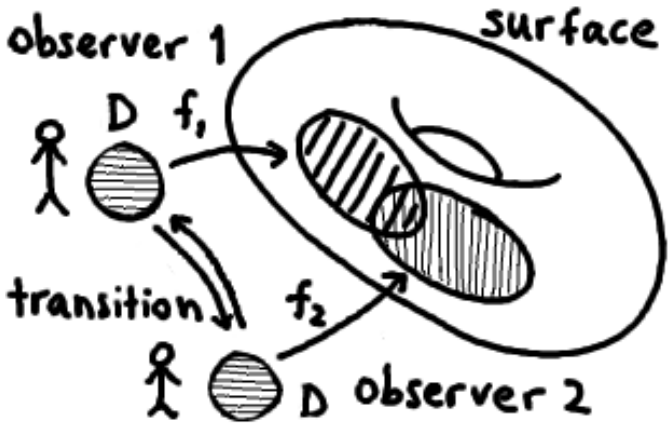
OXFORD MASTERCLASSES IN GEOMETRY 2014.
Part 1: Lectures on Geometry and Topology,
Prof. Alexander F. Ritter.

Comments and corrections are welcome: ritter@maths.ox.ac.uk



Date: This version of the notes was created on September 4, 2014.

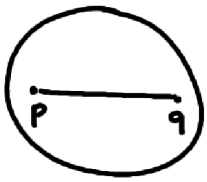
Contents

1. Spaces, distances, continuity and convergence

1.1. **Spaces, paths and distances.**

What do we mean by a space X which has a notion of distance between points? For example, if you know the lengths of paths between two points, you could call distance the least length you must travel if you start from point p and you end up at point q.

Physically the shortest path is the path of a light-ray, unless there is an obstacle in between.[1] In general, shortest paths need not look like straight lines. For example, suppose our world was a disc. If we use the usual Euclidean distance, then the shortest paths are straight lines:

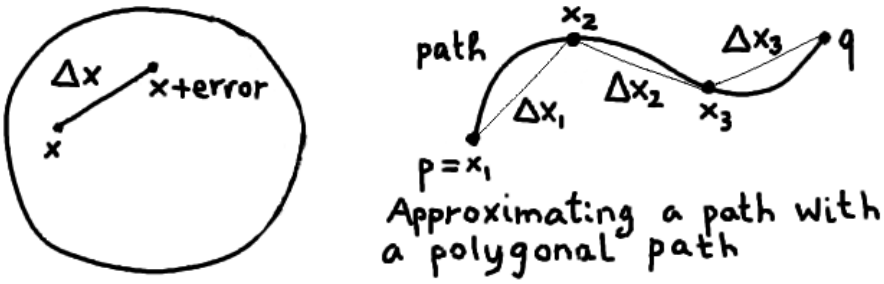


But if instead distances near the centre 0 of the disc were much longer than we imagined with Euclidean eyes, short paths would try to avoid passing through 0:

The picture on the right is what your world may look like inside 3-space if you take into account lengths (notice, it is still just a disc): so near 0 there is a mountain, and it is more convenient to walk around the mountain when going from p to q.

1.2. **Density functions.**

How would we measure distances in this disc world? One approach, is to consider a tiny Euclidean-straight line segment Δx in the disc with end-points $x, x +$ small error.



Then we compare the actual length $\ell(\Delta x)$ with the usual Euclidean length $\|\Delta x\|$:

$$\ell(\Delta x) \approx f(x)\,\|\Delta x\|,$$

[1] Shortest paths need not exist, for example take the plane with the usual Euclidean length, but remove the point 0. Then two opposite points $x, -x$ cannot be joined by a shortest path since the straight line through 0 is not a legitimate path as 0 does not actually belong to the space! Also, shortest paths, when they exist, need not be unique. For example on the sphere (the surface of the Earth) there are infinitely many great arcs joining the North Pole to the South Pole all of minimal length.

for some value $f(x)$ (which is rigorously defined as the value that the ratio $\ell(\Delta x)/\|\Delta x\|$ approaches as we shrink Δx to the point x). The function

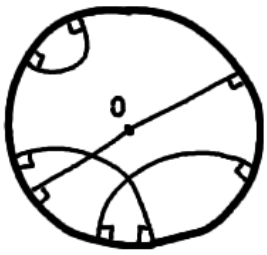
$$f : \text{Disc} \to \mathbb{R}$$

is called *density function* and it depends on the point x.

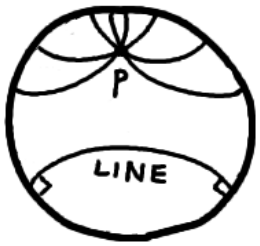
In the above example of the mountain at 0, a small Euclidean path near 0 has a very large length (since we actually walk over a mountain!) so $f(0)$ is very large.

1.3. **Hyperbolic geometry.**

There is a non-Euclidean geometry, called *hyperbolic geometry*, in which the universe is a unit disc $D = \{z \in \mathbb{C} : |z| < 1\}$. The circular boundary $\partial D = \{z \in \mathbb{C} : |z| = 1\}$ of the disc does not belong to D, indeed those boundary points should be thought of as being "at infinity". The shortest paths are arcs, which belong to Euclidean circles[2] that are perpendicular to ∂D.



You can check that all the axioms of Euclidean geometry are satisfied (with "line" meaning an arc as above), except for the *parallel axiom*. That axiom says that given a line and a point outside it, there is exactly one line through that point which does not intersect the given line.[3] This fails here because we can draw many such parallel lines:



The density function for the hyperbolic disc D is defined as:

$$f(z) = \frac{2}{1 - |z|^2}.$$

Notice that near the boundary $|z| = 1$, the ratio $f(z)$ between hyperbolic lengths and Euclidean lengths blows up. Indeed, a path from 0 to a boundary point $z \in \partial D$ has infinite hyperbolic length: so the intuition that the points at ∂D lie at infinity is correct.

Remark 1 (Curvature)**.** *The numerator* 2 *in* $f(z)$ *is artificial: you could have used any positive constant. The reason for having* 2*, is that you want another quantity, called* curvature*, to be* -1 *in this case. The idea of curvature, is that it measures how much two light-rays move apart when shot out from an observer:*

[2]Including circles of infinite radius, which give rise to diameters: these are also shortest paths.

[3]Two lines which do not intersect are called *parallel lines*.

On the sphere S^2 the curvature is positive: light-rays move towards each other. On the hyperbolic disc D the curvature is negative: light-rays move apart. On the Euclidean plane $\mathbb{R}^2$ the curvature is zero.

1.4. **How the choice of distance affects the geometry. Introduction to tilings.**

Later, we will study *topological* properties of spaces: properties which do not depend on exactly how we measure distances as long as a vague notion of "closeness" is preserved.

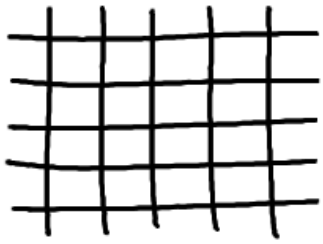
Example. The Euclidean plane $\mathbb{R}^2$ can be identified with the hyperbolic disc D by rescaling $z \in D$ to $\frac{z}{1-|z|} \in \mathbb{C} = \mathbb{R}^2$. This preserves a vague notion of ''closeness'', but it does not preserve the distance function.

Once we take into account the different distance functions on the Euclidean plane $\mathbb{R}^2$ and on the hyperbolic "plane" D, the two spaces have quite different geometrical properties. For example, in the Exercises you will see that the tilings they admit are rather different.

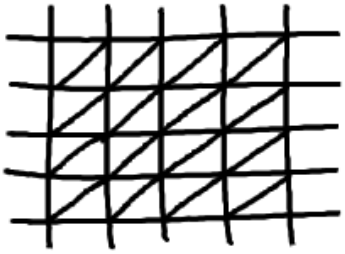
A *tiling* (or *tessellation*) of the plane by polygons is a covering of the plane by polygons, so that every point of the plane lies in some polygon, and the polygons do not overlap except possibly along their boundaries (that is, along edges or vertices).

Example.

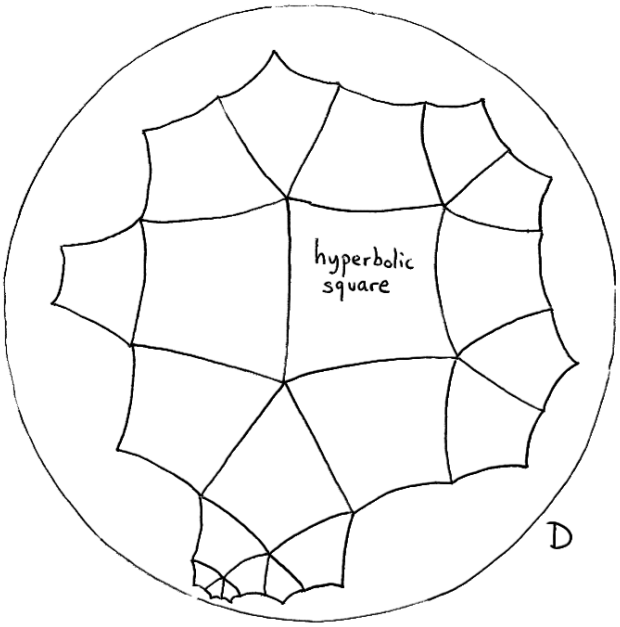
(1) *The tiling of the plane by unit squares:*



(2) *By decomposing the squares into triangles, the above turns into a tiling by isosceles triangles:*

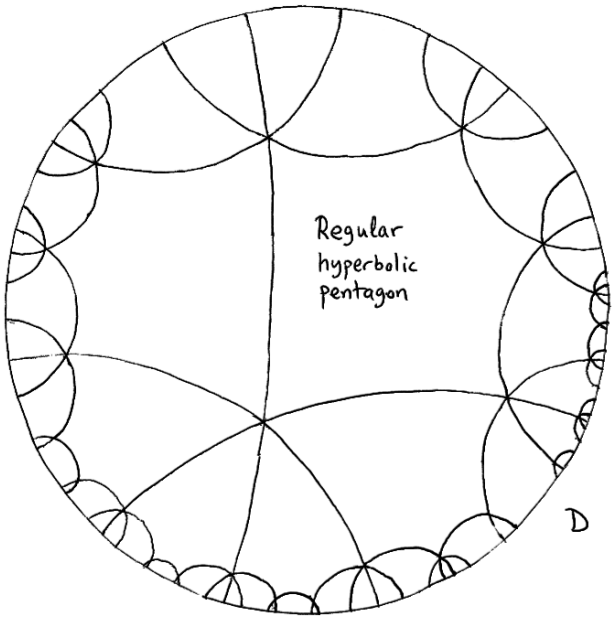


(3) *Tiling of the hyperbolic plane D using one hyperbolic square:*



Assuming that a regular polygon can tile the plane, you obtain the tiling from one polygon by repeatedly reflecting the given polygon in the edges you built so far inductively. To understand what reflection means in the hyperbolic world, first think about how you reflect a point in a line in Euclidean geometry (using only ruler and compass), then the same procedure will work[4] *for hyperbolic geometry since you never use the parallel axiom.*

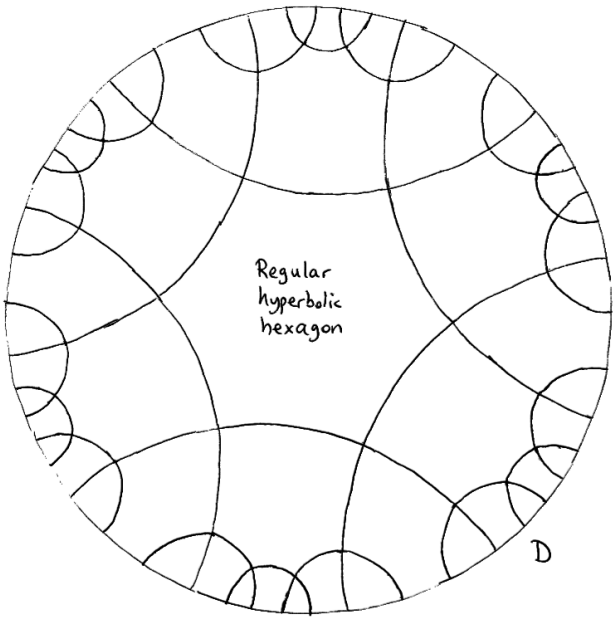
(4) *Tiling of the hyperbolic plane D using one regular pentagon:*



Can you see how the construction works?

(5) *Tiling of the hyperbolic plane D by using one regular hexagon:*

[4]In the hyperbolic world, you will of course use a "hyperbolic compass" which traces out hyperbolic circles. As an exercise, try showing that hyperbolic circles are in fact Euclidean circles except that the hyperbolic centre is usually not the same as the Euclidean centre. Start the exercise by first checking what hyperbolic circles are when the hyperbolic centre is 0 (*Hint.* the density function has a symmetry). Then use the symmetry maps from the Exercises to find all hyperbolic circles.

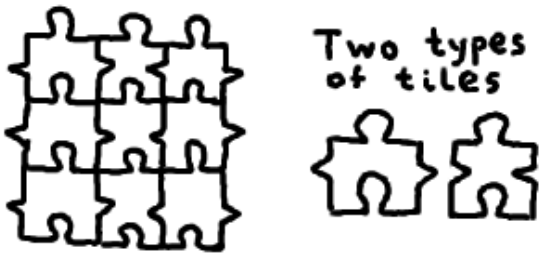


Can you see how the construction works?

In the Exercises you will show that the Euclidean plane can be tiled by a regular n-sided polygon (with fixed side length) if and only if $n = 3, 4, 6$ (so regular triangles, squares or hexagons). Whereas the hyperbolic "plane" D can be tiled for any n, provided that one chooses the regular polygon to be of the correct size.[5]

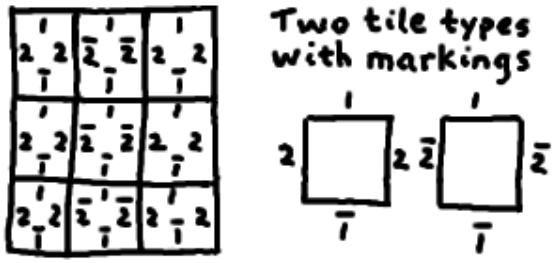
Usually, we care about tilings using tiles from a certain finite collection of possible polygons of given sizes. In the above example: by a particular square, or a particular triangle. One can also allow more general shapes than polygons, such as any set in the plane bounded by a closed curve with no self-intersections (so topologically it looks like a deformed disc).

For example:



This tiling is obtained from the tiling by squares after deforming the tiles. In this case we use two types of tiles. Since drawing these squiggly edges is tiring, it is often more convenient to put *markings* on the edges, which tell you how the tiles must fit together:

[5]A curious fact about hyperbolic geometry, which you will notice from those pictures, is that the sum of the interior angles of a triangle, or more generally of a regular n-gon, is not fixed: it depends on the size. For example, the area of a hyperbolic triangle is in fact $\pi -$ (sum of interior angles). For example, small hyperbolic triangles almost look like Euclidean triangles, so the angles almost sum up to π.



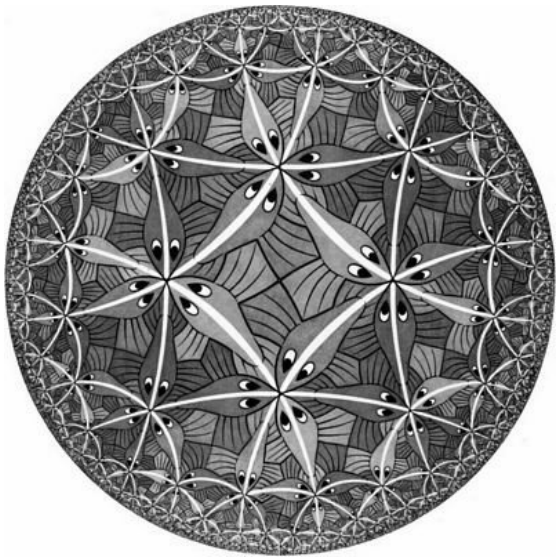
Above, we use labels $1, 2$ which are required to always be glued onto $\overline{1}, \overline{2}$ respectively. You can also use arrows to prescribe in which direction you want them to be glued.

The key property you want from your markings is that: for any tiling by squiggly tiles there is a *unique* tiling using marked tiles, and vice-versa from a tiling by marked tiles you can reconstruct uniquely a tiling by the squiggly tiles.

Markings can also come in the form of artistic decorations on top of the tile types, and the matching condition is that the decorations fit together nicely.

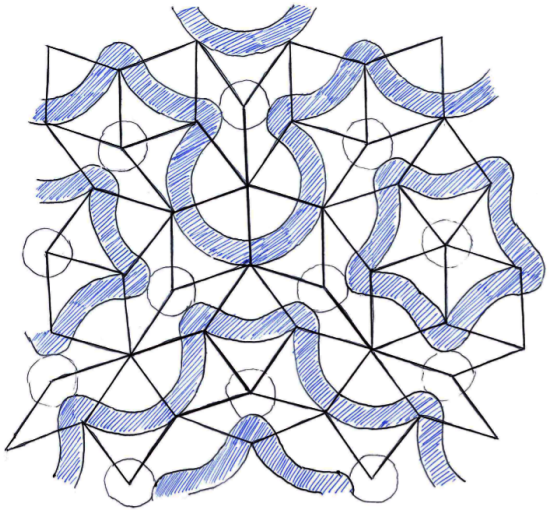
Examples.

(1) *M. C. Escher's Circle Limit III (1959) is a tiling of the hyperbolic plane:*



Exercise. Can you find the underlying tiling by regular hyperbolic hexagons?

(2) *A Penrose tiling using Penrose rhombi with decorations:*



Regarding the general definition of tiling. One needs some care in saying exactly what a tiling is, because we want to avoid bad tiles (e.g. a tile made up of disconnected pieces, or a tile having holes, or tiles that have parts which become infinitesimally thin), we want to avoid tiles that are too large (e.g. unbounded tiles, like infinite strips), and nasty things can happen if we allow infinitely many tile types (e.g. you usually do not want there to be infinitely many tiles covering a finite region, for example this can happen if you use the collection of tile types given by squares of any side length).

Exercise 2. *Consider the infinite collection of tile types consisting of discs of any positive radius. Can you tile the whole plane using only copies of tiles taken from this collection?*

1.5. **Distance function.**

Let's try to make the notion of distance more abstract. Given a set X, we want a function d, called *distance*, which eats two points $p, q \in X$ and spits out a number

$$d(p,q) = d(q,p) \geq 0 \qquad \text{for } p, q \in X.$$

We want two basic properties:

- zero distance $\Rightarrow p = q$, and vice-versa: $d(p,p) = 0$;
- $d(p,q) \leq d(p,x) + d(x,q)$ for any $p, x, q \in X$ (the *triangle inequality*).
  Example: $d(\text{Boston}, \text{Oxford}) \leq d(\text{Boston}, \text{The Moon}) + d(\text{The Moon}, \text{Oxford})$.

