GEOMETRY OF SURFACES

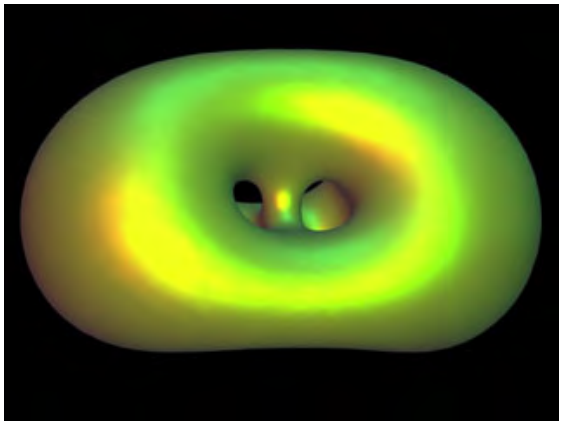
b3 course 2004

Nigel Hitchin

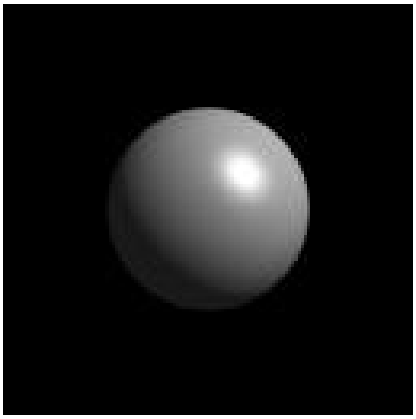
hitchin@maths.ox.ac.uk

1 Introduction

This is a course on surfaces. Your mental image of a surface should be something like this:



or this

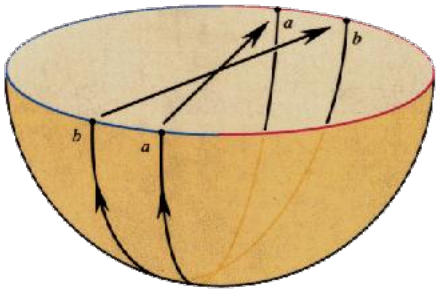


However we are also going to try and consider surfaces intrinsically, or abstractly, and not necessarily embedded in three-dimensional Euclidean space like the two above. In fact lots of them simply can't be embedded, the most notable being the projective plane. This is just the set of lines through a point in $\mathbf{R}^3$ and is as firmly connected with familiar Euclidean geometry as anything. It *is* a surface but it doesn't sit in Euclidean space.

If you insist on looking at it, then it maps to Euclidean space like this

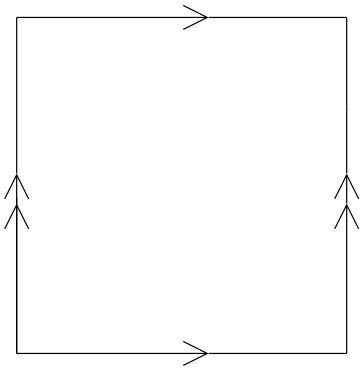
– called *Boy's surface*. This is not one-to-one but it does intersect itself reasonably cleanly.

A better way to think of this space is to note that each line through 0 intersects the unit sphere in two opposite points. So we cut the sphere in half and then just have to identify opposite points on the *equator*:

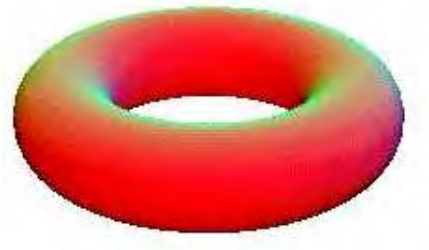


... and this gives you the projective plane.

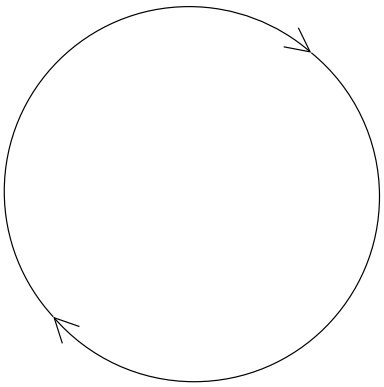
Many other surfaces appear naturally by taking something familiar and performing identifications. A doubly periodic function like $f(x, y) = \sin 2\pi x \cos 2\pi y$ can be thought of as a function on a surface. Since its value at (x, y) is the same as at $(x+m, y+n)$ it is determined by its value on the unit square but since $f(x, 0) = f(x, 1)$ and $f(0, y) = f(1, y)$ it is really a continuous function on the space got by identifying opposite sides:



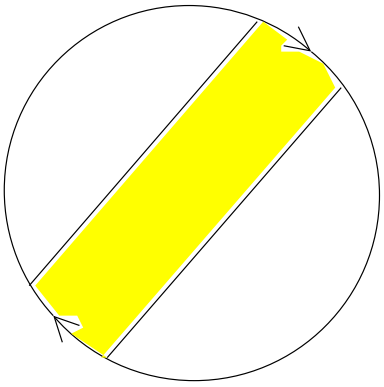
and this is a torus:



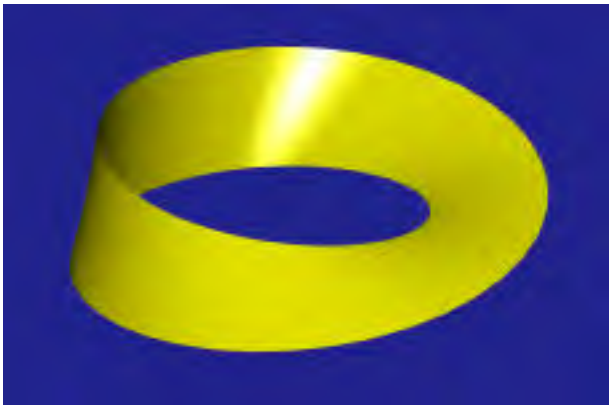
We shall first consider surfaces as topological spaces. The remarkable thing here is that they are completely classified up to homeomorphism. Each surface belongs to two classes – the orientable ones and the non-orientable ones – and within each class there is a non-zero integer which determines the surface. The orientable ones are the ones you see sitting in Euclidean space and the integer is the number of holes. The non-orientable ones are the "one-sided surfaces" – those that contain a Möbius strip – and projective space is just such a surface. If we take the hemisphere above and flatten it to a disc, then projective space is obtained by identifying opposite points on the boundary:



Now cut out a strip:



and the identification on the strip gives the Möbius band:



As for the integer invariant, it is given by the *Euler characteristic* – if we subdivide a surface A into V vertices, E edges and F faces then the Euler characteristic $\chi(A)$ is defined by

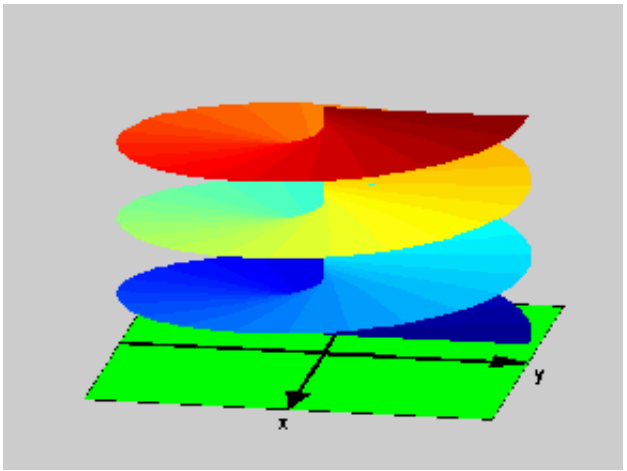
$$\chi(A) = V - E + F.$$

For a surface in Euclidean space with g holes, $\chi(A) = 2 - 2g$. The invariant χ has the wonderful property, like counting the points in a set, that

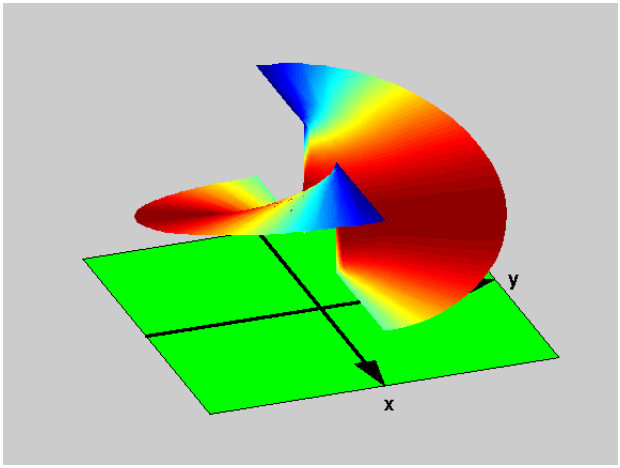
$$\chi(A \cup B) = \chi(A) + \chi(B) - \chi(A \cap B)$$

and this means that we can calculate it by cutting up the surface into pieces, and without having to imagine the holes.

One place where the study of surfaces appears is in complex analysis. We know that $\log z$ is not a single valued function – as we continue around the origin it comes back to its original value with $2\pi i$ added on. We can think of $\log z$ as a single valued function on a surface which covers the non-zero complex numbers:



The Euclidean picture above is in this case a reasonable one, using the third coordinate to give the imaginary part of $\log z$: the surface consists of the points $(re^{i\theta}, \theta) \in \mathbf{C} \times \mathbf{R} = \mathbf{R}^3$ and $\log z = \log r + i\theta$ is single-valued. But if you do the same to $\sqrt{z(z-1)}$ you get

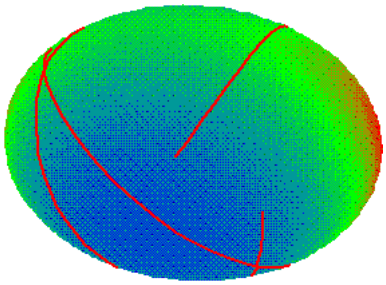


a surface with self-intersections, a picture which is not very helpful. The way out is to leave $\mathbf{R}^3$ behind and construct an abstract surface on which $\sqrt{z(z-1)}$ is single-valued. This is an example of a *Riemann surface*. Riemann surfaces are always orientable, and for $\sqrt{z(z-1)}$ we get a sphere. For $\sqrt{z(z-1)(z-a)}$ it is a torus, which amongst other things is the reason that you can't evaluate

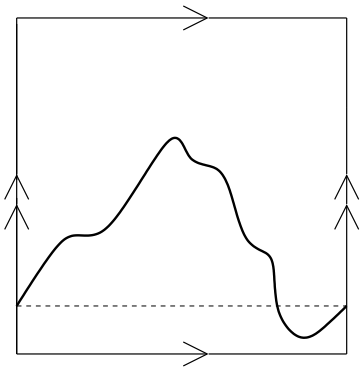
$$\int \frac{dx}{x(x-1)(x-a)}$$

using elementary functions. In general, given a multi-valued meromorphic function, the Euler characteristic of the Riemann surface on which it is defined can be found by a formula called the Riemann-Hurwitz formula.

We can look at a smooth surface in Euclidean space in many ways – as a topological space as above, or also as a *Riemannian manifold*. By this we mean that, using the Euclidean metric on $\mathbf{R}^3$, we can measure the lengths of curves on the surface.



If our surface is not sitting in Euclidean space we can consider the same idea, which is called a Riemannian metric. For example, if we think of the torus by identifying the sides of a square, then the ordinary length of a curve in the plane can be used to measure the length of a curve on the torus:



A Riemannian metric enables you to do much more than measure lengths of curves: in particular you can define areas, curvature and geodesics. The most important notion of curvature for us is the Gaussian curvature which measures the deviation of formulas for triangles from the Euclidean ones. It allows us to relate the differential geometry of the surface to its topology: we can find the Euler characteristic by integrating the Gauss curvature over the surface. This is called the *Gauss-Bonnet theorem*. There are other analytical ways of getting the Euler characteristic – one is to count the critical points of a differentiable function.

Surfaces with constant Gaussian curvature have a special role to play. If this curvature is zero then locally we are looking at the Euclidean plane, if positive it is the round sphere, but the negative case is the important area of hyperbolic geometry. This has a long history, but we shall consider the concrete model of the upper half-plane as a surface with a Riemannian metric, and show how its geodesics and isometries provide the axiomatic properties of non-Euclidean geometry and also link up with complex analysis. The hyperbolic plane is a surface as concrete as one can imagine, but is an abstract one in the sense that it is not in $\mathbf{R}^3$.

2 The topology of surfaces

2.1 The definition of a surface

We are first going to consider surfaces as topological spaces, so let's recall some basic properties:

Definition 1 *A topological space is a set X together with a collection $\mathcal{T}$ of subsets of X (called the 'open subsets' of X) such that*

- $\emptyset \in \mathcal{T}$ *and* $X \in \mathcal{T}$;
- *if* $U, V \in \mathcal{T}$ *then* $U \cap V \in \mathcal{T}$;
- *if* $U_i \in \mathcal{T} \quad \forall i \in I$ *then* $\bigcup_{i \in I} U_i \in \mathcal{T}$.

- X is called Hausdorff if whenever $x, y \in X$ and $x \neq y$ there are open subsets U, V of X such that $x \in U$ and $y \in V$ and $U \cap V = \emptyset$.
- A map $f : X \to Y$ between topological spaces X and Y is called continuous if $f^{-1}(V)$ is an open subset of X whenever V is an open subset of Y.
- $f : X \to Y$ is called a homeomorphism if it is a bijection and both $f : X \to Y$ and its inverse $f^{-1} : Y \to X$ are continuous. Then we say that X is homeomorphic to Y.
- X is called compact if every open cover of X has a finite subcover.

Subsets of $\mathbf{R}^n$ are Hausdorff topological spaces where the open sets are just the intersections with open sets in $\mathbf{R}^n$. A surface has the property that near any point it looks like Euclidean space – just like the surface of the spherical Earth. More precisely:

Definition 2 *A topological surface (sometimes just called a* surface*) is a Hausdorff topological space X such that each point x of X is contained in an open subset U which is homeomorphic to an open subset V of $\mathbf{R}^2$.*

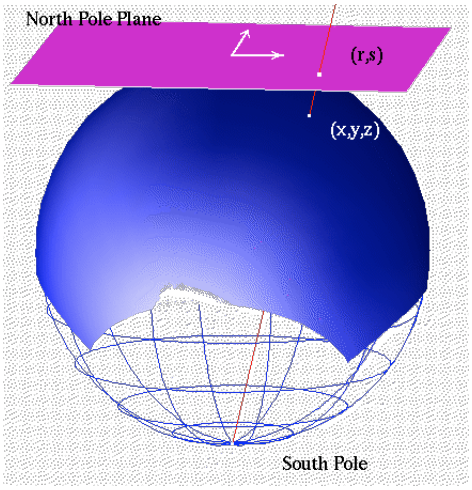
X is called a closed surface if it is compact.

A surface is also sometimes called a 2-*manifold* or a manifold of dimension 2. For any natural number n a topological n-manifold is a Hausdorff topological space X which is locally homeomorphic to $\mathbf{R}^n$.

Remark: (i) The Heine-Borel theorem tells us that a subset of $\mathbf{R}^n$ is compact if and only if it is *closed* (contains all its limit points) and *bounded*. Thus the use of the terminology 'closed surface' for a compact surface is a little perverse: there are plenty of surfaces which are closed subsets of $\mathbf{R}^3$, for example, but which are not 'closed surfaces'.

(ii) Remember that the image of a compact space under a continuous map is always compact, and that a bijective continuous map from a compact space to a Hausdorff space is a homeomorphism.

Example: The sphere. The most popular way to see that this is a surface according to the definition is stereographic projection:



Here one open set U is the complement of the South Pole and projection identifies it with $\mathbf{R}^2$, the tangent plane at the North Pole. With another open set the complement of the North Pole we see that all points are in a neighbourhood homeomorphic to $\mathbf{R}^2$.

We constructed other surfaces by identification at the boundary of a planar figure. Any subset of the plane has a topology but we need to define one on the space obtained by identifying points. The key to this is to regard identification as an *equivalence relation*. For example, in constructing the torus from the square we define $(x, 0) \sim (x, 1)$ and $(0, y) \sim (1, y)$ and every other equivalence is an equality. The torus is the set of *equivalence classes* and we give this a topology as follows:

Definition 3 *Let $\sim$ be an equivalence relation on a topological space X. If $x \in X$ let $[x]_\sim = \{y \in X : y \sim x\}$ be the equivalence class of x and let*

$$X/\sim = \{[x]_\sim : x \in X\}$$

be the set of equivalence classes. Let $\pi : X \to X/\sim$ be the 'quotient' map which sends an element of X to its equivalence class. Then the quotient topology on $X/\sim$ is given by

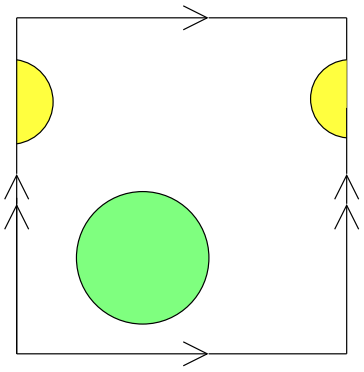
$$\{V \subseteq X/\sim : \pi^{-1}(V) \text{ is an open subset of } X\}.$$

In other words a subset V of $X/\sim$ is an open subset of $X/\sim$ (for the quotient topology) if and only if its inverse image

$$\pi^{-1}(V) = \{x \in X : [x]_\sim \in V\}$$

is an open subset of X.

So why does the equivalence relation on the square give a surface? If a point lies inside the square we can take an open disc around it still in the interior of the square. There is no identification here so this neighbourhood is homeomorphic to an open disc in $\mathbf{R}^2$. If the chosen point lies on the boundary, then it is contained in two half-discs D_L, D_R on the left and right:



We need to prove that the quotient topology on these two half-discs is homeomorphic to a full disc. First take the closed half-discs and set $B = D_L \cup D_R$. The map $x \mapsto x + 1$ on D_L and $x \mapsto x$ on D_R is a continuous map from B (with its topology from $\mathbf{R}^2$) to a single disc D. Moreover equivalent points go to the same point so it is a composition

$$B \to X/\sim \to D.$$

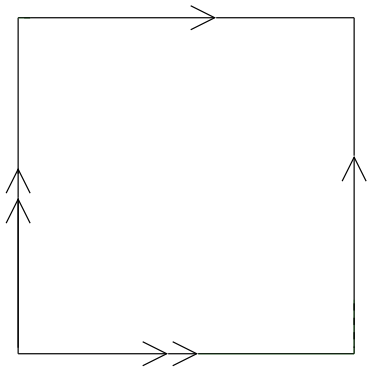
The definition of the quotient topology tells us that $B/\sim \to D$ is continuous. It is also bijective and $B/\sim$, the continuous image of the compact space B, is compact so this is a homeomorphism. Restrict now to the interior and this gives a homeomorphism from a neighbourhood of a point on the boundary of the square to an open disc.

If the point is a corner, we do a similar argument with quadrants.

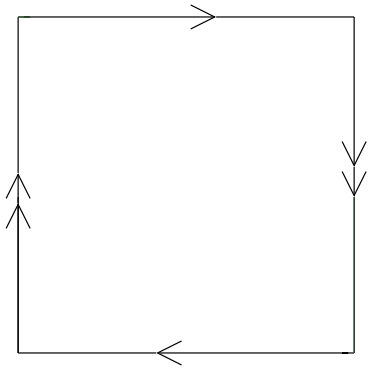
Thus the torus defined by identification is a surface. Moreover it is closed, since it is the quotient of the unit square which is compact.

Here are more examples by identification of a square:

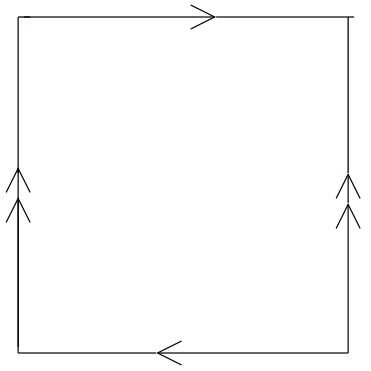
- *The sphere*



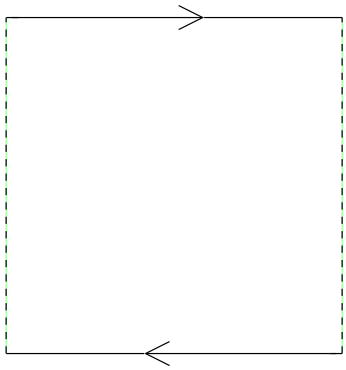
- *Projective space*



- *The Klein bottle*



- *The Möbius band*



The Möbius band is not closed, as the dotted lines suggest. Here is its rigorous definition:

Definition 4 *A Möbius band (or Möbius strip) is a surface which is homeomorphic to*

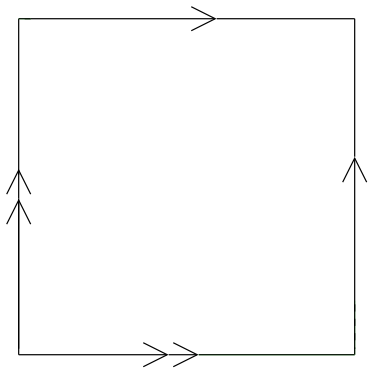
$$(0,1) \times [0,1]/\sim$$

with the quotient topology, where $\sim$ is the equivalence relation given by

$$(x,y) \sim (s,t) \text{ iff } (x = s \text{ and } y = t) \text{ or } (x = 1 - s \text{ and } \{y,t\} = \{0,1\}).$$

2.2 Planar models and connected sums

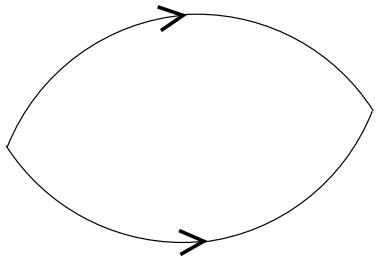
The examples above are obtained by identifying edges of a square but we can use any polygon in the plane with an even number of sides to construct a closed surface so long as we prescribe the way to identify the sides in pairs. Drawing arrows then becomes tiresome so we describe the identification more systematically: going round clockwise we give each side a letter a say, and when we encounter the side to be identified we call it a if the arrow is in the same clockwise direction and a^{-1} if it is the opposite. For example, instead of



we call the top side a and the bottom b and get

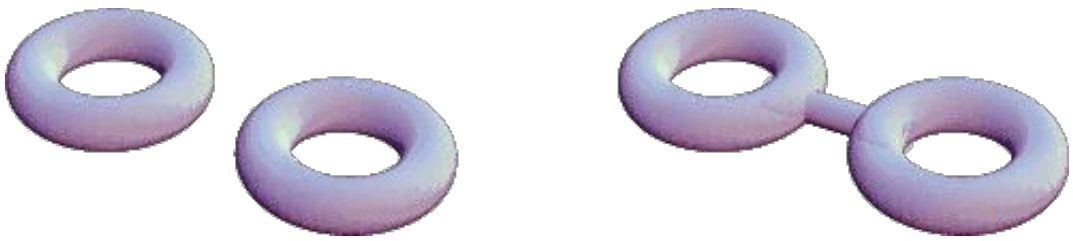
$$aa^{-1}bb^{-1}.$$

This is the sphere. Projective space is then $abab$, the Klein bottle $abab^{-1}$ and the torus $aba^{-1}b^{-1}$. Obviously the cyclic order is not important. There are lots of planar models which define the same surface. The sphere for example can be defined not just from the square but also by aa^{-1}, a 2-sided polygon:



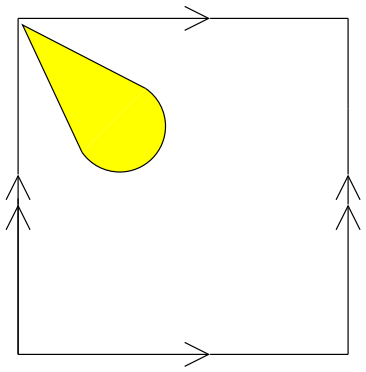
and similarly the projective plane is aa.

