Surfaces: (almost) everything you wanted to know about them

lectures by Anatole Katok

notes taken and TeXed by Vaughn Climenhaga

Contents

CHAPTER 1

Various Ways of representing Surfaces

1.1. Lecture 2: Wednesday, Aug. 29

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

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

To obtain the surface of revolution, we introduce y into the equation by making the substitution $x \mapsto \sqrt{x^2 + y^2}$, and obtain

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

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 $\mathcal{C}^1$ 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

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

It can be checked that this new choice of F does not introduce any extraneous points to the solution set, and now F is $\mathcal{C}^1$ on all of $\mathbb{R}^3$.

This argument cannot be applied literally to the homework problem N3 (see optional exercise below) but the general method can be used; namely one should start from the intersection of a sphere with two handles with a horizontal plane which may look e.g. as the union of three circles and build the rest of the surface symmetrically above and below.

Exercise 1. Prove that a sphere with $m \geq 2$ handles cannot be represented as a surface of revolution.

What good is all this? What benefit do we gain from representing the torus, or any other surface, by an equation? To answer this, we first backtrack a bit and discuss graphs of functions. In particular, given a function

$f : \mathbb{R}^2 \to \mathbb{R}$, the graph of f is

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

Assuming 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, contains 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 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

$$\begin{aligned} f_1(x, y) &= \sqrt{1 - x^2 - y^2} \\ f_2(x, y) &= -\sqrt{1 - x^2 - y^2} \end{aligned}$$

The graph of f_1 is the northern hemisphere, the graph of f_2 the southern. However, we run into problems at the equator $z = 0$, where the derivatives of f_1 and f_2 become infinite; we permit ourselves to use only functions with bounded derivatives. 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.

This approach has wide validity; if we have any surface S given as the zero set of a function $F : \mathbb{R}^3 \to \mathbb{R}$, we say that a point on S is *regular* if the gradient of F is nonzero at that point. At any regular point, then, 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 comprises only regular points, this allows us to describe the entire surface in terms of these local coordinates.

Recall that if the gradient (or, equivalently all partial derivatives) vanish at at point, the set of solutions may not look like a nice surface: a trivial example is the sphere or radius zero $x^2 + y^2 + z^2 = 0$; a more interesting one is the cone

$$x^2 + y^2 - z^2 = 1$$

near the origin.

b. Other ways of introducing local coordinates. From the geometric point of view, however, this choice of planes is somewhat arbitrary and unnatural; for example, projecting the northern hemisphere of S^2 to the $x - y$ coordinate 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 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.

Projections form a subset of the class of parametric representations of surfaces, but there are many other members of this class besides. In the case of a sphere, one well-known example is stereographic projection, which gives a homeomorphism between the plane and the sphere minus a point. Another example is given by the familiar system of longitude and latitude to locate points on the surface of the earth; these give a sort of polar coordinates, mapping the sphere minus a point onto the open disk. The north pole is the centre of the disk, while the (deleted) south pole is its boundary; lines of longitude become radii of the disk, while lines of latitude become concentric circles around the origin.

However if we want to measure distances on the sphere in these coordinates we cannot use usual Euclidean distance in the disc. This gives us our first example of a *Riemannian metric* on $\mathbb{D}^2$, that is, a notion of distance, apart from the usual Euclidean one. Along lines of longitude (radii), distance is preserved; however, distances along lines of latitude (circles) are distorted, particularly as we approach the boundary of the disk, where the actual distance between points is much less than the Euclidean distance (since every point on the boundary is identified).

c. Metrics on surfaces. We must now, of course, 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, we also have a formula; the distance between two points is simply the angle they make with the centre of the sphere. (Properties of the 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).

This is not, of course, how we typically measure distance along the surface of the earth; we are far more likely to simply count how many steps it takes to get from point A to point B, or something along those lines. This is more in line with the definition of distance we make for an arbitrary surface; the distance between two points is simply the length of the shortest path connecting them, or since we do not know yet whether such a shortest path exists, the infimum of lengths among all paths connecting the two points. 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 in a later lecture.

1.2. Lecture 3: Friday, Aug. 31

a. Remarks concerning the problem set. *Problem 5.* The most common way of introducing the Möbius strip 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 following equivalent model:

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 opposed 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 $0 \leq t \leq 1$.

Problem 6. A geodesic is a curve γ with the property that given any two points $\gamma(a)$ and $\gamma(b)$ which are sufficiently close together, any other curve from a to b will have length at least as great as the portion of γ from a to b. The question of whether such a curve always exists for sufficiently close points, and whether it is unique, will be dealt with in a later lecture.

Problem 7. The projective plane will be one of the star exhibits of this course. We can motivate its definition by considering the sphere as a geometric object, with the notion of a line in Euclidean space replaced by the concept of a geodesic, in this case a great circle. This turns out to have some not-so-nice features, however; 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, 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 opposed 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. Alternatively, we can consider the map $I : (x,y,z) \mapsto (-x,-y,-z)$, which is the only isometry of the sphere without any fixed points. Considering all members of a particular orbit of I to be the same point, we obtain the quotient space S^2/I, which is again the projective plane.

In the projective 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 central angle they make, and the notions of angle between geodesics and length of geodesics are still well-defined.

b. 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. The unit circle lying in the $x - y$ plane is now the solution set of

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

This is a simple example, for which these equations 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

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

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

$$\begin{aligned} x &= \cos t \\ y &= \sin t \\ z &= 0 \end{aligned}$$

with $0 \leq t \leq 2\pi$. If we replace the third equation with $z = t$, we obtain not a circle, but a helix; the implicit representation of this curve as the solution set of two equations is rather more complicated. We could also multiply the expressions for x and y by t to describe a spiral on the cone, which would be similarly difficult to write implicitly.

Exercise 2. 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, we must impose the further requirement that this tangent vector be nonvanishing, that is, that $(f_1')^2 + (f_2')^2 + (f_3')^2 \neq 0$ holds everywhere on the curve. This requirement is sufficient, but not necessary, in order to guarantee smoothness of the curve.

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 in the picture appears as a *cusp* at the origin. So in this case, even though f_1 and f_2 are perfectly smooth functions, the curve itself is not smooth.

To see that the nonvanishing condition is not necessary, consider the curve $x = t^3$, $y = t^3$. At $t = 0$ the tangent vector vanishes, but the curve

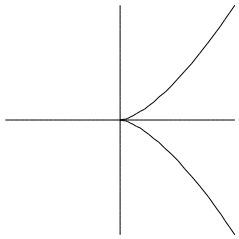

FIGURE 1. The curve $(x, y) = (t^2, t^3)$

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 nonvanishing.

c. Other means of representation. 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 $\mathbb{R}^3$? The projective plane, for example, cannot be embedded in $\mathbb{R}^3$, so if we are to compute the length of curves in the projective 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. We shall return to this later.

It is also possible for a particular 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 sorts of points on the surface; a given point can lie either at a vertex, along an edge, or on a face. From our point of view as three-dimensional beings, these are three distinct classes of points.

Now imagine that we are two-dimensional creatures wandering around the surface of the dodecahedron. We can tell whether or not we are at a vertex, because the angles will add up to less than 2π, whereas elsewhere they add up to exactly 2π. However, we cannot tell whether or not we are at an edge; 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, draw the line between the two points in question, and then fold the surface back up. As far as our two-dimensional selves are concerned, points on an edge and points on a face are indistinguishable.

An example of a surface which is most easily represented without reference to a particular embedding is the *Klein bottle*, which is a *factor space* of the square, or rectangle. As with the Möbius strip, the left and right edges are identified with direction reversed, but in addition, the top and bottom edges are identified (without reversing direction). The Klein bottle cannot be embedded into $\mathbb{R}^3$; 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.

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 into $\mathbb{R}^4$. This allows for the four-dimensional analogue of lifting a section of string off the surface of a table in order to avoid having it touch a line drawn on the table. No such manoeuvre is possible 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 American Mathematical Society. My two models were bought there.

We have already mentioned that the projective plane cannot be embedded in $\mathbb{R}^3$; even surfaces which can be so embedded, such as the torus, lose some of their nicer properties in the process. The usual embedding of the torus destroys the symmetry between meridians and parallels; all of the meridians are the same size, but the size 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 $s, t \in [0, 2\pi]$. We can also obtain the 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 $T_{m,n} : (x, y) \mapsto (x + m, y + n)$, 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 $\mathbb{R}/\mathbb{Z}$.

The flat torus along with the projective planes will be one of our star exhibits throughout this course.

d. Regularity conditions for parametric surfaces. A parametrisation of a surface 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 $(x, y, z) = (f_1(t, s), f_2(t, s), f_3(t, s))$.

EXERCISE 3. Give parametrisations of the sphere (using geographic coordinates) and of the torus (viewed as a surface of revolution in $\mathbb{R}^3$).

Just as in the case of parametric representations of curves, we need a regularity condition to ensure that our surface is in fact smooth, without cusps or singularities. As before, we require that the functions f_i be continuously differentiable, but now it is insufficient to simply require that the matrix of derivatives Df be nonzero. 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 Implicit Function Theorem guarantees that parametric representation is locally bijective and its inverse is differentiable.

Projections discussed in the previous lecture are particular cases of *inverses* for parametric representations.

1.3. Lecture 4: Wednesday, Sept. 5

a. Review of metric spaces and topology. In any geometry, be it Euclidean or non-Euclidean, projective or hyperbolic, or any other sort we may care to describe, the essential concept is that of *distance*. 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, $d : X \times X \to \mathbb{R}_0^+$, which satisfies the following axioms (these must hold 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: $d(x, z) \leq d(x, y) + d(y, z)$

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

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

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.

EXERCISE 4. Fill in the details in the above argument.

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 equation $d(x, z) = d(x, y)+d(y, z)$ whenever y lies between two points x and z sufficiently close along the curve.

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

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

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 $f : X \to Y$ 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 5. 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 sort 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 semi-simple 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 sort of mathematical object, 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 $f : X \to \mathbb{R}$ 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, $X = A \cup B$, A and B open, $A \cap B = \emptyset$ 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 $f : [0, 1] \to X$ 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. Isometries. In the topological context, the natural notion of equivalence between two spaces is that of homeomorphism, which preserves all the topological features of a space. A stronger definition is necessary for

metric spaces, in which not only the topology of open and closed sets, but the distance function from which that topology comes, must be preserved.

A map $f : X \to Y$ 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 f is an *isometry*, otherwise it is what is known as an *isometric embedding*.

We are particularly interested in the set of isometries from X to itself,

$$\mathrm{Iso}(X, d) = \{f : X \to X | f \text{ is an isometry}\}$$

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

In fact, $\mathrm{Iso}(X, d)$ is not just a set; it has a natural binary operation given by composition, under which is becomes a group. This is a very natural and general sort of group to consider; all the bijections of some fixed set, 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.

1.4. Lecture 5: Friday, Sept. 7

a. Issues relating to a lecture by A. Kirillov. Given a complete metric space, the *Hausdorff metric* defines a distance function on the collection of compact subsets, which in turn gives a notion of convergence of compact sets. This proves to be useful when trying to give a proper mathematical definition of a fractal, which also leads to various definitions of the dimension of a set.

We usually think of dimension as a topological invariant; for example, two Euclidean spaces $\mathbb{R}^m$ and $\mathbb{R}^n$ are homeomorphic if and only if they have the same dimension. However, the dimension defined for general compact metric spaces is a metric invariant, rather than a topological one. The main idea is to capture the rate at which volume, or some sort of measure, grows 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 $C(\varepsilon)$ be the minimum number of ε-balls required to cover X; that is, the minimum number of points $x_1, \ldots, x_{C(\varepsilon)}$ in X such that every point in X lies within ε of some x_i. Then the *upper box dimension* of X is

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

We take the upper limit because the limit itself may not exist. Lower box dimension is defined similarly with the lower limit. These notions are sufficient for the study of fractals but they certainly fails for sets such as rational numbers: the upper (and lower as well) box dimension of this set is equal to one!

There is a more effective notion of *Hausdorff dimension* where there is in particular no need to distinguish between upper and lower limits and which is equal to zero for any countable set. For 'good' sets all three definitions coincide; the definition of Hausdorff dimension requires an understanding of measure theory, so we will not discuss it here.

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

If we fix three noncollinear 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 the intersection of two circles around distinct centres contains no more than two points, and all points equidistant from these two points lie on the line defined by the centres of the circles.

As a consequence of this, an isometry of $\mathbb{R}^2$ is completely determined by its action on three noncollinear points. In fact, if we have an isometry $I : \mathbb{R}^2 \to \mathbb{R}^2$, and three such points x, y, z, 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 corresponds to the case where I preserves orientation, the other to the case where orientation is reversed. So for two pairs of distinct points $a,\ b$ and $a',\ b'$ such that the (Euclidean) 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 orientation preserving, and one orientation reversing.

Passing to algebraic descriptions notice that isometries must carry lines to lines, and hence are affine maps, so any isometry I may be written as $I : x \mapsto Ax + b$, 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) $I : z \mapsto az + b$, where $a, b \in \mathbb{C}$ and $|a| = 1$. In the orientation reversing case, we have $I : z \mapsto a\bar{z}+b$.

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*. We can take x to be the fixed point, so $Ix = x$. 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.

Rotations are entirely determined by the centre of rotation and the angle of rotation, so we require three parameters to specify a rotation.

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 $I : z \mapsto z + b$ is a translation.

One can also make a purely synthetic argument for this case. Namely, we will show that, unless the image of every interval is parallel to it, there is a fixed point. Let (a, b) be an interval. If the interval (Ia, Ib) is not parallel to (a, b) the perpendiculars to the midpoints of those intervals intersect in a single point c. But then I must map the triangle $a\,b\,c$ into the triangle $Ia\,Ib\,Ic$ and hence c is a fixed point. Thus (Ia, Ib) is parallel and equal to (a, b), hence (a, Ia) is parallel and equal to (b, Ib) and 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 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 translation*. 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 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 $\mathrm{Iso}(\mathbb{R}^2)$ 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:

$$\mathrm{Iso}(\mathbb{R}^2) = \left\{ \begin{bmatrix} O(2) & \mathbb{R}^2 \\ 0 & 1 \end{bmatrix} : \begin{bmatrix} \mathbb{R}^2 \\ 1 \end{bmatrix} \to \begin{bmatrix} \mathbb{R}^2 \\ 1 \end{bmatrix} \right\}.$$

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

c. Isometries of the sphere and the projective plane. We will see now that the picture for the sphere is somewhat similar to that for the Euclidean plane: *any* orientation preserving isometry has a fixed point, and most orientation reversing ones have none. But for the projective plane 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 and every orientation reversing isometry other that reflection in a great circle has a point of period two which becomes a fixed point when we pass to the projective plane.[1]

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 always itself a rotation. This is slightly different from the case with $\mathrm{Iso}(\mathbb{R}^2)$, 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 $\mathrm{Iso}(S^2)$ as a group of 3×3 matrices in a much more natural way than we did for $\mathrm{Iso}(\mathbb{R}^2)$ above since any isometry of S^2 extends to a linear orthogonal map of $\mathbb{R}^3$ and we will be able to use linear algebra directly.

1.5. Lecture 6: Monday, Sept. 10

a. 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, formulas 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 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 area on the sphere enclosed by three geodesics (great

[1] At the lecture an erroneous statement was made at this point.

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' 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, 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 compute the area A of the triangle as follows:

$$\begin{aligned} 4(\alpha+\beta+\gamma)R^2 &= 4\pi R^2 + 4A \\ A &= R^2(\alpha+\beta+\gamma-\pi) \end{aligned}$$

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. 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*.

b. Isometries of the sphere. 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 $\mathrm{Iso}(\mathbb{R}^2)$. The first is the *synthetic* approach, which treats the problem using the tools of solid geometry; this is the approach used by Euclid and the other ancient Greek geometers in developing spherical geometry for use in astronomy.

The second approach, which we will follow below, is to use methods of linear algebra; transferring 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 matrices were only 2×2.

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

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

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 $(a_i, a_j) = \delta_{ij}$, where $(\cdot, \cdot)$ denotes inner product, and δ_{ij} is the Kronecker delta. 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 look at the images of three points not all lying on the same geodesic, as we did with $\mathrm{Iso}(\mathbb{R}^2)$; in particular, 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 $I : S^2 \to S^2$, we can define a linear map $A : \mathbb{R}^3 \to \mathbb{R}^3$ 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:

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

This is a useful trick to remember, allowing 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 projective 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, that is, $\mathbb{R}P^2$.

As an example of how all isometries become rotations in $\mathbb{R}P^2$, consider the map A given by reflection through the $x-y$ 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.

c. 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 tangent vectors. That is to say, if v is a 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 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 later in these lectures. It is quite remarkable that the

isometry group of the hyperbolic plane allows not one but three natural representations as a matrix group (or a factor of such group by its two-element center) In fact these four examples exhaust all surfaces for which isometries act transitively on tangent vectors.

There are of course higher–dimensional spaces with this property: euclidean spaces, spheres and projective spaces immediately come to mind. But there are many more of those.

As an example of a space for which this property fails, consider the flat torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$. However it holds locally The properties mentioned above hold locally, in the neighbourhood of a point. However, while $\mathrm{Iso}(\mathbb{T}^2)$ 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. Thus the flat torus is an example of a *locally symmetric space*. Another example is a cylinder; more examples are discussed in the homework problem set N3.

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 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.

1.6. Lecture 7: Wednesday, Sept. 12

a. Symmetric spaces. Given a point x on a surface X, we can define the *geodesic flip* through x as the map I_x which sends each point y on a geodesic γ through x to the point lying on γ which is the same 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.

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 neighborhoods 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 neighborhoods of those points. In other words, locally such a space looks the same near every point.

If 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 $\mathrm{Iso}(X)$ acts transitively on all of X.

In the previous lecture we discussed a related, but stronger, notion, in which we require $\mathrm{Iso}(X)$ to act transitively not only on points in X, but on 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 $\mathbb{R}^4$, which is given as the zero set of the two equations

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

b. 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):

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

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 in 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 $X \times Y$.

The direct product provides a common means by which we decompose objects of interest into simpler examples in order to gain a complete understanding. A good example is the case of 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 abelian groups, where we decompose the group of interest into p-groups until 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*.

c. Topology and combinatorial structure on surfaces. Let X be a topological space. To avoid pathological cases, assume that X is *metrisable*, that is, it is possible to place a metric d on X which induces the given topology. Note that there are in general 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.

We say that X as above is a *manifold* if for every point $x \in X$, there exists some open neighbourhood U_x containing x which is homeomorphic to $\mathbb{R}^n$; that is, there exists a homeomorphism $\phi_x : U_x \to \mathbb{R}^n$.

Thus a manifold is a topological space which locally looks like Euclidean space. 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.

Exercise 6. Show that given X satisfying the above conditions, connectedness implies path connectedness. (This is not true for arbitrary metric spaces).

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 $\phi_x : U_x \to \mathbb{R}^m$ and $\phi_y : U_y \to \mathbb{R}^n$, we can find a path from x to y in M. Then we can find points $x_1, \ldots, x_k$ along the path such that $x_1 = x$, $x_k = y$, $\phi_{x_i} : U_{x_i} \to \mathbb{R}^{n_i}$ 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 does give an example of a continuous map from $\mathbb{R}$ onto $\mathbb{R}^2$. This cannot be made into a homeomorphism, however; indeed, if $f : \mathbb{R} \to \mathbb{R}^2$ is a homeomorphism, then $f : \mathbb{R}\backslash\{0\} \to \mathbb{R}^2\backslash\{f(0)\}$ 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 discuss this theorem later.

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.

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 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 next time, 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 obtained by gluing along the edges. The strategy will be first to classify all surfaces which can be obtained that way or, equivalently, all surfaces which allow such a combinatorial structure, and, second, to show that *every* surface is like that.

The first part, which will occupy us primarily, is fun, including combinatorics and algebra and providing a good set of tools dealing with various examples and questions. The second involves hard general topology starting from Jordan curve Theorem and involving subtle approximation constructions. Fortunately, once this is established it can be taken for granted.

CHAPTER 2

Combinatorial Structure and Topological Classification of Surfaces

2.1. Lecture 8: Friday, Sept. 14

a. Triangulation. Because non-compact surfaces can be extremely complicated, we will fix our attention for the next while upon 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 characterizing 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 structure, a triangulation, which avoids local complexity by building the surface up from simple pieces.

