Geometry and Topology
SHP Spring '18

Henry Liu

Last updated: April 28, 2018

Contents

0 About Mathematics

The majority of this course aims to introduce all of you to the many wonderfully interesting ideas underlying the study of geometry and topology in modern mathematics. But along the way, I want to also give you a sense of what "being a mathematician" and "doing mathematics" is really about.

There are three main qualities that distinguish professional mathematicians:

1. the ability to correctly understand, make, and prove precise, unambiguous logical statements (often using *a lot* of jargon);
2. the ability to ask meaningful questions that lead to interesting problems, tell how difficult these problems will be, and generally identify the easiest and most efficient ways to try to solve them;
3. a deep understanding and knowledge of the area in which they work.

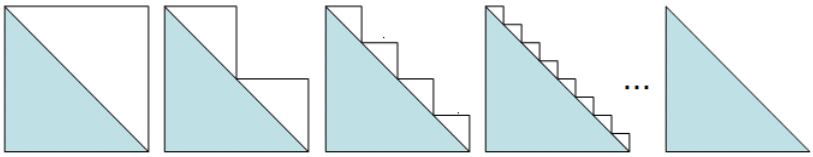
As we go through the content of the course, I'll point out every now and then how mathematicians apply these abilities to effectively solve problems. But right now, there is one immediate point I want to address.

Here is one of the problems that we will be able to answer later on in the course:

Can you fold a piece of paper over a sphere without crumpling it?

Some of you are probably thinking "well duh, obviously no," possibly followed by a fairly reasonable-sounding explanation. But how do you know your explanation is actually correct? There are many examples of **intuitive and reasonable explanations that produce completely false conclusions**. Here is one.

Example 0.1. Claim: $2 = \sqrt{2}$. Take a square of side length 1 and repeat the following process to approximate the length of the diagonal:



In each of the steps, the length of the jagged approximation is exactly 2. Since the approximation gets closer and closer to the actual diagonal, the length of the actual diagonal must also be 2. But by the Pythagorean theorem, the diagonal has length $\sqrt{2}$. So $\sqrt{2} = 2$.

What went wrong? It turns out we made the incorrect assumption that if the jagged line "approximates" the diagonal, its length also approximates the diagonal's length. Mathematicians use jargon to express ideas precisely enough that they can catch incorrect assumptions like that. If you asked a mathematician to accurately state exactly what went wrong in the example, he/she would probably say something like

the arclength function isn't continuous on the space of piecewise-linear paths with the L^1 norm.

This mathematical language might seem incomprehensible to you right now, but there is often a very simple and intuitive idea being expressed behind the jargon. It is just that in order to express it very precisely to avoid misunderstanding and ambiguity, mathematicians must resort to technical terms that have already been very carefully defined. (For example, you can already start nitpicking at my not-very-precise use of language: what does it mean for the jagged line to "approximate" the diagonal?)

I will introduce a minimal amount of jargon in this course. We only have so much time, and I want to focus on conveying the big ideas. That said, if at any point an argument I am presenting does not make sense to you, or does not sound believable, or you think you found a counterexample, ask! It is likely that either the concepts I used are too vague, or that you have actually discovered a technicality that I have tried to hide!

In these notes I will write little sidenotes like this one sometimes to make small comments on terminology and notation.

Let's get started.

1 Surfaces

Geometry and topology is the study of spaces. Spaces generally have a dimension. We live in a three-dimensional space, but it turns out three-dimensional spaces are much, much harder to study than two-dimensional spaces.

Definition 1.1. A **surface** is a space that, if you were to zoom in on any point, looks like a bent piece of the two-dimensional plane.

Formal definition. A **surface** is a space that is *locally homeomorphic* to $\mathbb{R}^2$.

We will actually see the idea of "being homeomorphic" in more detail later.

1.1 Examples of surfaces

The first thing mathematicians (should) do when given a definition is to find as many examples of it as possible! Examples help us develop intuition for what the definition is really saying. It is also important to find examples of objects that do **not** satisfy the definition, so that we develop intuition for what the definition is **not** saying.

Example 1.2. The simplest example of a surface is the two-dimensional plane itself, which we call $\mathbb{R}^2$. If we zoom in at any point, it clearly still looks like a two-dimensional plane.

Example 1.3. The **sphere** S^2 and the **torus** T^2 are also surfaces:

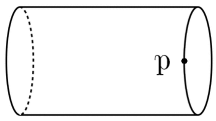
Here S^2 does not mean "S squared." It is just a label we use for the sphere. The 2 means "two-dimensional." The same goes for T^2.

Importantly, these are **hollow**; we only care about their surfaces, not what's inside.

If we zoom in on any point on the sphere, or if we lived on a very big sphere, the region around that point looks very much like a piece of the two-dimensional plane. In fact, the resemblance is so close that for a very long time, we thought we lived on a plane, not a sphere!

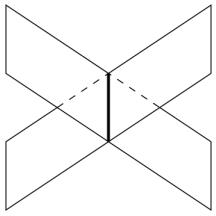
Example 1.4 (Non-example)**.** The **cylinder** is **not** a surface:

We usually say the cylinder is a "surface with boundary." Somewhat confusingly, surfaces with boundary are not surfaces.



If we zoom in at the point p, we get something that is almost a piece of $\mathbb{R}^2$, but it has a boundary and we cannot move past it. So it is not really a piece of $\mathbb{R}^2$, where we can move in any of the two directions we want.

Example 1.5 (Non-example)**.** The following space (which doesn't really have a name) is not a surface:



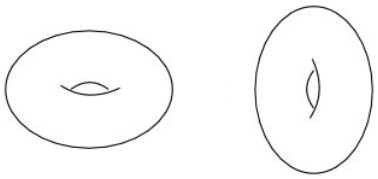
If we zoom in at the intersection point, it will always look like two planes intersecting, not a piece of $\mathbb{R}^2$.

1.2 Equivalence of surfaces

One major question that mathematicians always ask whenever they define a new type of object is:

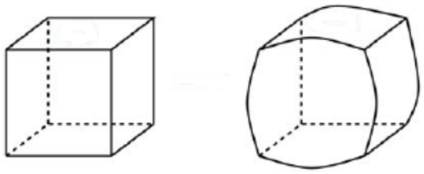
> Can we **classify** all the different objects of this type?

But what do we mean by *different*? In other words, when do we consider two surfaces to be **equivalent**? For example, clearly if we take a surface and rotate it, we should consider the result to be the same surface:



Definition 1.6 (Provisional)**.** Two surfaces are equivalent if we can get one from the other by rotating the surface.

The problem with this definition is that there are *too many* different surfaces and it becomes very hard to classify them! For example, we would have to describe exactly how the second cube bulges out on each side:



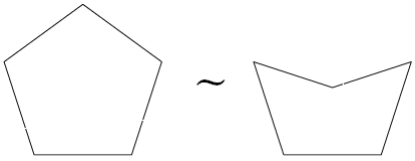
This is not easy to describe; we would have to specify what the curve of each side looks like. And that's just for a cube. There are many other shapes out there.

Example 1.7. To gain some insight into how we should really be defining equivalence, let's think about **polygons**.

> One easy way to classify them is by the *number of sides*, i.e. to say two polygons are **equivalent** if they have the same number of sides.

That way we can *stretch* and *deform* a polygon however we like, and the result will be equivalent to the original polygon:

We have to keep the sides straight though, so that we end up with a polygon!



We use the symbol $\sim$ to denote equivalence, i.e. we write $A \sim B$ to mean that "A is equivalent to B," for whatever definition of equivalence we are using.

We call $\sim$ an "equivalence relation."

Note that if we instead said two polygons are equivalent if one can be obtained from the other by rotation, classifying polygons becomes much harder: we would have to specify each angle and each side length. We use this insight to give a better definition of equivalence for surfaces.

Definition 1.8. Two surfaces are **equivalent**, or **homeomorphic**, if we can get one from the other by stretching and squeezing and deforming the surface as if it were made from a sheet of rubber.

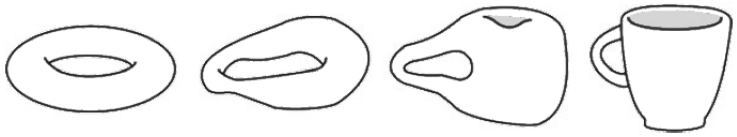
Formal definition. Two surfaces are **homeomorphic** if there is a *continuous function* from one to the other which has a continuous inverse.

Being "continuous" is a formalization of the idea that we can deform the surface, but cannot cut it or break it.

The notion of homeomorphism is actually not specific to two dimensions. We will see later on that it applies to spaces of any dimension. The field of **topology** is the study of spaces under this equivalence.

Example 1.9. An interesting consequence of our definition is that now a sphere and a cube are actually homeomorphic! The following diagram shows how:

Example 1.10. There is a famous mathematical saying that "topologists can't tell the difference between a donut and a coffee cup." This is because those two objects are homeomorphic too!



1.3 Distinguishing between surfaces

Homeomorphism is tricky to work with sometimes. How can we tell that two surfaces are **not** homeomorphic? For example, can you tell if the sphere S^2 and the torus T^2 are homeomorphic or not?

Ask a mathematician this question, and he/she will respond that

> In general, to distinguish between objects, define an **invariant** of your objects that assigns the same quantity to equivalent objects.

We will define an invariant called the **Euler characteristic** that assigns an integer to each surface, such that if two surfaces are homeomorphic, they have the same Euler characteristic. That way, if we compute the Euler characteristic for the sphere and the torus and get different integers, we know for sure they are not homeomorphic.

We say the Euler characteristic is "invariant under homeomorphism."

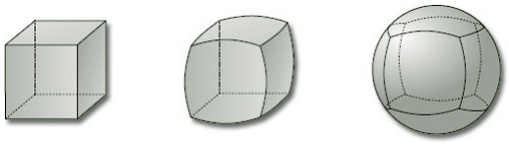
Definition 1.11. Here are the steps for calculating the Euler characteristic of a surface.

1. Deform the surface until it is a polyhedron, i.e. until it is made up of polygons.
2. Count the number of vertices V, edges E, and faces F on the polyhedron.
3. Calculate $V - E + F$, which is the **Euler characteristic**.

If X is the surface, its Euler characteristic is denoted by $\chi(X)$.

The symbol χ is the Greek letter chi, pronounced "kai."

Example 1.12. To compute the Euler characteristic $\chi(S^2)$ of the sphere, we first deform it into a cube:

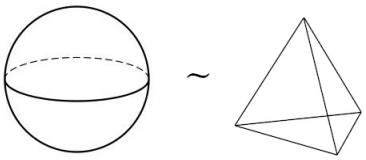


The cube has 8 vertices, 12 edges, and 6 faces, so

$$\chi(S^2) = 8 - 12 + 6 = 2.$$

Wait! What if we deformed the sphere into *some other polyhedron*? Since I said the Euler characteristic is the same for homeomorphic surfaces, it better be that calculating the Euler characteristic using this other deformation gives the same number. But that is not obvious at all.

Example 1.13. Let's compute the Euler characteristic $\chi(S^2)$ of the sphere by deforming it into a tetrahedron instead:



The tetrahedron has 4 vertices, 6 edges, and 4 faces, so

$$\chi(S^2) = 4 - 6 + 4 = 2.$$

Theorem 1.14. *If two surfaces are homeomorphic, they have the same Euler characteristic.*

A proof of this theorem is actually not hard! But it is long. As we go through a sketch of the proof, it is best to have an example surface in your head to which you can apply the steps in the proof. For example, I usually think of the sphere and the cube, which are two homeomorphic surfaces.

Proof. Let's sketch the general reason why this is true. Call the two surfaces X and Y. One key idea is that

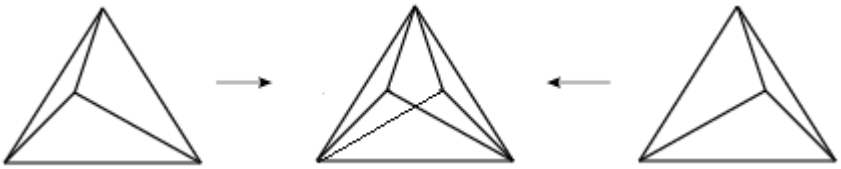
> If we want to compute the Euler characteristic of X, we don't actually need to do the deformation in step 1 of the process to computing the Euler characteristic.

Instead, we can just draw out where the vertices and edges are on X (e.g. the diagram above for the cube draws the cube on the sphere). We can do the same for Y. But since X and Y are homeomorphic, we can deform Y until it looks

identical to X. Now we have two copies of X, but they may have different drawings of vertices and edges on them.

So it suffices to show that no matter how we draw vertices and edges on X, we get the same Euler characteristic in the end. This involves the second key idea:

> Given two ways of drawing vertices and edges on X, we can "overlap" them to get another way of drawing vertices and edges, called the **refinement**.



The fact that Euler characteristic is unchanged by (or "invariant under") refinement is an important fact we will use later.

Why does this help? Well, you can check that if we take any two polygons sharing an edge and just remove that edge, the Euler characteristic does not change. Similarly, if we take a polygon and divide it into two adjacent polygons sharing an edge, the Euler characteristic does not change. In other words, *refinement does not change Euler characteristic*! So it must be that the original two different ways of drawing vertices and edges had the same Euler characteristic to start off with. □

1.4 Planar Models

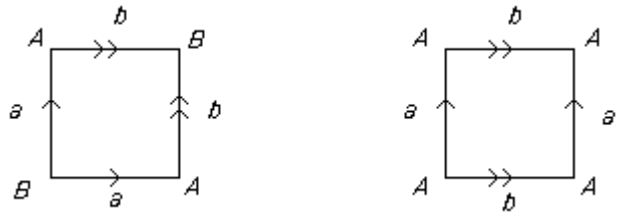
Now that we know we can compute Euler characteristic of a surface by deforming the surface however we want, let's compute $\chi(T^2)$, the Euler characteristic of the torus. This sounds simple now: just deform the torus into a polyhedron and count! But it's hard to keep track of how many vertices and edges and faces there are, and for more complicated surfaces we would have to work really hard to compute the Euler characteristic this way.

It turns out there is a better way, using **planar models**. Planar models will also help us study surfaces more easily in general.

Definition 1.15. A **planar model** of a surface is a polygon whose vertices and edges are **identified**, or "glued together," in some specified way. To specify that two edges are "glued together," we label them with the same type of arrow.

The best way to understand a definition is through examples, so let's look at some planar models for the surfaces we've seen already. Then we'll see some new planar models corresponding to surfaces we haven't seen yet.

Example 1.16. Here are the planar models for the sphere and the torus, respectively:



So now instead of imagining the surfaces sitting inside three-dimensional space, we can work with the planar models instead.

Example 1.17. Let's use the planar model of the torus to compute its Euler characteristic.

1. The square has 4 vertices, but in the planar model of the torus, every vertex is actually the same vertex, labeled A. So on the torus, there is 1 vertex.

   We say the four vertices are "identified" to the same vertex.

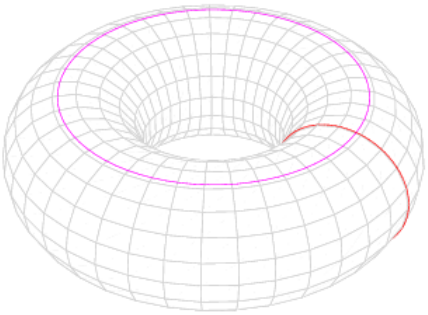
2. The square has 4 edges, but in the planar model, edges a and a are glued together and become the same edge on the torus, and the same for edges b and b. So on the torus, there are 2 edges.

3. There is still only one face; that doesn't change.

Hence $\chi(T^2) = 1 - 2 + 1 = 0$.

Warning: if you try to do a similar calculation for the Euler characteristic of the sphere, note that there are actually *two faces*: draw the planar model on the sphere and look! Thankfully, the sphere is the only surface we will see whose planar model has this problem.

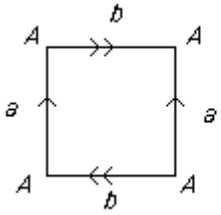
Some of you may be confused at this point, because we haven't actually deformed the torus into a polygon of any sort. The planar model, when we draw it out on the torus, corresponds to the red and pink lines below:

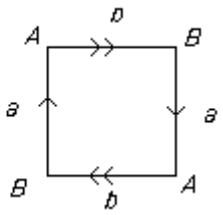


But that's alright, because just as in the image above, we can perform a **refinement** by adding in new gray edges and vertices (and consequently, new faces). Now of course this deforms into some sort of polyhedron. Because Euler characteristic is unchanged by refinement, we know this procedure doesn't affect our Euler characteristic calculation: $\chi(T^2)$ is still 0.

We defined a refinement earlier, in the proof of Theorem 1.14.

Example 1.18. We can draw arrows in different directions on the square to get new planar models:





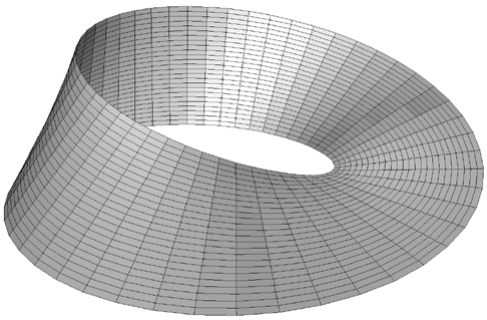
Neither of these seem to be the sphere or the torus, but we don't seem to be able to visualize them easily either. In fact they are new surfaces: the **Klein bottle** Kl^2, and the **projective plane** $\mathbb{P}^2$.

We can check that the Klein bottle and projective plane are surfaces: around any point, we can draw a little circle which looks like a bent piece of $\mathbb{R}^2$. In fact, any planar model with no boundary will be a surface. The problem with the Klein bottle and the projective plane is that they don't "live inside" three-dimensional space; they are actually objects in four-dimensional space!

Later we'll see the deeper reason is Kl^2 and $\mathbb{P}^2$ are both "non-orientable."

Here it is best to draw an analogy: just like a two-dimensional being has problems visualizing how the **Möbius strip** below does not intersect itself, we have problems visualizing how the Klein bottle and projective plane don't intersect themselves.

The Möbius strip is actually a surface with boundary.

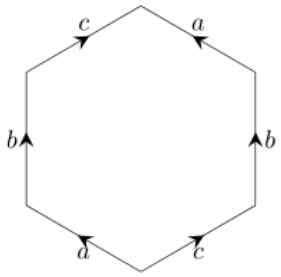


Exercise. Compute the Euler characteristics $\chi(\mathrm{Kl}^2) = 0$ and $\chi(\mathbb{P}^2) = 1$ using the planar models.

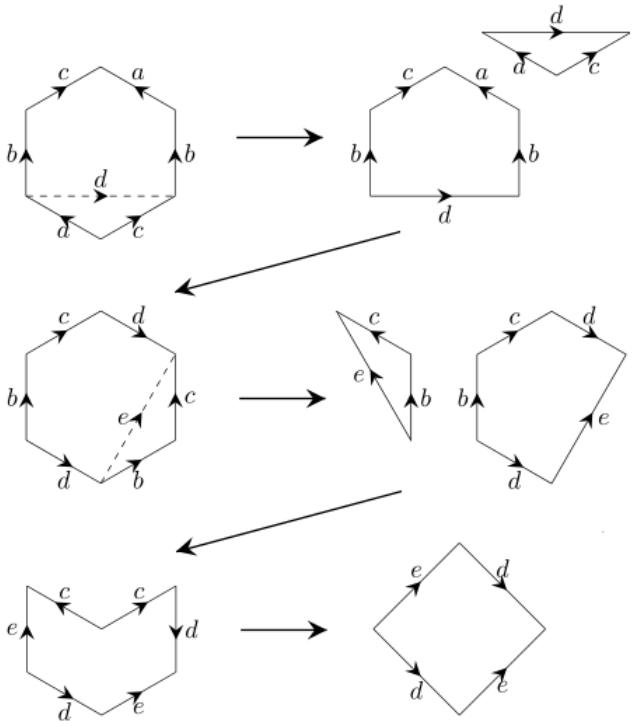
1.5 Cutting and Pasting

We have a problem: the Euler characteristic can't tell apart the Klein bottle and the torus, which both have Euler characteristic 0. Maybe they are actually the same surface? The Euler characteristic allows us to tell apart different surfaces; now we will develop a tool that allows us to show two surfaces are equivalent (aside from just "seeing" the deformation).

Example 1.19. Let's look at the following planar model (it is obvious which vertices are identified just by looking at the edge identifications, so from now on I'll start omitting vertex labels):



It has Euler characteristic 0. We'll show it is actually the torus by **cutting and pasting** edges, as follows.



Exercise. The sequence of steps above is not the fastest way to get a torus. Can you find a shorter way?

Cutting and pasting techniques can also construct new surfaces! The idea is to somehow glue two surfaces together in a way that produces a new surface.

Definition 1.20. Given two surfaces X and Y, their **connected sum** $X\#Y$ is the surface produced by the following steps:

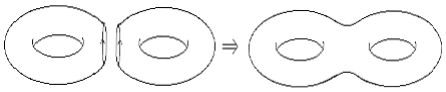
1. remove a small circle from anywhere on X, and another small circle from anywhere on Y;

Think of the # as "addition" for surfaces. The notation $X\#Y$ is read "X connected sum with Y."

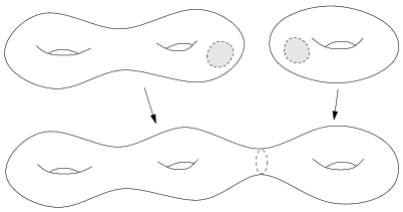
2. take the boundaries where the small circles used to be and glue them together.

Technical point: the small circles must be homeomorphic to disks. E.g. cutting out a strip around the entirety of the torus is not allowed!

Example 1.21. The connected sum $T^2\#T^2$ of two torii is a **two-holed torus**:



The connected sum $T^2\#T^2\#T^2$ of three torii is a **three-holed torus**:



Of course, we can produce **n-holed torii** for all positive integers n.

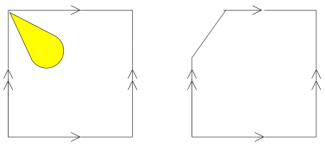
The notation $(T^2)^{\#n}$ is sometimes used for the n-holed torus: it stands for $T^2\#\cdots\#T^2$, i.e. T^2 connected sum with itself n times. The idea is that the n is an exponent, and the # tells us which operation is being repeated n times.