When we say *(metric) space* we will mean a set X together with a choice of distance d.

Examples.

(1) *The real line* $\mathbb{R}$*, with the usual Euclidean distance* $d(p,q) = |p - q|$.

(2) *The plane* $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$*, with the usual Euclidean distance*

$$d(p,q) = \|p - q\|_2 = \sqrt{|p_1 - q_1|^2 + |p_2 - q_2|^2}$$

where $p = (p_1, p_2)$ *are the* (x, y) *coordinates of* p.

(3) $\mathbb{R}^2$ *with* $d(p,q) = \|p - q\|_1 = |p_1 - q_1| + |p_2 - q_2|$,

(4) $\mathbb{R}^2$ *with* $d(p,q) = \|p - q\|_\infty = max\{|p_1 - q_1|, |p_2 - q_2|\}$.

(5) *Any set* X *with* $d(p,q) = 1$ *unless* $p = q$ *(in which case* $d(p,p) = 0$*).*

(6) $X =$ *Wadham College,* $d(p,q) =$ *the least length of a piece of string that you can have in Wadham College joining the points* p, q *(going through the windows or staircases as necessary).*

(7) *In a set* X *where you know how to measure lengths of curves, and assuming that any two points can be connected by some curve, you can define a distance function*

$$d(p,q) = min\{\ell \in \mathbb{R} : \text{ all curves from } p \text{ to } q \text{ have length at least } \ell\}.$$

(8) *Hyperbolic geometry: the disc* $D = \{z \in \mathbb{C} : |z| < 1\}$*. Then it turns out*[6] *that*

$$d(0,z) = \log \frac{1 + |z|}{1 - |z|}.$$

You can obtain a formula for $d(z_1, z_2)$ *for general points* z_1, z_2*, by using the above formula and using the symmetries of the hyperbolic disc*

$$\{z \mapsto \frac{az + b}{\overline{b}z + \overline{a}} : a, b \in \mathbb{C}, |a|^2 + |b|^2 \neq 0\}$$

mentioned in the Exercises. These symmetries preserve hyperbolic distances and they preserve angles.

[6]You could try to prove this by approximating any curve by a polygonal curve, then using the density function to obtain a sum of lengths of straight-line segments which approximates the length of the curve, and finally taking a "limit" as the polygonal approximation becomes better and better. However, pedagogically this is not so reasonable. That limit process is called *integration*, and calculus develops tools to calculate these limits very easily using *integrals*. So you should try out this calculation once you know about integrals.

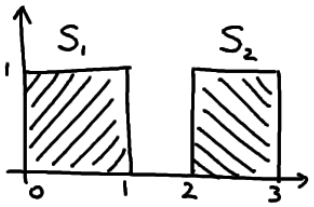
(9) *If X, Y are spaces, with distance functions d_X, d_Y, the product set of all pairs (x, y),*

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

has an natural distance function: the sum of the distances:

$$d((x, y), (x', y')) = d_X(x, x') + d_Y(y, y').$$

(10) *The* Hausdorff distance *is a notion of distance between certain subsets of $\mathbb{R}^n$, which you will look at more closely in the Exercises. Let's consider $\mathbb{R}^2$, and the set $X = \{$closed bounded subsets $S \subset \mathbb{R}^2\}$. ''Closed'' means the subset contains its ''boundary points''. For example the square $[0, 1] \times [0, 1] \subset \mathbb{R}^2$ is closed, whereas $(0, 1] \times (0, 1]$ is not since it does not contain the boundary point $(0, 0)$. ''Bounded'' means that the subset is contained in some sufficiently large disc (of finite radius). For example, what would* you like *the distance to be, between the following two squares?*



Ask yourself: should it be 1, 2, 3 or $\sqrt{10}$? how do you define the distance without violating one of the distance axioms?[7]

The answer, that works well, is to define:

$$d(S_1, S_2) = \min \left\{ \delta \geq 0 : \begin{array}{c} \text{every point } s_1 \in S_1 \text{ has } d(s_1, s_2) \leq \delta \text{ for some } s_2 \in S_2, \\ \text{and every point } s_2 \in S_1 \text{ has } d(s_1, s_2) \leq \delta \text{ for some } s_1 \in S_1 \end{array} \right\}$$

where the distance $d(s_1, s_2)$ between points s_1, s_2 refers to the Euclidean distance. In the example of the two squares, check that $d(S_1, S_2) = 2$.

1.6. **Topology and balls.**

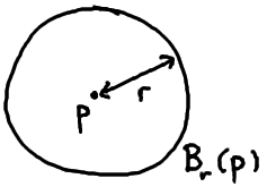
Without a notion of distance on a space, you cannot tell whether two points are close or far away. There is a very weak notion of distance, much weaker than the above, called a *topology* on a space. To define a *topology* on a set X, for each point p you need to choose a collection of "neighbourhoods": subsets surrounding the point p. For neighbourhoods U, V of p, you imagine that the points in U are "closer" to p than the points in V if

$$U \subset V.$$

But since you haven't assigned a numerical measurement to U, V, this is just a vague notion of "closeness", and it is difficult to work with. However, for a metric space, there is an obvious choice of neighbourhoods around p: the *(open) ball of radius $r > 0 \in \mathbb{R}$ around* p consisting of the points q within distance r of p,

$$B_r(p) = \{q \in X : d(q, p) < r\}.$$

[7] If you have trouble, try first asking: if you have a tiny small disc S_1 representing our spaceship, that has almost reached the surface of an enormous disc S_2 representing the Death-star, then do you want these to be considered close or not? How close? If Darth Vader is on the opposite side of the Death Star, does he feel like he's close or far away from our puny spaceship?



Examples.

(1) *For* $\mathbb{R}$ *with the usual distance,* $B_r(p) = (p-r, p+r)$ *is an open interval.*
(2) *For* $\mathbb{R}^2$ *with the usual distance,*

$$B_r(0) = \{(x, y) \in \mathbb{R}^2 : \sqrt{|x-0|^2 + |y-0|^2|} < r\}$$

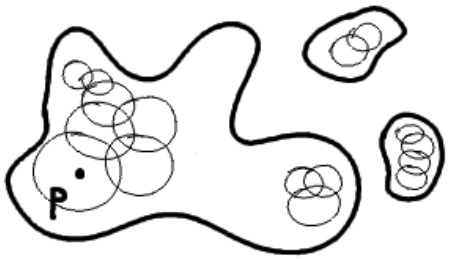
is the usual ball $x^2 + y^2 < r^2$ *around zero.*

1.7. **Open sets, topology, and neighbourhoods.**

A subset $U \subset X$ is called an *open subset* if it is a union of balls (possibly using different radii and different centres). Notice that the whole set X is open (the union of all possible balls), and by convention the empty set $\emptyset$ is open (a union over nothing).

The *topology* of X is, by definition, the collection of all possible open sets.[8]

A *neighbourhood* of p is any open subset containing p (so it is a union of balls, one of which must contain p).



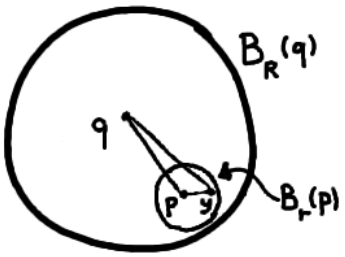
So general neighbourhoods can look very complicated indeed, since you are allowed to take unions over infinitely many balls if you want.

Exercise 3. *Show that any open set of* $\mathbb{R}$*, with the usual distance, is a countable union of disjoint open intervals (an interval is open if it does not contain its end-points).*

The key observation:

Observation 4. *Any neighbourhood of* p *contains some ball* $B_r(p)$ *centred at* p*.*

Proof. The neighbourhood $U = \bigcup B_{r_i}(p_i)$ is some union of balls. Since $p \in U$, we have $p \in B_{r_i}(p_i)$ for some index i. But now in general, if $p \in B_R(q)$ then $B_r(p) \subset B_R(q)$ for small enough $r > 0$.



[8]More precisely, a collection C of subsets of a set X is a *topology* if they satisfy: $\emptyset \in C$, $X \in C$, any union of any $S_i \in C$ is also in X, any *finite* intersection of any $S_i \in X$ is also in X. These subsets $S \in C$ of X are called *open sets*.

Indeed, take $r = R - d(p,q) > 0$: then for $y \in B_r(p)$, using the triangle inequality,

$$d(y,q) \leq d(y,p) + d(p,q) < r + d(p,q) = R,$$

so $y \in B_R(q)$. Hence $B_r(p) \subset B_R(q)$. □

1.8. **Subspaces.**

A subset $S \subset X$ of a metric space (X,d) is also a metric space: just use the same distance function! In that case, you call S a *subspace* of X.

Notice that the balls in S may not look like those in X because points may be missing:

Example. Take $X = \mathbb{R}$ with the usual distance. Then $S = [0,\infty)$ is a subspace. What is the open ball in S with centre 0 and radius 1?

$$B_1(0) = \{q \in S : d_S(q,0) < 1\} = \{q \in [0,\infty) : d(q,0) < 1\} = [0,1).$$

whereas in $X = \mathbb{R}$ we have $B_1(0) = (-1,1)$. It just so happens that the space S no longer contains the negative real numbers that $\mathbb{R}$ used to contain. On the other hand, for any point $p \in (0,\infty)$, for small enough radii r the ball $B_r(p)$ in S will be the same as that in $\mathbb{R}$, since S will not be missing any points.

When small enough balls in S around $p \in S$ are the same as balls in X, then p is called an *interior point*.[9] In the example, the *interior* (the set of interior points) is $\operatorname{Int}[0,\infty) = (0,\infty)$.

If a point p in X (which may or may not belong to S) is a *boundary point* of S if any ball around p contains points both from S and from $X \setminus S$. The subspace of boundary points is denoted ∂S.

Examples.

(1) $0, 1$ *are the boundary points of the intervals* $(0,1)$, $[0,1)$, $(0,1]$, $[0,1]$. *The interior is always* $(0,1)$.
(2) *For* $\mathbb{R}$ *with the usual distance,* $\partial\mathbb{Q} = \mathbb{R}$ *and* $\operatorname{Int}\mathbb{Q} = \emptyset$.
(3) *The boundary* ∂S *depends on* X*: for example, the disc*

$$\mathbb{D} = \{z \in \mathbb{C} : |z| \leq 1\} \subset \mathbb{C} = \mathbb{R}^2$$

has $\partial\mathbb{D} = S^1$, $\operatorname{Int}\mathbb{D} = D$ *(the disc without the points satisfying* $|z| = 1$*). But viewing* $\mathbb{D}$ *inside* $\mathbb{R}^3$ *(taking zero as the third coordinate) we get* $\partial\mathbb{D} = \mathbb{D}$, $\operatorname{Int}\mathbb{D} = \emptyset$. *Another mind-twister: viewing* S *as a subspace of* S *itself one always gets* $\partial S = \emptyset$, $\operatorname{Int} S = S$.

Exercise 5. *Show that the interior* $\operatorname{Int} S$ *of a subspace* $S \subset X$ *is always an open set. Deduce that* $S \subset X$ *is open if and only if* $S = \operatorname{Int} S$.

1.9. **Closed sets.**

A subset $C \subset X$ is called *closed* if it is the complement $X \setminus S$ of an open set S.

Examples.

(1) $\emptyset$, X *are open sets, therefore their complements* $X \setminus \emptyset = X$ *and* $X \setminus X = \emptyset$ *are closed sets. So* $\emptyset$, X *are both open and closed.*
(2) $[0,1] \subset \mathbb{R}$ *is closed since* $[0,1] = \mathbb{R} \setminus ((-\infty,0) \cup (1,\infty))$ *and* $(-\infty,0) \cup (1,\infty)$ *is open.*

Exercise 6. *Show that S is closed if and only if $\partial S \subset S$. Show that* $\operatorname{Int} S = S \setminus \partial S$.

Remark. For any subset $S \subset X$ you always have

$$(\operatorname{Int} S = S \setminus \partial S) \ \subset \ S \ \subset (\ S \cup \partial S = \overline{S}).$$

The interior is the largest open set contained in S, and the closure $\overline{S} = S \cup \partial S$ *is the smallest closed set containing S.*

[9] Precise definition: $p \in S$ is an interior point if there exists some $r > 0$ such that the ball $B_r(p)$ in X is contained inside S.

1.10. **Maps between spaces.**

Given two spaces X, Y, a *map* $f\colon X \to Y$ eats a point $x \in X$ and spits out a point $y = f(x) \in Y$. When $Y = \mathbb{R}$ we often call f a *function*.

Examples.

(1) *The familiar functions* $f : \mathbb{R} \to \mathbb{R}$ *such as straight lines* $f(x) = mx + c$ *and the standard parabola* $f(x) = x^2$.

(2) *You cannot*[10] *take* $X = \mathbb{R}$ *when the function is not defined everywhere on* $\mathbb{R}$*, for example the positive square root function is*

$$f\colon [0, \infty) \to \mathbb{R},\ f(x) = \sqrt{x}.$$

(3) *You also do not allow* f *to spit out more than one value, so*

$$f\colon [-1, 1] \to \mathbb{R},\ f(x) = \sqrt{1 - x^2}$$

only traces out the upper half of the unit circle.

(4) *The standard ellipse with parameters* $a, b \in \mathbb{R}$*,*

$$f\colon \mathbb{R} \to \mathbb{R}^2,\ f(t) = (a\cos t,\, b\sin t).$$

Notice that the (x, y) *coordinates of the image points* $f(t)$ *trace out the set of solutions of the equation*

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

For $a = b = 1$ *you get the unit circle. Notice that* f *is not injective (one-to-one) since* $f(t + 2\pi) = f(t)$*, but we could make it injective by taking* $X = [0, 2\pi)$ *instead of* $\mathbb{R}$.

1.11. **Test maps out of X and into X.**

You can study a space X by considering the functions $f\colon X \to \mathbb{R}$ *out of* X.

Physically, X may be the surface of the Earth and f is a measurement such as temperature, altitude, pressure, etc.

You can also study a space X by considering maps *into* X. Such as paths

$$\mathbb{R} \to X \qquad \text{or} \qquad [0, 1] \to X;$$

or loops

$$S^1 = \{z \in \mathbb{C} : |z| = 1\} \to X;$$

or discs

$$\mathbb{D} = \{z \in \mathbb{C} : |z| \leq 1\} \to X,$$

or higher dimensional discs

$$\mathbb{D}^n = \{x \in \mathbb{R}^n : \|x\| \leq 1\} \to X.$$

In each case, the space $T = \mathbb{R}, [0, 1], S^1, \mathbb{D}, \mathbb{D}^n, \ldots$ is a *test space* you use, and the test maps into X from T, or out of X into T, will give you various information about the geometry and the topology of X.

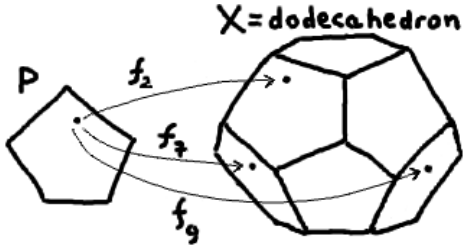
Physically, $f : \mathbb{R} \to X$ may be the motion of a particle in the universe X, so $f(t) \in X$ is the position at time $t \in \mathbb{R}$.

Examples.

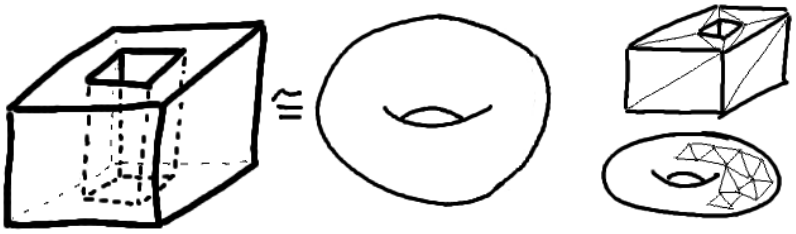
(1) *The previous example of an ellipse describes the motion of a planet around the Sun, and the parameters* a, b *of the ellipse depend on the planet.*

[10] In physics one often uses sentences such as "the function $\sqrt{x}$" or "the mass m". The mathematician always checks where these things exist: so $x \in [0, \infty)$ and $m \in (0, \infty)$. There's a joke: if you ask a mathematician to get in the car, the first thing the mathematician does is to check whether the car actually exists.

(2) *Take X to be the dodecahedron, the regular polyhedron with twelve pentagonal faces. Let P be a regular pentagon in the plane. Then you can define twelve maps $f_1, \ldots, f_{12} : P \to X$ which map the pentagon to the various faces of X.*



(3) *Take X to be a torus T^2: the surface of an American doughnut.*
A triangulation *of X is an identification $f : P \to X$ of X with a polyhedron made up of triangular faces. This means that we are ''dividing'' X into curved triangles, so that any edge belongs to exactly two curved triangles, and at any vertex the curved triangular faces fit together around the vertex in a cycle. For example, we can take P to be a cube, then puncture through it a small parallelepiped hole, and finally we divide the square/rectangular faces into triangles:*



In the last picture above, we show that you can draw a triangulation directly on the picture of the torus: just draw a tiling by curved triangles.