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

$$\sigma^n = \left\{ (x_0, \ldots, x_n) \in \mathbb{R}^{n+1} : x_i \geq 0 \ \forall i, \ \sum_{i=0}^{n} x_i = 1 \right\}$$

We also use σ^n to denote any homeomorphic image of the standard n-simplex along with the barycentric coordinates *$(x_0, \ldots, x_n)$, 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 $t_1A + t_2B + t_3C$, 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 $t_1A + t_2B + t_3C$ gives the location of the center of mass of the triangle.

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.

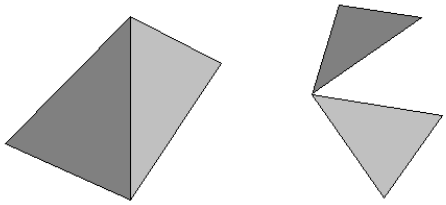

FIGURE 1. Properly attached 2-simplices

Figure 1 gives examples of properly attached simplices; the requirement that the simplices be properly attached forbids arrangments such as those in figure 2.

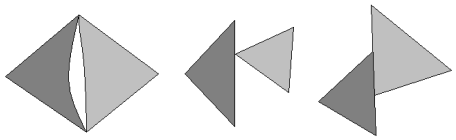

FIGURE 2. Improperly attached 2-simplices

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. *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 $\emptyset$.*

(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 $\sigma_{i_1}^2, \ldots, \sigma_{i_k}^2$ 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 configurations seen in figure 3, 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.

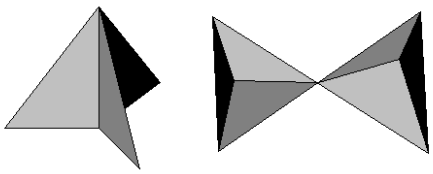

FIGURE 3. Forbidden configurations

EXERCISE 7. 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 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.

EXERCISE 8. Find a triangulation of the torus which uses the fewest possible simplices.

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.

b. Euler Characteristic.

DEFINITION 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*

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

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 1. *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, which inherit their barycentric coordinates from an appropriate scaling of the coordinates on σ^2.

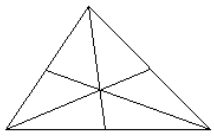

FIGURE 4. Barycentric subdivision of a 2-simplex

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 Euler characteristic is preserved by barycentric subdivision.

2.2. Lecture 9: Monday, Sept. 17

a. Continuation of the proof of Theorem 1. 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 1. *Any polygon can be triangulated.*

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.

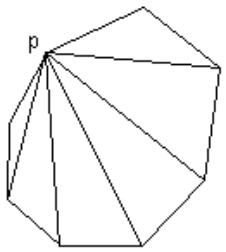

FIGURE 5. Triangulating a convex polygon

If the polygon P is nonconvex, 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 in the convex case, it may no longer happen that q can be taken to be adjacent to a neighbour of p.

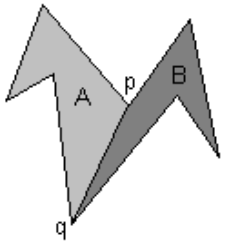

FIGURE 6. Triangulating a non-convex polygon

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 4. *Two triangulations $\mathcal{T}_1$ and $\mathcal{T}_2$ are* affinely equivalent *if there exists a bijection $f : \mathcal{T}_1 \to \mathcal{T}_2$ 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. *Any triangulated polygon is affinely equivalent to a convex triangulated polygon.*

Proof: Again 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.

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.

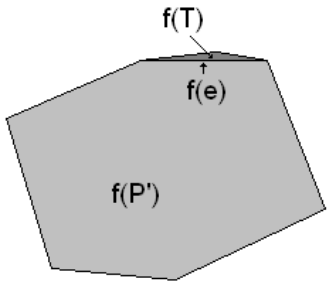

FIGURE 7. Obtaining a convex triangulation

□

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 5. *A* map[1] *of a surface S is a partition of S into properly attached polygons. 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, such as the placement of vertices in the middle of edges.

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

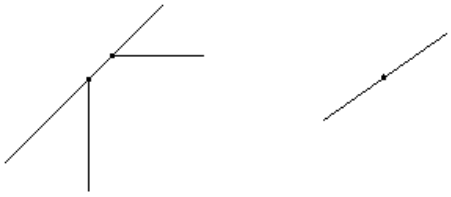

FIGURE 8. Permissible configurations for maps

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 9.

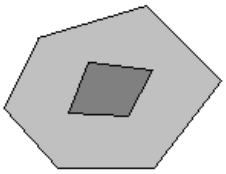

FIGURE 9. A forbidden configuration, even for a map

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 $\chi(\mathcal{T}_2) = \chi(\mathcal{M}) = \chi(\mathcal{T}_1)$.

Because $\mathcal{T}_2$ contains a finite number of vertices, we can find $\varepsilon_1 > 0$ such that the distance between any two vertices is at least $2\varepsilon_1$. Further, we can take θ to be the measure of the smallest angle in any triangle in $\mathcal{T}_2$, and set $\varepsilon_2 = \theta\varepsilon_1$.

[1]As in "geographic map"

These requirements guarantee that if we take B_i to be the ε_1-ball around the i^{th} vertex, and T_i to be the ε_2-'tube' around the i^{th} edge, as indicated in figure 10, then $B_i \cap B_j = \emptyset$ for $i \neq j$, and similarly for T_i.

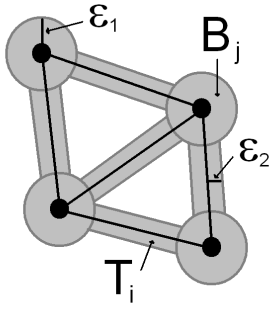

FIGURE 10. The neighbourhood within which we will approximate $\mathcal{T}_2$

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 11. For sufficiently large n, the diameter of $\mathcal{T}_1^n$ 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 $\mathcal{T}_2$.

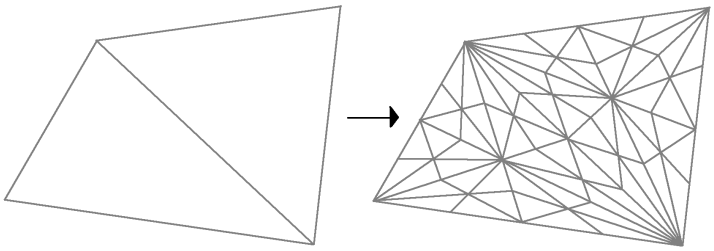

FIGURE 11. A refinement of $\mathcal{T}_1$

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. Although all the pictures are drawn with e appearing as a straight line, the proof is made for the case where e is any continuous curve $\gamma : [0,1] \to S$ with $\gamma(0) = v_1$ and $\gamma(1) = v_2$.

As the parameter t 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.

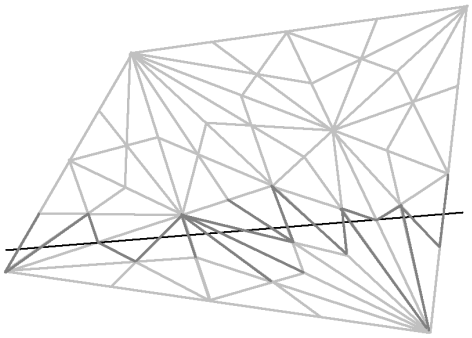

FIGURE 12. Determining a sequence of edges in the mesh

Now the sequence $x_0, x_1, \ldots, x_N$ 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 $e_1, \ldots, e_N$, each of which shares an endpoint with each of its neighbours. In order to obtain a proper path, we must eliminate two sorts of configurations, which may be thought of as fans and loops; the former are illustrated in figure 13.

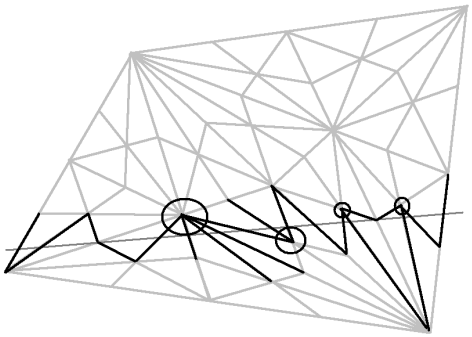

FIGURE 13. Eliminating 'fans' from the sequence of edges

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.

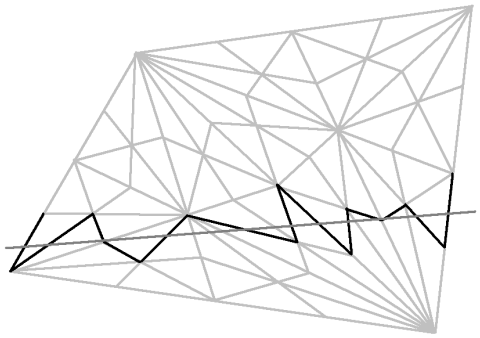

FIGURE 14. A true path along the mesh

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

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

□

The sort of technical drudgery involved in the above proof is common in *point set topology*. In *algebraic topology* one often is able to bypass considerations of this type by considering a coarser equivalence relation than

homeomorphism, namely that of *homotopy equivalence*. Homotopy equivalence makes no distinction between, for example, the unit disc and a single point set, 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 fundamental invariants studied in algebraic topology, first of of all homotopy and homology groups, are indeed invariants of homotopy equivalence, other important topological invariants such as dimension are not. 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 in hand, we can now state unequivocally that $\chi(S^2) = 2$.

Consider the given triangulation of the torus. 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 $\chi = V - E + F = 9 - 27 + 18 = 0$.

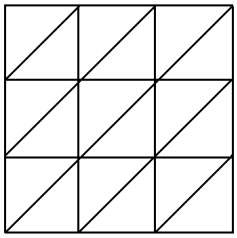

FIGURE 15. A triangulation of the torus

EXERCISE 9. 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 $\chi(\mathbb{R}P^2) = 1$.

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 $\chi(\mathbb{T}^2) = 0$.

2.3. Lecture 10: Wednesday, Sept. 19

a. From triangulations to maps. Given two surfaces M_1 and M_2 equipped with triangulations $\mathcal{T}_1$ and $\mathcal{T}_2$, if $f : M_1 \to M_2$ 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 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 admits a triangulation. This is in fact the case, but for the time being we will defer the proof of the result which requires both techniques of point set topology at a higher level than used in the proof of Theorem 1 and considerable combinatorial ingenuity.

Rather, we shall turn our attention from triangulations to maps, which we introduced briefly in the proof of Theorem 1. We will see, in particular, that the proof given there really shows a 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 may be thought of, for instance, as the regular n-gons lying in the complex plane $\mathbb{C}$ with vertices at the n^{th} roots of unity $\exp(2\pi ik/n)$, $1 \leq k \leq n$. As was mentioned before, we allow the case $n = 2$, which may be thought of as the unit circle $\{z \in \mathbb{C} : |z| = 1\}$ 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 $n = 1$, which may be thought of as having 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 $f : S_n \to M$ 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 6. *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 $f_i : S_{n_i} \to P_i$, 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. *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.*

Proof: Proceed exactly as in the proof of Theorem 1, with $\mathcal{T}_1 = \mathcal{T}$ and $\mathcal{T}_2 = \mathcal{M}$, noting 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 16, 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 the image of S_4 under a function identifying opposite faces, as in figure 17. The usual parametric embedding of the flat torus in $\mathbb{R}^4$ would give a concrete realisation of this map.

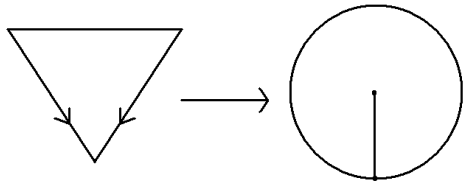

FIGURE 16. A map with a 'spike'

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

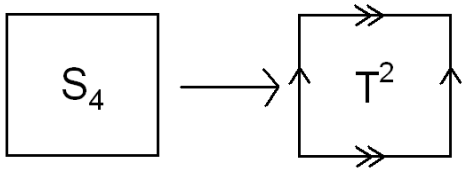

FIGURE 17. A map of the torus using only a single face

THEOREM 3. *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 1, 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 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.

Notice also that unlike the case of surfaces, not every higher-dimensional topological manifold admits a simplicial decomposition. Existence of such manifolds, which defies straightforward intuition, is among the most striking results of topology.

We also note that all our consideartions 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 $\mathbb{R}^2_+ = \{(x, y) \in \mathbb{R}^2 : y \geq 0\}$. Such surfaces may be usefully thought of as compact surfaces without boundary which have had holes removed.

b. Examples. We now know from Corollary 1 that we can classify compact triangulable surfaces by examining the quotient spaces of 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?

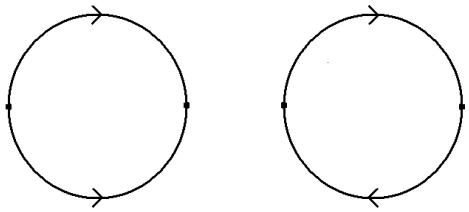

FIGURE 18. The two possible models on S_2

The two cases are shown in figure 18. In the first case, we have the quotient space of $D^2 \subset \mathbb{C}$ 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 projection to the equatorial plane we obtain the disc with antipodal boundary points identified, which is the picture in figure 18. Note that the vertices ± 1 are identified, so the model has $V = E = F = 1$ and hence $\chi = 1$, as in 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, there are two possible orientations, which we may think of as forward and backward, so given a

choice of how to pair off 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 19. 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.

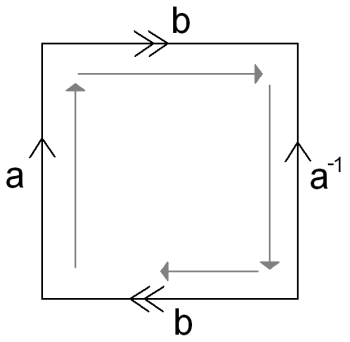

FIGURE 19. Notation for models on S_4

Thus we would write figure 19, 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

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

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 19 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 Euler characteristic uniquely determines the surface.

EXERCISE 10. 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.

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.

2.4. Lecture 11: Friday, Sept. 21

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 last time:

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 one end of C to each hole. 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 will be homeomorphic to a disc with two holes, the so-called 'pair of pants' surface. Gluing one end of C to each hole, we obtain a torus with a hole; 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.

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 20, we cut the hole in a corner, so that the torus with a hole has a planar model on a pentagon.

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 21 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 $F = 1$, so the net result of this process is to increase the number of edges by two, and hence to decrease the Euler characteristic by two.

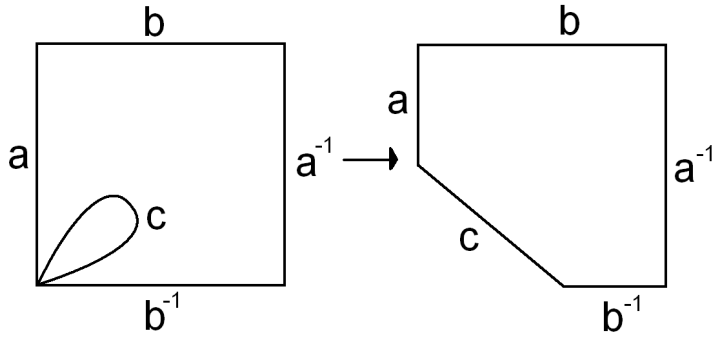

FIGURE 20. Cutting a hole in a torus

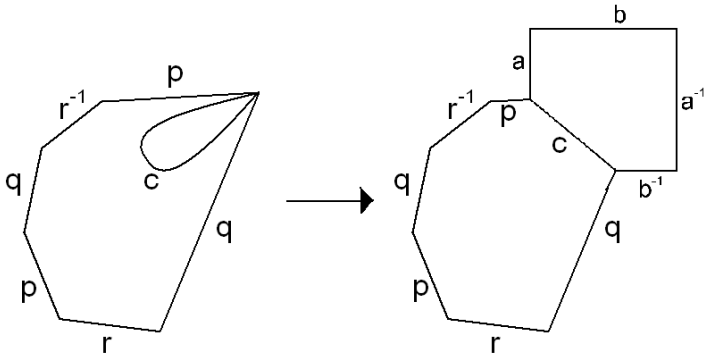

FIGURE 21. Attaching a handle to a planar model

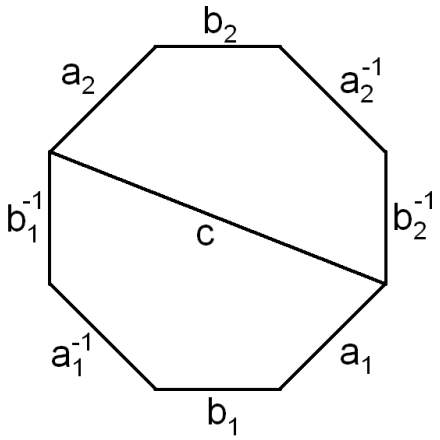

FIGURE 22. A sphere with two handles

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 22); 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*.

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 study of differentiable manifolds, rather than topological ones. Orientability is in fact a topological invariant, and so we proceed as we did for 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 23.

DEFINITION 7. *A triangulation $\mathcal{T}$ of a surface M is* orientable *if its 2-simplices admit a coherent collection of orientations.*

EXERCISE 11. Show that no triangulation of the Möbius strip is orientable.

THEOREM 4. *Given two triangulations $\mathcal{T}_1$ and $\mathcal{T}_2$ of a surface M, $\mathcal{T}_1$ is orientable if and only if $\mathcal{T}_2$ is orientable.*

Proof: Left as an exercise. The key points are that orientability is inherited by barycentric subdivision, and that it is preserved under coarsening.

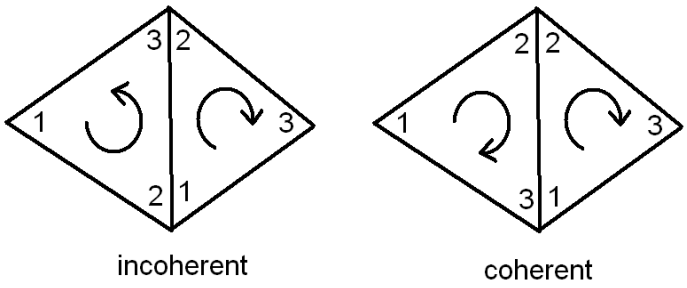

FIGURE 23. Coherent and incoherent orientations of 2-simplices

This last requires a definition of orientation for maps, rather than simply triangulations; the proof proceeds along the lines of Theorem 1. □

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.

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

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 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 Euler characteristic by two. Since no 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.

2.5. Lecture 12: Monday, Sept. 24

a. Non-orientable surfaces and Möbius caps. As indicated in figure 24, 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.

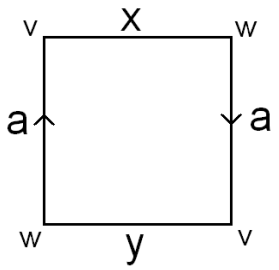

FIGURE 24. A planar model of the Möbius strip

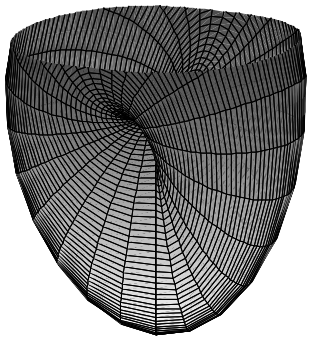

FIGURE 25. A cross cap; the Möbius strip immersed in $\mathbb{R}^3$ with self-intersection

If we immerse the Möbius strip in $\mathbb{R}^3$ as shown in figure 25, then the boundary circle xy 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.
(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 last time. The reason for this will shortly be made clear; first we ask what effect each of these has on the Euler characteristic χ. Let us see what each does to a map of the surface.

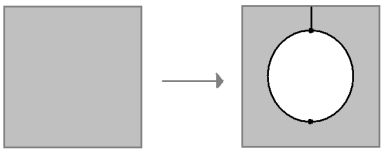

FIGURE 26. Cutting a hole in a surface

As shown in figure 26, cutting a single hole has the effect of adding two vertices, three edges, and leaving F constant, 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 preserves 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 24), 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 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 Euler characteristic by 2.

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 27 suggests a proof that each of aa and bb is equivalent to attaching a Möbius cap. Then the attachment shown in figure 28 may be seen to be equivalent to attaching two Möbius caps.