Can we get new surfaces by taking more sides? Certainly, but first let's consider another construction of surfaces. If X and Y are two closed surfaces, remove a small closed disc from each. Then take a homeomorphism from the boundary of one disc to the boundary of the other. The topological space formed by identifying the two circles is also a surface called the *connected sum* $X\#Y$. We can also think of it as joining the two by a cylinder:

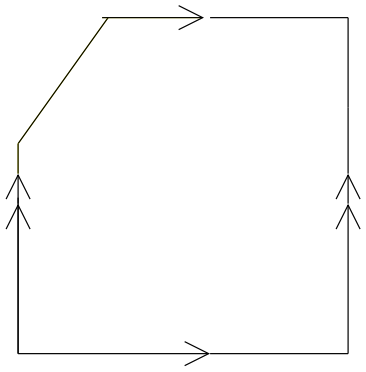


The picture shows that we can get a surface with two holes from the connected sum of two tori. Let's look at this now from the planar point of view.

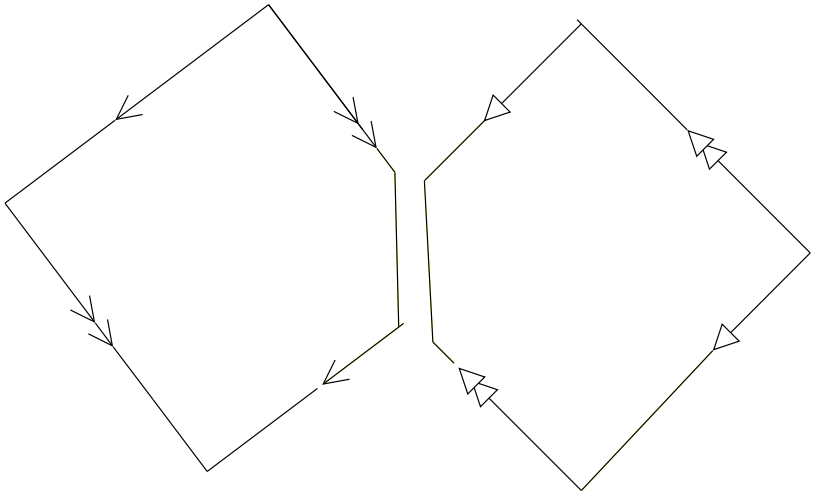
First remove a disc which passes through a vertex but otherwise misses the sides:



Now open it out:



and paste two copies together:



This gives an octagon, and the identification is given by the string of letters:

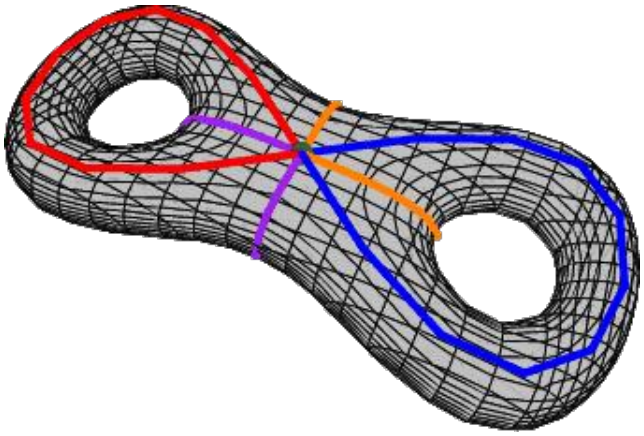
$$aba^{-1}b^{-1}cdc^{-1}d^{-1}.$$

It's not hard to see that this is the general pattern: a connected sum can be represented by placing the second string of letters after the first. So in particular

$$a_1b_1a_1^{-1}b_1^{-1}a_2b_2a_2^{-1}b_2^{-1}\dots a_gb_ga_g^{-1}b_g^{-1}$$

describes a surface in $\mathbf{R}^3$ with g holes.

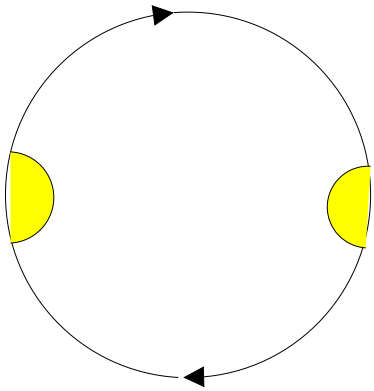
Note that when we defined a torus from a square, all four vertices are equivalent and this persists when we take the connected sum as above. The picture of the surface one should have then is $2g$ closed curves emanating from a single point, and the complement of those curves is homeomorphic to an open disc – the interior of the polygon.



If S is a sphere, then removing a disc just leaves another disc so connected sum with S takes out a disc and replaces it. Thus

$$X\#S = X.$$

Connected sum with the projective plane P is sometimes called *attaching a cross-cap.* In fact, removing a disc from P gives the Möbius band

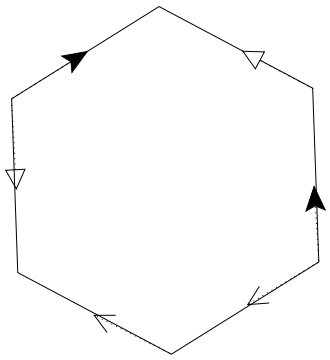


so we are just pasting the boundary circle of the Möbius band to the boundary of the disc. It is easy to see then that the connected sum $P\#P$ is the Klein bottle.

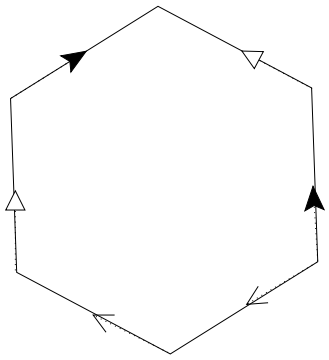
You can't necessarily cancel the connected sum though: it is not true that $X\#A = Y\#A$ implies $X = Y$. Here is an important example:

Proposition 2.1 *The connected sum of a torus T and the projective plane P is homeomorphic to the connected sum of three projective planes.*

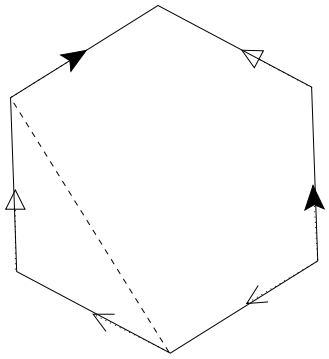
Proof: From the remark above it is sufficent to prove that $P\#T = P\#K$ where K is the Klein bottle. Now since P can be described by a 2-gon with relation aa and the Klein bottle is $bcbc^{-1}$, $P\#K$ is defined by a hexagon and the relation $aabcbc^{-1}$.



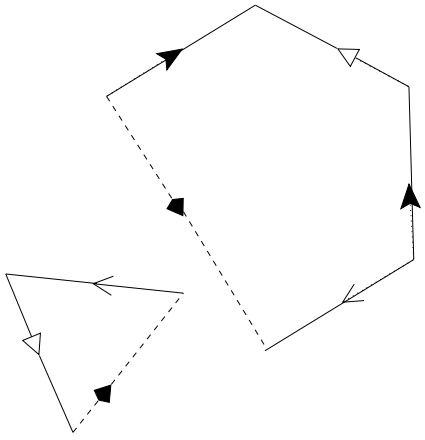
Now $P\#T$ is $aabcb^{-1}c^{-1}$:



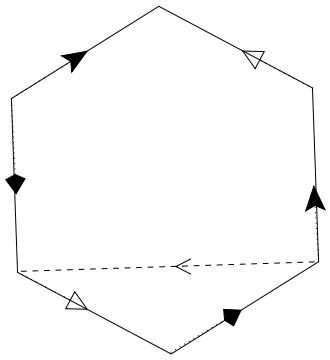
Cut along the dotted line...



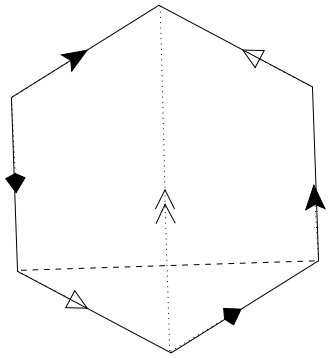
... detach the triangle and turn it over...



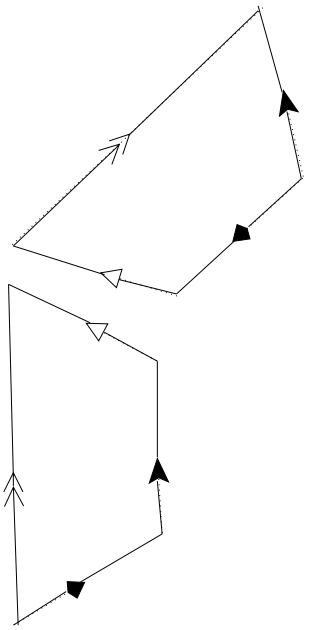
... reattach...



... cut down the middle...



... turn the left hand quadrilateral over and paste together again...



...and this is $aabcbc^{-1}$.

□

2.3 The classification of surfaces

The planar models allow us to classify surfaces. We shall prove the following

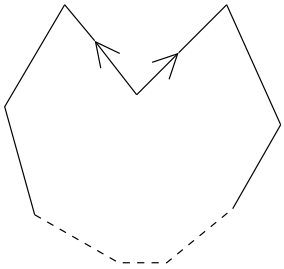
Theorem 2.2 *A closed, connected surface is either homeomorphic to the sphere, or to a connected sum of tori, or to a connected sum of projective planes.*

We sketch the proof below (*this is not examinable*) and refer to [3] or [1] for more details. We have to start somewhere, and the topological definition of a surface is quite general, so we need to invoke a theorem beyond the scope of this course: any closed surface X has a *triangulation*: it is homeomorphic to a space formed from the disjoint union of finitely many triangles in $\mathbf{R}^2$ with edges glued together in pairs. For a Riemann surface (see next section), we can directly find a triangulation so long as we have a meromorphic function, and that is also a significant theorem, so we can't escape this starting point.

We shall proceed by using a planar model but there is also an alternative proof in [2] (or download from here: **new.math.uiuc.edu/zipproof/zipproof.pdf**) if you don't object to surfaces covered with zip fasteners.

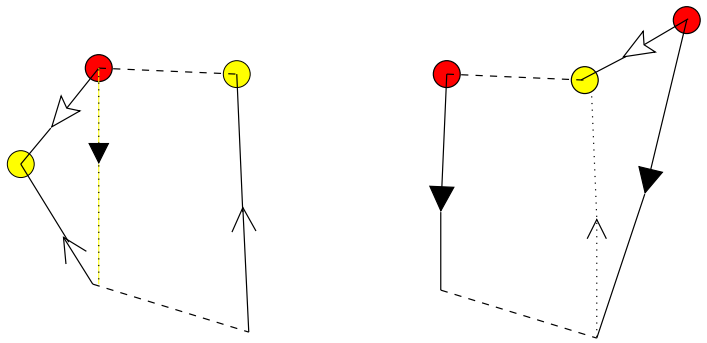
Now take one triangle on the surface, and choose a homeomorphism to a planar triangle. Take an adjacent one and the common edge and choose a homeomorphism to another plane triangle and so on... Since the surface is connected the triangles form a polygon and thus X can be obtained from this polygon with edges glued together in pairs. It remains to systematically reduce this, without changing the homeomorphism type, to a standard form.

Step 1: Adjacent edges occurring in the form aa^{-1} or $a^{-1}a$ can be eliminated.



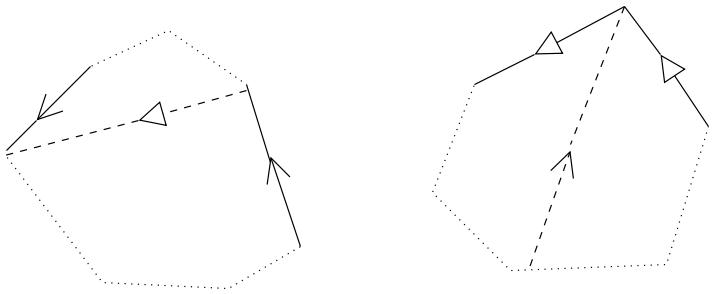
Step 2: We can assume that all vertices must be identified with each other. To see this, suppose Step 1 has been done, and we have two adjacent vertices in different

equivalence classes: red and yellow. Because of Step 1 the other side going through the yellow vertex is paired with a side elsewhere on the polygon. Cut off the triangle and glue it onto that side:

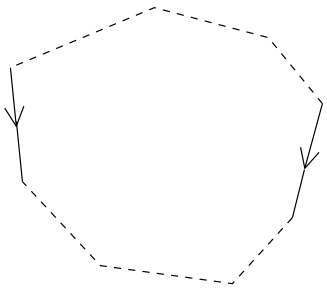


The result is the same number of sides but one less yellow and one more red vertex. Eventually, applying Step 1 again, we get to a single equivalence class.

Step 3: We can assume that any pair of the form a and a are adjacent, by cutting and pasting:



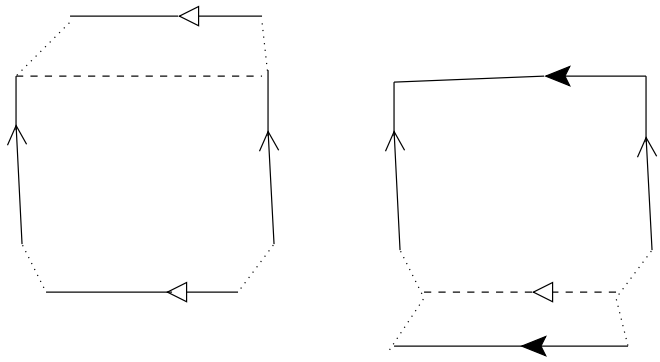
We now have a single equivalence class of vertices and all the pairs a, a are adjacent. What about a pair a, a^{-1}? If they are adjacent, Step 1 gets rid of them, if not we have this:



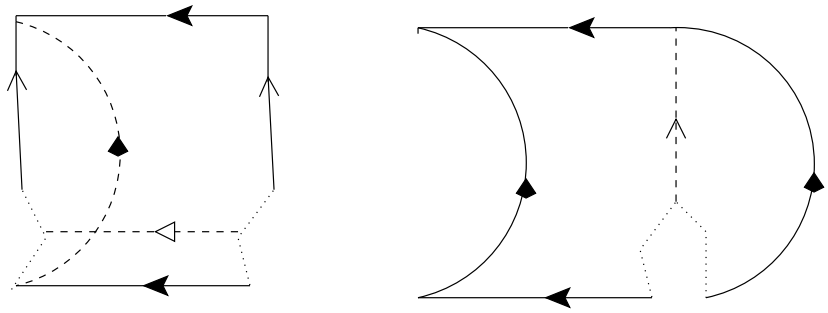
If all the sides on the top part have their partners in the top part, then their vertices will never be equivalent to a vertex in the bottom part. But Step 2 gave us one equivalence class, so there is a b in the top half paired with something in the bottom.

It can't be b because Step 3 put them adjacent, so it must be b^{-1}.

Step 4: We can reduce this to something of the form $cdc^{-1}d^{-1}$ like this. First cut off the top and paste it to the bottom.



Now cut away from the left and paste it to the right.



Finally our surface is described by a string of terms of the form aa or $bcb^{-1}c^{-1}$: a connected sum of projective planes and tori. However, if there is at least one projective plane we can use Proposition 2.1 which says that $P\#T = P\#P\#P$ to get rid of the tori.

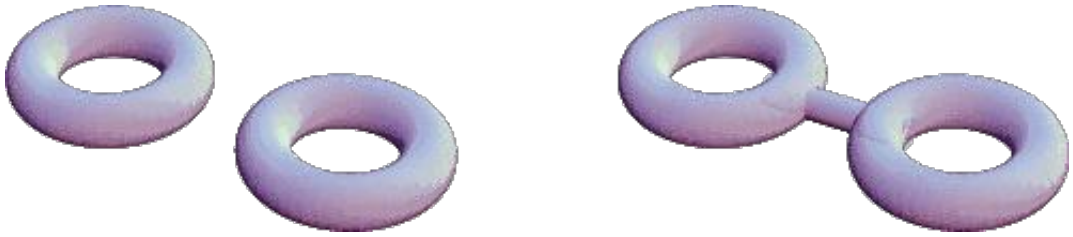
2.4 Orientability

Given a surface, we need to be able to decide what connected sum it is in the Classification Theorem without cutting it into pieces. Fortunately there are two concepts, which are invariant under homeomorphism, which do this. The first concerns orientation:

Definition 5 *A surface X is orientable if it contains no open subset homeomorphic to a Möbius band.*

From the definition it is clear that if X is orientable, any surface homeomorphic to X is too.

We saw that taking the connected sum with the projective plane means attaching a Möbius band, so the surfaces which are connected sums of P are non-orientable. We need to show that connected sums of tori are orientable. For this, we observe that the connected sum operation works for tori in $\mathbf{R}^3$ embedded in the standard way:



so a connected sum of tori can also be embedded in $\mathbf{R}^3$. The sketch proof below assumes our surfaces are differentiable – we shall deal with these in more detail later.

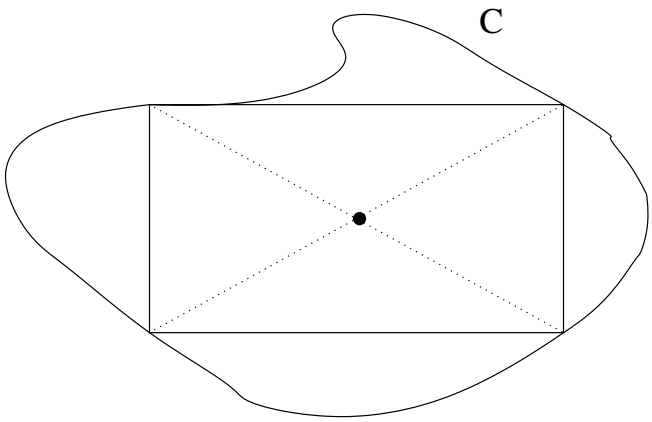
Suppose for a contradiction that X is a non-orientable compact smooth surface in $\mathbf{R}^3$. Then X has an open subset which is homeomorphic to a Möbius band, which means that we can find a loop (i.e. a closed path) in X such that the normal to X, when transported around the loop in a continuous fashion, comes back with the opposite direction. By considering a point on the normal a small distance from X, moving it around the loop and then connecting along the normal from one side of X to the other, we can construct a closed path $\gamma : [0,1] \to \mathbf{R}^3$ in $\mathbf{R}^3$ which meets X at exactly one point and is *transversal* to X at this point (i.e. the tangent to γ at x is not tangent to X). It is a general fact about the topology of $\mathbf{R}^3$ that any closed differentiable path $\gamma : [0,1] \to \mathbf{R}^3$ can be 'filled in' with a disc; more precisely there is a differentiable map $f : D \to \mathbf{R}^3$, where $D = \{(x,y) \in \mathbf{R}^2 | x^2 + y^2 \leq 1\}$, such that

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

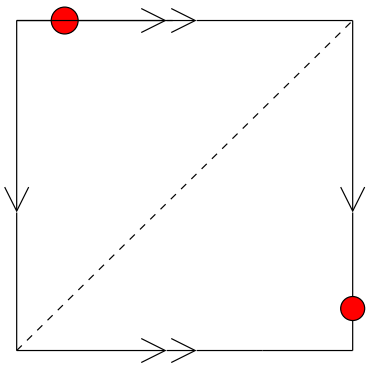
for all $t \in [0,1]$. Now we can perturb f a little bit, without changing γ or the values of f on the boundary of D, to make f transversal to X (i.e. the image of f is not tangent to X at any point of intersection with X). But once f is transversal to X it can be shown that the inverse image $f^{-1}(X)$ of X in D is very well behaved: it consists of a disjoint union of simple closed paths in the interior of D, together with paths meeting the boundary of D in exactly their endpoints (which are two distinct points on the boundary of D). Thus $f^{-1}(X)$ contains an even number of points on the boundary of D, which contradicts our construction in which $f^{-1}(X)$ has exactly one point on the boundary of D. The surface must therefore be orientable.

This argument shows why the projective plane in particular can't be embedded in $\mathbf{R}^3$. Here is an amusing corollary:

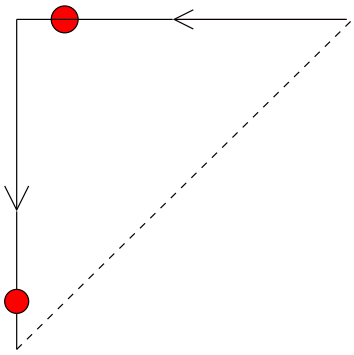
Proposition 2.3 *Any simple closed curve in the plane contains an inscribed rectangle.*



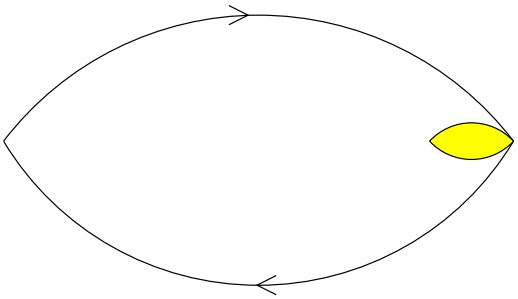
Proof: The closed curve C is homeomorphic to the circle. Consider the set of pairs of points (x, y) in C. This is the product of two circles: a *torus*. We now want to consider the set X of *unordered* pairs, so consider the planar model of the torus. We identify (x, y) with (y, x), which is reflection about the diagonal. The top side then gets identified with the right hand side, and under the torus identification with the left hand side.



The set of unordered points is therefore obtained by identification on the top triangle:



and this is the projective plane with a disc removed (the Möbius band):



Now define a map $f : X \to \mathbf{R}^3$ as follows:

$$(x, y) \mapsto (\frac{1}{2}(x + y), |x - y|) \in \mathbf{R}^2 \times \mathbf{R}$$

The first term is the midpoint of the line xy and the last is the distance between x and y. Both are clearly independent of the order and so the map is well-defined. When $x = y$, which is the boundary circle of the Möbius band, the map is

$$x \mapsto (x, 0)$$

which is the curve C in the plane $x_3 = 0$. Since the curve bounds a disc we can extend f to the surface obtained by pasting the disc to X and extending f to be the inclusion of the disc into the plane $x_3 = 0$. This is a continuous map (it can be perturbed to be differentiable if necessary) of the projective plane P to $\mathbf{R}^3$. Since P is unorientable it can't be an embedding so we have at least two pairs $(x_1, y_1), (x_2, y_2)$ with the same centre and the same separation. These are the vertices of the required rectangle. □

2.5 The Euler characteristic

It is a familar fact (already known to Descartes in 1639) that if you divide up the surface of a sphere into polygons and count the number of vertices, edges and faces then

$$V - E + F = 2.$$

This number is the Euler characteristic, and we shall define it for any surface. First we have to define our terms:

Definition 6 *A subdivision of a compact surface X is a partition of X into*

i) vertices *(these are finitely many points of X),*

ii) edges *(finitely many disjoint subsets of X each homeomorphic to the open interval $(0,1)$), and*

iii) faces *(finitely many disjoint open subsets of X each homeomorphic to the open disc $\{(x,y) \in \mathbf{R}^2 : x^2 + y^2 < 1\}$ in $\mathbf{R}^2$,*

such that

a) the faces are the connected components of $X \setminus \{$vertices and edges$\}$,

b) no edge contains a vertex, and

c) each edge 'begins and ends in a vertex' (either the same vertex or different vertices), or more precisely, if e is an edge then there are vertices v_0 and v_1 (not necessarily distinct) and a continuous map

$$f : [0,1] \to e \cup \{v_0, v_1\}$$

which restricts to a homeomorphism from $(0,1)$ to e and satisfies $f(0) = v_0$ and $f(1) = v_1$.

Definition 7 *The Euler characteristic (or Euler number) of a compact surface X with a subdivision is*

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

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

The fact that a closed surface has a subdivision follows from the existence of a triangulation. The most important fact is

Theorem 2.4 *The Euler characteristic of a compact surface is independent of the subdivision*

which we shall sketch a proof of later.