(4) *Most spaces X can be built up by gluing together discs. Consider, for example, the n-dimensional sphere. This is the boundary of the unit ball in $\mathbb{R}^{n+1}$:*

$$S^n = \partial \mathbb{D}^{n+1} = \{x \in \mathbb{R}^{n+1} : \|x\| = 1\}$$

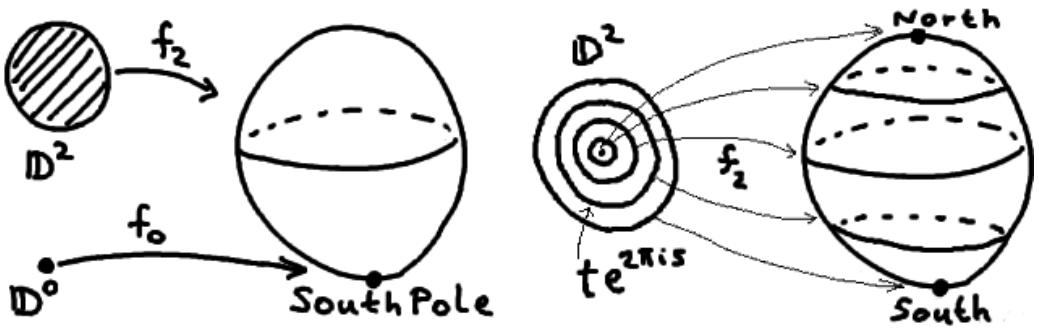
(for example, $S^1 = \partial \mathbb{D}^2$). It can be built up from two maps:

$$f_0 : \mathbb{D}^0 = \{point\} \to S^n \qquad f_n : \mathbb{D}^n \to S^n,$$

where f_0 maps the point to the South Pole $(0, \ldots, 0, 1)$ of S^n, and f_n wraps around the sphere by sending the boundary $\partial \mathbb{D}^n$ of the disc to the South Pole (''collapse the boundary to a point''). More explicitly, for example, in the case $n = 2$: think of the disc $\mathbb{D}^2$ as a family of circular paths

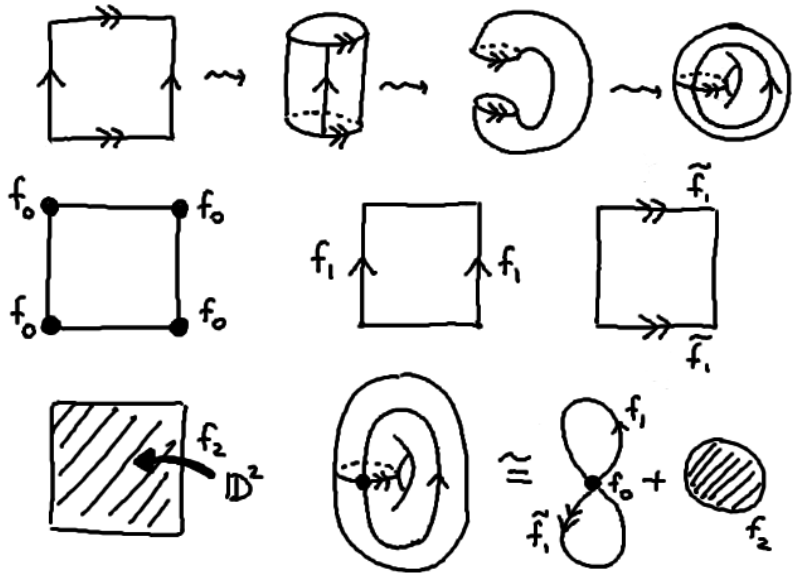
$$[0, 1] \to \mathbb{D},\ s \mapsto te^{2\pi is}$$

of radius $0 \leq t \leq 1$. For $t = 0$ you get a point, and you send that to the North Pole; for $t > 0$ you send these circles to the horizontal latitude circles of the sphere S^2 (parallel to the equator). For example, for $t = 1/2$ send the circle to the equator. As you approach $t = 1$ you send the circles to the small horizontal latitude circles around the South Pole. Finally for $t = 1$ you send the entire circle $\partial \mathbb{D}$ to the South Pole.



(5) *In general, a* cellular decomposition *of a space X involves producing a collection of such maps f_n from n-dimensional discs (with possibly several maps in the same dimension n), such that f_n maps the boundary $\partial\mathbb{D}^n$ into the $n-1$ dimensional space you have built so far (so you build X inductively by dimensions). You also want each map f_n to be injective on the interior $\operatorname{Int}\mathbb{D}^n$ (but it need not be injective on $\partial\mathbb{D}^n$, see the example above).*

For example, the torus T^2 is obtained by gluing a disc $f_2 : \mathbb{D}^2 \to T^2$ onto a figure 8 loop (which consists of one 0-cell $f_0 : \mathbb{D}^0 \to T^2$ and two 1-cells $f_1 : \mathbb{D}^1 \to T^2$ and $\tilde{f}_1 : \mathbb{D}^1 \to T^2$).



In the above picture, we first show how the torus arises from a square by gluing together opposite parallel sides (the arrows tell us in which direction to glue). Then we show on the square where the cells $f_0, f_1, \tilde{f}_1, f_2$ are. For example, the 4 vertices of the square all get glued together to give one point in T^2 which is our zero cell f_0.

(6) *This is the beginning of modern geometry. For example, consider the* Euler characteristic *χ, which you may have encountered for regular polyhedra as being defined as $V-E+F$, where V, E, F is the number of vertices, edges, faces in the polyhedron. This number turns out always to equal 2 because all regular polyhedra can be deformed*[11] *into the sphere S^2, which has $\chi(S^2) = 2$. Indeed, any triangulation of S^2 will give $\chi(S^2) = 2$.*

In general, for any space X, $\chi(X)$ is the alternating sum of the number

[11] If you have a balloon in the shape of a regular polyhedron, and you keep blowing it up, then it will deform into a sphere.

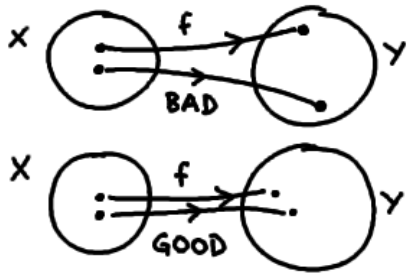
of cells in each dimension. So for the sphere: one 0-cell, zero 1-cells, one 2-cell, gives: $\chi = 1-0+1 = 2$. For the torus, above: one 0-cell, two 1-cells, one 2-cell, gives: $\chi = 1-2+1 = 0$. An important theorem is to show that χ is an invariant*: no matter how you express X in terms of a cellular decomposition, $\chi(X)$ will be the same integer.*

Exercise 7. *By drawing a triangulation, show that a doughnut with g holes instead of* 1 *hole, will have $\chi = 2 - 2g$.*

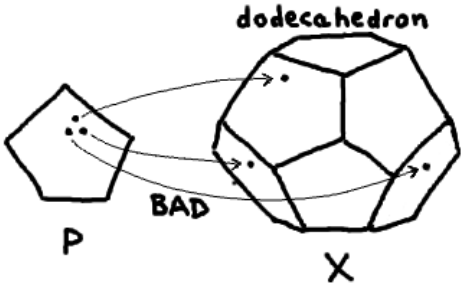
1.12. **Continuous maps.**

So far when defining maps $f : X \to Y$ we only used that X, Y were sets. We have not yet used that X, Y have a notion of distance. So which maps are good and which are bad?

If a map takes two points which are very close in X and tears them apart, spitting out two points which are very distant in Y, then of course the map f is bad. For a good map we expect that "close points map to close points":



Example. Let P be a regular pentagon in the plane, and X a dodecahedron. Let $f : P \to X$ be a map which sends each point of P to a randomly chosen point in one of the twelve pentagonal faces of X.[12]



Then f is a bad map: two points which are very close in P may end up far apart in two different faces of X.

So how do we decide precisely when a map is good? After all,

$$f : \mathbb{R} \to \mathbb{R},\ f(x) = 10^7 x$$

is a very good map (it's a straight line!), but the two close points $p = 0$, $q = 1/10$ map to $f(p) = 0$ and $f(q) =$ one million, which seem very far apart.

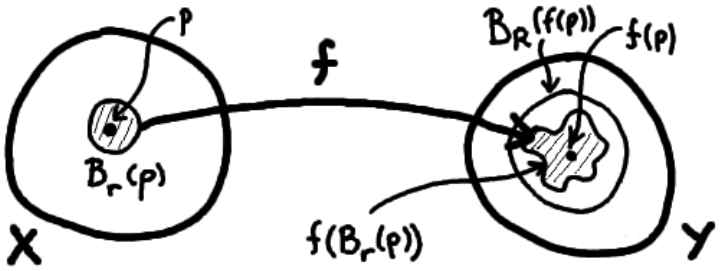
This issue arises because there is no precise notion of 'big' and 'small'. What we really want is that $f(p), f(q)$ are "close" in Y whenever we impose that p, q are "very close" in X. So in the above example: if we want $f(0) = 0$, $f(x) = 10^7 x$ to be within distance $1/100 = 10^{-2}$, it is enough if we require that $0, x$ are within distance 10^{-9}. We do not actually care about making the optimal requirement: we could also require that $0, x$ are within 10^{-10000}. As long as there is some requirement that works, we're happy.

So good map means: $f(x)$ is close to $f(p)$, if x is very close to p. Mathematically:

[12]More precisely, for each $p \in P$ we randomly assign a number $n(p) \in \{1, \dots, 12\}$ and we define $f(p) = f_{n(p)}(p)$ using the twelve maps $f_1, \dots, f_{12}$ mentioned in a previous example.

Definition 8. *A map $f : X \to Y$ is* continuous *if for any ball $B_R(f(p))$ in Y we can find a radius $r > 0$ such that the image in Y of the ball $B_r(p)$ in X fits inside $B_R(f(p))$:*

$$\boxed{f(B_r(p)) \subset B_R(f(p))}$$



Remarks 9. *If we want to show that a map is continuous, notice that: R is given to us, and we must find an r which works. In general, r will depend on R and on the points $p, f(p)$. However r is not allowed to depend on $x, f(x)$ for $x \neq p$.*

Examples.

(1) *A constant map $f : X \to Y$, $f(x) = z$ (for some fixed choice of point $z \in Y$) is always continuous. For any R we can take $r = 1$. Indeed:*

$$f(B_r(p)) = \{z\} \subset B_R(z) = B_R(f(p)).$$

(2) *The identity map $f : X \to X, f(x) = x$ is continuous. Indeed for $r = R$:*

$$f(B_r(p)) = B_r(p) \subset B_R(p) = B_R(f(p))$$

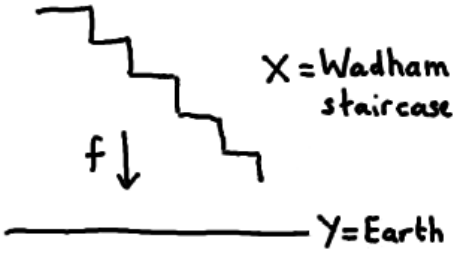
(3) *Linear functions $f : \mathbb{R} \to \mathbb{R}, f(x) = mx + c$ are continuous. Take $r = R/m$:*

$$f(B_r(p)) = B_{mr}(mp + c) \subset B_R(mp + c) = B_R(f(p)).$$

For example, $f(x) = 10^7 x$ is continuous since linear.

(4) *$f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^2+1$, then $f(x)-f(p) = x^2-p^2 = (x-p)(x+p)$. We want: $|x - p| < r$ implies $|x - p| \cdot |x + p| < R$. Exercise: what r do you pick? Careful: r must not depend on x. As you can see, already this simple example can become messy.*[13]

(5) *If X is a staircase in Wadham College (using as distance between two points, the least length of a piece of string connecting the points), and Y is the surface of the Earth underneath Wadham College, then the projection map $f : X \to Y$, which projects a point of the staircase vertically onto the surface of the Earth, is continuous.*



[13] *Hint.* How big can $|x + p|$ be? Here is a trick. First, find r_1 such that if $|x - p| < r_1$ then $|x + p| < m$ (for example take $m = 1$). Then, as in the previous example, $r_2 = R/m$ ensures that: if $|x - p| < r_2$ then $|x - p| \cdot m < R$. Finally, take $r = \min\{r_1, r_2\}$.

(6) *The operation of rescaling* $g_k : \mathbb{R} \to \mathbb{R}$, $g(x) = kx$ *by some* $k \neq 0 \in \mathbb{R}$ *is continuous. Indeed, take* $r = R/k$*:*

$$g_k(B_r(a)) = kB_r(a) = B_{kr}(ka) \subset B_R(ka).$$

(7) *The operation of addition* $g_+ \colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$, $g_+(a_1, a_2) = a_1+a_2$ *is continuous. Take* $r = R/2$ *then*

$$g_+(B_r(a_1), B_r(a_2)) = B_r(a_1) + B_r(a_2) \subset B_R(a_1 + a_2),$$

where the inclusion follows by the triangle inequality:

$$|x_1 + x_2 - (a_1 + a_2)| \leq |x_1 - a_1| + |x_2 - a_2| < r + r = R.$$

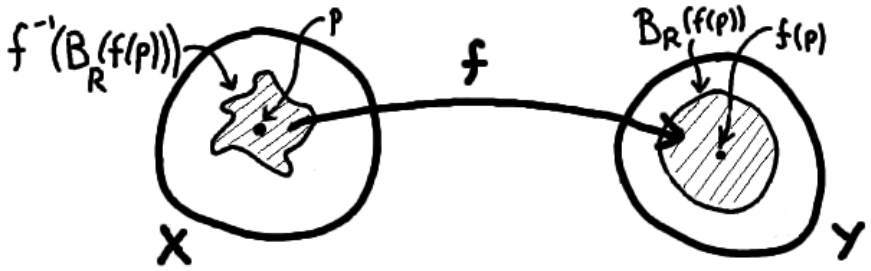
(8) *The operation of multiplication* $g_\times : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $g_\times(a_1, a_2) = a_1 \cdot a_2$ *is continuous (Exercise).*

(9) $f : \mathbb{R} \to \mathbb{R}$, $f(x) = 1$ *for* $x \neq 0$, *and* $f(0) = 0$. *Then* f *is not continuous near* $p = 0$. *Indeed:* $f(0) = 0$, *so* $B_{1/2}(f(0)) = (-\frac{1}{2}, \frac{1}{2})$ *but* $f(B_r(0))$ *will never fit inside* $(-\frac{1}{2}, \frac{1}{2})$ *since* $1 \in f(B_r(0))$ *for any* $r > 0$.

(10) *Intuitively speaking, a function* $f : \mathbb{R} \to \mathbb{R}$ *is continuous if you can draw its graph without lifting the pen off the paper (there are no ''jumps''). However, when* f *is not defined on all of* $\mathbb{R}$, *then this intuition may be deceptive: for example,* $f(x) = \frac{1}{x}$ *is a continuous function* $f : \mathbb{R}\backslash 0 \to \mathbb{R}$ *(you cannot pick* $p = 0$ *since it does not belong to the domain of definition).*

(11) *The map* $f : (0, 1) \to S^1, x \mapsto e^{2\pi ix}$ *is continuous: given a ball in* S^1 *around* $f(p)$ *we can fit* $f($*tiny interval around* $p)$ *inside it.*

Theorem 10 (See Exercises)**.** *A map* $f : X \to Y$ *is continuous if and only if the preimage* $f^{-1}(B_R(f(p)))$ *of any ball around* $f(p)$ *is a neighbourhood of* p *in* X*.*



Lemma 11. *If* $f : X \to Y$, $g : Y \to Z$ *are continuous, then the composition* $g \circ f : X \to Y \to Z$ *is continuous.*

Proof. Since g is continuous, the preimage via g of any ball around $g(f(p))$ in Z is a neighbourhood of $f(p)$ in Y. By the above Observation, this neighbourhood contains a ball around $f(p)$. Since f is continuous, the preimage via f of this new ball is a neighbourhood of p in X. Which again by the observation contains a ball $B_r(p)$. Thus $g \circ f(B_r(p)) \subset B_R(g \circ f(p))$. □

Example. When you have two maps, $f : X \to Y$, $g : X \to Z$, *you sometimes want to consider both maps simultaneously. In that case, you consider*

$$h : X \to Y \times Z,\ h(x) = (f(x), g(x)),$$

In the exercises you will show the following:

- *If* f, g *are continuous then so is* h*;*
- *Using the maps* $g_k, g_+, g_\times, h$ *above, you can view* kf, $f+g$, $f\cdot g$ *as compositions of continuous functions, so they are continuous by the above Lemma;*

- *Inductively this shows that taking sums/differences, products, and rescalings of finitely many continuous functions gives a continuous function;*
- *In particular any polynomial*

$$f : \mathbb{R} \to \mathbb{R}, f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

is continuous.

This abstract approach avoids the messy argument involved in finding r in terms of R and p (see the example $f(x) = x^2+1$ above). Mathematicians should always aim to find an elegant route, if the normal route is messy.

Warning. One often says "map" to mean "continuous map", to save ink. In the rare situations where one really does need to use a discontinuous map, one typically emphasizes that it is not continuous.

1.13. **An introduction to Limits via examples.**

Write $x \to p$ as an abbreviation for the sentence "the point x approaches the point p" (inside a space X), meaning that the distances $d(x, p)$ are approaching zero as we vary x in some family of points. Often we say "x *converges* to p".

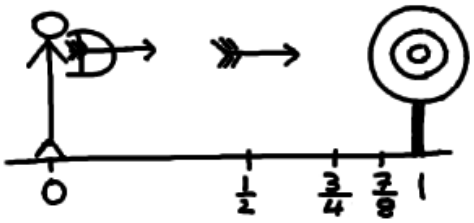
Examples.

(1) $x = \frac{1}{n} \to 0$ *in* $\mathbb{R}$ *as* n *grows to infinity (abbreviated:* $n \to \infty$*).*

(2) *As* $x \to 3$, $f(x) = x^2 + 1 \to f(3) = 3^2 + 1 = 10$. *It is a general feature of* continuous *maps that* $f(x) \to f(p)$ *as* $x \to p$.

(3) $f : \mathbb{R} \to \mathbb{R}$, $f(x) = 1$ *for* $x \neq 0$, *and* $f(0) = 0$ *(recall this f is not continuous). As* $x \to 0$, $f(x) = 1 \to 1$ *which does not equal* $f(0) = 0$.

(4) *An archer shoots an arrow towards a target. Assume that no air resistance exists... and that no gravity exists!*



The physicist says, ''let's take as unit of length the distance between the archer and the target, then the arrow will travel distance 1*''. A philosopher says, ''I'm not sure the arrow will ever hit the target, because the arrow must first travel half the distance,* $1/2$*, then it will travel half of the remaining distance,* $1/4$*, then half of what remains thereafter,* $1/8$*, and so on. So at each stage, the arrow has half of the remaining distance left to travel, so it never reaches the target''. The physicist scratches his head in perplexity, so the philosopher removes the target and yells at the archer ''Shoot! You can't possibly hit me!''. After the inevitable accident, the mathematician explains to the philospher in hospital: ''actually the infinite sum* $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^n} + \cdots$ *makes sense: consider the values you get when you take the partial sums:*

$$\frac{1}{2}, \frac{3}{4}, \frac{7}{8}, \frac{15}{16}, \frac{31}{32}, \cdots$$

At the n-th stage you get $1-\frac{1}{2^n}$ *and this converges to* 1 *as* $n \to \infty$. *So the infinite sum equals* 1. *That's why you got hit by the arrow''.*

(5) *Not all infinite sums make sense. For example* $1+1+1+1+\cdots$ *does not converge to a number (infinity is not a number). Another example:* $1-$

$1+1-1+1-1+\cdots$ *does not converge since you cannot decide whether the limit should be* 1 *or* 0*.*

Exercise 12. *Find a sequence of real numbers $x_1, x_2, x_3, \ldots$ such that: given any real number r you can rearrange the sequence: $x_{i_1}, x_{i_2}, x_{i_3}, \ldots$ (so each x_j appears exactly once), so that the infinite sum converges to r:*

$$x_{i_1} + x_{i_2} + x_{i_3} + \cdots \to r.$$

1.14. **Continuity in terms of limits.**

Theorem 13. *$f : X \to Y$ is continuous if and only if $f(x_n) \to f(p)$ for any sequence $x_n \to p$.*

Remark 14. *We clarify the definition of $x_n \to p$. It means that the sequence x_n is eventually inside any ball around p for large enough n. Precisely: for any $r > 0$, there is some $N \in \mathbb{N}$ so that $x_n \in B_r(p)$ for all $n \geq N$. Usually N depends on r. Similarly $f(x_n) \to f(p)$ means: for any $R > 0$, we can find $N \in \mathbb{N}$ so that $f(x_n) \in B_R(f(p))$ for all $n \geq N$.*

Proof of the Theorem. Let's now prove the direction "$\Rightarrow$" of the Lemma. If we want $f(x_n)$ to be within distance R from $f(p)$, then we know (by continuity) that for some r we just need to ensure that x_n is within distance r from p, and this holds for $n \geq N$ for some large N since $x_n \to p$.