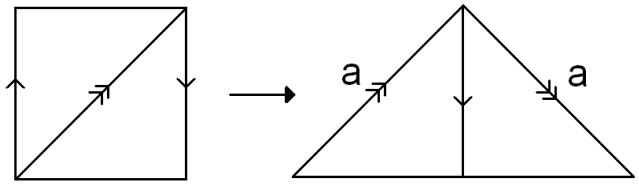

FIGURE 27. Another planar model of the Möbius strip

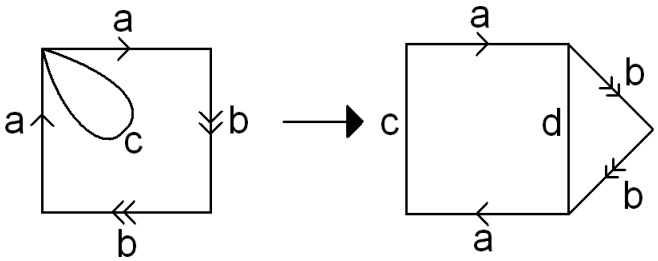

FIGURE 28. Attaching an inverted handle

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 Euler characteristic.

c. Covering non-orientable surfaces.

DEFINITION 8. *A (finite)* covering space *of a surface S is a connected surface $\tilde{S}$ together with a map $f : \tilde{S} \to S$ 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 $f^{-1}(U_x) = \cup_{i=1}^{n} V_i$, 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 . There are also *infinite covering spaces* where pre-image of a neighborhood of any point consists of a countable collection of homeomorphic images of the neighborhood. 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.

Going back to finite coverings notice that 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 $\tau : (x, y) \mapsto (x + 2, y)$. Then the quotient space of X by the action of τ is the cylinder; that is, we identify each point with all of its images under iterates of τ. Then a square root of τ is given by $\sigma : (x, y) \mapsto (x + 1, -y)$, and the quotient space of X by the action of σ is the Möbius strip. The covering map arises naturally as the canonical projection

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

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 1. *Every non-orientable surface has an orientable double cover.*

Proof: We follow an approach which is of wide applicability 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 the points of $\tilde{S}$ to be points of S along with a particular orientation at that point.

Locally, this looks like taking the direct product $S \times \{\pm 1\}$, where each point in S appears twice in $\tilde{S}$, once with a positive orientation and once with a negative orientation. Of course, this is not true globally, for precisely the reason that 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 $f : \tilde{S} \to S$. 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 $\gamma : [0,1] \to S$ 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. □

d. Classification of orientable surfaces.

THEOREM 5. *Given a compact, closed (without boundary), orientable surface M which has a map, 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 3, our surface admits a map with a single face, so we can consider a model on a $2n$-gon.

Next we 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 will show how to modify our map by making it have a single vertex. This will also decrease the number of edges to preserve the Euler characteristic.

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.

Now 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.

2.6. Lecture 13: Wednesday, Sept. 26

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 $a_1b_1a_1^{-1}b_1^{-1}\dots a_mb_ma_m^{-1}b_m^{-1}$.

By Theorem 3, we may take the map on S to consist of one 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 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 9. *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 10. *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 29 illustrates a common construction in algebraic topology; we consider a graph G, a maximal tree T (the lighter 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.

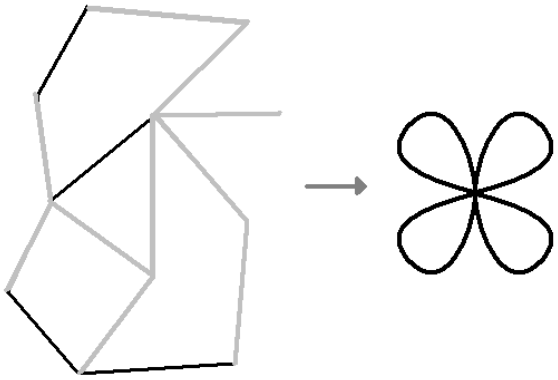

FIGURE 29. Collapsing a maximal tree

LEMMA 3. *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.*

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 30, take every other edge (besides e) which is attached to v, and attach it instead to w. This is a homeomorphism, and results in v being a spike in the map, which may then be eliminated.

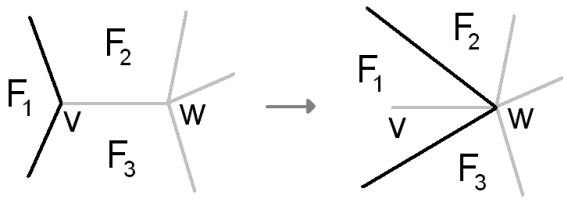

FIGURE 30. Turning a leaf into a spike

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 appears twice (as $\ldots a \ldots a \ldots$, rather than $\ldots a \ldots a^{-1} \ldots$, since a is a closed curve it may be seen that moving a 'clock' once around this curve reverses its orientation. Since S is orientable, this cannot happen, so 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 $\mathrm{dist}(a)$; 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, in the sequence aa^{-1}, we have $\mathrm{dist}(a) = 0$, and in $abcb^{-1}a^{-1}$, $\mathrm{dist}(a) = 2$.

Now choose a such that $\mathrm{dist}(a)$ is minimal; that is, $\mathrm{dist}(a) \leq \mathrm{dist}(b)$ for any other symbol b. Because we have eliminated spikes, we have $\mathrm{dist}(a) \geq 1$, so some symbol b lies between a and a^{-1}. Further, since $\mathrm{dist}(a)$ 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 31. The region labeled H is homeomorphic to a torus with a hole, as shown in figure 32.

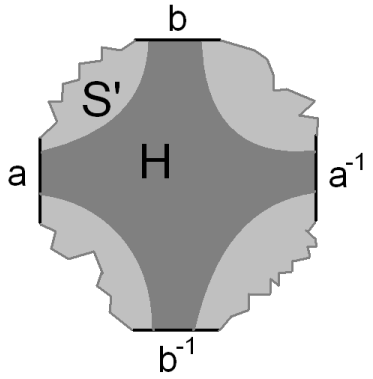

FIGURE 31. The configuration $aba^{-1}b^{-1}$

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 32 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 the case of higher dimensions, a complete classification along these lines is much more difficult to accomplish; indeed, one of the great achievements in recent mathematics has been an essential completion of the classification of 3-manifolds, which was achieved by Perelman's proof

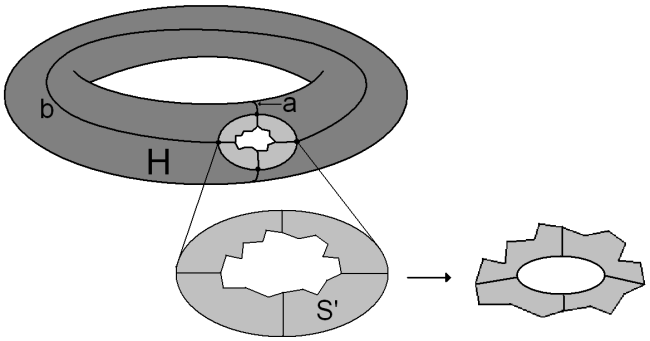

FIGURE 32. A visualisation of H and S'

of the Thurston geometrization conjecture which in particular 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, but are sometimes useful for understanding various geometric structures which will appear later in this course. For example, for this model identifications can be effected by parallel translations and thus one obtains a Euclidean structure at the surface everywhere except for vertices which become"super-conic" points with the total angle being a multiple of 2π.

EXAMPLE 1. For $m = 2$ all eight vertices of the regular octagon are identified producing a sphere with two handles and with the Euclidean structure everywhere except of the single point where the total angle is 6π. Later in this course we will see this as a particular limit case of the celebrated Gauss–Bonnet Theorem.

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 33. 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'.

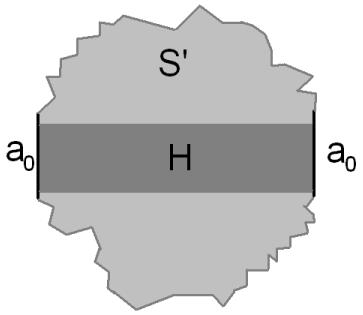

FIGURE 33. The configuration a_0a_0

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

$$a_1a_1a_2a_2\dots a_nb_n$$

which is 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.

2.7. Lecture 14: Friday, Sept. 28

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 11. *A* chain complex $\mathcal{C}$ *is a sequence of finite dimensional linear spaces* C_k *over some field (or more generally, modules over a ring) with linear maps* $\partial_k : C_k \to C_{k-1}$*, 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 12. *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. *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 $\{d, p_1^{k_1}, \ldots, p_n^{k_n}\}$ 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. Naturally then in applications the $\mathbb{Z}$-homology provides more information that 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 we will postpone 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, 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 $C_1(\mathcal{T})$? 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 34. 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.

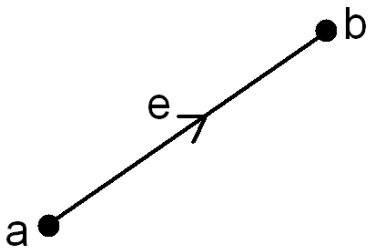

FIGURE 34. The boundary operator on an edge

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 $\partial_2 : C_2(\mathcal{T}) \to C_1(\mathcal{T})$. 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 35, 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 35, for example, we have

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

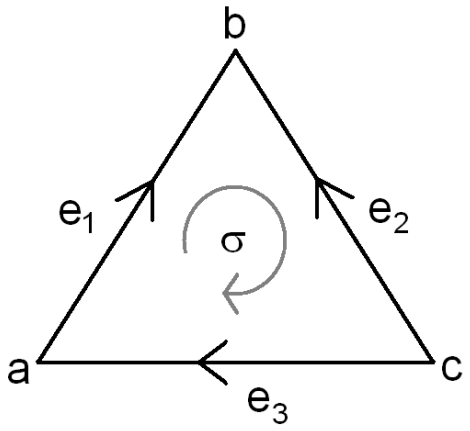

FIGURE 35. The boundary operator on an face

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 36. 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 $e_3 + e_4 \in C_1(\mathcal{T})$.

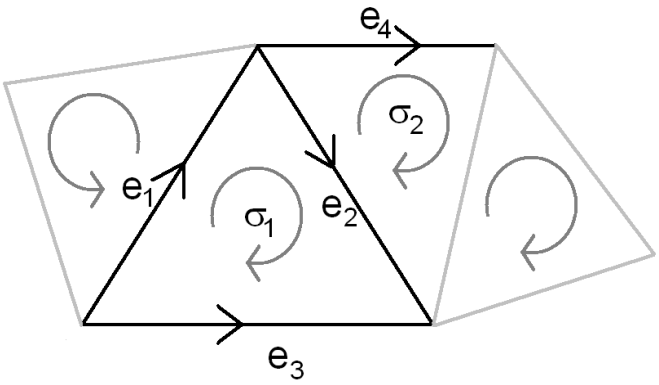

FIGURE 36. A chain of edges in $C_1(\mathcal{T})$

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 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 proceed 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 triangulation we obtain as a corollary that the alternate sum of Betti number does not depend of the triangulation either. as we will see soon the same applies to each Betti number separately.

d. 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 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 $n-1$ nonzero 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 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 what 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 13. *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 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 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 1; 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 calculation with maps which have few vertices and faces are much less cumbersome that those with triangulations.

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

e. Torsion in the first homology and non-orientable surfaces. The above treatment has glossed over some subtle points that arise for non-orientable surfaces. On the projective plane, for instance, we have $\beta_1 = 0$, suggesting that every cycle of edges may be the boundary of a chain of faces. This is however not true literally and here the difference between $\mathbb{R}$-homology and the richer $\mathbb{Z}$-homology shows up. 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. This is reflected in the fact the the the first homology group of projective plane is the group of two elements. Its presence is not noticed by the Betti number.

2.8. Lecture 15: Monday, Oct. 1

a. Alternate method for deriving 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 5, 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

$$H_0 = \ker \partial_0 / \mathrm{Im}\, \partial_1 = \mathbb{R}/\{0\} = \mathbb{R}$$

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. 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}, so $\ker \partial_2 = \mathbb{R} = C_2$, hence

$$H_2 = \ker \partial_2 / \mathrm{Im}\, \partial_3 = \mathbb{R}/\{0\} = \mathbb{R}$$

and we have $\beta_2 = 1$. In this case $\partial_2 = 0$ implies that

$$H_1 = \ker \partial_1 / \mathrm{Im}\, \partial_2 = \mathbb{R}^{2m}/\{0\} = \mathbb{R}^2 m$$

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

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

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

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 is true, but we will not prove it here.

CHAPTER 3

Differentiable (Smooth) Structure on Surfaces.

Material of lectures 15,16 and 17 has been rearranged and expanded.

3.1. Continuation of lecture 15: Monday, Oct. 1 and lecture 16: Monday, Oct. 8

a. Charts and atlases. Thus far we have considered primarily the topological properties of surfaces. The basic definition has been that of a topological manifold as something locally homeomorphic to Euclidean space, with triangulations and maps entering as auxiliary tools. These give the surface some extra structure which has proved useful in our programme of classification, but come with two drawbacks. In the first place, we have not yet established that they are universally applicable, that every surface admits a triangulation. From a more aesthetic point of view, the extra 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 path (set of coordinates) to another.

Definition 14. *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) $\phi_\alpha : U_\alpha \to \mathbb{R}^2$ such that*

(1) *The charts cover S; that is, $\cup_\alpha U_\alpha = S$.*
(2) *ϕ_α is a homeomorphism for every α.*

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

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

We say that an atlas $\mathcal{A}$ is differentiable *(or* smooth*) if every transition map is differentiable and has nonvanishing 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 nonvanishing 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 differentiable? The usual meaning of the word 'smooth', and the one which we will for the most part use, is $\mathcal{C}^\infty$, that is derivatives of all order exist and are continuous. We could also consider $\mathcal{C}^r$-manifolds, for which the transition maps are only required to have continuous derivatives up to order r.

The definition above can be generalized by replacing $\mathbb{R}^2$ with $\mathbb{R}^n$, in which case the manifold is said to be an n-dimensional *differentiable manifold* (or sometimes *smooth manifold*).

DEFINITION 15. *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 many different, incompatible, smooth structures. For example, $\mathbb{R}$ is a one-dimensional smooth manifold with atlas $\mathcal{A}$ given by a single map, the identity $\mathrm{Id} : \mathbb{R} \to \mathbb{R}$. We may consider an atlas $\mathcal{B} = \{\phi\}$ which is also given by a single piecewise linear map

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

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.

These differentiable structures on the line as well as similarly obtained structures on other manifolds, although incompatible, 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.

In higher dimensions, however, 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 nonvanishing reduces to the condition that $f'(x) \neq 0$. Given a map $f : \mathbb{R} \to \mathbb{R}$ 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 $f : x \mapsto x^3$, for which the derivative of f^{-1} has a singularity at 0.

The two-dimensional version of the Inverse Function Theorem states that if $F = (f, g) : \mathbb{R}^2 \to \mathbb{R}^2$ has $f, g \in \mathcal{C}^1$ and has nonvanishing Jacobian determinant, then 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 $\Phi = (\phi, \psi) : U \to \mathbb{R}^2$ such that $\Phi \circ F(x, y) = (x, y)$ for every $(x, y) \in \Phi(U)$, and that in addition,

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

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 1 shows an example of a continuously differentiable map from the unit square to itself which has nonvanishing 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.

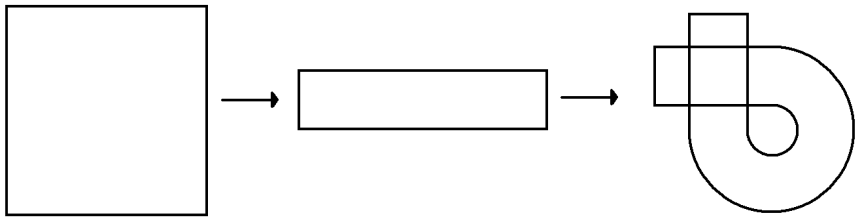

FIGURE 1. A map with no global inverse

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 noncompact surfaces, and it is natural to consider such cases since individual charts are already noncompact.

EXAMPLE 2. The open disc $D^2 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$ is a noncompact surface on which we can place a smooth atlas with a single

chart. Using polar coordinates, we may write the chart as

$$\begin{array}{rccc} \phi & : & D^2 & \to & \mathbb{R}^2 \\ & & (r,\theta) & \mapsto & (\rho(r),\theta) \end{array}$$

where $\rho : [0,1) \to [0,\infty)$ 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. There are several ways of doing that. Here is one, inspired by elementary considerations from complex analysis. First, here is a map form the D^2 to the upper half-plane given in the complex form

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

or in rectangular coordinates

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

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

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

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 that most of the time.

Example 3. Another useful method is to use open rectangles as patches. In this case again a single chart is sufficient since the map

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

maps the rectangle $a < x < b, c < y < d$ onto the whole plane.

Example 4. Any open subset $U \subset \mathbb{R}^2$ is a noncompact surface on which we can place a smooth atlas with infinitely many charts. In 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), which compose to form the affine map

$$\phi_{x_0,r} : x \to \frac{x - x_0}{r}$$

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, and the result follows.

It follows 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.

This 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 5. By considering the annulus, we see 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 2, and polar coordinates give a homeomorphism from each region to a rectangle in the (r, θ)-plane.

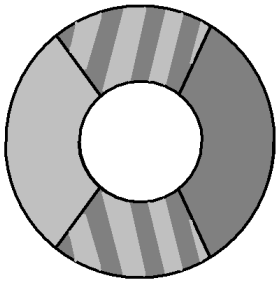

FIGURE 2. Two charts on an annulus

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, but not with one.

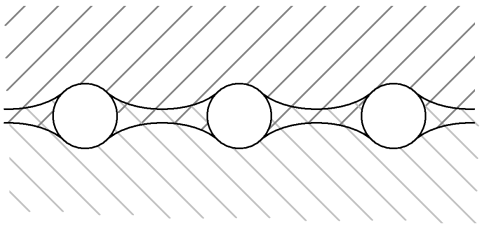

FIGURE 3. Two charts on the plane with several holes

3.2. Lectures 17: Wednesday, Oct. 10 and 18: Friday October 12

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 $\phi : U \to D^2$. The local coordinates on U are given by $\phi^{-1} : D^2 \to U$, as shown in figure 4, 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.

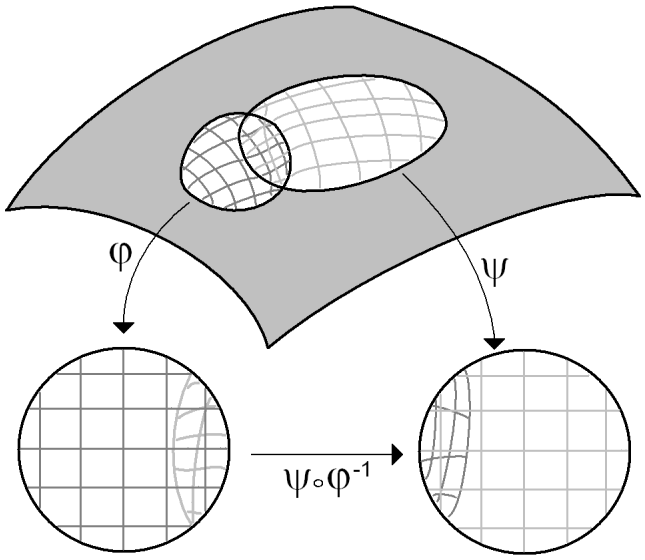

FIGURE 4. Two charts and their transition map

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 above, 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.

Now we consider specific case of differentiable manifolds.

DEFINITION 16. *Given two smooth surfaces S and S' with atlases $\mathcal{A}$ and $\mathcal{A}'$, respectively, and a homeomorphism $f : S \to S'$, any chart $\phi : U \to D^2$ in $\mathcal{A}$ can be carried to a chart $\phi \circ f^{-1} : f(U) \to D^2$ 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 say that $f : S \to S'$ 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.

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 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; 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 earlier discussion (several weeks ago) about 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.

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

c. More examples of charts and atlases. Now we can interpret some of the examples from the previous lecture as construction of diffeomorphisms between various manifolds. Specifically we proved that the disc and rectangles are diffeomorphic to the whole plane $\mathbb{R}^2$ and formulas (1) and (2) provide 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 $\mathbb{R}^2$. This observation motivates looking for more examples of this kind.