We need to make sure the n-holed torii are all inequivalent surfaces, though. We'd like to compute their Euler characteristics using planar models, but we don't know yet what the planar models for connected sums look like. The key idea is that

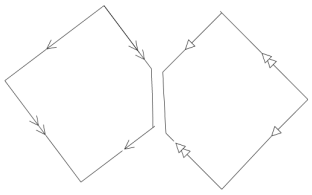
> We can do the connected sum construction *directly on the planar models.*

The steps are as follows (illustrated with $T^2\#T^2$):

1. Remove a little circle by choosing any vertex, drawing a loop around it, and "opening up" the loop;



2. Glue the resulting two planar models together along the newly created edges.



So the planar model for $T^2\#T^2$ is an octagon. Now we can compute its Euler characteristic: $\chi(T^2\#T^2) = -2$. In general, we have the following theorem to help us compute Euler characteristics of connected sums.

Theorem 1.22. *If X and Y are two surfaces, then $\chi(X\#Y) = \chi(X)+\chi(Y)-2$.*

Proof. We just have to keep track of how many vertices, edges, and faces are added or removed in the process of constructing the connected sum.

1. We added two extra edges (one on X and one on Y), which are the boundaries of the small circles we remove. Note that we **did not** add extra vertices: the "two" vertices at the ends of the added edge are actually the same vertex!

2. We glued these two new edges (and therefore the vertices at their endpoints) together, i.e. removed two vertices and two edges. By doing this gluing, we also merged the face of X with the face of Y, i.e. we removed a face.

Hence in total we removed one vertex and one face from $\chi(X) + \chi(Y)$, i.e. $\chi(X\#Y) = \chi(X) + \chi(Y) - 2$. □

Corollary 1.23. *The Euler characteristic of the n-holed torus is $2 - 2n$.*

Proof. We know $\chi(T^2) = 0$. Then

$$\begin{aligned}\chi(T^2\#T^2) &= 0 + 0 - 2 = -2\\ \chi((T^2\#T^2)\#T^2) &= -2 + 0 - 2 = -4\\ \chi((T^2\#T^2\#T^2)\#T^2) &= -4 + 0 - 2 = -6\\ &\cdots .\end{aligned}$$

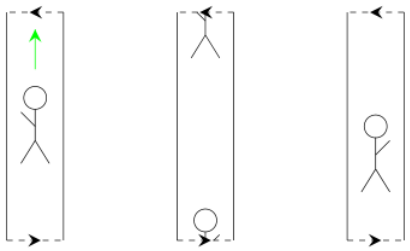
□

1.6 Orientability

After all that work with cutting and pasting, we still don't seem to be able to tell apart the Klein bottle Kl^2 and the torus T^2. (Try cutting and pasting their planar models; it doesn't seem like we can get from one to the other.) We need another invariant.

Definition 1.24. A surface is **non-orientable** if it contains a Möbius strip. It is **orientable** otherwise.

This is a confusing definition: why a Möbius strip, and why call it "orientable"? Here's the idea. Imagine a person with one arm living on the Möbius strip who starts from one point, walks around the entire Möbius strip, and returns to his starting point.

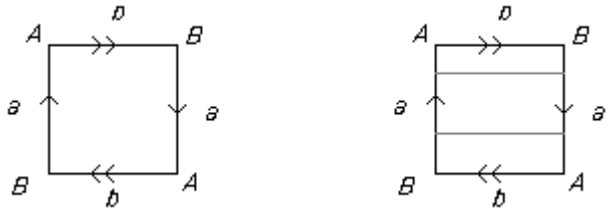


Unexpectedly, he is now flipped! What used to be "his left side" is now "his right side." We see that

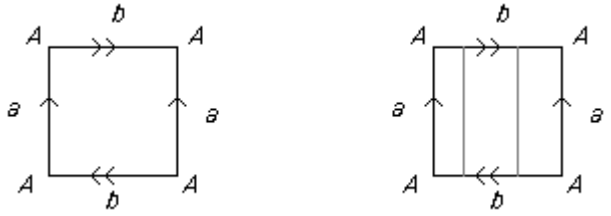
> On a Möbius strip, the concepts of "right" and "left" do not make sense, because walking around the strip once interchanges them.

It turns out that if a surface does *not contain* a Möbius strip, then "left" and "right" always make sense. Therefore we can pretend **non-orientable** means we cannot distinguish between "right" and "left."

Example 1.25. The projective plane $\mathbb{P}^2$ is not orientable. To see this, we look at its planar model, which obviously contains a Möbius strip (whose boundary is marked by the two gray lines) by visual inspection!



Similarly, the Klein bottle Kl^2 is not orientable:



But for more complicated surfaces, we want an easier way of determining whether there is a Möbius strip contained in the surface.

Theorem 1.26. *A surface is non-orientable if and only if its planar model contains two edges glued "in opposite directions," i.e. two edges both going clockwise, or both going counterclockwise, around the boundary of the polygonal planar model.*

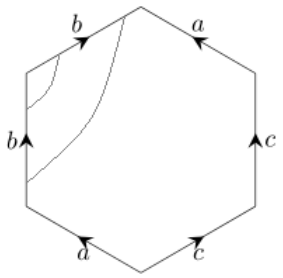
"If and only if" is actually a technical term. We say "A if and only if B" to mean that "(if A is true, then B is true) **and** (if B is true, then A is true)."

Proof. We need to prove the following two things in order to prove the theorem:

1. if the planar model of a surface contains two edges glued in "opposite directions," then the surface is non-orientable;

2. if a surface is non-orientable, its planar model contains two edges glued "in opposite directions".

The first statement is easy. If a planar model contains two edges glued "in opposite directions," we can draw a Möbius strip using those two edges. The diagram below is an example, again with gray edges marking the boundary of the Möbius strip.



Conversely, if a surface S is non-orientable, then by definition it contains a Möbius strip. The planar model of the Möbius strip contains two edges glued "in opposite directions," and these two edges have to be somewhere in the planar model of S. □

The "converse" of the statement "if A then B" is "if B then A," which is what we're going to prove now.

Example 1.27. The torus T^2 is orientable, because in its planar model, every pair of edges that are glued together are glued "in the same direction." So we can finally distinguish between the Klein bottle and the torus: the Klein bottle is non-orientable, but the torus is orientable. They are *different surfaces*!

1.7 Word representations

Although we could state and prove the classification theorem for surfaces right now, we actually want to develop one more tool. The problem right now is that working with planar models, e.g. cutting and gluing, is still a little cumbersome, and we would like a better way to represent planar models without having to draw them out every time. This is an important strategy mathematicians use:

> When working with complicated objects, try to represent them using symbols in a way that operations involving the complicated objects translate to very simple operations on the symbols.

(In fact one can claim that this is the strategy underlying *all* of mathematics.)

Definition 1.28. Given a planar model, its **word representation** is obtained as follows:

1. label each distinct edge of the polygon with a different letter (and of course two identified edges have the same label);

2. starting at any vertex, go clockwise around the polygon and read out the letters of each edge, and if an edge goes in the opposite direction, add an inverse sign, e.g. a^{-1} means an edge labeled a but going counterclockwise.

Example 1.29. Here are the word representations of the planar models we've been using for a few surfaces:

We choose to use this inverse notation because in many ways, a^{-1} should be treated as "$1/a$," because, for example, we will see that a^{-1} cancels with a when they are next to each other.

1. (sphere) $abb^{-1}a^{-1}$;

2. (torus) $aba^{-1}b^{-1}$;

3. (Klein bottle) $aba^{-1}b$;

4. (projective plane) $abab$.

Of course, a surface can have many different word representations. For example, for the torus, $ba^{-1}b^{-1}a$ is also a word representation, where we started reading clockwise from the top left (instead of the bottom left) corner of the planar model. As another example, we can also relabel edges arbitrarily, e.g. $aba^{-1}b^{-1} \sim cdc^{-1}d^{-1}$. We will need a list of operations we can perform on a word without changing the surface. Before we make this list, we need to make a small observation about word representations.

Lemma 1.30. *In any word representation of a surface, each letter that is used appears exactly twice.*

Proof. If a letter appears more than twice, then more than two edges of the polygon are being glued together. The resulting space is not a surface, because at any point on the edges being glued together, there are more than two directions in which we can move. □

Now we can make a list of allowed operations on word representations. For example, the operation of "relabel the edge x as y" can be written $AxBxC \sim AyByC$, where A, B, C represent some clumps of letters, and the $\sim$ means the resulting surface is equivalent to the original surface. By the lemma, x appears only twice, so A, B, and C do not contain x and therefore are unaffected by the relabeling. More intuitively, we write $\cdots x \cdots x \cdots \sim \cdots y \cdots y \cdots$.

The dots are parts of the word representation that remain unchanged by the operation.

Theorem 1.31. *The following operations on a word representation will give a word representation of an equivalent surface.*

1. **Relabeling edges***: replace every edge named x with some other unused symbol y, i.e. $\cdots x \cdots x \cdots \sim \cdots y \cdots y \cdots$.*

2. **Cycle the word***: move the first letter to the end, i.e. $x \cdots \sim \cdots x$.*

3. **Merge edges***: if the same sequence of letters appears twice, merge them into one letter, e.g. $\cdots xy \cdots xy \cdots \sim \cdots z \cdots z \cdots$.*

4. **Cancel xx^{-1} pairs**: *if xx^{-1} occurs in a word for any letter x, then we can remove xx^{-1} from the word, i.e. $Axx^{-1}B \sim AB$.*

5. **Swap edges**: *this is really the two different operations*

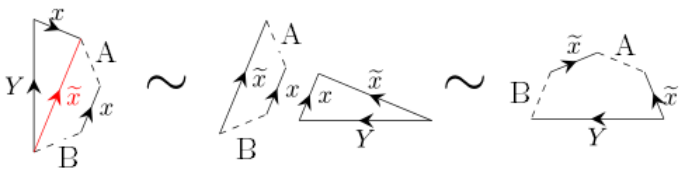
$$\cdots Yxx\cdots \sim \cdots xY^{-1}x\cdots$$
$$\cdots Yx\cdots x^{-1}\cdots \sim \cdots x\cdots x^{-1}Y\cdots .$$

Here Y^{-1} means "take the entire clump of letters Y and flip them," e.g. $(ab)^{-1} = b^{-1}a^{-1}$.

Also, the word representation of the connected sum of two surfaces is just the concatenation of their word representations, e.g. $T^2\#T^2$ has word representation $aba^{-1}b^{-1}cdc^{-1}d^{-1}$.

We have to relabel the edges of the second torus so that we don't accidentally glue them onto the edges of the first torus.

Proof. It is fairly straightforward that most of these operations give an equivalent surface, but the last operation of moving edges around needs some explanation. Both are obtained via some clever cutting and pasting. For example, the second swapping edges operation, which says $YxAx^{-1}B \sim xAx^{-1}YB$, is proved by the cutting and pasting diagram



followed by renaming $\tilde{x}$ back to x. The other equivalence is left as an exercise. □

The key thing to keep in mind is that all these operations on word representations are just "abbreviations" for operations (generally cutting and pasting operations) on the planar models corresponding to the word representations. But it is *much easier* for us, somehow, to manipulate word representations, as the following examples show.

Example 1.32. The word representation for $\mathbb{P}^2$ we have right now is $abab$. We write this as $\mathbb{P}^2 \sim abab$. But

$$\mathbb{P}^2 \sim abab \overset{3}{\sim} cc \overset{1}{\sim} aa,$$

so aa is also a word representation of $\mathbb{P}^2$. (The numbers above the $\sim$ specify which operation is being used.)

From now on we will use the word representation $\mathbb{P}^2 \sim aa$.

Example 1.33. The Klein bottle Kl^2 and the connected sum $\mathbb{P}^2\#\mathbb{P}^2$ of two projective planes both have Euler characteristic 0. (Check this for yourself using the formula in Theorem 1.22 for $\chi(A\#B)$.) Are they equivalent surfaces?

$$\mathrm{Kl}^2 \sim aba^{-1}b \overset{5}{\sim} baab \overset{2}{\sim} aabb \sim \mathbb{P}^2\#\mathbb{P}^2$$

by swapping edges followed by cycling the word.

Example 1.34. Here is another equivalence that will be useful later on:

$$\begin{aligned}\mathbb{P}^2\#T^2 &\sim (aab)cb^{-1}c^{-1} \overset{5}{\sim} a(b^{-1}acb^{-1})c^{-1}\\ &\overset{5}{\sim} a(c^{-1}a^{-1}b^{-1}b^{-1}c^{-1}) \overset{5}{\sim} abbac^{-1}c^{-1}\\ &\overset{1}{\sim} abba\#\mathbb{P}^2 \overset{2}{\sim} aabb\#\mathbb{P}^2 \sim \mathbb{P}^2\#\mathbb{P}^2\#\mathbb{P}^2.\end{aligned}$$

1.8 Classification of surfaces

Now we have all the necessary tools to state and prove the classification theorem for surfaces, which will conclude our study of surfaces as an introduction to topology. (We'll go on to study higher-dimensional spaces afterward.) The classification theorem is, in my opinion, one of the most beautiful results in topology, and, indeed, all of mathematics.

Theorem 1.35. *Every (compact, connected) surface is equivalent to one of the following three types of surfaces:*

1. *a sphere;*
2. *a connected sum of projective planes (if it is non-orientable);*
3. *a connected sum of torii (if it is orientable, and not a sphere).*

Being "compact" is a technical detail. All surfaces constructed using planar models are "compact." Essentially, "compact" means the surface area of the surface is finite. For example, $\mathbb{R}^2$ is not compact. "Connected" just means the surface is "one piece." For example, two spheres S^2 sitting side by side is a surface, but is not connected.

Proof. The main idea of the proof is, given a word representation, to apply the operations of Theorem 1.31 in a certain order, so that the end result is some sort of "standard form" for the word representation.

1. **Collect like terms into projective planes**: apply the (first) swapping edges operation to move any pair of letters to the beginning of the word, e.g.
$$X(aYa)Z \overset{5}{\sim} X(Y^{-1}aa)Z \overset{2}{\sim} aaZXY^{-1} \sim \mathbb{P}^2\#ZXY^{-1}.$$
Now ignore the $\mathbb{P}^2$ and work with the remaining surface ZXY^{-1}.
2. Repeat the previous step until there are no more pairs of letters xx available (i.e. if x appears, the other edge is x^{-1}, not x). Hence we have written the original word representation as $\mathbb{P}^2\#\cdots\#\mathbb{P}^2\#W$ where if a appears in the word representation W, then it is glued with a^{-1}. Now ignore the $\mathbb{P}^2$ and work with the remaining surface W.
3. **Collect interlinked pairs into torii**: apply the (second) swapping edges operation to move letters $\cdots a\cdots b\cdots a^{-1}\cdots b^{-1}\cdots$ to the beginning of the word, e.g.

The whole point of this tedious computation is to show we can move $aba^{-1}b^{-1}$ to the front of the word.

$$\begin{aligned} V(aWbXa^{-1})(Yb^{-1}Z) &\overset{5}{\sim} (aWbXa^{-1})V(Yb^{-1}Z) = aW(bXa^{-1}VYb^{-1})Z \\ &\overset{5}{\sim} a(bXa^{-1}VYb^{-1})WZ = (abXa^{-1}VYb^{-1})WZ \\ &\overset{2}{\sim} (bXa^{-1}VYb^{-1}a)WZ = bX(a^{-1}VYb^{-1}a)WZ \\ &\overset{5}{\sim} b(a^{-1}VYb^{-1}a)XWZ = (ba^{-1}VYb^{-1}a)XWZ \\ &\overset{2}{\sim} (b^{-1}aba^{-1}VY)XWZ = (b^{-1}aba^{-1})VYXWZ \\ &\overset{2}{\sim} (aba^{-1}b^{-1})VYXWZ \sim T^2\#VYXWZ. \end{aligned}$$

Now ignore the T^2 and work with the remaining surface $VYXWZ$.

4. Repeat the previous step until there are no more interlinked pairs of letters $\cdots a \cdots b \cdots a^{-1} \cdots b^{-1} \cdots$. Hence we have written the original word representation as $\mathbb{P}^2\# \cdots \#\mathbb{P}^2\#T^2\# \cdots \#T^2\#W$ where all the letters that remain in W are already matched up in pairs $xx^{-1}yy^{-1}zz^{-1}$.

5. **Cancel** xx^{-1}: using rule 4, cancel all the pairs xx^{-1} in the remaining part W of the word representation.

6. **Conclude**: we have turned the original word representation into something of the form $\mathbb{P}^2\# \cdots \#\mathbb{P}^2\#T^2\# \cdots \#T^2$. Now there are three cases:

   (a) if everything canceled and there are zero $\mathbb{P}^2$ and T^2, then the result is a sphere;

   (b) if there are only T^2 and no $\mathbb{P}^2$, the result is a connected sum of torii;

   (c) otherwise if there are both $\mathbb{P}^2$ and T^2, use that $\mathbb{P}^2\#T^2 \sim \mathbb{P}^2\#\mathbb{P}^2\#\mathbb{P}^2$ to turn all the torii T^2 into projective planes $\mathbb{P}^2$, so the result is a connected sum of projective planes only. □

Check that the word representation of a sphere is $S^2 \sim aa^{-1} \sim 1$, where 1 represents the "empty word" consisting of no letters at all.

Of course, we should check that these three types of surfaces are not equivalent to each other. This can be done using Euler characteristic along with orientability, and is left as an exercise.

2 Manifolds

Let's return to theory. The classification theorem for surfaces shows that we understand surfaces, which are two-dimensional objects, very well. The natural question that a mathematician asks at this point (about any result, after proving it) is

> Can we generalize this result somehow, and what is the most general statement we can make?

Of course, there are many things to potentially generalize. The classification theorem is a statement about two-dimensional spaces. So:

1. we could try to prove a classification theorem for two-dimensional surfaces *with boundary*;

2. we could also try to prove a classification theorem for *three or higher dimensional* spaces (maybe even with boundary).

It turns out (1) is quite straightforward using the same tools we used for the boundary-less case. But (2) is not. In order to start investigating (2), we must generalize the definition of a surface to higher dimensions.

Definition 2.1. A (topological) **manifold of dimension** n is a space that, if you were to zoom in on any point, looks like a bent piece of the n-dimensional space $\mathbb{R}^n$.

$\mathbb{R}^n$ is like $\mathbb{R}^2$, but instead of two coordinate axes, it has n, all in "different directions."

Actually, now that we are giving the general definition, it is time to point out that even in our formal definition of a surface, there were several missing technical details.

Formal definition. A (topological) manifold of dimension n is a space that is *locally homeomorphic* to $\mathbb{R}^n$, and also *Hausdorff* and *second-countable*.

Being "Hausdorff" and "second-countable" are technical details designed to exclude really weird spaces. It is very hard to be locally homeomorphic to $\mathbb{R}^n$ and not be Hausdorff and second-countable, so we will ignore these details.

As for surfaces, we have the usual notion of **homeomorphism** for manifolds of dimension n as well.

2.1 Examples of manifolds

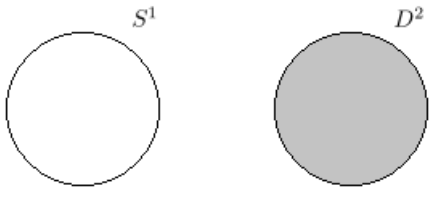
As we did for surfaces, we immediately look for examples of manifolds. We start with two obvious examples.

Example 2.2. Every surface is a manifold of dimension 2.

Example 2.3. The space $\mathbb{R}^n$ is a manifold of dimension n, called the **n-dimensional affine space** or **affine n-space** for short. Just like points in $\mathbb{R}^2$ are (x, y) for any real numbers x and y, the points in $\mathbb{R}^n$ are $(x_1, \ldots, x_n)$ for any real numbers $x_1, \ldots, x_n$.

We must also look for non-examples, so that we can tell what *isn't* a manifold. Of course, one thing that can go wrong is that we have a boundary. Another thing that can go wrong is that some parts of the space have lower or higher dimension than we want.

Example 2.4 (Non-example)**.** The **2-dimensional disk** D^2 is the space obtained by "filling in" the inside of the 1-dimensional circle S^1:

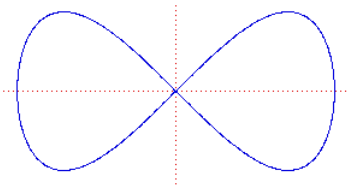


There is a boundary here, formed by what used to be S^1. However, aside from the boundary, D^2 is a perfectly valid manifold, so we call it a **manifold with boundary**. The **boundary** of a manifold X with boundary is denoted ∂X.

Just as with surfaces, a manifold without boundary is *not* a manifold, by our definition!

Example 2.5 (Non-example)**.** Affine n-space $\mathbb{R}^n$ is **not** an m-dimensional manifold for $m \neq n$. This is actually very hard for us to prove right now.

Example 2.6 (Non-example)**.** The following is not a 1-dimensional manifold:



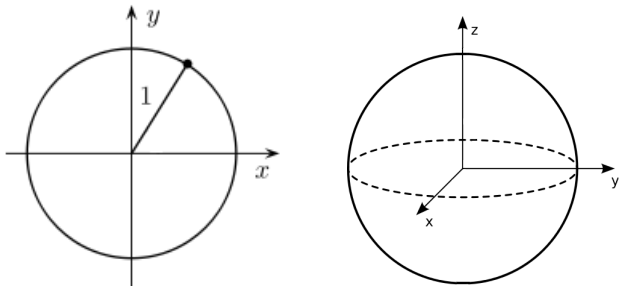
No matter how much we zoom in on the origin, there will be four "branches" coming out of it, and it will look like an X.

The problem with higher-dimensional examples is that higher-dimensional manifolds are very hard to visualize. Generally, higher-dimensional manifolds are described as graphs of equations, which, although still hard to visualize, are easier to work with and to manipulate. The following example demonstrates this.

Example 2.7. We will construct the **n-dimensional sphere**, or "n-sphere," for every $n \geq 0$. It is denoted S^n.