A planar model provides a subdivision of a surface. We have one face – the interior of the polygon – and if there are $2n$ sides to the polygon, these get identified in pairs so there are n edges. For the vertices we have to count the number of equivalence classes, but in the normal form of the classification theorem, we created a single equivalence class. In that case, the Euler characteristic is

$$1 - n + 1 = 2 - n.$$

The connected sum of g tori had $4g$ sides in the standard model $a_1 b_1 a_1^{-1} b_1^{-1} \dots a_g b_g a_g^{-1} b_g^{-1}$ so in that case $\chi(X) = 2 - 2g$. The connected sum of g projective planes has $2g$ sides so we have $\chi(X) = 2 - g$. We then obtain:

Theorem 2.5 *A closed surface is determined up to homeomorphism by its orientability and its Euler characteristic.*

This is a very strong result: nothing like this happens in higher dimensions.

To calculate the Euler characteristic of a given surface we don't necessarily have to go to the classification. Suppose a surface is made up of the union of two spaces X and Y, such that the intersection $X \cap Y$ has a subdivision which is a subset of the subdivisions for X and for Y. Then since V, E and F are just counting the number of elements in a set, we have immediately that

$$\chi(X \cup Y) = \chi(X) + \chi(Y) - \chi(X \cap Y).$$

We can deal with a connected sum this way. Take a closed surface X and remove a disc D to get a space X^o. The disc has Euler characteristic 1 (a polygon has one face,

n vertices and n sides) and the boundary circle has Euler characteristic 0 (no face). So applying the formula,

$$\chi(X) = \chi(X^o \cup D) = \chi(X^o) + \chi(D) - \chi(X^o \cap D) = \chi(X^o) + 1.$$

To get the connected sum we paste X^o to Y^o along the boundary circle so

$$\chi(X\#Y) = \chi(X^o)+\chi(Y^o)-\chi(X^o \cap Y^o) = \chi(X)-1+\chi(Y)-1-0 = \chi(X)+\chi(Y)-2.$$

In particular, $\chi(X\#T) = \chi(X) - 2$ so this again gives the value $2 - 2g$ for the connected sum of g tori.

To make all this work we finally need:

Theorem 2.6 *The Euler characteristic $\chi(X)$ of a compact surface X is a topological invariant.*

We give a sketch proof (which is not examinable).

Proof:

The idea is to give a different definition of $\chi(X)$ which makes it clear that it is a topological invariant, and then prove that the Euler characteristic of any subdivision of X is equal to $\chi(X)$ defined in this new way.

For each continuous path $f : [0, 1] \to X$ define its boundary ∂f to be the formal linear combination of points $f(0) + f(1)$. If g is another map and $g(0) = f(1)$ then, with coefficients in $\mathbf{Z}/2$, we have

$$\partial f + \partial g = f(0) + 2f(1) + g(1) = f(0) + g(1)$$

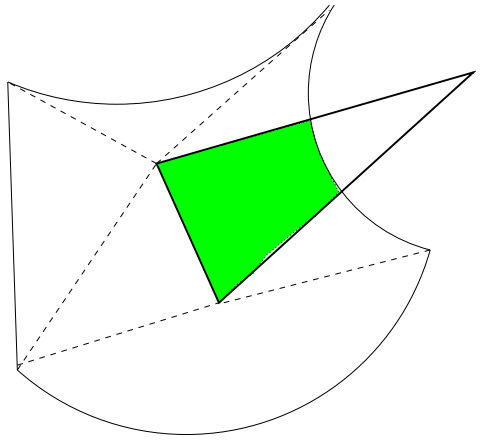
which is the boundary of the path obtained by sticking these two together. Let C_0 be the vector space of finite linear combinations of points with coefficients in $\mathbf{Z}/2$ and C_1 the linear combinations of paths, then $\partial : C_1 \to C_0$ is a linear map. If X is connected then any two points can be joined by a path, so that $x \in C_0$ is in the image of ∂ if and only if it has an even number of terms.

Now look at continuous maps of a triangle $ABC = \Delta$ to X and the space C_2 of all linear combinations of these. The boundary of $F : \Delta \to X$ is the sum of the three paths which are the restrictions of F to the sides of the triangle. Then

$$\partial\partial F = (F(A) + F(B)) + (F(B) + F(C)) + (F(C) + F(A)) = 0$$

so that the image of $\partial : C_2 \to C_1$ is contained in the kernel of $\partial : C_1 \to C_0$. We define $H_1(X)$ to be the quotient space. This is clearly a topological invariant because we only used the notion of continuous functions to define it.

If we take X to be a surface with a subdivision, one can show that because each face is homeomorphic to a disc, any element in the kernel of $\partial : C_1 \to C_0$ can be replaced by adding on something in ∂C_2 by a linear combination of edges of the subdivision:



Now we let $\mathcal{V}$, $\mathcal{E}$ and $\mathcal{F}$ be vector spaces over $\mathbf{Z}/2$ with bases given by the sets of vertices, edges and faces of the subdivision, then define boundary maps in the same way

$$\partial : \mathcal{E} \to \mathcal{V} \text{ and } \partial : \mathcal{F} \to \mathcal{E}.$$

Then

$$H_1(X) \cong \frac{\ker(\partial : \mathcal{E} \to \mathcal{V})}{\operatorname{im}(\partial : \mathcal{F} \to \mathcal{E})}.$$

By the rank-nullity formula we get

$$\dim H_1(X) = \dim \mathcal{E} - \operatorname{rk}(\partial : \mathcal{E} \to \mathcal{V}) - \dim \mathcal{F} + \dim \ker(\partial : \mathcal{F} \to \mathcal{E}).$$

Because X is connected the image of $\partial : \mathcal{E} \to \mathcal{V}$ consists of sums of an even number of vertices so that

$$\dim \mathcal{V} = 1 + \operatorname{rk}(\partial : \mathcal{E} \to \mathcal{V}).$$

Also $\ker(\partial : \mathcal{F} \to \mathcal{E})$ is clearly spanned by the sum of the faces, hence

$$\dim \ker(\partial : \mathcal{F} \to \mathcal{E}) = 1$$

so

$$\dim H_1(X) = 2 - V + E - F.$$

This shows that $V - E + F$ is a topological invariant. □

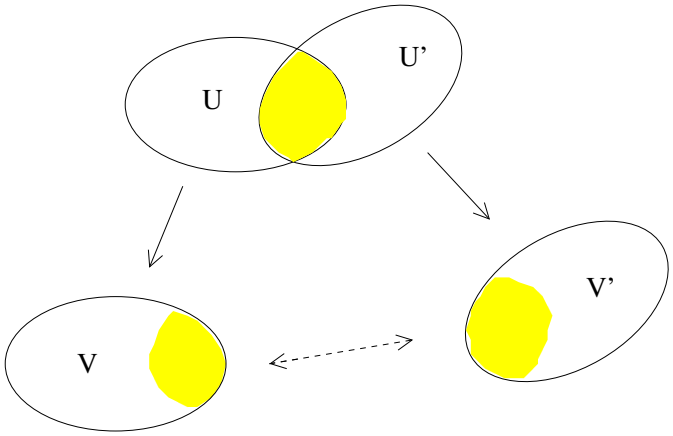
3 Riemann surfaces

3.1 Definitions and examples

From the definition of a surface, each point has a neighbourhood U and a homeomorphism φ_U from U to an open set V in $\mathbf{R}^2$. If two such neighbourhoods U, U' intersect, then

$$\varphi_{U'}\varphi_U^{-1} : \varphi_U(U \cap U') \to \varphi_{U'}(U \cap U')$$

is a homeomorphism from one open set of $\mathbf{R}^2$ to another.



If we identify $\mathbf{R}^2$ with the complex numbers $\mathbf{C}$ then we can define:

Definition 8 *A Riemann surface is a surface with a class of homeomorphisms φ_U such that each map $\varphi_{U'}\varphi_U^{-1}$ is a* holomorphic *(or* analytic*) homeomorphism.*

We call each function φ_U a holomorphic coordinate.

Examples:

1. Let X be the extended complex plane $X = \mathbf{C} \cup \{\infty\}$. Let $U = \mathbf{C}$ with $\varphi_U(z) = z \in \mathbf{C}$. Now take

$$U' = \mathbf{C}\backslash\{0\} \cup \{\infty\}$$

and define $z' = \varphi_{U'}(z) = z^{-1} \in \mathbf{C}$ if $z \neq \infty$ and $\varphi_{U'}(\infty) = 0$. Then

$$\varphi_U(U \cap U') = \mathbf{C}\backslash\{0\}$$

and

$$\varphi_U\varphi_{U'}^{-1}(z) = z^{-1}$$

which is holomorphic.

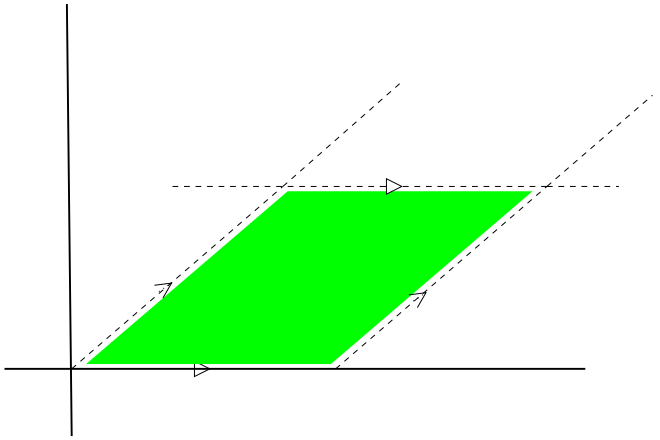
In the right coordinates this is the sphere, with ∞ the North Pole and the coordinate maps given by stereographic projection. For this reason it is sometimes called the *Riemann sphere*.

2. Let $\omega_1, \omega_2 \in \mathbf{C}$ be two complex numbers which are linearly independent over the reals, and define an equivalence relation on $\mathbf{C}$ by $z_1 \sim z_2$ if there are integers m, n such that $z_1 - z_2 = m\omega_1 + n\omega_2$. Let X be the set of equivalence classes (with the quotient topology). A small enough disc V around $z \in \mathbf{C}$ has at most one representative in each equivalence class, so this gives a local homeomorphism to its projection U in X. If U and U' intersect, then the two coordinates are related by a map

$$z \mapsto z + m\omega_1 + n\omega_2$$

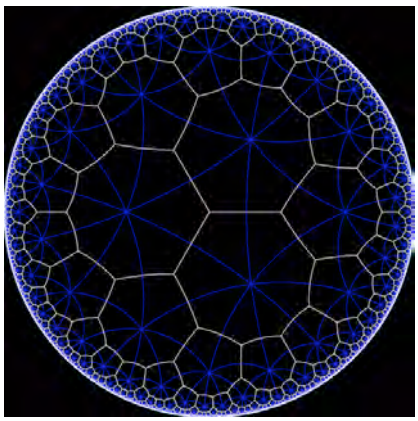
which is holomorphic.

This surface is topologically described by noting that every z is equivalent to one inside the closed parallelogram whose vertices are $0, \omega_1, \omega_2, \omega_1 + \omega_2$, but that points on the boundary are identified:



We thus get a torus this way. Another way of describing the points of the torus is as *orbits* of the action of the group $\mathbf{Z} \times \mathbf{Z}$ on $\mathbf{C}$ by $(m, n) \cdot z = z + m\omega_1 + n\omega_2$.

3. The parallelograms in Example 2 fit together to tile the plane. There are groups of holomorphic maps of the unit disc into itself for which the interior of a polygon plays the same role as the interior of the parallelogram in the plane, and we get a surface X by taking the orbits of the group action. Now we get a tiling of the disc:



In this example the polygon has eight sides and the surface is homeomorphic by the classification theorem to the connected sum of two tori.

4. A *complex algebraic curve* X in $\mathbf{C}^2$ is given by

$$X = \{(z, w) \in \mathbf{C}^2 : f(z, w) = 0\}$$

where f is a polynomial in two variables with complex coefficients. If $(\partial f/\partial z)(z, w) \neq 0$ or $(\partial f/\partial w)(z, w) \neq 0$ for every $(z, w) \in X$, then using the implicit function theorem (see Appendix A) X can be shown to be a Riemann surface with local homeomorphisms given by

$$(z, w) \mapsto w \text{ where } (\partial f/\partial z)(z, w) \neq 0$$

and

$$(z, w) \mapsto z \text{ where } (\partial f/\partial w)(z, w) \neq 0.$$

Definition 9 *A holomorphic map between Riemann surfaces X and Y is a continuous map $f : X \to Y$ such that for each holomorphic coordinate φ_U on U containing x on X and ψ_W defined in a neighbourhood of $f(x)$ on Y, the composition*

$$\psi_W \circ f \circ \varphi_U^{-1}$$

is holomorphic.

In particular if we take $Y = \mathbf{C}$, we can define holomorphc functions on X.

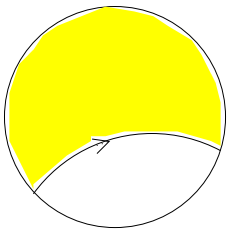
Before proceeding, recall some basic facts about holomorphic functions (see [4]):

- A holomorphic function has a convergent power series expansion in a neighbourhood of each point at which it is defined:
$$f(z) = a_0 + a_1(z-c) + a_2(z-c)^2 + \dots$$
- If f vanishes at c then
$$f(z) = (z-c)^m(c_0 + c_1(z-c) + \dots)$$
where $c_0 \neq 0$. In particular zeros are isolated.
- If f is non-constant it maps open sets to open sets.
- $|f|$ cannot attain a maximum at an interior point of a disc ("maximum modulus principle").
- $f : \mathbf{C} \mapsto \mathbf{C}$ preserves angles between differentiable curves, both in magnitude and sense.

This last property shows:

Proposition 3.1 *A Riemann surface is orientable.*

Proof: Assume X contains a Möbius band, and take a smooth curve down the centre: $\gamma : [0,1] \to X$. In each small coordinate neighbourhood of a point on the curve $\varphi_U\gamma$ is a curve in a disc in $\mathbf{C}$, and rotating the tangent vector γ' by $90°$ or $-90°$ defines an upper and lower half:



Identification on an overlapping neighbourhood is by a map which preserves angles, and in particular the sense – anticlockwise or clockwise – so the two upper halves agree on the overlap, and as we pass around the closed curve the strip is separated into two halves. But removing the central curve of a Möbius strip leaves it connected:

which gives a contradiction. □

From the classification of surfaces we see that a closed, connected Riemann surface is homeomorphic to a connected sum of tori.

3.2 Meromorphic functions

Recall that on a closed (i.e. compact) surface X, any continuous real function achieves its maximum at some point x. Let X be a Riemann surface and f a holomorphic function, then $|f|$ is continuous, so assume it has its maximum at x. Since $f\varphi_U^{-1}$ is a holomorphic function on an open set in $\mathbf{C}$ containing $\varphi_U(x)$, and has its maximum modulus there, the maximum modulus principle says that f must be a constant c in a neighbourhood of x. If X is connected, it follows that $f = c$ everywhere.

Though there are no holomorphic functions, there do exist meromorphic functions:

Definition 10 *A meromorphic function f on a Riemann surface X is a holomorphic map to the Riemann sphere $S = \mathbf{C} \cup \{\infty\}$.*

This means that if we remove $f^{-1}(\infty)$, then f is just a holomorphic function F with values in $\mathbf{C}$. If $f(x) = \infty$, and U is a coordinate neighbourhood of x, then using the coordinate z', $f\varphi_U^{-1}$ is holomorphic. But $\tilde{z} = 1/z$ if $z \neq 0$ which means that $(F \circ \varphi_U^{-1})^{-1}$ is holomorphic. Since it also vanishes,

$$F \circ \varphi_U^{-1} = \frac{a_0}{z^m} + \dots$$

which is usually what we mean by a meromorphic function.

Example: A rational function

$$f(z) = \frac{p(z)}{q(z)}$$

where p and q are polynomials is a meromorphic function on the Riemann sphere S.

The definition above is a geometrical one. Algebraically it is clear that the sum and product of meromorphic functions is meromophic – they form a field.

Here is an example of a meromorphic function on the torus in Example 2.

Define

$$\wp(z) = \frac{1}{z^2} + \sum_{\omega \neq 0} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right)$$

where the sum is over all non-zero $\omega = m\omega_1 + n\omega_2$. Since for $2|z| < |\omega|$

$$\left| \frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right| \leq 10 \frac{|z|}{|\omega|^3}$$

this converges uniformly on compact sets so long as

$$\sum_{\omega \neq 0} \frac{1}{|\omega|^3} < \infty.$$

But $m\omega_1 + n\omega_2$ is never zero if m, n are real so we have an estimate

$$|m\omega_1 + n\omega_2| \geq k\sqrt{m^2 + n^2}$$

so by the integral test we have convergence. Because the sum is essentially over all equivalence classes

$$\wp(z + m\omega_1 + n\omega_2) = \wp(z)$$

so that this is a meromorphic function on the surface X. It is called the Weierstrass P-function.

It is a quite deep result that any closed Riemann surface has meromorphic functions. Let us consider them in more detail. So let

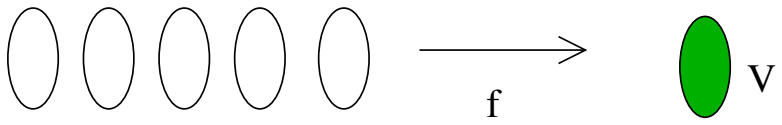
$$f : X \to S$$

be a meromorphic function. If the inverse image of $a \in S$ is infinite, then it has a limit point x by compactness of X. In a holomorphic coordinate around x with $z(x) = 0$, f is defined by a holomorphic function $F = f\varphi_U^{-1}$ with a sequence of points $z_n \to 0$ for which $F(z_n) - a = 0$. But the zeros of a holomorphic function are isolated, so we deduce that $f^{-1}(a)$ is a finite set. By a similar argument the points at which the derivative F' vanishes are finite in number (check that this condition is independent of the holomorphic coordinate). The points of X at which $F' = 0$ are called *ramification points*.

Now recall another result from complex analysis: if a holomorphic function f has a zero of order n at $z = 0$, then for $\epsilon > 0$ sufficiently small, there is $\delta > 0$ such that for all a with $0 < |a| < \delta$, the equation $f(z) = a$ has exactly n roots in the disc $|z| < \epsilon$.

This result has two consequences. The first is that if $F'(x) \neq 0$, then f maps a neighbourhood U_x of $x \in X$ homeomorphically to a neighbourhood V_x of $f(x) \in S$.

Define V to be the intersection of the V_x as x runs over the finite set of points such that $f(x) = a$, then $f^{-1}V$ consists of a finite number d of open sets, each mapped homeomorphically onto V by f:



The second is that if $F' = 0$, we have

$$F(z) = z^n(a_0 + a_1 z + \ldots)$$

for some n and F has a zero of order n at 0, where $z(x) = 0$. In that case there is a neighbourhood U of x and V of a such that $f(U) = V$, and the inverse image of $y \neq x \in V$ consists of n distinct points, but $f^{-1}(a) = x$. In fact, since $a_0 \neq 0$, we can expand

$$(a_0 + a_1 z + \ldots)^{1/n} = a_0^{1/n}(1 + b_1 z + \ldots)$$

in a power series and use a new coordinate

$$w = a_0^{1/n} z(1 + b_1 z + \ldots)$$

so that the map f is locally

$$w \mapsto w^n.$$

There are then two types of neighbourhoods of points: at an ordinary point the map looks like $w \mapsto w$ and at a ramification point like $w \mapsto w^n$.

Removing the finite number of images under f of ramification points we get a sphere minus a finite number of points. This is connected. The number of points in the inverse image of a point in this punctured sphere is integer-valued and continuous, hence constant. It is called the *degree* d of the meromorphic function f.

With this we can determine the Euler characteristic of the Riemann surface S from the meromorphic function:

Theorem 3.2 *(Riemann-Hurwitz) Let $f : X \to S$ be a meromorphic function of degree d on a closed connected Riemann surface X, and suppose it has ramification points $x_1, \ldots, x_n$ where the local form of $f(x) - f(x_k)$ is a holomorphic function with a zero of multiplicity m_k. Then*

$$\chi(X) = 2d - \sum_{k=1}^{n}(m_k - 1)$$

Proof: The idea is to take a triangulation of the sphere S such that the image of the ramification points are vertices. This is straighforward. Now take a finite subcovering of S by open sets of the form V above where the map f is either a homeomorphism or of the form $z \mapsto z^m$. Subdivide the triangulation into smaller triangles such that each one is contained in one of the sets V. Then the inverse images of the vertices and edges of S form the vertices and edges of a triangulation of X.

If the triangulation of S has V vertices, E edges and F faces, then clearly the triangulation of X has dE edges and dF faces. It has fewer vertices, though — in a neighbourhood where f is of the form $w \mapsto w^m$ the origin is a single vertex instead of m of them. For each ramification point of order m_k we therefore have one vertex instead of m_k. The count of vertices is therefore

$$dV - \sum_{k=1}^{n}(m_k - 1).$$

Thus

$$\chi(X) = d(V - E + F) - \sum_{k=1}^{n}(m_k - 1) = 2d - \sum_{k=1}^{n}(m_k - 1)$$

using $\chi(S) = 2$. □

Clearly the argument works just the same for a holomorphic map $f : X \to Y$ and then

$$\chi(X) = d\chi(Y) - \sum_{k=1}^{n}(m_k - 1).$$

As an example, consider the Weierstrass P-function $\wp : T \to S$:

$$\wp(z) = \frac{1}{z^2} + \sum_{\omega \neq 0}\left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2}\right)$$

This has degree 2 since $\wp(z) = \infty$ only at $z = 0$ and there it has multiplicity 2. Each $m_k \leq d = 2$, so the only possible value at the ramification points here is $m_k = 2$. The Riemann-Hurwitz formula gives:

$$0 = 4 - n$$

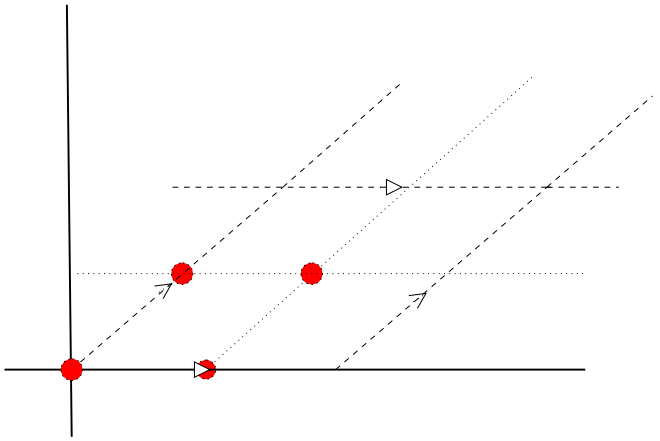
so there must be exactly 4 ramification points. In fact we can see them directly, because $\wp(z)$ is an even function, so the derivative vanishes if $-z = z$. Of course at $z = 0$, $\wp(z) = \infty$ so we should use the other coordinate on S: $1/\wp$ has a zero of

multiplicity 2 at $z = 0$. To find the other points recall that $\wp$ is doubly periodic so $\wp'$ vanishes where

$$z = -z + m\omega_1 + n\omega_2$$

for some integers m, n, and these are the four points

$$0, \omega_1/2, \omega_2/2, (\omega_1 + \omega_2)/2 :$$



3.3 Multi-valued functions

The Riemann-Hurwitz formula is useful for determining the Euler characteristic of a Riemann surface defined in terms of a multi-valued function, like

$$g(z) = z^{1/n}.$$

We look for a closed surface on which z and $g(z)$ are meromorphic functions. The example above is easy: if $w = z^{1/n}$ then $w^n = z$, and using the coordinate $z' = 1/z$ on a neighbourhood of ∞ on the Riemann sphere S, if $w' = 1/w$ then $w'^n = z'$.

Thus w and w' are standard coordinates on S, and $g(z)$ is the identity map $S \to S$. The function $z = w^n$ is then a meromorphic function f of degree n on S. It has two ramification points of order n at $w = 0$ and $w = \infty$, so the Riemann-Hurwitz formula is verified:

$$2 = \chi(S) = 2n - 2(n - 1).$$

The most general case is that of a complex algebraic curve $f(z, w) = 0$. This is a polynomial in w with coefficients functions of z, so its "solution" is a multivalued

function of z. We shall deal with a simpler but still important case $w^2 = p(z)$ where p is a polynomial of degree n in z with n distinct roots. We are looking then for a Riemann surface on which

$$\sqrt{p(z)}$$

can be interpreted as a meromorphic function.