EXAMPLE 6. Consider any bounded convex region U in the plane. It 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 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 p.

The above argument 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 the natural smooth structure is provided by any covering by discs to other surfaces we face the problem of defining a natural smooth structure.

Consider 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 neighborhood of the point. Compatibility 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 as consider specific example of the round sphere.

EXAMPLE 7. Following the above recipe one can try to project the sphere to each of the coordinate plane. Without loss of generality we can take the plane to be horizontal. Each hemisphere projects bijectively onto the unit disc and is thus covered by a chart. Equator is not covered, but it is covered by four charts arising from the projections to remaining coordinate planes.

Let as see how this coordinate systems are related. Consider without loss of generality intersection of two charts $y > 0$ and $z > 0$. In one chart x and z may serve as coordinates, in the other x and y. Since

$$y = \sqrt{1 - x^2 - z^2} \quad \text{and} \quad z = \sqrt{1 - x^2 - y^2},$$

$\frac{\partial y}{\partial z} = \frac{-2z}{\sqrt{1-x^2-z^2}} < 0$ and $\frac{\partial z}{\partial y} = \frac{-2y}{\sqrt{1-x^2-y^2}} < 0$. Notice that those derivatives are equal to corresponding Jacobians since obviously $\frac{\partial x}{\partial x} = 1$ and $\frac{\partial x}{\partial z} = 0$ and similarly for the inverse map. Thus transition $(x, z) \mapsto (x, y)$ from one coordinate system to the other satisfies the compatibility condition.

But there are other more economical atlases on the sphere which generate the same differentiable structure. The sphere cannot obviously be covered

with a single chart since it is compact and hence is not homeomorphic to the plane but can be covered with two via stereographic projection from two antipodal points. Notice that the stereographic projection maps the complement to a pole onto the plane thus providing a chart on the sphere *with a single point removed*. Explicit calculations checking that two stereographic projections are compatible and also compatible with the hemispheric charts described above are left to the reader.

EXAMPLE 8. The standard flat torus provides another way to introduce a natural differentiable structure which is somewhat less visual but 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 to the torus. To do that notice that every disc or radius less than $1/2$ projects to the torus injectively and thus defines a chart. As in the case of open subsets of the plane the transition between local coordinates coming from different discs are given by affine maps, hence the charts are compatible.

One can address right away the question of the minimal number of charts required. Four is obviously sufficient: the discs of radius $2/5$ centered in the center of the square, at midpoints of two non-identified sides and at the vertex obviously cover the torus. 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 in this case but the proof is far from straightforward.

3.3. Lecture 19: Monday, Oct. 15

Prior to the last few examples of the preceding lecture, we had dealt primarily with smooth manifolds in the abstract. The examples illustrated some possible techniques for defining smooth structures on particular 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. Consider first embedded surfaces in $\mathbb{R}^3$. That is, let $F : \mathbb{R}^3 \to \mathbb{R}$ be a smooth function, and let $S = \{(x, y, z) \in \mathbb{R}^3 : F(x, y, z) = 0\}$ 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. We have that 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. Smoothness of the transition maps is guaranteed by the Inverse Function Theorem, 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 is carried to the plane best fitted with the surface. A disadvantage of this method is that it is more complicated computationally.

b. Gluing surfaces. A second method, as outlined in this week's homework assignment, 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 glue along those holes. This is illustrated in figure 6; a sphere with a hole is simply a disc, so a sphere with n holes may be drawn as a disc with $n-1$ holes (figure 5). There is a natural smooth structure on the disc, coming from the plane, and so each of the two pieces in figure 6 has a smooth structure; it may be checked that these give rise to a smooth structure on the union, which in this case is a sphere with 3 handles.

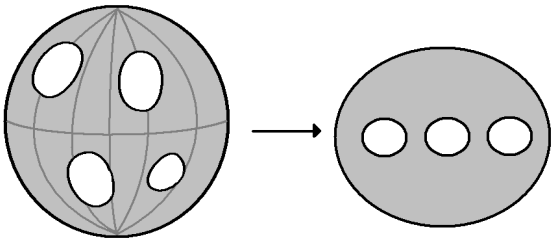

FIGURE 5. A sphere with 4 holes

Since we saw last time 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 applicable any time we are considering a quotient space and and already have a smooth structure on the covering surface. This for example was the case in the homework assignment to cover the torus with three charts.

Suppose $\pi : \tilde{S} \to S$ 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 $\pi : U \to \pi(U)$ may not be injective. However, because open subsets of patches can also be used as

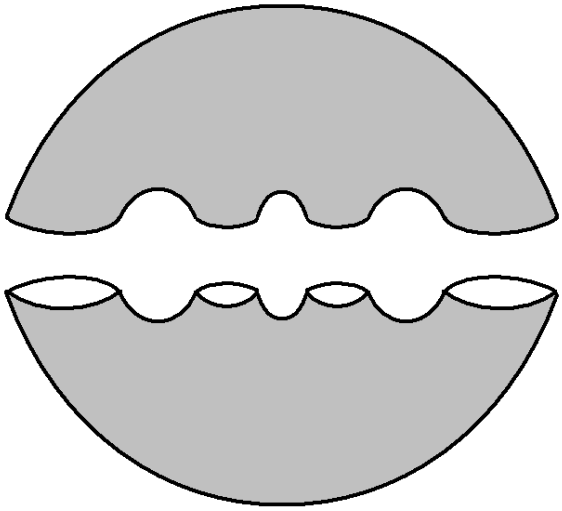

FIGURE 6. Gluing to get a sphere with 3 handles

patches, we can guarantee injectivity by only considering images of charts whose patches are 'small enough' in precisely that sense.

In the previous lecture, we saw this construction applied to the case of the flat torus; there 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.

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 optimal except in the case of the sphere, which can be covered with just two charts, for example, via stereographic projection.

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 7, 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.

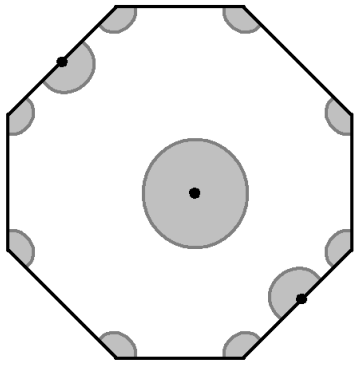

FIGURE 7. Patches on a planar model on an octagon

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 ordered cyclically giving a neighborhood homeomorphic to a disc but standard smooth structures in the eight sectors do not match 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 when a natural structure possesses isolated singularities. One quite 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

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

is mapped to the wedge

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

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). 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 interior and edge charts are obtained from the standard Euclidean coordinates by affine transformations and the transition functions

to and from the vertex chart vertex chart are given by the coordinate expression of the function z^3 and its inverse outside of the origin which satisfy both differentiability and Jacobian inevitability conditions.

e. Riemann surfaces. In fact, 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 suspect. We must begin our (brief) foray into the subject of complex manifolds with a definition from basic complex analysis.

DEFINITION 17. *Given an open domain $U \subset \mathbb{C}$, a map $f : U \to \mathbb{C}$ 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 a *conformal* map.

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, g on separated domains $U, V \subset \mathbb{R}^2$, we can 'glue' them together to obtain a smooth function $h : W \to \mathbb{R}^2$, 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 18. *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 an obvious complex structure on $\mathbb{R}^2$ making it into the Riemann surface $\mathbb{C}$; another example is any flat torus since transition maps are given

by translations which in complex notations have a form $z \mapsto z + c$ and are obviously differentiable as complex functions.

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

$$\begin{array}{rcl} S^2 - \{0\} & \to & \mathbb{C} \\ z & \mapsto & \dfrac{1}{z} \end{array}$$

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 atlas on the octagon discussed in the previous subsection, as well as it immediate generalizations to the orientable surfaces of higher genus, provide a structure of Riemann surface on the remaining orientable compact surfaces.

Having equipped the sphere with a complex structure, it is natural to ask whether this structure is inherited by its quotient space, the projective plane. Is $\mathbb{R}P^2$ a Riemann surface? In fact, it is not; because holomorphic maps preserve *signed* angles, 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. 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.

Remark . Similarly to the notion of diffeomorphism for real differentiable surfaces there is a notion of holomorphic equivalence for complex 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. Another aspect of this complexity is that on a given surface there are often many non-equivalent structures of a Riemann surface.

3.4. Lecture 20: Wednesday, Oct. 17

a. More on Riemann surfaces. As examples of Riemann surfaces (one-dimensional complex manifolds), we have seen the complex plane $\mathbb{C}$ and its open domains, as well as the Riemann sphere $S^2 = \mathbb{C} \cup \{\infty\}$. The latter is arguably the most important example of a Riemann surface, even more so than $\mathbb{C}$ itself; as justification for this claim, consider fractional linear

transformations of the form

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

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

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

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 nonlinear polynomials it will no longer be one-to-one.

Another example of a Riemann surface is the torus $\mathbb{T}^2 = \mathbb{C}/\mathbb{Z}^2$. In fact, we may consider *any* lattice in the complex plane given by

$$L = \{nu + mv : m, n \in \mathbb{Z}\}$$

where $u, v \in \mathbb{C} = \mathbb{R}^2$ are linearly independent over $\mathbb{R}$. Then $\mathbb{C}/L$ gives a Riemann surface which is diffeomorphic to the standard flat torus, but which may carry a different complex structure; in general, given two lattices $L_1, L_2 \subset \mathbb{C}$, the tori $\mathbb{C}/L_1$ and $\mathbb{C}/L_2$ will *not* be equivalent as Riemann surfaces. That is, there is no homeomorphism $f : \mathbb{C}/L_1 \to \mathbb{C}/L_2$ such that $\phi \circ f \circ \psi^{-1}$ is holomorphic for every pair of charts ϕ, ψ.

This last example may be used to show that surfaces which are diffeomorphic may not be holomorphically equivalent. We will return to it later in this course when we will consider Riemann surfaces more systematically.

As more striaghtforward 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, 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.[1]

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 a consequence of Cauchy's integral formula, one of the fundamental results in complex analysis; once this theorem is known, we

[1]The reader with some knowledge of hyperbolic geometry may wish to consider the ramifications of this paragraph in that context.

can simply observe that any polynomial $p(z)$ is holomorphic and bounded on D^2, so that if $\psi : \mathbb{C} \to D^2$ were a holomorphic equivalence, then $\psi \circ p$ would be a bounded holomorphic function on $\mathbb{C}$, contradicting the theorem.

b. Conformal property of holomorphic functions and invariance of angles on Riemann surfaces. It is instructive to consider the question of 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 are holomorphic; geometrically, they must preserve signed angles, so we can now speak of *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 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 $f : \mathbb{R} \to \mathbb{R}$ 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 $\phi : \mathbb{C} \to \mathbb{C}$ 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 an 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 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.

c. Differentiable functions on real surfaces.

DEFINITION 19. *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. 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

$$f_U = f \circ \phi^{-1}$$

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 20. *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 nonzero 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 nonzero determinant and hence is invertible, so $\nabla f_V = 0$ if and only if $\nabla f_U = 0$. We have proved the following:

LEMMA 4. *The critical points of a differentiable function are independent of the particular choice of coordinate chart.* □

We now have a lemma which shows that away from its critical points, any function can be made to assume a standard form by choosing an appropriate coordinate chart.

LEMMA 5. *Given a differentiable function $f : S \to \mathbb{R}$ and a regular point $p \in S$, there exists a chart $\phi : U \to D^2$ around p in which $f_U(x, y) = f(\phi^{-1}(x, y)) = x$.*

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. □

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 optimize 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.

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

$$\begin{aligned} \begin{pmatrix} x \\ y \end{pmatrix} &\mapsto A \begin{pmatrix} x \\ y \end{pmatrix} \\ &= \begin{pmatrix} ax + by \\ bx + cy \end{pmatrix} \end{aligned}$$

or to define a quadratic form $\mathbb{R}^2 \to \mathbb{R}$ by

$$\begin{aligned} \begin{pmatrix} x \\ y \end{pmatrix} &\mapsto \begin{pmatrix} x \\ y \end{pmatrix}^T A \begin{pmatrix} x \\ y \end{pmatrix} \\ &= \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\ &= ax^2 + 2bxy + cy^2 \end{aligned}$$

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^TAC.

Now because $\det C^T = \det C$, we have $\det(C^TAC) = \det(C)^2 \det A$, and so the sign of the determinant is preserved by changes of coordinates. Assuming the matrix $D^2f(p)$ is nondegenerate, we have three possibilities:

(1) $\det D^2f(p) > 0$ and $D^2f(p)$ is positive definite. Then p is a local minimum for f.
(2) $\det D^2f(p) > 0$ and $D^2f(p)$ is negative definite. Then p is a local maximum for f.
(3) $\det D^2f(p) < 0$. Then p is neither a minimum nor a maximum.

The *Morse lemma*, which we will be proved later, states that

(1) In the first case, there exists a local coordinate system in which $f(x, y) = f(0, 0) + x^2 + y^2$.
(2) In the second case, there exists a local coordinate system in which $f(x, y) = f(0, 0) - (x^2 + y^2)$.
(3) in the third case, there exists a local coordinate system in which $f(x, y) = f(0, 0) + x^2 - y^2$.

3.5. Lecture 21: Friday, Oct. 19

a. More about smooth functions on surfaces. Given a compact surface S and a smooth function $f : S \to \mathbb{R}$, 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 $S^2 = \{(x, y, z) \in \mathbb{R}^3 : x^2+y^2+z^2 = 1\}$ and the height function $f : (x, y, z) \mapsto z$. Then 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 8, 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

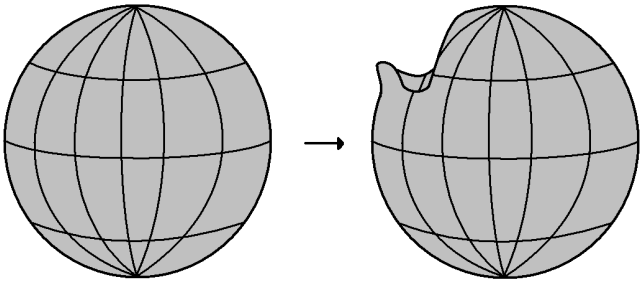

FIGURE 8. Two spheres with different height functions

perturbed sphere has six critical points; two maxima, two saddles, and two minima.

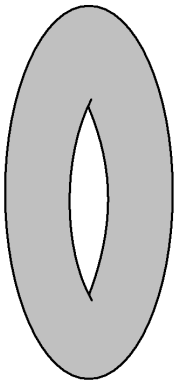

FIGURE 9. The torus of revolution $x^2 + (\sqrt{y^2+z^2} - 2)^2 = 1$

Another interesting example is given by the standard torus of revolution standing sideways as shown in figure 9, again with the height function $f : (x, y, z) \to z$. 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; figure 10 shows the setup for $m = 3$.

DEFINITION 21. *Let S be a smooth surface and $f : S \to \mathbb{R}$ a smooth function. f is called a* Morse function *if every critical point p of f satisfies the following:*

(1) *p is* isolated*; there exists some open neighbourhood U containing p such that p is the only critical point of f within U.*
(2) *p is* nondegenerate*; 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, a minimum, or a saddle; as a consequence of the

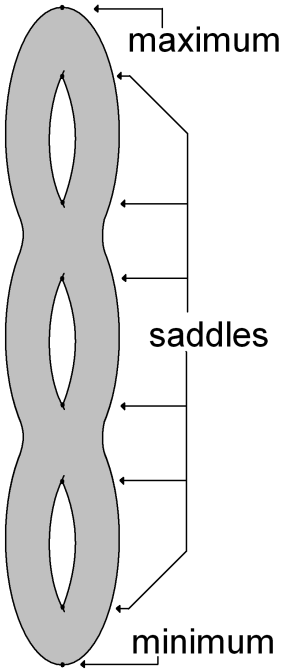

FIGURE 10. Height function on the sphere with three handles

Morse lemma (which we stated last time without proof), f can be put into a standard form around each critical point by a suitable choice of local coordinates.

We will find that looking at the level sets of a Morse function $f : S \to \mathbb{R}$ 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 $F : \mathbb{R}^3 \to \mathbb{R}$, 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. There are three possibilities:

(1) p 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 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.

(3) p 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 11; as we pass through c, two circles may join and become one, or one circle may split and become two.

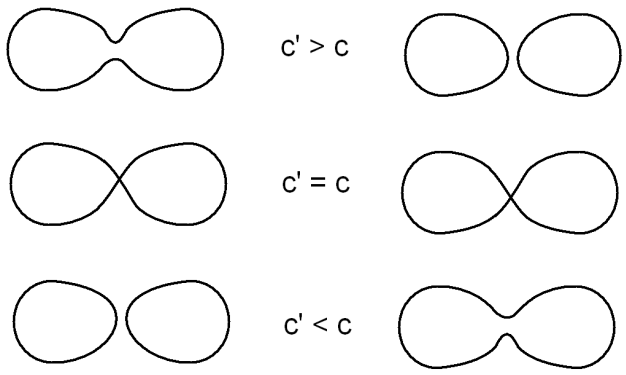

FIGURE 11. Level sets $f^{-1}(c')$ passing through a critical point

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 generalization 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.[2] 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.

[2]This brought a fields medal to Stephen Smale in 1966; solution of the conjecture in two remaining dimensions, four and the original three, resulted in two more Fields Medals later.

b. The third incarnation of Euler characteristics. At a more down-to-earth level, we will show now how to use Morse functions to describe a third incarnation of the Euler characteristic χ for surfaces. 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 6. *For any Morse function $f : S \to \mathbb{R}$, the Euler characteristic is related to the number of critical points by the formula*

$$\chi = (\#\text{ of maxima}) - (\#\text{ of saddles}) + (\#\text{ of minima})$$

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*

$$S_c = f^{-1}((-\infty, c]) = \{x \in S : f(x) \leq c\}$$

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 $M\ldots$

The next observation to make is that nothing interesting happens at non-critical levels. This is accomplished by the following lemma, which intuitively looks quite plausible. A rigorous proofs requires certain tools which we will develop later.

LEMMA 6. *Given a Morse function $f : S \to \mathbb{R}$ 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 $c, c' \in (a, b)$.* □

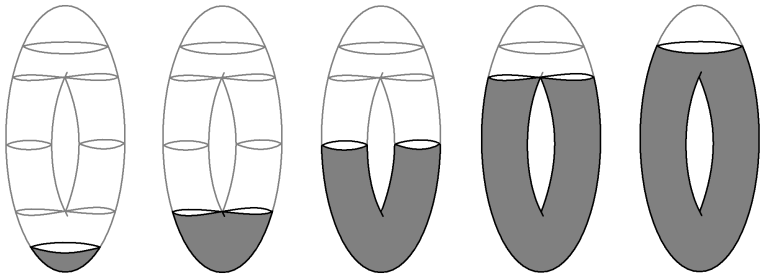

FIGURE 12. Sublevel sets on the vertical torus

Thus for the torus shown in figure 12, 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 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 6. Between critical levels, the lemma shows that the changes in S_c are only qualitative, not quantitative, 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.

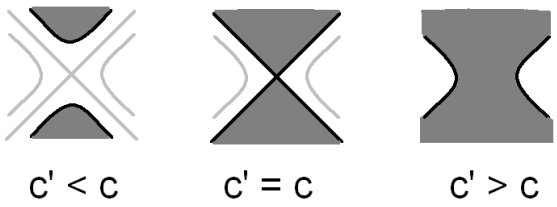

FIGURE 13. Changing sublevel sets passing through a saddle

It remains only to show that passing through a saddle point decreases χ by one. Figure 13 shows the sublevel sets $S_{c'}$ (viewed from above) for values of c' near the critical value c. Recall that subdividing edges does not change Euler characteristic; if we subdivide the two edges in the first picture by adding two vertices to each as shown in figure 14, then passing through the critical level has the effect of adding two edges and a face. This decreases Euler characteristic by one, and establishes the result. □

If we carry out this construction a bit more carefully, we can actually obtain a complete classification of smooth surfaces using Morse functions

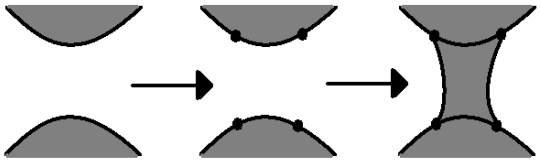

FIGURE 14. Decreasing Euler characteristic by passing through a saddle

as our tool; this was in fact the inspiration for the proof we gave of the classification theorem as well as a "baby version" of the arguments used in higher dimension, like those on which the above-mentioned proof of the Poincaré conjecture in dimensions five and higher is based.