If a space has a name, e.g. "sphere," and is n-dimensional, we usually put the n before the name, e.g. "n-sphere" or "affine n-space."

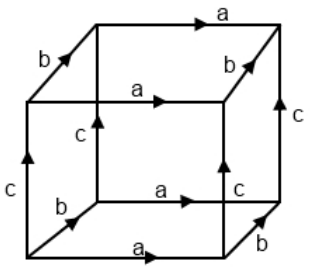
1. $(n = 1)$ The 1-sphere S^1 is the circle $x^2 + y^2 = 1$ inside $\mathbb{R}^2$.
2. $(n = 2)$ The 2-sphere S^2 is the sphere $x^2 + y^2 + z^2 = 1$ inside $\mathbb{R}^3$.



Following the pattern, for n in general, we define the **n-sphere** S^n to be the sphere $x_1^2 + x_2^2 + \cdots + x_{n+1}^2 = 1$ inside $\mathbb{R}^{n+1}$, whose points are of the form $(x_1, \ldots, x_{n+1})$.

Another method of describing higher-dimensional manifolds is via gluing. Just like we could glue sides of a square to get a 2-dimensional torus (via its planar model), we can glue sides of a cube to get a 3-dimensional torus.

Example 2.8. The following filled-in cube in $\mathbb{R}^3$ with the given edge (and therefore face) identifications is the **3-torus**.



Exercise. A square lives in $\mathbb{R}^2$, and a cube lives in $\mathbb{R}^3$. How would you describe the higher-dimensional analogues, called **hypercubes**, in $\mathbb{R}^n$? Using hypercubes, how would you define the "n-torus" for larger n? How about using n-spheres to define the "n-disk"?

2.2 Constructing new manifolds

We are not done coming up with examples of manifolds! Just like with surfaces, we can also cut and paste with manifolds. But now, because we don't have to restrict ourselves to dimension 2, we can define more general constructions. These constructions also make it easier to describe higher-dimensional manifolds.

Definition 2.9. The **product** of two manifolds M and N is called $M \times N$. It is the manifold obtained by taking every point in M and replacing it with a copy of N (or by taking every point in N and replacing it by a copy of M).

Although this is called a product, we never write MN to mean $M \times N$.

Formal definition. The product $M \times N$ of two manifolds consists of the points

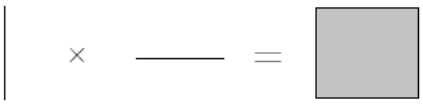
$$M \times N = \{(x, y) : x \in M \text{ and } y \in N\},$$

which is just fancy notation that means "all the points (x, y) such that x is a point in X and y is a point in Y."

It is not obvious that $M \times N$ is still a manifold! We would have to check that if we zoom on any point in $M \times N$, it looks like a bent piece of $\mathbb{R}^k$ for some k. What should k be? The following examples may shed some light on this.

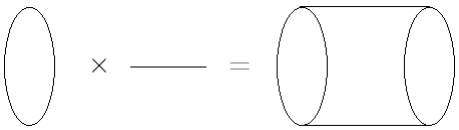
Example 2.10. The product $\mathbb{R} \times \mathbb{R}$ is just $\mathbb{R}^2$. We can see this by taking the product of two line segments, and then imagining that the segments actually extend forever in both directions.

This example gives a good reason for calling affine n-space $\mathbb{R}^n$: it is literally the product of n copies of $\mathbb{R}$.



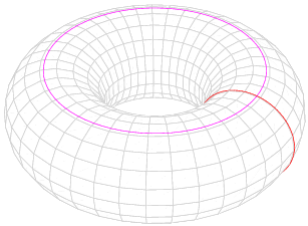
Analogously, the product $\mathbb{R}^2 \times \mathbb{R}$ is $\mathbb{R}^3$. In general, $\mathbb{R}^m \times \mathbb{R}^n = \mathbb{R}^{m+n}$, although this is harder to visualize.

Example 2.11. The product $S^1 \times \mathbb{R}$ is an infinitely long cylinder.



Example 2.12. The product $S^1 \times S^1$ is the torus! In fact, we call the product of n copies of S^1 the **n-torus**, and denote it $(S^1)^n$ or T^n.

Even though we never write MN for the product $M \times N$, it is very common to write M^n to mean the product of n copies of M.



Theorem 2.13. *Let M be a manifold of dimension m, and N be another manifold of dimension n. Then $M \times N$ is a manifold of dimension $m + n$.*

This is the reason we didn't see the product construction when we covered surfaces: the product of two surfaces is a 4-dimensional manifold.

Proof idea. At every point of $M \times N$, we have m directions to move which come from M, and n extra directions to move which come from N. □

Another very useful construction is a generalization of the idea of gluing. However, we must be very careful with this construction, because it does not always produce a manifold! In fact, it should actually be defined for any space whatsoever, not just manifolds, which is what we will do.

Definition 2.14. Let X be any space. The idea is to obtain a new space by gluing certain points of X together. The gadget that tells us which points to glue is called an **equivalence relation**, written $\sim$. If we want to glue two points x and y in X, we set x to be equivalent to y, which we write as $x \sim y$. The **quotient** $X/\sim$ is the space obtained from X by gluing together every pair of points x, y that are equivalent.

In general, we use $\sim$ to mean "equivalent." We already did this with surfaces, saying $X \sim Y$ if X and Y are homeomorphic.

Example 2.15. Take D^2, the 2-dimensional disk. If we squish all of its boundary into a point, we get S^2, the 2-dimensional sphere:

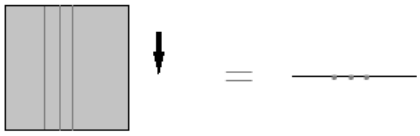
We say that S^2 is the quotient of D^2 by the equivalence relation where $x \sim y$ if and only if x and y are both in the boundary D^2.

In the above example, the quotient was particularly simple: we take a part of our space X and squish it into a point. Because this situation is so common, there is a special name and a special notation for it.

Definition 2.16. Let X be any space, and A be any piece of X, called a **subspace**. The **quotient** X/A is the space obtained from X by squishing everything in A into a point. So the above example shows $D^2/\partial D^2 = S^2$.

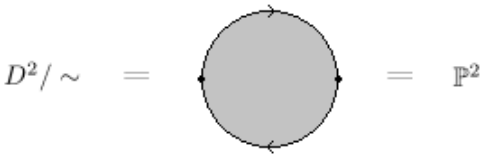
In other words, X/A is the quotient $X/\sim$ where $x \sim y$ if and only if x and y are both in A.

Example 2.17. Take $\mathbb{R}^2$, the 2-dimensional plane. Let the equivalence relation be $(x, y_1) \sim (x, y_2)$ for all real numbers x and y_1 and y_2. What is the quotient $X/\sim$? Well, any two points on the same vertical line are equivalent according to the equivalence relation. So the effect of quotienting is to squish each vertical line into a point.



But then we are just left with an infinite line of points, which is $\mathbb{R}$.

Example 2.18. Take D^2, the 2-dimensional disk. View it inside $\mathbb{R}^2$, centered at the origin $(0, 0)$. Let the equivalence relation be $x \sim -x$ for every point x on the boundary of D^2. We have already seen the quotient $D^2/\sim$. It is exactly the planar model for the projective plane $\mathbb{P}^2$!



Generalizing, we define the **n-dimensional projective space** $\mathbb{P}^n$ to be the quotient of D^n by the equivalence relation $x \sim -x$ for every point x on the boundary of D^n.

2.3 Distinguishing between manifolds

Now we have a bunch of manifolds (e.g. $\mathbb{R}^n$, S^n, T^n, $\mathbb{P}^n$), and just like with surfaces, we would like to have tools to tell them apart. We had two tools for surfaces: Euler characteristic and orientability. Both of them generalize, but not easily, and they are *not the most useful tools anymore*. For example, Euler characteristic for surfaces used the idea of vertices, edges, and faces, but now we can have higher-dimensional objects.

Definition 2.19. We define higher-dimensional analogues of "faces" by analogy. The idea is to see what makes an edge an edge, and a face a face, and then to extend those properties.

We actually need all these higher-dimensional analogues to be "convex," which roughly means that they have "no holes."

1. A polygon, also called a "2-polytope," is a 2-dimensional object whose boundary consists of lines (called 1-polytopes).

2. A polyhedron, also called a "3-polytope," is a 3-dimensional object whose boundary consists of polygons, i.e. 2-polytopes.

3. A 4**-polytope** is a 4-dimensional object whose boundary consists of polyhedra.

4. In general, an n**-polytope** is an n-dimensional space whose boundary consists of $(n-1)$-polytopes. Vertices are 0-polytopes.

The **Euler characteristic** $\chi(M)$ of a manifold M of dimension n is found by deforming the manifold M into a collection of n-polytopes, and then calculating

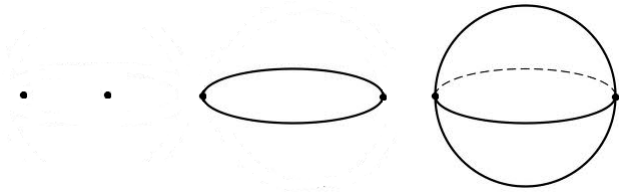
$$\chi(M) = c_0 - c_1 + c_2 - c_3 + \cdots + (-1)^n c_n$$

where c_k is the number of k-polytopes. (For surfaces, c_0 is the number of vertices, c_1 the number of edges, and c_2 the number of faces.)

The process of deforming M into a collection of n-polytopes is (roughly) also known as "putting a CW-complex structure" on M.

Just like Euler characteristic for surfaces using this barebones definition is hard to compute, it is also hard to compute for manifolds.

Example 2.20. We will compute $\chi(S^n)$ for all $n \geq 0$. First, we look at small n to get some intuition. Here are S^0, S^1, and S^2:



In other words, to get S^1 from S^0, we add two 1-polytopes (the edges) joining the two 0-polytopes (the vertices). To get S^2 from S^1, we add two 2-polytopes (the faces, i.e. upper and lower hemispheres) joining the two 1-polytopes (the edges). So in general, to get S^n from S^{n-1}, we just need to add two n-polytopes joining the two $(n-1)$-polytopes in S^{n-1}.

This shows that in S^n, there are exactly two k-polytopes for every k. So the Euler characteristic is

$$\chi(S^n) = 2 - 2 + 2 - 2 + \cdots + (-1)^n 2 = \begin{cases} 2 & n \text{ even} \\ 0 & n \text{ odd.} \end{cases}$$

This is one possible CW complex structure that we can put on S^n. There are others. A standard one is to use the fact that $D^n/\partial D^n = S^n$, so we can take one n-polytope (homeomorphic to D^n) and glue everything on its boundary to one 0-polytope (a point), so in total there are only two polytopes: one 0-polytope and one n-polytope.

Exercise (Challenge)**.** This is hard: compute the Euler characteristics

$$\chi(\mathbb{R}^n) = 0, \quad \chi(T^n) = 0, \quad \chi(\mathbb{P}^n) = \begin{cases} 1 & n \text{ even} \\ 0 & n \text{ odd.} \end{cases}$$

We see that in higher dimensions, the Euler characteristic is somewhat useless: it can't even tell apart S^2 and S^4, and it is also hard to compute. Worse, it is even harder to define orientability in an intuitive way, and it is harder still to check for orientability using only the tools we have so far.

Definition 2.21 (Rough definition)**.** The correct way to generalize "being orientable" from surfaces to n-dimensional manifolds is something like the following. A manifold M is **orientable** if at each point we can consistently define what "clockwise" and "anti-clockwise" mean, so that regardless of how we move around M, it never happens that "clockwise" becomes "anti-clockwise."

We will not focus on Euler characteristic or orientability anymore. The reason is that there are more powerful invariants that we have yet to talk about.

2.4 Homotopy and homotopy groups

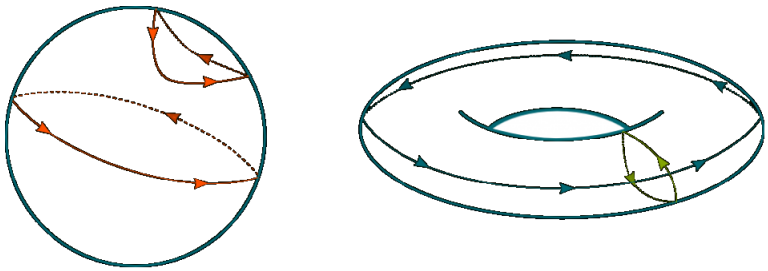
There are two very powerful invariants of spaces in general: homotopy groups, and homology groups. Homotopy groups are hard to calculate but are relatively easier to define; homology groups are easier to calculate but are harder to define. So we begin with homotopy groups, which will enable us to tell apart many of the manifolds we have so far. First, let's see the underlying idea.

Definition 2.22. A **loop** on a manifold M is a path, with direction and possibly self-intersecting, that begins and ends at the same point, called the **basepoint** of the loop. A loop is **homotopic** to another loop if we can deform the first loop into the second loop while staying inside the manifold and without breaking the loop.

Formal definition. A loop on a manifold M is a continuous function $\gamma(t)$, for $0 \leq t \leq 1$, to the manifold M such that $\gamma(0) = \gamma(1)$, i.e. the start and end points are the same point, the basepoint, in M. A loop γ_0 is homotopic to another loop γ_1 if they are part of a collection of loops γ_s, for every $0 \leq s \leq 1$, such that $\gamma_s(t)$ is a continuous function in (s, t).

The Greek letter γ, called "gamma," is often used to denote paths.

Example 2.23. Here are some loops on the sphere S^2 and the torus T^2:



The two loops drawn on the sphere are homotopic. But the two loops drawn on the torus are not homotopic: there is no way to deform one loop into the other without breaking it. In fact, with a bit more thought, we can see that *any two loops on the sphere are homotopic*!

Since we are trying to construct an *invariant*, an important observation to make at this point is that

> The statement "two loops are homotopic" is invariant under homeomorphism. In other words, given two homotopic loops on a manifold M, if we deform it (along with the loops drawn on it) into an equivalent manifold N, the two resulting loops on the manifold N are still homotopic.

Immediately, this gives us a way to distinguish the sphere S^2 from the torus T^2: any two loops on the sphere are homotopic, but not every two loops on the torus are homotopic. This idea underlies the construction of the homotopy groups. We will define the first homotopy group in the following very long and technical definition. Hopefully the examples that follow it will help.

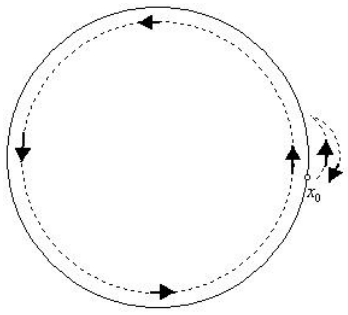
Definition 2.24. Let M be a manifold and fix a point x_0 in M. The **first homotopy group** (or **fundamental group**) $\pi_1(M, x_0)$ of M is the collection of all non-homotopic loops on M with basepoint x_0, with the following operations on them:

1. if γ and η are two loops, then we can **concatenate** them to get a loop $\gamma\eta$, which is the loop where we traverse the path given by γ and then traverse the path given by η;
2. if γ is a loop, we can **reverse** it to get a loop γ^{-1}, which is the loop γ traversed backward.

Warning: $\gamma\eta$ is in general not homotopic to $\eta\gamma$, so order matters! (In other words, concatenation is not commutative.) Can you find an example?

Example 2.25. Fix a point x_0 in S^2. There is a really silly loop which is the path that goes from x_0 to x_0 by staying fixed at x_0; we call this the **constant loop**. Since any two loops on the sphere S^2 are homotopic, every loop is homotopic to the constant loop. So "the collection of all non-homotopic loops on S^2" is just the constant loop.

Example 2.26. Let's try to figure out what the first homotopy group of the circle S^1 is. First of all, we should try to get a sense of what loops on the circle are like. Here are two loops:



Again, we can consider the constant loop. The loop in the diagram above that goes all the way around the circle to return to x_0 is not homotopic to the constant loop, but the other loop in the diagram is indeed homotopic to the constant loop. So we have identified two non-homotopic loops:

1. the "go once around the circle counterclockwise" loop, which we call γ, and

2. the constant loop, which we call γ^0 (for a very good reason that we'll see soon).

But we need to find *all* the non-homotopic loops! It turns out we can do that by thinking about the concatenation of loops.

What is the concatenation $\gamma\gamma$, also written γ^2? Well, it is the loop that goes around the circle twice. Some thought should convince you that γ^2 is not homotopic to γ or to 1. Similarly, γ^3 is not homotopic to γ^2, γ, or 1. We can continue this pattern, to get γ^n for every $n > 0$.

Just like with word representations, we pretend the operation of concatenation is like multiplication, and write γ^2 to mean γ concatenated with γ.

What about γ^{-1}? It is just γ, but traversed backward, so it is the "go once around the circle clockwise" loop. Again, some thinking should convince you that γ^{-1} is not homotopic to any γ^n. Now we can concatenate to get $\gamma^{-n} = \gamma^{-1} \cdots \gamma^{-1}$ for any $n > 0$.

Summary: we have constructed non-homotopic loops γ^n for every integer n. It is true that any loop on the circle S^1 is homotopic to one of these loops. So they are all the non-homotopic loops on S^1. Hence

> The first homotopy group $\pi_1(S^1, x_0)$ consists of all the loops γ^n for integers n, with the operation
>
> $$\gamma^n \gamma^m = \gamma^{n+m} \quad \text{for integers } n, m.$$

But we can give a better description. Note that it really doesn't matter what we call the loops γ^n. So let's give them different names; call γ^n just by the integer n. Then what is $\pi_1(S^1, x_0)$? It is the collection of integers, with concatenation being precisely addition of integers: if we concatenate the loops n and m, then we get the loop $n + m$. So $\pi_1(S^1, x_0)$ is secretly just the **integers**, called $\mathbb{Z}$, in disguise! Mathematicians like to write $\pi_1(S^1, x_0) = \mathbb{Z}$. (More precisely, this is a **group isomorphism**. See Appendix A for details.)

The letter $\mathbb{Z}$ denotes the integers because of the German "Zahlen," meaning "numbers."

That was harder than computing Euler characteristic! But at least we can tell apart S^1 from S^2 and $\mathbb{R}^n$ now, since S^1 has $\mathbb{Z}$ as its first homotopy group, but the first homotopy groups of S^2 and (as an exercise) $\mathbb{R}^n$ both consist of only the constant loop.

Exercise. Convince yourself that $\pi_1(S^2, x_0)$ and $\pi_1(\mathbb{R}^n, x_0)$ consists of just the constant loop (for any choice of basepoint x_0). If the first homotopy group consists of just the constant loop, we say it is **trivial**, or that it is the **trivial group**.

Exercise (Challenge)**.** Convince yourself that $\pi_1(S^n, x_0)$ is trivial for every $n \geq 2$. (Hint: use that $S^n = D^n / \partial D^n$.)

So the first homotopy group can only tell apart S^1 from the rest of the higher-dimensional spheres. Let's see an outline of how to define higher homotopy groups, which can be used to tell apart S^n from S^m for every $n \neq m$.

Definition 2.27 (Rough definition)**.** We can rephrase the first homotopy group $\pi_1(M, x_0)$ of a manifold M as: the collection of non-homotopic ways to draw S^1 in M such that the north pole of S^1 is at the basepoint x_0. The generalization is the **n-th homotopy group** $\pi_n(M, x_0)$ of M, which is the collection of non-homotopic ways to draw the n-dimensional sphere S^n in M such that the north pole of S^n is at the basepoint x_0.

A **major unsolved problem** in topology is to find tools that let us compute the n-th homotopy group of S^m for every n and m.

Mathematicians call this the problem of "computing higher homotopy groups of spheres."

2.5 Classification of manifolds

Let's think a little about what we've done so far. We started off with classifying surfaces, and this we did very effectively using planar models and their word representations to glue together, using the connected sum, the "building blocks" of surfaces: spheres, torii, and projective planes. We have not focused on this approach in the setting of manifolds. Why is that?

The *fundamental issue* with trying to find "building blocks" for manifolds of dimension 3 or higher is the following.

> For surfaces, if we cut out disks D^2 from two surfaces and glue along the resulting boundary circles S^1, there is essentially only one way to glue the two boundary circles together. For higher dimensions, if we cut out disks D^n and try to glue along the resulting boundary circles S^{n-1}, there are many (often infinite) different ways to glue the two boundary spheres S^{n-1} together.

This makes the classification problem for manifolds very difficult in higher dimensions. In fact, it is *impossible* for dimensions 4 or higher, by the following result.

Theorem 2.28 (Proved by Markov, 1958)**.** *There is no algorithm, i.e. series of steps, that can decide whether two given manifolds of dimension $n \geq 4$ are homeomorphic.*

This theorem immediately shows there is no hope of classifying manifolds of dimension n where $n \geq 4$: if there were such a classification of n-dimensional manifolds into a list of different types, given two manifolds we can

1. check which types they are, and
2. declare they are homeomorphic if they are of the same type, or
3. declare they are not homeomorphic if they are not of the same type.

But this is an algorithm for deciding whether two given n-manifolds are homeomorphic, contradicting the theorem! So such a classification cannot exist in dimensions $n \geq 4$.

However, we understand low-dimensional cases very well:

1. ($n = 1$) any connected 1-dimensional manifold is homeomorphic to S^1 or $\mathbb{R}$;

2. ($n = 2$) any connected compact 2-dimensional manifold is homeomorphic to a connected sum of S^n, T^n, and $\mathbb{P}^n$ (this is the classification theorem for surfaces that we proved).

Remember that "compact" and "connected" are technical conditions and just mean "finite volume" and "consists of one piece" respectively.

Naturally, mathematicians were curious about what happens in $n = 3$. In 1904, Poincaré decided to tackle a simplified version of the classification problem in $n = 3$: he wanted just to be able to see if a given 3-dimensional manifold M is homeomorphic to S^3.

Theorem 2.29 (Poincaré conjecture; posed in 1904)**.** *Every connected compact 3-dimensional manifold which has trivial first homotopy group is actually homeomorphic to S^3.*

Even though we call this a "conjecture," it is actually a proven result.

Mathematicians worked on this problem for a long time, and it quickly became famous as a particularly tricky problem in the field of topology.