Now we prove the direction "$\Leftarrow$". Suppose f is not continuous, by contradiction. Then for some $R > 0$, no matter how small we pick $r > 0$, the following inclusion **fails**:

$$f(B_r(p)) \subset B_R(f(p)).$$

So there is a bad point $x_r \in B_r(p)$ with $f(x_r) \notin B_R(f(p))$. As we let r vary in a family which decreases to zero (for example, take the sequence $r = 1/n$), then by construction $x_r \to p$. So, by assumption, $f(x_r) \to f(p)$. But $f(x_r) \notin B_R(f(p))$ so it cannot get close to p! Contradiction. □

2. When do we want to think of two spaces as being the same?

2.1. **Various notions of equality.**

There are several options for what it means for two spaces X, Y to be the same:

(1) **Equality**: if they are equal $X = Y$. But this is already too harsh for sets: for example, do we really want to think of $\{A, B, C\}$ and $\{a, b, c\}$ as different?
(2) **Bijection**: if there is a bijection[14] $f : X \to Y$. For example, above, send A, B, C to a, b, c respectively. This is a good notion for sets, but for spaces it ignores distances.
(3) **Continuous bijection**: if there is a continuous bijection $f : X \to Y$. But then we would be identifying some spaces which should be thought of as different. For example: $[0, 2\pi) \to S^1, x \mapsto e^{ix}$ is a continuous bijection, but we don't want to think of an interval as being the same as a circle. Another bad example: $[0, 1] \cup (2, 3] \to [0, 2]$ defined on the first interval by $x \mapsto x$ and on the second interval by $x \mapsto x - 1$. This is a continuous bijection.
(4) **Homeomorphism**: if there is a continuous bijection $f : X \to Y$ such that the inverse $f^{-1} : Y \to X$ is also continuous. This is called a *homeomorphism*. The previous two examples will fail, because the inverse functions $S^1 \to [0, 2\pi)$ and $[0, 2] \to [0, 1] \cup (2, 3]$ are not continuous near 1. This notion is good if you allow your space to be continuously deformed in a way that you can also continuously

[14] A map $f : X \to Y$ is a *bijection* if it is injective (one-to-one: no two points in X map to the same point in Y) and surjective (onto: each point of Y is the image of some point from X). Exercise: Show that $f : X \to Y$ is a bijection if and only if there are two maps $f : X \to Y$, $g : Y \to X$ such that the composites $f \circ g$ and $g \circ f$ are the identity maps. The map g is called the inverse of f, and is written f^{-1}.

undo the deformation. However, notice that distances do not need to be preserved. For example, D with the hyperbolic distance and D with the Euclidean distance are homeomorphic via the identity map $D \to D, z \mapsto z$.

(5) **Isometry**: homeomorphisms $f : X \to Y$ which preserve lengths, so $d_Y(f(x), f(x')) = d_X(x, x')$. These are called *isometries*. They arise often in geometry and physics, and are sometimes called *symmetries*. This notion is good when working for example with tilings. This is quite a rigid property: there are often very few isometries. For example, we mentioned in Section 1.5 the symmetries of the hyperbolic disc D.

Examples.

(1) *In $\mathbb{R}$ with the usual distance, the isometries have the form*

$$f(x) = x + \textit{constant} \quad \textit{or} \quad f(x) = -x + \textit{constant}.$$

(2) *Any strictly increasing (or strictly decreasing) continuous function $f : \mathbb{R} \to \mathbb{R}$ is a homeomorphism. For example $f(x) = x^3$ is a homeomorphism (but not an isometry).*

(3) *In $\mathbb{R}^2$ with the usual distance, all the isometries are obtained by composing a translation*[15] *with a rotation or a reflection. So, identifying $\mathbb{R}^2 = \mathbb{C}$ and writing $\overline{z}$ for the complex conjugate, the isometries of $\mathbb{R}^2$ are:*

$$z \mapsto e^{i\alpha} z + c \quad \textit{or} \quad z \mapsto e^{i\alpha}\overline{z} + c \qquad (\textit{for constants } \alpha \in \mathbb{R}, c \in \mathbb{C})$$

2.2. **Homeomorphisms.**

For our purposes, the notion of isometry is too strong. We want to allow continuous deformations (like stretching a rubber band). So we will study *homeomorphisms* $f : X \to Y$, namely bijections f such that f, f^{-1} are both continuous.

Examples.

(1) $[0, 1] \to [0, 2], x \mapsto 2x$ *is a homeomorphism;*

(2) $[0, 1] \to [0, 1], x \mapsto x^2$ *is a homeomorphism, and its inverse is the homeomorphism $x \mapsto \sqrt{x}$;*

(3) *Every polygon (including the inside) is homeomorphic to a disc.*[16] *The picture to have in mind is like that of blowing up a balloon: you can stretch and pull, you can even make corners, as long as you don't tear.*

(4) *A sphere is homeomorphic to a tetrahedron (and to any regular polyhedron). Just stretch the tetrahedron outwards, towards a big sphere that contains the tetrahedron, like blowing up a balloon.*

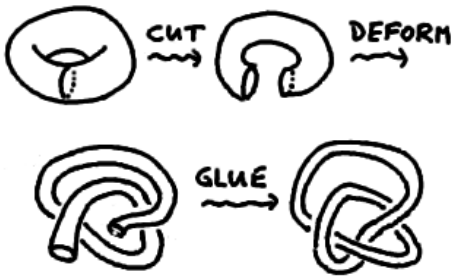
(5) *If you take a surface S, make a cut, then continuously deform, and later glue the cut up exactly as it was before, then the result is homeomorphic*[17] *to S. For example, the following knotted doughtnut is homeomorphic to the usual doughnut:*

---

[15] adding constants to the x, y coordinates: $f(x, y) = (x + a, y + b)$.

[16] This uses the following ideas. First notice that the map $[0, a] \to [0, 1], x \mapsto x/a$ is a homeomorphism (for $a > 0$ fixed). Similarly in $\mathbb{R}^2$ the map $(x, y) \to \frac{1}{a}(x, y)$ is a homeomorphism from the line segment joining $(0, 0)$ and $a \cdot (x_0, y_0)$ to the line segment joining $(0, 0)$ and (x_0, y_0). To build the homeomorphism from the polygon to the disc, along each ray you use such a stretching map. The value of a will depend on the angle that the ray makes with the positive real axis, but a will depend continuously on the angle, so that's not a problem.

[17] check this: pick any point p on S, draw a little disc around it, and follow how the disc changes under the transformation – at the end, you should get another deformed disc.



Indeed, the ''knotting'' of the surface is a property of the ambient in which we view S (here $S \subset \mathbb{R}^3$): it is not something intrinsic about S. That's why homeomorphisms do not detect it. More existentially, as an ant living in this doughnut universe, you wouldn't know that someone has messed with your universe since you are not aware that there is more space ''outside'' of your universe.

Exercise 15. *In that last example, can you prove that you cannot knot the doughnut above without making a cut (that is, you cannot knot it by just using a continuous deformation)?*[18]

3. Using continuous maps to understand the topology

3.1. **Homeomorphic spaces cannot be distinguished using continuous functions.**

Lemma 16. *If $f : X \to Y$ is a homeomorphism, then the continuous maps into X and out of X are the "same" as those for Y.*

Proof. Any map $\mathrm{In} : T \to X$ from a test space T into X gives rise to a map into Y by composition:

$$f \circ \mathrm{In} : T \to X \to Y.$$

Similarly, any map $\mathrm{Out} : X \to T$ from X into a test space T gives rise to a map from Y by composition:

$$\mathrm{Out} \circ f^{-1} : Y \to X \to T.$$

So, phrasing the claim more precisely, we are saying that there are natural bijections between the spaces of continuous maps from/to a test space T:

$$C(T, X) \to C(T, Y), \mathrm{In} \mapsto f \circ \mathrm{In}, \qquad C(X, T) \to C(Y, T), \mathrm{Out} \mapsto \mathrm{Out} \circ f^{-1}.$$

You can easily check that these are bijections.[19] □

Example. For $f : [0, 1] \to [0, 2\pi], f(x) = 2\pi x$, the map $\mathrm{Out} : [0, 2\pi] \to T = S^1$, $\mathrm{Out}(x) = e^{ix}$ corresponds to the map $\mathrm{Out} \circ f : [0, 1] \to T = S^1$, $\mathrm{Out}(f(x)) = \mathrm{Out}(2\pi x) = e^{2\pi i x}$. Physicists like their angles to go from 0 to 2π, but mathematicians like to think of the circle as $[0, 1]$ with endpoints identified. The above shows that it won't matter what you prefer: you will detect the ''same'' continuous maps.

3.2. **Connected spaces.**

Let's study the simplest case of maps out of a space X: consider maps into the test space $\mathbb{Z}$ (with the usual Euclidean distance).

A space X is called *connected* if any continuous integer-valued function

$$f : X \to \mathbb{Z}$$

[18] Hint. Look up "trefoil knot" in Wikipedia.

[19] *Hint.* Guess what the inverse maps are supposed to be, then prove that they are the inverse maps by showing that certain compositions are the identity map. *More hints.* $C(T, X) \to C(T, Y) \to C(T, X), \mathrm{In} \mapsto f \circ \mathrm{In} \mapsto f^{-1} \circ (f \circ \mathrm{In}) = \mathrm{In}$, which is the identity map.

is always constant.[20] Intuitively, this is saying that locally constant functions must actually be globally constant.

Examples.

(1) *The unit interval* $[0,1]$ *is connected. The key idea to prove this, is that two integers can only be very close if they are equal!*[21]
(2) *A circle is connected. Proof: Given* $f : S^1 \to \mathbb{Z}$*, remove a tiny interval around* 1 *from* S^1*. Call that* X'*. Notice* X' *is homeomorphic to an interval. So the restriction* $f|_{X'} : X' \to \mathbb{Z}$ *to* X' *must be constant since the interval is connected. So* f *is constant everywhere on* S^1 *except possibly near* 1*. Similarly, if you removed a tiny interval around* -1 *you would deduce that* f *is constant everywhere except possibly near* -1*. Combining, we deduce that* f *is constant.*
(3) *The disjoint union of two circles is disconnected: the function which takes value* 0 *on the first circle and* 1 *on the second circle is continuous.*
(4) *If you join two circles at one point, you obtain a ''figure eight''. This is connected. Proof. Given* f : (*figure 8*) $\to \mathbb{Z}$*, the restriction of* f *to each of the two circles must be constant since the circle is connected. The two constants you get must agree since* f *needs to take the same value at the point where the circles are joined.*
(5) *If you remove the joining point in the figure eight, you obtain a disconnected space (it's a disjoint union of two open intervals).*

3.3. **Path-connected spaces.**

Let's consider a simple case of maps into X: consider maps from the test space $[0,1]$ (with the usual Euclidean distance).

A space X is called *path-connected* if for any two given points p, q there is a continuous map

$$f : [0,1] \to X,\ f(0) = p,\ f(1) = q.$$

Intuitively, f is a continuous path joining p to q.

Lemma 17. *If a space is path-connected, then it must be connected.*

Proof. Suppose by contradiction that there is some path-connected space X which is not connected. So there is a non-constant continuous function $g : X \to \mathbb{Z}$. Now consider a continuous path $f : [0,1] \to X$ joining two points $p = f(0), q = f(1)$ for which $g(p) \neq g(q)$. Then the composite $g \circ f : [0,1] \to X \to \mathbb{Z}$ is continuous and non-constant, contradicting that $[0,1]$ is connected. □

Examples.

(1) *The above examples of connected spaces are also path-connected.*
(2) *Can you think of a connected space which is not path-connected? (Exercise)*

3.4. **Fillability of maps, and simply connected spaces.**

Notice that the 0-dimensional sphere consists of two points:

$$S^0 = \{\pm 1\} = \{x \in \mathbb{R}^1 : |x| = 1\} \subset \mathbb{D}^1 = \{x \in \mathbb{R} : |x| \leq 1\} = [-1,1].$$

[20] You could also replace $\mathbb{Z}$ by two points $\{0,1\}$ in the definition: can you see why?

[21] Proof: given $f : [0,1] \to \mathbb{Z}$, if x' is sufficiently close to x then by continuity $f(x')$ is arbitrarily close to $f(x)$. But $f(x), f(x')$ are integers, so if they are within distance $1/2$ then they are equal. Now we use a trick: consider the set $S = \{s \in [0,1] : f(x) = f(0) \text{ for all } x \in [0,s]\}$. We know that small enough $s > 0$ are in S by continuity of f near $x = 0$. Check that S is an interval. Finally check that the boundary of S must be 0 and 1 (*Hint. if instead of* 1 *you had the boundary point* $x > 0$*, then use continuity at* x *to deduce that* S *contains a small interval around* x *contradicting that* $x \neq 1$ *is a boundary point*).

So if $f : S^0 \to X$ sends $f(-1) = p$ and $f(+1) = q$, then asking whether there is a path $F : [-1, 1] \to X$ joining p to q is the same as asking whether there is a *filling* $F : \mathbb{D}^1 \to X$ which restricts to f on the boundary $S^0 = \partial\mathbb{D}^1$. So a space is *path-connected* if and only if every map $f : S^0 \to X$ can be filled by a continuous map $F : \mathbb{D}^1 \to X$ with $F|_{S^0} = f$.

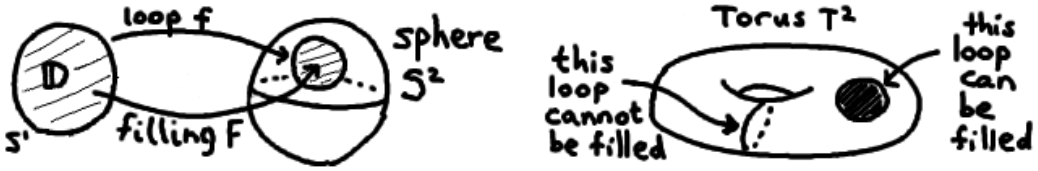
Similarly, whenever you have a continuous map

$$f : S^1 \to X,$$

you can ask whether there is a continuous map $F : \mathbb{D} = \mathbb{D}^2 \to X$ which restricts to f along the boundary $S^1 = \partial\mathbb{D}$. In symbols:

$$\boxed{F : \mathbb{D} \to X,\ F|_{S^1} = f}$$

Geometrically, you are asking whether there is a disc in X which fills in the circle $f(S^1) \subset X$.



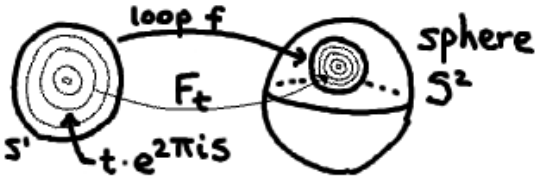
Notice we do not require f or F to be injective. For example, a constant loop $f = p \in X$ is always fillable by the constant disc $F = p$.

One can view the disc $\mathbb{D} = \{z \in \mathbb{C} : |z| \leq 1\}$ as a family of circles of shrinking radii:

$$[0, 1] \to \mathbb{D},\ s \mapsto te^{2\pi is}$$

is a circular subset of $\mathbb{D}$ of radius t, with $0 \leq t \leq 1$, and centre 0. At time $t = 0$ this is just a point (a circle of zero radius).

Fillability is therefore asking whether the loop $f(S^1) \subset X$ can be continuously deformed (through loops) to a point.



The family $(F_t)_{0 \leq t \leq 1}$ of loops is given by

$$F_t : S^1 \to X,\ F_t(e^{2\pi is}) = F(te^{2\pi is})$$

and these loops contract down to the point $F(0) \in X$ obtained for $t = 0$. Notice $F_1 = f$ is the original loop we wanted to fill.

A space is called *simply connected* if it is connected and the above filling property always holds (that is, any loop can be continuously deformed to a point).

Example.

(1) $[0, 1]$ *is simply connected: take*

$$F_t(s) = t \cdot f(s).$$

Notice $F_1 = f$ *and* $F_0 =$ *the point* 0. *The same works for the discs* $\mathbb{D}, \mathbb{D}^n$ *and for* $\mathbb{R}, \mathbb{R}^n$.

(2) *Any convex subset*[22] *of* $S \subset \mathbb{R}^n$, *is simply connected: take*

$$F_t(z) = tf(z) + (1 - t)y$$

[22] $S \subset \mathbb{R}^n$ is convex if the straight line segment $(tp + (1-t)q)_{0 \leq t \leq 1}$ joining any two points $p, q \in S$ always lies entirely in S.

where y is a fixed chosen point of S. Again notice $F_1 = f$ and $F_0 =$ the point y.

