\dot{\theta}. \end{aligned}$$

We only need the first of these; since the action functional can only be minimised when γ is parametrised by arc length, we have $ds = dt$, and (4.21) gives

$$gg_r\left(\frac{d\theta}{ds}\right)^2 = gg_r\dot{\theta}^2 = \frac{d^2r}{ds^2} = -g\frac{d\theta}{ds}\frac{d\psi}{ds},$$

from which the lemma follows. □

Using this lemma, we may continue the above computations and write the integral as

$$\alpha + \int_{\theta=0}^{\theta=\alpha} d\psi = \alpha + \psi(\theta)\Big|_{\theta=0}^{\theta=\alpha} = \alpha + \gamma - (\pi - \beta) = \alpha + \beta + \gamma - \pi,$$

which completes the proof. □

Of course, so far we have not shown that the definition (4.18) of the curvature $\kappa(q)$ is independent of our choice of the origin p, and so we really should replace the word 'curvature' in the above theorem with the phrase 'curvature in the coordinates centred at A'. However, it is not too hard to show that the result holds for any choice of origin p, provided that p is close enough to the vertices of the triangle to guarantee the existence of a unique shortest geodesic from p to each vertex. Then if p is inside ABC, we can decompose the triangle ABC into three smaller triangles pAB, pBC, and pCA; using p as the origin, the theorem holds for each, and summing the resulting formulae gives (4.20) for the coordinates centred at p. The case where p lies outside ABC can be handled similarly.

Now we can finally show that the value in (4.18) at a given point q does not depend on the choice of the centre point p for the geodesic polar coordinate system. This is because (4.20) implies that curvature can be defined intrinsically—even though this intrinsic definition is an immediate corollary of Theorem 4.13, it is an important enough fact to warrant formulation as a separate proposition.

Proposition 4.18. *The curvature at a point p of a surface with a Riemannian metric is equal to the limit of the ratio of the angular excess of a small geodesic triangle Δ (that is, the difference between the sum of its angles and π) and the area of Δ, taken as all vertices of Δ converge to p.*

This generalises to geodesic polygons in the natural way.

In order to tie up all the loose ends of our exposition, it remains only to justify our derivation of (4.19) by showing that $\lim_{r\to 0} -g_{rr}/g$ exists. But we have just shown that the quantity in the limit is independent of our choice of p, and so choosing some other origin, the limit in question becomes

$$\lim_{(r,\theta)\to(r_0,\theta_0)} -\frac{g_{rr}}{g},$$

where $r_0 \neq 0$, and this limit obviously exists.

Returning to the statement of Theorem 4.13, if we consider the boundary of the triangle as a single closed curve, then it is a piecewise smooth curve which is a geodesic at all but three points where it has a corner. The content of the theorem is that the integral of the curvature is equal to the sum of the angles at these corners minus π; there is a more general version of this theorem which deals with curves with more than three corners, and even with curves which are not geodesics. In the latter case, we must include a term accounting for the *geodesic curvature* of the boundary, as well as any angles where the curve is not smooth.

Now we give an example, in the form of two exercises, which shows that Proposition 4.18 can sometimes be used to calculate curvature. Here $\mathcal{H}$ is the one-sheeted hyperboloid in $\mathbb{R}^3$ given by the equation $x^2 + y^2 - z^2 = 1$.

Exercise 4.27. Prove that through every point of $\mathcal{H}$ pass two straight lines which lie in $\mathcal{H}$. Find the equations of these lines using the coordinate z as a parameter, and prove that the lines are geodesics in $\mathcal{H}$.

Exercise 4.28. Prove that $\mathcal{H}$ has negative curvature at every point.

As an important corollary of Theorem 4.13, we obtain another classical description of the Euler characteristic.

Theorem 4.15 (Gauss-Bonnet). *For any Riemannian metric on a compact surface S,*

$$\int_S \kappa \, dS = 2\pi\chi(S).$$

The Gauss-Bonnet Theorem is deduced from Theorem 4.13 in the same way as in the case of constant curvature. In that case we added areas of triangles, while here we add integrals over triangles, but the rest of the proof is verbatim. We only need to make sure that there exists a triangulation of the surface into geodesic triangles—for this we take a finite but sufficiently dense set of points and connect pairs of points from the set which are sufficiently close to each other by unique short geodesic segments. Looking at the part of the picture which lies inside any particular coordinate chart, we obtain a decomposition of the patch into geodesic polygons, which then can be further triangulated.

Figure 4.28 shows a sphere with two handles, shaded by curvature; darker areas have positive curvature, lighter areas have negative

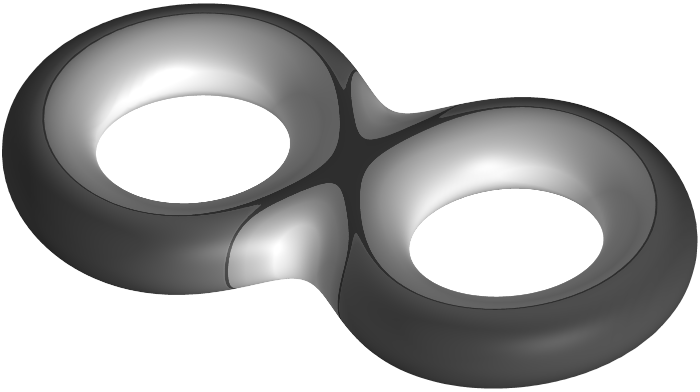


Figure 4.28. Gaussian curvature on a sphere with two handles.

curvature, and points with zero curvature are indicated by the dark curves and the ‘X’-shaped region in the middle of the figure eight. Because the Euler characteristic of the surface is negative, Theorem 4.15 implies that the average curvature is negative as well; a calculation of the values of the curvature (which we do not carry out here) shows that $|\kappa|$ is roughly twice as large in the lighter regions of Figure 4.28 as in the darker regions, and hence the negative values predominate upon integration over the whole surface.

d. Comparison with traditional approach. The path that we have taken to reach this point is somewhat different from the traditional approach to differential geometry. One of the fundamental difficulties of the subject is the lack of a preferred coordinate system in which to make definitions, perform calculations, etc. In our treatment of curvature, we used geodesic polar coordinates as our preferred system, but these still suffer from two drawbacks. In the first place, as we remarked above, they depend on the choice of origin, and so are not completely general; in the second place, they are singular at that origin, and so cannot be used on the tangent space of the point in which we are most interested!

The traditional approach to the difficulty of coordinate systems is to consider a surface which is embedded in $\mathbb{R}^3$, for then we do indeed have the preferred coordinates (x, y, z) which are inherited from the

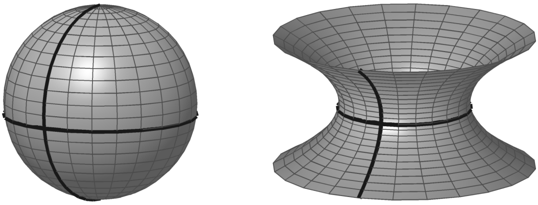


Figure 4.29. Directions of principal curvature on a sphere and a hyperboloid.

ambient space. Given a particular chart $\phi\colon (x, y, z) \mapsto (u, v)$, we may do our calculations of curvature and other geometric properties in terms of $x(u, v)$, $y(u, v)$, and $z(u, v)$, and then derive their forms in terms of u and v from the coordinates in $\mathbb{R}^3$.

In this philosophy, the approach to curvature is as follows. At each point of $S \subset \mathbb{R}^3$, we have a unit normal vector n. Given a tangent vector v at a point $p \in S$, we may consider the plane spanned by n and v; this plane intersects S in a curve γ through the point p. Since γ lies in a plane, we know how to compute its curvature (osculating circles), and we say that this is the curvature of S in the direction v.

In the course of these calculations, a 2×2 matrix arises which determines how the curvature changes as v changes; the eigenvalues of this matrix are the *principal curvatures* of S at p. On a positively curved surface such as a sphere or an ellipsoid, both principal curvatures have the same sign, which is reflected in the fact that the two curves on the sphere which are highlighted in Figure 4.29 open in the same direction. On a negatively curved surface, the principal curvatures have different signs from each other, and so the two highlighted curves on the hyperboloid pictured open in different directions.

The punchline of all of this is that while all of the definitions are completely extrinsic, being dependent on the particular choice of embedding into $\mathbb{R}^3$, the product of the principal curvatures, the so-called *Gaussian curvature*, is in fact completely *intrisically* determined (this is our κ). That is, the embedding of S into $\mathbb{R}^3$ induces a Riemannian metric on S from the metric on $\mathbb{R}^3$, and the Gaussian curvature depends only on this metric, and not on the embedding; this is Gauss' Theorema Egregium.

In our treatment here, we have eschewed the traditional approach, avoiding the technical discussions and computations it inevitably entails; for example, we have made no mention of Christoffel symbols, which the reader will encounter in any more in-depth studies of differential geometry. This has allowed us to cover more ground than we would have otherwise, but the reader ought to be aware that certain common topics have been omitted, as they will undoubtedly appear in any further studies of this material.

Chapter 5

Topology and Smooth Structure Revisited

Lecture 33

a. Back to degree and index. In examining vector fields, curves, etc. on a smooth surface S, there is a natural ambiguity in the terminology surrounding the notion of 'index'—do we speak of the index of a vector field at a critical point, or the index of a critical point of a vector field? In terms of curves on S, do we speak of the index of a curve with respect to a point, or the index of a point with respect to a curve?

Both options make perfect sense, and indeed both are legitimate. For the sake of concreteness, though, we shall choose the former for the time being, and refer to the index of a curve γ with respect to a point x, denoted $\operatorname{ind}_x \gamma$.

As we have seen, this index is independent of parametrisation; nevertheless, in order to work with the curve and determine properties of the index, we fix a parametrisation

$$\gamma\colon S^1 \to \mathbb{R}^2.$$

It is worth pointing out at this juncture that if the curve is not smooth and is allowed to have self-intersections, it may have certain pathological properties. Smooth curves, even with self-intersections, behave more or less according to our intuition; even curves which are merely continuous exhibit many nice properties, provided γ is injective and the curve does not intersect itself.

In the most general case, however, our intuition fails; it turns out that we can find a continuous curve γ such that $\gamma(S^1)$ has a non-empty interior. The classic example is the Peano curve, a continuous surjective map from the unit interval $[0, 1]$ onto the unit square $[0, 1] \times [0, 1]$. This is usually constructed via an inductive geometric procedure, but could also be given explicitly in terms of the binary expansion of the parameter $t \in [0, 1]$.

Recall that given a curve γ and a point $x \in \mathbb{R}^2 \setminus \gamma(S^1)$, we define the index of γ around x by means of a circle map $\phi_{x,\gamma}$. This map is defined by

$$\begin{aligned} \phi_{x,\gamma} \colon S^1 &\to S^1, \\ t &\mapsto \frac{\gamma(t) - x}{\|\gamma(t) - x\|}, \end{aligned}$$

where the fact that we subtract x from $\gamma(t)$ also illustrates the need to fix a choice of local coordinates. The index is given by

$$\operatorname{ind}_x \gamma = \deg \phi_{x,\gamma}.$$

Note that the quantity $\|\gamma(t) - x\|$ is non-vanishing because $x \notin \gamma(S^1)$. Further, by compactness of S^1, this quantity attains its minimum, and hence is bounded away from zero; that is, there exists $\varepsilon > 0$ such that $\|\gamma(t) - x\| \geq \varepsilon$ for every $t \in S^1$.