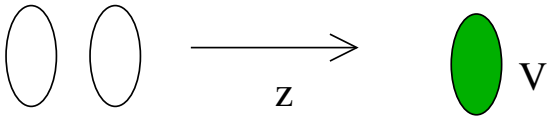
We proceed to define coordinate neighbourhoods on each of which w is a holomorphic function. First, if $p(a) \neq 0$ then

$$p(z) = p(a)(1 + a_1(z-a) + \ldots + a_n(z-a)^n).$$

For each choice of $\sqrt{p(a)}$ we have w expressed as a power series in z in a neigbourhood of a:

$$w = \sqrt{p(a)}(1 + a_1(z-a) + \ldots + a_n(z-a)^n)^{1/2} = 1 + a_1 z/2 + \ldots.$$

So we can take z as a coordinate on each of two open sets, and w is holomorphic here.



If $p(a) = 0$, then since p has distinct roots,

$$p(z) = (z-a)(b_0 + b_1(z-a) + \ldots)$$

where $b_0 \neq 0$. Put $u^2 = (z-a)$ and $p(z) = u^2(b_0 + b_1 u^2 + \ldots)$ and so, choosing $\sqrt{b_0}$, w has a power series expansion in u:

$$w = u\sqrt{b_0}(1 + b_1 u^2/b_0 + \ldots).$$

(The other choose of $\sqrt{b_0}$ is equivalent to taking the local coordinate $-u$.) This gives an open disc, with u as coordinate, on which w is holomorphic.

For $z = \infty$ we note that

$$\frac{w^2}{z^n} = a_n + \frac{a_{n-1}}{z} + \ldots$$

so if $n = 2m$,

$$\left(\frac{w}{z^m}\right)^2 = a_n + \frac{a_{n-1}}{z} + \ldots$$

and since $a_n \neq 0$, putting $w' = 1/w$ and $z' = 1/z$ we get

$$w' = a_n^{-1/2} z'^m (1 + a_{n-1} z'/a_n + \ldots)^{-1/2}$$

which is a holomorphic function. If $n = 2m + 1$, we need a coordinate $u^2 = z'$ as above.

The coordinate neighbourhoods defined above give the set of solutions to $w^2 = p(z)$ together with points at infinity the structure of a compact Riemann surface X such that

- z is a meromorphic function of degree 2 on X
- w is a meromorphic function of degree n on X
- the ramification points of z are at the points $(z = a, w = 0)$ where a is a root of $p(z)$, and if n is odd, at $(z = \infty, w = \infty)$

The Riemann-Hurwitz formula now gives

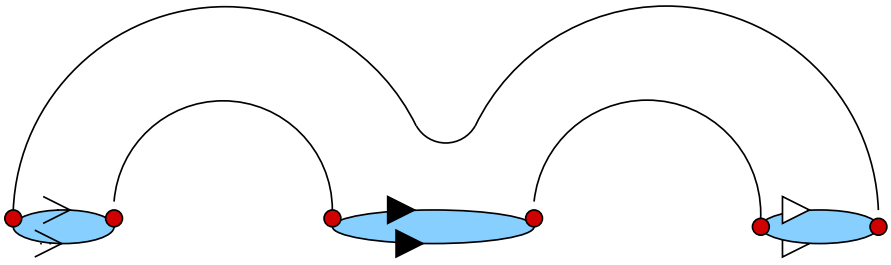
$$\chi(X) = 2 \times 2 - n = 4 - n$$

if n is even and

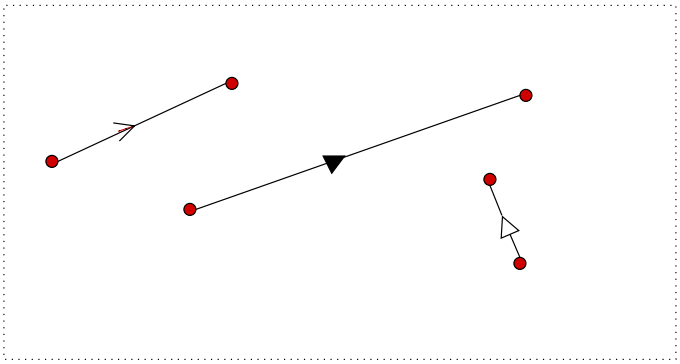
$$\chi(X) = 4 - (n + 1) = 3 - n$$

if n is odd.

This type of Riemann surface is called *hyperelliptic*. Since the two values of $w = \sqrt{p(z)}$ only differ by a sign, we can think of $(w, z) \mapsto (-w, z)$ as being a holomorphic homeomorphism from X to X, and then z is a coordinate on the space of orbits. Topologically we can cut the surface in two – an "upper" and "lower" half – and identify on the points on the boundary to get a sphere:



It is common also to view this downstairs on the Riemann sphere and insert cuts between pairs of zeros of the polynomial $p(z)$:



As an example, consider again the P-function $\wp(z)$, thought of as a degree 2 map $\wp : T \to S$. It has 4 ramification points, whose images are ∞ and the three finite points e_1, e_2, e_3 where

$$e_1 = \wp(\omega_1/2), \quad e_2 = \wp(\omega_2/2), \quad e_3 = \wp((\omega_1 + \omega_2)/2).$$

So its derivative $\wp'(z)$ vanishes only at three points, each with multiplicity 1. At each of these points $\wp$ has the local form

$$\wp(z) = e_1 + (z - \omega_1/2)^2(a_0 + \ldots)$$

and so

$$\frac{1}{\wp'(z)^2}(\wp(z) - e_1)(\wp(z) - e_2)(\wp(z) - e_3)$$

is a well-defined holomorphic function on T away from $z = 0$. But $\wp(z) \sim z^{-2}$ near $z = 0$, and so $\wp'(z) \sim -2z^{-3}$ so this function is finite at $z = 0$ with value $1/4$. By the maximum argument, since T is compact,the function is a constant, namely $1/4$.

Thus the meromorphic function $\wp'(z)$ on T can also be considered as

$$2\sqrt{(u - e_1)(u - e_2)(u - e_3)}$$

setting $u = \wp(z)$.

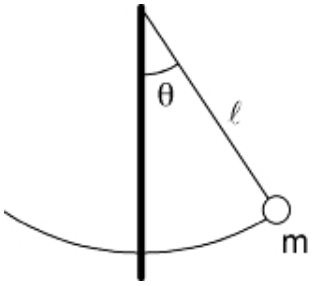
Note that, substituting $u = \wp(z)$, we have

$$\frac{du}{2\sqrt{(u - e_1)(u - e_2)(u - e_3)}} = dz.$$

By changing variables with a Möbius transformation of the form $u \mapsto (au+b)/(cu+d)$ any integrand

$$\frac{du}{\sqrt{p(u)}}$$

can be brought into this form if p is of degree 3 or 4. This can be very useful, for example in the equation for a pendulum:



$$\theta'' = -(g/\ell)\sin\theta$$

which integrates once to

$$\theta'^2 = 2(g/\ell)\cos\theta + c.$$

Substituting $v = e^{i\theta}$ we get

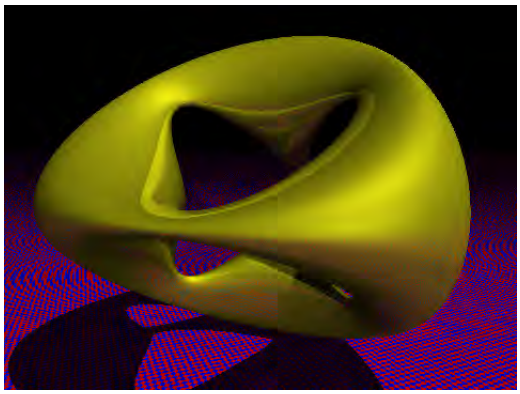
$$v' = i\sqrt{2(g/\ell)(v^3 + v) + cv^2}.$$

So time becomes (the real part of) the parameter z on $\mathbf{C}$. In the torus this is a circle, so (no surprise here!) the solutions to the pendulum equation are periodic.

4 Surfaces in $\mathbf{R}^3$

4.1 Definitions

At this point we return to surfaces embedded in Euclidean space, and consider the differential geometry of these:



We shall not forget the idea of an abstract surface though, and as we meet objects which we call *intrinsic* we shall show how to define them on a surface which is not sitting in $\mathbf{R}^3$. These remarks are printed in a smaller typeface.

Definition 11 *A smooth surface in* $\mathbf{R}^3$ *is a subset* $X \subset \mathbf{R}^3$ *such that each point has a neighbourhood* $U \subset X$ *and a map* $\mathbf{r} : V \to \mathbf{R}^3$ *from an open set* $V \subseteq \mathbf{R}^2$ *such that*

- $\mathbf{r} : V \to U$ *is a homeomorphism*
- $\mathbf{r}(u, v) = (x(u, v), y(u, v), z(u, v))$ *has derivatives of all orders*
- *at each point* $\mathbf{r}_u = \partial\mathbf{r}/\partial u$ *and* $\mathbf{r}_v = \partial\mathbf{r}/\partial v$ *are linearly independent.*

Already in the definition we see that X is a topological surface as in Definition 2, since $\mathbf{r}$ defines a homeomorphism $\varphi_U : U \to V$. The last two conditions make sense if we use the *implicit function theorem* (see Appendix 1). This tells us that a local invertible change of variables in $\mathbf{R}^3$ "straightens out" the surface: it can be locally defined by $x_3 = 0$ where (x_1, x_2, x_3) are (nonlinear) local coordinates on $\mathbf{R}^3$. For any two open sets U, U', we get a smooth invertible map from an open set of $\mathbf{R}^3$ to another which takes $x_3 = 0$ to $x_3' = 0$. This means that each map $\varphi_{U'}\varphi_U^{-1}$ is a smooth invertible homeomorphism. This motivates the definition of an abstract smooth surface:

Definition 12 *A smooth surface is a surface with a class of homeomorphisms φ_U such that each map $\varphi_{U'}\varphi_U^{-1}$ is a smoothly invertible homeomorphism.*

Clearly, since a holomorphic function has partial derivatives of all orders in x, y, a Riemann surface is an example of an abstract smooth surface. Similarly, we have

Definition 13 *A smooth map between smooth surfaces X and Y is a continuous map $f : X \to Y$ such that for each smooth coordinate system φ_U on U containing x on X and ψ_W defined in a neighbourhood of $f(x)$ on Y, the composition*

$$\psi_W \circ f \circ \varphi_U^{-1}$$

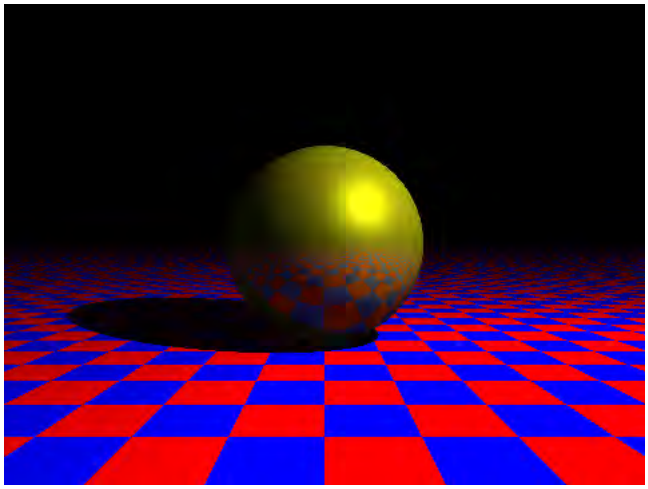
is smooth.

We now return to surfaces in $\mathbf{R}^3$:

Examples:

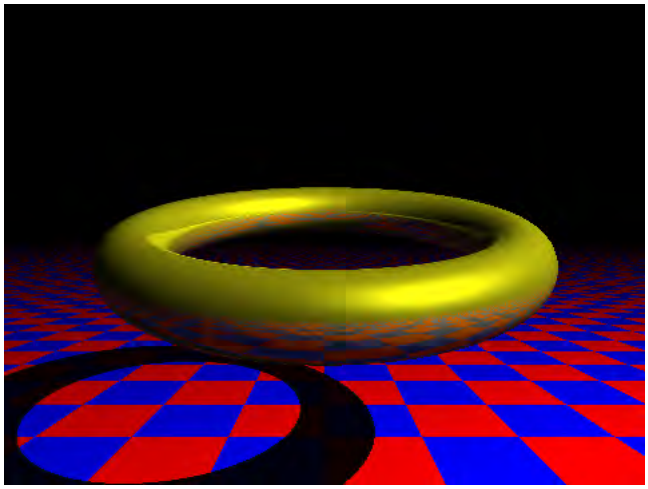
1) A sphere:

$$\mathbf{r}(u, v) = a \sin u \sin v\, \mathbf{i} + a \cos u \sin v\, \mathbf{j} + a \cos v\, \mathbf{k}$$



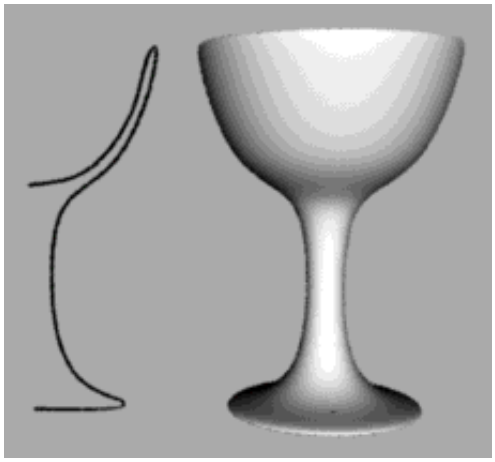
2) A torus:

$$\mathbf{r}(u, v) = (a + b \cos u)(\cos v\, \mathbf{i} + \sin v\, \mathbf{j}) + b \sin u\, \mathbf{k}$$



3) A surface of revolution:

$$\mathbf{r}(u,v) = f(u)(\cos v\,\mathbf{i} + \sin v\,\mathbf{j}) + u\mathbf{k}$$



These are the only compact surfaces it is easy to write down, but the following non-compact ones are good for local discussions:

Examples:

1) A plane:

$$\mathbf{r}(u,v) = \mathbf{a} + u\mathbf{b} + v\mathbf{c}$$

for constant vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ where $\mathbf{b}, \mathbf{c}$ are linearly independent.

2) A cylinder:

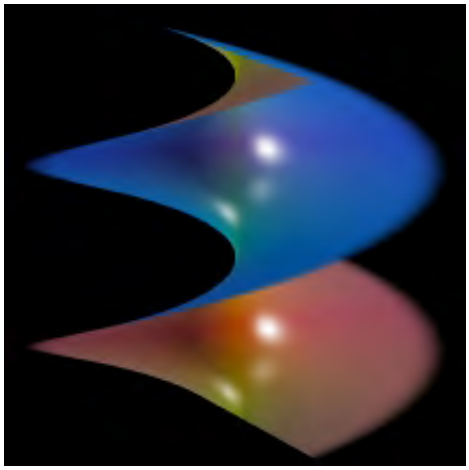
$$\mathbf{r}(u, v) = a(\cos v\,\mathbf{i} + \sin v\,\mathbf{j}) + u\mathbf{k}$$

3) A cone:

$$\mathbf{r}(u, v) = au\cos v\,\mathbf{i} + au\sin v\,\mathbf{j} + u\mathbf{k}$$

4) A helicoid:

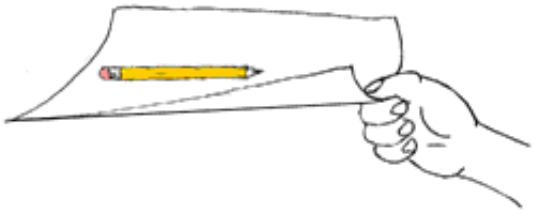
$$\mathbf{r}(u, v) = au\cos v\,\mathbf{i} + au\sin v\,\mathbf{j} + v\mathbf{k}$$



5) A developable surface: take a curve $\boldsymbol{\gamma}(u)$ parametrized by arc length and set

$$\mathbf{r}(u,v) = \boldsymbol{\gamma}(u) + v\boldsymbol{\gamma}'(u)$$

This is the surface formed by bending a piece of paper:



A change of parametrization of a surface is the composition

$$\mathbf{r} \circ f : V' \to \mathbf{R}^3$$

where $f : V' \to V$ is a *diffeomorphism* – an invertible map such that f and f^{-1} have derivatives of all orders. Note that if

$$f(x,y) = (u(x,y), v(x,y))$$

then by the chain rule

$$\begin{aligned} (\mathbf{r} \circ f)_x &= \mathbf{r}_u u_x + \mathbf{r}_v v_x \\ (\mathbf{r} \circ f)_y &= \mathbf{r}_u u_y + \mathbf{r}_v v_y \end{aligned}$$

so

$$\begin{pmatrix} (\mathbf{r}\circ f)_x \\ (\mathbf{r}\circ f)_y \end{pmatrix} = \begin{pmatrix} u_x & v_x \\ u_y & v_y \end{pmatrix} \begin{pmatrix} \mathbf{r}_u \\ \mathbf{r}_v \end{pmatrix}.$$

Since f has a differentiable inverse, the Jacobian matrix is invertible, so $(\mathbf{r}\circ f)_x$ and $(\mathbf{r}\circ f)_y$ are linearly independent if $\mathbf{r}_u, \mathbf{r}_v$ are.

Example: The (x, y) plane

$$\mathbf{r}(x, y) = x\mathbf{i} + y\mathbf{j}$$

has a different parametrization in polar coordinates

$$\mathbf{r}\circ f(r, \theta) = r\cos\theta\,\mathbf{i} + r\sin\theta\,\mathbf{j}.$$

We have to consider changes of parametrizations when we pass from one open set V to a neighbouring one V'.

Definition 14 *The tangent plane (or* tangent space*) of a surface at the point a is the vector space spanned by $\mathbf{r}_u(a), \mathbf{r}_v(a)$.*

Note that this space is independent of parametrization. One should think of the origin of the vector space as the point a.

Definition 15 *The vectors*

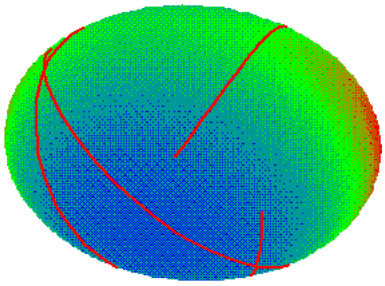
$$\pm\frac{\mathbf{r}_u\wedge\mathbf{r}_v}{|\mathbf{r}_u\wedge\mathbf{r}_v|}$$

are the two unit normals ("inward and outward") to the surface at (u, v).

4.2 The first fundamental form

Definition 16 *A smooth curve lying in the surface is a map $t \mapsto (u(t), v(t))$ with derivatives of all orders such that $\boldsymbol{\gamma}(t) = \mathbf{r}(u(t), v(t))$ is a parametrized curve in $\mathbf{R}^3$.*

A parametrized curve means that $u(t), v(t)$ have derivatives of all orders and $\boldsymbol{\gamma}' = \mathbf{r}_u u' + \mathbf{r}_v v' \neq 0$. The definition of a surface implies that $\mathbf{r}_u, \mathbf{r}_v$ are linearly independent, so this condition is equivalent to $(u', v') \neq 0$.



The arc length of such a curve from $t = a$ to $t = b$ is:

$$\begin{aligned}\int_a^b |\boldsymbol{\gamma}'(t)|dt &= \int_a^b \sqrt{\boldsymbol{\gamma}' \cdot \boldsymbol{\gamma}'}dt \\ &= \int_a^b \sqrt{(\mathbf{r}_u u' + \mathbf{r}_v v') \cdot (\mathbf{r}_u u' + \mathbf{r}_v v')}dt \\ &= \int_a^b \sqrt{Eu'^2 + 2Fu'v' + Gv'^2}dt\end{aligned}$$

where

$$E = \mathbf{r}_u \cdot \mathbf{r}_u, \quad F = \mathbf{r}_u \cdot \mathbf{r}_v, \quad F = \mathbf{r}_v \cdot \mathbf{r}_v.$$

Definition 17 *The first fundamental form of a surface in* $\mathbf{R}^3$ *is the expression*

$$Edu^2 + 2Fdudv + Gdv^2$$

where $E = \mathbf{r}_u \cdot \mathbf{r}_u, F = \mathbf{r}_u \cdot \mathbf{r}_v, G = \mathbf{r}_v \cdot \mathbf{r}_v$.

The first fundamental form is just the quadratic form

$$Q(\mathbf{v}, \mathbf{v}) = \mathbf{v} \cdot \mathbf{v}$$

on the tangent space written in terms of the basis $\mathbf{r}_u, \mathbf{r}_v$. It is represented in this basis by the symmetric matrix

$$\begin{pmatrix} E & F \\ F & G \end{pmatrix}.$$

So why do we write it as $Edu^2 + 2Fdudv + Gdv^2$? At this stage it is not worth worrying about what exactly du^2 is, instead let's see how the terminology helps to manipulate the formulas.

For example, to find the length of a curve $u(t), v(t)$ on the surface, we calculate

$$\int \sqrt{E\left(\frac{du}{dt}\right)^2 + 2F\frac{du}{dt}\frac{dv}{dt} + G\left(\frac{dv}{dt}\right)^2}\,dt$$

– divide the first fundamental form by dt^2 and multiply its square root by dt.