(3) *It turns out that S^1 is not simply connected (we will prove this when we study winding numbers). However, the 2-sphere S^2 and more generally $S^n = \partial\mathbb{D}^{n+1}$ for $n \geq 2$ are simply connected.*

Sketch Proof that S^2 is simply connected: if you can find a point $p \in S^2$ which is not on the loop, then you can fill the loop. Indeed, $S^2 \setminus p$ is homeomorphic to $\mathbb{R}^2$ (Exercise), and we know that loops in $\mathbb{R}^2$ can be filled! However, as you will see in the Exercises, there are (continuous!) loops which pass through every point of S^2. The trick around this involves three ideas:

(a) *if two loops are very close*[23] *then you can continuously deform one into the other;*[24]

(b) *continuous functions can be approximated arbitrarily well by polynomials, similarly continuous loops $f : S^1 \to S^2$ can be approximated arbitrarily well by a polygonal path $g : S^1 \to S^2$;*

(c) *thus we can deform any loop f into a polygonal loop g, and since polygonal loops obviously don't pass through every point of S^2, we can fill g, meaning we can further deform g into a point.*

(4) *The annulus $A = \{z \in \mathbb{C} : 1 \leq |z| \leq 2\}$ is not simply connected. Here is a trick to prove this using the previous example. Notice that the unit circle $S^1 = \{z \in \mathbb{C} : |z| = 1\}$ is a subset of A. So there is an inclusion of the circle into the annulus:*

$$\mathrm{In} : S^1 \to A,\ z \mapsto z.$$

There is also a map from the annulus into the circle, by ''squashing'' the width of the annulus so that it becomes a circle:

$$\mathrm{Out} : A \to S^1,\ z \mapsto \frac{z}{|z|}.$$

They key property that we will use about the In-and-Out maps, is that their composite is the identity map on the circle:

$$\mathrm{Out} \circ \mathrm{In} = \mathrm{Id} : S^1 \to A \to S^1,\ z \mapsto z \mapsto \frac{z}{|z|} = z.$$

Now suppose that $f : S^1 \to S^1$ is any loop inside the circle. We can produce a loop in the annulus by composition:

$$\mathrm{In} \circ f : S^1 \to A.$$

Suppose by contradiction that A is simply connected. Therefore the loop $\mathrm{In} \circ f$ can be filled. So there is a filling map

$$F : \mathbb{D} \to A \text{ with } F|_{S^1} = \mathrm{In} \circ f.$$

Now we can use the Out-map to produce a filling in the circle:

$$\mathrm{Out} \circ F : \mathbb{D} \to S^1.$$

Let's check this really is a filling of f:

$$\mathrm{Out} \circ F|_{S^1} = \mathrm{Out} \circ \mathrm{In} \circ f = \mathrm{Id} \circ f = f.$$

[23] Two loops $f : S^1 \to S^2$, $g : S^1 \to S^2$ are close if $\|f(e^{2\pi i s}) - g(e^{2\pi i s})\|$ is small for all $s \in [0, 1]$.

[24] Indeed just move the point $f(e^{2\pi i s})$ to the nearby point $g(e^{2\pi i s})$ by following the shortest arc that connects them.

Thus, if A really was simply connected, then the In-and-Out trick would show that every loop in S^1 can be filled. But the previous example says that this is not true. So A is not simply connected.

(5) *The torus $T^2 = S^1 \times S^1 = \{(z_1, z_2) \in \mathbb{C} \times \mathbb{C} : |z_1| = 1, |z_2| = 1\}$ (the surface of an American doughnut[25]) is not simply connected. You can prove this again by the In-and-Out trick, using* $\mathrm{In} : S^1 \to T^2, z \mapsto (z, 1)$ *and* $\mathrm{Out} : T^2 \to S^1, (z_1, z_2) \mapsto z_1$. *You can use this trick also to find explicit examples of non-contractible loops: any latitude circle a and any longitude circle b cannot be shrunk to a point.*

Exercise. How about the loop $aba^{-1}b^{-1}$ (meaning: latitude circle, followed by longitude circle, then go along the latitude circle in reverse, and the longitude circle in reverse.) Can you show that this is fillable?[26]

3.5. **Fillability of higher-dimensional spheres, Brower's fixed point theorem.** More generally, for

$$f : S^n = \partial \mathbb{D}^{n+1} \to X$$

you can ask if there is a filling $F : \mathbb{D}^{n+1} \to X$, that is a continuous map which restricts to f along the boundary. Geometrically, you are asking if you can deform the sphere $f(S^n)$ through spheres $F_t(S^n)$ down to a point.

Exercise 18. *Show that you can view the disc $\mathbb{D}^{n+1}$ as a family of spheres of varying radius $0 \leq t \leq 1$ (for $t = 0$ you a sphere of zero radius, which is just a point). Deduce that asking whether $f : S^n \to X$ has filling $F : \mathbb{D}^{n+1} \to X$ is the same as asking whether you can continuously deform the sphere f down to a point.*

Example.

(1) *As usual, for $[0, 1]$, $\mathbb{D}$, $\mathbb{D}^n$, $\mathbb{R}$, $\mathbb{R}^n$ and in general any convex subset of $\mathbb{R}^n$, you can always fill. Use the usual formula $F_t(z) = tf(z) + (1 - t)y$, for any fixed chosen point y in the space, where z is the varying point in S^n.*

(2) *It turns out that any map $f : S^n \to S^1$ for $n \geq 2$ can be filled by some $F : \mathbb{D}^{n+1} \to S^1$. It also turns out that any map $f : S^n \to T^2$ for $n \geq 2$ can be filled by some $F : \mathbb{D}^{n+1} \to T^2$.*

(3) *It turns out that there are maps $S^n \to S^n$, such as the identity map, which cannot be filled by any $F : \mathbb{D}^{n+1} \to S^n$.*

Exercise 19 (Brower's Fixed Point Theorem)**.** *Brower's theorem states that any continuous map $c : \mathbb{D}^{n+1} \to \mathbb{D}^{n+1}$ must have a fixed point, that is a point x for which $c(x) = x$.*

Any $n \geq 0$ works, but it helps to draw pictures in the plane (so $n = 1$). Prove the theorem as follows. Suppose by contradiction that the theorem is false. Deduce that you can then define a map $\mathrm{Out} : \mathbb{D}^{n+1} \to S^n$ *by letting* $\mathrm{Out}(x) \in S^n$ *be the point where the straight ray from $c(x)$ through x intersects the boundary $S^n = \partial \mathbb{D}^{n+1}$. Check that* Out *is the identity map on $S^n \subset \mathbb{D}^{n+1}$. Finally, define an obvious inclusion* $\mathrm{In} : S^n \to \mathbb{D}^{n+1}$. *Use the In-and-Out trick, and the previous examples, to deduce that any map $S^n \to S^n$ can then be filled by $F : \mathbb{D}^{n+1} \to S^n$. But as mentioned in the above examples, this is false for the identity map $S^n \to S^n$. Contradiction. So Brower's theorem must be true.*

[25] Recall we showed that T^2 can be obtained from a square by gluing together opposite sides. The two sides are intervals $[0, 1]$ and gluing their ends gives rise to a circle S^1, that is where the two S^1 factors in $T^2 = S^1 \times S^1$ come from. More geometrically, pretend that the Earth was a doughnut: then the circles of latitude are the circles $S^1 \times$ constant, and the circles of longitude are constant $\times S^1$ (in the square which glues to become the torus, these are respectively the horizontal and the vertical line segments inside the square). Any point on the Earth is specified uniquely by a latitude and a longitude coordinate (in the square which gets glued, any point is uniquely specified by the (x, y) coordinates modulo $\mathbb{Z} \times \mathbb{Z}$).

[26] *Hint.* We discussed an interesting disc $f_2 : \mathbb{D} \to T^2$ in the a cellular decomposition of the torus.

3.6. **Applications of fillability: telling spaces apart, interactions of strings.**

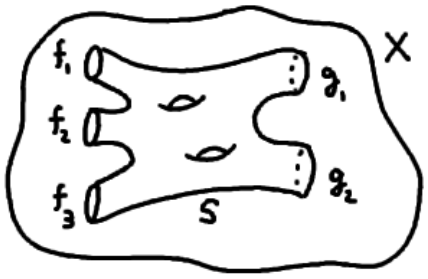
One reason for caring about fillings is that it allows you to tell spaces apart.

Example. The 2-sphere S^2 and the torus T^2 are not homeomorphic, because we saw that loops $f : S^1 \to S^2$ are fillable (by discs) but there are loops $f : S^1 \to T^2$ which are not. But homeomorphic spaces have the ''same'' continuous maps. So S^2, T^2 cannot be homeomorphic.

Asking whether circles and spheres can be filled with dics and balls is the beginning of *Homotopy Theory*.

One can also ask about more complicated fillings. For example, remove a small disc around a point of the torus. The surface you obtain, $S = T^2 \setminus \mathbb{D}$, has boundary S^1. So you can ask whether a loop $f : S^1 \to X$ in a space X can be filled by a map $F : S \to X$, with $F|_S = f$. Thinking about such more general fillings is the beginning of *Homology Theory* and of *Cobordism Theory*.

If you think of the universe as being made up of strings (that is, loops $S^1 \to X$), you may ask whether two strings $f_0 \sqcup f_1 : S^1 \sqcup S^1 \to X$ can be filled by a cylinder $F : [0,1] \times S^1 \to X$ (so $F(0,z) = f_0(z)$ and $F(1,z) = f_1(z)$). This is physically relevant: it answers whether f_0 can evolve in time to become f_1. You think of $z \mapsto F(t,z)$ as the loop at time t in your evolution cylinder.

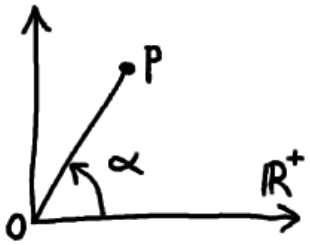


More generally, you may consider surfaces S (possibly with doughnut holes) which have n incoming circles and m outgoing circles. So you ask whether strings $f_1, \dots, f_n, g_1, \dots, g_m$ are fillable by $F : S \to X$. This is physically relevant: you are asking whether the strings $f_1, \dots, f_n$ can evolve and interact so as to turn into the strings $g_1, \dots, g_m$.

These ideas arise in *String Topology* and in *Topological Quantum Field Theory*.

4. Winding numbers

4.1. **The angle functions $\alpha_n^{\pm}$.** In the plane $\mathbb{R}^2$, for points[27] $p \neq 0$, one can define the angle between the positive x-axis and the line segment joining 0 to p.



However, it is less clear how you define $\alpha(p(t))$ for a moving point $p(t)$. The problem is that there does not exist a well-defined continuous angle function

$$\alpha : \mathbb{R}^2 \setminus 0 \to \mathbb{R}.$$

Example. Consider a point moving around the circle $p(t) = (\cos 2\pi t, \sin 2\pi t)$ for $0 \leq t \leq 1$. How do you define $\alpha(p(t))$? The natural choice is $\alpha(p(t)) = 2\pi t$. Now

[27]For $p = 0$ we cannot define an angle.

$p(0) = p(1) = (1,0) \in \mathbb{R}^2$, so α should take the same value at $p(0)$ and at $p(1)$. But $\alpha(p(0)) = 0$ and $\alpha(p(1)) = 2\pi$. So $\alpha(p(t))$ is badly defined.

There are several options to deal with this problem:

(1) **Multivalued functions**: the angle $\alpha : \mathbb{R}^2 \setminus 0 \to \mathbb{R}$ is not a function, rather we assign multiple values. For example $\alpha(1,0)$ is $2\pi\mathbb{Z}$, any integer multiple of 2π. So the values of α are subsets of $\mathbb{R}$, not points in $\mathbb{R}$.

(2) **Discontinuous functions**: you make a choice, say $\alpha(\text{positive } x\text{-axis } \mathbb{R}^+) = 0$. Then $\alpha : \mathbb{R}^2 \setminus 0 \to [0, 2\pi)$ will not be continuous near $\mathbb{R}^+$ since the angle there jumps by 2π.

(3) **Cut the domain**: we remove enough of $\mathbb{R}^2$ to ensure that α becomes continuous. For example, this works if we remove the positive x-axis

$$\mathbb{R}^+ = \{(x,0) \in \mathbb{R}^2 : x \geq 0\}.$$

and we define

$$\alpha_0^+ : \mathbb{R}^2 \setminus \mathbb{R}^+ \to (0, 2\pi) \subset \mathbb{R}.$$

We could also have cut the plane along $\mathbb{R}^- = \{(x,0) \in \mathbb{R}^2 : x \leq 0\}$ to obtain

$$\alpha_0^- : \mathbb{R}^2 \setminus \mathbb{R}^- \to (-\pi, \pi) \subset \mathbb{R}.$$

Notice that $\alpha_0^\pm$ are both continuous.
Notice that we can also shift those angle functions by multiples of 2π:

$$\alpha_n^+ : \mathbb{R}^2 \setminus \mathbb{R}^+ \to \mathbb{R}, \quad \alpha_n^+ = \alpha_0^+ + 2n\pi$$
$$\alpha_n^- : \mathbb{R}^2 \setminus \mathbb{R}^- \to \mathbb{R}, \quad \alpha_n^- = \alpha_0^- + 2n\pi.$$

4.2. Defining a continuous angle function α_p for a path p.

Observation 20. *For any continuous path $p : [a,b] \to \mathbb{R}^2 \setminus 0$ we can define a continuous angle function $\alpha_p : [a,b] \to \mathbb{R}$ by patching together the functions $\alpha_n^\pm$ as required by continuity. The only freedom of choice is the initial value $\alpha_p(a)$, and shifting $\alpha_p(a)$ by $2n\pi$ causes the same shift by $2n\pi$ in the final value $\alpha_p(b)$. So the difference $\alpha_p(b) - \alpha_p(a)$ does not depend on choices.*

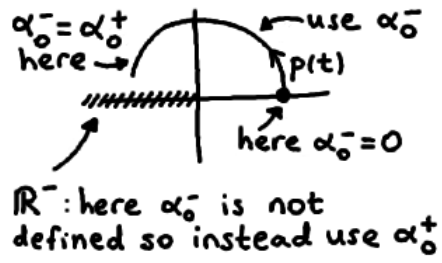
Examples.

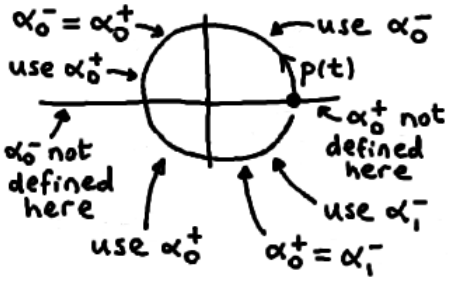
(1) *Consider the loop $p : [0,1] \to \mathbb{R}^2 \setminus 0$, $p(t) = (\cos(2\pi t), \sin(2\pi t))$. We can pick the initial angle value at $p(0) = (1,0)$ to be zero,*

$$\alpha_p(0) = 0 = \alpha_0^-(1,0) = (\alpha_0^- \circ p)(0).$$

Notice we are forced to use α_0^- since the other $\alpha_n^\pm$ are either not defined at $p(0)$ or do not equal zero at $p(0)$. So we are forced to define

$$\alpha_p(t) = (\alpha_0^- \circ p)(t) \text{ for small } t \geq 0.$$





We can use this same α_0^- as we increase t until we approach $t = 1/2$: at that point $p(1/2) = (-1,0)$ has reached the boundary $\mathbb{R}^-$ of the domain of definition of α_0^-. On $\mathbb{R}^-$, α_0^- is not defined. Also we cannot use α_0^- for

$t > 1/2$ *since* α_0^- *jumps from* π *to* $-\pi$ *when we cross* $\mathbb{R}^-$. *So we cannot use* $\alpha_0^- \circ p$ *for* $t \geq 1/2$. *However, for* $t < 1/2$ *close to* $1/2$, *we have*

$$\alpha_0^-(p(t)) = \alpha_0^+(p(t)) \quad (\textit{both values are close to } \pi),$$

so we can ''patch them together'': take $\alpha_p = \alpha_0^+ \circ p$ *for* t *close to* $1/2$, *so we stop using* α_0^- *and instead we use* α_0^+. *We can keep using* α_0^+ *until we approach* $t = 1$ *(when* p *reaches the boundary of the domain of* α_0^+*). For* $t < 1$ *close to* 1 *we ''patch'' with* α_1^- *since* $\alpha_0^+(p(t)) = \alpha_1^-(p(t))$ *(both close to* 2π*). Thus we take* $\alpha_p = \alpha_1^- \circ p$ *for* $t \leq 1$ *close to* 1. *The value of* α_p *at the endpoint is therefore* $\alpha_p(1) = \alpha_1^-(1,0) = 2\pi$.

(2) *If we had chosen instead* $\alpha_p(0) = 2n\pi$, *then we would have used* $\alpha_n^-, \alpha_n^+, \alpha_{n+1}^-$ *instead of* $\alpha_0^-, \alpha_0^+, \alpha_1^-$. *So the endpoint would also be shifted by* $2n\pi$ *since* $\alpha_p(1) = \alpha_{n+1}^-(1,0) = 2\pi + 2n\pi$. *But the difference does not change:*

$$\alpha_p(1) - \alpha_p(0) = 2\pi = \textit{Total angle } p \textit{ has rotated by.}$$

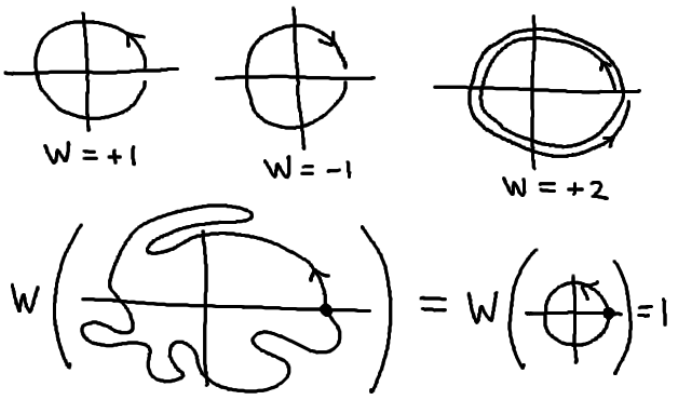
(3) *For the path that goes around zero* k *times:* $p : [0,k] \to \mathbb{R}^2 \setminus 0$, $p(t) = (\cos(2\pi t), \sin(2\pi t))$, *we would use* $\alpha_0^-, \alpha_0^+, \alpha_1^-, \alpha_1^-, \alpha_1^+, \alpha_2^-, \ldots, \alpha_{k-1}^-, \alpha_{k-1}^+, \alpha_k^-$, *so*

$$\alpha_p(k) - \alpha_p(0) = 2k\pi = \textit{Total angle } p \textit{ has rotated by.}$$

4.3. **Definition of the winding number of a loop.**

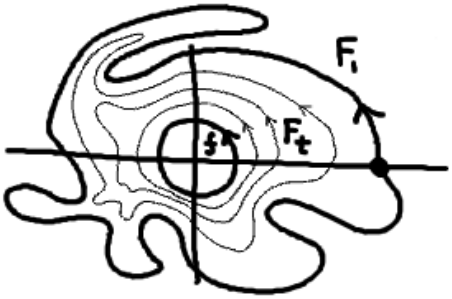
In general, for a continuous path $p : [a,b] \to \mathbb{R}^2 \setminus 0$, such that $p(a) = p(b)$ (so p is a loop), the *winding number* of p is the integer

$$W(p) = \frac{\alpha_p(b) - \alpha_p(a)}{2\pi} \in \mathbb{Z}$$



Notice in the examples that $W(p)$ is just the total number of full rotations of p around the origin (counted with signs depending on whether we go around anti-clockwise or clockwise). It also seems that $W(p)$ does not change if we "deform" p without crossing 0: but what does deforming mean exactly?