What properties does the index have? How does it behave if we vary x or γ? We begin by investigating what happens as x varies—note that the complement $\mathbb{R}^2 \setminus \gamma(S^1)$ is an open set, and so upon decomposing it into connected components, we find that these components must themselves be open, and hence are path-connected. This allows us to prove the following:

Proposition 5.19. *Let C be a connected component of $\mathbb{R}^2 \setminus \gamma(S^1)$. Then $\operatorname{ind}_x \gamma$ is constant on C as a function of x.*

Proof. Given $x_0, x_1 \in C$, the above discussion shows the existence of a curve $\delta \colon [0, 1] \to C$ such that $\delta(0) = x_0$, $\delta(1) = x_1$. Now define

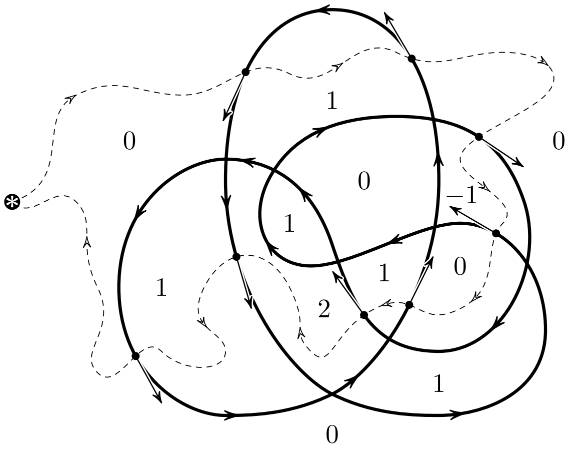


Figure 5.1. Index of points w.r.t. a curve.

$f\colon [0,1] \to \mathbb{Z}$ by

$$f(t) = \operatorname{ind}_{\delta(t)} \gamma.$$

Because the circle map $\phi_{x,\gamma}$ depends continuously on x, the function f is continuous, and hence constant since the integers are discrete. It follows that

$$\operatorname{ind}_{x_0} \gamma = f(0) = f(1) = \operatorname{ind}_{x_1} \gamma. \qquad \square$$

How does the index change if x passes from one connected component to another? In the simplest case, we consider passing through a point at which the curve is smooth, regular, and injective (see Figure 5.1). Formally, we assume that $t \in S^1$ is such that γ is smooth at t and $\gamma(t)$ is non-critical; that is, $\gamma'(t) \neq (0,0)$ using local coordinates. It is worth noting that the Implicit Function Theorem then guarantees the existence of some system of local coordinates in which $\gamma(t) = (t,0)$, so γ is just one of the coordinate axes.

Under the further assumption of injectivity, that there does not exist any parameter value $s \neq t$ with $\gamma(s) = \gamma(t)$, we will show in the next lecture that as x passes from one connected component to another through the point $\gamma(t)$, the index $\operatorname{ind}_x \gamma$ changes by one. Whether it increases or decreases depends on the direction of the parametrisation relative to the direction in which x moves across the curve.

This gives us a sense of how the index responds to variations in the point x. What happens if the curve γ changes? It turns out that we find a similar continuous dependence; here the topology on the space of all possible curves is the $\mathcal{C}^0$ topology, which is generated by the following metric:[1]

$$d(\gamma_1, \gamma_2) = \max_{t \in S^1} d(\gamma_1(t), \gamma_2(t)).$$

Let $\varepsilon > 0$ be as before, so that $\|\gamma(t) - x\| > \varepsilon$ for all t; then for any curve $\tilde{\gamma}$ with $d(\gamma, \tilde{\gamma}) < \varepsilon$ we have

$$\operatorname{ind}_x \tilde{\gamma} = \operatorname{ind}_x \gamma.$$

It follows that the index remains constant under continuous deformations. To be precise, suppose γ_s is a continuous one-parameter family of curves, with $x \notin \gamma_s(S^1)$ for every $s \in [0, 1]$. Then the value of $\operatorname{ind}_x \gamma_s$ is constant.

Exercise 5.1. Let $\gamma \colon S^1 \to \mathbb{R}^2$ be a closed curve and $f \colon S^1 \to S^1$ a continuous map. Let $\gamma_f(t) = \gamma(f(t))$. Prove that

$$\operatorname{ind}_x \gamma_f = \deg f \cdot \operatorname{ind}_x \gamma.$$

b. The Fundamental Theorem of Algebra. The fact that continuous deformation of the curve γ does not change its index with respect to the point x is central to one proof of the Fundamental Theorem of Algebra, which states that every polynomial with complex coefficients has a complex root. It is a somewhat odd fact that despite the completely algebraic nature of this statement, there is no purely algebraic proof known.

The name is also belied by the fact that modern algebra has followed a direction in which the complex numbers are no longer the most important objects, and so the theorem is not so fundamental to algebra anymore. Classically, however, it forms the capstone of the steady progression from the natural numbers to the integers, from

[1] Note that under this definition, reparametrisations of the same curve are considered to be different curves, lying a positive distance from each other. If we want to consider reparametrisations as equivalent, then we must take the infimum over all parametrisations.

the integers to the rationals, from the rationals to the reals, and from the reals to the complex numbers, each step of which may be seen as being motivated by the desire to include roots of more and more polynomials.

Theorem 5.1. *Let $p \in \mathbb{C}[z]$ be a polynomial with complex coefficients. Then there exists $z_0 \in \mathbb{C}$ such that $p(z_0) = 0$.*

Corollary 5.2. *Every polynomial map $p\colon \mathbb{C} \to \mathbb{C}$ is surjective—for every $c \in \mathbb{C}$ there exists $z \in \mathbb{C}$ such that $p(z) = c$.*

Corollary 5.3. *Every polynomial with complex coefficients factors as a product of linear terms; given any $p \in \mathbb{C}[z]$ there exist $a, z_1, \dots, z_n \in \mathbb{C}$ such that*

$$p(z) = a(z - z_1) \dots (z - z_n).$$

Proof of the theorem. Consider the circle of radius r around the origin:

$$C_r = \{ z \in \mathbb{C} \mid |z| = r \}.$$

C_r is homeomorphic to S^1 via a simple homothety, and so the restriction of p to C_r defines a curve

$$\gamma_r \colon S^1 \to \mathbb{C}.$$

Now γ_r gives a continuous family of curves with $r \in [0, \infty)$. We consider the index $\operatorname{ind}_0 \gamma_r$ of these curves around the origin as r varies, and proceed by contradiction. Suppose $p(z) \neq 0$ for every $z \in \mathbb{C}$. Then in particular, $z \notin \gamma_r(S^1)$ for all values of r, and so our previous result implies that $\operatorname{ind}_0 \gamma_r$ is constant. Since $\gamma_0(S^1)$ is just a point, the associated circle map is a constant map, and we have

$$\operatorname{ind}_0 \gamma_r = 0$$

for every $r \geq 0$.

We now claim that this fails for very large values of r:

Lemma 5.2. *For sufficiently large values of r, we have*

$$\operatorname{ind}_0 \gamma_r = \deg p.$$

Proof of the lemma. Let $n = \deg p$, and write

$$p(z) = a_n z^n + q(z),$$

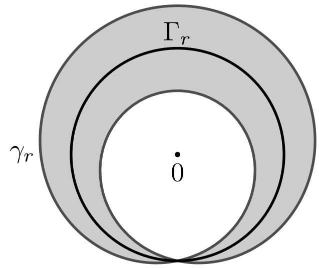


Figure 5.2. γ_r has non-zero index around 0 when r is large.

where $\deg q \leq n-1$. Then we have

$$\gamma_r(t) = p(re^{2\pi it}) = a_n r^n e^{2\pi int} + q(re^{2\pi it}).$$

Let Γ_r be the curve given by the leading term:

$$\Gamma_r(t) = a_n z^n = a_n r^n e^{2\pi int}$$

where $z = re^{2\pi int}$. The circle map associated to Γ_r is just the expanding map $t \mapsto nt$, which has degree n, and hence $\operatorname{ind}_0 \Gamma_r = n$. It remains to show that γ_r and Γ_r have the same index around the origin (see Figure 5.2).

Consider the family of curves

$$\gamma_r^s(t) = a_n z^n + (1-s)q(z)$$

which has $\gamma_r^0 = \gamma_r$ and $\gamma_r^1 = \Gamma_r$. Since q has degree at most $n-1$, there exists some constant $C > 0$ such that $|q(z)| \leq C|z|^{n-1}$ whenever $|z| > 1$, and so we have for any $t \in S^1$ that

$$|\gamma_r^s(t)| \geq |a_n|r^n - Cr^{n-1}.$$

For sufficiently large values of r (in particular, $r > C/|a_n|$), this is always positive, and hence all the curves γ_r^s avoid the origin. It follows that $\operatorname{ind}_0 \gamma_r^s$ is constant in s, and so $\operatorname{ind}_0 \gamma_r = \operatorname{ind}_0 \Gamma_r = n$. □

From this contradiction, we deduce that there must exist some $z \in \mathbb{C}$ with $p(z) = 0$, which completes the proof of the theorem. □

Lecture 34

a. Jordan Curve Theorem. Common sense tells us that a circle has an inside and an outside—if we draw a circle in the dirt and then stand at a point which is not part of the circle, then we are either inside the circle or outside of it. Mathematically, this may be rephrased as the statement that the plane with a circle removed has exactly two connected components.

The generalisation of this assertion from circles to arbitrary continuous closed curves without self-intersection is known as the *Jordan Curve Theorem*, which we will state momentarily and then proceed to prove in the course of this and the next lecture. As is often the way of things in topology, this innocuous-looking theorem is rather more difficult to prove than naïve intuition would lead us to expect, due in part to the fact that the homeomorphic image of a circle (that is, a continuous closed curve without self-intersection) may have a fantastically complicated local structure, even taking the form of a fractal.

Recall that the plane is homeomorphic to the sphere with a point removed, and hence we have a correspondence between curves in $\mathbb{R}^2$ and curves on S^2 (via stereographic projection, for example). In the prototypical example where our curve is the unit circle, the interior of the curve is a disc, and the exterior of the curve is homeomorphic to a disc if we include the point at infinity. This may readily be seen by considering the form this curve takes on the sphere, where it is simply the equator. The equator separates S^2 into two connected components, the northern and southern hemispheres, each of which is homeomorphic to a disc.

Theorem 5.3 (Jordan Curve Theorem)**.** *Let $\gamma\colon S^1 \to \mathbb{R}^2$ be a homeomorphism onto its image. Then $\mathbb{R}^2 \setminus \gamma(S^1)$ consists of two connected components.*

As discussed above, the same result holds if we replace $\mathbb{R}^2$ with S^2. In fact, we can strengthen this theorem (in $\mathbb{R}^2$ or S^2) and relate it directly to the prototypical case.