Furthermore if we change the parametrization of the surface via $u(x,y), v(x,y)$ and try to find the length of the curve $(x(t), y(t))$ then from first principles we would calculate

$$u' = u_x x' + u_y y' \qquad v' = v_x x' + v_y y'$$

by the chain rule and then

$$\begin{aligned} Eu'^2 + 2Fu'v' + Gv'^2 &= E(u_x x' + u_y y')^2 + 2F(u_x x' + u_y y')(v_x x' + \ldots \\ &= (Eu_x^2 + 2Fu_x v_x + Gv_x^2)x'^2 + \ldots \end{aligned}$$

which is heavy going. Instead, using du, dv etc. we just write

$$\begin{aligned} du &= u_x dx + u_y dy \\ dv &= v_x dx + v_y dy \end{aligned}$$

and substitute in $Edu^2 + 2Fdudv + Gdv^2$ to get $E'dx^2 + 2F'dxdy + G'dy^2$. Using matrices, we can write this transformation as

$$\begin{pmatrix} u_x & u_y \\ v_x & v_y \end{pmatrix} \begin{pmatrix} E & F \\ F & G \end{pmatrix} \begin{pmatrix} u_x & v_x \\ u_y & v_y \end{pmatrix} = \begin{pmatrix} E' & F' \\ F' & G' \end{pmatrix}$$

Example: For the plane

$$\mathbf{r}(x,y) = x\mathbf{i} + y\mathbf{j}$$

we have $\mathbf{r}_x = \mathbf{i}, \mathbf{r}_y = \mathbf{j}$ and so the first fundamental form is

$$dx^2 + dy^2.$$

Now change to polar coordinates $x = r\cos\theta, y = r\sin\theta$. We have

$$\begin{aligned} dx &= dr\cos\theta - r\sin\theta d\theta \\ dy &= dr\sin\theta + r\cos\theta d\theta \end{aligned}$$

so that

$$dx^2 + dy^2 = (dr\cos\theta - r\sin\theta d\theta)^2 + (dr\sin\theta + r\cos\theta)^2 = dr^2 + r^2 d\theta^2$$

Here are some examples of first fundamental forms:

Examples:

1. The cylinder

$$\mathbf{r}(u,v) = a(\cos v\,\mathbf{i} + \sin v\,\mathbf{j}) + u\mathbf{k}.$$

We get

$$\mathbf{r}_u = \mathbf{k}, \quad \mathbf{r}_v = a(-\sin v\,\mathbf{i} + \cos v\,\mathbf{j})$$

so

$$E = \mathbf{r}_u \cdot \mathbf{r}_u = 1, \quad F = \mathbf{r}_u \cdot \mathbf{r}_v = 0, \quad G = \mathbf{r}_v \cdot \mathbf{r}_v = a^2$$

giving

$$\boxed{du^2 + a^2 dv^2}$$

2. The cone

$$\mathbf{r}(u,v) = a(u\cos v\,\mathbf{i} + u\sin v\,\mathbf{j}) + u\mathbf{k}.$$

Here

$$\mathbf{r}_u = a(\cos v\,\mathbf{i} + \sin v\,\mathbf{j}) + \mathbf{k}, \quad \mathbf{r}_v = a(-u\sin v\,\mathbf{i} + u\cos v)$$

so

$$E = \mathbf{r}_u \cdot \mathbf{r}_u = 1 + a^2, \quad F = \mathbf{r}_u \cdot \mathbf{r}_v = 0, \quad G = \mathbf{r}_v \cdot \mathbf{r}_v = a^2u^2$$

giving

$$\boxed{(1+a^2)du^2 + a^2u^2dv^2}$$

3. The sphere

$$\mathbf{r}(u,v) = a\sin u\sin v\,\mathbf{i} + a\cos u\sin v\,\mathbf{j} + a\cos v\,\mathbf{k}$$

gives

$$\mathbf{r}_u = a\cos u\sin v\,\mathbf{i} - a\sin u\sin v\,\mathbf{j}, \quad \mathbf{r}_v = a\sin u\cos v\,\mathbf{i} + a\cos u\cos v\,\mathbf{j} - a\sin v\,\mathbf{k}$$

so that

$$E = \mathbf{r}_u \cdot \mathbf{r}_u = a^2\sin^2 v, \quad F = \mathbf{r}_u \cdot \mathbf{r}_v = 0, \quad G = \mathbf{r}_v \cdot \mathbf{r}_v = a^2$$

and so we get the first fundamental form

$$\boxed{a^2dv^2 + a^2\sin^2 v du^2}$$

4. A surface of revolution

$$\mathbf{r}(u,v) = f(u)(\cos v\,\mathbf{i} + \sin v\,\mathbf{j}) + u\mathbf{k}$$

has

$$\mathbf{r}_u = f'(u)(\cos v\,\mathbf{i} + \sin v\,\mathbf{j}) + \mathbf{k}, \quad \mathbf{r}_v = f(u)(-\sin v\,\mathbf{i} + \cos v\,\mathbf{j})$$

so that

$$E = \mathbf{r}_u \cdot \mathbf{r}_u = 1 + f'(u)^2, \quad F = \mathbf{r}_u \cdot \mathbf{r}_v = 0, \quad G = \mathbf{r}_v \cdot \mathbf{r}_v = f(u)^2$$

gives

$$\boxed{(1 + f(u)'^2)du^2 + f(u)^2 dv^2}$$

5. A developable surface

$$\mathbf{r}(u,v) = \boldsymbol{\gamma}(u) + v\mathbf{t}(u).$$

here the curve is parametrized by arc length $u = s$ so that

$$\mathbf{r}_u = \mathbf{t}(u) + v\mathbf{t}'(u) = \mathbf{t} + v\kappa\mathbf{n}, \quad \mathbf{r}_v = \mathbf{t}$$

where $\mathbf{n}$ is the normal to the curve and κ its curvature. This gives

$$\boxed{(1 + v^2\kappa^2)du^2 + 2dudv + dv^2}$$

The analogue of the first fundamental form on an abstract smooth surface X is called a *Riemannian metric*. On each open set U with coordinates (u,v) we ask for smooth functions E, F, G with $E > 0, G > 0, EG - F^2 > 0$ and on an overlapping neighbourhood with coordinates (x,y) smooth functions E', F', G' with the same properties and the transformation law:

$$\begin{pmatrix} u_x & u_y \\ v_x & v_y \end{pmatrix} \begin{pmatrix} E & F \\ F & G \end{pmatrix} \begin{pmatrix} u_x & v_x \\ u_y & v_y \end{pmatrix} = \begin{pmatrix} E' & F' \\ F' & G' \end{pmatrix}$$

A smooth curve on X is defined to be a map $\gamma : [a,b] \to X$ such that $\varphi_U \gamma$ is smooth for each coordinate neighbourhood U on the image. The length of such a curve is well-defined by a Riemannian metric.

Examples:

1. The torus as a Riemann surface has the metric

$$dzd\bar{z} = dx^2 + dy^2$$

as the local holomorphic coordinates are z and $z + m\omega_1 + n\omega_2$ so that the Jacobian matric is the identity. We could also multiply this by any positive smooth doubly-periodic function.

2. The hyperelliptic Riemann surface $w^2 = p(z)$ where $p(z)$ is of degree $2m$ has Riemannian metrics given by

$$\frac{1}{|w|^2}(a_0 + a_1|z|^2 + \ldots + a_{m-2}|z|^{2(m-2)})dzd\bar{z}$$

where the a_i are positive constants.

3. The upper half-space $\{x + iy \in \mathbf{C} : y > 0\}$ has the metric

$$\frac{dx^2 + dy^2}{y^2}.$$

None of these have anything to do with the first fundamental form of the surface embedded in $\mathbf{R}^3$.

We introduced the first fundamental form to measure lengths of curves on a surface but it does more besides. Firstly if two curves $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2$ on the surface intersect, the angle θ between them is given by

$$\cos\theta = \frac{\boldsymbol{\gamma}_1' \cdot \boldsymbol{\gamma}_2'}{|\boldsymbol{\gamma}_1'||\boldsymbol{\gamma}_2'|} \tag{1}$$

But $\boldsymbol{\gamma}_i' = \mathbf{r}_u u_i' + \mathbf{r}_v v_i'$ so

$$\begin{aligned} \boldsymbol{\gamma}_i' \cdot \boldsymbol{\gamma}_j' &= (\mathbf{r}_u u_i' + \mathbf{r}_v v_i') \cdot (\mathbf{r}_u u_j' + \mathbf{r}_v v_j') \\ &= Eu_i'u_j' + F(u_i'v_j' + u_j'v_i') + Gv_i'v_j' \end{aligned}$$

and each term in (1) can be expressed in terms of the curves and the coefficients of the first fundamental form.

We can also define *area* using the first fundamental form:

Definition 18 *The area of the domain $\mathbf{r}(U) \subset \mathbf{R}^3$ in a surface is defined by*

$$\int_U |\mathbf{r}_u \wedge \mathbf{r}_v| dudv = \int_U \sqrt{EG - F^2} dudv.$$

The second form of the formula comes from the identity

$$|\mathbf{r}_u \wedge \mathbf{r}_v|^2 = (\mathbf{r}_u \cdot \mathbf{r}_u)(\mathbf{r}_v \cdot \mathbf{r}_v) - (\mathbf{r}_u \cdot \mathbf{r}_v)^2 = EG - F^2.$$

Note that the definition of area is independent of parametrization for if

$$\mathbf{r}_x = \mathbf{r}_u u_x + \mathbf{r}_v v_x, \quad \mathbf{r}_y = \mathbf{r}_u u_y + \mathbf{r}_v v_y$$

then

$$\mathbf{r}_x \wedge \mathbf{r}_y = (u_x v_y - v_x u_y)\mathbf{r}_u \wedge \mathbf{r}_v$$

so that

$$\int_U |\mathbf{r}_x \wedge \mathbf{r}_y| dxdy = \int_U |\mathbf{r}_u \wedge \mathbf{r}_v||u_x v_y - v_x u_y| dxdy = \int_U |\mathbf{r}_u \wedge \mathbf{r}_v| dudv$$

using the formula for change of variables in multiple integration.

Example: Consider a surface of revolution

$$(1 + f'(u)^2)du^2 + f(u)^2 dv^2$$

and the area between $u = a, u = b$. We have

$$EG - F^2 = f(u)^2(1 + f'(u)^2)$$

so the area is

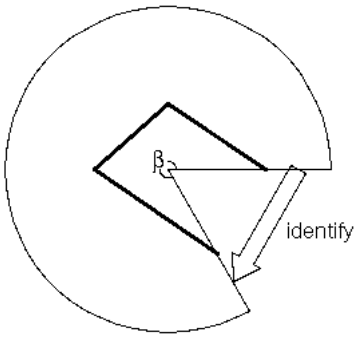
$$\int_a^b f(u)\sqrt{1 + f'(u)^2} dudv = 2\pi \int_a^b f(u)\sqrt{1 + f'(u)^2} du.$$

If a closed surface X is triangulated so that each face lies in a coordinate neighbourhood, then we can define the area of X as the sum of the areas of the faces by the formula above. It is independent of the choice of triangulation.

4.3 Isometric surfaces

Definition 19 *Two surfaces X, X' are isometric if there is a smooth homeomorphism $f : X \to X'$ which maps curves in X to curves in X' of the same length.*

A practical example of this is to take a piece of paper and bend it: the lengths of curves in the paper do not change. The cone and a subset of the plane are isometric this way:



Analytically this is how to tell if two surfaces are isometric:

Theorem 4.1 *The coordinate patches of surfaces U and U' are isometric if and only if there exist parametrizations $\mathbf{r} : V \to \mathbf{R}^3$ and $\mathbf{r}' : V \to \mathbf{R}^3$ with the same first fundamental form.*

Proof: Suppose such a parametrization exists, then the identity map is an isometry since the first fundamental form determines the length of curves.

Conversely, suppose X, X' are isometric using the function $f : V \to V'$. Then

$$\mathbf{r}' \circ f : V \to \mathbf{R}^3, \qquad \mathbf{r} : V \to \mathbf{R}^3$$

are parametrizations using the same open set V, so the first fundamental forms are

$$\tilde{E}du^2 + 2\tilde{F}dudv + \tilde{G}dv^2, \quad Edu^2 + 2Fdudv + Gdv^2$$

and since f is an isometry

$$\int_I \sqrt{\tilde{E}u'^2 + 2\tilde{F}u'v' + \tilde{G}v'^2}dt = \int_I \sqrt{Eu'^2 + 2Fu'v' + Gv'^2}dt$$

for all curves $t \mapsto (u(t), v(t))$ and *all intervals*. Since

$$\frac{d}{dt}\int_a^{a+t} h(u)du = h(t)$$

this means that

$$\sqrt{\tilde{E}u'^2 + 2\tilde{F}u'v' + \tilde{G}v'^2} = \sqrt{Eu'^2 + 2Fu'v' + Gv'^2}$$

for all $u(t), v(t)$. So, choosing u, v appropriately:

$$\begin{aligned} u = t, v = a &\Rightarrow \tilde{E} = E \\ u = a, v = t &\Rightarrow \tilde{G} = G \\ u = t, v = t &\Rightarrow \tilde{F} = F \end{aligned}$$

and we have the same first fundamental form as required. □

Example:

The cone has first fundamental form

$$(1 + a^2)du^2 + a^2u^2dv^2.$$

Put

$$r = \sqrt{1 + a^2}u$$

then we get

$$dr^2 + \left(\frac{a^2}{1 + a^2}\right) r^2dv^2$$

and now put

$$\theta = \sqrt{\frac{a^2}{1 + a^2}}v$$

to get the plane in polar coordinates

$$dr^2 + r^2d\theta^2.$$

Note that as $0 \leq v \leq 2\pi$, $0 \leq \theta \leq \beta$ where

$$\beta = \sqrt{\frac{a^2}{1 + a^2}}2\pi < 2\pi$$

as in the picture.

Example: Consider the unit disc $D = \{x + iy \in \mathbf{C} | x^2 + y^2 < 1\}$ with first fundamental form

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

and the upper half plane $H = \{u + iv \in \mathbf{C} | v > 0\}$ with the first fundamental form

$$\frac{du^2 + dv^2}{v^2}.$$

We shall show that there is an isometry from H to D given by

$$w \mapsto z = \frac{w - i}{w + i}$$

where $w = u + iv \in H$ and $z = x + iy \in D$.

We write $|dz|^2 = dx^2 + dy^2$ and $|dw|^2 = du^2 + dv^2$. If $w = f(z)$ where $f : D \to H$ is holomorphic then

$$f'(z) = u_x + iv_x = v_y - iu_y$$

and so

$$|f'(z)|^2|dz|^2 = (u_x^2+v_x^2)(dx^2+dy^2) = (u_x dx+u_y dy)^2+(v_x dx+v_y dy)^2 = du^2+dv^2 = |dw|^2.$$

Thus we can substitute

$$|dw|^2 = \left|\frac{dw}{dz}\right|^2 |dz|^2 \tag{2}$$

to calculate how the first fundamental form is transformed by such a map.

The Möbius transformation

$$w \mapsto z = \frac{w - i}{w + i} \tag{3}$$

restricts to a smooth bijection from H to D because $w \in H$ if and only if $|w - i| < |w + i|$, and its inverse is also a Möbius transformation and hence is also smooth. Substituting (3) and (2) with

$$\frac{dw}{dz} = \frac{1}{w + i} - \frac{(w - i)}{(w + i)^2} = \frac{2i}{(w + i)^2}$$

into $v^{-2}|dw|^2$ gives $4(1 - |z|^2)^{-2}|dz|^2$, so this Möbius transformation gives us an isometry from H to D as required.

4.4 The second fundamental form

The first fundamental form describes the intrinsic geometry of a surface – the experience of an insect crawling around it. It is this that we can generalize to abstract surfaces. The second fundamental form relates to the way the surface sits in $\mathbf{R}^3$, though as we shall see, it is not independent of the first fundamental form.

First take a surface $\mathbf{r}(u,v)$ and push it inwards a distance t along its normal to get a one-parameter family of surfaces:

$$\mathbf{R}(u,v,t) = \mathbf{r}(u,v) - t\mathbf{n}(u,v)$$

with

$$\mathbf{R}_u = \mathbf{r}_u - t\mathbf{n}_u, \quad \mathbf{R}_v = \mathbf{r}_v - t\mathbf{n}_v.$$

We now have a first fundamental form $Edu^2 + 2Fdudv + Gdv^2$ depending on t and we calculate

$$\frac{1}{2}\frac{\partial}{\partial t}(Edu^2 + 2Fdudv + Gdv^2)|_{t=0} = -(\mathbf{r}_u \cdot \mathbf{n}_u du^2 + (\mathbf{r}_u \cdot \mathbf{n}_v + \mathbf{r}_v \cdot \mathbf{n}_u)dudv + \mathbf{r}_v \cdot \mathbf{n}_v dv^2).$$

The right hand side is the second fundamental form. From this point of view it is clearly the same type of object as the first fundamental form — a quadratic form on the tangent space.

In fact it is useful to give a slightly different expression. Since $\mathbf{n}$ is orthogonal to $\mathbf{r}_u$ and $\mathbf{r}_v$,

$$0 = (\mathbf{r}_u \cdot \mathbf{n})_u = \mathbf{r}_{uu} \cdot \mathbf{n} + \mathbf{r}_u \cdot \mathbf{n}_u$$

and similarly

$$\mathbf{r}_{uv} \cdot \mathbf{n} + \mathbf{r}_u \cdot \mathbf{n}_v = 0, \quad \mathbf{r}_{vu} \cdot \mathbf{n} + \mathbf{r}_v \cdot \mathbf{n}_u = 0$$

and since $\mathbf{r}_{uv} = \mathbf{r}_{vu}$ we have $\mathbf{r}_u \cdot \mathbf{n}_v = \mathbf{r}_v \cdot \mathbf{n}_u$. We then define:

Definition 20 *The second fundamental form of a surface is the expression*

$$Ldu^2 + 2Mdudv + Ndv^2$$

where $L = \mathbf{r}_{uu} \cdot \mathbf{n}, M = \mathbf{r}_{uv} \cdot \mathbf{n}, N = \mathbf{r}_{vv} \cdot \mathbf{n}$.

Examples:

1) The plane

$$\mathbf{r}(u,v) = \mathbf{a} + u\mathbf{b} + v\mathbf{c}$$

has $\mathbf{r}_{uu} = \mathbf{r}_{uv} = \mathbf{r}_{vv} = 0$ so the second fundamental form vanishes.

2) The sphere of radius a: here with the origin at the centre, $\mathbf{r} = a\mathbf{n}$ so

$$\mathbf{r}_u \cdot \mathbf{n}_u = a^{-1}\mathbf{r}_u \cdot \mathbf{r}_u, \quad \mathbf{r}_u \cdot \mathbf{n}_v = a^{-1}\mathbf{r}_u \cdot \mathbf{r}_v, \quad \mathbf{r}_v \cdot \mathbf{n}_v = a^{-1}\mathbf{r}_v \cdot \mathbf{r}_v$$

and

$$Ldu^2 + 2Mdudv + Ndv^2 = a^{-1}(Edu^2 + 2Fdudv + Gdv^2).$$

The plane is characterised by the vanishing of the second fundamental form:

Proposition 4.2 *If the second fundamental form of a surface vanishes, it is part of a plane.*

Proof: If the second fundamental form vanishes,

$$\mathbf{r}_u \cdot \mathbf{n}_u = 0 = \mathbf{r}_v \cdot \mathbf{n}_u = \mathbf{r}_u \cdot \mathbf{n}_v = \mathbf{r}_v \cdot \mathbf{n}_v$$

so that

$$\mathbf{n}_u = \mathbf{n}_v = 0$$

since $\mathbf{n}_u, \mathbf{n}_v$ are orthogonal to $\mathbf{n}$ and hence linear combinations of $\mathbf{r}_u, \mathbf{r}_v$. Thus $\mathbf{n}$ is constant. This means

$$(\mathbf{r} \cdot \mathbf{n})_u = \mathbf{r}_u \cdot \mathbf{n} = 0, \quad (\mathbf{r} \cdot \mathbf{n})_v = \mathbf{r}_v \cdot \mathbf{n} = 0$$

and so

$$\mathbf{r} \cdot \mathbf{n} = const$$

which is the equation of a plane. □

Consider now a surface given as the graph of a function $z = f(x, y)$:

$$\mathbf{r}(x, y) = x\mathbf{i} + y\mathbf{j} + f(x, y)\mathbf{k}.$$

Here

$$\mathbf{r}_x = \mathbf{i} + f_x\mathbf{k}, \qquad \mathbf{r}_y = \mathbf{j} + f_y\mathbf{k}$$

and so

$$\mathbf{r}_{xx} = f_{xx}\mathbf{k}, \quad \mathbf{r}_{xy} = f_{xy}\mathbf{k}, \quad \mathbf{r}_{yy} = f_{yy}\mathbf{k}.$$

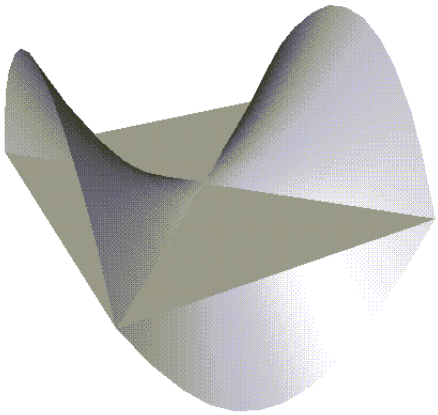
At a critical point of f, $f_x = f_y = 0$ and so the normal is $\mathbf{k}$. The second fundamental form is then the *Hessian* of the function at this point:

$$\begin{pmatrix} L & M \\ M & N \end{pmatrix} = \begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix}.$$

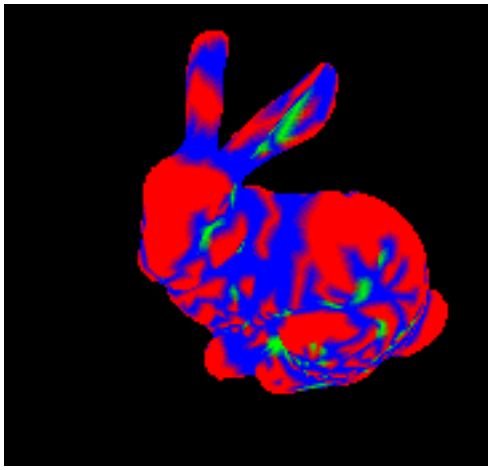
We can use this to qualitatively describe the behaviour of the second fundamental form at different points on the surface. For any point P parametrize the surface by its projection on the tangent plane and then $f(x, y)$ is the height above the plane. Now use the theory of critical points of functions of two variables.

If $f_{xx}f_{yy} - f_{xy}^2 > 0$ then the critical point is a local maximum if the matrix is negative definite and a local minimum if it is positive definite. For the surface the difference is only in the choice of normal so the local picture of the surface is like the sphere – it lies on one side of the tangent plane at the point P.

If on the other hand $f_{xx}f_{yy} - f_{xy}^2 < 0$ we have a saddle point and the surface lies on both sides of the tangent plane:



A general surface has points of both types, like this rabbit:



In fact any closed suface X in $\mathbf{R}^3$, not just rabbit-shaped ones, have both types of points.

Proposition 4.3 *Any closed surface X in $\mathbf{R}^3$ has points at which the second fundamental form is positive definite.*

Proof: Since X is compact, it is bounded and so can be surrounded by a large sphere. Gradually deflate the sphere until at radius R it touches X at a point. With X described locally as the graph of a function f we then have

$$f - (R - \sqrt{R^2 - x^2 - y^2}) \geq 0$$

and the first nonzero term in the Taylor series of this is

$$\frac{1}{2}(f_{xx}x^2 + 2f_{xy}xy + f_{yy}y^2) - \frac{1}{2R}(x^2 + y^2)$$

so

$$Lx^2 + 2Mxy + Ny^2 \geq \frac{1}{R}(x^2 + y^2) > 0.$$

□

It is easy to understand qualitatively the behaviour of a surface from whether $LN - M^2$ is positive or not. In fact there is a closely related function called the Gaussian curvature which we shall study next.

4.5 The Gaussian curvature

Definition 21 *The Gaussian curvature of a surface in* $\mathbf{R}^3$ *is the function*

$$K = \frac{LN - M^2}{EG - F^2}$$

Note that under a coordinate change

$$\begin{pmatrix} u_x & u_y \\ v_x & v_y \end{pmatrix} \begin{pmatrix} E & F \\ F & G \end{pmatrix} \begin{pmatrix} u_x & v_x \\ u_y & v_y \end{pmatrix} = \begin{pmatrix} E' & F' \\ F' & G' \end{pmatrix}$$

so taking determinants

$$(u_x v_y - u_y v_x)^2 (EG - F^2) = (E'G' - F'^2).$$

Since the second fundamental form is a quadratic form on the tangent space just like the first, it undergoes the same transformation, so the ratio $(LN - M^2)/(EG - F^2)$ is independent of the choice of coordinates.

Examples:

1. For a plane, $L = M = N = 0$ so $K = 0$

2. For a sphere of radius a, the second fundamental form is a^{-1} times the first so that $K = a^{-2}$.

We defined K in terms of the second fundamental form which we said describes the extrinsic geometry of the surface. In fact it only depends on E, F, G and its derivatives, and so is intrinsic – our insect crawling on the surface could in principle work it out. It was Gauss who showed this in 1828, a result he was particularly pleased with.

What it means is that if two surfaces are locally isometric, then the isometry maps the Gaussian curvature of one to the Gaussian curvature of the other – for example the Gaussian curvature of a bent piece of paper is zero because it is isometric to the plane. Also, we can define Gaussian curvature for an abstract Riemannian surface.

We prove Gauss's "egregious theorem", as he proudly called it, by a calculation. We consider locally a smooth family of tangent vectors

$$\mathbf{a} = f\mathbf{r}_u + g\mathbf{r}_v$$

where f and g are functions of u, v. If we differentiate with respect to u or v this is no longer necessarily tangential, but we can remove its normal component to make it so, and call this the *tangential derivative*:

$$\begin{aligned} \nabla_u \mathbf{a} &= \mathbf{a}_u - (\mathbf{n}\cdot\mathbf{a}_u)\mathbf{n} \\ &= \mathbf{a}_u + (\mathbf{n}_u\cdot\mathbf{a})\mathbf{n} \end{aligned}$$

since $\mathbf{a}$ and $\mathbf{n}$ are orthogonal.

The important thing to note is that this tangential derivative only depends on E, F, G and their derivatives, because we are taking a tangent vector like $\mathbf{r}_u$, differentiating it to get $\mathbf{r}_{uu}$ and $\mathbf{r}_{uv}$ and then projecting back onto the tangent plane which involves taking dot products like $\mathbf{r}_{uu}\cdot\mathbf{r}_u = (\mathbf{r}_u\cdot\mathbf{r}_u)_u/2 = E_u/2$ etc.