4.4. **What precisely is a continuous deformation?** Intuitively, imagine you have a small piece of string f, and you throw it to a friend in the class-room. At time $t = 0$, you are holding your string in your hand, that's position $F_0 = f$. At time $0 < t < 1$ it is flying through the air, that's position F_t. Finally at time $t = 1$ the string is lying in the hand of your friend, that's position F_1. The string may have wiggled, stretched and contracted (but not ripped) during it's journey F, which you can think of as a movie $F = (F_t)_{0 \leq t \leq 1}$ depending on the time variable t.



Let's turn this into precise mathematics. A *continuous deformation* of a loop $f : S^1 \to X$ is a continuous map

$$F : [0,1] \times S^1 \to X$$

satisfying $F(0, e^{2\pi is}) = f(e^{2\pi is})$. Geometrically, you have a family of loops

$$F_t : S^1 \to X,\ e^{2\pi is} \mapsto F_t(e^{2\pi is}) = F(t, e^{2\pi is}).$$

At time $t = 0$ you have the original loop $F_0 = f$, and at time $t = 1$ you have the new deformed loop F_1.

4.5. **The winding number does not change if you deform the loop.**

Given a loop $f : S^1 \to \mathbb{R}^2 \setminus 0$ we produce a corresponding path:

$$p_f : [0,1] \to \mathbb{R}^2 \setminus 0, \quad p_f(s) = f(e^{2\pi is})$$

and we define $W(f) = W(p_f)$. Since mathematicians often think of S^1 as $[0,1]$ with endpoints $0, 1$ identified, one sometimes blurs the distinction between f and p_f.

The last picture above, on winding numbers, suggests that $W(f)$ does not change when we continuously deform f, meaning $W(F_t)$ is constant in $t \in [0,1]$.

Theorem 21. *The winding number does not change under continuous deformations of the loop (provided we do not cross* 0*).*

Proof. Notice that $W(p) = W(\frac{p}{\|p\|})$ since the angle function $\alpha_p = \alpha_{p/\|p\|}$ does not care about the length $\|p\|$ of p (the distance of p from 0). So we can always assume that $p : [0,1] \to \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\} = S^1$ lands inside the circle.

We now prove the claim in four steps:

(1) We will show that $W(f)$ depends continuously on the continuous loop $f \in C(S^1, S^1)$ (recall $C(S^1, S^1)$ denotes the set of continuous maps $S^1 \to S^1$).
(2) A continuous deformation of f is a continuous map $F : [0,1] \times S^1 \to S^1$.
(3) Thus we can produce a continuous integer-valued function:

$$w : [0,1] \xrightarrow{F_\bullet} C(S^1, S^1) \xrightarrow{W} \mathbb{Z},\ w(t) = W(F_t).$$

This is continuous since it's a composition of continuous maps.
(4) Since $[0,1]$ is connected, $w : [0,1] \to \mathbb{Z}$ is constant, so we deduce the required result:

$$W(f) = W(F_0) = W(F_1).$$

The only difficulty is to explain the first step: in what metric space do loops live in? Define the distance function on the set $C(S^1, S^1)$ by taking the maximum of the Euclidean distances between the two given loops $f, g : S^1 \to S^1$ evaluated at each point of S^1:

$$d(f,g) = \max_{e^{2\pi is} \in S^1} d_{S^1}(f(e^{2\pi is}), g(e^{2\pi is})).$$

Exercise: check that this satisfies the requirements in the definition of a distance function.

In the construction of α_{p_f}, we can stipulate that we only use $\alpha_n^{\pm}$ when $f(e^{2\pi is})$ is at least a distance $1/10$ away from $(\pm 1, 0)$ (a boundary point of the domain of $\alpha_n^{\pm}$). So once we get within $1/10$ distance, we switch to the other relevant angle function (which has the opposite

domain boundary $\mathbb{R}^{\mp}$). So f never gets within $1/10$ of the domain boundary of the angle function we are using.

If g is a loop close to f, say $d(f,g) < 1/99$, then in fact we can use the same angle functions as those we used for f at each time s. This is possible because g cannot get closer than distance $\frac{1}{10} - \frac{1}{99}$ from the boundary point ± 1 of the domain of $\alpha_n^{\pm}$ unless f gets within $1/10$ from it (which is not allowed by construction).

Thus, at any time s, we are using the same $\alpha_n^{\pm}$:

$$\alpha_{p_f}(s) = (\alpha_n^{\pm} \circ p_f)(s) \qquad \alpha_{p_g}(s) = (\alpha_n^{\pm} \circ p_g)(s).$$

If f, g are close, say $d(f,g) < 1/99$, then p_f, p_g are close and hence, since $\alpha_n^{\pm}$ is continuous, also $\alpha_{p_f}(s)$, $\alpha_{p_g}(s)$ are close.

If we replace $1/99$ by a very small number, then we can make $|\alpha_{p_f}(s) - \alpha_{p_g}(s)|$ arbitrarily small (for each s). Thus, taking $s = 1$ and $s = 0$,

$$|\alpha_{p_f}(1) - \alpha_{p_{g_m}}(1)| \to 0 \qquad |\alpha_{p_f}(0) - \alpha_{p_{g_m}}(0)| \to 0$$

for any sequence of loops g_m approaching f. Hence:

$$W(f) - W(g_m) = \frac{1}{2\pi}\left(\alpha_{p_f}(1) - \alpha_{p_{g_m}}(1)\right) + \frac{1}{2\pi}\left(\alpha_{p_f}(0) - \alpha_{p_{g_m}}(0)\right) \to 0.$$

Since $W(g_m) \to W(f)$ whenever $g_m \to f$, we deduce by Theorem 13 that the winding number $W : C(S^1, S^1) \to \mathbb{R}$ is continuous. □

4.6. **The circle is not simply connected.**

Corollary 22. *S^1 is not simply connected.*

Proof. Suppose S^1 is simply connected, by contradiction. Then any loop f can be continuously deformed to a point, so $W(f) = W(\text{point}) = 0$. But we showed above that the loop $f : S^1 \to S^1$, $f(e^{2\pi i s}) = e^{2\pi i s}$ has $W(p) = 1 \neq 0$. Contradiction. □

4.7. **The fundamental theorem of algebra.**

Theorem 23. *Any non-constant complex polynomial*

$$F(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_2 z^2 + a_1 z + a_0 \qquad (a_0, a_1, \ldots, a_{n-1} \in \mathbb{C})$$

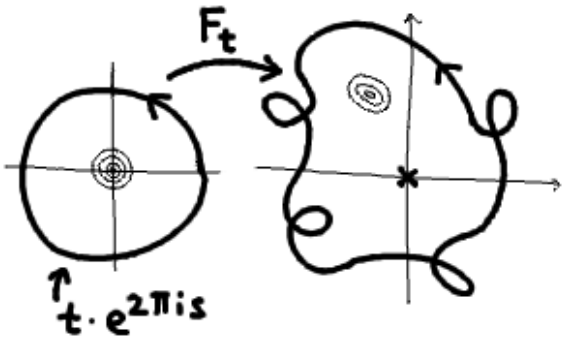
has at least one root, that is a solution $z = z_1$ of the equation $F(z) = 0$.

Remark 24. *From this we can deduce that the polynomial F factorizes completely as $F(z) = (z - z_1)(z - z_2)\cdots(z - z_n)$, so there are exactly n roots $z_1, \ldots, z_n$ (some of which may be repeated). Indeed Euclidean division by $z - z_1$ gives $F(z) = (z - z_1) \cdot G(z) + r(z)$, where $r(z)$ is the remainder polynomial. The remainder always has degree strictly less than what we divide by, which in our case is $z - z_1$. So $r(z)$ has degree 0, so it is a constant. But this constant is zero since plugging in $z = z_1$ gives: $0 = F(z_1) = (z_1 - z_1)G(z_1) + r(z_1) = r(z_1)$. Thus $F(z) = (z - z_1)G(z)$. Since G has lower degree than F, we can factorize it completely by induction on the degree of the polynomial.*

Proof of the Theorem. Suppose by contradiction that F is never zero. Then from F we can produce loops:

$$F_t : S^1 \to \mathbb{C} \setminus 0,\ F_t(e^{2\pi i s}) = F(t \cdot e^{2\pi i s}).$$

where the positive real number $t > 0$ is the radius of the circle $t \cdot e^{2\pi i s}$. Notice that F_t lands in $\mathbb{R}^2 \setminus 0$ because we assumed that F is never zero.



The picture above shows what F_t may look like for general t. The small circles show what F_t looks like for very small t. For $t = 0$, F_0 is just a constant: the point $F(0)$.

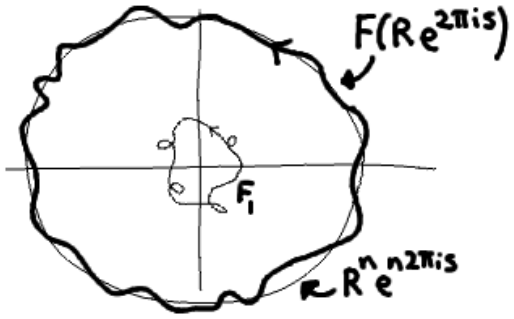
Since the winding number does not change under continuous deformations, $W(F_t)$ is constant in t. Thus

$$W(F_t) = W(F_0) = W(\text{point}) = 0.$$

But now consider a large radius $t = R$. Then

$$\begin{aligned} F(Re^{2\pi is}) &= R^n e^{n2\pi is} + a_{n-1}R^{n-1}e^{(n-1)2\pi is} + \cdots + a_1 Re^{2\pi is} + a_0 \\ &= R^n e^{n2\pi is}\left(1 + \tfrac{1}{R}a_{n-1}e^{(n-1)2\pi is} + \cdots + \tfrac{1}{R^{n-1}}a_1 e^{2\pi is} + \tfrac{1}{R^n}a_0\right). \end{aligned}$$

Notice the a_j are constants, and the $e^{i\cdot\text{real}}$ have length 1, so for large R the round bracket $1 + \cdots$ is approximately equal to 1. Indeed, for large R we can assume the bracket is within distance $1/100$ from 1. Therefore F_R is a small deformation of the loop $S^1 \to \mathbb{C} \setminus 0$, $e^{2\pi is} \mapsto R^n e^{n2\pi is}$, which (dividing by the length R^n) has the same winding number as our favourite loop $S^1 \mapsto S^1$, $z \mapsto z^n$, which goes around zero n times (so $W = n$).



Thus they have the same winding number. So

$$W(F_t) = W(F_R) = W(S^1 \to S^1, z \mapsto z^n) = n.$$

But above we got $W(F_t) = 0$, so $n = 0$, so F is a constant polynomial. Contradiction. □

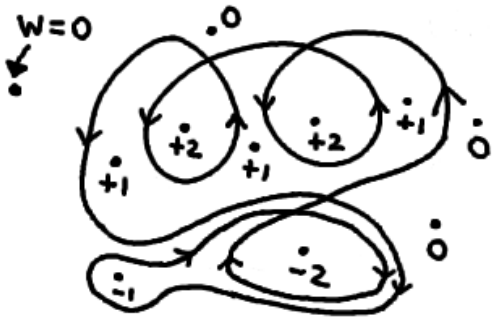
4.8. **Winding number around a point.**

There is nothing special about the point 0 when we worked with loops $f : S^1 \to \mathbb{C} \setminus 0$. We could pick any point $z \in \mathbb{C}$ not on the loop, so $f : S^1 \to \mathbb{C} \setminus z$, then since we can identify

$$\mathbb{C} \setminus z \to \mathbb{C} \setminus 0,\ w \mapsto w - z,$$

we know how to calculate the winding number of f around z, denoted by $W(f; z)$. Explicitly:

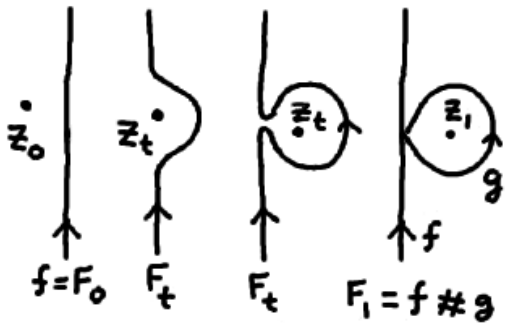
$$\boxed{W(f; z) = W(f - z; 0)}$$



By deformation invariance, the integer $W(f;z)$ does not change if we continuously deform f without crossing z. Similarly, $W(f;z)$ does not change if we continuously move z without crossing f, you will prove this in the Exercises.[28]

4.9. **How does $W(f;z)$ change when z crosses the loop?**

In the picture above, notice that $W(f;z)$ can only jump by $+1$ or -1 when z crosses the loop once.[29] If we can prove this (and find a recipe to decide the sign $\pm$) then this gives an easy way to calculate winding numbers: just start from a point z near infinity (there $W(f;z)=0$) then move z towards any other point and add the contributions ± 1 as you cross the loop.



In the picture, we find out how W changes as z crosses the path f. Call these starting positions z_0, F_0, and suppose the final position of z after crossing f is z_1. Let F_t be a deformation from $F_0 = f$ to $F_1 = f\#g$ (a copy of f attached to an extra loop g going around z_1 once). Let z_t be the motion of the point from z_0 to z_1, ensuring that z_t never lies on the loop F_t. Thus, by invariance[30] of W we deduce that $W(f;z_0)$ equals:

$$W(F_0;z_0) = W(F_t;z_t) = W(F_1;z_1) = W(f\#g;z_1) = W(f;z_1) + W(g;z_1) = W(f;z_1) + 1.$$

where we used the obvious additivity of winding numbers (see also the Exercises on concatenations of loops), and we used that $W(g;z_1) = +1$ since the loop g goes once around z_1 anti-clockwise. Thus

$$W(f;z_1) - W(f;z_0) = -1,$$

so the winding number changes by -1 if "we" (the point z) cross a loop that passes in front of us in the anticlockwise sense. Similarly, if the arrow on f had been pointing in

[28] *Hint. You can view this as being the same as fixing z but deforming f: $W(f;z_t) = W(f - z_t; 0)$.*

[29] In the picture, when z crosses a vertex, that is a self-intersection point of the loop, then $W(f;z)$ jumps by ± 2. But that is because z is actually crossing the loop twice. The two small arcs of the image loop in $\mathbb{C} \setminus 0$ which intersect actually correspond to two completely different arcs on the domain of the map f. Indeed, there are two different points $e^{2\pi i s_1}, e^{2\pi i s_2}$ giving rise to the self-intersection $f(e^{2\pi i s_1}) = f(e^{2\pi i s_2})$.

[30] If you are uncomfortable with moving both simultaneously, you could do the deformations in stages, at each stage either moving F_t or moving z_t. However, moving both simultaneously is legitimate since $W(F_t;z_t) = W(F_t - z_t; 0)$ and $F_t - z_t$ is never zero.

the opposite direction then also the arrow on the loop g would be opposite, so the winding number changes by $+1$ if we cross a loop that passes in front of us in the clockwise sense.

Exercise 25. *Given a point z not lying on the loop $f : S^1 \to \mathbb{C}$, show that you can quickly determine $W(f; z)$ by drawing a ray*[31] *from z towards infinity, and appropriately counting by ± 1 the intersections of the ray with the loop.*

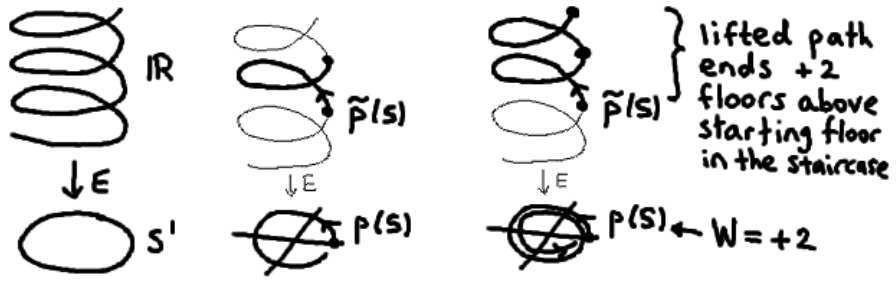
4.10. **The exponential map, and lifts of loops.**

Our goal, in the next Section, will be to show that a loop in $\mathbb{R}^2 \setminus 0$ can be shrunk to a point if and only if the winding number is zero. In one direction, the proof is easy: if a loop $p : S^1 \to \mathbb{R}^2 \setminus 0$ can be continuously shrunk to a point then $W(p) = W(\text{point}) = 0$. The proof of the converse is tricky. We will prove it by our usual In-and-Out trick, where the "Out" map will be the following exponential map.

Define the "exponential map"

$$E : \mathbb{R} \to S^1,\ E(s) = e^{2\pi i s} = \text{point on the circle forming the angle } 2\pi s \text{ with } \mathbb{R}^+.$$

Think of $\mathbb{R}$ as a spiralling staircase, and E as the vertical projection to the ground floor S^1.



Then, for any path $p : [a, b] \to S^1$, $\alpha_p : [a, b] \to \mathbb{R}$ satisfies

$$E \circ \tfrac{\alpha_p}{2\pi} = p : [a, b] \to S^1$$

by construction.[32] For this reason, $\widetilde{p} = \frac{\alpha_p}{2\pi}$ is often called a *lift* of p to $\mathbb{R}$.

When $p : [a, b] \to S^1$ is a loop ($p(a) = p(b)$), the winding number is

$$W(p) = \widetilde{p}(b) - \widetilde{p}(a).$$

If you think of $\mathbb{R}$ as a staircase, then the winding number tells you how many floors you have gone up along the lifted path.

4.11. **A loop can be shrunk to a point if and only if it has winding number zero.** We first consider loops $S^1 \to S^1$:

Theorem 26. *If a loop $p : S^1 \to S^1$ has zero winding number, then it can be shrunk down to a point (there is a continuous deformation $P : [0, 1] \times S^1 \to S^1$).*

Proof. $W(p) = \widetilde{p}(b) - \widetilde{p}(a)$ for any choice of lift $\widetilde{p} = \frac{\alpha_p}{2\pi}$. Thus $W(p) = 0$ if and only if the lift $\widetilde{p} : [a, b] \to \mathbb{R}$ is also a loop: $\widetilde{p}(b) = \widetilde{p}(a)$. But $\mathbb{R}$ is simply connected, so there is a filling $\widetilde{P} : \mathbb{D} \to \mathbb{R}$ of $\widetilde{p}$, thus $E \circ \widetilde{P} : \mathbb{D} \to S^1$ is a filling of $E \circ \widetilde{p} = p$ as required. □

Corollary 27. *A loop $f : S^1 \to \mathbb{R}^2 \setminus 0$ can be shrunk down to a point if and only if $W(f) = 0$.*

Proof. You will show in the Exercises that any loop $S^1 \to \mathbb{R}^2 \setminus 0$ can be continuously deformed to a loop $S^1 \to S^1$. The deformed loop in S^1 still has $W = 0$, so by the Theorem above it can be shrunk down to a point (in fact within S^1). □

[31] We did not take this approach in the notes, because this does not always work. Even in the simplest case, when f is a polygonal path, there are problems if the ray intersects the path at a vertex.