3.6. Lecture 22: Monday, Oct. 22

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 now would like to extend the same ideas and techniques to the case where our function $f : S \to \mathbb{R}$ may fail to be Morse by having degenerate critical points.

We begin by noting that in the nondegenerate Morse case, we obtained the Euler characteristic by giving each critical point a 'weight' of $+1$ (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 $\mathrm{ind}_f(p)$ in such a way that the following formula holds:

$$\chi = \sum_{\nabla f(p)=0} \mathrm{ind}_f(p)$$

It is instructive to begin by considering degenerate critical points in one dimension. Given a smooth function $f : \mathbb{R} \to \mathbb{R}$, nondegenerate 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 $f : x \mapsto x^3$, 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 nondegenerate critical points. This is in sharp contrast to the case where the critical point is already nondegenerate; because the condition $\det(D^2 f) \neq 0$ is an *open* condition, sufficiently small perturbations will not effect any qualitative changes. As we have seen, perturbing f will result in some sort of *bifurcation* near a degenerate critical point.

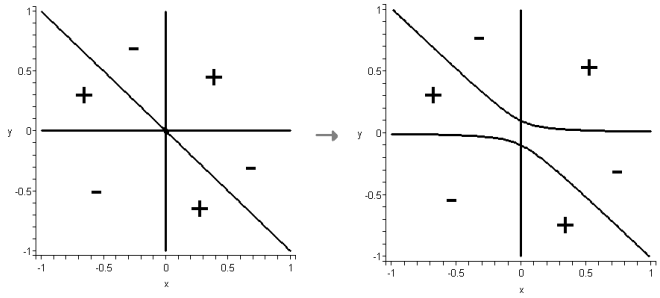

FIGURE 15. Perturbing f in the neighbourhood of a degenerate critical point

Let us examine a two-dimensional example. The function $f : (x, y) \mapsto xy(x + y)$ has the level set $f^{-1}(0)$ shown in the first graph in figure 15. 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; 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 15 shows the situation for $\varepsilon < 0$, where the single degenerate critical point has bifurcated into two nondegenerate 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.

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 adding *three* edges and one face, 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 to every smooth function $f : S \to \mathbb{R}$ is associated 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.

b. Degree of a circle map. Given a map $f : S^1 \to S^1$ from the circle to itself, we can think of the circle as being wrapped around itself a number of times by f; this number is the degree of the map. We can make this 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 $F : \mathbb{R} \to \mathbb{R}$ (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 $x+\mathbb{Z} = \{\ldots, x-1, x, x+1, \ldots\}$). 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 lifst and any two lifts differ by an integer constant.

DEFINITION 22. *Given $f : S^1 \to S^1$ and $F : \mathbb{R} \to \mathbb{R}$ defined as above, the integer $F(1) - F(0)$ is the* degree *of the circle map f.*

The first half of figure 16 shows a circle map f (actually, a graph of the lift F) with degree 2.

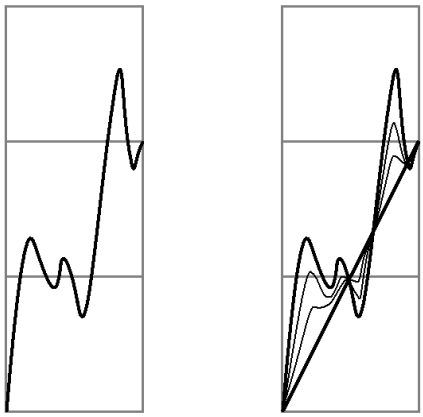

FIGURE 16. A map of the circle with degree two is homotopic to a linear map

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 $f_0, f_1 : S^1 \to S^1$ 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 16, we can construct a linear homotopy from any circle map with degree k to the standard linear map $x \mapsto kx$ via the family of functions

$$F_t(x) = (1-t)F(x) + tkx$$

with $F_0 = F$, $F_1 : x \mapsto kx$. This can be used to show that two circle maps with the same degree are homotopic to each other, so degree classifies circle maps up to homotopy.

c. Zeroes of a vector field and their indices. We now apply this to a continuous vector field; that is, a map that assigns to each point on the surface a vector tangent to the surface at that point. For our present purposes, it suffices to consider a vector field in terms of its representation in a particular coordinate system; in other words a vector field is a function $(x, y) \mapsto (u, v)$. We postpone the coordinate-free definition for later.

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 nonzero, 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 (x_0, y_0) be an isolated zero of (u, v), and consider a small circle around (x_0, y_0). To each point (x, y) on the circle the vector field assigns a vector (u, v), and by normalising (u, v) to

$$\left(\frac{u}{||(u,v)||}, \frac{v}{||(u,v)||}\right)$$

we obtain a unit vector, which is just a point on the unit circle. In this way the vector field near (x_0, y_0) defines a circle map.

DEFINITION 23. *The* index *of a critical point of a vector field is the degree of the circle map defined above.*

By continuously deforming the circle into a circle of a different radius, or any other simple closed curve around (x_0, y_0), we vary the induced circle map, and hence the index, continuously, provided (x_0, y_0) is the only point within or on the curve at which the vector field vanishes. 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.

We have developed this theory with the goal of applying it to the gradient vector field ∇f to define the index of any isolated critical point of f, which we are now in a position to do. In fact, 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.

3.7. Lecture 23: Wednesday, Oct. 24

a. More on degrees. Last time we introduced the notion of the degree of a circle map $f : S^1 \to S^1$, which counts how many times the circle wraps

around itself under the map f. The standard examples are the linear maps E_n for $n \in \mathbb{Z}$; if we write the circle additively as $\mathbb{R}/\mathbb{Z}$, then

$$E_n(x) = nx \mod 1$$

and if we choose to write it multiplicatively as the unit circle $\{z \in \mathbb{C} : |z| = 1\}$, then

$$E_n(z) = z^n$$

This map wraps the circle around itself n times, and serves as the standard representative of maps with degree n.

For a general map $f : S^1 \to S^1$, we defined the *lift* of f as a function $F : \mathbb{R} \to \mathbb{R}$ which makes the following diagram commute:

$$\begin{array}{ccc} \mathbb{R} & \xrightarrow{F} & \mathbb{R} \\ \pi \downarrow & & \downarrow \pi \\ S^1 & \xrightarrow{f} & S^1 \end{array}$$

Here π is the covering map; in additive notation, $\pi(x) = x \mod 1 \in \mathbb{R}/\mathbb{Z}$, and in multiplicative notation, $\pi(x) = \exp(2\pi i x) \in S^1 \subset \mathbb{C}$.

The degree of f is given by $\deg f = F(1) - F(0)$, which must be an integer since the diagram commutes. It must be checked that $\deg f$ is well-defined; that is, that we might not obtain some other value for the degree by choosing a different lift $\tilde{F}$.

The lift F is not unique; given $m \in \mathbb{Z}$, the function $\tilde{F} : \mathbb{R} \to \mathbb{R}$ defined by $\tilde{F}(x) = F(x) + m$ is also a lift of f; however, this is the only ambiguity in the definition. That is, given any other lift $\tilde{F}$, we have $\pi F = f\pi = \pi\tilde{F}$, hence $F(x) - \tilde{F}(x) \in \mathbb{Z}$ for every x, and since $F, \tilde{F}$ are continuous, the difference must be constant. Then

$$\tilde{F}(1) - \tilde{F}(0) = (F(1) + m) - (F(0) + m) = F(1) - F(0)$$

and hence the degree is well-defined.

It is not immediately obvious what the higher-dimensional generalisation of this ought to be. For a map f from the n-dimensional torus to itself, we could follow a similar procedure by noting that the n-torus is the quotient space $\mathbb{R}^n/\mathbb{Z}^n$, and then lifting the map f to $\mathbb{R}^n$; no such idea will work on the n-sphere, however, which cannot be covered by any other manifold. 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.

It turns out that we can generalise the concept of degree to higher dimensions, and this is in fact a fundamental definition in algebraic and differential topology. However, it requires us to use the homology groups of the sphere, and all in all would get us into deeper waters than we are prepared for at the moment.

One last comment about degree is in order; we may think of a circle map f as giving the progress of a runner around a track. At time $t = 0$, he is at

the start line, and the proceeds around the track with direction and speed determined by f. 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.

With this analogy in mind, 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 direction of f at that point, and then summing over these values; the sum will be the degree of f.

b. More on indices. We have not yet given a proper definition of a vector field on an arbitrary surface. Provided we restrict ourselves to a particular choice of coordinates, however, the idea is simple enough; if $U \subset \mathbb{R}^2$ is the image of a patch, we consider a map

$$\begin{array}{rccc} X & : & U & \to & \mathbb{R}^2 \\ & & (x,y) & \mapsto & (u,v) \end{array}$$

which assigns to each point $(x, y) \in U$ a vector (u, v) which is to be thought of as specifying a direction and magnitude at (x, y). The vector field is continuous, smooth, etc. if the map X is continuous, smooth, etc.

Now given a critical point (x_0, y_0) of X, that is, a point such that $u(x_0, y_0) = v(x_0, y_0) = 0$, we consider a simple closed curve γ which 'goes around (x_0, y_0) once' in a sense which we will make precise below. We must put a parameter t on the curve, that is, fix a map $\gamma : [0, 1] \to U$ with $\gamma(0) = \gamma(1)$; because the circle S^1 is homeomorphic to the quotient space $[0, 1]/\sim$ where $0 \sim 1$, this corresponds to fixing a homeomorphism between the circle and γ. (Note the abuse of notation in using γ for both the curve and its parametrisation).

If the region enclosed by γ is star-shaped from the point (x_0, y_0), we may take as our parameter on γ the angle made by the line from (x_0, y_0) to (x, y) with the positive horizontal direction; for curves enclosing more complicated regions, the process is somewhat more delicate.

It is vital to our definition that γ does not go through any critical points of X; that is the vector field must be nonvanishing along the curve. 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.

Under these assumptions, we can define a circle map $\phi_\gamma : S^1 \to S^1$ by

$$\phi_\gamma : t \mapsto \frac{(u(\gamma(t)), v(\gamma(t))}{||(u(\gamma(t)), v(\gamma(t))||}$$

Then the index of the vector field X at the critical point (x_0, y_0) is the degree of ϕ_γ.

Continuously changing the curve γ corresponds to continuously changing the map ϕ_γ, which leaves the degree, and hence the index, constant. Last time we used this process to achieve a homotopy between any circle map with degree n and the standard linear map E_n; in the case of a critical point for a vector field the standard model to keep in mind is a small circle around the critical point, to which any simple closed curve can be deformed without passing through any critical points, provided it 'goes around (x_0, y_0) once'.

In order to make this last statement precise, we must also define the notion of index for a curve. 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

$$\phi : t \mapsto \frac{\gamma(t) - p}{||\gamma(t) - p||}$$

DEFINITION 24. *Given γ, p, ϕ as above, the* index *of γ around p, also called the* winding number*, is the degree of ϕ.*

This definition 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. For our purposes above, the rather vague statement that γ 'goes around (x_0, y_0) once' ought to be replaced by the requirement that the index of γ around (x_0, y_0) is one.

c. Tangent vectors, tangent spaces, and the tangent bundle. 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 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 union of all the tangent spaces.

We begin by considering a single chart $\phi : U \to \mathbb{R}^2$; 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 behaves 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 ϕ^{-1} applied to 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

$$T_pS = \left\{ t\frac{\partial}{\partial x} + s\frac{\partial}{\partial y} : (t,s) \in \mathbb{R} \right\}$$

What happens to this definition under a change of coordinates $F : (x,y) \mapsto (u,v)$? We will see next time 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 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

$$\lim_{t\to 0} \frac{||\phi(\gamma(t)) - \phi(\eta(t))||}{t} = 0$$

Then the tangent space T_pS is given as the set of equivalence classes of smooth curves up to tangency.

In the next chapter we will look in detail at one particular structure on a surface which is the most important from the point of view of geometry: Riemannian metric. However our business with other aspects of smooth structure is far form over.

First, we only mentioned some important results such as Morse lemma and expression of Euler characteristic as the sum of indices of zeroes for a vector field (the fourth incarnation of Euler characteristic).

Second, we have not developed our understanding of vector fields far enough. Two principal topics here are: (i) integration of vector fields to produce *flows*, one-parameter groups of diffeomorphisms of surfaces, and (ii) structure of vector fields near non-degenerate zeroes. Study of orbits of flows on surfaces is the first chapter of qualitative theory of ordinary differential equations (ODE).

We will turn to those subjects later, in the second chapter dedicated to smooth structure on surfaces.

CHAPTER 4

Riemannian Metrics on Surfaces

4.1. Lecture 24: Friday, Oct. 26

a. Definition of a Riemannian metric. The definition given last time of the tangent space 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 define the notion 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's core from one to the other, which is what the above 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 $\gamma : [0, 1] \to S$ with $\gamma(0) = x$, $\gamma(1) = x'$. For a surface embedded 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 noncompact two-dimensional manifold, as the tangent bundle of S^1. The tangent bundle of a surface will be a noncompact 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

$$\begin{array}{rccc} \phi \times \mathrm{Id} & : & U \times \mathbb{R}^2 & \to & \mathbb{R}^4 = \mathbb{R}^2 \times \mathbb{R}^2 \\ & & (p, u\partial_x + v\partial_y) & \mapsto & (x, y, u, v) \end{array}$$

where $\phi : p \mapsto (x, y)$ is a chart on U, and we use the notation

$$\begin{aligned} \partial_x &= \frac{\partial}{\partial x} \\ \partial_y &= \frac{\partial}{\partial y} \end{aligned}$$

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 25. *An* inner product *(or* scalar product*) on T_pS is a function*

$$\begin{array}{rccc} \langle \cdot, \cdot \rangle_p & : & T_pS \times T_pS & \to & \mathbb{R} \\ & & (u, v) & \mapsto & \langle u, v \rangle_p \end{array}$$

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 26. *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.*

What does 'smooth' mean in this context? By way of answer, 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 $a, b, c : \mathbb{R}^2 \to \mathbb{R}$ 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; the reader is invited to confirm that the sufficient condition is that in addition the matrix in the above calculations has positive determinant. In particular, note that this is an open condition; that is, 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^TAC, 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 $\phi : (x,y) \mapsto (x',y')$. 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

$$\begin{aligned} u'\partial_{x'} + v'\partial_{y'} &= u\partial_x + v\partial_y \\ &= u\left(\frac{\partial x'}{\partial x}\partial_{x'} + \frac{\partial y'}{\partial x}\partial_{y'}\right) + v\left(\frac{\partial x'}{\partial y}\partial_{x'} + \frac{\partial y'}{\partial y}\partial_{y'}\right) \\ &= \left(\frac{\partial x'}{\partial x}u + \frac{\partial x'}{\partial y}v\right)\partial_{x'} + \left(\frac{\partial y'}{\partial x}u + \frac{\partial y'}{\partial y}v\right)\partial_{y'} \\ \begin{pmatrix} u' \\ v' \end{pmatrix} &= D\phi \begin{pmatrix} u \\ v \end{pmatrix} \end{aligned}$$

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 $\langle u, v\rangle_p = u^T A v$, 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.

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 real number λ, the linear combination $\lambda f_1 + f_2$ 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 by a negative constant will not result in a positive definite matrix, so we must restrict ourselves to multiplication by positive valus of λ. 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 27. *Let $\{U_1, \ldots, U_N\}$ be a finite cover of S by coordinate patches. A* smooth partition of unity *is a collection $\{\rho_1, \ldots, \rho_N\}$ of smooth functions $S \to \mathbb{R}$ which satisfy the following conditions:*

(1) $\text{supp}(\rho_i) = \overline{\{x : \rho_i(x) \neq 0\}} \subset U_i$
(2) $\rho_i \geq 0$
(3) $\sum_{i=1}^N \rho_i \equiv 1$

We will defer until next time a proof that any finite cover of S by coordinate patches admits a smooth partition of unity, and content ourselves for now 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:

$$A = \sum_{i=1}^N \rho_i A_i$$

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.

This method of gluing together locally defined things which have no *a priori* relation to each other is often the only way of defining 'good' global objects, and has wide applicability.

4.2. Lecture 25: Monday, Oct. 29

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 last time.

THEOREM 7. *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*

$$\phi_i : U_i \to D^2 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$$

are coordinates charts. Then there exists a smooth partition of unity subordinate to $\mathcal{U}$, that is, smooth functions $\rho_i : S \to \mathbb{R}$ such that

(1) $\text{supp}(\rho_i) \subset U_i$
(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 7. *Given a finite smooth atlas $\mathcal{U}$ as above, there exists $\varepsilon > 0$ such that the sets*

$$U_i^\varepsilon = \phi_i^{-1}(D^2_{1-\varepsilon})$$

still cover S, where $D^2_r = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < r^2\}$ is the disc of radius r.

Proof of lemma. We proceed by contradiction; if no such ε exists, then for every $\varepsilon > 0$ we have

$$\bigcup_{i=1}^{N} U_i^\varepsilon \subsetneqq S$$

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 $(t_1, \ldots, t_k)$. In the case of a noncompact manifold and an infinite cover, the lemma is not true as stated, but a 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 $\rho_i : S \to \mathbb{R}$ 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 $F_\varepsilon : \mathbb{R} \to \mathbb{R}$ whose graph is as shown in figure 1. Assuming all the derivatives of F_ε vanish at 0, we can then try to 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 2) of a smooth function for which all derivatives

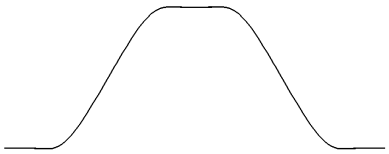

FIGURE 1. The radial profile of our desired bump function

vanish at 0, but which is not identically zero on any neighbourhood of the origin. Define $f : \mathbb{R} \to \mathbb{R}$ piecewise by

$$f(x) = \begin{cases} 0 & x \leq 0 \\ e^{-1/x^2} & x > 0 \end{cases}$$

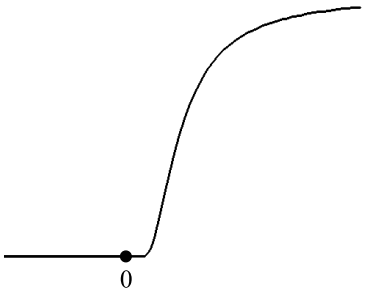

FIGURE 2. A smooth function which is not analytic

EXERCISE 12. Using the fact that the exponential function grows faster than any polynomial, show that $f^{(n)} = 0$ for all $n \geq 0$.

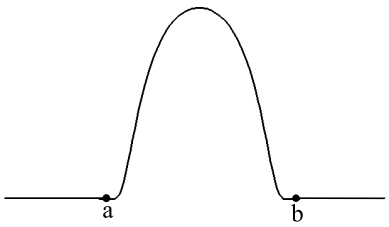

FIGURE 3. A smooth function with support $[a, b]$

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 3. Since $\mathrm{supp}(f_{a,b}) = [a, b]$, we would like to simply define our bump function by

$$\rho_i(\phi_i^{-1}(x, y)) = f_{-1+\varepsilon,1-\varepsilon}(\sqrt{x^2 + y^2})$$

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, nonnegative, 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 on the dramatic difference between smoothness and analyticity for real function; 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. As described last time, we can 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 *locally* defined components.

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 with 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 at such approximation since they depend oh higher derivatives. Quadratic approximation recovers this at least if the second derivative does not vanish. 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, e.g. the radius of the sphere in the case of a spherical metric. The recipe is clear: taking first and, if necessary, higher derivatives of the coefficients into account. We will come to this in a systematic way later.

But there is also another aspect in the relationships between global and infinitesimal properties. Let us look at the basic calculus example again.