Theorem 5.4 (Schoenflies)**.** *Let $\gamma\colon S^1 \to \mathbb{R}^2$ be a homeomorphism onto its image. Then there exists a homeomorphism $h\colon \mathbb{R}^2 \to \mathbb{R}^2$ such that $h(\gamma(S^1))$ is the unit circle.*

If γ is a smooth regular curve, then h can be chosen to be a diffeomorphism.

By adding the point at infinity, one again obtains the corresponding result for the sphere.

An important corollary of the Schoenflies Theorem is that all handles are the same—in other words, the procedure of attaching a handle is uniquely defined up to a homeomorphism. The same holds true for Möbius caps. Since we do not plan to give a complete proof of the Schoenflies Theorem, we omit the details of these deductions.

The proofs of both the Jordan Curve and Schoenflies Theorems rest upon the technique of approximating an arbitrary continuous curve γ with smooth or piecewise smooth curves, so we begin by restricting our attention to such curves. The notion of index of a curve (which will be fixed) with respect to a point (which will change) plays a central role in this argument.

Theorem 5.5 (Smooth Jordan Curve Theorem)**.** *Let $\gamma\colon S^1 \to \mathbb{R}^2$ be smooth, regular (which in this case means that the derivative does not vanish), and without self-intersection. Then $\mathbb{R}^2 \setminus \gamma(S^1)$ consists of two connected components.*

Proof. First note that the result is true locally, as a consequence of the Implicit Function Theorem. That is, given a neighbourhood $U \subset \mathbb{R}^2$ such that $\gamma(S^1) \cap U$ is homeomorphic to a line (in other words, γ passes through U exactly once), we can find coordinates on U such that $\gamma(S^1) \cap U$ is the x-axis. Thus $U \setminus \gamma(S^1)$ has exactly two components, corresponding to the upper and lower half-planes.

This picture allows us to prove the claim from the last lecture that passing over such a segment of $\gamma(S^1)$ changes the index $\operatorname{ind}_x \gamma$ by exactly one. Consider two points x_1 and x_2 in U which lie just above and just below the x-axis, respectively, in our coordinate system, such that the distance between them is small compared with the distance to the edge of U (Figure 5.3). Then for points $\gamma(t) \notin U$, the vectors

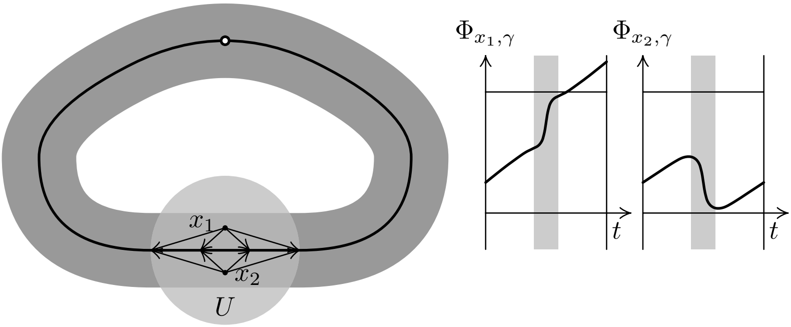


Figure 5.3. Crossing a curve changes index by one.

$\gamma(t) - x_1$ and $\gamma(t) - x_2$ are nearly identical, and so the circle maps associated with x_1 and x_2 differ substantially only on the interval (a, b), where $\gamma(S^1) \cap U = \gamma((a, b))$; this is the shaded area in the graphs of $\Phi_{x_i,\gamma}$ in Figure 5.3.

As t goes from a to b, the direction of the vector $\gamma(t)-x_i$ changes by an amount nearly equal to π. The difference between x_1 and x_2 is that the direction in which $\gamma(t)-x_i$ moves on that interval is different for each one, and hence the degrees of the circle maps differ by one.

Returning to our proof of the theorem, we observe that every $x \in \mathbb{R}^2 \setminus \gamma(S^1)$ belongs to a connected component which contains points arbitrarily close to the curve. This follows by considering a line ℓ connecting x and some point on the curve, then taking the point of $\ell \cap \gamma(S^1)$ which lies nearest x.

In order to complete the proof, we need the idea of a *tubular neighbourhood*, which is important in differential topology. We state and prove a lemma for curves on surfaces, an analogue of which holds for submanifolds of higher-dimensional smooth manifolds. We will use a Riemannian metric as a convenient auxiliary tool—in the plane, of course, one can use the standard Euclidean metric, and the geodesics in question become simply line segments.

Lemma 5.6. *Given a smooth regular curve $\gamma\colon S^1 \to S$ without self-intersections on an orientable smooth surface S, there exists a neighbourhood $U \supset \gamma(S^1)$ and a diffeomorphism $\Gamma\colon A \to U$, where $A \subset \mathbb{R}^2$*

is an annulus with coordinates (r, θ)*,* $\theta \in S^1$*,* $r \in (1 - \varepsilon, 1 + \varepsilon)$*, such that* $\Gamma(1, \theta) = \gamma(\theta)$*.*

Proof of the lemma. We use *Fermi geodesic coordinates*, which are an analogue of the geodesic polar coordinates we used in our discussion of curvature. At each point $\gamma(t)$ on the curve, there exists a unique geodesic η_t which intersects the curve orthogonally; along each such geodesic, we introduce an arc length parametrisation such that $\eta_t(1) = \gamma(t)$ and the positive direction $\eta_t'(1)$ varies continuously with t (this is possible because S is orientable).

Defining Γ by $\Gamma(r, \theta) = \eta_\theta(r)$, it remains only to show that Γ is a diffeomorphism for a sufficiently small value of ε. This holds because $\gamma(S^1)$ is compact—for each geodesic η_t, we may consider the minimal value of s such that either of $\eta_t(1 + s)$ or $\eta_t(1 - s)$ lies on some other geodesic η_τ. This value is continuous with respect to t, and is always positive, hence is bounded away from zero. □

Note the analogy with our discussion of the isometries of the hyperbolic plane—for fixed values of r, the curves $\Gamma(r, \theta)$ are equidistant curves from γ, which intersect the one-parameter family of geodesics η_θ orthogonally.

Fixing a tubular neighbourhood of $\gamma(S^1)$, we see that it has exactly two components, which are the images under Γ of $(1 - \varepsilon, 1) \times S^1$ and $(1, 1 + \varepsilon) \times S^1$. Then since as we observed before, any point $x \in \mathbb{R}^2 \setminus \gamma(S^1)$ lies in the same component as points arbitrarily near $\gamma(S^1)$, Theorem 5.5 follows. □

The index argument depends on the global structure of the plane; on compact orientable surfaces other than the sphere, existence of a tubular neighbourhood does not guarantee that the curve separates the surface into two different components, since points in the two halves of the neighbourhood may be connected through the outside of it. A simple example is given by the curve $\gamma(t) = (t, 1/2)$ on the flat torus $[0, 1] \times [0, 1]/\sim$.

On a non-orientable surface, Lemma 5.6 holds for some curves and fails for others. Whether it holds or not depends on what happens to a vector in the normal direction when it is carried around the curve. If it changes orientation (consider, for example, the middle circle of the Möbius strip), the neighbourhood remains connected after the curve itself is removed. We will see that this has something to do with the different relations between the genus g and the Euler characteristic χ for orientable ($\chi = 2 - 2g$) and non-orientable ($\chi = 2 - g$) surfaces.

Exercise 5.2. Let $\gamma\colon S^1 \to \mathbb{R}^2$ be a smooth regular closed curve with one transversal (non-tangential) self-intersection, i.e. the curve intersects itself in just one point at a non-zero angle. Prove that the complement of γ consists of three connected components, and list (with a proof) all possibilities for the indices of points in those components with respect to γ.

Exercise 5.3. Prove Theorem 5.5 for piecewise smooth curves.

Notice that any polygonal broken line without self-intersections is a piecewise smooth curve. Hence it separates the plane into two parts, one compact and one not. Now one can apply Lemmas 2.5 and 2.7 to deduce that the compact part is indeed homeomorphic to a closed disc. One can also consider the non-compact part as a polygonal domain by adding a point at infinity and tinkering with coordinates a bit. Thus one may construct a triangulation which agrees with the triangulation of the first domain along the boundary, and thus prove the Schoenflies theorem for a polygonal curve on the sphere, which of course implies the similar statement for the plane. Finally, one can deduce the theorem for a smooth or even piecewise smooth curve by using polygonal approximation.

Exercise 5.4. Given a piecewise smooth closed curve $\gamma\colon S^1 \to \mathbb{R}^2$ without self-intersections, show that there exists a polygonal curve $\tilde{\gamma}$ and a homeomorphism $h\colon \mathbb{R}^2 \to \mathbb{R}^2$ such that $h(\gamma(S^1)) = \tilde{\gamma}(S^1)$.

b. Another interpretation of genus. Thanks to our classification of surfaces admitting triangulations, which implies that any such surface S is homeomorphic to a sphere with some number of handles and/or Möbius caps, we know that S admits a smooth structure. The converse is also true; given a smooth surface, taking an appropriate

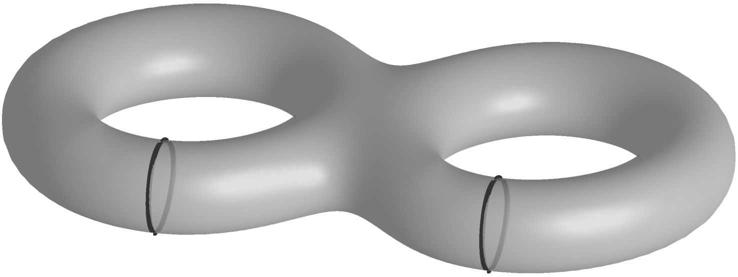


Figure 5.4. Two disjoint curves which do not disconnect a surface of genus two.

set of points and drawing geodesics between them yields a triangulation.[2] Hence the class of surfaces admitting triangulations is the same as the class of surfaces admitting smooth structures—this allows us to give an interpretation of the genus of a surface in terms of smooth closed curves.

Theorem 5.7. *The genus g of a smooth surface S is equal to the maximum number of pairwise disjoint smooth regular curves without self-intersection which may be found on S such that the complement of their union is connected.*

Proof. Consider orientable surfaces first. Let N be the maximum number of such curves. By considering a sphere with N handles and drawing a curve on each handle as shown in Figure 5.4 for the case $g = 2$, we see that $N \geq g$.

To obtain the reverse inequality, consider a collection of $g + 1$ pairwise disjoint smooth regular curves without self-intersection on S. Let $\mathcal{T}$ be a triangulation of S such that each curve γ is a union of edges of the $\mathcal{T}$. Upon removing γ from S, we are left with a surface of genus $g - 1$ with 2 holes (boundary components). Filling these holes in gives a surface in which the number of edges and the number of vertices have both been changed by the same amount, while the number of faces has increased by 2, and hence χ has increased by 2.