Theorem 2.30 (Generalized Poincaré conjecture for $n \geq 5$; proved by Smale, 1961)**.** *For $n \geq 5$, every connected compact n-dimensional manifold whose homotopy groups agree with those of the sphere S^n is actually homeomorphic to S^n.*

Theorem 2.31 (Generalized Poincaré conjecture for $n = 4$; proved by Freedman, 1982)**.** *Every connected compact 4-dimensional manifold whose homotopy groups agree with those of the sphere S^4 is actually homeomorphic to S^4.*

Yet Poincaré's original conjecture, for $n = 3$, remained unsolved, and became one of the 7 Millennium Prize Problems posed by the Clay Mathematics Institute in 2000. As of today (April 28, 2018), only one of these 7 problems has been solved: in 2003, Perelman proved the Poincaré conjecture (in $n = 3$) by actually solving the following harder problem.

Theorem 2.32 (Thurston's geometrization conjecture; proved by Perelman, 2003)**.** *Every connected compact 3-dimensional manifold can be split into pieces that fall into one of eight different types.*

This solves the classification problem in $n = 3$ as well.

3 Differential Topology

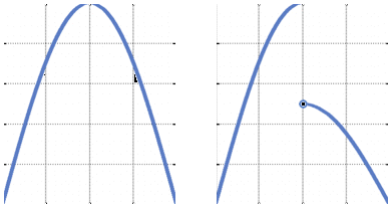
A major theme in modern mathematics is to study not only an object, but also continuous functions defined on the object. For example, instead of studying a manifold M, we can study continuous functions from S^1 to M.

Definition 3.1. A **continuous function** from a manifold M to another manifold N sends each point in M to a point in N, in a way such that if the point x in M is sent to y in N, then points "close to" x in M are sent to points "close to" y in N.

We generally call functions f or g or h. If f maps from M to N, we write $f\colon M \to N$.

Formal definition (ϵ-δ definition of continuity). A function $f\colon M \to N$ is **continuous** if at every point x in M, given a small number $\epsilon > 0$, you can find another small number $\delta > 0$ such that every point within a distance δ of x is sent to a point within a distance ϵ of y.

Example 3.2. A continuous function f from $\mathbb{R}$ to $\mathbb{R}$ is essentially any function that can be graphed on the two-dimensional plane "without lifting the pencil."



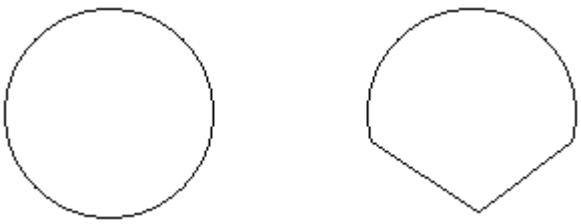
For example, the function on the left is continuous, but the function on the right is discontinuous.

Example 3.3. A continuous function f from the circle S^1 to a manifold M is just a loop inside M! Note that we implicitly assumed loops to be continuous when we first started talking about them.

We have already seen that studying loops inside manifolds is very useful, by defining the first homotopy group and using it as an invariant. In general, studying continuous functions on manifolds is very useful. However, we often want to restrict ourselves to a specific kind of continuous function.

Definition 3.4. A **differentiable function** is a continuous function where at each point it makes sense to talk about the slope at that point.

Example 3.5. Here are functions from S^1 to $\mathbb{R}^2$, i.e. two loops in the plane $\mathbb{R}^2$.



The first one is differentiable, because at every point the slope is well-defined. The second one is not differentiable at the tip of the spike, because there is no meaningful "slope" there. However note that both loops are still continuous.

The reason we want to work with differentiable functions instead of just continuous functions is that calculus works for differentiable functions and not for continuous functions. Being able to do calculus with differentiable functions

on manifolds is very helpful in investigating properties of manifolds in general. In fact, there is an entire subfield of topology, called **differential topology**, where we investigate differentiable functions on (differentiable) manifolds.

For the idea of "differentiable function" to make sense on a manifold, we actually need that the manifold itself is "differentiable," which roughly means it has no spikes or jumps or disconnections.

In this section, we will develop some basic tools of differential topology and use them to prove some cool results. For example, one of the results we will prove, called Brouwer's fixed point theorem, can be used to show that

> If I take two pieces of paper of equal size, crumple one up and put it on top of the other one, then there is at least one point on the crumpled sheet of paper that lies directly above the same point on the flat sheet.

3.1 Topological degree

The first tool we will develop is called the topological degree of a continuous function. Since this will be hard to define for arbitrary continuous functions, we'll focus on continuous functions from S^1 to S^1. These will be the functions we need to use in the proof of Brouwer's fixed point theorem.

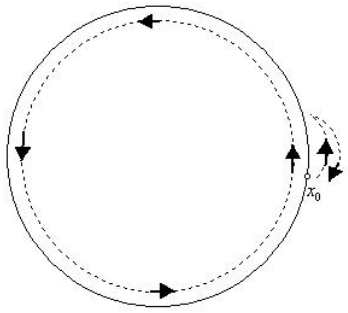
Example 3.6. We have already looked at continuous functions from S^1 to S^1: these are precisely loops on S^1! When we computed the first homotopy group $\pi_1(S^1)$ of the circle, we identified many such functions. For example, there were:

1. the constant loop γ^0;
2. the loop γ^1 that goes once around the circle counterclockwise;
3. the loop γ^{-1} that goes once around the circle clockwise;
4. concatenations of these loops.

Definition 3.7. We computed earlier that the first homotopy group of the circle is $\pi_1(S^1) = \mathbb{Z}$. This shows that given a loop $f\colon S^1 \to S^1$, it is homotopic to some γ^n for exactly one integer n. The **(topological) degree** of the function f is this integer n. We write $\deg f = n$. Note that two functions of different degree cannot be homotopic.

Topological degree can actually defined for more arbitrary continuous functions. In the special case of functions $S^1 \to S^1$, it is also commonly called "winding number," because it is the number of times a loop winds around the circle.

Example 3.8. The degrees of the two loops in the image below are 1 and 0:



This is because the loop on the left is exactly γ^1, while the loop on the right is homotopic to γ^0.

Although this should be a familiar concept by now because we are used to working with loops, it is sometimes hard to interpret functions from S^1 to S^1 as loops and to compute their degree. Let's do two very important examples.

Example 3.9. Treat S^1 as the unit (i.e. radius 1) circle centered at the origin in the two-dimensional plane.

1. Consider the **identity function** $f(x) = x$ on S^1. This is the loop γ^1, and therefore has degree 1.
2. Consider the **antipodal function** $f(x) = -x$ on S^1. This is again the loop γ^1, and therefore has degree 1.
3. Consider the **constant function** $f(x) = c$ on S^1, for some constant c. This is the constant loop γ^0, and therefore has degree 0.
4. (Represent points on the circle as $e^{i\pi\theta}$. Then the function $f(e^{i\pi\theta}) = e^{2i\pi\theta}$ has degree 2.)

The identity function is so commonly used that it has its own symbol: we denote it by id.

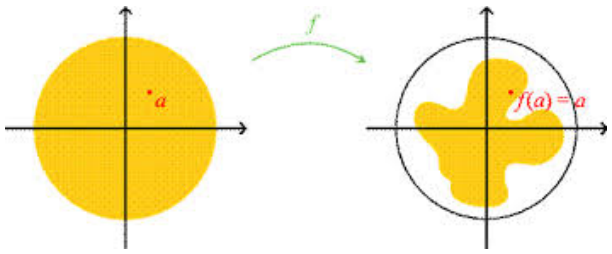
In particular, the identity function is not homotopic to the antipodal function, because they have different degree.

Note that we have actually done nothing new so far; all this is just a rephrasing of what we did in our computation of $\pi_1(S^1) = \mathbb{Z}$. But topological degree can be defined more generally, and so it is important to know it exists as a concept.

3.2 Brouwer's fixed point theorem

Using the idea of topological degree, we can prove Brouwer's fixed point theorem, one of the fundamental theorems in differential topology. Throughout this subsection, it helps to think of D^2 inside $\mathbb{R}^2$ centered at the origin.

Theorem 3.10 (Brouwer's fixed point theorem)**.** *Any continuous function $f\colon D^2 \to D^2$ from the disk D^2 to itself has a fixed point, i.e. there exists a point x in D^2 such that $f(x) = x$.*

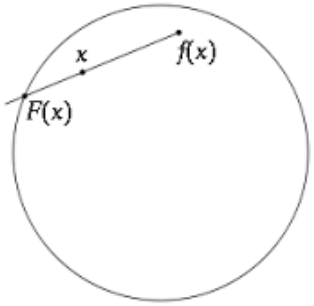


Proof. This proof is a bit more complicated than the ones we have seen so far. So, first, here is a rough outline of how the proof goes.

1. Suppose that there exists a function $f\colon D^2 \to D^2$ that has no fixed point.
2. Use it to construct a function $F\colon D^2 \to \partial D^2$ that is the identity function on ∂D^2, i.e. $F(x) = x$ for every x on the boundary of D^2.
3. Prove, using topological degree, that such a function F cannot exist.
4. Since this is a contradiction, our initial assumption that the function f exists must be false. Then we are done the proof.

Okay, let's do it!

Suppose there exists $f\colon D^2 \to D^2$ with no fixed point. That means $f(x) \neq x$ for every x in D^2. Then we can draw a line segment from $f(x)$ to x, and continue drawing until it hits the boundary of D^2.



Call the point where it hits the boundary $F(x)$. Hence we have constructed a function $F\colon D^2 \to \partial D^2$, sending every point in D^2 to some point on its boundary ∂D^2. In particular, we see that $F(x) = x$ for every point x on the boundary of D^2.

Now we show this function F cannot exist. First, since $\partial D^2 = S^1$, we can write F as $F\colon D^2 \to S^1$. If we look at F only on ∂D^2, it is the identity function $S^1 \to S^1$, and therefore has degree 1. But ∂D^2 can be continuously deformed into, i.e. is homotopic to, a constant loop inside D^2. If we look at F only on this constant loop, it is the constant function $S^1 \to S^1$ and therefore has degree 0. Hence F gives a way to continuously deform a degree 1 function into a degree 0 function. But this is impossible, since functions of different degree cannot be homotopic.

We have reached a contradiction. So our initial assumption that f exists must be incorrect. Hence every continuous function $f\colon D^2 \to D^2$ must have a fixed point. □

We used a key fact in this proof that is an important technical tool in its own right. These technical tools, to be used later in proofs of bigger things, are generally called **lemmas**. Here is the important lemma we used in the above proof.

Lemma 3.11. *If a function $F\colon S^1 \to S^1$ actually comes from looking at a part of a function $F\colon D^2 \to S^1$, then the degree of F must be zero.*

Proof. Here's a brief summary of how we proved this lemma in the above proof: such a function $F\colon D^2 \to S^1$ tells us how to make a homotopy from $F\colon S^1 \to S^1$ to the constant loop, which is degree 0. Since degree is unchanged by homotopy and the constant loop has degree 0, the function F has degree 0 as well. □

Note that Brouwer's fixed point theorem is true in general for D^n, the n-dimensional disk. The proof in this more general setting is analogous to the proof above, except using a generalization of the notion of degree.

3.3 Application: Nash's equilibrium theorem

An interesting application of Brouwer's fixed point theorem is to the Nash equilibrium theorem, which very roughly says that given any game with multiple players, there exists a strategy for each player such that, assuming the other players don't change their strategies, each player cannot gain anything by changing their own strategy, i.e. their own strategy is in some sense "the best possible." Let's first define a Nash equilibrium via an example.

Example 3.12. The **prisoner's dilemma** is a common example used to illustrate the concept of a Nash equilibrium. Two prisoners are in prison, being suspected of some crime, and are not allowed to communicate with each other. The prison warden offers each prisoner, individually, a choice: either stay silent, or betray the other prisoner by testifying the other committed the crime.

1. If both prisoners betray each other, they both get 2 years in prison.
2. If prisoner A betrays prisoner B but B remains silent, then A is set free and B gets 3 years in prison (and vice versa).
3. If both prisoners remain silent, they both get 1 year in prison.

We can write down the **payoff matrix**: a diagram that shows the "payoff," or "score," for each prisoner depending on both of their choices.

	B: silent	B: betray
A: silent	-1 -1	-3 0
A: betray	0 -3	-2 -2

The top right number in each box is the payoff for A, and the bottom left is the payoff for B. For example, -3 means three years in prison.

What should A and B do? Suppose A stays silent. Then the best strategy for B is to betray A, and then A is worse off than if A had betrayed B. So A should betray B. The same reasoning shows B should also betray A. Such a strategy for A and B, where both obtain the best outcome possible taking into account the other's choices, is called a **Nash equilibrium**.

Definition 3.13. A **pure strategy** for a player A is a complete specification of how A will play the game in any situation arising in the game. A **mixed strategy** is a collection of pure strategies each of which player A uses with some specified probability.

So a pure strategy is just a mixed strategy where one strategy has a 100% probability of being picked.

Theorem 3.14 (Nash's equilibrium theorem)**.** *Every game with a finite number of players has a Nash equilibrium among all possible mixed strategies for each player.*

We will prove Nash's equilibrium theorem for two-player games only, but the general case is completely analogous. However, before we proceed, it will be helpful to formalize some notation for how we mathematically represent a mixed strategy. The idea is as follows: if in total there are n possible pure strategies, let $0 \leq x_k \leq 1$ be the probability we pick pure strategy k, for every $k = 1, \ldots, n$. For convenience, sometimes we write $\vec{x}$ to mean the vector $(x_1, \ldots, x_n)$, and we say $\vec{x}$ is the mixed strategy.

Example 3.15. In the prisoner's dilemma, the pure strategies (for either player) are: stay silent (strategy 1), or betray the other prisoner (strategy 2). So $\vec{x} = (1/4, 3/4)$ is the mixed strategy where with probability $1/4$ we stay silent, and with probability $3/4$ we betray the other prisoner.

We also want to represent the payoff matrix in a more mathematical form. Let $A_{i,j}$ be the payoff for player A if A picks pure strategy i and B picks pure strategy j. Similarly, let $B_{i,j}$ be the payoff for player B if A picks pure strategy i and B picks pure strategy j. For convenience, sometimes we write A to mean the **matrix** (a grid of numbers) where the entry in the i-th row and j-th column is $A_{i,j}$.

This is very much like collecting $x_1, \ldots, x_n$ into a vector.

Example 3.16. The two payoff matrices (for player A and B) for the prisoner's dilemma are as follows:

$$A = \begin{pmatrix} -1 & -3 \\ 0 & -2 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 0 \\ -3 & -2 \end{pmatrix}.$$

For example, if player A picks pure strategy 1 (stay silent) and player B picks pure strategy 2 (betray), then we look at row 1 column 2 of matrix A to see that prisoner A gets 3 years in prison, and we look at row 1 column 2 of matrix B to see that prisoner B gets 0 years in prison.

What if player A picks a mixed strategy $\vec{x}$, and player B picks a mixed strategy $\vec{B}$? What are the expected payoffs for player A and player B? Recall how the concept of expected value works: if you have a $1/4$ chance of winning \$20 and a $3/4$ chance of winning \$8, your expected winning is $(1/4)(\$20) + (3/4)(\$8) = \$11$. In general, we multiply each payoff with the probability of obtaining it, and add them all together.

> Given that player A uses mixed strategy $\vec{x}$ and player B uses mixed strategy $\vec{y}$, the expected payoff for player A is
>
> $$\sum_{i=1}^{n} \sum_{j=1}^{n} x_i y_j A_{i,j} (= x_1 y_1 A_{1,1} + x_1 x_2 A_{1,2} + \cdots + x_n y_n A_{n,n})$$
>
> and likewise the expected payoff for player B is $\sum_{i=1}^{n} \sum_{j=1}^{n} x_i y_j B_{i,j}$.

Proof of Nash's equilibrium theorem. Let Δ be the set of all possible mixed strategies for player A and player B, so that an element in Δ is a pair $(\vec{x}, \vec{y})$ where $\vec{x}$ is a mixed strategy for player A, and $\vec{y}$ is a mixed strategy for player B. We will define a continuous function $f\colon \Delta \to \Delta$ such that

> given mixed strategies $(\vec{x}, \vec{y})$, if $\vec{x}$ does not give the highest possible payoff for player A, then f produces $(\vec{x}', \vec{y})$ such that $\vec{x}'$ now gives a higher possible payoff for player A (and likewise for $\vec{y}$ and player B).

Then a fixed point of f must be a Nash equilibrium, because $f(\vec{x}, \vec{y}) = (\vec{x}, \vec{y})$ means neither player A or player B can improve their payoffs by changing their strategies. Now we want to apply Brouwer's fixed point theorem to f. It is indeed continuous. The set Δ is just the set of all points $(x_1, \ldots, x_n)$ in $\mathbb{R}^n$ such that $x_1 + \cdots + x_n = 1$. The region consisting of these points is convex, and therefore homeomorphic to D^{n-1}. (It really doesn't matter what dimension the disk is.) So by Brouwer's fixed point theorem, f has a fixed point. So a Nash equilibrium must exist.

How do we construct the function f? This is conceptually easy: given $(\vec{x}, \vec{y})$, we just check if increasing x_i makes the payoff for A higher. If it does, increase x_i. Do the same for y_i and B. This produces new strategies $(\vec{x}', \vec{y}')$ that have higher payoffs. However, writing this down mathematically is a little difficult:

$$x_i' = \frac{x_i + c_i(\vec{x}, \vec{y})}{1 + \sum_{j=1}^n c_j(\vec{x}, \vec{y})} \text{ with } c_i(\vec{x}, \vec{y}) = \max\left(0, \sum_{k=1}^n y_k A_{i,k} - \sum_{j=1}^n \sum_{k=1}^n x_j y_k A_{j,k}\right)$$

$$y_i' = \frac{y_i + d_i(\vec{x}, \vec{y})}{1 + \sum_{j=1}^n d_j(\vec{x}, \vec{y})} \text{ with } d_i(\vec{x}, \vec{y}) = \max\left(0, \sum_{j=1}^n x_j B_{j,i} - \sum_{j=1}^n \sum_{k=1}^n x_j y_k B_{j,k}\right).$$

A small technical point: the resulting vectors $\vec{x}'$ and $\vec{y}'$ may not have entries that sum to 1, so we have to "normalize" them by dividing by the sum of all the entries.

Here $c_i(\vec{x}, \vec{y})$ is the amount we want to increase x_i by in order to make the payoff increase for player A. (This is why we take the maximum of 0 and the other value: if increasing x_i actually makes the payoff decrease, we set $c_i = 0$.) The denominators of the fractions are for "normalization" (see note in margin). It is a fact that the composition of continuous functions is continuous, so since all the pieces of f are continuous, so is f. □

As a concluding remark, note that the prisoner's dilemma has an interesting property: even though each player is doing what is best for themselves, the resulting strategy for both players is not actually the best one. Instead of both getting 2 years in prison, they could have both gotten just 1 year in prison by both staying silent. But that is not a Nash equilibrium. So a Nash equilibrium is not necessarily "the best strategy for both players." Rather, it is something closer to "the best strategy for each individual player when each he/she acts selfishly to maximize only his/her payoff." Thus the concept of Nash equilibrium is useful in game theory for studying cooperation.

3.4 Borsuk–Ulam theorem

Next we will use topological degree to prove the Borsuk–Ulam theorem.

Theorem 3.17 (Borsuk–Ulam)**.** *If f is a continuous function from S^2 to $\mathbb{R}^2$, then there exists a point x on the sphere S^2 such that $f(x) = f(-x)$.*

Here is a more intuitive way to see what the Borsuk–Ulam theorem says. Imagine the surface of the Earth; this is a two-dimensional sphere S^2. At each point on the surface, we can measure two values: temperature T, and pressure P. Together, (T, P) is a point in $\mathbb{R}^2$. So we have a continuous function that, for each point on the surface of the Earth, gives the temperature and pressure at that point. This is a continuous function from S^2 to $\mathbb{R}^2$. The Borsuk–Ulam theorem then says that

> At any moment, there exist two points on the surface of the Earth, directly opposite each other, where the temperature and pressure are the same.

This is somewhat non-intuitive! We will see, however, that the same ideas that go into proving Brouwer's fixed point theorem can also be used to prove the Borsuk–Ulam theorem. In fact, the proof technique will be very similar.

Proof of Borsuk–Ulam. As with the proof of Brouwer's fixed point theorem, here is an outline of the proof of Borsuk–Ulam.

1. Suppose that there exists a function $f\colon S^2 \to \mathbb{R}^2$ such that $f(x) \neq f(-x)$ for every x on S^2.
2. Use it to construct a function $F\colon S^1 \to S^1$ such that $F(-x) = -F(x)$ for every x on S^1.
3. Prove, using topological degree, that such a function F cannot exist.
4. Since this is a contradiction, our initial assumption that the function f exists must be false. Then we are done the proof.

Okay, let's do it!

Suppose that there exists a function $f\colon S^2 \to \mathbb{R}^2$ such that $f(x) \neq f(-x)$ for every x on S^2. Then let

$$h(x) = \frac{f(x) - f(-x)}{\|f(x) - f(-x)\|}.$$

Note that we really require $f(x) \neq f(-x)$ for every x, so that the denominator of $h(x)$ is never zero; we can't divide by zero.

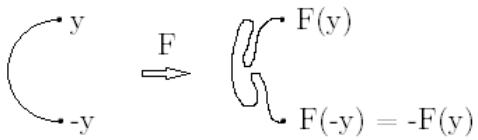
Note that $h(x)$ has two important properties:

1. the value of $h(x)$ for any x is on the circle S^1;
2. $h(-x) = -h(x)$, since

$$h(-x) = \frac{f(-x) - f(x)}{\|f(-x) - f(x)\|} = -\frac{f(x) - f(-x)}{\|f(x) - f(-x)\|} = -h(x).$$

But we wanted a function $F\colon S^1 \to S^1$, and right now we only have a function $h\colon S^2 \to S^1$. This is easy to solve: just look at h only on the equator of S^2. The equator is just a circle S^1, so if we focus only on the equator, we obtain a p function $S^1 \to S^1$ which we will call F. Since $h(-x) = -h(x)$ is true everywhere on S^2, it is true in particular on the equator, so $F(-x) = -F(x)$ as well.