Now differentiate $\nabla_u\mathbf{a}$ tangentially with respect to v:

$$\nabla_v\nabla_u\mathbf{a} = \mathbf{a}_{vu} - (\mathbf{n}\cdot\mathbf{a}_{vu})\mathbf{n} + \nabla_v((\mathbf{n}_u\cdot\mathbf{a})\mathbf{n}).$$

But since we are taking the tangential component, we can forget about differentiating the coefficient of $\mathbf{n}$. Moreover, since $\mathbf{n}$ is a unit vector, $\mathbf{n}_v$ is already tangential, so we get:

$$\nabla_v\nabla_u\mathbf{a} = \mathbf{a}_{vu} - (\mathbf{n}\cdot\mathbf{a}_{vu})\mathbf{n} + (\mathbf{n}_u\cdot\mathbf{a})\mathbf{n}_v$$

Interchanging the roles of u and v and using the symmetry of the second derivative $\mathbf{a}_{uv} = \mathbf{a}_{vu}$ we get

$$\nabla_v\nabla_u\mathbf{a} - \nabla_u\nabla_v\mathbf{a} = (\mathbf{n}_u\cdot\mathbf{a})\mathbf{n}_v - (\mathbf{n}_v\cdot\mathbf{a})\mathbf{n}_u = (\mathbf{n}_u\wedge\mathbf{n}_v)\wedge\mathbf{a}.$$

Now

$$\mathbf{n}_u\wedge\mathbf{n}_v = \lambda\mathbf{n} \tag{4}$$

so we see that $\nabla_v\nabla_u - \nabla_u\nabla_v$ acting on $\mathbf{a}$ rotates it in the tangent plane by 90° and multiplies by λ, where λ is intrinsic. Now from (4),

$$\lambda\mathbf{n}\cdot\mathbf{r}_u\wedge\mathbf{r}_v = (\mathbf{n}_u\wedge\mathbf{n}_v)\cdot(\mathbf{r}_u\wedge\mathbf{r}_v) = (\mathbf{n}_u\cdot\mathbf{r}_u)(\mathbf{n}_v\cdot\mathbf{r}_v) - (\mathbf{n}_u\cdot\mathbf{r}_v)(\mathbf{n}_v\cdot\mathbf{r}_u) = LN - M^2$$

but also

$$\mathbf{n} \cdot \mathbf{r}_u \wedge \mathbf{r}_v = \sqrt{EG - F^2}$$

which gives

$$\lambda = (LN - M^2)/\sqrt{EG - F^2}. \tag{5}$$

It follows that $LN - M^2$ and hence K depends only on the first fundamental form.

4.6 The Gauss-Bonnet theorem

One of the beautiful features of the Gaussian curvature is that it can be used to determine the topology of a closed orientable surface – more precisely we can determine the Euler characteristic by integrating K over the surface. We shall do this by using a triangulation and summing the integrals over the triangles, but the boundary terms involve another intrinsic invariant of a curve in a surface:

Definition 22 *The geodesic curvature κ_g of a smooth curve in X is defined by*

$$\kappa_g = \mathbf{t}' \cdot (\mathbf{n} \wedge \mathbf{t})$$

where $\mathbf{t}$ is the unit tangent vector of the curve, which is parametrized by arc length.

This is the tangential derivative of the unit tangent vector $\mathbf{t}$ and so is intrinsic.

The first version of Gauss-Bonnet is:

Theorem 4.4 *Let γ be a smooth simple closed curve on a coordinate neighbourhood of a surface X enclosing a region R, then*

$$\int_\gamma \kappa_g ds = 2\pi - \int_R K dA$$

where κ_g is the geodesic curvature of γ, ds is the element of arc-length of γ, K is the Gaussian curvature of X and dA the element of area of X.

Proof: Recall Stokes' theorem in $\mathbf{R}^3$:

$$\int_C \mathbf{a} \cdot d\mathbf{s} = \int_S \operatorname{curl} \mathbf{a} \cdot d\mathbf{S}$$

for a curve C spanning a surface S. In the xy plane with $\mathbf{a} = (P, Q, 0)$ this becomes Green's formula

$$\int_\gamma (Pu' + Qv')dt = \int_R (Q_u - P_v) du dv \tag{6}$$

Now choose a unit length tangent vector field, for example $\mathbf{e} = \mathbf{r}_u/\sqrt{E}$. Then $\mathbf{e}, \mathbf{n}\wedge\mathbf{e}$ is an orthonormal basis for each tangent space. Since $\mathbf{e}$ has unit length, $\nabla_u\mathbf{e}$ is tangential and orthogonal to $\mathbf{e}$ so there are functions P, Q such that

$$\nabla_u\mathbf{e} = P\,\mathbf{n}\wedge\mathbf{e}, \qquad \nabla_v\mathbf{e} = Q\,\mathbf{n}\wedge\mathbf{e}.$$

In Green's formula, take $\mathbf{a} = (P, Q, 0)$ then the left hand side of (6) is

$$\int_\gamma (u'\nabla_u\mathbf{e} + v'\nabla_v\mathbf{e})\cdot(\mathbf{n}\wedge\mathbf{e}) = \int_\gamma \mathbf{e}'\cdot(\mathbf{n}\wedge\mathbf{e}) \tag{7}$$

Let $\mathbf{t}$ be the unit tangent to γ, and write it relative to the orthonormal basis

$$\mathbf{t} = \cos\theta\,\mathbf{e} + \sin\theta\,\mathbf{n}\wedge\mathbf{e}.$$

So

$$\mathbf{t}'\cdot(\mathbf{n}\wedge\mathbf{e}) = \cos\theta\,\mathbf{e}'\cdot(\mathbf{n}\wedge\mathbf{e}) + \cos\theta\,\theta'.$$

The geodesic curvature of γ is defined by $\kappa_g = \mathbf{t}'\cdot(\mathbf{n}\wedge\mathbf{t})$ so

$$\mathbf{t}' = \alpha\mathbf{n} + \kappa_g\mathbf{n}\wedge\mathbf{t} = \alpha\mathbf{n} + \kappa_g(\cos\theta\,\mathbf{n}\wedge\mathbf{e} - \sin\theta\,\mathbf{e})$$

and so

$$\kappa_g = \mathbf{e}'\cdot(\mathbf{n}\wedge\mathbf{e}) + \theta'.$$

We can therefore write (7) as

$$\int_\gamma (\kappa_g - \theta')ds = \int_\gamma \kappa_g ds - 2\pi.$$

To compute the right hand side of (6), note that

$$\nabla_v\nabla_u\mathbf{e} = \nabla_v(P\,\mathbf{n}\wedge\mathbf{e}) = P_v\,\mathbf{n}\wedge\mathbf{e} + P\,\mathbf{n}\wedge\nabla_v\mathbf{e} = P_v\,\mathbf{n}\wedge\mathbf{e} + PQ\,\mathbf{n}\wedge(\mathbf{n}\wedge\mathbf{e})$$

since $\mathbf{n}_v\wedge\mathbf{e}$ is normal. Interchanging the roles of u and v and subtracting we obtain

$$(\nabla_v\nabla_u - \nabla_u\nabla_v)\mathbf{e} = (P_v - Q_u)\mathbf{n}\wedge\mathbf{e}$$

and from (5) this is equal to $K\sqrt{EG - F^2}$.

Applying Green's theorem and using $dA = \sqrt{EG - F^2}dudv$ gives the result. □

Note that the extrinsic normal was only used to define $\mathbf{n}\wedge\mathbf{e}$ which is one of the two unit tangent vectors to X orthogonal to $\mathbf{e}$. If the surface is orientable we can systematically make a choice and then the proof is intrinsic.

If the curve γ is piecewise smooth – a curvilinear polygon – then θ jumps by the external angle δ_i at each vertex, so the integral of θ' which is 2π in the theorem is replaced by

$$\int_\gamma \theta' ds = 2\pi - \sum_i \delta_i = \sum_i \alpha_i - (n-2)\pi$$

where α_i are the internal angles. The Gauss-Bonnet theorem gives in particular:

Theorem 4.5 *The sum of the angles of a curvilinear triangle is*

$$\pi + \int_R K dA + \int_\gamma \kappa_g ds.$$

Examples:

1. In the plane, a line has constant unit tangent vector and so $\kappa_g = 0$. Since the Gaussian curvature is zero too this says that the sum of the angles of a triangle is π.

2. A great circle on the unit sphere also has κ_g zero, for example if $\gamma(s) = (\cos s, \sin s, 0)$, then $\mathbf{t} = (-\sin s, \cos s, 0)$ and $\mathbf{t}' = -(\cos s, \sin s, 0)$ which is normal to the sphere. Since here $K = 1$, we have, for the triangle Δ with angles A, B, C

$$\alpha + \beta + \gamma = \pi + \text{Area}(ABC).$$

Here is the most interesting version of Gauss-Bonnet:

Theorem 4.6 *If X is a smooth orientable closed surface with a Riemannian metric, then*

$$\int_X K dA = 2\pi\chi(X)$$

Proof: Take a smooth triangulation so that each triangle is inside a coordinate neighbourhood and apply Theorem 4.5 and add. The integrals of κ_g on the edges cancel because the orientation on the edge from adjacent triangles is opposite (this is for Green's theorem – we use the anticlockwise orientation on γ). The theorem gives the total sum of internal angles as

$$\pi F + \int_X K dA.$$

But around each vertex the internal angles add to 2π so we have

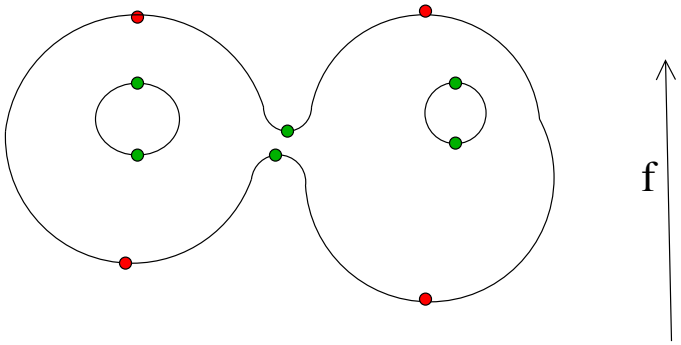
$$2\pi V = \pi F + \int_X K dA$$

and as our faces are triangles whose sides meet in pairs there are $3F/2$ edges. Hence

$$2\pi\chi(X) = 2\pi(V - E + F) = \pi F + \int_X K dA - 3\pi F + 2\pi F = \int_X K dA.$$

□

The Gauss-Bonnet theorem and its method of proof give another formula for the Euler characteristic, involving smooth real-valued functions $f : X \to \mathbf{R}$ on a closed surface X. Since X is compact, f certainly has a maximum and a minimum, but may have other critical points too. Think of a surface in $\mathbf{R}^3$ and the function f given by its height above a plane:



This has 2 maxima, 2 minima and 6 saddle points. We shall be able to calculate the Euler characteristic from these numbers.

First recall that a smooth function $f(u, v)$ has a critical point at a if

$$f_u(a) = f_v(a) = 0.$$

Because of the chain rule, this condition is independent of coordinates: if $u = u(x, y), v = v(x, y)$ then

$$f_x = f_u u_x + f_v v_x, \qquad f_y = f_u u_y + f_v v_y$$

so f_u and f_v vanish if and only if f_x and f_y vanish. This means we can unambiguously talk about the critical points of a smooth function on a surface X.

The Hessian matrix

$$\begin{pmatrix} f_{uu} & f_{uv} \\ f_{uv} & f_{vv} \end{pmatrix}$$

at a critical point transforms like

$$\begin{pmatrix} u_x & u_y \\ v_x & v_y \end{pmatrix} \begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix} \begin{pmatrix} u_x & v_x \\ u_y & v_y \end{pmatrix} = \begin{pmatrix} f_{uu} & f_{uv} \\ f_{uv} & f_{vv} \end{pmatrix}$$

and so

$$(f_{uu}f_{vv} - f_{uv}^2) = (u_x v_y - u_y v_x)^2 (f_{uu}f_{vv} - f_{uv}^2)$$

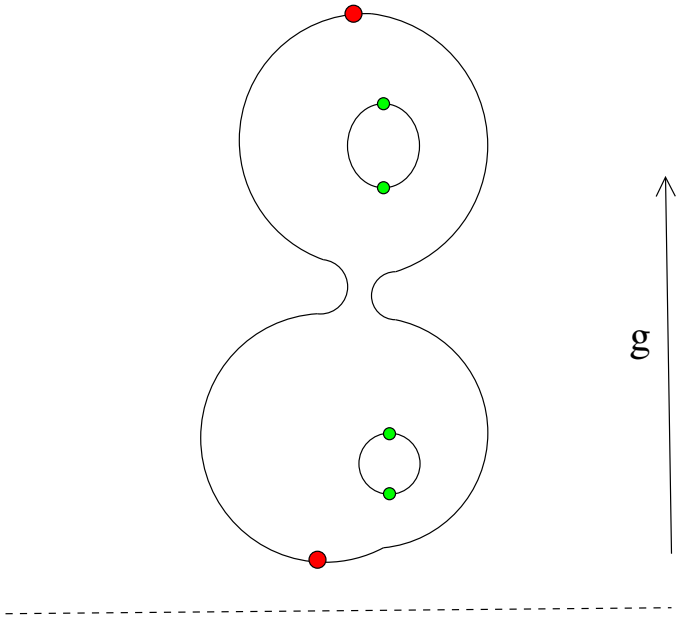
therefore to say that the determinant of the Hessian is non-zero, or positive or negative, is again independent of the choice of coordinate.

Definition 23 *A function f on a surface X has a nondegenerate critical point at $a \in X$ if its Hessian at a is invertible.*

We know from calculus that if $f_{uu}f_{vv} - f_{uv}^2 > 0$ and $f_{uu} > 0$ we have a local minimum, if $f_{uu} < 0$ a local maximum and if $f_{uu}f_{vv} - f_{uv}^2 < 0$ a saddle point. The theorem is the following:

Theorem 4.7 *Let f be a smooth function on a closed surface X with nondegenerate critical points, then the Euler characteristic $\chi(X)$ is the number of local maxima and minima minus the number of saddle points.*

In the picture, we have $\chi(X) = 4 - 6 = -2$ which is correct for the connected sum of two tori. If we turn it on its side we get one maximum, one minimum and 4 saddle points again giving the same value: $2 - 4 = -2$.



Proof: Given a function f on X we can define its gradient vector field:

$$\mathbf{a} = \frac{1}{EG - F^2}[(Gf_u - Ff_v)\mathbf{r}_u + (Ef_v - Ff_u)\mathbf{r}_v]$$

which is normal to the contour lines of f. Away from the critical points we can normalize it to get a unit vector field $\mathbf{e}$. Surround each critical point by a small closed curve γ_i enclosing a disc R_i. Let Y be the complement of the discs, then from the argument of Theorem 4.4

$$\int_Y K dA = -\sum_i \int_{\gamma_i} \mathbf{e}' \cdot (\mathbf{n} \wedge \mathbf{e}) ds$$

using the negative sign because Y is outside R_i.

Inside R_i we choose a unit vector field $\mathbf{f}$ and then we get

$$\int_{R_i} K dA = \int_{\gamma_i} \mathbf{f}' \cdot (\mathbf{n} \wedge \mathbf{f}) ds$$

so adding gives

$$\int_X K dA = \sum_i \int_{\gamma_i} [\mathbf{f}' \cdot (\mathbf{n} \wedge \mathbf{f}) - \mathbf{e}' \cdot (\mathbf{n} \wedge \mathbf{e})] ds.$$

From the proof of the theorem we had

$$\kappa_g = \mathbf{e}' \cdot (\mathbf{n} \wedge \mathbf{e}) + \theta' = \mathbf{f}' \cdot (\mathbf{n} \wedge \mathbf{f}) + \phi'$$

where θ is the angle between γ' and $\mathbf{e}$ and ϕ between γ' and $\mathbf{f}$. So the contribution is just the change in angle between the vector field $\mathbf{e}$ and a fixed one $\mathbf{f}$ which extends. This is an integer multiple of 2π so we can evaluate it by deforming to the standard Euclidean case. A local minimum is $f = x^2 + y^2$ which gives

$$\mathbf{e} = (\cos\theta, \sin\theta)$$

and contributes $+1$, as does the local minimum $-(\cos\theta, \sin\theta)$. For a saddle point $f = x^2 - y^2$ which gives

$$\mathbf{e} = (\cos\theta, -\sin\theta) = (\cos(-\theta), \sin(-\theta)$$

and contributes -1. □

4.7 Geodesics

Geodesics on a surface are curves which are the analogues of straight lines in the plane. Lines can be thought of in two ways:

- shortest curves
- straightest curves

The first point of view says that a straight line minimizes the distance between any two of its points. Conceptually this leads to the idea of stretching a string between two points on a surface until it tightens, and this certainly is one approach to geodesics. The second approach is however generally easier. A line is straightest because its tangent vector doesn't change – it is constant along the line. We generalize this to a curve on a surface by insisting that the component of $\mathbf{t}'$ tangential to the surface should vanish. Or....

Definition 24 *A geodesic on a surface X is a curve $\boldsymbol{\gamma}(s)$ on X such that $\mathbf{t}'$ is normal to the surface.*

From Definition 22 this is the same as saying that the geodesic curvature vanishes.

The general problem of finding geodesics on a surface is very complicated. The case of the ellipsoid is a famous example, needing hyperelliptic functions to solve it – integrals of $dz/\sqrt{p(z)}$ where $p(z)$ is a polynomial of degree 6. But there are cheap ways to find some of them, as in these examples:

Examples:

1) The normal to a curve in the plane is parallel to the plane, so the condition that $\mathbf{t}'$ is normal to the plane means $\mathbf{t}' = 0$ which integrates to $\mathbf{r} = s\mathbf{a} + \mathbf{c}$, the equation of a straight line. Geodesics in the plane really are straight lines, then.

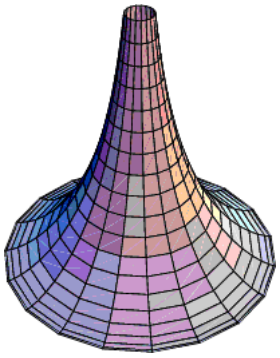
2) Take the unit sphere and a plane section through the origin. We saw earlier that $\kappa_g = 0$ here.

3) Similarly, any plane of symmetry intersects a surface in a geodesic, because the normal to the surface at such a point must be invariant under reflection in the plane of symmetry and hence lie in that plane. It is orthogonal to the tangent vector of the curve of intersection and so $\mathbf{t}'$ points normally.

A useful class of examples is provided by a surface of revolution

$$\mathbf{r}(u, v) = f(u)(\cos v\, \mathbf{i} + \sin v\, \mathbf{j}) + u\mathbf{k}$$

The reflection $(x, y, z) \mapsto (x, -y, z)$ maps the surface to itself, as, by symmetry, does any reflection in a plane containing the z-axis. So the *meridians* $v = const.$ are geodesics:



To find the geodesics in general we need to solve a nonlinear system of ordinary differential equations:

Proposition 4.8 *A curve* $\boldsymbol{\gamma}(s) = (u(s), v(s))$ *on a surface parametrized by arc length is a geodesic if and only if*

$$\begin{aligned} \frac{d}{ds}(Eu' + Fv') &= \frac{1}{2}(E_u u'^2 + 2F_u u'v' + G_u v'^2) \\ \frac{d}{ds}(Fu' + Gv') &= \frac{1}{2}(E_v u'^2 + 2F_v u'v' + G_v v'^2) \end{aligned}$$

Proof: We have for the curve $\boldsymbol{\gamma}$

$$\mathbf{t} = \mathbf{r}_u u' + \mathbf{r}_v v'$$

and it is a geodesic if and only if $\mathbf{t}'$ is normal i.e.

$$\mathbf{t}' \cdot \mathbf{r}_u = \mathbf{t}' \cdot \mathbf{r}_v = 0.$$

Now

$$\mathbf{t}' \cdot \mathbf{r}_u = (\mathbf{t} \cdot \mathbf{r}_u)' - \mathbf{t} \cdot \mathbf{r}_u'$$

so the first equation is

$$(\mathbf{t} \cdot \mathbf{r}_u)' = \mathbf{t} \cdot \mathbf{r}_u'.$$

The left hand side is

$$\frac{d}{ds}((\mathbf{r}_u u' + \mathbf{r}_v v') \cdot \mathbf{r}_u) = \frac{d}{ds}(Eu' + Fv')$$

an the right hand side is

$$\begin{aligned}\mathbf{t}\cdot(\mathbf{r}_{uu}u'+\mathbf{r}_{uv}v') &= \mathbf{r}_u\cdot\mathbf{r}_{uu}u'^2+(\mathbf{r}_v\cdot\mathbf{r}_{uu}+\mathbf{r}_u\cdot\mathbf{r}_{uv})u'v'+\mathbf{r}_v\cdot\mathbf{r}_{uv}v'^2\\ &= \frac{1}{2}E_uu'^2+(\mathbf{r}_v\cdot\mathbf{r}_u)_uu'v'+\frac{1}{2}G_uv'^2\\ &= \frac{1}{2}(E_uu'^2+2F_uu'v'+G_uv'^2)\end{aligned}$$

The other equation follows similarly. □

It is clear from 4.8 that geodesics only depend on the first fundamental form, so that geodesics can be defined for abstract surfaces and moreover an isometry takes geodesics to geodesics.

Examples:

1) The plane: $E=1, F=0, G=1$ in Cartesian coordinates, so the geodesic equations are

$$x''=0=y''$$

which gives straight lines

$$x=\alpha_1 s+\beta_1, \quad y=\alpha_2 s+\beta_2.$$

2) The cylinder

$$\mathbf{r}(u,v)=a(\cos v\,\mathbf{i}+\sin v\,\mathbf{j})+u\mathbf{k}$$

has first fundamental form

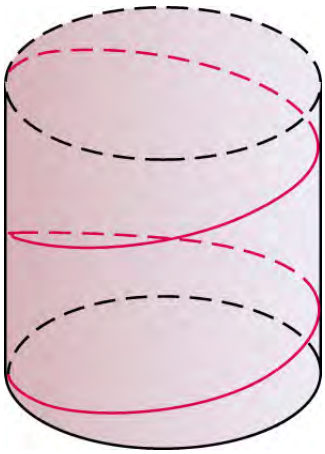
$$du^2+a^2dv^2=du^2+d(av)^2.$$

This is isometric to the plane so the geodesics are of the form

$$u=\alpha_1 s+\beta_1, \quad v=\alpha_2 s+\beta_2$$

which gives a helix

$$\boldsymbol{\gamma}=a(\cos(\alpha_2 s+\beta_2)\,\mathbf{i}+\sin(\alpha_2 s+\beta_2)\,\mathbf{j})+(\alpha_1 s+\beta_1)\mathbf{k}$$



The differential equation for geodesics gives us the following general fact:

Proposition 4.9 *Through each point P on a surface and in each direction at P there passes a unique geodesic.*

Proof: We are solving a differential equation of the form

$$u'' = a(u, v, u', v'), \quad v'' = b(u, v, u', v')$$

or equivalently a first order system

$$\begin{aligned} u' &= p \\ v' &= q \\ p' &= a(u, v, p, q) \\ q' &= b(u, v, p, q) \end{aligned}$$

and the Cauchy existence theorem (see Appendix B) gives a unique solution with initial conditions (u, v, p, q), namely the point of origin and the direction. □

Example: Given a point **a** on the unit sphere and a tangential direction **b** the span of **a**, **b** is a plane through the origin which meets the sphere in a great circle through **a** with tangent **b**. Thus *every* geodesic is a great circle.