In order to find 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 values of the function at those point 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* will lead us to the second-order *Euler-Lagrange differential equation* for a geodesic parametrized by the arc-length, and will allow us to restrict our search to a much smaller class of paths. We will see that solution is uniquely defined by initial condition and initial "velocity", i.e the tangent vector of length one at the initial point. A counterpart of the second derivative test will allow us to conclude that for any two sufficiently close points the solution is indeed has minimal length. And, unlike the case of Euclidean plane, the situation becomes more complicated if the endpoints are far away or if a geodesic comes back to or close to the initial point. The latter is inevitable on compact surfaces even if the geometry local looks Euclidean as in 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

$$||u||_p = \sqrt{\langle u, u \rangle_p}$$

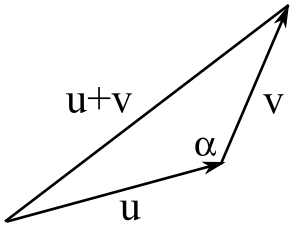

FIGURE 4. Calculating the angle between two vectors

Now consider the triangle shown in figure 4. The law of cosines states that

$$||u+v||_p^2 = ||u||_p^2 + ||v||_p^2 + 2||u||_p||v||_p \cos\alpha$$

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 $\gamma, \eta : (-\varepsilon, \varepsilon) \to S$ 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 $\gamma : [a,b] \to S$ is given by the integral 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 $s : [c,d] \to [a,b]$, the curve $\tilde{\gamma}$ defined by

$$\tilde{\gamma}(s) = \gamma(s(t))$$

should have the same length as γ. This immediately follows from the change of variable under the sign of the integral formula from calculus of one variable.

Now we will 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 coordinate patch. So we will consider 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 $\frac{\partial}{\partial y}$. The area of a domain D is defined as

$$a(D) = \int_D \rho(x, y) dx dy.$$

The change of variable formula from the calculus of two variables shows that this definition is independent of a coordinate change and hence the area of a large domain can be defined as the sum of the areas of its pieces which lie inside coordinate patches.

4.3. Lecture 26: Wednesday, Oct. 31

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. We saw last time 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

$$ds^2 = a(x, y)dx^2 + 2b(x, y)dxdy + c(x, y)dy^2$$

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, the metric is given by 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 $\rho^2(x,y)ds^2$, where $\rho(x,y)$ is a nonvanishing smooth function. That is, the length of each tangent vector in the tangent bundle is scaled by a factor which depends on the base point (x,y) but 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.

b. Differential equations. We recall briefly 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

$$\ddot{x}_i = f_i(x, \dot{x})$$

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

$$\begin{aligned} \dot{x}_i &= v_i(t) \\ \dot{v}_i &= f_i(x, v, t) \end{aligned}$$

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, straight lines. To this end, we fix two points a and b on the surface and look for shortest curves between them.

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 $\gamma : [0, s] \to S$ 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

$$\ell : \gamma \mapsto \int_0^s ||\gamma'(t)||_{\gamma(t)} dt$$

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, we can obtain a criterion for a curve γ to be a critical point of the action functional α. We will not carry out the details at this time, but the end result is a second order ODE; assuming the Riemannian metric is $\mathcal{C}^2$, the existence and uniqueness theorem discussed above applies, and we have the following result:

PROPOSITION 7. *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 $\gamma_v : [0, \infty) \to S$ satisfying*

(1) $\gamma_v'(0) = v$
(2) $||\gamma_v'|| \equiv 1$
(3) *If $|t_1 - t_2| \leq \varepsilon$ then $\gamma : [t_1, t_2] \to S$ 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 8. *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. This is 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.

4.4. Lecture 27: Friday, Nov. 2

a. First glance at curvature. We now turn our attention to what is perhaps the most important 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*.

Two of the standard examples to keep in mind during our discussion of curvature are the Euclidean plane and the 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 independent of a point, respectively; what about negative curvature? We will shortly see such an example, the hyperbolic plane, which cannot be isometrically embedded into $\mathbb{R}^3$. So we first make preliminary remarks concerning definition of curvature without reference to such an embedding. We will return to this problem in more thorough way after having this third invaluable example at out disposal.

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. H.S.M. Coxeter, *Introduction to Geometry*, Section 20.1), which gives a characterisation of one of the several curvature characteristics of the embedded surface the curvature of a 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 θ 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.

Turning to the question of determining curvature on a surface, we will see that intrinsic properties are sufficient. 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 $\kappa : S \to \mathbb{R}$ which is invariant under isometries; that is, which is intrinsically determined. Further, 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 $\tilde{\kappa} : \tilde{S} \to \mathbb{R}$ 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. Further, 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 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 the plane with radius r. We will see that

$$\text{circumference} = 2\pi r - cr^3$$

where c is a constant related to the curvature, and that for the area of the circle we have

$$\text{area} = \pi r^2 - cr^4$$

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

$$\begin{aligned} H^2 &= \{(x,y) \in \mathbb{R}^2 : y > 0\} \\ &= \{z \in \mathbb{C} : \operatorname{Im} z > 0\} \end{aligned}$$

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:

$$ds^2 = \frac{dx^2 + dy^2}{y^2}$$

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.

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 tangent bundle. That is, given any map $f : S \to S$, the Jacobian Df_p at a point p defines a linear transormation between the tangent spaces T_pS and $T_{f(p)}S$, so that the pair (f, Df) acts on the tangent bundle as

$$\begin{array}{rccc} (f, Df) & : & TS & \to & TS \\ & & (p, v) & \mapsto & (f(p), Df_p v) \end{array}$$

For both S^2 and $\mathbb{R}^2$, this action is transitive; given any two points $p, q \in S$ and tangent vectors $v \in T_pS$, $w \in T_qS$, there exists an isometry $f : S \to S$

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 considered the orientation reversing isometry $(x, y) \mapsto (-x, y)$.

We will prove later that every orientation preserving isometry of H^2 has the form

$$\phi : z \mapsto \frac{az + b}{cz + d}$$

where $a, b, c, d \in \mathbb{R}$. (Proposition 10). This condition guarantees that ϕ 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 ϕ is a single point; in fact, we must have $ad - bc > 0$, otherwise ϕ swaps the upper and lower half planes.

Observe that as given, ϕ depends 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 ϕ 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 ϕ 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

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

In fact, if given such a matrix A we denote the transformation given above by ϕ_A, then a little algebra verifies that

$$\phi_{AB} = \phi_A \circ \phi_B$$

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 $\phi_I = \phi_{-I} = Id$, 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 ϕ_A is an isometry for every $A \in SL(2, \mathbb{R})$. The details of this are left to the reader, but the idea is to show that every such ϕ 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

$$\left\{\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} : t \in \mathbb{R}\right\} \bigcup \left\{\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} : \lambda \in \mathbb{R}^+\right\} \bigcup \left\{\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\right\}$$

To see that $z \mapsto -1/z$ is an isometry, one must suffer through a small amount of algebra and use the fact that for $z = x + iy$ we have

$$-\frac{1}{z} = \frac{-1}{x+iy} = -\frac{x-iy}{x^2+y^2} = \frac{-x+iy}{x^2+y^2}$$

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 calculate it for the hyperbolic plane, we will find that our life is made easier by the introduction of a different model 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

$$ds^2 = \frac{4(dx^2+dy^2)}{(1-x^2-y^2)^2} \tag{3}$$

and we may see that this model is the image of H^2 under a conformal transformation, for example

$$z \mapsto \frac{iz+1}{z+i}.$$

An advantage of this model is that rotation around the origin is an isometry, and so circles around the origin are simply circles in the 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 $x_1 = x_2 = x$;

[1] There are of course plenty of other reasons: it is sufficient to recall that the geometry of hyperbolic plane is the original non-Euclidean geometry where all standard axioms except for the fifth postulate hold.

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

$$\begin{aligned} \ell(\gamma) &= \int_{z_1}^{z_2} ||\frac{d}{dt}(x+it)||_{x+it} dt \\ &= \int_{z_1}^{z_2} \frac{1}{t} dt \\ &= \log z_2 - \log z_1 \end{aligned}$$

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. 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 $z \mapsto -\frac{1}{z}$ behaves differently. This map is the composition of reflection about the imaginary axis with the map $z \mapsto -\frac{1}{\bar{z}}$, and the latter is simply inversion in the unit circle. We encountered this map on a homework assignment 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; 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.

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 5.

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 be additive along geodesics. We may construct a geodesic connecting z_1 and z_2 by drawing the perpendicular bisector of the line segment they determine 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.

[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.

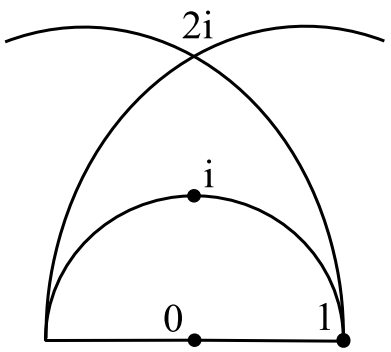

FIGURE 5. Failure of the parallel postulate in H^2

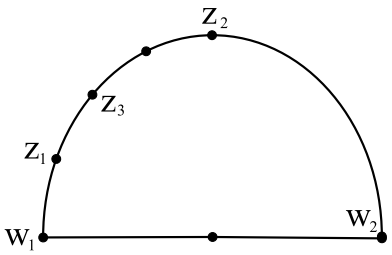

FIGURE 6. Determining the cross ratio of two points

As shown in figure 6, let w_1 and w_2 be the points at which this circle intersects the real line; it turns out that the *cross ratio*

$$\left| \frac{z_1 - w_1}{z_2 - w_1} \div \frac{z_1 - w_2}{z_2 - w_2} \right|$$

is preserved by all isometries of H^2. However, this is 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 6, 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, we must take the logarithm of the cross ratio, and obtain

$$d(z_1, z_2) = \log \left| \frac{z_1 - w_1}{z_2 - w_1} \right| - \log \left| \frac{z_1 - w_2}{z_2 - w_2} \right|$$

4.5. Lecture 28: Monday, Nov. 5

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 today, 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:

$$H^2 = \{(x, y) \in \mathbb{R}^2 : y > 0\} = \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$$

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

$$X = \{(x, y) \in H^2 : 0 \leq x \leq 1\}$$

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.

In order to determine a distance function from the Riemannian metric on H^2, we begin by considering the prototypical example of 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 $\gamma : [y_1, y_2] \to H^2$ given by

$$\gamma(t) = x + it$$

has length given by

$$\begin{aligned} \ell(\gamma) &= \int_{y_1}^{y_2} ||\gamma'(t)|| dt \\ &= \int_{y_1}^{y_2} \frac{1}{t} dt \\ &= \log y_2 - \log y_1 \end{aligned}$$

To see that this is in fact minimal, let $\eta : [a,b] \to H^2$ 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 $x'(t) \equiv 0$. 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. This means we must first find some isometries; these turn out to be given by *fractional linear transformations*

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

where $a, b, c, d \in \mathbb{R}$ are such that $ad - bc = 1$. 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 generally not an attractive specimen (although we will soon be able to compute the distance formula for H^2), and so this approach is something less than ideal.

Rather, we take the infinitesimal point of view and examine the action of f on tangent vectors. That is, we recall that given a map $f : \mathbb{R}^2 \to \mathbb{R}^2$, 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 $v \in T_z H^2$ 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 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 an explicit formula 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 28. *Given $z_1, z_2, z_3, z_4 \in \mathbb{C}$, the* cross-ratio *is the complex number*

$$(z_1, z_2; z_3, z_4) = \frac{z_1 - z_3}{z_2 - z_3} \div \frac{z_1 - z_4}{z_2 - z_4}$$

It turns out that *any* fractional linear transformation, whether the coefficients lie in $\mathbb{R}$ or not, preserves the cross-ratio.

LEMMA 8. *Given any $a, b, c, d \in \mathbb{C}$ with $ad - bc \neq 0$ and $z_1, z_2, z_3, z_4 \in \mathbb{C}$, 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. Substitute expressions of w_i into the cross-ratio formula; notice that constant terms cancel out additively and coefficients in front of z's multiplicatively leaving the cross-ratio of z's. 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 and the *simple ratio*

$$\frac{z_1 - z_3}{z_2 - z_3}$$

then this ratio 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 the linear transformations are characterized by the property of preserving the simple ratio as can be easily seen by assuming $f(z) = w$, fixing two points z_1, z_2 and expressing w through z from the equality

$$\frac{z_1 - z}{z_2 - z} = \frac{f(z_1) - w}{f(z_2) - w}.$$

Later we will use the same argument to show that fractional linear transformations are characterized by the property of preserving the cross-ratio, see Lemma 9.

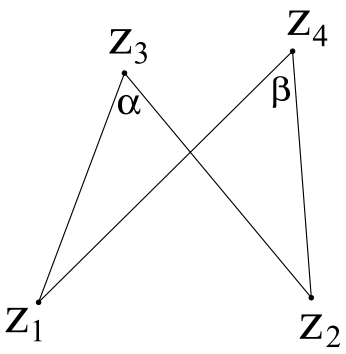

FIGURE 7. Interpreting the cross-ratio of four numbers

What then is the geometric interpretation of the cross-ratio? 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 7. 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 the following theorem:

THEOREM 8. *If γ is a line or a circle in $\mathbb{C}$ and $f : \mathbb{C} \to \mathbb{C}$ 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

$$\begin{aligned} f(\infty) &= \frac{a}{c} \\ f\left(-\frac{d}{c}\right) &= \infty \end{aligned}$$

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.

Thus half circles whose centre lies in $\mathbb{R}$ are also geodesics; given a fractional linear transformation f which maps the vertical half-line $\{z \in H^2 : \operatorname{Re} z = 0\}$ to the half-circle $\{z \in H^2 : |z - a_0| = r\}$ and two points z_1, z_2 lying on the half-circle, 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.

Further, 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 circle whose centre is in $\mathbb{R}$, and that any such half-line or half-circle can be obtained as the image of the imaginary axis under a fractional linear transformation.

The former assertion is straightforward, as described last time. 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 half-circle 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 $1/2$.

4.6. Lectures 29: Wednesday, Nov. 7 and 30: Friday, Nov. 9

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.

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 second approach is the *analytic* one, which we have made some use of thus far; one derives and then makes use of formulas 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 symmetric cases of the Euclidean, hyperbolic and to a certain extent elliptic (projective) planes. Hyperbolic trigonometry can be associated with this approach too.

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. Linear algebra here becomes an invaluable and powerful tool.

b. Characterization of isometries. First, then, we must obtain a complete description of the isometries of H^2. We saw last time that linear

fractional 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 isometries, 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

$$z \mapsto \frac{a\bar{z}+b}{c\bar{z}+d} \tag{4}$$

where $ad - bc = -1$.

Now we claim that these are in fact all of the isometries of H^2. rather that presenting a specific argument for the hyperbolic plane we must first establish a rather general result which says that for any surface the isometry group is not "too big".

PROPOSITION 9. *If S is a surface endowed with a Riemannian metric such that any two points determine a unique geodesic connecting them, then any isometry I of S is uniquely determined by the images of three points which do not lie on the same geodesic.*

Proof. 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 at two points on 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, 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 $I(y)$. □

This proposition establishes uniqueness, but says nothing about the *existence* of such an isometry. 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 regarding fractional linear transformations in general.

LEMMA 9. *Given two triples of distinct points in the extended complex plane, $\mathbb{C} \cup \{\infty\}$ (the Riemann sphere z_1, z_2, z_3 and w_1, w_2, w_3 there exist unique coefficients $a, b, c, d \in \mathbb{C}$ such that the fractional linear transformation*

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

satisfies $f(z_j) = w_j$ for $j = 1, 2, 3$. Furthermore, fractional linear transformations are characterized by the property of preserving the cross-ratio.

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

$$(z_1, z_2; z_3, z) = (w_1, w_2; w_3, w) \tag{5}$$

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:

$$w = \frac{w_1(z_1 - z_3)(w_2 - w_3)(z_2 - z) - w_2(z_2 - z_3)(w_1 - w_3)(z_1 - z)}{(z_1 - z_3)(w_2 - w_3)(z_2 - z) - (z_2 - z_3)(w_1 - w_3)(z_1 - z)}. \tag{6}$$

Since (5) implies (6) we also get the second statement. □

PROPOSITION 10. *Given $z_1, z_2, z_3, w_1, w_2, w_3 \in H^2$ 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 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).

REMARK . The first part of this proposition states that given two triangles in H^2, with corresponding sides of equal length, there 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 9 gives us a fractional linear transformation which is a candidate to be an isometry, this candidate is the desired isometry only if the orientation 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 points 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 11. *Given $z_1, z_2, w_1, w_2 \in H^2$ 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.

By lemma 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 in H^2, 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 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 $\mathbb{R}$.

Thus $f(\mathbb{R}) = \mathbb{R}$, so f has real coefficients and is an isometry of H^2. □

In order to obtain proposition 10, 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 $R : z \mapsto -\bar{z}$ 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 $g^{-1} \circ R \circ g$.

c. Classification of isometries. Now we turn to the task of classifying these isometries and understanding what they look like geometrically.

c.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

$$\begin{aligned} az+b &= cz^2+dz \\ cz^2+(d-a)z-b &= 0 \end{aligned}$$

the roots of which are given by

$$\begin{aligned} z &= \frac{1}{2c}(a-d \pm \sqrt{(a-d)^2+4bc}) \\ &= \frac{1}{2c}(a-d \pm \sqrt{(a+d)^2-4(ad-bc)}) \\ &= \frac{1}{2c}(a-d \pm \sqrt{(a+d)^2-4}) \end{aligned}$$

Note that the quantity $a+d$ is just the trace of the matrix of coefficients $X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, 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 $z = f(z)$:

- **(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 $z_1, z_2 \in \mathbb{R}$.

c.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

$$f : z \mapsto \frac{(\cos\alpha)z + \sin\alpha}{(-\sin\alpha)z + \cos\alpha}$$

which is rotation by α around the point i; the geodesics passing through i are shown in figure 8. Also pictured 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.

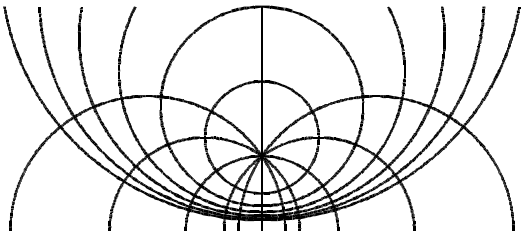

FIGURE 8. Geodesics passing through i and hyperbolic circles centred at i

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, and
(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.

c.3. *Parabolic isometries.* Case **(P)**. This can be considered as a limit case of the previous situation where the point p goes to infinity. Let $t \in \mathbb{R} \cup \{\infty\}$ be the unique fixed point in the Riemann sphere. We can consider now similarly to the family of rotations the family of all orientation preserving isometries preserving t; notice that as is the case with rotation this family is a *one-parameter group* which we will denote $p_s^{(t)},\ s \in \mathbb{R}$. Accordingly one can see two invariant families of curves as above:

(1) The pencil of all geodesics passing through t; each element of this family maps to another and the group $p_s^{(t)}$ acts transitively on this family, and
(2) the family of *limit circles* more commonly called *horocycles* 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 this group acts transitively on each horocycle.

A useful (but visually somewhat misleading) example is given by the case $t = \infty$ with

$$p_s^{(\infty)} z = z + s.$$

Notice that in the process we lost the angle as a distinct invariant of an ordinary rotation; contrary to a natural guess s does not have properties similar to that of the rotation angle.

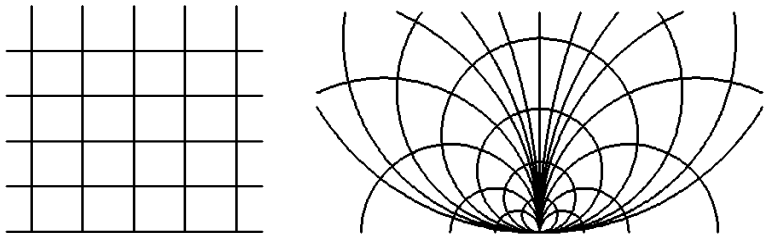

FIGURE 9. Parallel geodesics and horocycles

c.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 half-circle 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 center of rotation in the elliptic case; 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, as shown in figure 10. The effect of f on a member η of this family is determined by the effect of f on the point where η intersects γ and
(2) the family of curves orthogonal to those geodesic which are fixed by f. Those are *equidistant curves* (or *hypercircles*), which are also shown in the picture. 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.