Now we show such a function $F\colon S^1 \to S^1$ satisfying $F(-x) = -F(x)$ cannot exist. First we show the degree of F must be odd. This means we have to interpret F as a loop, and see how many times it goes around the circle. Consider F on half the circle S^1, starting from y and ending at $-y$:



Since $F(-y) = -F(y)$, whatever the resulting path is, it ends halfway across the circle. But also because $F(-x) = -F(x)$, the other half of the loop, starting from $F(-y)$ and ending at $F(y)$, is just a mirror image of the first half of the loop. Hence F, as a loop, goes around the circle an odd number of times, i.e. it has odd degree. However, we can also apply lemma 3.11 to show F has degree 0. This is because we can look at $h\colon S^2 \to S^1$ only on the upper hemisphere of the sphere S^2, so that it becomes a function $g\colon D^2 \to S^1$. The equator is part of the upper hemisphere (and also the lower hemisphere, for that matter), so $F\colon S^1 \to S^1$ actually arises as a little part of the function $g\colon D^2 \to S^1$. Then the lemma shows us that the degree of F is zero. Hence the degree of F is both odd and zero at the same time.

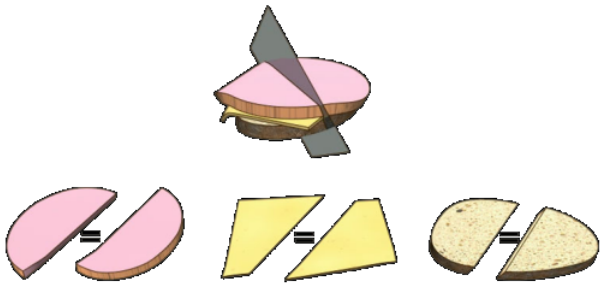
We have reached a contradiction. So our initial assumption that f exists must be incorrect. Hence for every function $f\colon S^2 \to S^1$, there exists a point x on S^2 such that $f(x) = f(-x)$. □

There are some interesting consequences of the Borsuk–Ulam theorem once we write down its most general form, in n dimensions (instead of in 2). Here is one such consequence, in two different flavors.

Consequences of theorems are called "corollaries."

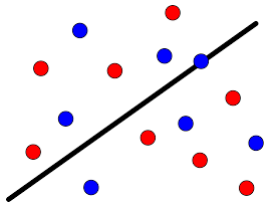
Corollary 3.18 (Ham sandwich theorem)**.** *Given* 3 *finite-volume (closed) subsets of* $\mathbb{R}^3$*, there exists a plane dividing the volume of the three pieces exactly in half simultaneously.*

Here, "closed" is a very technical restriction that we are going to ignore.



This is called the ham sandwich theorem because we can imagine two of the subsets being pieces of bread, and the other subset being a slice of ham. Then the theorem says there always exists a way we can slice the sandwich in half, so that there is exactly the same amount of bread and ham in the two pieces.

Corollary 3.19 (Discrete ham sandwich theorem)**.** *Given any number of red dots and blue dots arranged in any way in the plane, there exists a line dividing the plane into two regions such that the number of red dots on each side is equal, and the number of blue dots on each side is equal.*



It turns out, actually, that Brouwer's fixed point theorem is also a consequence of the Borsuk–Ulam theorem: that's how powerful the Borsuk–Ulam theorem is!

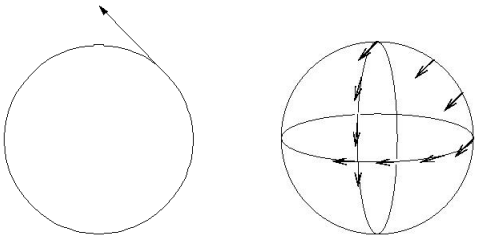
3.5 Poincaré–Hopf theorem

Let's think about the surface of the Earth some more. The Borsuk–Ulam theorem shows that there always exist antipodal points on the surface of the Earth with the same temperature and pressure. Temperature and pressure are both functions that assign a *number* to each point on the surface of the Earth. What if we wanted to model something like wind? We would need a function that assigns a direction to each point. More precisely, we want to assign a *vector* to each point, so that we know both the direction and the magnitude of the wind.

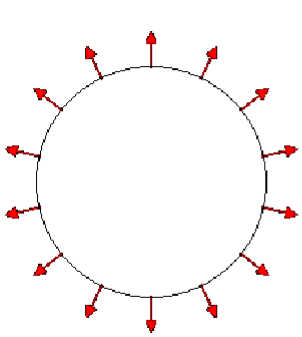
Two points on the sphere are "antipodal" if they are directly opposite each other.

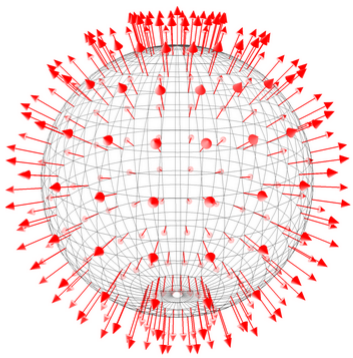
Definition 3.20. Let M be a manifold. A vector is **tangent** to the manifold M, if the direction in which it points is a direction in which you can move and stay in M. A **vector field** on a manifold M is a function that assigns a tangent vector to each point in M in a continuous fashion. (In other words, given a point x in M, the vectors at points around x should not be too different from the vector at x.)

Example 3.21. Here are examples of tangent vectors on the circle S^1 and the sphere S^2:



Here are examples of non-tangent vectors on the circle S^1 and the sphere S^2:





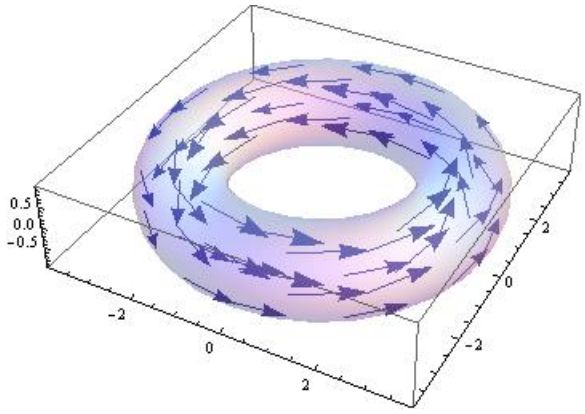
A vector field on the sphere S^2 therefore models wind on the surface of the Earth: at each point, there is a vector telling us which direction the wind is blowing at that point, and how strongly it is blowing. So one question we can ask is:

> on a given manifold M, does there exist a vector field that is non-zero everywhere?

If our vector field represents wind, then this question is: can there be wind everywhere on the surface of the Earth at the same time?

Of course, the important (and realistic) assumption is that wind is continuous.

Example 3.22. There does exist a vector field on the torus T^2 that is non-zero everywhere, shown (somewhat crudely) by the following diagram.



What about the sphere S^2? After a few tries, it may start to seem impossible to draw a vector field on S^2 that is non-zero everywhere. It turns out that it is indeed impossible. So there exists a vector field that is non-zero everywhere on the torus T^2, but there does not exist such a vector field on the sphere S^2. Whether or not such a vector field exists actually depends on the topology of the manifold!

Theorem 3.23 (Hairy ball theorem)**.** *Every vector field on the sphere S^2 is zero somewhere.*

This is called the "hairy ball theorem" because we can imagine a ball with lots of hair on its surface. If we can comb all the hair flat, then each hair can be interpreted as a tangent vector, so we would have a vector field that is non-zero everywhere. So sometimes people state the hairy ball theorem as "you can't comb a hairy ball flat."

Proof sketch. Since we have already seen two very long proofs (of Brouwer's fixed point theorem and of the Borsuk–Ulam theorem), let's just go through the outline of this proof.

1. Suppose there exists a vector field on the sphere that is non-zero everywhere.
2. Define a function $f\colon S^2 \to S^2$ which takes a point x and sends it to the point y given by moving a little bit, from x in the direction of the vector at x.
3. Show that f is homotopic to the identity function.
4. Since the vector field is non-zero everywhere, $f(x) \neq x$ for every x, i.e. f has no fixed points. Use this to construct a homotopy from f to the antipodal function.
5. Use topological degree to show this is a contradiction: the degree of the identity function is 1, but the degree of the antipodal function is -1. □

Corollary 3.24. *At any moment in time, there is a point on the surface of the Earth with no wind.*

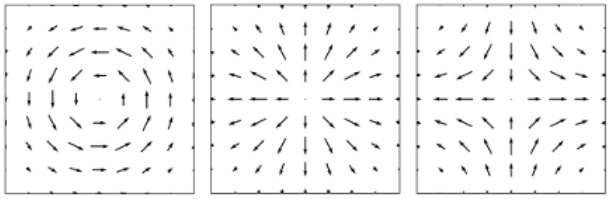
Actually, the hairy ball theorem is a special case of a much more general theorem. To state this more general theorem, we need to look more carefully at the points where the vector field is zero; it turns out there are "different ways" in which the vector field can be zero.

Definition 3.25. Consider a vector field V on a manifold M. Suppose that at p, the vector field is zero. The **index of V at p** can be calculated as follows:

1. draw a small circle around p, so that at each point of the circle is a vector;
2. count the number of times N the vector "turns" as we move counterclockwise around the circle;
3. if the vector turned N times counterclockwise, the index is N; otherwise if it turned N times clockwise, the index is $-N$.

In fact, this definition works even if the vector field is not zero at the point p. But it turns out that for such p (where the vector field is non-zero), the index is always 0, so we generally don't care about these points.

Example 3.26. The indices of the following vector fields are 1, 1, and -1.



Theorem 3.27 (Poincaré–Hopf)**.** *Let V be a vector field on M. The sum of the indices at every point where V is zero is equal to the Euler characteristic $\chi(M)$.*

This is usually written as $\sum_p \operatorname{ind}_p(V) = \chi(M)$.

Corollary 3.28 (Hairy ball theorem)**.** *Every vector field on the sphere S^2 is zero somewhere.*

Proof. Assume V is a vector field on S^2 which is non-zero everywhere, then the sum specified in the Poincaré–Hopf theorem is zero. So $\chi(S^2) = 0$. This is a contradiction, because we know $\chi(S^2) = 2$. So such a vector field V cannot exist: it must be zero somewhere. □

But Poincaré–Hopf can give us more than just the hairy ball theorem. By the same argument as above, we have proved the following.

Corollary 3.29. *If $\chi(M) \neq 0$, then every vector field on M is zero somewhere.*

In particular, let's think about surfaces. By the classification theorem for surfaces, there are only two surfaces with Euler characteristic zero: the torus T^2, and the Klein bottle Kl^2. Vector fields on every other surface must all be zero somewhere.

Exercise. Find a vector field on the Klein bottle that is non-zero everywhere.

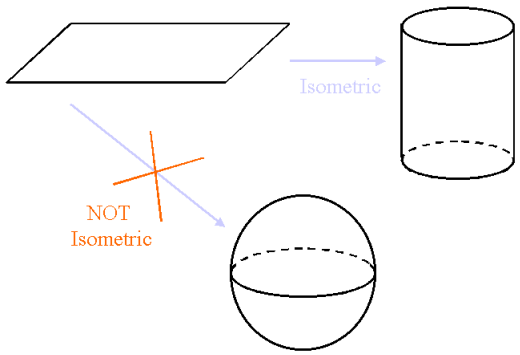
4 Geometry

We are now going to upgrade the idea of "homeomorphism." Remember that two spaces are homeomorphic if one can be deformed into the other. However, this process of deformation may change the *distance* between two points in the space, e.g. a big sphere is homeomorphic to a smaller sphere, but the distance between any two given points changes.

> Sometimes we care about the distances between points, and don't want the distance to change when we consider "equivalent" spaces. So we need a stricter notion of whether two spaces are "equivalent."

Definition 4.1. In the field of **geometry**, two spaces are equivalent if they are homeomorphic in a way that preserves distances. Such a homeomorphism is called an **isometry**.

Example 4.2. Let's look at isometries of the two-dimensional plane $\mathbb{R}^2$. The best way to visualize whether a deformation is an isometry for $\mathbb{R}^2$ is to imagine $\mathbb{R}^2$ as a flat piece of paper. Then any transformation we can imagine applying to the paper must preserve the distances between points, because the paper itself can't stretch or shrink.



So rolling up the paper to get a cylinder $S^1 \times \mathbb{R}$ is an isometry, but intuitively it does not seem possible to get a sphere S^2.

Although we are calling it a cylinder, there is no actual gluing happening; we are not allowed to glue for isometries. Similarly, the sphere is not actually a sphere.

Before we continue, note that distance is not *intrinsic* to a space. If I give you an arbitrary space by telling you what the points in the space are, what is the distance between two given points? This information is not intrinsically contained in the space. Instead, there are generally two ways to specify distances on a space X.

1. The first way is to embed X into $\mathbb{R}^n$ for some n, and then use the usual idea of distance on $\mathbb{R}^n$ to measure distances on X. The **distance** between two points p and q on X is the length of the shortest curve in X from p to q. For example, the distance between the north and south poles on a sphere S^2 of radius 1 is exactly π.

2. The second way is to specify a **distance function** $d(p, q)$ on X that explicitly tells you the distance from point p to point q for any two points p and q. Such a distance function must satisfy some properties:

   (a) $d(p, q) \geq 0$ for any points p, q, and equality holds if and only if $p = q$ ("the distance between any two points cannot be negative, and is zero only when the two points are the same");

   (b) $d(p, q) = d(q, p)$ ("the distance from p to q is the same as the distance from q to p");

   (c) $d(p, r) \leq d(p, q) + d(q, r)$ ("the distance from p to r is at most the distance from p to an intermediate point q plus the additional distance from q to r").

Actually, for a manifold, it is more appropriate to specify an "infinitesimal distance function" called a *metric* that acts like a dot product on tangent vectors. Then the length of a path is the integral along the path of the lengths of the tangent vectors of the path, and the distance is the length of the shortest path.

We will stick with the first method, which is much simpler to work with.

4.1 Geodesics

From now on, M is a manifold embedded into $\mathbb{R}^N$. So the distance between two points p and q on M is the length of the shortest curve in M connecting p and q. However, such a shortest path is often very hard to find, and very hard

to work with. Imagine if you were an ant on a surface. You don't know the shortest path to a point a million miles away, but at least you can try to walk "in a straight line" toward it.

Definition 4.3. A **geodesic** on M is a curve on M that is *locally* distance-minimizing. In other words, if we zoom in far enough, the geodesic is the shortest path between two points (which are close to each other).

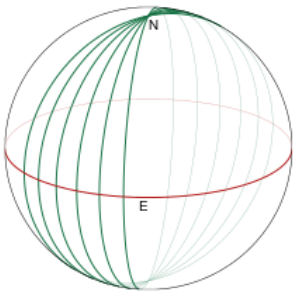
This definition illustrates the main philosophy behind doing things with manifolds:

> Because manifolds locally look like $\mathbb{R}^n$, and we understand $\mathbb{R}^n$ very well, we try to define objects and concepts only locally as much as possible.

A good way to imagine geodesics on a surface is to imagine taking a rubber band and nailing it to the surface at p and q. The path it shrinks to is always a geodesic.

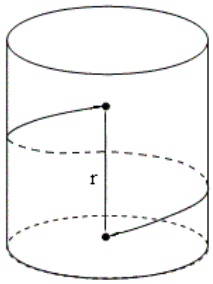
Example 4.4. Geodesics on $\mathbb{R}^n$ are just straight lines. So a good way to think about geodesics is that they are generalizations to arbitrary manifolds of the idea of a straight line in $\mathbb{R}^n$.

Example 4.5. Geodesics on the sphere S^2 are **great circles**.



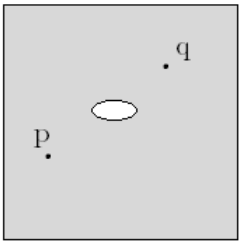
This is why airplanes fly in "curved" paths around the Earth. Though on a flat map they may seem like longer paths, these "curved" paths are actually the shortest paths!

Example 4.6. Here are two different kinds of geodesics on the cylinder $S^1 \times \mathbb{R}$:



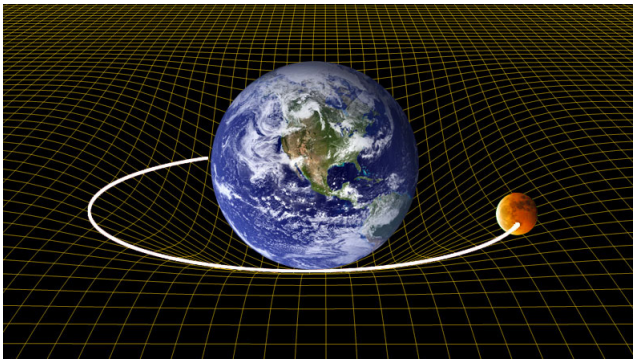
Note that the one that goes around the cylinder is indeed locally distance-minimizing, but is clearly longer than the distance r between the two points. The straight line geodesic between the two points is the shortest path.

Example 4.7. Note that geodesics may not always exist! Imagine the plane $\mathbb{R}^2$ with a disk D^2 removed:



Then there is no geodesic between p and q, since any locally distance-minimizing path must cross the hole.

In general relativity, geodesics are very important: the paths that objects take through spacetime are always geodesics (as long as no external forces other than gravity act on them), with the distance function determined by gravity.



You may have seen pictures of the "spacetime fabric": these are visualizations of what spacetime would look like if it were embedded into $\mathbb{R}^n$. The gravity of massive objects like planets and stars "curve" the spacetime around them. Objects orbit other objects because in those curved regions of spacetime, orbits are the geodesics, not straight lines.

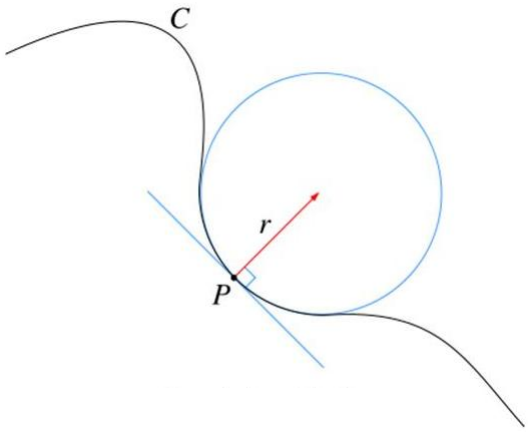
4.2 Gaussian curvature

From what we have seen so far, the geometry of a space is affected by how "curved" it is, while in topology it really didn't matter, because all the curves could be made flat via deformations. So it makes sense that the idea of "curvature" is a big deal in geometry. In fact, we will see soon that curvature is an isometry invariant, so it helps us distinguish non-isometric manifolds.

We want to somehow assign a number to each point of a manifold that says how curved the manifold is at that point. There are actually many ways this can be done. We will see one of them, for surfaces only. From now on, M is an oriented surface in $\mathbb{R}^3$.

Definition 4.8. To define curvature for surfaces, we must first define curvature for curves, i.e. 1-dimensional manifolds, say inside $\mathbb{R}^3$. Let C be a curve and p be a point on C. The **curvature** of C at p calculated using the following steps.

1. Draw the **osculating circle** of C at p: this is the circle which is tangent to the curve at p, and which "hugs" the curve the tightest.



   If the curve is a line, then the osculating circle has "infinite" radius.

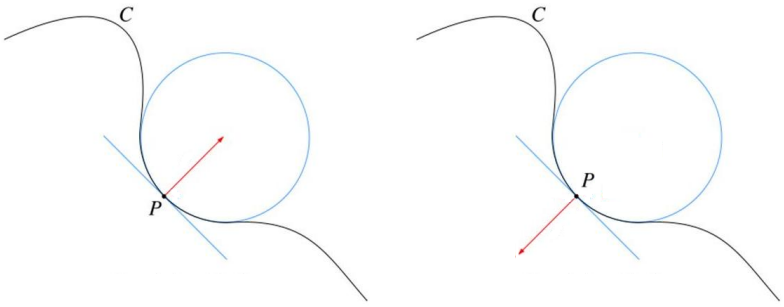
2. Let r be the radius of the osculating circle. The curvature is $1/r$. (If the osculating circle has "infinite" radius and the curvature is 0.)

Example 4.9. Let M be a circle with radius r. Then the curvature of M at any point p is just $1/r$, because the osculating circle at p is precisely M itself and therefore has radius r.

Definition 4.10. Let p be a point on M. The **Gaussian curvature** of M at p is calculated using the following steps.

1. Pick a **normal vector** at every point close to p. A normal vector at the point p is a vector perpendicular to the surface at p; more precisely, it is perpendicular to every tangent vector at the point p.

2. Consider a cross-section of the surface M that contains p and the normal vector at p. In this cross-section, M looks like a curve. Let κ be the curvature of this curve.



The **normal curvature** of M with respect to this cross-section is κ if the normal vector points "into" the osculating circle, and $-\kappa$ otherwise.

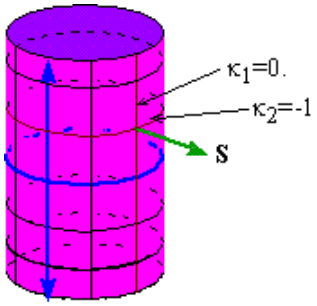
3. Let κ_1 and κ_2 be the maximal and minimal normal curvatures at p, across all possible cross-sections.

4. The Gaussian curvature is the product $K = \kappa_1\kappa_2$.

Example 4.11. The Gaussian curvature K of a sphere S^2 with radius r is $1/r^2$. This is because every normal curvature at any point is $\pm 1/r$, because every cross-section is a (portion of a) circle of radius r. In particular, $K > 0$ for any sphere.

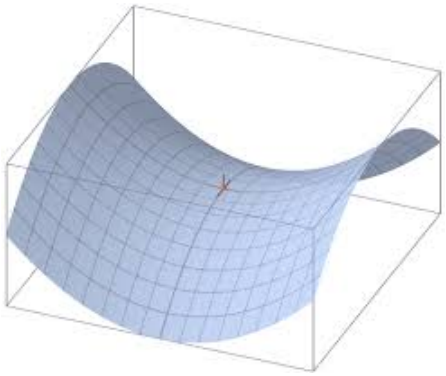
Example 4.12. The Gaussian curvature of the plane $\mathbb{R}^2$ is 0. This is because every normal curvature at any point is 0, because every cross-section is a straight line. So $K = 0$.