[32] Proof that $E \circ \frac{\alpha_p}{2\pi} = p$: Given $p(s) = e^{2\pi i s}$, the angle function $\alpha_p(s) = 2\pi s + n2\pi$ for some integer n, hence $E \circ \frac{\alpha_p(s)}{2\pi} = e^{2\pi i s + n2\pi i} = e^{2\pi i s} = p(s)$.

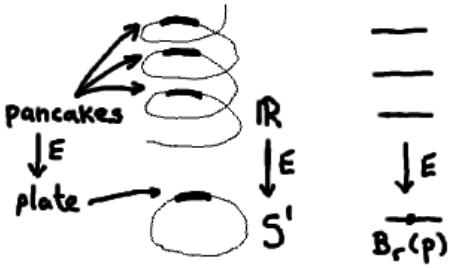
4.12. **The fundamental group and universal covers.**

In the Exercises, you will show that the space of loops $S^1 \to S^1$ modulo continuous deformations forms a group. The group operation is just concatenation of loops ("first go around one loop, then go around the other"). This group is called the *fundamental group* $\pi_1(S^1)$ and in the Exercises you will show that $\pi_1(S^1)$ can be identified with the group $\mathbb{Z}$ (with addition). The identification is given by the winding number:

$$W : \pi_1(S^1) \to \mathbb{Z}.$$

More generally, you can define the fundamental group $\pi_1(X)$ for any space X.

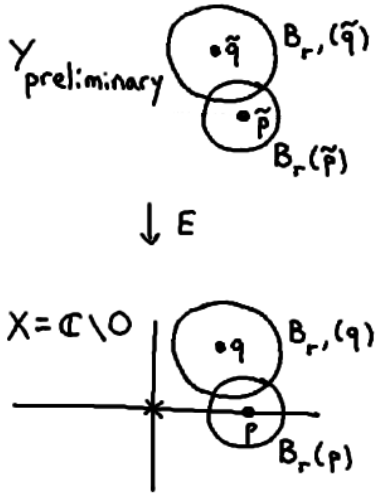
The map $E : \mathbb{R} \to S^1$ is called the *universal cover* of S^1. In general, a universal cover of X is a simply connected space Y together with a continuous map $E : Y \to X$ such that for any point $p \in X$ there is a sufficiently small ball $B_r(p)$ such that the preimage $E^{-1}(B_r(p))$ is a disjoint union of balls in Y, and each such ball maps homeomorphically to $B_r(p)$ via E. Intuitively, you should think of this union of balls as a "stack of pancakes" over the plate $B_r(p)$:



Example: how to build the universal cover of the plane with a hole

Suppose you have a space X and you want to start building a universal cover Y. In the pictures, we will think of the simple case of $X = \mathbb{C} \setminus 0$, the plane with a hole where 0 used to be.

The first key property, is that Y should locally look like X. So let's start building Y piece by piece as follows. Start with a point $p \in X$, near p. We need to have a copy of p in Y. So start with $Y_{\text{preliminary}} = \widetilde{p}$ (just a point, and we put a twiddle so we don't confuse Y with X). Now the spaces X, Y should be the same near p, so we might as well *define* a small ball $B_r(\widetilde{p})$ around $\widetilde{p}$ in Y to be the same (with the same distance function) as a small ball $B_r(p)$ around p in X. So now, $Y_{\text{preliminary}} = B_r(\widetilde{p})$ has become larger.



Now start walking away from p in X towards the boundary $\partial B_r(p)$ of the ball: there, you find another ball $B_{r'}(q)$ overlapping with $B_r(p)$. Therefore, upstairs in $Y_{preliminary}$, we also want a copy $B_{r'}(\widetilde{q})$ of that ball with centre $\widetilde{q}$ lying over q via E, and overlapping with $B_r(\widetilde{p})$

just like $B_r(p) \cap B_{r'}(q)$ overlap in X. Continue inductively: pick a small ball $B_{r'}(q) \subset X$, make $Y_{\text{preliminary}}$ bigger by gluing a copy of this ball onto what we built so far. So

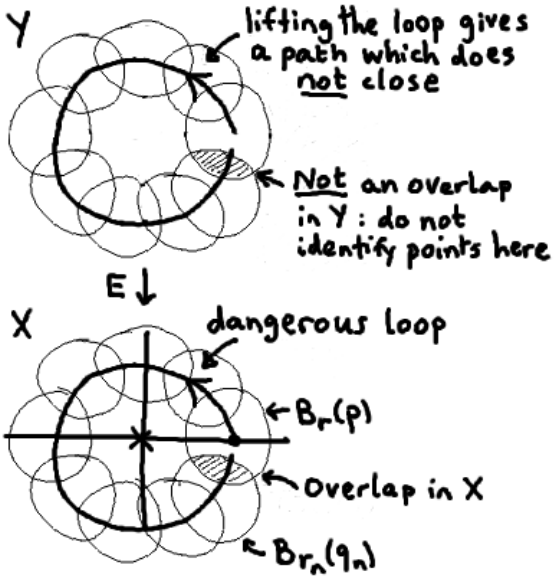
$$Y_{\text{preliminary}} = B_r(\widetilde{p}) \cup B_{r'}(\widetilde{q})$$

(with the same overlap as downstairs, in X, between the balls $B_r(p), B_{r'}(q)$).

Keep building. So aren't we just building a copy of X? Yes, because we have not yet imposed the second key property of a universal cover Y. We want Y to be simply connected (loops are contractible). As long as the $Y_{\text{preliminary}}$ we built so far is simply connected, the $Y_{\text{preliminary}}$ is an exact copy (homeomorphic) to a subspace of X. So far, we can therefore define our covering map $E : Y_{\text{preliminary}} \to X$ as being the "identity" which identifies these two exact copies.

Example. If the space X is already simply-connected (such as $\mathbb{D}, \mathbb{R}, S^2$, and also $\mathbb{D}^n, \mathbb{R}^n, S^n$ for $n \geq 2$), then you really are just building X again so $Y = X$. In other words, simply connected spaces are universal covers of themselves via the identity map, so it's not so interesting (only one pancake per plate!).

But now, consider our example of $X = \mathbb{C} \setminus 0$ (which we know is not simply connected). If we pick discs $B_r(p), B_{r_2}(q_2), \ldots, B_{r_n}(q_n)$ going around a circle about 0, then there is the dangerous loop in $\mathbb{C} \setminus 0$ which we know cannot be shrunk to a point.



We do not want that loop to exist in Y since we want Y to be simply connected. So in Y, you declare that those discs in the picture, $B_r(p)$ and $B_{r_n}(q_n)$ which overlap in X give rise to discs $B_r(\widetilde{p})$ and $B_{r_n}(\widetilde{p_n})$ in $Y_{\text{preliminary}}$ which do *not* overlap.

It is easiest to visualize what $Y_{\text{preliminary}}$ looks like if you think of it as a spiralling staircase (with stairs of infinite width). So think: X = ground floor, Y = Wadham staircase, and the map $E : Y \to X$ is the vertical projection.

More abstractly, you don't actually need to think of your staircase Y as existing inside a bigger universe (Wadham College). To define a space you just need to specify a set and you

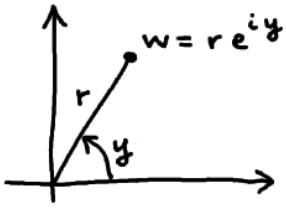
need to specify a distance function. You do not need to think of the points of Y as being part of a larger space that you are more familiar with.

In the Exercises you will construct the universal cover of a torus $T^2 = S^1 \times S^1$ and of a figure "8" loop.

Exercise 28. *In the above construction, how do we define distances in Y? Locally we know that Y is the same as X, but how do we define distances between "far-away" points?*

4.13. **Riemann surfaces: how these ideas arise in Analysis.**

Recall that the complex number $w = re^{iy}$ lies at distance r from 0 and forms an angle y with the real-axis:



For $z = x+iy$, $e^z = e^x e^{iy} = e^x(\cos y + i \sin y)$. By definition, we want the complex logarithm to be an "inverse" to the complex exponential function, so we want

$$\mathrm{Log}(e^z) = z.$$

We will now use this condition, to guess how $\mathrm{Log}(z)$ needs to be defined in general.

Since $|e^z| = e^x$ and the real logarithm[33] satisfies $\log e^x = x$, and since $\alpha(e^z) = y$ is the angle between e^z and the $\mathbb{R}^+$ axis, we deduce that we want:

$$\mathrm{Log}(e^z) = x + iy = \log|e^z| + i\alpha(e^z).$$

Therefore, we define:

$$\mathrm{Log}\, z = \log|z| + i\alpha(z),$$

where $\alpha(z)$ is only defined up to adding integer multiples of 2π since $e^{z+2\pi i} = e^z$.

Riemann's idea, is that one should not think of $\mathrm{Log}\, z$ as being defined on $\mathbb{C} \setminus 0$ as a multi-valued "function". Instead, one should build a new surface, called *Riemann surface*, and use that as domain of definition for a (single-valued) function. For $\mathrm{Log}\, z$ this is the infinite spiral staircase above:

$$\mathrm{Log} : (\text{infinite spiral staircase}) \to \mathbb{C}.$$

Exercise 29. *Define* Log *on the spiral staircase using the angle functions $\alpha_n^{\pm}$.*

Thus, the angle functions $\alpha_n^{\pm}$ that we defined before, are also part of the construction of a well-defined complex logarithm. Physicists usually just think of Log as a "multi-valued function" (since it is only defined up to adding integer multiples of $2\pi i$); whereas a mathematician instead says: "if you're not happy with the universe, because Log can't be defined on $\mathbb{C}$, then just change your universe!".

The functions considered by Riemann were *analytic functions*: that is, complex-valued functions which locally around a point $p \in \mathbb{C}$ can be expressed as a power series in $z - p$, that is a convergent infinite sum involving powers of $z - p$ (think of it as a "polynomial" of infinite degree in $z - p$).

Examples.

(1) *Complex polynomials are analytic functions.*

(2) *The exponential is analytic. For example* $e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots$ *is the power series around* $p = 0$*, and that series is correct for any* $z \in \mathbb{C}$*.*

[33] our logarithms are always in base e, so natural logarithms: $\log(e) = 1$.

(3) *The complex logarithm is analytic. For example*

$$\mathrm{Log}(z) = \mathrm{Log}(1 + (z-1)) = (z-1) - \frac{(z-1)^2}{2} + \frac{(z-1)^3}{3} - \frac{(z-1)^4}{4} + \cdots$$

is the power series around $p = 1$*, but the series is only correct for* $z \in B_1(1)$ *(that is:* $|z-1| < 1$*). On each ball* $B_r(p)$ *which does not contain* 0 *(so* $r < |p|$*) one can find a series for* $\mathrm{Log}(z)$ *in powers of* $z - p$*.*

When you try to find the largest possible domain of definition of an analytic function, you therefore patch together balls (thus gluing together different locally defined series expansions of the same function). This process of extending the domain of definition by gluing together series is called *analytic continuation*. The famous problem called the *Riemann hypothesis*, for example, is about finding the zeros of the analytic continuation of a certain function, called the *Riemann zeta function*.

4.14. **The Jordan curve theorem.**

By *closed curve* in a space X we will mean a continuous map $f : S^1 \to X$. We call it a *simple closed curve* if it does not cross itself, meaning: f is injective.

Theorem 30. *A simple closed curve* $f : S^1 \to \mathbb{R}^2$ *in the plane divides the plane into two connected components.*[34] *More precisely,* $\mathbb{R}^2 \setminus f(S^1) = A \cup B$ *where* A, B *are disjoint connected components,* B *contains the points at infinity and is called the "outside" of the curve, and* A *is called the "inside" of the curve.*

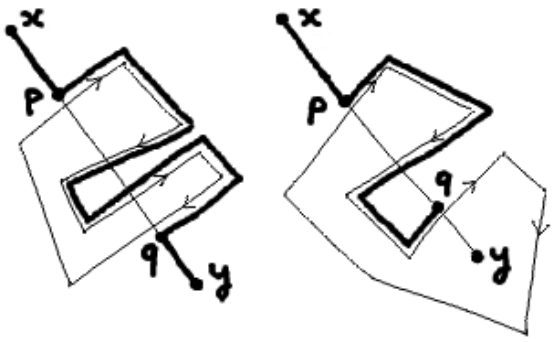
Proof for polygonal paths. Define

$$\begin{aligned} A &= \{z \in \mathbb{R}^2 \setminus f(S^1) : W(f; z) \text{ is odd}\} \\ B &= \{z \in \mathbb{R}^2 \setminus f(S^1) : W(f; z) \text{ is even}\}. \end{aligned}$$

Note that A, B are disjoint and $\mathbb{R}^2 \setminus f(S^1) = A \cup B$. So what is left to prove? We need to show that A, B are connected. For general continuous f, this requires more machinery than we have learnt so far. But we can prove this for polygonal paths f (that is, when f traces out a polygon). In this case, we will show that A, B are path-connected (hence connected).

Note that no path can connect a point $x \in A$ with a point $y \in B$ without crossing f because otherwise $W(f; x) = W(f; y)$ but the parity of these integers is different.

Suppose $x, y \in B$ (the following argument will work also for A, since we will only use that $W(f; x), W(f; y)$ have the same parity). Draw the straight line segment joining x to y. If this segment does not intersect f, then we are done: we have path-connected x to y. So suppose now that the segment does intersect f, and call p the first point of intersection and q the last point of intersection.



[34] A *connected component* $S \subset X$ means a maximal connected subset. Maximal means that there is no larger connected subset $S' \subset X$ containing S.

We will now explicitly build a path connecting x to y and avoiding f.

Imagine the game Tron (like in the movie) where you are driving a blue car and your enemy f is driving a red car. You start from x and you race along the segment in the direction of y. Just before crashing into your enemy's path at p you make a sharp turn in the same direction as your enemy. You follow closely your enemy's path (staying just a little distance away from it), until your enemy passes through q. At that stage you stop.

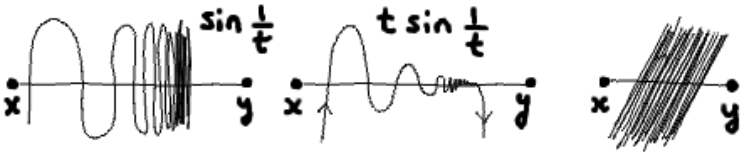
In the left picture, your blue car has reached the segment on the side of y, so you can now race along the segment towards y and we are done. In the picture on the right, we ended up at the wrong side of the segment: we cannot race towards y without crashing into our enemy's path f. But this situation cannot happen: since we would only intersect the enemy's path once, we would have built a path from x to y intersecting f only once, so the winding numbers $W(f;x), W(f;y)$ would have different parities. But we assumed that they had the same parity. □

Exercise 31. *Would the proof have been easier or harder, if instead of following the red car until q, you instead alternate between following the red car and following the line segment? (that is: you follow the red car between p and the next point where the red car intersects the line segment, then repeat: follow the segment until you almost crash into the red car, then follow the red car until you again come back to the segment, etc.)*

Exercise 32 (Jordan-Brower separation theorem)**.** *For any injective continuous map $f : S^n \to \mathbb{R}^{n+1}$, the complement $\mathbb{R}^{n+1} \setminus f(S^n)$ of the image consists of two connected components (the "inside" and the "outside"). Using the Tron game, prove this theorem for $n = 2$, when f is "polygonal": so $f(S^2)$ is a polyhedron in $\mathbb{R}^3$.*

4.15. **Non-polygonal paths, and Lie groups.**

What can go wrong in the above proof, when f is not a polygonal path? Here are some pictures of what we may be worried about:



We may be worried that f wiggles like $\sin(1/t)$ for t close to zero. But in that case, ask yourself: what happens eventually with f? Does this wiggly path get arbitrarily close to y? If yes, then by continuity it would intersect y, but we assumed that y does lie on the path f. If it does not get arbitrarily close to y, then there is a problem in making f continuous (since $\sin(1/t)$ is not continuous at $t = 0$).

So suppose instead it wiggles like $t \cdot \sin(1/t)$, which is continuous even near $t = 0$. This is not an issue in the proof: if our enemy's red car can wiggle like that, then our blue car is also allowed to wiggle like that! But we need some care not to crash into the red car at the end of the wiggles at $t = 0$, which can be achieved by a last minute swerve towards y.

So, in many situations, the above proof still works with minor modifications. What we are scared of, is whether there is enough space around the red car's path to say "follow closely your enemy's path". This was clear in the case of a polygonal path, but what if the enemy's path f involved a "dense" bunch of lines, like in the third picture above?

Continuous maps in general can cause such phenomena. The simplest example of this is a line of irrational slope inside the torus. More precisely, think of the torus $T^2 = S^1 \times S^1$ as arising from a square, with opposite parallel sides identified. Consider the usual lattice $\mathbb{Z}^2$ inside $\mathbb{R}^2$, this gives the vertices of our favourite tiling of the plane by squares. You can also build the torus by identifying all these squares via all translations of the form

$$x \mapsto x + n,\ y \mapsto y + m \quad (\text{for integers } n, m \in \mathbb{Z}).$$

This determines[35] a natural map $E : \mathbb{R}^2 \to T^2$. So any straight line in $\mathbb{R}^2$ gives, via E, a curve in T^2.

When does the curve close? Precisely when the slope of that line is rational. It turns out (try to prove it) that for an irrational slope you get a (non-closed) curve $\mathbb{R} \to T^2$ in the torus which inside the square (before identifying sides) looks like infinitely many parallel segments that get arbitrarily close to each other everywhere.

These ideas arise in the theory of *Lie groups*.[36] A Lie group is a group where there is a notion of "closeness" of the elements (a topology) and the group operations of multiplication and inversion are continuous. You also want the space to locally look like a disc $\mathbb{D}^n$.

Examples.