There is one case – a surface of revolution – where the geodesic equations can be "solved", or anyway, reduced to a single integration. We have

$$E = 1 + f'(u)^2, \quad F = 0, \quad G = f(u)^2$$

and the equations become

$$\begin{aligned} \frac{d}{ds}((1+f'^2)u') &= f'(f''u'^2 + fv'^2) \\ \frac{d}{ds}(f^2v') &= 0 \end{aligned}$$

We ignore the first equation – it is equivalent to a more obvious fact below. The second says that

$$f^2v' = c \tag{8}$$

where c is a constant. Now use the fact that the curve is parametrized by arc length (this is an "integral" of the equations), and we get

$$(1+f'^2)u'^2 + f^2v'^2 = 1 \tag{9}$$

Substitute for v' from (8) in (9) to get

$$(1+f'^2)u'^2 + \frac{c^2}{f^2} = 1$$

and then

$$s = \int f\sqrt{\frac{1+f'^2}{f^2-c^2}}du$$

which is "only" an integration. Having solved this by $u = h(s)$, v can be determined by a further integration from (8):

$$v(s) = \int \frac{c}{f(h(t))^2}dt.$$

If we are only interested in the curve and not its arclength parametrization, then (8) and (9) give

$$(1+f'(u)^2)\left(\frac{du}{dv}\right)^2 + f(u)^2 = \frac{f(u)^4}{c^2}$$

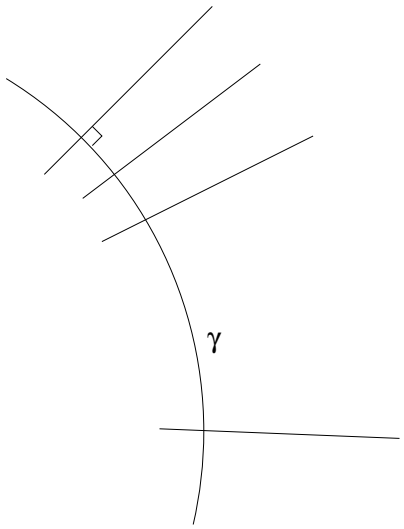
which reduces to the single integration

$$v = \int \frac{c}{f(u)}\sqrt{\frac{1+f'(u)^2}{f(u)^2-c^2}}du.$$

4.8 Gaussian curvature revisited

We may not be able to solve the geodesic equations explicitly, but existence of geodesics through a given point and in a given direction give rise to various natural coordinate systems, modelled on Cartesian coordinates. Here is one: choose a geodesic γ parametrized by arc length. Through the point $\gamma(v)$ take the geodesic $\gamma_v(s)$ which intersects γ orthogonally, and define

$$\mathbf{r}(u, v) = \gamma_v(u).$$



Since $\mathbf{r}_u$ and $\mathbf{r}_v$ are orthogonal at $u = 0$ they are linearly independent in a neighbourhood and so are good coordinates.

Now the curves $v = const.$ are parametrized by arc length, so $E = 1$. These curves are also geodesics and u is arc length so in the second geodesic equation

$$\frac{d}{ds}(Fu' + Gv') = \frac{1}{2}(E_v u'^2 + 2F_v u'v' + G_v v'^2)$$

we put $v = const.$ and $u = s$ which, with $E = 1$, gives $F_u = 0$. But F vanishes at $u = 0$ because the two geodesics are orthogonal there, hence $F = 0$ and the first fundamental form is

$$du^2 + G(u, v)dv^2.$$

In this form the Gaussian curvature is simple:

Proposition 4.10 *The Gaussian curvature of the metric $du^2 + G(u, v)dv^2$ is*

$$K = -G^{-1/2}(G^{1/2})_{uu}$$

Examples:

1. For the plane $dx^2 + dy^2$, $G = 1$ and $K = 0$.

2. For the unit sphere with first fundamental form $du^2 + \sin^2 u dv^2$, $G = \sin^2 u$ so

$$K = -\frac{1}{\sin u}(\sin u)_{uu} = \frac{1}{\sin u}\sin u = 1.$$

3. For the upper half-space with metric $(dx^2 + dy^2)/y^2$ put $u = \log y$ and $v = x$ and then we have $du^2 + e^{-2u}dv^2$, so that

$$K = -e^u(e^{-u})_{uu} = -e^u e^{-u} = -1.$$

Proof: Recall the tangential derivative ∇: the tangential component of the ordinary derivative. Then since by construction $\mathbf{r}_u$ is the unit tangent vector of a geodesic, by the definition of a geodesic its u-derivative is normal so $\nabla_u \mathbf{r}_u = 0$.

Consider now $\nabla_v \mathbf{r}_u = A\mathbf{r}_u + B\mathbf{r}_v$. The dot product with $\mathbf{r}_u$ gives

$$E_v/2 = \mathbf{r}_{vu} \cdot \mathbf{r}_u = A$$

but $E = 1$ so $A = 0$.

Using $E = 1$ and $F = 0$ the product with $\mathbf{r}_v$ gives

$$G_u/2 = \mathbf{r}_v \cdot \mathbf{r}_{vu} = BG.$$

Now from (5)

$$(\nabla_v \nabla_u - \nabla_u \nabla_v)\mathbf{r}_u = K\sqrt{EG - F^2}\mathbf{n} \wedge \mathbf{r}_u = KG^{1/2}(\mathbf{r}_v G^{-1/2}) = K\mathbf{r}_v$$

But the left hand side (using $\nabla_u \mathbf{r}_v = \nabla_v \mathbf{r}_u$ which follows from $\mathbf{r}_{uv} = \mathbf{r}_{vu}$) is

$$-\nabla_u(G_u/2G)\mathbf{r}_v = -((G_u/2G)_u + (G_u/2G)^2)\mathbf{r}_v$$

which gives the result. □

With this coordinate system we can characterize surfaces with *constant* Gaussian curvature:

Theorem 4.11 *A surface with $K = 0$ is locally isometric to the plane, with $K = 1$ locally isometric to the unit sphere and with $K = -1$ locally isometric to the upper half space with metric $(dx^2 + dy^2)/y^2$.*

Proof: Use the form $du^2 + Gdv^2$.

i) If $K = 0$ then $(G^{1/2})_{uu} = 0$ so $G = A(v)u + B(v)$. But at $u = 0$, $\mathbf{r}_u$ and $\mathbf{r}_v$ are unit so $B(v) = 1$. Also, the curve $u = 0$ is a geodesic – the initial curve γ – with v arc length. So the geodesic equation

$$\frac{d}{ds}(Eu' + Fv') = \frac{1}{2}(E_u u'^2 + 2F_u u'v' + G_u v'^2)$$

gives $0 = G_u(0, v)/2$ and this means in our case $A(v) = 0$. The first fundamental form is therefore $du^2 + dv^2$ and by 4.1 this is isometric to the plane.

ii) If $K = 1$, the equation for $G^{1/2}$ is

$$(G^{1/2})_{uu} + G^{1/2} = 1$$

which is solved by $G^{1/2} = A(v)\sin u + B(v)\cos u$. The boundary conditions give $G = \cos^2 u$ and the metric $du^2 + \cos^2 u\, dv^2$ – the sphere.

iii) If $K = -1$ we have $du^2 + \cosh^2 u\, dv^2$. The substitution $x = v\tanh u, y = v\,\mathrm{sech}\, u$ takes this into $(dx^2 + dy^2)/y^2$. □

5 The hyperbolic plane

5.1 Isometries

We just saw that a metric of constant negative curvature is modelled on the upper half space H with metric

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

which is called the *hyperbolic plane*. This is an abstract surface in the sense that we are not considering a first fundamental form coming from an embedding in $\mathbf{R}^3$, and yet it is concrete enough to be able to write down and see everything explicitly. First we consider the isometries from H to itself.

If $a, b, c, d \in \mathbf{R}$ and $ad - bc > 0$ then the Möbius transformation

$$z \mapsto w = \frac{az + b}{cz + d} \tag{10}$$

restricts to a smooth bijection from H to H with smooth inverse

$$w \mapsto z = \frac{dw - b}{-cw + a}.$$

If we substitute

$$w = \frac{az + b}{cz + d} \text{ and } dw = \left(\frac{a}{cz + d} - \frac{c(az + b)}{(cz + d)^2}\right) dz = \frac{(ad - bc)}{(cz + d)^2} dz$$

into

$$\frac{du^2 + dv^2}{v^2} = \frac{4|dw|^2}{|w - \bar{w}|^2}$$

we get

$$\frac{4(ad - bc)^2|dz|^2}{|(az + b)(c\bar{z} + d) - (a\bar{z} + b)(cz + d)|^2} = \frac{4(ad - bc)^2|dz|^2}{|(ad - bc)(z - \bar{z})|^2} = \frac{4|dz|^2}{|z - \bar{z}|^2} = \frac{dx^2 + dy^2}{y^2}.$$

Thus this Möbius transformation is an isometry from H to H. So is the transformation $z \mapsto -\bar{z}$, and hence the composition

$$z \mapsto \frac{b - a\bar{z}}{d - c\bar{z}} \tag{11}$$

is also an isometry from H to H. In fact (10) and (11) give all the isometries of H, as we shall see later.

In **4.3** we saw that the unit disc D with the metric

$$\frac{du^2 + dv^2}{(1 - u^2 - v^2)^2}$$

is isometric to H, so any statements about H transfer also to D. Sometimes the picture is easier in one model or the other. The isometries $f : D \to D$ of the unit disc model of the hyperbolic plane are also Möbius transformations, if they preserve orientations, or compositions of Möbius transformations with $z \mapsto \bar{z}$ if they reverse orientations. The Möbius transformations which map D to itself are those of the form

$$z \mapsto w = e^{i\theta}\left(\frac{z - a}{1 - \bar{a}z}\right)$$

where $a \in D$ and $\theta \in \mathbf{R}$. They are isometries because substituting for w and

$$dw = e^{i\theta}\frac{(1 - |a|^2)}{(1 - \bar{a}z)^2}dz$$

in $4(1 - |w|^2)^{-2}|dw|^2$ gives $4(1 - |z|^2)^{-2}|dz|^2$.

Notice that the group $\mathrm{Isom}(H)$ of isometries of H acts transitively on H because if $a + ib \in H$ then $b > 0$ so the transformation

$$z \mapsto bz + a$$

is an isometry of H which takes i to $a + ib$. Similarly the group $\mathrm{Isom}(D)$ of isometries of D acts transitively on D since if $a \in D$ then the isometry

$$z \mapsto \frac{z - a}{1 - \bar{a}z}$$

maps a to 0. Notice also that the subgroup of $\mathrm{Isom}(D)$ consisting of those isometries which fix 0 contains all the rotations

$$z \mapsto e^{i\theta}z$$

about 0 as well as $z \mapsto \bar{z}$.

5.2 Geodesics

The hyperbolic plane is a case where the geodesic equations can be easily solved: since $E = G = 1/y^2$ and $F = 0$, and these are independent of x, the first geodesic equation

$$\frac{d}{ds}(Eu' + Fv') = \frac{1}{2}(E_u u'^2 + 2F_u u'v' + G_u v'^2)$$

becomes

$$\frac{d}{ds}(\frac{x'}{y^2}) = 0$$

and so

$$x' = cy^2. \tag{12}$$

We also know that parametrization is by arc length in these equations so

$$\frac{x'^2 + y'^2}{y^2} = 1 \tag{13}$$

If $c = 0$ we get $x = const.$, which is a vertical line. Suppose $c \neq 0$, then from (12) and (13) we have

$$\frac{dy}{dx} = \sqrt{\frac{y^2 - c^2y^4}{c^2y^4}}$$

or

$$\frac{cydy}{\sqrt{1 - c^2y^2}} = dx$$

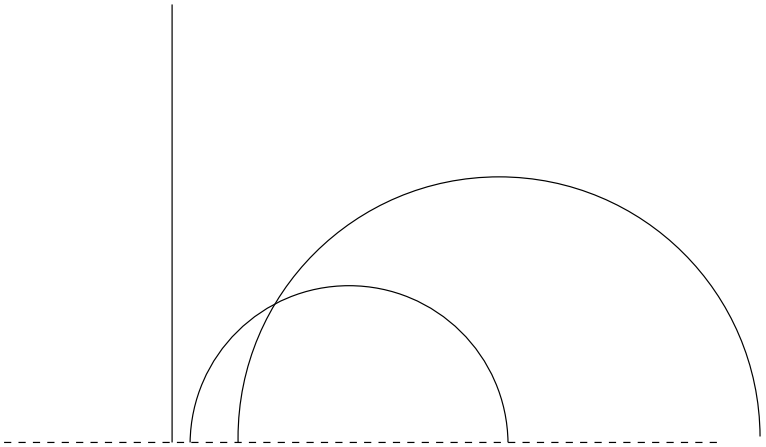
which integrates directly to

$$-c^{-1}\sqrt{1 - c^2y^2} = x - a$$

or

$$(x - a)^2 + y^2 = 1/c^2$$

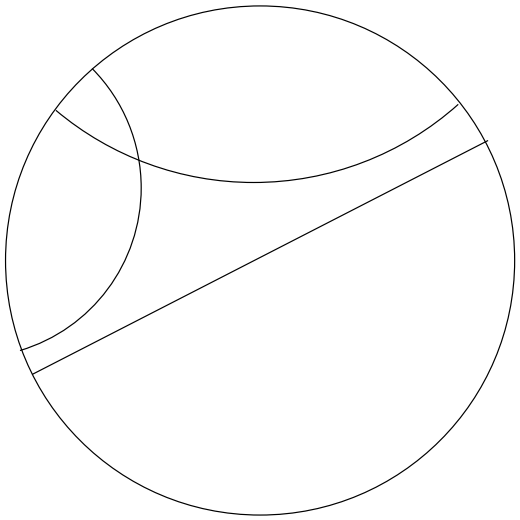
which is a semicircle centred on the real axis.



The isometry from H to D given by

$$w \mapsto z = \frac{w - i}{w + i}$$

takes geodesics to geodesics (since it is an isometry) and it is the restriction to H of a Möbius transformation $\mathbf{C} \cup \{\infty\} \to \mathbf{C} \cup \{\infty\}$ which takes circles and lines to circles and lines, preserves angles and maps the real axis to the unit circle in $\mathbf{C}$. It therefore follows that the geodesics in D are the circles and lines in D which meet the unit circle at right angles.



Using geodesics we can now show that any isometry is a Möbius transformation as above. So suppose that $F : D \to D$ is an isometry. Take a Möbius isometry G taking $F(0)$ to 0, then we need to prove that GF is Möbius. This is an isometry fixing 0, so it takes geodesics through 0 to geodesics through 0. It preserves angles, so acts on those geodesics by a rotation or reflection. It also preserves distance so it takes a point on a geodesic a distance r from the origin to another point at the same distance. However, as we noted above, each rotation $R : z \mapsto e^{i\theta}z$ is a Möbius isometry, so composing with this we see that $RGF = 1$ and $F = (RG)^{-1}$ is a Möbius isometry.

5.3 Angles and distances

Hyperbolic angles in H and in D are the same as Euclidean angles, since their first fundamental forms satisfy $E = G$ and $F = 0$. Distances between points are given by the lengths of geodesics joining the points. Since the interval $(-1, 1)$ is a geodesic in the unit disc D, the distance from 0 to any $x \in (0, 1)$ is given by the hyperbolic length of the line segment $[0, x]$, which is

$$\int_0^x \sqrt{Eu'^2 + 2Fu'v' + Gv'^2}dt = \int_0^x \frac{dt}{1-t^2} = 2\tanh^{-1}x$$

where $u(t) = t$ and $v(t) = 0$ and $E = G = (1 - u^2 - v^2)^{-2}$ and $F = 0$. Given any $a, b \in D$ we can choose $\theta \in \mathbf{R}$ such that

$$e^{i\theta} \frac{b-a}{1-\bar{a}b} = \left| \frac{b-a}{1-\bar{a}b} \right|$$

is real and positive, so its distance from 0 is

$$2 \tanh^{-1} \left| \frac{b-a}{1-\bar{a}b} \right|.$$

Since the isometry

$$z \mapsto e^{i\theta} \frac{z-a}{1-\bar{a}z}$$

preserves distances and takes a to 0 and b to $e^{i\theta}(b-a)/(1-\bar{a}b)$, it follows that the hyperbolic distance from a to b in D is

$$d_D(a,b) = 2 \tanh^{-1} \left| \frac{b-a}{1-\bar{a}b} \right|.$$

We can work out hyperbolic distances in H in a similar way by first calculating the distance from i to λi for $\lambda \in [1, \infty)$ as the length of the geodesic from i to λi given by the imaginary axis, which is

$$\int_1^\lambda \frac{dt}{t} = \log \lambda,$$

and then given $a, b \in H$ finding an isometry of H which takes a to i and b to λi for some $\lambda \in [1, \infty)$. Alternatively, since we have an isometry from H to D given by

$$w \mapsto z = \frac{w-i}{w+i},$$

the hyperbolic distance between points $a, b \in H$ is equal to the hyperbolic distance between the corresponding points $(a-i)/(a+i)$ and $(b-i)/(b+i)$ in D, which is

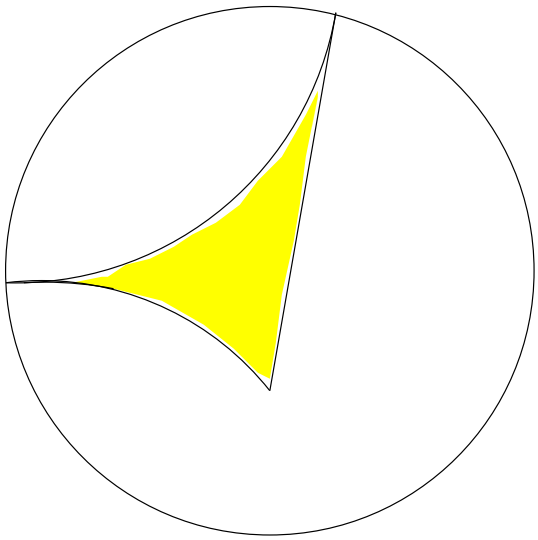
$$d_H(a,b) = d_D(\frac{a-i}{a+i}, \frac{b-i}{b+i}) = 2\tanh^{-1} \left| \frac{(b-i)(a+i) - (a-i)(b+i)}{\overline{(a+i)}(b+i) - \overline{(a-i)}(b-i)} \right| = 2\tanh^{-1} \left| \frac{b-a}{b-\bar{a}} \right|.$$

5.4 Hyperbolic triangles

A hyperbolic triangle Δ is given by three distinct points in H or D joined by geodesics. We see immediately from Gauss-Bonnet that the sum of the angles of a triangle is given by

$$A + B + C = \pi - \text{Area}(\Delta).$$

We can also consider hyperbolic triangles which have one or more vertices 'at infinity', i.e. on the boundary of H or D. These triangles are called *asymptotic*, *doubly (or bi-) asymptotic* and *triply (or tri-) asymptotic*, according to the number of vertices at infinity. The angle at a vertex at infinity is always 0, since all geodesics in H or D meet the boundary at right angles.



Theorem 5.1 *(*The cosine rule for hyperbolic triangles*) If Δ is a hyperbolic triangle in D with vertices at a, b, c and*

$$\alpha = d_D(b,c), \quad \beta = d_D(a,c) \text{ and } \gamma = d_D(a,b)$$

then

$$\cosh\gamma = \cosh\alpha \cosh\beta - \sinh\alpha \sinh\beta \cos\theta$$

where θ is the internal angle of Δ at c.

Proof: Because the group of isometries of D acts transitively on D we can assume that $c = 0$. Moreover, since the rotations $z \mapsto e^{i\phi}z$ are isometries which fix 0, we can also assume that a is real and positive. Then $\beta = 2\tanh^{-1}(a)$ so

$$a = \tanh(\beta/2)$$

and similarly

$$b = e^{i\theta}\tanh(\alpha/2)$$

while

$$\tanh(\gamma/2) = \left|\frac{b-a}{1-\bar{a}b}\right|.$$

Recall that

$$\frac{1+\tanh^2(\gamma/2)}{1-\tanh^2(\gamma/2)} = \cosh(\gamma)$$

so

$$\cosh(\gamma) = \frac{|1-\bar{a}b|^2+|b-a|^2}{|1-\bar{a}b|^2-|b-a|^2} = \frac{(1+|a|^2)(1+|b|^2)-2(\bar{a}b+a\bar{b})}{(1-|a|^2)(1-|b|^2)}.$$

Now

$$\frac{1+|a|^2}{1-|a|^2} = \frac{1+\tanh^2(\beta/2)}{1-\tanh^2(\beta/2)} = \cosh\beta$$

as above, and similarly

$$\frac{1+|b|^2}{1-|b|^2} = \cosh\alpha$$

while

$$\frac{2(\bar{a}b+a\bar{b})}{(1-|a|^2)(1-|b|^2)} = \frac{2\tanh(\alpha/2)\tanh(\beta/2)(e^{i\theta}+e^{-i\theta})}{\mathrm{sech}^2(\alpha/2)\mathrm{sech}^2(\beta/2)} = \sinh\alpha\,\sinh\beta\,\cos\theta.$$

This completes the proof. □

Theorem 5.2 *(*The sine rule for hyperbolic triangles) Let Δ be a hyperbolic triangle in D with internal angles A, B, C at vertices a, b, c and

$$\alpha = d_D(b,c), \quad \beta = d_D(a,c) \text{ and } \gamma = d_D(a,b).$$

Then

$$\frac{\sin A}{\sinh\alpha} = \frac{\sin B}{\sinh\beta} = \frac{\sin C}{\sinh\gamma}.$$

Proof: Two alternatives approaches:

1) Use the cosine rule to find an expression for $\sinh^2\alpha\,\sinh^2\beta\,\sin^2 C$ in terms of $\cosh\alpha$, $\cosh\beta$ and $\cosh\gamma$ which is symmetric in α, β and γ, and deduce that

$$\sinh^2\alpha\,\sinh^2\beta\,\sin^2 C = \sinh^2\alpha\,\sinh^2\gamma\,\sin^2 B = \sinh^2\gamma\,\sinh^2\beta\,\sin^2 A.$$

2) First prove that if $C = \pi/2$ then $\sin A\,\sinh\gamma = \sinh\alpha$ by applying the cosine rule to Δ in two different ways. Then deduce the result in general by dropping a perpendicular from one vertex of Δ to the opposite side. □

Gauss-Bonnet and its limits give the following:

Theorem 5.3 *(*Areas of hyperbolic triangles*)*

(i) The area of a triply asymptotic hyperbolic triangle Δ *is* π*.*

(ii) The area of a doubly asymptotic hyperbolic triangle Δ *with internal angle* θ *is* $\pi - \theta$*.*

(iii) The area of an asymptotic hyperbolic triangle Δ *with internal angles* θ *and* ϕ *is* $\pi - \theta - \phi$*.*

(iv) The area of a hyperbolic triangle Δ *with internal angles* θ*,* ϕ *and* ψ *is* $\pi - \theta - \phi - \psi$*.*

5.5 Non-Euclidean geometry

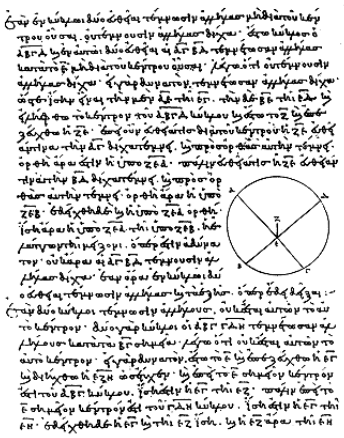
As we see above, the analogy between Euclidean geometry and its theorems and the geometry of the hyperbolic plane is very close, so long as we replace lines by geodesics, and Euclidean isometries (translations, rotations and reflections) by the isometries of H or D. In fact it played an important historical role.

For centuries, Euclid's deduction of geometrical theorems from self-evident common notions and postulates was thought not only to represent a model of the physical space in which we live, but also some absolute logical structure. One postulate caused some problems though – was it really self-evident? Did it follow from the other axioms? This is how Euclid phrased it:

"That if a straight line falling on two straight lines makes the interior angle on the same side less than two right angles, the two straight lines if produced indefinitely, meet on that side on which the angles are less than two right angles".

Some early commentators of Euclid's *Elements*, like Posidonius (1st Century BC), Geminus (1st Century BC), Ptolemy (2nd Century AD), Proclus (410 - 485) all felt that the parallel postulate was not sufficiently evident to accept without proof.

Here is a page from a medieval edition of Euclid dating from the year 888. It is handwritten in Greek. The manuscript, contained in the Bodleian Library, is one of the earliest surviving editions of Euclid.



The controversy went on and on with Greek and Islamic mathematicians puzzling over it. Johann Lambert (1728-1777) realized that if the parallel postulate did not hold then the angles of a triangle add up to less than 180°, and that the deficit was the area. He found this worrying in many ways, not least because it says that there is an absolute scale – no distinction between similar and congruent triangles. Finally Janos Bolyai (1802-1860) and Nikolai Lobachevsky (1793-1856) discovered non-Euclidean geometry simultaneously. It satisfies all of Euclid's axioms except the parallel postulate, and we shall see that it is the geometry of H or D that we have been studying.