[2] Recall that we used existence of a triangulation into geodesic triangles in our proof of the Gauss-Bonnet Theorem.

The Euler characteristic of an orientable surface is $\chi = 2 - 2g$, and so repeating this g times, we obtain a surface with $\chi = 2$, which must be the sphere. Thus the next curve disconnects the surface, by Theorem 5.5, and so $N \leq g$.

Now consider a sphere with q Möbius caps. The boundaries of these caps are disjoint closed curves (recall that the boundary of a Möbius strip is a circle), the removal of which still leaves a connected surface. Thus $N \geq q$.

Conversely, consider any collection of $q+1$ disjoint closed (smooth non-self-intersecting) curves. Once again we can assume there is a triangulation for which each curve is a collection of edges. Removing a curve makes either two holes (if a tubular neighbourhood exists) or one (otherwise), and filling each hole increases the Euler characteristic by one.

Recall that the Euler characteristic of a non-orientable surface is $\chi = 2-g$. Filling all the holes created in the previous step, we observe that since the maximal Euler characteristic of a connected surface is two, and the only surface with $\chi = 2$ is the sphere, Theorem 5.5 once again implies that $q + 1$ curves separate the surface. □

It is natural to try to prove the full Jordan Curve Theorem 5.3 by approximating a given continuous non-self-intersecting curve γ with a sequence of smooth curves, to which Theorem 5.5 may be applied. Since γ is given in local coordinates by a pair of continuous functions, which can be easily approximated by smooth functions, we may try to approximate γ globally by 'gluing together' the local approximations using a partition of unity.

Two problems appear, however—the resulting curves may not be regular, and they may have self-intersections. The first problem is technical and can be easily solved; the second is more serious. An indication of how it can be addressed has been given already in the proof of Theorem 2.4—in particular, see Figures 2.9 and 2.10.

Lecture 35

a. A remark on tubular neighbourhoods. One of the hypotheses in the statement of the lemma on tubular neighbourhoods was the

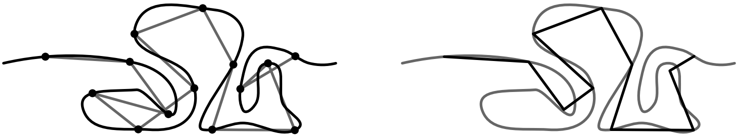


Figure 5.5. Approximating γ with $\bar{\gamma}$.

assumption that the surface in question is orientable. This was used to guarantee the existence of a continuous positive direction along the normal geodesics—recall that a surface is non-orientable precisely if it admits some curve along which no such continuous positive direction can be found.

In order to give a proper answer to the question of which curves on a non-orientable surface S admit tubular neighbourhoods and which do not, we would need to develop an understanding of the fundamental group, which we have not examined here. The key result is that the fundamental group, whose elements may be thought of as closed curves on S (technically, they are homotopy classes of such curves), contains a subgroup of index two with the property that one coset contains all curves which admit tubular neighbourhoods, while the other coset contains all curves which do not admit tubular neighbourhoods. The fact that the tubular neighbourhood lemma only applies to 'half' the curves on a non-orientable surface is related to the difference in the expressions $2 - 2g$ and $2 - g$ for the Euler characteristic of orientable and non-orientable surfaces.

b. Proving the Jordan Curve Theorem. We now present a proof of the Jordan Curve Theorem for arbitrary continuous curves without self-intersections. As mentioned in the previous lecture, the main idea is to approximate the curve with a piecewise linear, or polygonal, curve, for which the result is easier to obtain.

Proof of Theorem 5.3. *Step 1.* Because S^1 is compact, continuity of $\gamma\colon S^1 \to \mathbb{R}^2$ implies uniform continuity. Hence for every $\varepsilon > 0$ there exists $\delta > 0$ such that $|t_1 - t_2| < \delta$ implies $\|\gamma(t_1) - \gamma(t_2)\| < \varepsilon$. Choose N such that $1/N < \delta$, and let $\tilde{\gamma}$ be the piecewise linear curve,

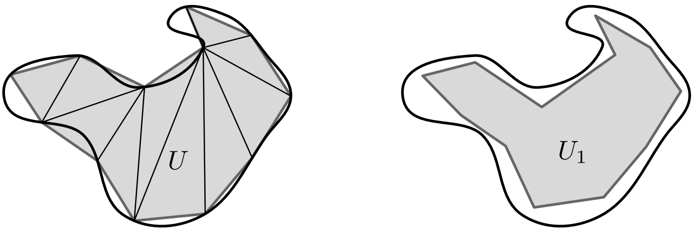


Figure 5.6. Approximating the interior of γ with a polygon.

or polygon, with vertices at $\gamma(k/N)$ for $k = 0, \dots, N$. That is,

$$\tilde{\gamma}(t) = (1-s)\gamma\left(\frac{k}{N}\right) + s\gamma\left(\frac{k+1}{N}\right),$$

where $t = \frac{k+s}{N}$ for $s \in [0,1]$.

Step 2. $\tilde{\gamma}$ may have self-intersections, so we must remove these before we continue. The idea will be to 'chop off' the loops created by these self-intersections, and the key observation is that we can only have $\tilde{\gamma}(t_1) = \tilde{\gamma}(t_2)$ if t_1 and t_2 are close to each other, so that we are not removing much of the curve when we do this. In particular, because γ itself is injective and S^1 is compact, for every $\delta > 0$ there exists $\epsilon > 0$ such that $\|\gamma(t_1) - \gamma(t_2)\| < \epsilon$ implies $|t_1 - t_2| < \delta$. Hence if $\tilde{\gamma}$ approximates γ to within ϵ (which can be accomplished by taking a sufficiently large value of N in Step 1), we can only have $\tilde{\gamma}(t_1) = \tilde{\gamma}(t_2)$ if $|t_1 - t_2| < \delta$.

Now beginning at $t = 0$, let t_1^a be the first parameter value such that $\tilde{\gamma}(t_1^a)$ is a point of self-intersection, and let t_1^b be the largest parameter value such that $\tilde{\gamma}(t_1^b) = \tilde{\gamma}(t_1^a)$. Then $t_1^a < t_1^b < t_1^a + \delta$, and we may similarly find $t_i^a < t_i^b < t_i^a + \delta$ for $i = 2, \dots, n$ such that $\tilde{\gamma}(t_i^a) = \tilde{\gamma}(t_i^b)$, and $\tilde{\gamma}$ has no self-intersections between t_i^b and t_{i+1}^a.

Thus we may define a new approximation, $\bar{\gamma}$, by taking only the pieces of $\tilde{\gamma}$ lying between t_i^b and t_{i+1}^a for $i = 0, \dots, n$ (Figure 5.5). $\bar{\gamma}(S^1)$ still lies in an ε-neighbourhood of $\gamma(S^1)$, and now we may construct a tubular neighbourhood of $\bar{\gamma}(S^1)$ as in the proof of Theorem 5.5, which allows us to use the same argument as in that proof to show that $\mathbb{R}^2 \setminus \bar{\gamma}(S^1)$ has two connected components, U and V. One of these (say U) is bounded, and the other (say V) is unbounded.

Step 3. Since $\bar{\gamma}$ is a polygonal curve, U is the interior of a polygon, and hence can be triangulated (Figure 5.6). Thus it is topologically a disc—there exists a homeomorphism $h\colon D^2 \to \bar{U} = U \cup \bar{\gamma}(S^1)$. Denote by D^2_r the disc with radius r—for $r < 1$, this is D^2 with a neighbourhood of the boundary removed. Take $r < 1$ as large as possible, but small enough that $h(D^2_r) \cap \gamma(S^1) = \emptyset$, that is, that the homeomorphic image of D^2_r under h does not intersect our *original* curve γ. This is possible since $\gamma(S^1)$ lies in an ε-neighbourhood of $\bar{\gamma}(S^1)$. Call this image U_1—then U_1 is a subset of some connected component of $\mathbb{R}^2 \setminus \gamma(S^1)$, and the boundary of U_1 lies near $\gamma(S^1)$.

By choosing a better approximation $\bar{\gamma}$ in the same way and following the same procedure, we may obtain a larger open set $U_2 \supset U_1$ which still lies in a single connected component of $\mathbb{R}^2 \setminus \gamma(S^1)$. Iterating, we obtain a sequence $U_1 \subset U_2 \subset \cdots$ such that every point x in each U_i has non-zero index with respect to γ. Taking the union of all the sets U_i and observing that their boundaries lie within arbitrarily small neighbourhoods of $\gamma(S^1)$, we see that the union U contains every such point, and this is one of our two connected components.

A similar procedure may be carried out for the sets V_i lying outside the curve (if we work on the sphere instead of the plane, the argument is exactly the same for both sides of the curve), and so we obtain a connected open set V which contains all points whose index with respect to γ is zero. This exhausts the possibilities, and so $\mathbb{R}^2 \setminus \gamma(S^1)$ has exactly two connected components. □

With a little care, this can be extended to a proof of Schoenflies Theorem. The key moment comes in step 3, when we are choosing a better refinement $\bar{\gamma}$ and obtaining $U_2, U_3, \dots$. If we proceed carefully and choose a triangulation of U which preserves the triangulation from the previous step, then each successive refinement simply extends the domain of the homeomorphism h, until in the limit the domain is the entire disc, and h is well defined.

The main idea of the above proof of the Jordan Curve Theorem was the fact that for every $\varepsilon > 0$, the set $\mathbb{R}^2 \setminus \bar{B}_\varepsilon(\gamma(S^1))$ has exactly two connected components, which we used to establish our result by letting ε go to zero. It is worth noting that the compactness of S^1 was

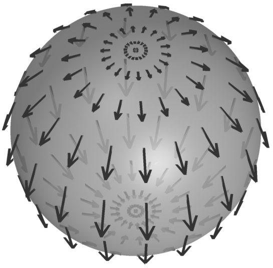


Figure 5.7. The vector field $-\nabla f$ associated with the height function f on a sphere.

crucial to our proof, since it allowed us to establish a uniform bound on how close to self-intersection γ could come for parameter valuess not near each other, and also that we made use of the *geometric* structure of the plane (drawing lines, etc.) even though the result is of a purely *topological* nature.

c. Poincaré-Hopf Index Formula. Consider a compact smooth surface S, and a continuous vector field V on S which has only isolated zeroes. The reader who has some knowledge of ordinary differential equations will notice that this condition on V is too weak to guarantee the existence and uniqueness of integral curves for the vector field, and so we should not use such curves in the proof of the formula we are about to state, which highlights yet another incarnation of the Euler characteristic.

Theorem 5.8 (Poincaré-Hopf)**.** *Under the conditions above, the Euler characteristic is the sum of the indices of the critical points:*

$$\sum_{V(x)=0} \operatorname{ind}_x V = \chi(S). \tag{5.1}$$

We postpone a proof of this result until the next lecture, and content ourselves for the time being with an example. Consider the unit sphere in $\mathbb{R}^3$ with vector field V running along the meridians from the north pole to the south pole, such that the magnitude of the vector at each point (x, y, z) is $\sqrt{x^2 + y^2} = \sqrt{1 - z^2}$, as shown in Figure 5.7. This vanishes at the poles and is non-zero everywhere else—the north pole is a *source*, since the vector field points away from it in all directions, and the south pole is a *sink*, since the vector field points toward it from all directions.