(1) $\mathbb{R}$ *is a Lie group with addition as group operation, and switching sign is inversion. Also* $\mathbb{R}$ *locally looks like an interval* $\mathbb{D}^1$.
(2) *The circle* S^1 *is a Lie group. You multiply by* $e^{ia}e^{ib} = e^{i(a+b)}$ *and invert by* $(e^{ia})^{-1} = e^{-ia}$. *Also* S^1 *locally looks like* $\mathbb{D}^1$ *(an interval).*
(3) *The torus* $T = S^1 \times S^1$ *is a Lie group. You multiply in each* S^1 *factor, and locally* T^2 *looks like a disc* $\mathbb{D}^2$.
(4) *Lines of rational slope in* $\mathbb{R}^2$ *modulo* $\mathbb{Z}^2$ *are Lie subgroups* $S^1 \to T^2$ *(circle subgroups) whereas lines of irrational slope are a Lie subgroup* $\mathbb{R} \to T^2$ *which is ''dense'' in* T^2 *(it gets arbitrarily close to any point of* T^2*). Exercise: Prove this using the grasshopper trick from the Exercises. Hint: first consider the points* $e^{n \cdot ia} \in S^1$ *for* $n \in \mathbb{N}$*, for* a *rational or irrational.*

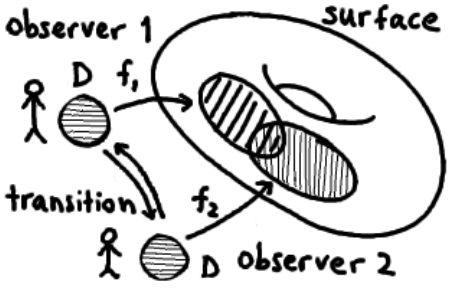
5. Classification of surfaces

5.1. **What is a surface?** We will conclude these notes with a taster of the theory of surfaces, which is the beginning of *Differential Geometry* and *Differential Topology*.

A *surface* S is a (metric) space which locally looks like a disc $D = \{z \in \mathbb{C} : |z| < 1\}$. This means that around each point $p \in S$, you can find a continuous map $f : D \to S$ which is a homeomorphism onto a neighbourhood of p.

In physics, you should think of f as a "frame of reference":[37] you have chosen certain coordinates (x, y) near p. Namely, $f(z) = f(x + iy) \in S$ has local coordinates (x, y) in the frame of reference f, where $z = x + iy \in D$.

In physics it's important that two observers, staring at the same phenomenon, both agree when a phenomenon is varying continuously or not.



[35] Indeed, this is just the exponential map in each entry: $E(x, y) = (e^{2\pi ix}, e^{2\pi iy}) \in S^1 \times S^1$. In the Exercises, you show that this is the universal cover of the torus.

[36] Pronounced "lee" (or "lii" in most other European languages).

[37] Reference frames are called *local parametrizations*, and their inverses $f^{-1} : f(D) \to D$ are called *charts*.

The first observer, $f_1 : D \to S$, and the second observer, $f_2 : D \to S$, have each chosen local coordinates. If a particle was moving in S in the shaded overlap of the two discs, then it will have coordinates $(x(t), y(t))$ for the first observer, and coordinates $(\tilde{x}(t), \tilde{y}(t))$ for the second observer, say. Are we sure that x, y are continuous if and only if $\tilde{x}, \tilde{y}$ are continuous?

Let's check. We have $f_1(x(t), y(t)) = f_2(\tilde{x}(t), \tilde{y}(t))$, therefore

$$(\tilde{x}(t), \tilde{y}(t)) = (f_2^{-1} \circ f_1)(x(t), y(t)).$$

Since f_1, f_2 are homeomorphisms onto their image, the map $f_2^{-1} \circ f_1$ is continuous (where defined). So, indeed,[38] $\tilde{x}, \tilde{y}$ are continuous in t if x, y are continuous in t.

Examples.

(1) $\mathbb{R}^2$ *is a surface;*
(2) *The sphere* S^2 *and the torus* T^2 *are surfaces. We saw that they are not homeomorphic to each other;*
(3) *The open disc* D *is a surface, and so is any open convex subset of* $\mathbb{R}^2$*. These are homeomorphic to each other.*
(4) *The disjoint union* $S_1 \sqcup \cdots \sqcup S_n$ *of several surfaces* $S_1, \ldots, S_n$ *is again a surface (but it is not connected).*
(5) *Given two surfaces* S_1, S_2*, remove a disc from each of the surfaces* S_1, S_2*, and glue the two surfaces together along the circular boundaries of the removed discs. This is called the* connect sum $S_1 \# S_2$*.*

To keep things simple, we will from now on only consider connected surfaces and we will assume that the surface is *bounded*. This means that the distance function is bounded.[39] However, bad things can still happen:

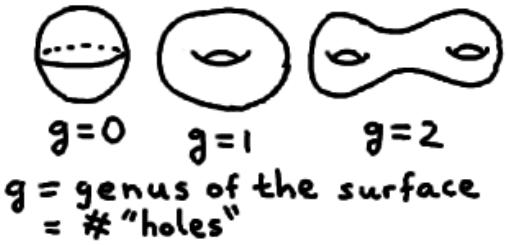
Example. Start with a doughnut. Connect sum with another doughnut, whose distance function we rescale by $1/2$*. Apply connect sum with another doughnut with distance function rescaled by* $1/2^2$*. Keep repeating using rescaling factors* $1/2^n$*. Eventually you get a bounded doughy surface with infinitely many holes!*

To avoid this, we will assume the surface is *compact*, meaning: if $f_i : D \to S$ is an infinite collection of reference frames covering $S = \cup f_i(D)$ then you can pick a finite subcollection $f_{i_1}, \ldots, f_{i_n}$ which also covers S, meaning: $S = f_{i_1}(D) \cup \ldots \cup f_{i_n}(D)$.

Exercise 33. *Show that a compact surface is always bounded.*

5.2. **Surfaces of genus g.**

Compact connected surfaces come in two big families. The first family consists of the sphere, the torus, and in general the surface of a doughnut with g holes:



Here g is called the *genus* of the surface. For this family, once you computed g, it determines the surface up to homeomorphism.

There are many ways of computing g. One way, is to pick a triangulation of the surface (or, more generally, pick a homeomorphism between the surface and a polyhedron), then

[38] If they had not been, we could have played the mathematician's favourite card: defined it to be so. In other words, we would have said that we only allow those frames of reference for which the transitions $f_2^{-1} \circ f_1$ are continuous. For example, when you define *smooth manifolds* you need to play this trick.

[39] That is, there is a large number R such that $d(p, q) \leq R$ for all points $p, q \in S$.

calculate the Euler characteristic $\chi(S) = V - E + F$ where V, E, F is the number of vertices, edges, faces. Then in general:

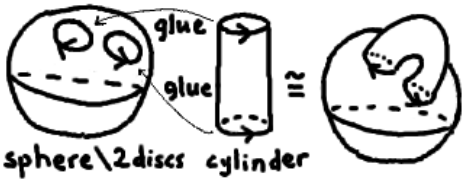
Theorem 34. $\chi(S) = 2 - 2g$.

Observe that the above surfaces can all be obtained from the sphere S^2 by *attaching handles*. To "attach a handle" means

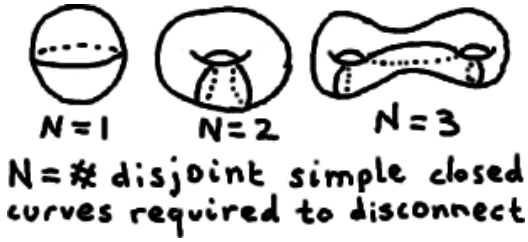
(1) remove two disjoint discs from S^2;
(2) glue the cylinder $[0,1] \times S^1$ onto the resulting surface by gluing the two boundary circles of the cylinder onto the circular boundaries of the two removed discs of S^2.

The genus g surface is obtained by attaching g handles to S^2.

In the attachment procedure, we want the cylinder to stick out on the outside of the surface. To ensure this, mark the boundaries of the removed discs with arrows pointing in opposite directions (indicated by arrows in the picture below). The cylinder also has arrows on the two bounding circles (the anticlockwise direction of $0 \times S^1$ and $1 \times S^1$). When gluing the cylinder to the surface, these arrow directions must match along the gluing circles.



Another way to find g is related to the Jordan curve theorem. You ask the question: "how many disjoint simple closed curves are required to disconnect the surface?". Call N this number.



Example. Consider the sphere. It turns out that a simple closed curve inside S^2 cannot pass through each point of S^2. So pick a point p not belonging to the loop, identify $S^2 \backslash p \cong \mathbb{R}^2$, and apply the Jordan curve theorem. We deduce that every simple closed curve in S^2 separates S^2 into two connected components. Thus $N = 1$ for the sphere.

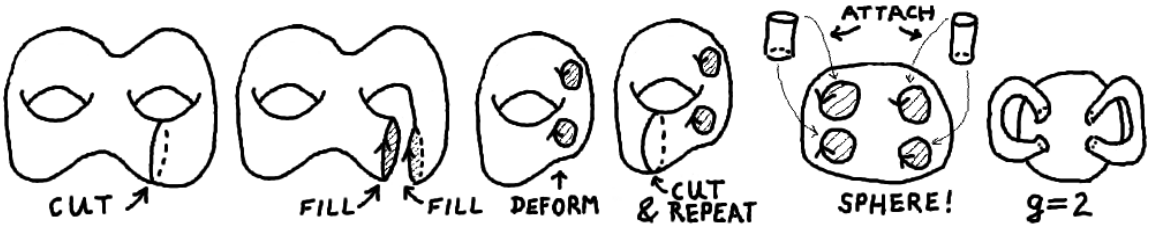
In general, for a genus g surface, it turns out that:

Theorem 35. $N = g + 1$.

Corollary 36. *A compact connected surface is homeomorphic to the sphere if and only if any simple closed curve divides the surface into two connected components.*

The above theorem/corollary (which we have not proved) allows you to detect whether a surface is a sphere or not. The proof of the classification of surfaces is to start with a surface S and to make a cut along a simple closed curve which does not disconnect the surface. The cut then gets filled up with a pair of discs. You keep repeating this cutting/filling procedure until it becomes impossible: any simple closed curve disconnects the surface. But by the

Corollary that means you've now got a sphere! To reconstruct the original surface S you now simply attach cylinders joining the pairs of filling discs (which you remove again!). This cylinder plays the role of identifying the bounding circles of the filling discs which formed the cut.

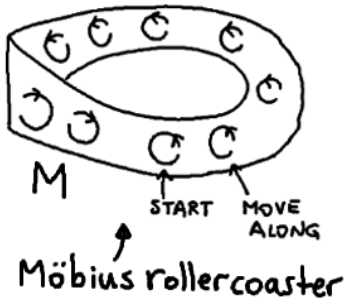


Knowing how to detect spheres is also important in higher dimensions, for example in the classification of 3-dimensional *manifolds* (the 3-dimensional analogue of surfaces, except now you use frames of reference $f : \mathbb{D}^3 \to X$). The *Poincaré Conjecture* states that every compact simply-connected 3-dimensional manifold is homeomorphic to the sphere S^3. This was only very recently proved by Grigori Perelman, in 2003, using foundational work by Richard Hamilton. For this, Perelman was awarded the fields medal in 2006, which he declined, and he was awarded a million dollar Millenium Prize by the Clay Mathematics Institute in 2010, which again he declined.

5.3. **Non-orientable surfaces.**

The above family of surfaces are called *orientable surfaces*. What does orientable mean? Consider a small disc $\mathbb{D} \subset S$ inside[40] a surface S. You can *orient* the boundary of the disc by drawing[41] an arrow along the circle bounding $\mathbb{D}$. Now suppose that you continuously move your oriented disc $\mathbb{D}$ around the surface S until you eventually come back[42] to $\mathbb{D}$. If, no matter how you move $\mathbb{D}$, the orientation of the boundary of $\mathbb{D}$ has not switched, then S is called *orientable*. Otherwise, S is called *non-orientable*.

An example of a space[43] where the orientation can flip, is the *Möbius strip* M:



When you move a small disc $\mathbb{D}$ along the "equator" of the Möbius strip M until you eventually get back to the starting disc, the boundary arrow will be reversed.

This proves that an orientable surface cannot contain a copy of the Möbius strip M, since following the "Möbius rollercoaster" you can switch the orientation of a small disc.

Recall that the Möbius strip M is obtained by taking a strip, let's say the rectangle $[0, 100] \times [0, 2]$, and then gluing the short ends together in the "wrong way": so $(0, t)$ gets

[40]that is, an injective continuous map $f : \mathbb{D} \to S$, and identify $\mathbb{D} = f(\mathbb{D})$. You can find such discs by considering a local frame of reference.

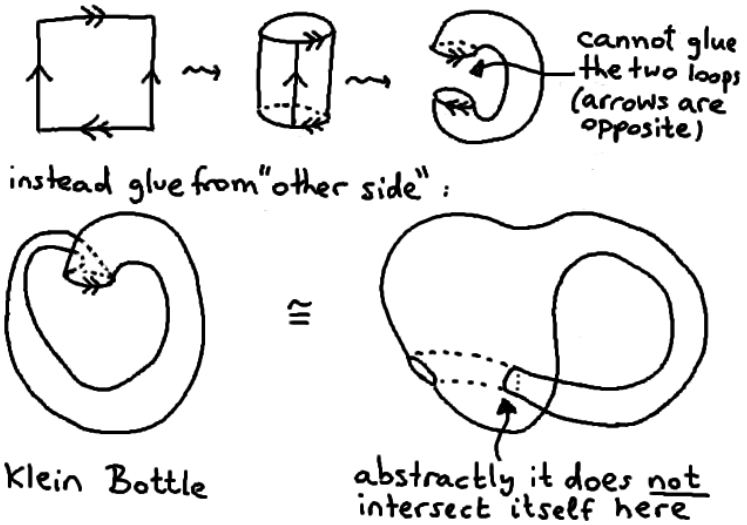
[41]More precisely, an injective map $f : \mathbb{D} \to S$ already specifies an arrow, by following (via f) the anti-clockwise direction of $S^1 = \partial\mathbb{D}$.

[42]that is, you consider a continuous family $F : [0, 1] \times \mathbb{D} \to S$, with $F(0, z) = F(1, z) = f(z)$ the initial and final position of $\mathbb{D}$.

[43]This is not a surface: near a point on the boundary of the strip the space does not look like $\mathbb{D}$, it looks like a half-disc $\{z \in \mathbb{D} : \text{ImaginaryPart}(z) \geq 0\}$. So M is actually called a *surface with boundary*.

identified with $(100, 2-t)$ (if you had identified it the "correct way" with $(100, t)$ then you would have obtained a cylinder). Unlike the cylinder whose boundary is two disjoint circles, the boundary of the Möbius strip M is just one circle.[44]

The *Klein bottle* is a famous example of a non-orientable surface. It is obtained from a square by gluing opposite parallel sides, just like the torus, except that two of those sides are glued together in the "wrong way" (just like for the Möbius strip):

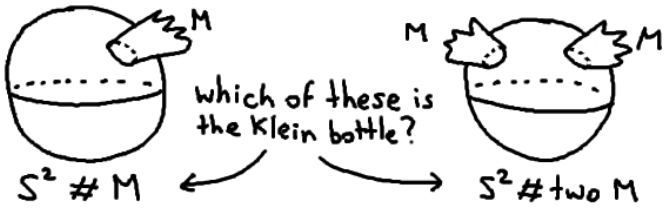


Exercise 37. *Show that the Klein bottle is non-orientable, by showing how to move a small disc around until you get back to the starting disc with reversed boundary orientation.*

The natural question to ask, therefore, is: what surfaces do you get from the sphere by "attaching Möbius strips"? Attaching M to a surface S means:

(1) remove a disc from S;
(2) attach M onto the resulting surface by gluing the boundary circle of M onto the circular boundary of the disc you removed from S.

It is difficult to draw this attachment in $\mathbb{R}^3$ (try!), so we will just draw wiggly caps denoted by M:



So, for any integer $h \geq 1$, you can build a non-orientable surface from the sphere by attaching h copies of M. This is the second family of surfaces, and again for this family, once you computed h, it determines the surface up to homeomorphism.

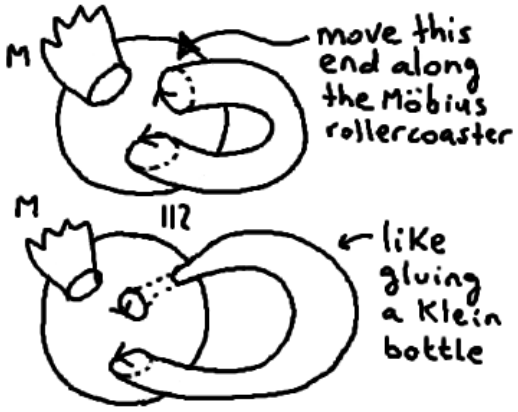
Exercise 38. *The Klein bottle is obtained from S^2 by attaching h Möbius strips: do you need $h = 1$ or $h = 2$ strips?*

It is natural to ask whether you can create a third family of surfaces by attaching both cylinders and Möbius strips to a sphere. We now show that you get nothing new.

Suppose we have a sphere with $h \geq 1$ Möbius strips attached, and let's now also attach a cylinder. We can continuously deform the surface by moving one of the two boundary

[44]Indeed, first run along $(s, 0)$ for $0 \leq s \leq 100$, then you reach the two identified points $(100, 0) \sim (0, 2)$, then run along $(s, 2)$ for $0 \leq s \leq 100$ until you reach the two identified points $(100, 2) \sim (0, 0)$, which was the starting point!

circles of the cylinder along one of the Möbius strips. As this boundary circle moves along the Möbius rollercoaster, the arrow which orients the boundary circle will flip. So we end up with a cylinder attached to the surface, however the attachment boundary circles on the sphere now have arrows pointing in the same direction (recall that the two circles on the sphere where cylinders get attached are supposed to have opposite arrow orientations). So how is the cylinder actually attached after the rollercoaster ride?



Exercise 39. *Check that one end of the cylinder is actually glued onto the sphere from the inside of the sphere! Deduce that the cylinder, attached in this "wrong way", corresponds to having a Klein bottle attached to the sphere. Deduce that attaching cylinders to a sphere with $h \geq 1$ Möbius strips attached gives (up to homeomorphism) a sphere with some other number $h' \geq 1$ of Möbius strips attached.*

5.4. **The classification of surfaces.**

Theorem 40 (Classification of Surfaces)**.**
Any compact connected surface is homeomorphic to precisely one of:

(1) *a sphere;*
(2) *a sphere with $g \geq 1$ cylinders attached;*
(3) *a sphere with $h \geq 1$ Möbius strips attached.*

A beautiful simple proof of this theorem, is explained in Zeeman's notes, "An introduction to topology" (google it), which I highly recommend (beautifully written, beautifully illustrated, and a pleasure to read).

6. Acknowledgements

Many thanks to the CMI-PROMYS scholars who attended the 2014 Oxford Masterclasses in Wadham College, and who contributed to shape these notes with their many insightful comments and questions.