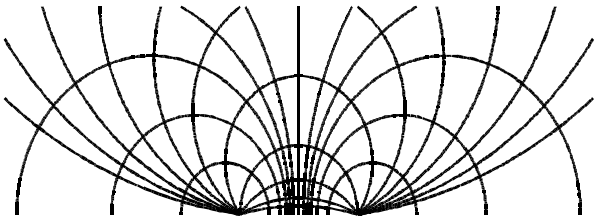

FIGURE 10. Orthogonal geodesics and equidistant curves for a half-circle

A good example of maps f falling into the case **(H)** are the maps which fix 0 and ∞:

$$f : z \mapsto \lambda^2 z$$

In this case the geodesic γ connecting the fixed points is the imaginary axis, which is the vertical line in figure 11, the geodesics intersecting γ orthogonally are the (Euclidean) circles centred at the origin, and the equidistant curves are the (Euclidean) lines emanating from the origin.

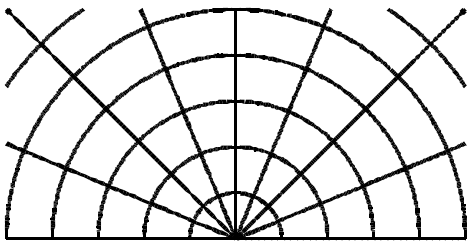

FIGURE 11. Orthogonal geodesics and equidistant curves for the imaginary axis

c.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 preserves orientation.

$$\begin{aligned} PSL(2,\mathbb{R}) = SL(2,\mathbb{R})/\pm \mathrm{Id} &\quad\longleftrightarrow\quad \mathrm{Isom}^+(H^2) \\ A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} &\quad\longleftrightarrow\quad f_A : z \mapsto \frac{az+b}{cz+d} \end{aligned}$$

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; an example of this in the Euclidean plane is given by considering any two rotations by an angle α, which will be conjugated by the translation taking the fixed point of one to the fixed point of the other.

Thus 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 12. *Every matrix in $SL(2,\mathbb{R})$ is conjugate to one of the following (up to sign):*

(E): *An* elliptic *matrix of the form*

$$\begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}$$

for some $\alpha \in \mathbb{R}$.

(P): *The* parabolic *matrix*

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

(H): *The* hyperbolic *matrix*

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

for some $t \in (0,\infty)$. □

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 is we consider only conjugacy by orientation preserving isometries, then α itself is an invariant in the elliptic case and two parabolic matrices $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$ are not conjugate. In contrast, conjugacy classes in the hyperbolic case do not change.

d. Geometric interpretation of isometries. In order to understand the geometric interpretation of each of these three categories of isometries, we begin by examining some basic concepts of hyperbolic geometry.

We have seen already that geodesics in H^2 (hyperbolic lines) are either vertical Euclidean lines or Euclidean circles which intersect $\mathbb{R}$ orthogonally. In fact, a similar statement is true regarding hyperbolic circles; given any point $p \in H^2$ and a radius $r > 0$, the circle

$$\{z \in H^2 : d(z, p) = r\}$$

is a Euclidean circle, which in general will have a different centre and radius. To see this, notice that in the half-disc model of H^2, the hyperbolic circles centred at the origin are simply Euclidean circles also centred at the origin. Because the half-disc model and the upper half-plane model are related by a fractional linear transformation, which takes circles to circles and maps 0 to i, it follows that the one-parameter family of hyperbolic circles centred at i is a one-parameter family of Euclidean circles in H^2. By considering all circles obtained from this one-parameter family via horizontal translation and homothety (both of take hyperbolic circles by hyperbolic circles by virtue of being isometries, and Euclidean circles to Euclidean circles by virtue of being fractional linear transformations), we obtain every hyperbolic circle and every Euclidean circle lying in H^2.

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 12.

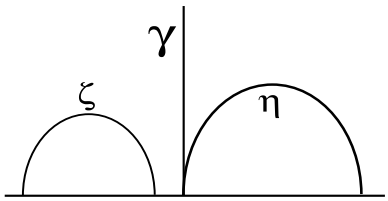

FIGURE 12. Parallels and ultraparallels

The curves γ, η, and ζ in figure 12 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 9. 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 9, 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 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. How can we distinguish between the two sorts of asymptotic behaviour without reference to the ideal boundary?

Notice that 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 13, 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 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.

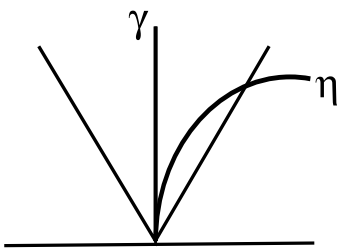

FIGURE 13. Distance between parallel geodesics

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, 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 13 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 8. Given $v \in T_pH^2$, denote by γ_v the unique geodesic passing through p with $\gamma'(p) = v$. Then we have

$$\begin{array}{ccccc} f & : & \{\gamma_v\}_{v\in T_pH^2} & \to & \{\gamma_v\}_{v\in T_pH^2} \\ & & \gamma_v & \mapsto & \gamma_w \end{array}$$

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

$$\begin{array}{ccccc} f & : & \{\gamma_t\}_{t\in\mathbb{R}} & \to & \{\gamma_t\}_{t\in\mathbb{R}} \\ & & \gamma_t & \mapsto & \gamma_{t+1} \end{array}$$

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 in $\mathbb{R}$ rather than at ∞; in this case, the geodesics and horocycles asymptotic to x are as shown in the second image in figure 9. The invariant family of geodesics consists of geodesics parallel to each other.

Finally, in the hyperbolic case *(H)* the standard form $f_A(z) = \lambda^2 z$ for $\lambda = e^t$, 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 11, 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 10. The invariant geodesic η_0 is the half-circle connecting the fixed points, and the equidistant curves are the other circles passing through those two points. The family of orthogonal curves are the geodesics intersecting η_0 orthogonally, as shown in the picture.

4.7. Lecture 31: Monday, Nov. 12

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 presence of similarity transformations - diffeomorphisms of $\mathbb{R}^2$ which expand or shrink the metric by a uniform constant - precluded the existence of any such formula.

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.

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 14. 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.

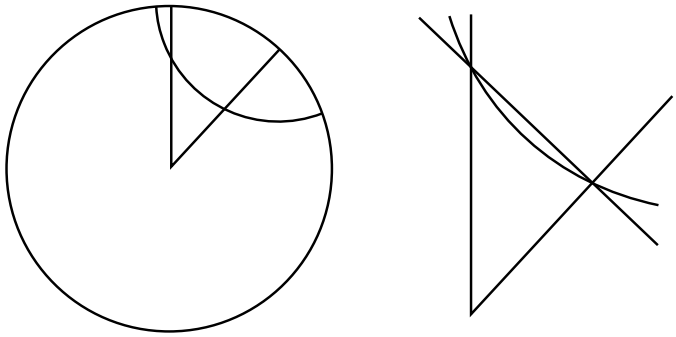

FIGURE 14. A hyperbolic triangle

b. Area and angular defect in hyperbolic geometry. Our proof of the area formula is due to Gauss, and follows the exposition in the 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 use drawings in the models.

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, among which 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 9. *Given a hyperbolic triangle Δ with angles α, β, and γ, the area A of Δ is given by*

$$A = \frac{1}{-\kappa}(\pi - \alpha - \beta - \gamma)$$

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 15, 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.

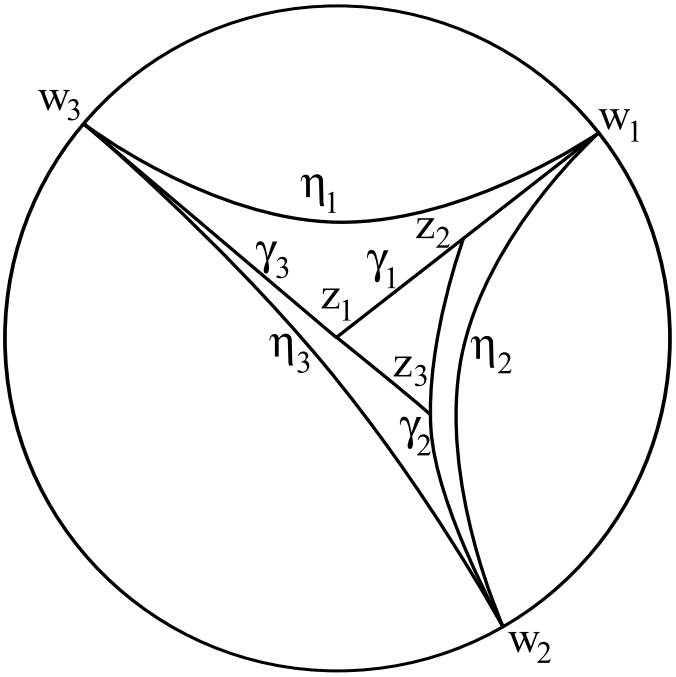

FIGURE 15. Computing the area of a hyperbolic triangle

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 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.

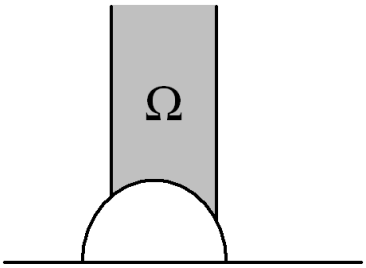

FIGURE 16. A singly asymptotic triangle

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 shown in figure 16, where we use the half-plane model and choose the point at infinity to be ∞, so two of the geodesics are vertical lines. The infinitesimal area element at each point is given by $\frac{1}{y^2}dxdy$ where dx and dy are Euclidean displacements, and so the area of the shaded region Ω is

$$A(\Omega) = \int_\Omega \frac{1}{y^2} dxdy$$

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. 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 nonzero 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

$$\begin{aligned} A(\Delta) &= T_0-T_1-T_2-T_3 \\ &= \mu-\frac{\mu}{\pi}(\theta_1+\theta_2+\theta_3) \\ &= \frac{\mu}{\pi}(\pi-\theta_1-\theta_2-\theta_3) \end{aligned}$$

and hence our formula is proved, 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

$$\begin{aligned} \mu &= \int_T \frac{1}{y^2}dxdy \\ &= \int_{-1}^{1}\int_{\sqrt{1-x^2}}^{\infty}\frac{1}{y^2}dydx \\ &= \int_{-1}^{1}\frac{1}{\sqrt{1-x^2}}dx \\ &= \int_{-\pi/2}^{\pi/2}d\theta \\ &= \pi \end{aligned}$$

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. We sketch a purely synthetic proof, 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.

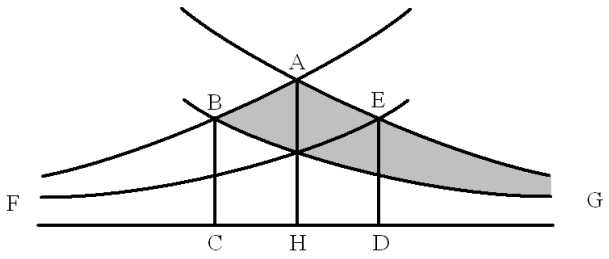

FIGURE 17. Decomposing an asymptotic triangle

Consider such a triangle, given by the shaded region in figure 17. 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. Reflecting AG in the line AH gives the geodesic EF, and 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 later in bisectors of angles obtained by intersecting the reflected lines with the side AG , the rest of the shaded region can be brought into the pentagon $ABCDE$. The details of this are left to the reader; once it is established that ABG can be decomposed into triangles whose isometric images fill $ABCDE$ disjointly, we have that the area of ABG is finite, and the proof is complete.

4.8. Lecture 32: Wednesday, Nov. 14

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 $\Gamma \subset \mathrm{Isom}(\mathbb{R}^2)$ 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 not be carried out for these planar models as well. As a concrete example, consider the octagon with opposite sides identified via the four translations

$$\begin{array}{rcccc} f_1 & : & (x,y) & \mapsto & (x+2,y) \\ f_2 & : & (x,y) & \mapsto & (x+\sqrt{2},y+\sqrt{2}) \\ f_3 & : & (x,y) & \mapsto & (x,y+2) \\ f_4 & : & (x,y) & \mapsto & (x-\sqrt{2},y+\sqrt{2}) \end{array}$$

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, as shown in figure 18, 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 (formally, 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

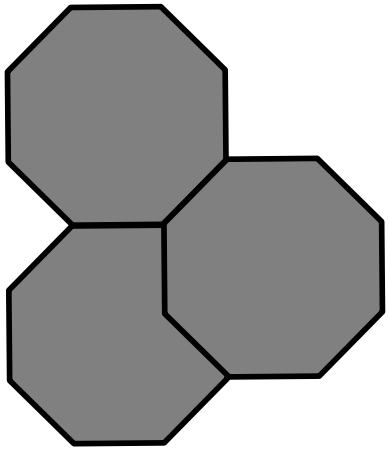

FIGURE 18. Impossibility of tiling the Euclidean plane with octagons

the area is proportional to the angular defect *vis-a-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 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 19.

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 20. As the radius is increased, neighbouring geodesics eventually become parallel and meet at

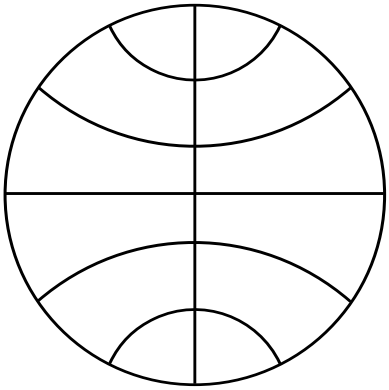

FIGURE 19. Geodesics for a hyperbolic translation

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.

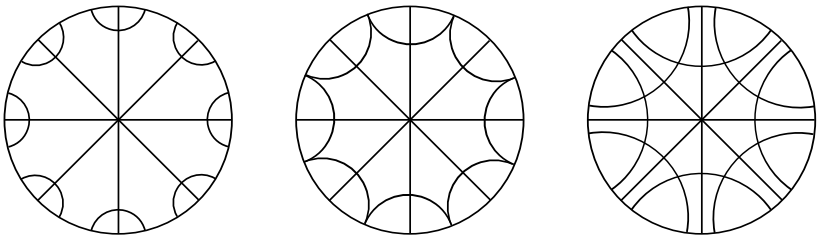

FIGURE 20. Various attempts at a hyperbolic octagon

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.

Now if we let Γ be the subgroup of $\mathrm{Isom}(H^2)$ generated by $\{f_1, f_2, f_3, f_4\}$ and consider the quotient space H^2/Γ whose points are orbits of Γ, we obtain the surface of genus two, as desired. 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

$$f_1 \circ f_2 \circ f_3 \circ f_4 \circ f_1^{-1} \circ f_2^{-1} \circ f_3^{-1} \circ f_4^{-1} = \mathrm{Id}$$

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 $\mathrm{Isom}(\mathbb{R}^2)$ 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 $g_1, g_2, g_3, g_4 \in \mathrm{Isom}(H^2)$ and take the quotient space of H^2 by the action of these isometries; provided the relation

$$g_1 \circ g_2 \circ g_3 \circ g_4 \circ g_1^{-1} \circ g_2^{-1} \circ g_3^{-1} \circ g_4^{-1} = \mathrm{Id}$$

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.

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*.

THEOREM 10. *Let S be a 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 surface with constant curvature κ, we have

$$\kappa A(S) = 2\pi\chi(S)$$

Note that the cases in which the curvature is positive or zero correspond to the sphere and the torus, respectively, where the above formula is already known, and that this relationship forbids the existence of a locally hyperbolic metric on the torus.

Proof. We use the angular defect formula for the area of a hyperbolic triangle, applied to a geodesic triangulation of S. The existence of 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 Euler characteristic:

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

Furthermore, as for any triangulation, counting edges gives

$$3F = 2E$$

which we will use as

$$F = 2E - 2F$$

Finally, for every triangle τ in the triangulation, the angular defect formula 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

$$\begin{aligned} A(S) &= \pi(F - 2V) \\ &= \pi(2E - 2F - 2V) \\ &= -2\pi\chi(S) \end{aligned}$$

which establishes the first formula.

Recall that if we scale the metric by a constant factor, area scales as the square of that factor, and curvature scales as the inverse of the area. Hence the product $\kappa A(S)$ remains constant and equal to $2\pi\chi(S)$, establishing the second formula. □

We originally defined Euler characteristic in terms of triangulations, and then found it crop up in homology via the Betti numbers, and in Morse theory via critical points of smooth functions. The above theorem illustrates yet another guise of 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 to that curvature is nearly constant on each triangle, and then apply the angular defect/excess formula to each triangle, which in its general form states that

$$A = \kappa(\alpha + \beta + \gamma - \pi)$$

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 $\int_S \kappa(x)dA(x)$, obtaining the general expression

$$\chi(S) = \frac{\int_S \kappa(x)dA(x)}{2\pi}$$

for the Euler characteristic.

4.9. Lecture 33: Friday, Nov. 16

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.

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 consider our three standard symmetric examples.

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

$$ds^2 = dr^2 + r^2 d\theta^2 \tag{7}$$

On the sphere with radius R, we may take p to be the north pole. Then the geodesics through p are the meridians, i.e. 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

$$ds^2 = dr^2 + R^2 \sin^2\left(\frac{r}{R}\right) d\theta^2 \tag{8}$$

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 (3) in geodesic polar coordinates becomes

$$ds^2 = dr^2 + \sinh^2 r d\theta^2. \tag{9}$$

In general, in the geodesic polar coordinates as described above, the curves $\theta = const$ are geodesics, while for small values of c the curves $r = c$ are circles centred at p, i.e. the loci of points at the distance precisely c from p. The circles intersect the geodesics $\theta =$const 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

$$ds^2 = dr^2 + (g(r, \theta))^2 d\theta^2$$

where $g : \mathbb{R}^2 \to \mathbb{R}$ is some smooth function positive at $r > 0$ i.e. outside of p and vanishing at p.

b. Curvature as an error term in the circle length formula. Our information about curvature express in geodesic polar coordinates must come from the properties of the function g. We make the following definition, and then offer some geometric justification.

DEFINITION 29. *If we define g as above, the curvature of S at a point $q = (r, \theta)$ is*

$$\kappa(q) = \kappa(r, \theta) = -\frac{g_{rr}}{g} = -\frac{1}{g}\frac{\partial^2 g}{\partial r^2} \tag{10}$$

Notice that in the three symmetric cases (7), (8) and (9) one obtains $\kappa \equiv 0, R$ and -1 correspondingly.

Notice also that since g vanishes at $r = 0$ curvature is not defined at the point p, the center of the geodesic polar coordinate system. In fact, the limit of the right hand part of the expression (10) as $r \to 0$ exists and 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 has the weakness of being dependent on our particular choice of 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?

In order to show that $\kappa(q)$ is in fact independent of the choice of coordinates, we give a coordinate-free interpretation of curvature at p in terms of the circumference of small circles around p. For that we assume that the limit of $-\frac{g_{rr}}{g}$ as $r \to 0$ exists and is finite. Denote this limit $\kappa(p)$. To avoid vicious circle we will show that after the proof of Theorem 11. Fix $r_0 > 0$ and let $C_p(r_0)$ be the circle of radius r_0 around p. Abusing notation slightly, we write $\ell(r_0)$ for the circumference of this circle, and we see that

$$\ell(r_0) = \ell(C_p(r_0)) = \int_{C_p(r_0)} ds = \int_0^{2\pi} g(r_0, \theta)\, d\theta$$

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.

As r_0 goes to zero, we approach the Euclidean case, and the circumference is $2\pi r_0$ plus some higher order terms. Thus we have

$$g_r|_{r=0} = 1$$

Since g is differentiable at 0 and $\lim_{r\to 0}$ is finite this implies that g_{rr} vanishes at $r = 0$. Fixing θ and writing g as a function of r we see that

$$g(r, \theta) = r - \frac{\kappa}{6} r^3 + o(r^3)$$

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:

$$\kappa(p) = 3\lim_{r\to 0}\frac{2\pi r - \ell(r)}{\pi r^3} \tag{11}$$

The beauty of this formula is that it is completely intrinsic, and makes no reference to a particular coordinate system. Thus we see that using the formula given above, the curvature is well defined and independent of our choice of origin for the coordinate system.

c. The Gauss-Bonnet Theorem. We are now in a position to state and prove the Gauss-Bonnet theorem for a geodesic triangle. In the following, we write dS for an infinitesimal area element.