Example 4.13. The Gaussian curvature of the cylinder $S^1 \times \mathbb{R}$ is also 0.



This is because at any point, the maximal normal curvature $\pm 1/r$ is given by the cross section along S^1, and the minimal normal curvature 0 is given by the cross section along $\mathbb{R}$. Hence $K = (\pm 1/r) \cdot 0 = 0$.

Example 4.14. The Gaussian curvature of the following surface at the marked point is negative.



This is because regardless of which direction the chosen normal vector points, the maximal normal curvature is positive, and the minimal normal curvature is negative. So $K < 0$.

Note that the cylinder and the plane have the same Gaussian curvature, and we know they are isometric (by "rolling up" the plane). This is some evidence for the following theorem.

Theorem 4.15 (Gauss's Theorema Egregium)**.** *Gaussian curvature is invariant under isometries, i.e. two isometric surfaces always have the same Gaussian curvature.*

"Theorema egregium" literally means "remarkable theorem" in Latin.

In particular, the theorem implies that the sphere is not isometric to the plane, just as we suspected. This means a piece of paper can't be made into a sphere without creasing it or ripping it somewhere.

> Equivalently, a sphere can't be somehow "flattened" without distorting some part of its surface. This is why maps of the Earth never accurately represent distance everywhere.

Mapmakers have many different systems for approximating this "flattening" process so that only certain regions (e.g. near the north or south pole) are distorted.

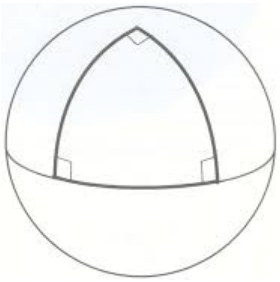
For example, in the Mercator projection system (shown above), areas around the poles seem much bigger than they should be; Greenland looks huge in comparison to Australia, even though in reality Australia is much bigger.

4.3 Gauss–Bonnet theorem

It is amazing that the Gaussian curvature is invariant under isometry, since it is not true that normal curvatures are invariant under isometry. It is even more amazing that the Gaussian curvature, which is a number associated to the geometry of the surface, is directly connected to the Euler characteristic, which is a number associated to the topology of the surface. This connection arises by looking at the "average value" of the Gaussian curvature over a polygon on a surface.

Definition 4.16. A polygon in the two-dimensional plane $\mathbb{R}^2$ is just a region bounded by straight lines. For a general orientable surface S, a **polygon** in S is a region bounded by geodesics, which are supposed to be the generalizations of straight lines to arbitrary manifolds.

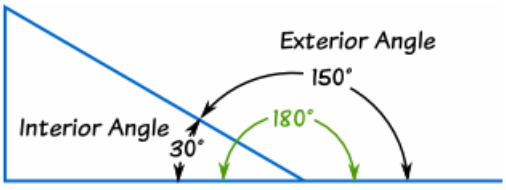
Example 4.17. What is a triangle on a sphere? Here is one example.



Note an interesting property of this triangle: each of its internal angles is $\pi/2$, so the sum of its internal angles is actually $3\pi/2$.

On the two-dimensional plane $\mathbb{R}^2$, we know from basic Euclidean geometry that the sum of internal angles for any triangle is π. The reason we can get more than π on the sphere seems to be because the sphere is curved. So perhaps we can find a relationship between curvature and the sum of some internal angles.

Definition 4.18. Given a polygon P, the **exterior angle** at a vertex of P is π minus the interior angle at P.



Theorem 4.19 (Local Gauss–Bonnet formula)**.** *Let S be a (compact) orientable surface, and $K(x)$ be the Gaussian curvature at the point x in S. If P is a polygon in S, then*

$$\int_P K(x)\,dx + (\textit{sum of exterior angles of } P) = 2\pi\chi(P).$$

The technical condition "compact" essentially just means "finite surface area."

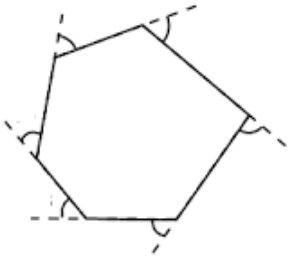
Example 4.20. Let's check the Gauss–Bonnet formula for polygons on the plane. Let $\mathbb{R}^2$ be the two-dimensional plane, and let P be an n-sided polygon on it. We will compute each of the terms in the formula.

1. (The $\int_P K(x)\,dx$ term) The normal curvatures at every point of the plane are 0, since every cross-section is just a straight line. Hence $K(x) = 0$ for any point on $\mathbb{R}^2$. So $\int_P K(x)\,dx = 0$.
2. (The (sum of exterior angles of P) term) We will leave this term as it is.
3. (The $2\pi\chi(P)$ term) An n-sided polygon has n vertices, n edges, and 1 face, so $2\pi\chi(P) = 2\pi(n - n + 1) = 2\pi$.

Hence the Gauss–Bonnet formula simplifies to

$$(\text{sum of exterior angles of } P) = 2\pi.$$

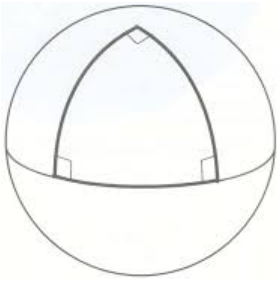
This is a true fact about polygons in the plane.



At each vertex, the interior angle plus the exterior angle is exactly π. So a consequence of this formula is

$$\begin{aligned}(\text{sum of interior angles of } P) &= n\pi - (\text{sum of exterior angles of } P)\\ &= (n-2)\pi.\end{aligned}$$

Example 4.21. Let's check the Gauss–Bonnet formula for triangles on spheres. Let S^2 be the sphere of radius 1, and let P be the triangle in the diagram below.



We will compute each of the terms in the formula.

1. (The $\int_P K(x)\,dx$ term) The normal curvatures at every point of the sphere are 1, since the osculating circle of every cross-section is of radius exactly 1. Hence $K(x) = 1$ for every x. In particular,

$$\int_P K(x)\,dx = \int_P 1\,dx = \text{area}(P).$$

   What is the area of P? Well, 8 copies of P would cover up the entire surface of the sphere, so the area of P must be 1/8-th of the surface area of the sphere, which is 4π. Hence $\int_P K(x)\,dx = \pi/2$.

2. (The (sum of exterior angles of P) term) Clearly the sum of exterior angles of P is $3\pi/2$.

3. (The Euler characteristic of P) The triangle has 3 vertices, 3 edges, and 1 face, giving $\chi(P) = 3 - 3 + 1 = 1$. Hence $2\pi\chi(P) = 2\pi$.

Since $\pi/2 + 3\pi/2 = 2\pi$, the Gauss–Bonnet formula works in this case!

In fact, this example shows we can use the Gauss–Bonnet formula to calculate areas of triangles on the sphere S^2: we know $2\pi\chi(P) = 2\pi$ for any triangle P, so as long as we know the sum of exterior angles of the triangle, we get

$$\text{area}(P) = \int_P 1\,dx = \int_P K(x)\,dx = 2\pi - (\text{sum of exterior angles of } P).$$

Theorem 4.22 (Global Gauss–Bonnet formula)**.** *Let S be a (compact) orientable surface, and $K(x)$ be the Gaussian curvature at the point x in S. Then*

$$\int_S K(x)\,dx = 2\pi\chi(S).$$

Note that this global Gauss–Bonnet formula isn't anything new: just imagine taking the polygon P to be the entire surface S, so that there are no exterior angles to sum. The result of the local Gauss–Bonnet formula in this case is precisely the global Gauss–Bonnet formula.

Example 4.23. Let S^2 be the sphere of radius r. Then $K(x) = 1/r^2$, so the Gauss–Bonnet formula simplifies to

$$\int_S \frac{1}{r^2}\,dx = \frac{1}{r^2}\text{area}(S) = 2\pi(2) = 4\pi.$$

Indeed, the surface area of a sphere of radius r is $4\pi r^2$.

Example 4.24. Since the torus T^2 has Euler characteristic 0, the global Gauss–Bonnet formula says $\int_{T^2} K(x)\,dx = 0$.

Rough proof of global Gauss–Bonnet formula. Let's see a brief proof outline for the global Gauss–Bonnet formula. The idea is to approximate the surface using finer and finer triangular meshes, i.e. triangulations of the surface. This is essentially what we did to prove that Euler characteristic is unchanged by homeomorphisms.

Draw a triangular mesh on the surface S and "flatten" each of the triangles. Then the resulting surface is homeomorphic but probably not isometric to the original surface. Let $\delta(v)$ be the "angle defect" at the vertex v in the mesh, i.e. 2π minus the sum of the angles around the vertex v. We are interested in the quantity $\sum_v \delta(v)$, the sum of all the angle defects for all the vertices of the mesh. It can be shown that as the mesh gets finer and finer, the sum $\sum_v \delta(v)$ gets closer and closer to $\int_S K(x)\,dx$. In particular, as the mesh gets finer and finer, it approximates the surface S better and better.

In other words, the *limit* of $\sum_v \delta(v)$ as the mesh becomes small is $\int_S K(x)\,dx$.

On the other hand, we can compute $\sum_v \delta(v)$ in terms of the number of vertices V, the number of edges E, and the number of faces F. Note that since we "flattened" each of the triangles in the mesh, their internal angles add up to π. So the sum of all the angles at every vertex in the mesh is just the sum of all the internal angles of every face, i.e. πF. But we can rewrite the total angle defect as

$$\begin{aligned}\sum_v \delta(v) &= \sum_v (2\pi - \text{actual sum of angles around } v) \\ &= 2\pi V - \sum_v (\text{actual sum of angles around } v) \\ &= 2\pi V - \pi F.\end{aligned}$$

We will now relate this expression $2\pi V - \pi F$ to the Euler characteristic $\chi(S) = V - E + F$. Let's try to get a relationship between the number of edges E and the number of faces F. Note that every face is bounded by three edges. So we want to say that the quantity $3F$ is exactly the number of edges. But this is not correct: each edge is involved in exactly two faces, so each edge is counted exactly twice. Hence $3F = 2E$. Rearranging, $F = 2(E - F)$. Hence

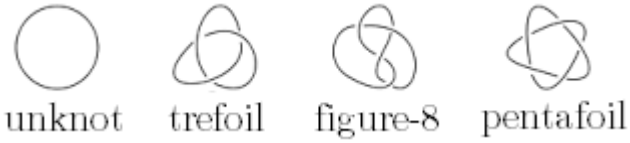
$$\sum_v \delta(v) = 2\pi V - \pi F = 2\pi V - 2\pi(E - F) = 2\pi(V - E + F) = 2\pi\chi(S). \quad \square$$

5 Knot theory

We have spent a lot of time developing theory. Now let's look at an interesting application of ideas from topology to "real world objects": knots. Knots may not seem like important or fundamental mathematical objects, like surfaces are, but in the last few decades they have appeared in many unexpected places, e.g. the study of quantum physics.

Definition 5.1. A **knot** should be thought of as a piece of string in $\mathbb{R}^3$ with its two ends attached together. Two knots are **equivalent** if we can change one

into the other without breaking the string or moving the string through itself. To draw knots, we use **knot diagrams**, like the following.



A **crossing** in a knot diagram is when two segments of the string cross over one another. For clarity, the rule when drawing knot diagrams is that every crossing can involve only *two* segments and no more.

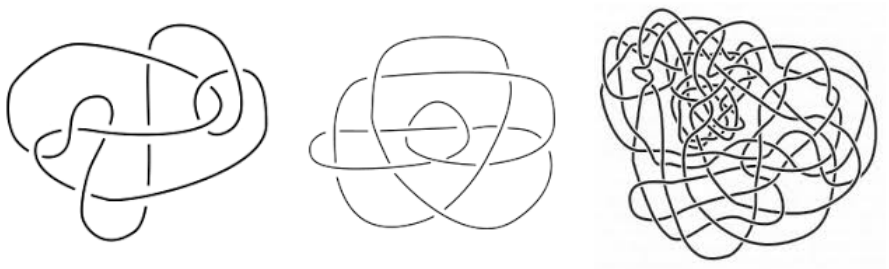
Formal definition. A **knot** is a (smooth) embedding of the circle S^1 into $\mathbb{R}^3$. Two knots are **equivalent** if there is a homotopy $h\colon S^1 \times [0,1] \to \mathbb{R}^3$ from one knot to the other inside $\mathbb{R}^3$ such that for every fixed $t \in [0,1]$, the function $f(x) = h(x,t)$ is a knot. (We will define homotopy later on!)

This restriction on the homotopy makes it an *isotopy*. It makes sure that the homotopy between the knots does not make the string pass through itself.

We are in a familiar situation now: we have defined a new type of mathematical object and a notion of "equivalence" for these objects. The natural thing to ask now is:

> Can we classify knots? That is, is there a good way to determine when two knots are equivalent?

In particular, if we can classify knots, we can efficiently determine whether a given knot is equivalent to the unknot, i.e. whether it can be untied! This seems hard. For example, which of the following knots are equivalent to the unknot?



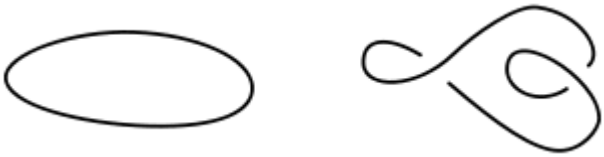
5.1 Invariants and Reidemeister moves

Whenever we are faced with a classification problem, the first instinct should be to find invariants, i.e. things that assign a number to each knot such that two equivalent knots get the same value.

Example 5.2. The **unknotting number** of a knot is the minimum number of times that the string must pass through itself in order to get the unknot. For example, it is fairly easy to see that the unknotting number of the trefoil is 1:

This is a knot invariant. If we can "untie" a knot K_1 by passing the string through itself n times, then we can do the same procedure on an equivalent knot K_2. The key here is that to get from a knot to an equivalent knot, we cannot ever change the unknotting number, by the definition of equivalence for knots: the string cannot pass through itself.

Example 5.3. Suppose we take a knot diagram and count the number of crossings in it. This number is *not* a knot invariant. For example, the unknot has many possible knot diagrams, with different numbers of crossings. Here are two:

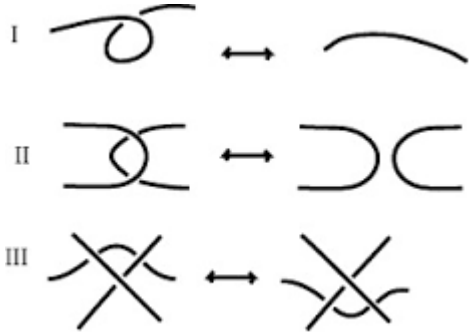


These have 0 and 2 crossings, respectively. However, if we look at the *minimum* number of crossings over all possible knot diagrams of a knot, then that is a knot invariant, called the **crossing number**. For example, the crossing number of the unknot is 0, and the crossing number of the trefoil is 3.

It is not hard to intuitively see that the crossing number of the trefoil is 3, but can you prove it?

The unknotting number and the crossing number are very hard to calculate, but it is easy to see that they are actually knot invariants. We would like a better invariant that we can easily calculate, like Euler characteristic for surfaces. In general, the trade-off is that it may be very hard to prove whether an easily-calculated quantity is actually an invariant. For knots, however, the following theorem, due to Reidemeister, gives three simple criteria for whether something is a knot invariant.

Theorem 5.4 (Reidemeister's theorem)**.** *Given two equivalent knot diagrams, it is always possible to go from one to the other via repeated applications of the following three* **Reidemeister moves***:*



It is clear that the Reidemeister moves produce equivalent knots, but showing that they are enough to get from a diagram to any other equivalent one is hard. We skip the proof.

Corollary 5.5. *To check whether a quantity is a knot invariant, it suffices to check that the quantity is unchanged by the three Reidemeister moves.*

5.2 The Jones polynomial

Even using Reidemeister's theorem, easily-calculated knot invariants are hard to find. In 1985, Vaughan Jones discovered the Jones polynomial, which was the start of a series of major advances in knot theory and physics. In this section, we'll see how to compute the Jones polynomial, and then we'll prove it is a knot invariant using Reidemeister's theorem.

Definition 5.6. The most direct way to define the Jones polynomial is to define a related object called the **bracket polynomial**. Given a knot K, the bracket polynomial is denoted $\langle K \rangle$, and is defined using the following three rules:

The bracket polynomial was discovered by Louis Kauffman in 1987, and hugely improved our understanding of the Jones polynomial.

1. $\left\langle \bigcirc \right\rangle = 1$
2. $\left\langle L \cup \bigcirc \right\rangle = (-A^2 - A^{-2})\langle L \rangle$
3. $\left\langle \times \right\rangle = A \left\langle \asymp \right\rangle + A^{-1} \left\langle) (\right\rangle$

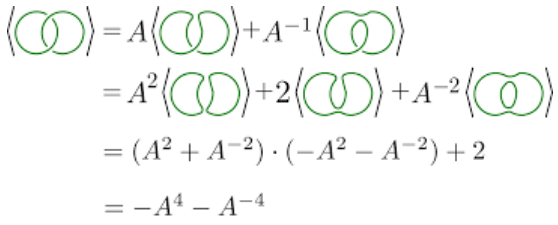
Let's go through what these rules mean. Keep in mind throughout that A is just a variable.

1. The first rule says that the bracket polynomial of the given knot diagram $\bigcirc$ of the unknot is just the value 1. Note that this does not mean the bracket polynomial of *any* knot diagram of the unknot is 1.

2. The expression $\langle L \cup \bigcirc \rangle$ means a knot diagram L with an extra unknot $\bigcirc$ that does not cross the rest of the diagram. The second rule then says we can find the bracket polynomial of L with this extra $\bigcirc$ by first finding the bracket polynomial of L, and then multiplying by $(-A^2 - A^{-2})$.

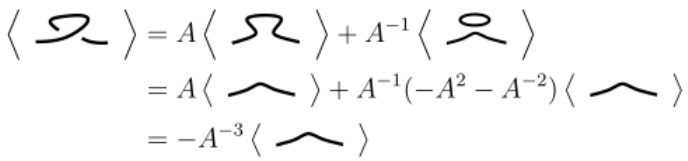
3. The third rule says that whenever we have a crossing, we can "get rid of it" by computing the bracket polynomials of the new knots obtained by the two different ways of replacing the crossing with non-intersecting segments. This is the key idea behind the bracket polynomial: each application of this rule removes one crossing, so because each knot diagram has only a finite number of crossings, eventually we get to some number of unknots $\bigcirc$ with no crossings.

Example 5.7. Here is a computation of the bracket polynomial for the **Hopf link**:

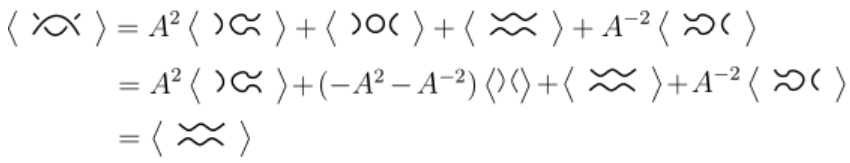


In order to determine whether the bracket polynomial is a knot invariant, we should see what happens to it under the three Reidemeister moves.

1. The bracket polynomial **is not** invariant under move I:



2. The bracket polynomial **is** invariant under move II:

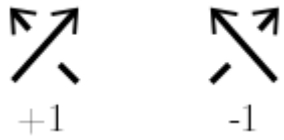


3. (Exercise) The bracket polynomial **is** invariant under move III.

So the bracket polynomial is almost a knot invariant; we just have to "fix" what happens to it under move I. The idea is to multiply extra factors into the bracket polynomial to compensate for the "extra" A^{-3} we get from one application of move I.

Definition 5.8. An **orientation** of a knot is a choice of direction to "go around" the knot. We indicate the orientation by an arrow.

When a knot has an orientation, we can distinguish between **positive** and **negative** crossings:



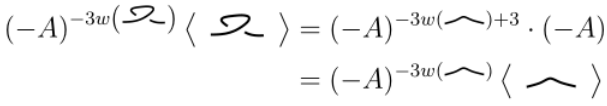
Let $n_+(D)$ and $n_-(D)$ be the number of positive and negative crossings, respectively, of a knot diagram D. The **writhe** of D is $w(D) = n_+(D) - n_-(D)$, their difference.

Note that for knot diagrams, we need an orientation to compute the writhe, but the writhe does not actually depend on which of the two possible orientations we pick.

Like the bracket polynomial, the writhe is unchanged by move II and move III. (Check this! This is a good exercise to make sure you understand writhe.) However, applying move I decreases the writhe by exactly 1. Hence we can use the writhe $w(D)$ to "cancel out" the extra factor of the bracket polynomial under move I.

Lemma 5.9. *Let K be a knot with knot diagram D. Then the polynomial $(-A)^{-3w(D)}\langle D\rangle$ is an invariant of the knot K.*

Proof. Both the writhe and the bracket polynomial are invariant under moves II and III, so it suffices to check move I:



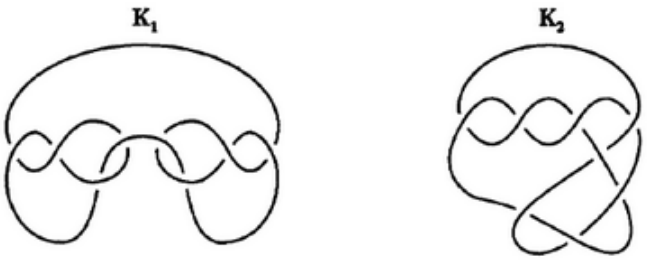
Hence the polynomial $(-A)^{3w(D)}\langle D\rangle$ is invariant under all three Reidemeister moves. By Reidemeister's theorem, it is a knot invariant. □

Definition 5.10. Let K be a knot with knot diagram D. If we take the polynomial $(-A)^{-3w(D)}\langle D\rangle$ and set $A = t^{-1/4}$, we get the **Jones polynomial** $V_K(t)$ of the knot K. (This substitution is not interesting; it is just to make sure our definition of the Jones polynomial via the bracket polynomial actually agrees with the original polynomial defined by Jones.)