Looking at a neighbourhood of the north pole in coordinates given by projection to the horizontal plane, we see that the vector field is given by $V(x, y) = (x, y)$, and so the associated circle map is the identity, which has degree 1. Thus the index of the north pole is 1.

Following the same approach at the south pole, we have $V(x, y) = (-x, -y)$, so the associated circle map is rotation by π, which also has degree 1, and the index here is 1 as well. Thus the indices sum to 2, which is the Euler characteristic of the sphere.

Lecture 36

a. Proving the Poincaré-Hopf Index Formula. We devote the final lecture in these notes to a proof of the Poincaré-Hopf Index Formula (5.1) and a few corollaries. As before, S is a compact surface, and V is a continuous vector field on S with isolated zeroes.

Proof of Theorem 5.8.

Step 1. It suffices to consider orientable surfaces, as follows; any non-orientable surface S has a standard orientable double cover $\pi\colon \tilde{S} \to S$. We have $\chi(\tilde{S}) = 2\chi(S)$, and V lifts to a vector field $\tilde{V}$ on $\tilde{S}$ with two zeroes for every zero of V, so that the left side of the equation is multiplied by two as well, and thus the formula for the non-orientable surface S will follow from the formula for the orientable surface $\tilde{S}$.

Step 2. One of the standard models for an orientable surface of genus g is as the quotient space of two discs with g holes identified appropriately along boundaries. For example, a disc with one hole is an annulus, or a cylinder, and gluing two cylinders together along their boundaries, we obtain a torus, the orientable surface of genus 1; the corresponding construction for $g = 2$ was illustrated in Figures 3.6 and 3.7.

We decompose S as such a union $D_1 \cup D_2/\sim$, where each D_i is a disc with g holes. By using this two-disc model of our surface, we can now work with vector fields in the plane. The vector field V

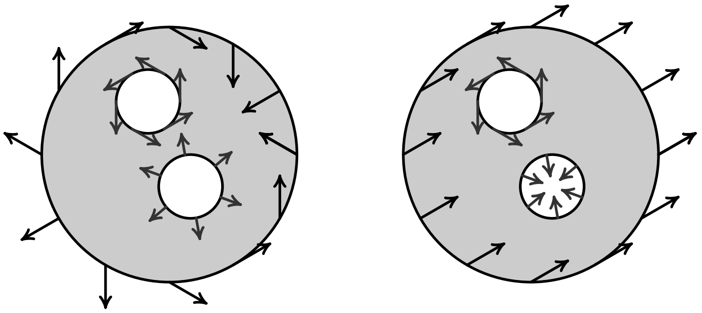


Figure 5.8. The decomposition of a surface of genus 2.

on our surface S passes to vector fields V_j on the two domains D_j, as shown in Figure 5.8. For simplicity of representation, V_j has only been drawn along the boundaries—in fact, it is defined on the entire domain, but we will be particularly interested in V_j on the boundaries of the domain. We denote these curves by $\gamma_0, \gamma_1, \dots, \gamma_g$, where γ_0 is the exterior boundary (the large circle), and $\gamma_1, \dots, \gamma_g$ are the smaller circles.

Technically, since we are interested in vector fields we must use a smooth atlas on S, while the above construction is merely topological. The solution is to extend each disc slightly to include a tubular neighbourhood of γ_i for each i—then instead of gluing along the curves γ_i we glue the two domains together along these 'collars'.

Step 3. Without loss of generality (by moving the boundary components a little, if necessary), we assume our decomposition to be such that all the zeroes of V_1 and V_2 lie in the interior of the two domains D_1 and D_2, so that the vector field is non-vanishing on each curve γ_j. Assign the positive orientation (counterclockwise) to γ_0, and the negative orientation (clockwise) to the other curves $\gamma_1, \dots, \gamma_g$—then we may define the index of the vector field with respect to the composite boundary as

$$\operatorname{ind}_{D_j} V_j = \sum_{i=0}^{g} \operatorname{ind}_{\gamma_i} V_j .$$

It remains to relate this sum to the indices of the zeroes of V, and to relate the values of $\operatorname{ind}_{\gamma_i} V_1$ and $\operatorname{ind}_{\gamma_i} V_2$, since as indicated

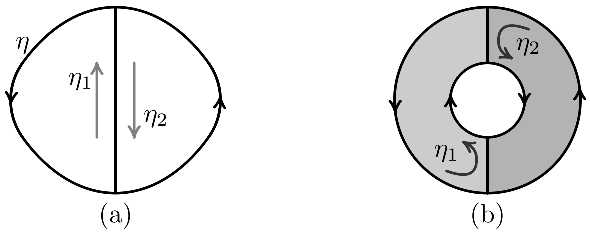


Figure 5.9. Decomposing curves and boundaries.

in Figure 5.8, V_1 and V_2 take different forms along the curves γ_i, which reflects that these domains lead us to view the curve from two different sides (think of the equator on the sphere, with stereographic projection from the poles).

Step 4. In fact, we find that $\operatorname{ind}_{D_j} V_j$ is the sum of the indices of the zeroes contained in D_j:

$$\operatorname{ind}_{D_j} V_j = \sum_{\substack{x \in D \\ V_j(x)=0}} \operatorname{ind}_x V_j.$$

To see this, consider a closed curve η, and decompose η as the composition of η_1 and η_2 (that is, following the first one, then the other) as shown in Figure 5.9(a). Here η is the boundary of the circle, η_2 is the 'D'-shape on the right, and η_1 is the reversed 'D'-shape on the left. If V is any non-vanishing vector field along η, an examination of the associated circle maps shows that $\operatorname{ind}_\eta V = \operatorname{ind}_{\eta_1} V + \operatorname{ind}_{\eta_2} V$.

We may carry out a similar decomposition on our domains D_j—Figure 5.9(b) shows an example of the case $g = 1$. Here γ_0 and γ_1 are as described before, and η_1 and η_2 are the boundaries of the left and right 'C'-shapes, respectively. We see that

$$\operatorname{ind}_D V = \operatorname{ind}_{\gamma_0} V + \operatorname{ind}_{\gamma_1} V = \operatorname{ind}_{\eta_1} V + \operatorname{ind}_{\eta_2} V.$$

By continuing this decomposition until each curve η_i surrounds exactly one zero of V, we obtain the formula claimed at the beginning of this step, and see that

$$\operatorname{ind}_{D_1} V_1 + \operatorname{ind}_{D_2} V_2 = \sum_{V(x)=0} \operatorname{ind}_x V.$$

Thus it only remains to examine the relationship between V_1 and V_2 along each curve γ_i.

Step 5. We claim that for $1 \leq i \leq g$, the indices of V_j along γ_i are related by the formula $\operatorname{ind}_{\gamma_i} V_1 + \operatorname{ind}_{\gamma_i} V_2 = -2$, while for $i = 0$ (the outer boundary in Figure 5.9), the sum is 2. Then summing over all values of i and applying the formula from step 4 will give

$$\sum_{V(x)=0} \operatorname{ind}_x V = 2 - 2g = \chi(S),$$

so it only remains to prove the claim. We see that the difference in sign is due to the different orientation of the curves, so it suffices to consider the exterior boundary γ_0.

In considering the relationship between V_1 and V_2, the example to keep in mind is the equator of the sphere, with tubular neighbourhood given by a small region of the tropics. Then the two vector fields V_1 and V_2 in the plane correspond to the representations of V under stereographic projection from the two poles, and are related by reflection in the line tangent to the circle at the given point, as shown in Figure 5.8.

To formalise this, we parametrise γ_0 by $(x, y) = (\cos\theta, \sin\theta)$, and let $v_j(\theta)$ denote the angle that the vector $V_j(x(\theta), y(\theta))$ makes with the positive x-axis. Then the tangent line to γ_0 at (x, y) makes an angle $\alpha = \theta + \pi/2$ with the horizontal, and reflection in this line is given by the map

$$v \mapsto 2\alpha - v$$

where again, v is the angle a vector makes with the positive x-axis. Because V_1 and V_2 are the images of each other under this reflection, we have

$$v_2(\theta) = 2(\theta + \pi/2) - v_1(\theta),$$

and so we see that

$$v_1(\theta) + v_2(\theta) = 2\theta + \pi.$$

It follows that the circle maps have degrees which sum to 2, and so

$$\operatorname{ind}_{\gamma_0} V_1 + \operatorname{ind}_{\gamma_0} V_2 = 2,$$

which completes our proof. □

One immediate application of the index formula (5.1) is the existence of zeroes for any vector field on a surface with non-zero Euler characteristic.

Corollary 5.4. *Let S be a smooth compact surface other than the torus or the Klein bottle. Then any continuous vector field on S has at least one zero.*

Proof. In order to apply the index formula we need to know that all zeroes are isolated. Then of course there is at least one zero since otherwise the sum in (5.1) is taken over an empty set. If the zeroes are not isolated then the index formula does not apply, but of course this can only happen if there are already a great many zeroes; in fact, infinitely many of them. □

This statement is sharp—a simple example of a non-vanishing vector field on the torus is given by $\frac{\partial}{\partial x}$ in the standard flat model (see also Exercise 3.22), and this projects to the Klein bottle under the covering map $(x, y) \mapsto (x + 1/2, -y)$.

On every other surface, the corollary implies the existence of at least one zero for any vector field, and in fact there always exists a vector field with a single zero which 'absorbs' all the index. This is easiest to see on the projective plane, since rotations have a single fixed point, and so one can take the vector field which generates the family of rotations around a point. If we try to find an example on the sphere by lifting this vector field, the fixed point becomes two fixed points, which must be merged somehow—this leads to one possible solution of Exercise 3.23.

On every orientable surface with negative Euler characteristic one can make the following construction. Take the standard $4n$-gon model with pairs of opposite sides identified by translations. The horizontal vector field (defined just as it was for the torus) is non-zero except for the vertex, where it is discontinuous. To make it continuous, or even smooth, one makes a *time change*, multiplying it by a non-negative smooth function which is positive away from the vertex and which decays quickly to zero near the vertex. A direct calculation using the smooth structure described in Lecture 18(d) confirms that this is indeed a smooth vector field.

Exercise 5.5. Construct a smooth vector field with a single zero on any non-orientable surface with negative Euler characteristic.

b. Gradients and index formula for general functions. With the Poincaré-Hopf Index Formula in hand, we can extend our earlier result connecting Morse functions and Euler characteristic to a result which is valid for any smooth function with isolated critical points, possibly degenerate, by considering the indices of the zeroes of a gradient vector field. One small problem here is that our gradient vector fields so far have been defined locally—although we showed that with this local definition, indices of critical points are independent of the choice of coordinates (Proposition 3.9), we will still need a global definition of a gradient vector field for a function in order to apply formula (5.1).