THEOREM 11. *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 excess:*

$$\int_\Delta \kappa\, dS = \alpha + \beta + \gamma - \pi \tag{12}$$

Proof. Choose geodesic polar coordinates centred at A; then 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, 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 10. *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. Exercise. □

Using this lemma, we may continue the above computations and write the integral as

$$\begin{aligned}\alpha + \int_{\theta=0}^{\theta=\alpha} d\psi &= \alpha + \psi(\theta)\Big|_{\theta=0}^{\theta=\alpha} \\ &= \alpha + \gamma - (\pi - \beta) \\ &= \alpha + \beta + \gamma - \pi\end{aligned}$$

which completes the proof. □

Now we can show that the definition of curvature (10) at a given point q does not depend on the choice of the center point p for the geodesic polar coordinate system as long as p is different from q. For, (12) implies that curvature can be defined intrinsically as the the limit of the ratio of the angular excess of a geodesic triangle to its area as all vertices of the triangle converge to q.

It remains to justify the definition (11) which uses the point p itself as the center. For this we only need to show that the second derivative g_{rr} vanishes at zero. But if it does not there would be points near p where g_{rr}/g has arbitrary large absolute value and hence by continuity using (12) we conclude that there are arbitrary small geodesic triangles with the ration of absolute value of the angular excess or defect arbitrary large. But this would contradict the fact that is we choose another point as the center the corresponding expression g_{rr}/g with be uniformly bounded in absolute value near p.

If we consider the boundary of the triangle as a single closed curve, then it is a smooth curve which is a geodesic at all but three points, where it has a corner. The content of the Gauss-Bonnet 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.

As in important corollary of Theorem (11) we obtain another classical description of the Euler characteristic.

THEOREM 12 (Gauss-Bonnet). *For any Riemannian metric on a compact surface S,*

$$\int_S \kappa \, dS = 2\pi\chi(S).$$

Gauss-Bonnet theorem is deduced from Theorem (11) 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 that 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 at the part of the picture inside a coordinate chart we obtain a decomposition into geodesic polygons which then can be further triangulated. A detailed
Do it here justification of this procedure will be given later.

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 ambient space. Given a particular chart $\phi : (x, y, z) \mapsto (u, v)$, we may do our calculations of curvature and other geometric properties in terms of $x(u, v)$, $y(u, v)$, and $z(u, v)$, 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 as p. On a positively curved surface such as a sphere or an ellipsoid, both principal curvatures are positive; on the other hand, a hyperboloid has one positive principal curvature, and one negative.

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

Smooth and Combinatorial Structure revisited

5.1. Lecture 34: Monday, Nov. 26

a. More on indices. In examining vector fields, curves, etc. on a smooth surface S, there is a natural ambiguity in our discussion of the 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 our present purposes, we shall choose the former and refer to the index of a curve γ with respect to a point x, denoted $\mathrm{ind}_x\gamma$.

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

$$\gamma : S^1 \to S$$

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 can 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 $x \in S \setminus \gamma(S^1)$, we define the index of the γ around x by means of a circle map $\phi_{x,\gamma}$. This map is defined by

$$\begin{array}{rccc} \phi_{x,\gamma} & : & S^1 & \to & S^1 \\ & & t & \mapsto & \frac{\gamma(t)-x}{\|\gamma(t)-x\|} \end{array}$$

where the difference is computed in some choice of local coordinates. The index is given by

$$\mathrm{ind}_x\gamma = \deg \phi_{x,\gamma}$$

Note that the quantity $\|\gamma(t) - x\|$ is nonvanishing 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? The answer to the latter question turns out to be very important. To begin, note that the complement $S \setminus \gamma(S^1)$ 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 13. *Let C be a connected component of $S \setminus \gamma(S^1)$. Then $\mathrm{ind}_x\gamma$ 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 $\delta : [0,1] \to C$ such that $\delta(0) = x_0$, $\delta(1) = x_1$. Now define $f : [0,1] \to \mathbb{Z}$ by

$$f(t) = \mathrm{ind}_{\delta(t)}\gamma$$

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

$$\mathrm{ind}_{x_0}\gamma = f(0) = f(1) = \mathrm{ind}_{x_1}\gamma \quad \square$$

How does the index change if x passes from one connected component to another? In the simplest case, we consider a point at which the curve is smooth, regular, and injective. Formally, we assume that $t \in S^1$ is such that γ is smooth at t and $\gamma(t)$ is noncritical; 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 $\gamma(s) = \gamma(t)$, it can be shown that as x passes from one connected component to another through the point $\gamma(t)$, the index $\mathrm{ind}_x\gamma$ 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 $\mathbb{C}^0$ topology, which is generated by the following metric:

$$d(\gamma_1, \gamma_2) = \max_{t \in S^1} d(\gamma_1(t), \gamma_2(t))$$

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

$$\mathrm{ind}_x\tilde{\gamma} = \mathrm{ind}_x\gamma$$

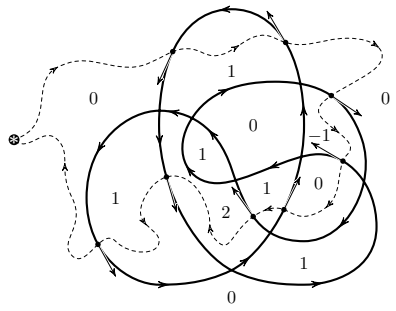

FIGURE 1. Index of points w.r.t. a curve

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 $\mathrm{ind}_x \gamma_s$ is constant.

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 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 13. *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 2. *Every polynomial map $p : \mathbb{C} \to \mathbb{C}$ is surjective—for every $c \in \mathbb{C}$ there exists $z \in \mathbb{C}$ such that $p(z) = c$.*

COROLLARY 3. *Every polynomial with complex coefficients factors as a product of linear terms—given any $p \in \mathbb{C}[z]$ there exist $a, z_1, \ldots, z_n \in \mathbb{C}$ such that*

$$p(z) = a(z - z_1) \ldots (z - z_n)$$

Proof of theorem. Consider the circle of radius r around the origin:

$$C_r = \{z \in \mathbb{C} : |z| = r\}$$

C_r is homeomorphic to S^1 via a simple homothety, and so the restriction of p to C_r defines a curve

$$\gamma_r : S^1 \to \mathbb{C}$$

Now γ_r gives a continuous family of curves with $r \in [0, \infty)$. We consider the index $\mathrm{ind}_0\gamma_r$ 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 $\mathrm{ind}_0\gamma_r$ is constant. Since $\gamma_0(S^1)$ is just a point, the associated circle map is a constant map, and we have

$$\mathrm{ind}_0\gamma_r = 0$$

for every $r \geq 0$.

We now claim that this fails for very large values of r:

LEMMA 11. *For sufficiently large values of r, we have*

$$\mathrm{ind}_0\gamma_r = \deg p$$

Proof of lemma. Let $n = \deg p$, and write

$$p(z) = a_n z^n + q(z)$$

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 $\mathrm{ind}_0\Gamma_r = n$. It remains to show that γ_r and Γ_r have the same index around the origin.

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 $|q(z)| \leq C|z|^{n-1}$, 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 $\mathrm{ind}_0\gamma_r^s$ is constant in s, and so $\mathrm{ind}_0\gamma_r = \mathrm{ind}_0\Gamma_r = n$. □

The lemma establishes our desired contradiction, and so there must exist some $z \in \mathbb{C}$ with $p(z) = 0$. □

5.2. Lecture 35: Wednesday, Nov. 28

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 and prove momentarily. 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 14 (Jordan Curve Theorem). *Let $\gamma : S^1 \mapsto \mathbb{R}^2$ 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 on the sphere. In fact a stronger result holds which we formulate for the sphere case.

THEOREM 15 (Schoenflies). *Let $\gamma : S^1 \mapsto S^2$ be a homeomorphism onto its image. Then $S^2 \setminus \gamma(S^1)$ consists of two connected components U_1 and U_2 such that there are homeomorphisms between the closed disc and both $U_1 \cup \gamma(S^1)$ and $U_2 \cup \gamma(S^1)$.*

The proofs of these theorems use approximation of an arbitrary continuous curve γ with smooth or piece-wise smooth curves, and now we only consider such curves.

THEOREM 16. *Let $\gamma : S^1 \mapsto \mathbb{R}^2$ 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 $\mathrm{ind}_x\gamma$ 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. Then for points $\gamma(t) \notin U$, the vectors $\gamma(t) - x_1$ 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 $\gamma(S^1) \cap U = \gamma((a, b))$. 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 to differential topology. We state and prove a lemma for curves on surfaces—in general, an analogous result holds for submanifolds of higher-dimensional smooth manifolds.

LEMMA 12. *Given a smooth regular curve $\gamma : S^1 \to S$ without self-intersections on an orientable smooth surface S, there exists a neighborhood $U \supset \gamma(S^1)$ and a diffeomorphism $\Gamma : A \to U$, where A 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 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 due to orientability the positive direction $\eta_t'(1)$ varies continuously with t.

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 $\gamma(S^1)$, the result follows. □

The index argument depends on the global structure of the plane. On compact orientable surfaces other than the sphere existence of the tubular neighborhood does not guarantee that that the curve separate the surface

since points in the two halves of the neighborhood 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 12 holds for some curves and does not hold for others. This of course depends on what happens with a normal direction when it is carries around the curve. If it changes orientation as, for example, for the middle circle of the Mobius strip the neighborhood remains connected after the curve itself is removed. We will see that this leads to different relations between the genus g and the Euler characteristic χ for orientable ($\chi = 2 - 2g$) and non-orientable ($\chi = 2 - g$) surfaces.

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 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 set of points and drawing geodesics between them yields a triangulation. In fact we used existence of a triangulation into geodesic triangles in our proof of Gauss-Bonnet theorem. 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 17. *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 2 for the case $g = 2$, we see that $N \geq 2$.

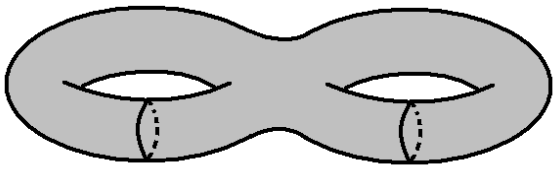

FIGURE 2. 2 disjoint curves which do not disconnect a surface of genus 2

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.

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 16, and so $N \leq g$.

Now consider a sphere with q Möbius caps. Those caps are disjoint closed curve and removal of all those curves still leaves a connected surface. Thus $N \geq q$. Consider any collection of $q + 1$ disjoint closed (smooth non self-intersecting) curves. again we can assume there is a triangulation where each curve is a collection of edges. Removing a curve makes either two holes (if a tubular neighborhood exists) or with one (otherwise). Filling each hole increase Euler characteristic by one and since the maximal Euler characteristic of a connected surface is two and the only such surface is the sphere we use Theorem 16 again and deduce that $q + 1$ curves divide the surface. □

Now it is natural to try to proof the full Jordan Theorem 14 by approximating a given continuous non self-intersecting curve γ by a sequence of smooth curves and apply Theorem 16 to those curves. Since γ is given in local coordinates by a pair of continuous functions those can be easily approximated by smooth functions and approximations "glued together" using partition of unity. Tho problems appear however: 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 1, see in particular Figure 13.

5.3. Lecture 36: Friday, Nov. 30

a. A remark on tubular neighbourhoods. One of the hypotheses in the statement of the lemma on tubular neighbourhoods was the 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.

A proper answer to the question of which curves on a non-orientable surface S admit tubular neighbourhoods and which do not requires an understanding of the fundamental group, which we will not examine in any detail 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.

b. Jordan Curve Theorem. We present now a proof of the Jordan Curve Theorem for arbitrary continuous curves without self-intersections. As mentioned last time, 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.

Step 1. Because S^1 is compact, continuity of $\gamma : S^1 \mapsto \mathbb{R}^2$ 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, or polygon, with vertices at $\gamma(k/N)$ for $k = 0, \ldots, N$. 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 $\varepsilon > 0$ there exists $\delta > 0$ (here ε and δ bear no relation to step 1) such that $\|\gamma(t_1) - \gamma(t_2)\| < \delta$ implies $|t_1 - t_2| < \varepsilon$. Hence if $\tilde{\gamma}$ approximates γ to within δ, we can only have $\tilde{\gamma}(t_1) = \tilde{\gamma}(t_2)$ if $|t_1 - t_2| < \varepsilon$.

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 $t_1^a < t_1^b < t_1^a + \varepsilon$, and we may similarly find $t_i^a < t_i^b < t_i^a + \varepsilon$ for $i = 2, \ldots, n$ 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 $i = 0, \ldots, n$. $\bar{\gamma}(S^1)$ 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 16, 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. Thus it is topologically a disc—there exists a homeomorphism $h : D^2 \to \bar{U} = U \cup \bar{\gamma}(S^1)$. Denote by D_r^2 the disc with radius r—for $r < 1$, this is D^2 with an neighbourhood of the boundary removed. Take $r < 1$ as large as possible, but small enough that $h(D_r^2) \cap \gamma(S^1) = \emptyset$, that is, that the homeomorphic image of D_r^2 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 $U_1 \subset U_2 \subset \ldots$ such that every point x in each U_i has nonzero 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 step comes in step 3, when we are choosing a better refinement $\bar{\gamma}$, to 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 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 formula. Consider now 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 18 (Poincaré-Hopf). *Under the conditions above, the Euler characteristic is the sum of the indices of the critical points:*

$$\sum_{V(x)=0} \operatorname{ind}_x V = \chi(S) \tag{13}$$

We postpone a proof of this result, and instead offer 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 $\sqrt{x^2 + y^2} = \sqrt{1 - z^2}$. This vanishes at the poles and is nonzero 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.

5.4. Lecture 37: Monday, Dec. 3

a. Proof of the Poincaré-Hopf Index Formula. We conclude these notes with a proof of the Poincaré-Hopf Index Formula (13). As before, S is a compact surface, and V is a continuous vector field on S with isolated zeroes.

Step 1. It suffices to consider orientable surfaces, because any non-orientable surface S has a standard orientable double cover $\pi : \tilde{S} \to S$. 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). Thus we decompose S as such a union $D_1 \cup D_2/\sim$, and obtain two discs with g holes, as shown in figure 3.

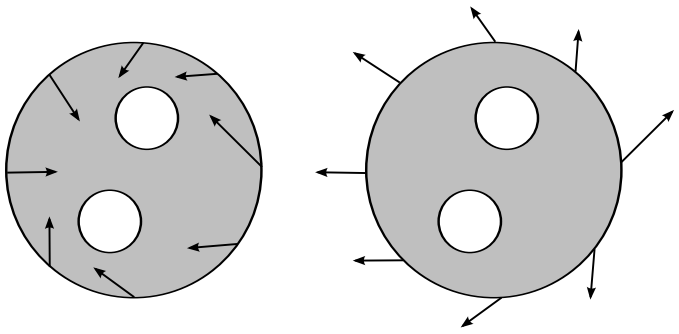

FIGURE 3. The decomposition of a surface of genus 2

By using the two-disc model of our surface, we can now work with vector fields in the plane. The vector field V on our surface S passes to vector fields V_j on the two domains D_j, as shown in figure 3. For simplicity of representation, V_j has only been drawn along the exterior boundary—in fact, it is defined on the entire domain, including, crucially, the boundaries of the holes. We will be particularly interested in V_j on the boundaries of the domain—we denote these curves by $\gamma_0, \gamma_1, \ldots, \gamma_g$, where γ_0 is the exterior boundary (the large circle), and $\gamma_1, \ldots, \gamma_g$ 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, in 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 nonvanishing on each curve γ_j. Assign the positive orientation (counterclockwise) to γ_0, and the negative orientation (clockwise) to the other curves $\gamma_1, \ldots, \gamma_g$—then we may define the index of the vector field with respect to the composite boundary as

$$\mathrm{ind}_{D_j} V_j = \sum_{i=0}^{g} \mathrm{ind}_{\gamma_i} V_j$$

It remains to relate this sum to the indices of the zeroes of V, and to relate the values of $\mathrm{ind}_{\gamma_i} V_1$ and $\mathrm{ind}_{\gamma_i} V_2$, since as indicated in figure 3, 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.

Step 4. In fact, we find that $\mathrm{ind}_{D_j} V_j$ is the sum of the indices of the zeroes contained in D_j:

$$\mathrm{ind}_{D_j} V_j = \sum_{\substack{x \in D \\ V_j(x)=0}} \mathrm{ind}_x V_j$$

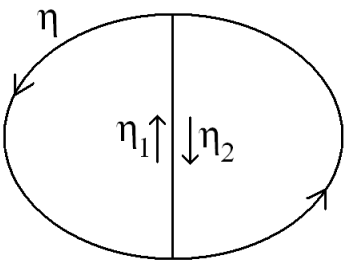

FIGURE 4. Decomposing a curve η

To see this, consider a closed curve η, and decompose η as the composition of η_1 and η_2 (that is, following first one, then the other) as shown in figure 4. 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 nonvanishing vector field along η, an examination of the associated circle maps shows that $\mathrm{ind}_\eta V = \mathrm{ind}_{\eta_1} V + \mathrm{ind}_{\eta_2} V$.

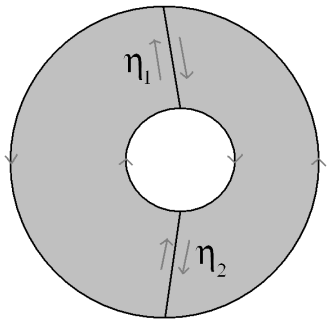

FIGURE 5. Decomposing the boundary of D

We may carry out a similar decomposition on our domains D_j. Figure 5 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

$$\mathrm{ind}_D V = \mathrm{ind}_{\gamma_0} V + \mathrm{ind}_{\gamma_1} V = \mathrm{ind}_{\eta_1} V + \mathrm{ind}_{\eta_2} V$$

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

$$\mathrm{ind}_{D_1} V_1 + \mathrm{ind}_{D_2} V_2 = \sum_{V(x)=0} \mathrm{ind}_x V$$

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 $\mathrm{ind}_{\gamma_i} V_1 + \mathrm{ind}_{\gamma_i} V_2 = -2$, while for $i = 0$, the sum is 2. Then summing over all values of i and applying the formula from step 4 will give

$$\sum_{V(x)=0} \mathrm{ind}_x V = 2 - 2g = \chi(S)$$

so it only remains to prove the claim. We see that the difference is 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 3.

To make this more formal, 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

$$\mathrm{ind}_{\gamma_0} V_1 + \mathrm{ind}_{\gamma_0} V_2 = 2$$

which completes our proof. □

As a corollary to this theorem, we can extend our earlier result connecting Morse functions and Euler characteristic to a result valid for any smooth function with isolated critical points, possibly degenerate, by considering the indices of the zeroes of the gradient vector field.

b. The ubiquitous Euler characteristic. Euler characteristic has appeared in many guises throughout this course from combinatorial, algebraic, differentiable (smooth functions and ODE), and geometric considerations.

Let us summarize different ways in which Euler characteristic appears for the surfaces. Let S be a compact closed surface which admits a map or, equivalently, a smooth structure, see Lecture 33, Section **??**.[1]. Then $\chi(S)$, the Euler characteristic of S, is equal to any of the following:

(1) $\#(faces) - \#(edges) + \#(vertices)$ *for any map (in particular, a triangulation) of* S*;*
(2) $\beta_2 - \beta_1 + \beta_0$ *where* β_i, $i = 0, 1, 2$ *are Betti numbers which appear from the chain complex associated with any triangulation or map of* S*;*
(3) $\#(maxima) - \#(saddles) + \#(minima)$ *for any Morse function on* S*;*
(4) *Sum of the indices of critical points for any differentiable function on* M *with finitely many critical points;*

[1] In fact, any surface admits those structures but we have not proved that in this course.

(5) *Sum of the indices of zeroes of any continuous vector field with finitely many zeroes;*
(6) *Integral of curvature with respect to any Riemannian metric on S divided by 2π.*

Euler characteristic is a prototype of a topological invariant for a manifold which can be expressed through various structures. In fact, on even-dimensional compact manifolds Euler characteristics can be defined the same way as in (2) and some but not all of its guises extend to this case. For orientable odd–dimensional manifolds expression (2) vanishes which is itself is a manifestation of one of the most remarkable facts in topology – Poincaré duality.