Example 5.11. We continue with the Hopf link, whose bracket polynomial we computed earlier to be $-A^4 - A^{-4}$. Pick either orientation to see that the Hopf link has two negative crossings. Hence

$$V_K(t) = (-A)^{3\cdot(-2)}(-A^4 - A^{-4}) = -A^{10} - A^2 = -t^{5/2} - t^{1/2}.$$

However, we have to be a little careful. It is true that the Jones polynomial is an invariant, which means two equivalent knots have the same Jones polynomial. It is not necessarily true, though, that if two knots have the same Jones polynomial, then they are equivalent. The following two knots have the same Jones polynomial, but another invariant called the Alexander polynomial shows they are actually inequivalent:



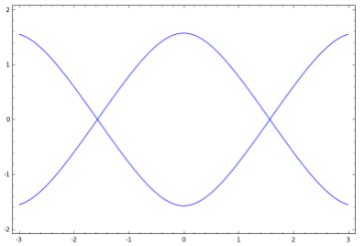
However, we can ask the following question: if the Jones polynomial of a knot K is 1, is K necessarily the unknot? The answer is still unknown, and attempts to create better invariants that fully distinguish the unknot from other knots have been very fruitful.

Many of these "better" invariants are "quantum knot invariants," which have interesting relationships with other areas of mathematics such as representation theory.

6 Algebraic geometry

So far, we have been working only with manifolds and objects defined on manifolds, and manifolds are always locally like $\mathbb{R}^n$. Now we will venture further and explore spaces which are not necessarily manifolds. The points at which they fail to be manifolds are called **singularities**, and spaces with such points are in general called **singular** spaces. For the sake of visualization, we will start off with curves which have singularities.

In particular, manifolds are "non-singular" spaces.



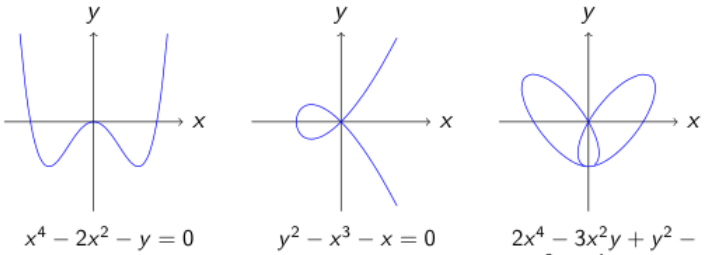
Here is the curve $\cos(y) = \sin(x)^2$. It is singular; in fact, it has infinitely many singularities, at each of the crossing points. However, it turns out that the theory of curves described by arbitrary equations is too hard, and we must restrict ourselves to only curves described by polynomials. The idea is as follows.

> To describe/define functions like cos, sin, exp, etc., we require infinitely many numbers to specify the coefficients of their power series expansions. To describe/define polynomials, we require only finitely many. In mathematics, finiteness often makes theory easier.

For example, to define $\cos(x) = 1 - x^2/2! + x^4/4! + \cdots$ as a function of x, we need to specify infinitely many coefficients $(1, 0, 1/2!, 0, 1/4!, \ldots)$ for the power series. But to define $f(x) = 1 + x + 2x^3$, it suffices to specify finitely many coefficients $(1, 1, 0, 2)$.

Definition 6.1. A **plane curve of degree** d is the zero set of a degree d polynomial $f(x, y)$ in two variables. The degree of $f(x, y)$ is written $\deg f$, and is the biggest total degree among all terms in f. For example, $\deg x^4y = 4+1 = 5$ and $\deg(y^4 - x^3y^3) = 3 + 3 = 6$.

Example 6.2. Some plane curves are non-singular, in which case they are manifolds of dimension 1; some are singular, in which case they are not.



In low degrees, we have names for plane curves.

1. A plane curve of degree 1 is a **line** $ax + by + c = 0$.
2. A plane curve of degree 2 is a **conic** $ax^2 + bxy + cy^2 + dx + ey + f = 0$.
3. Plane curves of degrees 3, 4, 5 are **cubics**, **quartics**, and **quintics** respectively.

In general, we can look at spaces described by any number of equations in any number of variables. Such spaces are called (affine) **algebraic varieties**. For example, the space given by $x^2 + y^2 + z^2 = 4$ is a sphere S^2 of radius 2, and the space given by

$$\begin{cases} (x_1 - 1)^2 + (x_2 - 1)^2 + (x_3 - 1)^2 + (x_4 - 1)^2 = 25 \\ (x_1 + 2)^2 + (x_2 + 2)^2 + (x_3 + 2)^2 + (x_4 + 2)^2 = 25 \end{cases}$$

is the intersection of two S^3's of radius 5, centered respectively at $(1, 1, 1, 1)$ and $(-2, -2, -2, -2)$. The theory and study of such spaces is called **algebraic geometry**.

Here the word algebraic means the equations are polynomials, as opposed to trig functions, for example. We call non-algebraic objects "transcendental".

Instead of delving into the machinery of algebraic geometry, i.e. the theory of algebraic varieties, we will instead develop theory to answer enumerative/counting questions regarding (plane) curves. For example, in order of difficulty:

1. how many points do two lines intersect at?
2. how many points do two plane curves of degree d and degree e intersect at?
3. how many points uniquely specify a conic, e.g. an ellipse?

4. how many lines intersect four given lines in $\mathbb{R}^3$?

5. how many circles are tangent to three given circles?

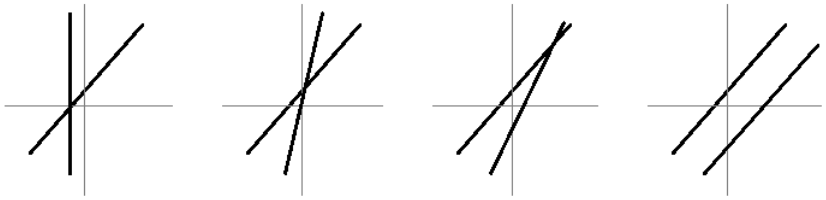
6. how many plane curves of degree d pass through $3d-1$ given points?

6.1 Projective geometry

We start with the first question: how many points do two lines intersect at? Clearly we have two cases: either the lines are parallel and don't intersect, or they intersect at exactly one point. This answer should be unsatisfying; why should there be two cases? We are going to force the answer to *always* be one.

This seems like an unnatural thing to want until we think about a related question: how many zeros does a quadratic ax^2+bx+c have? The naive answer is 0, 1, or 2, but of course if we extend the reals to the complex numbers $\mathbb{C}$ and account for double roots, then the correct statement is that a quadratic always has two zeros. The zeros may just be imaginary, or the apparently single zero actually has multiplicity two. The moral of the story is that quadratics do not behave nicely over $\mathbb{R}$, so we extended to $\mathbb{C}$.

How should we extend the plane $\mathbb{R}^2$ so that two lines in it always intersect at exactly one point? The better question is where two parallel lines intersect. If we approximate the situation with parallel lines as follows,



then the point of intersection is moving farther and farther away, off toward infinity. This suggests we should enlarge $\mathbb{R}^2$ by adding "points at infinity", one for each possible slope of two parallel lines.

Theorem 6.3. *This enlargement is exactly the projective plane $\mathbb{P}^2$.*

Proof sketch. The key step is to see how $\mathbb{R}^2$ fits inside $\mathbb{P}^2$; the rest will become clear. First recall that $\mathbb{P}^2$ is the disk D^2 with antipodal points on the boundary glued together. Say the disk has radius 1, i.e. it is the set of points

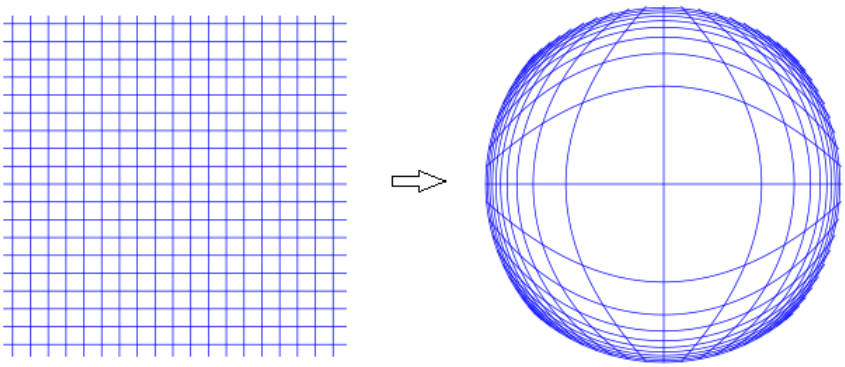
$$\{(x,y) : x^2+y^2 \leq 1\}.$$

Then in fact $\mathbb{R}^2$ is homeomorphic to the **open disk** in D^2, i.e.

$$\mathbb{R}^2 \cong \{(x,y) : x^2+y^2 < 1\}.$$

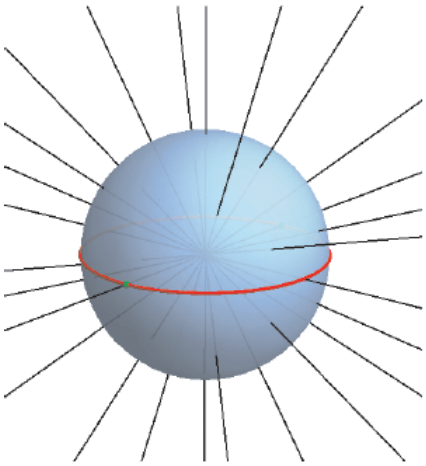
The idea is to shrink distances in $\mathbb{R}^2$ more and more the farther we get from the origin, as illustrated in the following diagram:

Formally, the homeomorphism is given by $\vec{x} \mapsto \vec{x}/(1+\|x\|)$. It is a good exercise to convince yourself that this is indeed a homeomorphism.



The open disk is just the disk D^2 without its boundary. If we add in the boundary, note that two parallel vertical lines intersect at the north and south poles, which are conveniently the same point after gluing antipodal points to get $\mathbb{P}^2$. Similarly, any two parallel lines will intersect at antipodal points on the boundary of the disk, and therefore at exactly one point in $\mathbb{P}^2$. □

To better understand $\mathbb{P}^2$, we want to put coordinates on it, just as coordinates (x, y) help us work with objects in $\mathbb{R}^2$. The way to do this is to realize that another description of $\mathbb{P}^2$ is the "space of all lines passing through the origin in $\mathbb{R}^3$".



Each line corresponds to a point in the upper hemisphere D^2, and antipodal points on the equator ∂D^2 are glued together. Hence a point in $\mathbb{P}^2$ can be thought of as a line in $\mathbb{R}^3$ passing through the origin.

Definition 6.4. A point in $\mathbb{P}^2$ has **homogeneous coordinates** $[x : y : z]$ if (x, y, z) lies on the line in $\mathbb{R}^3$ corresponding to the point in $\mathbb{P}^2$. In particular, this means that

$$[x : y : z] = [cx : cy : cz]$$

for any non-zero real number c.

The word "homogeneous" here refers to how if we scale x, we must scale y and z by the same amount.

Example 6.5. Recall that $\mathbb{P}^2$ is supposed to extend $\mathbb{R}^2$. If we start with a point (x, y) in $\mathbb{R}^2$ and carefully keep track of which point on the disk it is, and therefore which line corresponds to it, we see its homogeneous coordinates are $[x : y : 1]$. In general, there are three types of points in $\mathbb{P}^2$:

1. points of the form $[x : y : 1]$ for any real numbers x, y, which together form a copy of $\mathbb{R}^2$;
2. points of the form $[z : 1 : 0]$ for any real number z, which together form a copy of $\mathbb{R}^1$ on the boundary ∂D^2 "at infinity";
3. the point $[1 : 0 : 0]$, which is the remaining point on the boundary.

For example, $[-1 : 2 : 3] = [-1/3 : 2/3 : 1]$ is just a regular point in $\mathbb{R}^2$, while $[2 : \pi : 0] = [2/\pi : 1 : 0]$ is a point at infinity.

Exercise. Check that (x, y) in $\mathbb{R}^2$ corresponds to $[x : y : 1]$ in $\mathbb{P}^2$.

Note that since $\mathbb{P}^2$ extends the plane, lines in the plane naturally extend to lines in $\mathbb{P}^2$. For example, the line $y = 3x$ consists of points $(x, 3x)$, which in $\mathbb{P}^2$ are points

$$[x : 3x : 1] = [1/3 : 1 : 1/3x].$$

As x gets bigger, $1/3x$ tends toward zero, so in the limit at infinity we hit the point $[z : 1 : 0] = [1/3 : 1 : 0]$ in $\mathbb{P}^2$. Note that $1/z = 3$ is the slope of the line.

Exercise. Verify that the line $ax + by + c = 0$, once extended to $\mathbb{P}^2$, contains the point $[b/a : 1 : 0] = [b : a : 0]$ at infinity, which records the slope of the line. In particular,

1. the vertical line $x = 0$ which has "infinite slope" contains the point $[0 : 1 : 0]$, and
2. the horizontal line $y = 0$ which has zero slope contains the point $[1 : 0 : 0]$.

Note that $[b/a : 1 : 0]$ does not make sense when $a = 0$, but $[b : a : 0]$ does, so for $a = 0$ it only makes sense to write $[b : a : 0]$. Be careful when working in homogeneous coordinates!

We would like to extend the equation $ax+by+c = 0$ into an equation for the extended line in $\mathbb{P}^2$. It turns out there is a systematic way to extend equations in $\mathbb{R}^2$ to equations to $\mathbb{P}^2$ that works for all polynomials.

Definition 6.6. Given a polynomial $f(x, y) = cx^a y^b + \cdots$, its **homogenization** $f(x, y, z)$ is obtained by multiplying factors of z to each term until every term has the same degree.

Example 6.7. We homogenize some low degree polynomials.

1. The homogenization of $ax + by + c$ is $ax + by + cz$, because now each term has total degree 1.
2. The homogenization of $2x^2y - y + 3$ is $2x^2y - yz^2 + 3z^3$, because now each term has total degree 3.

Homogeneous polynomials are the correct notion of "polynomial" for homogeneous coordinates on $\mathbb{P}^2$. For example, the equation $y = x + 1$ does not even make sense on $\mathbb{P}^2$, because $[1:2:1] = [2:4:2]$ in $\mathbb{P}^2$ but the lhs satisfies the equation while the rhs does not. For homogeneous polynomials of degree d,

$$f(x, y, z) = 0 \text{ if and only if } f(cx, cy, cz) = c^d f(x, y, z) = 0.$$

From now on, when we write a polynomial $f(x, y)$, it refers to both the curve in $\mathbb{R}^2$ and the curve in $\mathbb{P}^2$ given by its homogenization $f(x, y, z)$.

Example 6.8. Let's examine conics in $\mathbb{P}^2$ by homogenizing them. In $\mathbb{R}^2$, there are three types of conics: parabolas, hyperbolas, and ellipses.

1. The parabola $y = x^2$ homogenizes to $yz = x^2$.
2. The hyperbola $xy = 1$ homogenizes to $xy = z^2$.
3. The circle $x^2 + y^2 = 1$ homogenizes to $x^2 + y^2 = z^2$, which is the same as $x^2 = (y - z)(y + z) = uv$ after the change of variables $u = y - z$ and $v = y + z$.

Here we only homogenize standard conics; any other parabola, hyperbola, or ellipse becomes a standard conic under some change of variables.

So we see that up to a change of variables, all conics are the same in $\mathbb{P}^2$, since they all look like $xy = z^2$.

6.2 Bézout's theorem

We have now realized that $\mathbb{P}^2$, not $\mathbb{R}^2$, is the correct setting to do **intersection theory**, e.g. study intersections of plane curves, just as the complex numbers are the correct setting to study zeros of polynomials. We also know now what plane curves in $\mathbb{P}^2$ are. So it is time to address the second problem on our list: how many points do two plane curves of degree d and e intersect at? When $d = e = 1$, these are two lines, and we have already established that the answer is 1.

Technically, we must work with complex numbers instead of real numbers to properly do intersection theory, and for most of the results in this section to be true. Concretely, this means we work in the *complex* projective plane $\mathbb{CP}^2$, i.e. homogeneous coordinates are complex numbers. But because we will not actually do any coordinate computations, this technical point is largely irrelevant for us right now.

Theorem 6.9 (Bézout's theorem for curves)**.** *Two plane curves in $\mathbb{P}^2$ of degrees d and e intersect at either infinitely many points (if they are the same curve), or at exactly de points (with multiplicity).*

In fact, Bézout's theorem is true more generally, in higher dimensions. In $\mathbb{P}^n$, the zero set of a homogeneous polynomial of degree-d is called a **degree-d hypersurface**, and Bézout's theorem works for hypersurfaces as well.

Theorem 6.10 (Bézout's theorem)**.** *A collection of n hypersurfaces in $\mathbb{P}^n$ of degrees $d_1, d_2, \ldots, d_n$ intersect at either infinitely many points, or at exactly $d_1 d_2 \cdots d_n$ points (with multiplicity).*

Proofs of Bézout's theorem involve some of the machinery of algebraic geometry, and we will not get into them here. Instead, Bézout's theorem will be our main tool for answering enumerative problems. We apply it to the third

problem on our list: how many points uniquely specify a conic? For conceptual clarity, let's reformulate this problem as: how many points do we have to specify so there is only a finite number of conics (hopefully exactly one) passing through all of them?

Lemma 6.11. *The collection of all conics is precisely the space $\mathbb{P}^5$, i.e. every conic corresponds to a point in $\mathbb{P}^5$.*

Proof. As with $\mathbb{P}^2$, we have homogeneous coordinates $[a : b : c : d : e : f]$ on $\mathbb{P}^5$. The correspondence we want is

$$ax^2 + bxy + cy^2 + dx + ey + f \leftrightarrow [a : b : c : d : e : f].$$

This is because if we rescale all the coefficients by the same factor, the conic defined by the equation does not change. For example, the point $[1 : 0 : 1 : 0 : 0 : -4]$ in $\mathbb{P}^5$ corresponds to the equation $x^2 + y^2 - 4 = 0$, which is the circle of radius 2 centered at $(0, 0)$. □

Let's suppose we want to look at the collection of conics that pass through the point $(2, 3)$. The coefficients a, b, c, d, e, f of such conics must therefore satisfy the equation

$$4a + 6b + 9c + 2d + 3e + f = 0,$$

by plugging $(x, y) = (2, 3)$ into the general equation $ax^2+bxy+cy^2+dx+ey+f = 0$. Hence the set of conics passing through $(2, 3)$ is a degree-1 hypersurface in $\mathbb{P}^5$, the space of all conics. Of course, this is not specific to the point $(2, 3)$; this argument works for any given point.

Theorem 6.12. *Given five points (in general position) in $\mathbb{P}^2$, there is exactly one conic passing through them.*

Proof. Call the five points $p_1, \ldots, p_5$. The set of conics passing through p_i is a degree-1 hypersurface in $\mathbb{P}^5$. The set of conics passing through all five points is the intersection of these five degree-1 hypersurfaces. So we want to know how many points are in the intersection. By Bézout's theorem, if we can prove the intersection does not contain infinitely many points, then it must actually contain exactly 1 point. This point in $\mathbb{P}^5$ corresponds to the unique conic passing through the five given points $p_1, \ldots, p_5$.

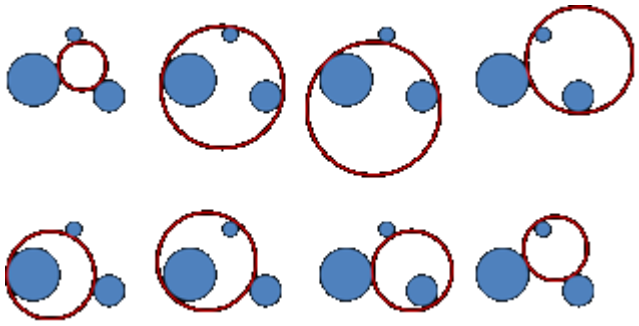
In general, checking that the desired intersection is finite is the hard part of applying Bézout's theorem. Generally, the condition that the given points are "in general position" ensures the intersection is finite. The phrase "in general position" means the points are not in some special configuration. In our case, we just need to avoid the case where four of them are collinear. □

For details on this, look up the term "transversal intersection" in relation to Bézout's theorem.

Exercise. Show that given four points and a line in $\mathbb{P}^2$, there are exactly two conics which pass through the points and are tangent to the line. Disregard issues with finiteness of intersections and general position. (Hint: show that the condition of "being tangent to a line" is a degree-2 hypersurface in the space of conics.)

Now we will look at a more complicated application of Bézout's theorem: the fifth problem in our list. This is a very classical problem dating from approximately 200BC.

Theorem 6.13 (Apollonius problem)**.** *Given three (generic) circles in the plane, there are exactly eight circles tangent all three.*



Proof. By now we know the general recipe:

1. identify a geometric space which is "the space of all circles in the plane";
2. formulate the condition "being tangent to a given circle" as some hypersurface;
3. apply Bézout's theorem to count the number of intersection points in the intersection of such hypersurfaces.

We already know the space of all conics is $\mathbb{P}^5$. What makes a conic a circle? Well, we require it to be of the form $(x-a)^2+(y-b)^2-r^2=0$, which after homogenization becomes

$$(x-az)^2+(y-bz)^2-r^2z^2=0.$$

From this, it is clear that every circle passes through $O_+=[1:i:0]$ and $O_-=[1:-i:0]$. (Here $i=\sqrt{-1}$ is the imaginary unit.) Conversely, if a conic passes through both O_+ and O_-, it satisfies

$$A+Bi-C=A-Bi-C=0,$$

i.e. $A=C$ and $B=0$. Then by completing the square, we can show such a conic must be a circle with center $(D/2,E/2)$ and some complicated radius:

$$\begin{aligned}&Ax^2+Ay^2+Dxz+Eyz+Fz^2\\&=A\left(x-\frac{D}{2}z\right)^2+A\left(y-\frac{E}{2}z\right)^2+\left(F-A\frac{D^2}{4}+A\frac{E^2}{4}\right)z^2.\end{aligned}$$

Hence: a conic is a circle iff it passes through O_+ and O_-, and the collection of circles is the intersection of two degree-1 hypersurfaces in $\mathbb{P}^5$.