There is one case, the torus, where the solution is easy—here we can choose local coordinate systems in such a way that the transition maps are translations, and hence derivatives of any function (in particular, the gradient) are defined coherently for all patches.

This fails to generalise to surfaces without an additive structure, however, and the most convenient solution in the general case is also the most elegant, making use of a Riemannian metric. Given a smooth function F on a surface S with Riemannian metric g, one defines the (Riemannian) *gradient* $\nabla_g F$ of F as follows:

At any non-critical point x, there is a unique direction of fastest increase of F—that is, a tangent vector $v \in T_x S$ such that the derivative $D_v F$ of F along v, which measures the rate of increase of F along any parametrised curve tangent to v, is maximal among all derivatives along tangent vectors of unit length. Define

$$\nabla_g F(x) = \begin{cases} D_{v(x)}F \cdot v(x) & \text{if } x \text{ is non-critical,} \\ 0 & \text{if } x \text{ is critical.} \end{cases}$$

Exercise 5.6. Prove that $\nabla_g F$ is a smooth vector field which is orthogonal to the level curves of the function F at all non-critical points.

Notice that the coordinatewise definition of gradient corresponds to this construction for the metric

$$ds^2 = dx^2 + dy^2, \tag{5.2}$$

and so the global gradient on the torus (as defined above) corresponds to this construction for the standard flat metric. Similarly, one can avoid the calculations needed for the solution of Exercise 5.6 and still be able to use the index formula (5.1) by picking a Riemannian metric which has the standard form (5.2) for some local coordinate system near each critical point. This can be done using a partition of unity for an atlas where every critical point belongs to a single chart.

A natural relationship between vector fields and continuous maps is given by the construction of solutions of ordinary differential equations. ODEs can be formulated as vector fields, and solutions of ODEs correspond to integral curves—the flow along these curves is a one-parameter group of continuous maps. Here the parameter is usually thought of as time, and fixing a time interval corresponds to choosing a particular map.

As a brief aside, we will use this idea to demonstrate the usefulness of both the existence and uniqueness result for ODEs and the construction of the gradient vector field for a smooth function by outlining a proof of Lemma 3.19.

Introduce a Riemannian metric on the surface, and consider the gradient vector field for the Morse function f. Notice that for any open set U, any smooth vector field V, and any $t \in \mathbb{R}$, the time-t shift ϕ_t along the orbits of V is a diffeomorphism between U and $\phi_t(U)$. This is essentially a reformulation of existence, uniqueness, and smooth dependence on the initial conditions for the solutions of an ODE. Now *reparametrise* the gradient vector field in such a way that the total time between $f^{-1}(c)$ and $f^{-1}(c')$ along any integral curve becomes constant, say t. This can be done using cutoff functions, and it is at this step that we use the key requirement that the interval (a, b) contains only regular values, since this implies that every integral curve which intersects $f^{-1}(c)$ will reach $f^{-1}(c')$ in finite time. Having done this, the time-t map for the reparametrised gradient flow maps S_c onto S'_c, and we are done.

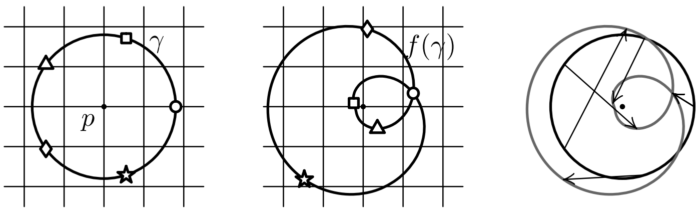


Figure 5.10. Associating a circle map to a continuous map near a fixed point.

c. Fixed points and index formula for maps. We used a degree argument in our proof of Brouwer's fixed point theorem in Lecture 21(c). Since degree and index are closely related, it is not too surprising that a similar result can also be derived from a close relative of the Poincaré-Hopf Index Formula—to achieve this, we must formulate the result in terms of continuous maps rather than vector fields.

We have just seen how to associate a continuous map to a vector field in a natural way, by flowing along its integral curves. In order to define the notion of index for a continuous map, we proceed in the opposite direction, taking a map in the neighbourhood of a fixed point (or more generally, a map close to the identity) and defining a related vector field.

To this end, let S be a surface, $f\colon S \to S$ a continuous map, and p an isolated fixed point. Take local coordinates in a neighbourhood of p, and consider a small closed curve γ around p which has no other fixed point of f in its interior—for example, a circle of small radius with centre at p, as shown in Figure 5.10. We proceed as in Lecture 22, replacing the vector field which was given to us in that case by the vector field $f(x) - x$, which of course depends on the coordinate system. Using this vector field, we define the circle map ϕ_γ as in (3.8); that is, we associate with each parameter $t \in S^1 = [0,1]/\sim$ (or, if you prefer, the point $\exp 2\pi i t \in S^1 \subset \mathbb{C}$) the normalised vector

$$\frac{f(\gamma(t)) - \gamma(t)}{\|f(\gamma(t)) - \gamma(t)\|}.$$

The degree of this map (Figure 5.11) is called *the index of the map f at p*, and is denoted $\operatorname{ind}_p f$. As before, this definition is invariant

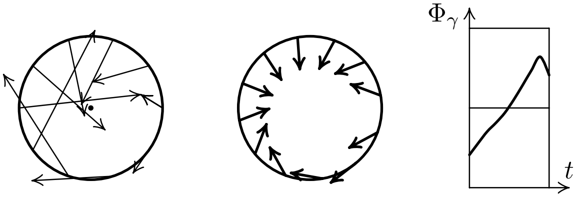


Figure 5.11. The degree of a continuous map at a fixed point.

under continuous changes of the curve γ, as long as γ is moved without touching other fixed points of the map f.

Now we must show that this definition is invariant under smooth coordinate changes, even though these will of course change the map ϕ_γ. First note that for a linear coordinate change, invariance follows from considering the effect of the coordinate change on $f(x) - x$. A non-linear coordinate change is a composition of a linear one (its derivative at p) with a smooth coordinate change whose derivative at p is the identity—the latter, however, changes both the curve γ and the direction of the vector $f(x)-x$ only slightly, hence by Proposition 3.9, the index does not change.

Once again, as for the gradient, this construction can be made global by using the Riemannian metric, at least for any map f which is 'not too far' from the identity. Specifically, it is sufficient that for any $x \in S$ there be a unique shortest geodesic between x and $f(x)$—this is the case, for instance, if the distance between any point and its image is less than the number ε from Proposition 4.12. Note that this number does not have to be small. For example, a sufficient condition on the sphere is that x and $f(x)$ are never diametrically opposite, while $\varepsilon = 1/2$ works on the standard flat torus. On any such surface, a compactness argument shows that the unique shortest geodesic depends continuously on x as long as x is not a fixed point.

Given any map f satisfying this assumption, we may define a vector field X_f as follows—if x is not a fixed point, take the unique shortest geodesic from x to $f(x)$, and let $X_f(x)$ be the tangent vector to this geodesic at x, of length $d(x, f(x))$. If x is a fixed point, put $X_f(x) = 0$. Then it is easy to see that for any fixed point p of the map f we have

$$\operatorname{ind}_p f = \operatorname{ind}_p X_f.$$

This follows from the simple observation that in a coordinate system containing two nearby points x and y, the angle between the line segment from x to y and the tangent vector to the shortest geodesic connecting x and y is small.

Thus as a corollary of the Poincaré-Hopf Index Formula (Theorem 5.8), we obtain the corresponding result for continuous maps:

Theorem 5.9. *Let S be a surface with a Riemannian metric, and $f\colon S \to S$ a continuous map such that for any $x \in S$ there is a unique shortest geodesic between x and $f(x)$. Then*

$$\sum_{x \in S : f(x) = x} \operatorname{ind}_x f = \chi(S). \tag{5.3}$$

As before, an immediate corollary is the existence of fixed points in certain situations.

Corollary 5.5. *Under the assumption of Theorem 5.9, if S is not the torus or the Klein bottle, then the map f has a fixed point.*

For the sphere, where the condition on f reduces to demanding that no point be mapped to its antipode (the point diametrically opposite), and hence also for its factor space the projective plane, we have an even more interesting conclusion.

Corollary 5.6. *Any continuous map $f\colon S^2 \to S^2$ has a fixed point or a point which maps to the point diametrically opposite it.*

Corollary 5.7. *Any continuous map $f\colon \mathbb{R}P^2 \to \mathbb{R}P^2$ has a fixed point.*

d. The ubiquitous Euler characteristic. The Euler characteristic has appeared in many guises throughout this course, from many different sorts of considerations—combinatorial, algebraic, differentiable (smooth functions and ODE), topological, and geometric.

We end these notes by summarising the different ways in which the Euler characteristic has appeared with relation to surfaces. Let S be a compact closed surface which admits a map or, equivalently, a smooth structure (see Lecture 32(c)).[3] Then $\chi(S)$, the Euler characteristic of S, is equal to any of the following:

(1) #(faces)−#(edges)+#(vertices) for any map (in particular, any triangulation) of S;

(2) $\beta_2 - \beta_1 + \beta_0$ where β_i, $i = 0, 1, 2$, are the Betti numbers which arise from the chain complex associated with any triangulation or map of S;

(3) #(maxima) − #(saddles) + #(minima) for any Morse function on S;

(4) The sum of the indices of critical points for any differentiable function on M with finitely many critical points;

(5) The sum of the indices of zeroes of any continuous vector field with finitely many zeroes;

(6) The sum of the indices of fixed points of any continuous map $f\colon S \to S$ with finitely many fixed points which is sufficiently close to (or rather 'not too far from') the identity;

(7) The integral of curvature with respect to any Riemannian metric on S divided by 2π.

The Euler characteristic is a prototypical example of a topological invariant for a manifold which can be expressed through various structures. The situation in higher dimensions is, at first sight, somewhat surprising—on even-dimensional compact manifolds, the Euler characteristic can be defined as the alternating sum of the Betti numbers, just as in (2), and some, but not all, of its guises extend to this case. For orientable odd-dimensional manifolds, however, the sum in (2) vanishes. This renders the obvious generalisation useless, but is itself a manifestation of one of the most remarkable facts in topology—Poincaré duality.

[3] In fact, any surface admits these structures, but we have not proved that in this course.

Suggested Reading

There are many books which cover various aspects of the subjects developed or touched upon in this book. We restrict our selection to a few titles which can be considered classical, or nearly so.

Concurrent reading

Euclidean geometry, projective geometry, and hyperbolic geometry all receive an excellent exposition, primarily from the synthetic point of view, in

> H. S. M. Coxeter, *Introduction to Geometry*, Wiley, New York, 1969,

which we have followed in several of our proofs. Coxeter also introduces curvature and related properties for curves and surfaces following the more traditional exposition via normal and principal curvatures, Christoffel symbols, etc.

For the topological side of things, we recommend

> Andrew H. Wallace, *Differential Topology: First Steps*, Dover, Mineola, NY, 2006,

which requires about the same background as the present book, but has a more narrow scope, allowing it to go considerably further in the direction of differential topology and Morse theory.