Bolyai became interested in the theory of parallel lines under the influence of his father Farkas, who devoted considerable energy towards finding a proof of the parallel postulate without success. He even wrote to his son:

"I entreat you, leave the doctrine of parallel lines alone; you should fear it like a sensual passion; it will deprive you of health, leisure and peace – it will destroy all joy in your life."

Another relevant figure in the discovery was Carl Friedrich Gauss (1777-1855), who as we have seen developed the differential geometry of surfaces. He was the first to consider the possibility of a geometry denying the parallel postulate. However, for fear of being ridiculed he kept his work unpublished, or maybe he never made the connection with the curvature of real world surfaces and the Platonic ideal of axiomatic geometry. Anyway, when he read Janos Bolyai's work he wrote to Janos's father:

"If I commenced by saying that I must not praise this work you would certainly be surprised for a moment. But I cannot say otherwise. To praise it, would be to praise

myself. Indeed the whole contents of the work, the path taken by your son, the results to which he is led, coincide almost entirely with my meditations, which have occupied my mind partly for the last thirty or thirty-five years."

Euclid's axioms were made rigorous by Hilbert. They begin with undefined concepts of

- "point"
- "line"
- "lie on" (a point lies on a line)
- "betweenness"
- "congruence of pairs of points"
- "congruence of pairs of angles".

Euclidean geometry is then determined by logical deduction from the following axioms:

EUCLID'S AXIOMS

I. AXIOMS OF INCIDENCE

1. Two points have one and only one straight line in common.
2. Every straight line contains a least two points.
3. There are at least three points not lying on the same straight line.

II. AXIOMS OF ORDER

1. Of any three points on a straight line, one and only one lies between the other two.
2. If A and B are two points there is at least one point C such that B lies between A and C.
3. Any straight line intersecting a side of a triangle either passes through the opposite vertex or intersects a second side.

III. AXIOMS OF CONGRUENCE

1. On a straight line a given segment can be laid off on either side of a given point (the segment thus constructed is congruent to the give segment).

2. If two segments are congruent to a third segment, then they are congruent to each other.

3. If AB and $A'B'$ are two congruent segments and if the points C and C' lying on AB and $A'B'$ respectively are such that one of the segments into which AB is divided by C is congruent to one of the segments into which $A'B'$ is divided by C', then the other segment of AB is also congruent to the other segment of $A'B'$.

4. A given angle can be laid off in one and only one way on either side of a given half-line; (the angle thus drawn is congruent to the given angle).

5. If two sides of a triangle are equal respectively to two sides of another triangle, and if the included angles are equal, the triangles are congruent.

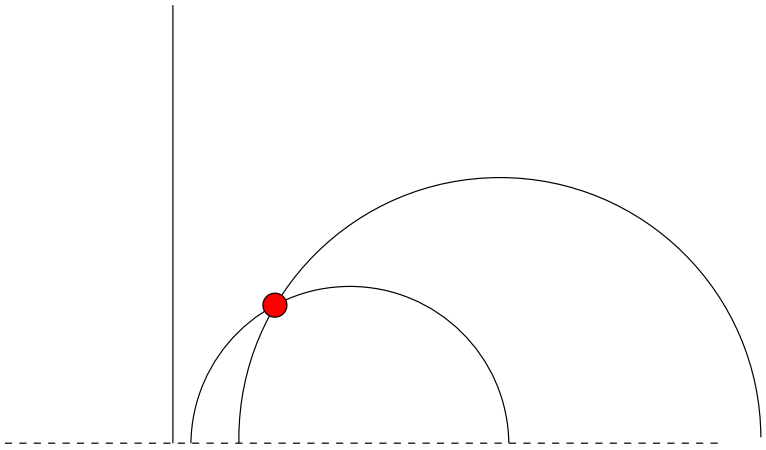
IV. AXIOM OF PARALLELS

Through any point not lying on a straight line there passes one and only one straight line that does not intersect the given line.

V. AXIOM OF CONTINUITY

1. If AB and CD are any two segments, then there exists on the line AB a number of points $A_1, \ldots, A_n$ such that the segments $AA_1, A_1A_2, \ldots, A_{n-1}A_n$ are congruent to CD and such that B lies between A and A_n

Clearly H does not satisfy the Axiom of Parallels:



The fact that hyperbolic geometry satisfies all the axioms except the parallel postulate is now only of historic significance and the reader is invited to do all the checking. Often one model is easier than another. Congruence should be defined through the action of the group of isometries.

5.6 Complex analysis and the hyperbolic plane

The intricate metric structure of the hyperbolic plane – geodesics, triangles and all – is actually determined purely by the holomorphic functions on it, so we could also think of hyperbolic geometry as a branch of complex analysis. Here is the theorem that makes it work:

Theorem 5.4 *Any holomorphic homeomorphism $f : D \to D$ is an isometry of the hyperbolic metric.*

Proof: The argument follows Schwartz's lemma. By applying an isometry we can assume that $f(0) = 0$, and since the image of f is D, we have $|f(z)| < 1$ if $z \in D$. Now since $f(0) = 0$, $f_1(z) = f(z)/z$ is holomorphic and applying the maximum principle to a disc of radius $r < 1$ we get

$$|f_1(z)| \leq \frac{1}{r}$$

and in the limit as $r \to 1$, $|f_1(z)| \leq 1$ or equivalently

$$|f(z)| \leq |z|.$$

Since f is a homeomorphism, its inverse satisfies the same inequality so

$$|z| \leq |f(z)|$$

and $|f_1(z)| = 1$ everywhere. Since this is true at an interior point the function must be a constant c so $f(z) = cz$ and since $|f(z)| = |z|$,

$$f(z) = e^{i\theta} z$$

which is an isometry. □

In fact a similar result holds for $\mathbf{C}$:

Theorem 5.5 *Any holomorphic homeomorphism $f : \mathbf{C} \to \mathbf{C}$ is of the form $f(z) = az + b$ with $a \neq 0$.*

If $|a| = 1$ this is an isometry of the Euclidean metric $dx^2 + dy^2$. The extra scaling $z \mapsto \lambda z$ is what gives rise in classical geometrical terms to similar but non-congruent triangles.

Proof: For $|z| > R$ consider the function $g(z) = f(1/z)$. Suppose g has an essential singularity at $z = 0$. Then the Casorati-Weierstrass theorem (Exercise 17.5 in [4]) tells us that $g(z)$ gets arbitrarily close to any complex number if z is small enough, and in particular to values in the image of $\{z : |z| \leq R\}$ under f. But we assumed f was bijective, which is a contradiction. It follows that g has at most a pole at infinity and so $f(z)$ must be a polynomial of some degree k.

However the equation $f(z) = c$ then has k solutions for most values of c, and again since f is bijective we must have $k = 1$ and

$$f(z) = az + b.$$

□

For completeness, we add the following

Theorem 5.6 *Any holomorphic homeomorphism f of the Riemann sphere to itself is a Möbius transformation $z \mapsto (az + b)/(cz + d)$.*

Proof: By using a Möbius transformation we can assume that $f(\infty) = \infty$ and then the previous theorem tells us that $f(z) = az + b$. □

These results are all about the complex plane and its subsets. In fact hyperbolic geometry has an important role to play in the study of compact Riemann surfaces. Recall that local holomorphic coordinates on a Riemann surface are related by holomorphic transformations and these preserve angles. Given two smooth curves on a Riemann surface, it makes good sense to define their angle of intersection and this is called a *conformal structure*. A metric also defines angles so we can consider metrics compatible with the conformal structure of a Riemann surface. In a local coordinate z such a metric is of the form

$$f dz d\bar{z} = f(dx^2 + dy^2).$$

The remarkable result is the following *uniformization* theorem:

Theorem 5.7 *Every closed Riemann surface X has a metric of constant Gaussian curvature compatible with its conformal structure.*

Note that by the Gauss-Bonnet theorem $K > 0$ implies $\chi(X) > 0$, i.e. X is a sphere, $K = 0$ implies $\chi(X) = 0$, i.e. X is a torus, and $K < 0$ gives $\chi(X) < 0$.

Proof: The proof is a corollary of a difficult theorem called the Riemann mapping theorem. Recall that a space is *simply-connected* if it is connected and every closed path can be shrunk to a point. The Riemann mapping theorem (proved by Poincaré and Koebe) says that every simply-connected Riemann surface is holomorphically homeomorphic to either the Riemann sphere, $\mathbf{C}$ or H.

If X is any reasonable topological space, one can form its *universal covering space* $\tilde{X}$ (see [3]) which is simply connected and has

- a projection $p : \tilde{X} \to X$
- every point $x \in X$ has a neighbourhood V such that $p^{-1}(V)$ consists of a disjoint union of open sets each of which is homeomorphic to V by p
- there is a group π of homeomorphisms of $\tilde{X}$ such that $p(gy) = p(y)$, so that π permutes the different sheets in $p^{-1}(V)$.
- no element of π apart from the identity has a fixed point
- X can be identified with the space of orbits of π acting on $\tilde{X}$.

The standard example of this is $X = S^1$, $\tilde{X} = \mathbf{R}$, $p(t) = e^{it}$ and $\pi = \mathbf{Z}$ acting by $t \mapsto t + 2n\pi$. It is easy to see that the universal covering of a Riemann surface is a Riemann surface, so applying the Riemann mapping theorem we see that $\tilde{X}$ is either the Riemann sphere, $\mathbf{C}$ or H.

So consider the cases:

- If $\tilde{X}$ is the sphere S, it is compact and so $p : \tilde{X} \mapsto X$ has only a finite number k of sheets. By counting vertices, edges and faces it is clear that $\chi(\tilde{X}) = k\chi(X)$. Since $\chi(S) = 2$, we must have $k = 1$ or 2, but if the latter $\chi(X) = 1$ which is not of the allowable form $2-2g$ for an orientable surface and a Riemann surface *is* orientable. So it is only the Riemann sphere in this case.
- If $\tilde{X} = \mathbf{C}$, we appeal to Theorem 5.5. The group π of covering transformations is holomorphic and so each element is of the form $z \mapsto az + b$. But π has no fixed points, so $az + b = z$ has no solution which means that $a = 1$. the transformations $z \mapsto z + b$ are just translations and are isometries of the metric $dx^2 + dy^2$ which has $K = 0$.
- If $\tilde{X} = H$, then from Theorem 5.4, the action of π preserves the hyperbolic metric.

□

So we see that these abstract metrics have a role to play in the study of Riemann surfaces – a long long way from surfaces in $\mathbf{R}^3$.

6 APPENDIX: Technical results

6.1 A: The inverse function theorem

Lemma 6.1 *(Contraction mapping principle) Let M be a complete metric space and suppose $T : M \to M$ is a map such that*

$$d(Tx, Ty) \leq kd(x, y)$$

where $k < 1$. Then T has a unique fixed point.

Proof: Choose any point x_0, then

$$\begin{aligned} d(T^m x_0, T^n x_0) &\leq k^m d(x_0, T^{n-m} x_0) \qquad \text{for} \quad n \geq m \\ &\leq k^m (d(x_0, Tx_0) + d(Tx_0, T^2 x_0) + \ldots + d(T^{n-m-1} x_0, T^{n-m} x_0)) \\ &\leq k^m (1 + k + \ldots + k^{n-m-1}) d(x_0, Tx_0) \\ &\leq \frac{k^m}{1-k} d(x_0, Tx_0) \end{aligned}$$

This is a Cauchy sequence, so completeness of M implies that it converges to x. Thus $x = \lim T^n x_0$ and so by continuity of T,

$$Tx = \lim T^{n+1} x_0 = x$$

For uniqueness, if $Tx = x$ and $Ty = y$, then

$$d(x, y) = d(Tx, Ty) \leq kd(x, y)$$

and so $k < 1$ implies $d(x, y) = 0$. □

Theorem 6.2 *(Inverse function theorem) Let $U \subseteq \mathbf{R}^n$ be an open set and $f : U \to \mathbf{R}^n$ a C^∞ function such that Df_a is invertible at $a \in U$. Then there exist neighbourhoods V, W of a and $f(a)$ respectively such that $f(V) = W$ and f has a C^∞ inverse on W.*

Proof: By an affine transformation $x \mapsto Ax + b$ we can assume that $a = 0$ and $Df_a = I$. Now consider $g(x) = x - f(x)$. By construction $Dg_0 = 0$ so by continuity there exists $r > 0$ such that if $\|x\| < 2r$,

$$\|Dg_x\| < \frac{1}{2}$$

It follows from the mean value theorem that

$$\|g(x)\| \leq \frac{1}{2}\|x\|$$

and so g maps the closed ball $\bar{B}(0,r)$ to $\bar{B}(0,r/2)$. Now consider

$$g_y(x) = y + x - f(x)$$

(The choice of g_y is made so that a fixed point $g_y(x) = x$ solves $f(x) = y$).

If now $\|y\| \leq r/2$ and $\|x\| \leq r$, then

$$\|g_y(x)\| \leq \frac{1}{2}r + \|g(x)\| \leq \frac{1}{2}r + \frac{1}{2}r = r$$

so g_y maps the complete metric space $M = \bar{B}(0,r)$ to itself. Moreover

$$\|g_y(x_1) - g_y(x_2)\| = \|g(x_1) - g(x_2)\| \leq \frac{1}{2}\|x_1 - x_2\|$$

if $x_1, x_2 \in \bar{B}(0,r)$, and so g_y is a contraction mapping. Applying Lemma 1 we have a unique fixed point and hence an inverse $\varphi = f^{-1}$.

We need to show first that φ is continuous and secondly that it has derivatives of all orders. From the definition of g and the mean value theorem,

$$\begin{aligned} \|x_1 - x_2\| &\leq \|f(x_1) - f(x_2)\| + \|g(x_1) - g(x_2)\| \\ &\leq \|f(x_1) - f(x_2)\| + \frac{1}{2}\|x_1 - x_2\| \end{aligned}$$

so

$$\|x_1 - x_2\| \leq 2\|f(x_1) - f(x_2)\|$$

which is *continuity* for φ. It follows also from this inequality that if $y_1 = f(x_1)$ and $y_2 = f(x_2)$ where $y_1, y_2 \in B(0,r/2)$ then $x_1, x_2 \in \bar{B}(0,r)$, and so

$$\begin{aligned} \|\varphi(y_1) - \varphi(y_2) - (Df_{x_2})^{-1}(y_1 - y_2)\| &= \|x_1 - x_2 - (Df_{x_2})^{-1}(f(x_1) - f(x_2))\| \\ &\leq \|(Df_{x_2})^{-1}\|\|Df_{x_2}(x_1 - x_2) - f(x_1) + f(x_2)\| \\ &\leq A\|x_1 - x_2\|R \end{aligned}$$

where A is a bound on $\|(Df_{x_2})^{-1}\|$ and the function $\|x_1 - x_2\|R$ is the remainder term in the definition of differentiability of f. But $\|x_1 - x_2\| \leq 2\|y_1 - y_2\|$ so as $y_1 \to y_2$, $x_1 \to x_2$ and hence $R \to 0$, so φ is differentiable and moreover its derivative is $(Df)^{-1}$.

Now we know the derivative of φ:

$$D\varphi = (Df)^{-1}$$

so we see that it is continuous and has as many derivatives as f itself, so φ is C^∞. □

6.2 B: Existence of solutions of ordinary differential equations

Lemma 6.3 *Let M be a complete metric space and $T : M \to M$ a map. If T^n is a contraction mapping, then T has a unique fixed point.*

Proof: By the contraction mapping principle, T^n has a unique fixed point x. We also have

$$T^n(Tx) = T^{n+1}x = T(T^n x) = Tx$$

so Tx is also a fixed point of T^n. By uniqueness $Tx = x$. □

Theorem 6.4 *Let $f(t, x)$ be a continuous function on $|t - t_0| \leq a, \|x - x_0\| \leq b$ and suppose f satisfies a Lipschitz condition*

$$\|f(t, x_1) - f(t, x_2)\| \leq \|x_1 - x_2\|.$$

If $M = \sup |f(t, x)|$ and $h = \min(a, b/M)$, then the differential equation

$$\frac{dx}{dt} = f(t, x), \qquad x(t_0) = x_0$$

has a unique solution for $|t - t_0| \leq h$.

Proof: Let

$$(Tx)(t) = x_0 + \int_{t_0}^{t} f(s, x(s))ds$$

Then Tx is differentiable since f and x are continuous and if $Tx = x$, x satisfies the differential equation (differentiate the definition). We use the metric space

$$X = \{x \in C([t_0 - h, t_0 + h], \mathbf{R}^n) : \|x(t) - x_0\| \leq Mh\}$$

with the uniform metric

$$d(x_1, x_2) = \sup_{|t-t_0|\leq h} \|x_1(t) - x_2(t)\|$$

which makes it complete. If $x \in M$, then $Tx \in M$ and we claim

$$|T^k x_1(t) - T^k x_2(t)\| \leq \frac{c^k}{k!}|t - t_0|^k d(x_1, x_2)$$

For $k = 0$ this is clear, and in general we use induction to establish:

$$\begin{aligned}
\|T^k x_1(t) - T^k x_2(t)\| &\leq \int_{t_0}^{t} \|f(s, T^{k-1}x_1(s) - f(s, T^{k-1}x_2(s)\| ds \\
&\leq c \int_{t_0}^{t} \|T^{k-1}x_1(s) - T^{k-1}x_2(s)\| ds \\
&\leq (c^k/(k-1)!) \int_{t_0}^{t} |s - t_0|^{k-1} ds \; d(x_1, x_2) \\
&\leq (c^k/k!)|t - t_0|^k d(x_1, x_2)
\end{aligned}$$

So T^n is a contraction mapping for large enough N, and the result follows. □

Theorem 6.5 *The solution above depends continuously on the initial data x_0.*

Proof: Take $h_1 \leq h$ and $\delta > 0$ such that $Mh + \delta \leq b$, and let

$$Y = \{y \in C([t_0 - h_1, t_0 + h_1] \times \bar{B}(x_0, \delta); \mathbf{R}^n : \|y(t, x) - x\| \leq Mh, y(t_0, x) = x\}$$

which is a complete metric space as before. Now set

$$(Ty)(t, x) = x + \int_{t_0}^{t} f(s, y(s, x)) ds$$

Since $Mh_1 + \delta \leq b$, T maps Y to Y and just as before T^n is a contraction mapping with a unique fixed point which satisfies

$$\frac{\partial y}{\partial t} = f(t, y), \qquad y(t_0, x) = x$$

Since y is continuous in t and x this is what we need. □

If $f(t, x)$ is smooth then we need more work to prove that the solution to the equation is smooth and smoothly dependent on parameters.

6.3 B': Smooth dependence

Lemma 6.6 *Let $A(t, x), B(t, x)$ be continuous matrix-valued functions and take $M \geq \sup_{t,x} \|B\|$. The solutions of the linear differential equations*

$$\begin{aligned}
\frac{d\xi(t, x)}{dt} &= A(t, x)\xi(t, x), \qquad \xi(t_0, x) = a(x) \\
\frac{d\eta(t, x)}{dt} &= B(t, x)\eta(t, x), \qquad \eta(t_0, x) = b(x)
\end{aligned}$$

satisfy

$$\sup_x \|\xi(t,x) - \eta(t,x)\| \leq C\|A-B\|\frac{e^{M|t-t_0|}-1}{M} + \|a-b\|e^{M|t-t_0|}$$

where C is a constant depending only on A and a.

Proof: By the existence theorem we know how to find solutions as limits of ξ_n, η_n where

$$\begin{aligned} \xi_k &= a + \int_{t_0}^t A\xi_{k-1} ds \\ \eta_k &= b + \int_{t_0}^t B\eta_{k-1} ds \end{aligned}$$

Let $g_k(t) = \sup_x \|\xi_k(t,x) - \eta_k(t,x)\|$ and $C = \sup_{k,x,t} \|\xi_k\|$. Then

$$g_n(t) \leq \|a-b\| + C\|A-B\||t-t_0| + M\int_{t_0}^t g_{n-1}(s)ds$$

Now define f_n by $f_0(t) = \|a-b\|$ and then inductively by

$$f_n(t) = \|a-b\| + C\|A-B\||t-t_0| + M\int_{t_0}^t f_{n-1}(s)ds$$

Comparing these two we see that $f_n \geq g_n$. This is a contraction mapping, so that $f_n \to f$ with

$$f(t) = \|a-b\| + C\|A-B\||t-t_0| + M\int_{t_0}^t f(s)ds$$

and solving the corresponding differential equation we get

$$f(t) = \|a-b\|e^{M|t-t_0|} + C\|A-B\|\frac{e^{M|t-t_0|}-1}{M}$$

As $g_n(t) \leq f_n(t)$,

$$\sup_x \|\xi_n(t,x) - \eta_n(t,x)\| \leq f_n(t)$$

and the theorem follows by letting $n \to \infty$. □

Theorem 6.7 *If f is C^k and*

$$\frac{d}{dt}\alpha(t,x) = f(t,\alpha(t,x)), \qquad \alpha(0,x) = x$$

then α is also C^k.

Proof: The hardest bit is $k = 1$. Assume f is C^1 so that $\partial f/\partial t$ and $\partial f/\partial x_i$ exist and are continuous. We must show that α is C^1 in all variables. If that were true, then the matrix valued function λ where $(\lambda_i = \partial\alpha/\partial x_i)$ would be the solution of the differential equation

$$\frac{d\lambda}{dt} = D_x f(t, \alpha)\lambda \tag{14}$$

so we shall solve this equation by the existence theorem and prove that the solution is the derivative of α. Let $F(s) = f(t, a + s(b - a))$. Then

$$\frac{dF}{ds} = D_x f(t, a + s(b - a))(b - a)$$

so

$$f(t, b) - f(t, a) = \int_0^1 D_x f(t, a + s(b - a))(b - a)ds$$

But then

$$\begin{aligned}\frac{d}{dt}(\alpha(t, x + y) - \alpha(t, x)) &= f(t, \alpha(t, x + y)) - f(t, \alpha(t, x)) \\ &= \int_0^1 D_x f(t, \alpha(t, x) + s(\alpha(t, x + y) - \alpha(t, x)))(\alpha(t, x + y) - \alpha(t, x))ds\end{aligned}$$

Let $A(t, x) = D_x f(t, \alpha(t, x))$ and $\xi(t, x) = \lambda(t, x)y$ and

$$B_y(t, x) = \int_0^1 D_x f(t, \alpha(t, x)+s(\alpha(t, x+y)-\alpha(t, x)))ds, \quad \eta_y(t, x) = \alpha(t, x+y)-\alpha(t, x)$$

Apply the previous lemma and we get

$$\sup_{|t|\leq\epsilon} \|\lambda(t, x)y - (\alpha(t, x + y) - \alpha(x))\| = o(\|y|)$$

and so $D_x\alpha = \lambda$, which is continuous in (t, x). Since also $d\alpha/dt = f(t, \alpha)$ this means that α is C^1 in all variables.

To continue, suppose inductively that the theorem is true for $k - 1$, and f is C^k. Then $A(t, x) = D_x f(t, \alpha(t, x))$ is C^{k-1} but since

$$\frac{d\lambda}{dt} = A\lambda$$

we have λ is C^{k-1}. Now $D_x\alpha = \lambda$ so the x_i-derivatives of α are C^{k-1}. But also $d\alpha/dt = f(t, \alpha)$ is C^{k-1} too, so α is C^k. □

References

[1] P A Firby and C F Gardiner, "Surface topology". Second edition. *Ellis Horwood Series: Mathematics and its Applications. Ellis Horwood, New York* distributed by Prentice Hall, Inc., Englewood Cliffs, NJ, 1991, ISBN 0-13-855321-1

[2] G K Francis and J R Weeks, Conway's ZIP proof. *Amer. Math. Monthly* **106** (1999), 393–399.

[3] W S Massey, "A basic course in algebraic topology." *Graduate Texts in Mathematics*, **127**. Springer-Verlag, New York, 1991. ISBN 0-387-97430-X

[4] H A Priestley, "Introduction to complex analysis". Revised second edition. Oxford University Press, Oxford, 2003. ISBN 0-19-852562-1