Let's now examine the conditions on $[A:0:A:D:E:F]$ in order for the corresponding circle to be tangent to a given circle. Suppose for simplicity the given circle is $x^2+y^2=1$. Our circle intersects with this given circle at points (x,y) satisfying

$$\begin{cases} Ax^2+Ay^2+Dx+Ey+F &= 0 \\ x^2+y^2=1. \end{cases}$$

Substituting the second equation into the first, we get $A+Dx+Ey+F=0$. Substituting this back into the second, we get

$$E^2x^2+(Dx+A+F)^2=E^2.$$

Since this is a quadratic in x, there will be exactly two intersection points. This is as expected, since by Bézout's theorem, two circles intersect at exactly 4 points: the two points O_+ and O_-, and two more points P and Q. If $P=Q$, then the two circles are tangent! But $P=Q$ only when the quadratic above has discriminant 0. This condition, worked out, is precisely the equation

$$A^2-D^2-E^2+2AF+F^2=0.$$

Hence: the condition "$[A:0:A:D:E:F]$ is tangent to a given circle" is a hypersurface of degree 2.

Now we just apply Bézout's theorem. Restricting to circles among conics means we take the intersection of two hypersurfaces of degree 1, and restricting to the circles tangent to three given circles means intersecting with three more hypersurfaces of degree 2. Bézout's theorem says the intersection contains exactly $1\cdot 1\cdot 2\cdot 2\cdot 2=8$ points. □

Because we just need to figure out the degree of the equation constraining A,D,E,F, we can work with a specific circle without loss of generality, and assume the intersections happen in $\mathbb{R}^2\subset\mathbb{P}^2$ so we can set $z=1$ for simplicity.

The condition actually has an extra factor of E^2, but if $E=0$ then the resulting circle is actually a line $(ax+by+c)^2=0$; such "circles" are called *degenerate* and we do not consider them as actual circles.

The technical caveat is that this theorem is true only when we extend to $\mathbb{P}^2$, and we use complex numbers instead of real numbers (so that a degree n polynomial always has exactly n solutions). If we return to the real number case, and $\mathbb{R}^2$, it is still true that there are *at most* eight circles tangent to all three given circles (because a degree n polynomial still always has *at most* n solutions).

6.3 Schubert calculus

So far we have only been able to answer questions about conics in the plane, because there is a very simple description of the "space of all plane conics" as $\mathbb{P}^5$. Now we will look at the intersection theory of objects in higher dimensions. For example, the fourth question in our list involves lines in $\mathbb{R}^3$, and we don't have a good "space of lines in $\mathbb{R}^3$" yet. However, in the same way that we extended $\mathbb{R}^2$ to $\mathbb{P}^2$ when we did intersection theory in the plane, now we will also extend $\mathbb{R}^3$ to $\mathbb{P}^3$. So the actual question we should ask is:

4. how many lines intersect four given lines in $\mathbb{P}^3$?

We will not be rigorous in the process of answering this and related questions. In particular, we will rely on the following principle, called Schubert's "principle of conservation of number", without proof:

> as long as the number of solutions remains finite, we can move the given objects around without changing the total number of solutions (counted with multiplicity) to the enumerative problem.

(As with Bézout's theorem, this principle requires us to work in $\mathbb{P}^n$, and to use complex coordinates.) Let's see this principle in action.

Proposition 6.14 (Schubert, 1879)**.** *There are exactly two lines intersecting four given (generic) lines in* $\mathbb{P}^3$*.*

Proof. We will show this the same way Schubert did. Let:

1. g_{line} be the condition of "passing through a given line";
2. g_{plane} be the condition of "being contained in a given plane";
3. g_{point} be the condition of "passing through a given point".

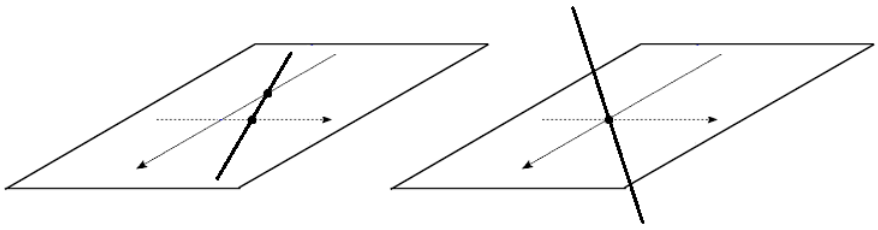
Products of these symbols mean imposing conditions simultaneously. For example:

1. g_{line}^2 is the condition of "passing through two given lines";
2. g_{plane}^2 is the condition of "being contained in two given planes";
3. g_{point}^2 is the condition of "passing through two given points".

Some products will evaluate to actual numbers. For example, through two given points is a unique line, so $g_{\text{point}}^2 = 1$, and the intersection of two planes is a unique line, so $g_{\text{plane}}^2 = 1$. Others, like g_{plane} on its own, do not evaluate to an actual number, because there are infinitely many lines contained in a given plane.

The advantage of this notation is that we can now work algebraically. The number we want to compute is g_{line}^4. We first compute g_{line}^2 as follows. Using Schubert's principle, we can move the two given lines until they intersect. Then for a line ℓ to pass through both given lines, ℓ must either

1. be contained in the plane spanned by the two given lines, or
2. pass through the point where the two given lines intersect.



In other words, we have shown $g_{\text{line}}^2 = g_{\text{plane}} + g_{\text{point}}$. This immediately implies that

$$\begin{aligned} g_{\text{line}}^4 = (g_{\text{plane}} + g_{\text{point}})^2 &= g_{\text{plane}}^2 + 2g_{\text{plane}}g_{\text{point}} + g_{\text{point}}^2 \\ &= 1 + 2g_{\text{plane}}g_{\text{point}} + 1. \end{aligned}$$

So it remains to figure out what $g_{\text{plane}}g_{\text{point}}$ is, i.e. how many lines are contained in a given (generic) plane and a given (generic) point. But a generic point in $\mathbb{R}^3$ will not lie on the given plane, so the line cannot lie on the plane and pass through the point simultaneously, and $g_{\text{plane}}g_{\text{point}} = 0$. It follows that

$$g_{\text{line}}^4 = 2,$$

i.e. there are exactly two lines passing through four given (generic) lines. □

In general, **Schubert calculus** is a set of rules for taking products of symbols g_λ, which are generalizations of conditions such as "being contained in a plane" to conditions involving k-dimensional planes in $\mathbb{R}^n$. Using Schubert's principle of conservation of number, we can move these k-dimensional planes to special positions. For example, when we talk about "being contained in a given two-dimensional plane in $\mathbb{R}^4$", we may as well assume the given plane is $\{(x_1, x_2, 0, 0) \in \mathbb{R}^4\}$, i.e. the plane spanned by the x_1 and x_2 axes.

Definition 6.15. From now on, we abbreviate "k-dimensional plane" as "**k-plane**".

1. The **Grassmannian** $\text{Gr}(k, n)$ is the space of all k-planes in $\mathbb{R}^n$ that pass through the origin.
2. The **projective Grassmannian** $\mathbb{G}(k, n)$ is the space of all k-planes in $\mathbb{P}^n$.

The Grassmannian is named after Hermann Grassmann, who first introduced the space. It is, in fact, a manifold!

Since a point in $\mathbb{P}^n$ is a line in $\mathbb{R}^{n+1}$ through the origin, it follows that a k-plane in $\mathbb{P}^n$ is a $(k+1)$-plane in $\mathbb{R}^{n+1}$ through the origin, i.e.

$$\mathbb{G}(k, n) = \text{Gr}(k+1, n+1).$$

When we ask enumerative questions about $\mathbb{G}(k, n)$, we do the actual intersection theory in $\text{Gr}(k+1, n+1)$, for reasons that will become clear shortly.

Example 6.16. The space of all lines in $\mathbb{P}^3$ is by definition $\mathbb{G}(1, 3)$. But $\mathbb{G}(1, 3) = \text{Gr}(2, 4)$, the space of 2-planes in $\mathbb{R}^4$ passing through the origin. The space $\text{Gr}(2, 4)$ will be our primary source of examples from now on.

The best way (for us) to represent a k-plane in $\mathbb{R}^n$ is to pick k vectors in $\mathbb{R}^n$ which span the plane, and put them into a $k \times n$ matrix row by row. For example, the 2-plane spanned by x_1 and x_2 axes in $\mathbb{R}^4$ is given by

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

Note that given any k-plane as a matrix, row-reducing the matrix does not change which plane it describes. For example,

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

are both the same 2-plane in $\mathbb{R}^4$, but the second is in row-reduced form.

Example 6.17. In general, given an arbitrary 2-plane spanned by v_1, v_2 in $\mathbb{R}^4$,

$$\begin{pmatrix} \text{—} & v_1 & \text{—} \\ \text{—} & v_2 & \text{—} \end{pmatrix},$$

there are only a few cases for what the row-reduced form looks like. (Here we use $*$ to denote an entry that can be any number.)

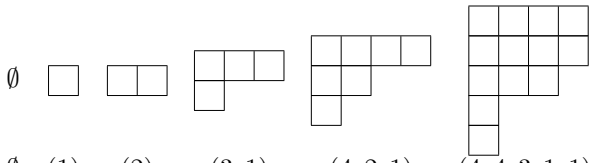
1. $\begin{pmatrix} 1 & 0 & * & * \\ 0 & 1 & * & * \end{pmatrix}$, a generic 2-plane in $\mathbb{R}^4$ (or a generic line in $\mathbb{P}^3$);
2. $\begin{pmatrix} 1 & * & 0 & * \\ 0 & 0 & 1 & * \end{pmatrix}$, a 2-plane which intersects the given 2-plane
$$\{(0, 0, x_3, x_4) \in \mathbb{R}^4\}$$
(or a line intersecting a given line in $\mathbb{P}^3$);
3. $\begin{pmatrix} 1 & * & * & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$, a 2-plane which intersects the given line
$$\{(0, 0, 0, x_4) \in \mathbb{R}^4\}$$
(or a line intersecting a given point in $\mathbb{P}^3$);
4. $\begin{pmatrix} 0 & 1 & 0 & * \\ 0 & 0 & 1 & * \end{pmatrix}$, a 2-plane which is contained in the given 3-plane
$$\{(0, x_2, x_3, x_4) \in \mathbb{R}^4\}$$
(or a line contained in a given 2-plane in $\mathbb{P}^3$);
5. $\begin{pmatrix} 0 & 1 & * & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$, a 2-plane contained in the given 3-plane
$$\{(0, x_2, x_3, x_4) \in \mathbb{R}^4\}$$
and intersecting a given line in that 3-plane (or a line contained in a given 2-plane passing through a point in $\mathbb{P}^3$);
6. $\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$, the specific 2-plane
$$\{(0, 0, x_3, x_4) \in \mathbb{R}^4\}$$
(or a specific line in $\mathbb{P}^3$).

Each of these possibilities is a different condition we can put on the space of lines in $\mathbb{P}^3$. For example, possibilities (2), (3), and (4) are what we called g_{line}, g_{point}, and g_{plane} earlier, respectively. But we clearly need a more systematic way of naming these conditions. The naming system we use involves objects called Young diagrams.

Definition 6.18. A **Young diagram** λ is a collection of boxes arranged in a grid, with rows left-justified and with row lengths in non-increasing order. The **shape** of a Young diagram is a list of its row lengths.

It is customary to use the Greek letter "lambda" λ to denote Young diagrams.

Example 6.19. Here are some valid Young diagrams with their shapes below them:



$\emptyset$ (1) (2) (3, 1) (4, 2, 1) (4, 4, 3, 1, 1).

(We use $\emptyset$ to denote the empty diagram.) The shape of a Young diagram always has non-increasing entries.

The idea is that to each distinct possibility for the row-reduced matrix M we can associate a distinct Young diagram. The method is simple: compare the matrix M with the generic matrix, and for each row in M, count the number of extra 0s on the left of the first 1. These numbers, read from last row to first row, specify the shape of a Young diagram.

Example 6.20. We transform each of the possibilities for $\mathrm{Gr}(2,4)$ into Young diagrams, marking extra zeros in each row in bold.

Matrix	Shape	Young diagram
$\begin{pmatrix} 1 & 0 & * & * \\ 0 & 1 & * & * \end{pmatrix}$	$\emptyset$	$\emptyset$
$\begin{pmatrix} 1 & * & 0 & * \\ \mathbf{0} & 0 & 1 & * \end{pmatrix}$	(1)	
$\begin{pmatrix} 1 & * & * & 0 \\ \mathbf{0} & \mathbf{0} & 0 & 1 \end{pmatrix}$	(2)	
$\begin{pmatrix} \mathbf{0} & 1 & 0 & * \\ \mathbf{0} & 0 & 1 & * \end{pmatrix}$	(1, 1)	
$\begin{pmatrix} \mathbf{0} & 1 & * & 0 \\ \mathbf{0} & \mathbf{0} & 0 & 1 \end{pmatrix}$	(2, 1)	
$\begin{pmatrix} \mathbf{0} & \mathbf{0} & 1 & 0 \\ \mathbf{0} & \mathbf{0} & 0 & 1 \end{pmatrix}$	(2, 2)	

Exercise. Convince yourself of the following facts:

1. every row-reduced matrix for $\mathrm{Gr}(k,n)$ gives a valid Young diagram;
2. the resulting Young diagram fits into a $k \times (n-k)$ grid, i.e. row lengths are never $> (n-k)$;
3. every such Young diagram corresponds to a distinct row-reduced matrix for $\mathrm{Gr}(k,n)$.

In other words, the collection of constraints for objects in $\mathrm{Gr}(k,n)$ are in bijection with Young diagrams fitting into a $k \times (n-k)$ grid.

For example, we can rewrite the constraints we had earlier in terms of Young diagrams.

$$g_{\text{line}} = \square, \quad g_{\text{point}} = \square\square, \quad g_{\text{plane}} = \begin{array}{|c|}\hline\ \\\hline\ \\\hline\end{array}.$$

The relations we computed earlier can be written

$$\left(\square\right)^2 = \square\square + \begin{array}{|c|}\hline\ \\\hline\ \\\hline\end{array}, \quad \left(\square\square\right)^2 = \left(\begin{array}{|c|}\hline\ \\\hline\ \\\hline\end{array}\right)^2 = 1, \quad \square\square \cdot \begin{array}{|c|}\hline\ \\\hline\ \\\hline\end{array} = 0.$$

In fact, there are general formulas for the product of Young diagrams.

Theorem 6.21 (Pieri rule)**.** *Let (1^m) denote the shape $(1,1,\ldots,1)$, where there are m ones. Let σ_λ denote the constraint associated to the Young diagram λ. Then*

$$\sigma_\lambda \cdot \sigma_{(1^k)} = \sum_{\mu \supset \lambda} \sigma_\mu$$

where the sum ranges over all Young diagrams μ such that:

1. *μ still fits into a $k \times (n-k)$ box,*
2. *μ contains m more boxes than λ,*
3. *no two of the m newly added boxes are in the same row.*

Example 6.22. The Pieri rule already allows us to compute $\square^4$ step-by-step. Boxes with dots denote newly added boxes.

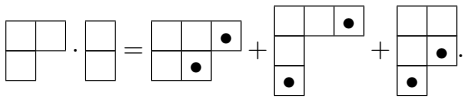
$$\left(\square\right)^2 = \square\bullet + \begin{array}{|c|}\hline\ \\\hline\bullet\\\hline\end{array}$$

$$\left(\square\right)^3 = \left(\square\square + \begin{array}{|c|}\hline\ \\\hline\ \\\hline\end{array}\right) \cdot \square = \begin{array}{|c|c|}\hline\ & \ \\\hline\bullet & \\\hline\end{array} + \begin{array}{|c|c|}\hline\ & \bullet\\\hline\ & \\\hline\end{array}$$

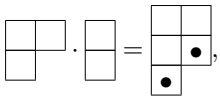
$$\left(\square\right)^4 = \left(2 \cdot \begin{array}{|c|c|}\hline\ & \ \\\hline\ & \\\hline\end{array}\right) \cdot \square = 2 \cdot \begin{array}{|c|c|}\hline\ & \ \\\hline\ & \bullet\\\hline\end{array}.$$

Since $\begin{array}{|c|c|}\hline\ & \ \\\hline\ & \\\hline\end{array}$ represents one specific line, the final answer is 2 lines.

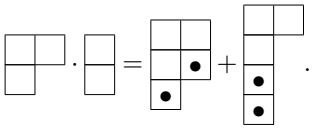
Example 6.23. In $\mathrm{Gr}(3,6)$, i.e. all Young diagrams fit into a 3×3 grid, we can compute that



However, in $\mathrm{Gr}(3,5)$, we would have



and in $\mathrm{Gr}(4,6)$ we would have



So when doing calculations with Young diagrams, it is very important to keep in mind which Grassmannian we are working in.

Exercise. Using the Pieri rule in $\mathbb{G}(1, 2n+1)$, show that there are exactly $n+1$ lines in $\mathbb{P}^{2n+1}$ which intersect four given (generic) n-planes. (Note that for $n = 1$ this is the result that in $\mathbb{P}^3$ there are exactly two lines which intersect four given lines.)

In general, the product of two arbitrary Young diagrams is calculated using the **Littlewood–Richardson rule**, of which the Pieri rule is a special case. We will not describe this general rule here.

A Appendix: Group theory

Sometimes mathematical objects have too much structure, and we would like to restrict our attention to only a portion of the structure they possess. For example, surfaces are potentially very complicated objects, but when we only look at them topologically, i.e. up to homeomorphism, we obtain a beautiful classification theorem. The concept of a "group" can be similarly motivated. For example, we can look at the set of all integers, which we denote by $\mathbb{Z}$. This set has an addition operation (the addition of integers) and a multiplication operation (the multiplication of integers). When we forget about the multiplication operation and only look at the set of integers with an addition operation, we are looking at $\mathbb{Z}$ as a "group."

Formal definition. A **group** G is a set that has a **group operation** $\star$. More precisely, this means that for any two elements a and b in G, we can apply the operation $\star$ to them to obtain an element $a \star b$. This operation must satisfy some axioms:

1. there must exist an **identity element** e of G such that $e \star x = x$ for all x;
2. the group operation must be **associative**, i.e. $(a \star b) \star c = a \star (b \star c)$;
3. every element x must have an **inverse**, i.e. an element x^{-1} such that $x \star x^{-1} = e$.

It is common to call the inverse x^{-1} because we often pretend the group operation is "multiplication."

Example A.1. It is straightforward to check that the integers $\mathbb{Z}$ form a group using addition as the group operation. Clearly given two integers x and y, their sum $x + y$ is still an integer.

1. The identity element is 0, because $0 + x = x$ for any integer x.
2. The operation of addition is associative, because $(x + y) + z = x + (y + z)$ for all integers x, y, z.
3. The inverse of an integer x is the integer $-x$ (which always exists), because $x + (-x) = 0$.

Example A.2. Let $\mathbb{Z}/n$ denote the group of **integers modulo** n, using addition as the group operation. In other words, it is the set $\{0, 1, 2, \ldots, n-2, n-1\}$ with addition modulo n as the group operation. Because of the "modulo n," given two elements x and y of $\mathbb{Z}/n$, the result $x + y \bmod n$ is still an element of $\mathbb{Z}/n$. That this group operation satisfies the axioms of a group operation is easy to check.

Example A.3. The first homotopy group $\pi_1(M, x_0)$ is a group using the concatenation of loops as the group operation. Clearly the concatenation of two loops is still a loop.

1. The identity element is the constant loop, which we will denote 1, because concatenation of any loop γ with the constant loop is still γ. We write this as $1 \cdot \gamma = \gamma$.
2. The operation of concatenation is associative, because $(\gamma_1 \cdot \gamma_2) \cdot \gamma_3 = \gamma_1 \cdot (\gamma_2 \cdot \gamma_3)$.
3. The inverse of a loop γ is the same loop traversed backward, which we denoted γ^{-1}, because $\gamma \cdot \gamma^{-1}$ is homotopic, and therefore equal in $\pi_1(M, x_0)$, to the constant loop.

Definition A.4. A group G is **abelian** if $x \star y = y \star x$ for every x and y in G. For example, $\mathbb{Z}$ is abelian, but $\pi_1(M, x_0)$ in general is not.

In other words, a group is abelian if the group operation is "commutative".

Just like we consider two surfaces to be equivalent when they are homeomorphic, we need to have a notion of when two groups are "equivalent." For example, if we take a group and just rename each of its elements while keeping the same group operation, this renamed group should clearly be considered equivalent to the original group. The general idea is that equivalent objects should have the "same group operation" when objects in one group are appropriately relabeled.

Formal definition. A **group homomorphism** from a group G to a group H is a function $f\colon G \to H$ such that $f(x \star_G y) = f(x) \star_H f(y)$. Here $\star_G$ denotes the group operation on G, and $\star_H$ denotes the group operation on H. We say a group homomorphism $f\colon G \to H$ is a **group isomorphism** if there exists an **inverse** group homomorphism $g\colon H \to G$ such that $f \circ g = g \circ h = \mathrm{id}$, the identity function (which sends every element to itself).

Example A.5. The group $\pi_1(S^1)$ is isomorphic to $\mathbb{Z}$ (as groups). Explicitly, the group isomorphism is given by sending the loop γ^n in $\pi_1(S^1)$ to the integer n in $\mathbb{Z}$. This is a group homomorphism because $\gamma^n \cdot \gamma^m$ is sent to $n + m$. The inverse homomorphism sends the integer n to the loop γ^n, and it is easy to check that this is also a homomorphism.

In fact, just like with (compact, connected) surfaces, there is a classification of (finitely generated) abelian groups. In order to describe this classification, we need to define the product of groups.

"Finitely generated" means every element in the group can be written as $x_1 \star x_2 \star \cdots \star x_n$ for some n, and for $x_1, \ldots, x_n$ picked from a *finite* set of prescribed elements.

Definition A.6. Given two groups G and H, their **product** $G \times H$ is the group whose elements are pairs (g, h) with $g \in G$ and $h \in H$, and group operation given by

$$(g_1, h_1) \star (g_2, h_2) = (g_1 \star_G g_2, h_1 \star_H h_2).$$

Theorem A.7 (Classification of finitely generated abelian groups)**.** *Any finitely generated abelian group is isomorphic to*

$$\mathbb{Z}^r \times \mathbb{Z}/n_1 \times \cdots \times \mathbb{Z}/n_k$$

for some integers $r, k \geq 0$ and $n_1, \ldots, n_k \geq 2$.