A good complement to Wallace's book is the short classic

> John W. Milnor, *Topology from the Differentiable Viewpoint*, Princeton University Press, Princeton, NJ, 1997,

written with fewer details but with excellent insights.

Further reading

The fundamental group, covering spaces, homology and homotopy theory, and other aspects of the subject are covered in

> Allen Hatcher, *Algebraic Topology*, Cambridge, 2001,

which has the added benefit of being available as a free download from the author's web page. This is the standard text on algebraic topology at present.

A good introduction to curvature (as the title implies) and many other topics that we cover in Chapter 4, which explores the attendant definitions and their generalisations to higher dimensions rather more thoroughly than we have done here, may be found in

> John M. Lee, *Riemannian Manifolds: An Introduction to Curvature*, Graduate Texts in Mathematics **176**, Springer, New York, 1997.

A more encyclopedic treatment of these topics, which includes (among other things) an introduction to the calculus of variations and a derivation of the Euler-Lagrange equations, is

> B. A. Dubrovin, A. T. Fomenko, S. P. Novikov, *Modern Geometry—Methods and Applications: Part 1: The Geometry of Surfaces, Transformation Groups, and Fields*, Graduate Texts in Mathematics **93**, Springer, Berlin, 1991.

Discussions of Riemann surfaces, symmetric spaces and Lie theory, Morse theory, higher-dimensional versions of degree and index, algebraic topology (homology and homotopy groups), and much more besides, may be found in the remaining two volumes of that work:

B. A. Dubrovin, A. T. Fomenko, S. P. Novikov, *Modern Geometry—Methods and Applications: Part 2: The Geometry and Topology of Manifolds*, Graduate Texts in Mathematics **104**, Springer, Berlin, 1985.

B. A. Dubrovin, A. T. Fomenko, S. P. Novikov, *Modern Geometry—Methods and Applications: Part 3: Introduction to Homology Theory*, Graduate Texts in Mathematics **124**, Springer, Berlin, 1990.

The reader is warned that unlike the other books quoted above, the standards of rigor in these last three books are less uniform, and sometimes rather lax. Hence one should not expect to understand all the proofs just by following the text, without consulting other sources or making a considerable mental effort.

Background reading

The basic concepts of point set topology (open, closed, compact, connected, etc.), along with the Inverse Function Theorem, the Implicit Function Theorem, and a host of other basic results and techniques, may be found in any of the following:

Jerrold E. Marsden and Michael J. Hoffman, *Elementary Classical Analysis*, W.H. Freeman, New York, 1993.

Charles C. Pugh, *Real Mathematical Analysis*, Springer-Verlag, New York, 2002.

Halsey L. Royden, *Real Analysis*, Macmillan, New York, 1988.

Walter Rudin, *Principles of Mathematical Analysis*, McGraw-Hill, New York-Auckland-Düsseldorf, 1976.

More details regarding holomorphic functions, conformal mappings, fractional linear transformations with arbitrary complex coefficients, and other results from complex analysis may be found in either of these two books:

> Jerrold E. Marsden and Michael J. Hoffman, *Basic Complex Analysis*, W.H. Freeman, New York, 1999.
>
> Walter Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1987.

The existence and uniqueness theorems for ODEs referred to in the text, and other related results, can be found in a convenient form in

> Vladimir I. Arnold, *Ordinary Differential Equations*, Springer-Verlag, Berlin, 1992.

Jordan normal form, of which the classification of 2×2 matrices in Proposition 4.16 is a special case, along with any other concepts from linear algebra of which the reader may need reminding, is presented in

> Kenneth Hoffman and Ray Kunze, *Linear Algebra*, Prentice Hall, Upper Saddle River, NJ, 1971.

Hints

Chapter 1

1.3. Visualise how the square can be bent or folded to make the identifications specified.

1.4. Modify the isometries in the definition of $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$.

1.5. Consider the intersection of the generating curve with the axis of revolution.

1.6. Find a function of two variables whose zero set is the intersection of the two-handled sphere with one of its planes of symmetry.

1.7. This will turn out to be, in some sense, a reflection in the equator.

1.8. Write a parametric equation for the original circular cross-section, then add a second parameter to incorporate revolution.

1.9. This time begin with a parametrised line segment, which will both rotate (around its centre) and revolve (around the z-axis) as the second parameter varies.

1.11. The length of a planar curve is given by integrating the arc length $ds^2 = dx^2 + dy^2$. Begin by finding parametric coordinates $(x(u,v), y(u,v), z(u,v))$ on the cylinder and cone in which the arc length of a curve on the surface is $ds^2 = dx^2 + dy^2 + dz^2 = du^2 + dv^2$, and then use the fact that geodesics in the plane are straight lines.

1.13. Construct a self-intersecting projective plane as the image of a map $f\colon S^2 \to \mathbb{R}^3$ such that $f(-x,-y,-z) = f(x,y,z)$, and then compose f with the usual parametric representation of the sphere. Follow a similar procedure in $\mathbb{R}^4$, using the extra coordinate to avoid self-intersections.

1.15. To show that this is a torus, cut and paste to turn the hexagon into a parallelogram or a rectangle. To show that it is not isometric to the standard flat torus, find some property which is invariant under isometries but which differs for the two surfaces; consider properties related to diameter, area, geodesics, etc.

1.16. Consider separately the four types of isometries, and do this in conjunction with the next exercise.

1.18. Look for fixed points of this isometry or of its composition with some translation.

1.19. Consider the lifts of x, y, and any geodesics connecting them, to the sphere.

1.20. Decompose it into triangles.

1.21. Each of these is $\mathbb{R}^2$ modulo some group of isometries $\Gamma \subset \mathrm{Iso}(\mathbb{R}^2)$, and so each point $x \in \mathbb{R}^2$ is identified with its orbit under Γ. Each isometry of the quotient space will lift to an isometry of $\mathbb{R}^2$, which must map orbits to orbits in order to be well defined. (Algebraically, it must lie in the normaliser of Γ.)

1.22. For (3), observe that translations of $\mathbb{R}$ preserve equivalence classes.

1.23. Lifting to $\mathbb{R}^2$, recall that an isometry with a fixed point is a rotation or a reflection. The latter has order 2. For rotation by θ, use the fact that the lifted isometry maps L to L. One way is to consider the matrix representation, and obtain restrictions on θ by showing that the trace of the matrix must be an integer. Another way uses discreteness of L; argue that if L contains a regular n-gon for $\theta = 2\pi/n$, then it contains arbitrarily small n-gons unless $n = 2$, 3, 4, or 6.

Chapter 2

2.1. (a) Given x, show that the set of points z for which there is a path from x to z is both open and closed. (b) For example, begin with the graph of $\sin(1/x)$.

2.2. Find a fundamental domain and determine the edge identifications.

2.3. Find a neighbourhood of each point homeomorphic to $\mathbb{R}^2$. Consider separately points within faces, points on edges, and vertices.

2.4. By counting edges between faces and edges between vertices, establish that $3F = 2E = \sum$ degree of vertices $\leq V(V-1)$. Use the fact that $\chi = 0$ to obtain $V \geq 7$, then find a triangulation which achieves this bound, using the planar model on either the square or the hexagon.

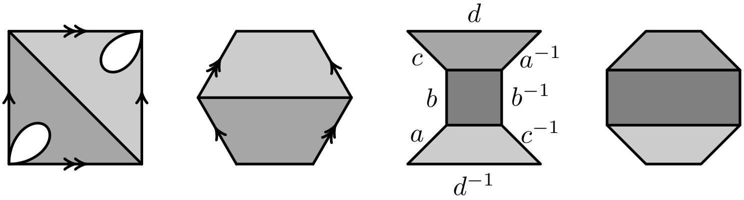
2.5. The argument from the previous exercise may be modified to give a potential lower bound.

2.6. Find a covering space whose Euler characteristic is known.

2.7. Cut and paste with wild abandon.

2.8. By relabeling two adjacent adges of the hexagon as a single edge, it is sometimes possible to reduce to a planar model on the square; for example, writing $d = bc$, we see that $abcac^{-1}b^{-1} = adad^{-1}$ is a Klein bottle.

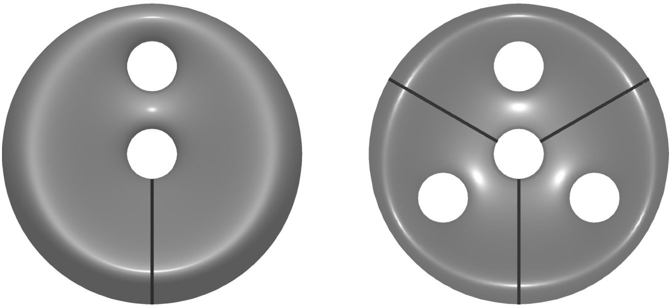
2.9. Given a regular $2n$-gon with opposite sides identified by translations, cut two holes at opposite vertices to obtain a $2n+2$-gon with one pair of opposite sides left free. Finish adding a handle by gluing a cylinder between these sides; cut and paste to obtain a $2n+4$-gon with opposite sides identified. Induction does the rest.



2.10. Follow the centre circle and see what happens.

2.11. What does it mean to choose a positive direction of rotation at a point within a face? on an edge? at a vertex?

2.12.



2.13. Use the standard planar model for a sphere with m Möbius caps as a $2m$-gon with identifications $a_1a_1\dots a_ma_m$, and choose a nice triangulation. What happens if you choose a 'pseudo-triangulation'—for example, by drawing lines from the centre of the $2m$-gon to each vertex? This fails to be a triangulation because all vertices of the $2m$-gon are identified.

Chapter 3

3.1. Show that the atlas which comes from the usual parametrisation of the torus of revolution is compatible with the atlas which comes from the Implicit Function Theorem via the embedding of the torus in $\mathbb{R}^3$ as a level set.

3.2. Fix an orientation on one patch and show that this determines orientations on all the others (assuming the surface is connected) in a coherent fashion.

3.3. Write down the formulae of the transition maps.

3.4. Find a smooth change of coordinates $\phi\colon (x,y,z) \to (\tilde{x},\tilde{y},\tilde{z})$ such that $\tilde{F}(\tilde{x},\tilde{y},\tilde{z}) := F\circ\phi^{-1}(\tilde{x},\tilde{y},\tilde{z}) = \tilde{x}^2+\tilde{y}^2+\tilde{z}^2$.

3.5. Use a fractional linear transformation $z \mapsto (az+b)/(cz+d)$ to go from the open disc to the upper half-plane, and use various elementary functions to obtain the strips, perhaps via the first quadrant if necessary. For the open square, one needs an *elliptic integral*, which is related to the *Schwarz-Christoffel formula* (see Marsden and Hoffman's book for details). For the ellipse, consider the *Zhukovsky map* $z \mapsto z + 1/z$.

3.6. Find a holomorphic map which takes any two given points to 0 and ∞.

3.7. Consider transformations of the form $z \mapsto (az + b)/(cz + d)$.

3.8. Given a holomorphic equivalence, use the fact that its lift to $\mathbb{C}$ is conformal.

3.9. Rotations and homotheties are holomorphic equivalences.

3.10. Given z near p, consider $w(z) = f(z) - f(p)$.

3.11. Compute Df in a suitable local coordinate system.

3.12. If we write $r(x, y)$ for the remainder term of order $o(x^2 + y^2)$, then the level set is $\{(x, y) \mid x^2 - y^2 + r(x, y) = 0\}$.

3.13. Find the necessary coordinate change along the level sets, and extend it smoothly to a neighbourhood.

3.14. Along the curve $y = 0$, the origin is a local minimum, so on some neighbourhood of p, this curve intersects each curve $f = c$ exactly twice, once for $x > 0$ and once for $x < 0$. Parametrising these curves by arc length, find a smooth map which takes each $f = c$ to $x'y' = c$.

3.15. For each angle θ, f is at first strictly increasing as we move away from p along the ray which makes an angle θ with the positive x-axis. Let $g(\theta)$ be its value at the first point when it is no longer strictly increasing along the ray. Show that g is continuous and use compactness of S^1.

3.16. Recall Exercise 1.6.

3.18. Try perturbing it with a quadratic polynomial.

3.19. (1) Argue as in Exercise 3.12. (2) Perturb the function as in the example.

3.20. The lift of $f \circ g$ is $F \circ G$.

3.21. x is a fixed point of f iff the lift F has $F(x) = x + k$ for some $k \in \mathbb{Z}$. Draw a graph.

3.22. Consider constant vector fields.

3.23. Via stereographic projection, the sphere is the plane with a point at infinity. Construct a non-vanishing vector field on the plane and pull it back to the sphere.

3.24. Find a vector field whose associated circle map lifts to the map $x \mapsto kx$, $k \in \mathbb{Z}$.

Chapter 4

4.1. Given two tangent vectors in the coordinate plane, find their preimages as tangent vectors to the embedding in $\mathbb{R}^3$, and use the usual Euclidean inner product.

4.3. First put such a metric on each patch, and then use a partition of unity to build a metric on the whole surface.

4.5. Use the law of cosines again.

4.7–8. Take p and q as in Proposition 4.12 to lie on a plane of symmetry, and use the uniqueness result.

4.9. Use the previous exercise.

4.11. Use Proposition 4.11.

4.12. Prove that this formula is invariant under isometries, and then show that it holds when z_1 and z_2 have the same real part.

4.13. Perform a change of coordinates to determine what form these maps take in the upper half-plane model.

4.14. Show that the vertical line passing through the Euclidean centre passes through the hyperbolic centre as well.

4.15. Argue as in Exercise 4.13.

4.16. One way is to do it first for a particular choice of z_1 and z_2, and then use the fact that isometries take horocycles to horocycles.

4.17. The centre of a (Euclidean) circle containing two points must lie on their perpendicular bisector, and any such circle is either a hyperbolic circle, a geodesic, a horocycle, or an r-equidistant curve for some r. Investigate how r varies as the centre of the circle moves along the bisector.

4.18. Find the points which lie on the ideal boundary, and find the geodesic connecting those two points.

4.19. Proposition 4.16 partitions $SL(2,\mathbb{R})$ into conjugacy classes; a normal subgroup must be the union of some or all of these classes. Show that each non-trivial conjugacy class generates the whole group.

4.20. Look for points or curves which are invariant under the product of the reflections.

4.21. Recall that a fractional linear transformation is determined by its action on three points.

4.22. Take one of the infinite vertices to be at infinity, so that two of the sides are vertical lines.

4.23. Note that, as for the three standard examples, g depends only on r and is independent of θ, so that an expression for the circumference of a circle of radius r centred at $(0, 0, \phi(0))$ will yield an expression for $g(r)$.

4.24. Integrate the formula (4.19) for the circumference.

4.25. Write the points on the (geodesic) circle of radius r as the points on a (Euclidean) circle of radius r in the tangent plane plus a small error term, and show that $\ell(r) = 2\pi r + O(r^4)$.

4.27. Straight lines through the point (x, y, z) may be parametrised as $(x, y, z) + t(a, b, c)$, where a, b, c are fixed and t is the parameter. Find conditions on a, b, c to guarantee that points of this form lie on $\mathcal{H}$.

4.28. Proposition 4.18 may be generalised to geodesic polygons, including quadrilaterals.

Chapter 5

5.1. Recall Exercise 3.20.

5.2. Modify the statement and proof of the lemma on tubular neighbourhoods to accomodate the case where γ intersects itself. The proof in the text still shows that Γ is a *local* diffeomorphism.

5.3. In the tubular neighbourhood lemma, Γ will be diffeomorphic on each smooth segment of the curve, but the images will overlap near the corners. Modify Γ so that it is a homeomorphism.

5.4. Take a very fine triangulation so that on each triangle, γ is nearly a straight line. Then define h piecewise as a change of coordinates on each triangle.

5.5. Use the orientable double cover (modifying one of the standard planar models from Exercise 2.9) and follow the construction in the text.

5.6. Work in local coordinates.

Index

Titles in This Series

46 **Anatole Katok and Vaughn Climenhaga,** Lectures on surfaces: (Almost) everything you wanted to know about them, 2008

45 **Harold M. Edwards,** Higher arithmetic: An algorithmic introduction to number theory, 2008

44 **Yitzhak Katznelson and Yonatan R. Katznelson,** A (terse) introduction to linear algebra, 2008

43 **Ilka Agricola and Thomas Friedrich,** Elementary geometry, 2008

42 **C. E. Silva,** Invitation to ergodic theory, 2007

41 **Gary L. Mullen and Carl Mummert,** Finite fields and applications, 2007

40 **Deguang Han, Keri Kornelson, David Larson, and Eric Weber,** Frames for undergraduates, 2007

39 **Alex Iosevich,** A view from the top: Analysis, combinatorics and number theory, 2007

38 **B. Fristedt, N. Jain, and N. Krylov,** Filtering and prediction: A primer, 2007

37 **Svetlana Katok,** p-adic analysis compared with real, 2007

36 **Mara D. Neusel,** Invariant theory, 2007

35 **Jörg Bewersdorff,** Galois theory for beginners: A historical perspective, 2006

34 **Bruce C. Berndt,** Number theory in the spirit of Ramanujan, 2006

33 **Rekha R. Thomas,** Lectures in geometric combinatorics, 2006

32 **Sheldon Katz,** Enumerative geometry and string theory, 2006

31 **John McCleary,** A first course in topology: Continuity and dimension, 2006

30 **Serge Tabachnikov,** Geometry and billiards, 2005

29 **Kristopher Tapp,** Matrix groups for undergraduates, 2005

28 **Emmanuel Lesigne,** Heads or tails: An introduction to limit theorems in probability, 2005

27 **Reinhard Illner, C. Sean Bohun, Samantha McCollum, and Thea van Roode,** Mathematical modelling: A case studies approach, 2005

26 **Robert Hardt, Editor,** Six themes on variation, 2004

25 **S. V. Duzhin and B. D. Chebotarevsky,** Transformation groups for beginners, 2004

24 **Bruce M. Landman and Aaron Robertson,** Ramsey theory on the integers, 2004

23 **S. K. Lando,** Lectures on generating functions, 2003

22 **Andreas Arvanitoyeorgos,** An introduction to Lie groups and the geometry of homogeneous spaces, 2003

21 **W. J. Kaczor and M. T. Nowak,** Problems in mathematical analysis III: Integration, 2003

20 **Klaus Hulek,** Elementary algebraic geometry, 2003

TITLES IN THIS SERIES

19 **A. Shen and N. K. Vereshchagin,** Computable functions, 2003

18 **V. V. Yaschenko, Editor,** Cryptography: An introduction, 2002

17 **A. Shen and N. K. Vereshchagin,** Basic set theory, 2002

16 **Wolfgang Kühnel,** Differential geometry: curves – surfaces – manifolds, second edition, 2006

15 **Gerd Fischer,** Plane algebraic curves, 2001

14 **V. A. Vassiliev,** Introduction to topology, 2001

13 **Frederick J. Almgren, Jr.,** Plateau's problem: An invitation to varifold geometry, 2001

12 **W. J. Kaczor and M. T. Nowak,** Problems in mathematical analysis II: Continuity and differentiation, 2001

11 **Mike Mesterton-Gibbons,** An introduction to game-theoretic modelling, 2000

10 **John Oprea,** The mathematics of soap films: Explorations with Maple®, 2000

9 **David E. Blair,** Inversion theory and conformal mapping, 2000

8 **Edward B. Burger,** Exploring the number jungle: A journey into diophantine analysis, 2000

7 **Judy L. Walker,** Codes and curves, 2000

6 **Gérald Tenenbaum and Michel Mendès France,** The prime numbers and their distribution, 2000

5 **Alexander Mehlmann,** The game's afoot! Game theory in myth and paradox, 2000

4 **W. J. Kaczor and M. T. Nowak,** Problems in mathematical analysis I: Real numbers, sequences and series, 2000

3 **Roger Knobel,** An introduction to the mathematical theory of waves, 2000

2 **Gregory F. Lawler and Lester N. Coyle,** Lectures on contemporary probability, 1999

1 **Charles Radin,** Miles of tiles, 1999

Photo courtesy of S. Katok

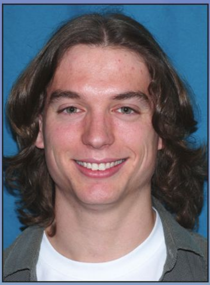


Photo courtesy of Eric M. Sauder

Surfaces are among the most common and easily visualized mathematical objects, and their study brings into focus fundamental ideas, concepts, and methods from geometry, topology, complex analysis, Morse theory, and group theory. At the same time, many of those notions appear in a technically simpler and more graphic form than in their general "natural" settings.

The first, primarily expository, chapter introduces many of the principal actors—the round sphere, flat torus, Möbius strip, Klein bottle, elliptic plane, etc.—as well as various methods of describing surfaces, beginning with the traditional representation by equations in three-dimensional space, proceeding to parametric representation, and also introducing the less intuitive, but central for our purposes, representation as factor spaces. It concludes with a preliminary discussion of the metric geometry of surfaces, and the associated isometry groups. Subsequent chapters introduce fundamental mathematical structures—topological, combinatorial (piecewise-linear), smooth, Riemannian (metric), and complex—in the specific context of surfaces.

The focal point of the book is the Euler characteristic, which appears in many different guises and ties together concepts from combinatorics, algebraic topology, Morse theory, ordinary differential equations, and Riemannian geometry. The repeated appearance of the Euler characteristic provides both a unifying theme and a powerful illustration of the notion of an invariant in all those theories.

The assumed background is the standard calculus sequence, some linear algebra, and rudiments of ODE and real analysis. All notions are introduced and discussed, and virtually all results proved, based on this background.

This book is a result of the MASS course in geometry in the fall semester of 2007.

ISBN 978-0-8218-4679-7

9 780821 846797

STML/46

For additional information and updates on this book, visit

www.ams.org/bookpages/stml-46